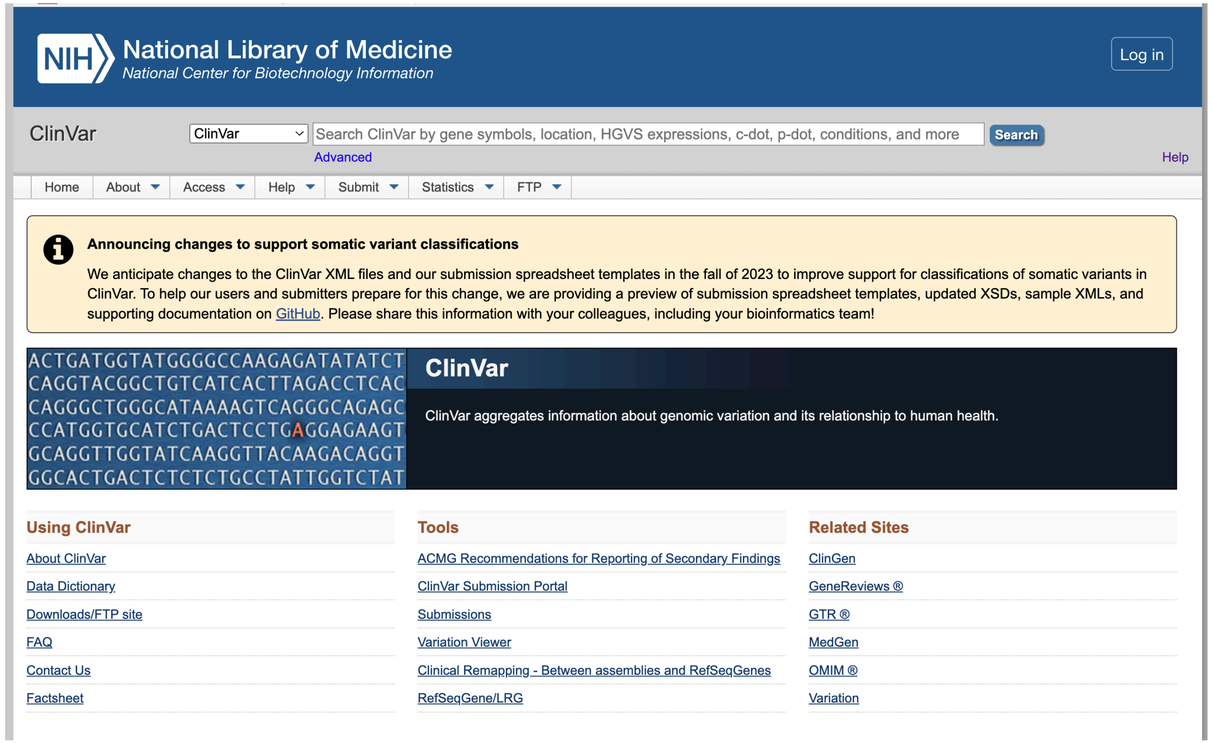
BIO392: Variants & diseases

Exploring ClinGen and ClinVar resources to find out relationships between genetic diseases and genes/ variants implicated.

Hangjia Zhao
hangjia.zhao@uzh.ch
2025.04.11

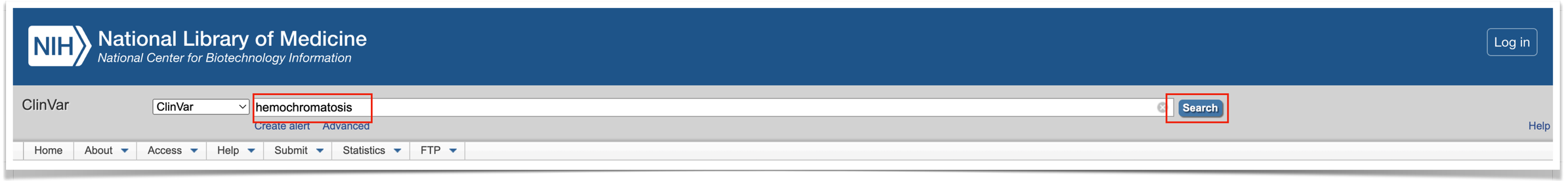
ClinVar

A database of genomic variants and the interpretation of their relevance to disease



Search in different ways:

- gene symbols, e.g. PTEN
- gene symbol and c. or p., e.g., mutyh c.1103g>a
- location / chromosome coordinates, e.g., chr1:11102837-11267747
- HGVS expressions, e.g. NM_005228.5:c.2573T>G
- protein changes, e.g. G132V
- SNP identifier, e.g. rs56045941
- diseases, e.g. cystic fibrosis
- submitters, e.g. Invitae
- a ClinVar accession number (VCV, RCV, or SCV)



Clinical significance clear
- Conflicting interpretations (0)
- Benign (0)
- Likely benign (0)
- Uncertain significance (0)
- Likely pathogenic (19)
- ✓ **Pathogenic** (28)

Molecular consequence
- Frameshift (6)
- Missense (14)
- Nonsense (5)
- Splice site (1)
- ncRNA (1)
- Near gene (0)
- UTR (4)

Variation type
- Deletion (5)
- Duplication (1)
- Indel (0)
- Insertion (1)
- Single nucleotide (21)

Variation size
- Short variant (< 50 bps) (28)
- Structural variant (>= 50 bps) (0)

Variant length
- < 1kb, single gene (26)
- > 1kb, single gene (0)
- > 1kb, multiple genes (0)

Review status clear
- Practice guideline (0)
- Expert panel (0)
- ✓ **Multiple submitters** (28)
- Single submitter (0)
- At least one star (28)
- Conflicting interpretations (0)

Clear all

Show additional filters

Search results

Display options | Sort by Relevance | Download

Items: 28

Filters activated: Pathogenic, Multiple submitters. Clear all to show 1415 items.

The following term was not found in ClinVar: clinsig established risk allele[Properties].

	Variation *Location*	Gene(s)	Protein change	Condition(s)	Clinical significance (Last reviewed)	Review status
1.	NM_213653.4(HJV):c.187C>T (p.Arg63Ter) *GRCh37:* Chr1:145415368 *GRCh38:* Chr1:146019645	HJV	R63*	**Hemochromatosis** type 2A	Pathogenic/Likely pathogenic (Oct 19, 2021)	**criteria provided, multiple submitters, no conflicts**
2.	NM_000410.4(HFE):c.892G>T (p.Glu298Ter) *GRCh37:* Chr6:26093188 *GRCh38:* Chr6:26092960	HFE	E298*, E118*, E196*, E275*, E284*, E192*, E210*, E206*, E295*	**Hemochromatosis** type 1, Hereditary **hemochromatosis**	Pathogenic/Likely pathogenic (Apr 25, 2023)	**criteria provided, multiple submitters, no conflicts**
3.	NM_014585.6(SLC40A1):c.626C>T (p.Ser209Leu) *GRCh37:* Chr2:190430214 *GRCh38:* Chr2:189565488	SLC40A1	S209L	**Hemochromatosis** type 4	Pathogenic/Likely pathogenic (Sep 20, 2022)	**criteria provided, multiple submitters, no conflicts**
4.	NM_003227.4(TFR2):c.2101C>T (p.Arg701Ter) *GRCh37:* Chr7:100224421 *GRCh38:* Chr7:100626798	LOC113687175, TFR2	R530*, R701*	Hereditary **hemochromatosis**, **Hemochromatosis** type 1	Pathogenic (Mar 10, 2022)	**criteria provided, multiple submitters, no conflicts**
5.	NM_213653.4(HJV):c.59dup (p.Ser21fs) *GRCh37:* Chr1:145414839-145414840 *GRCh38:* Chr1:146020172-146020173	HJV	S21fs	**Hemochromatosis** type 2A, not provided	Pathogenic/Likely pathogenic (Aug 20, 2022)	**criteria provided, multiple submitters, no conflicts**
6.	NM_003227.4(TFR2):c.313C>T (p.Arg105Ter) *GRCh37:* Chr7:100238469 *GRCh38:* Chr7:100640846	TFR2	R105*	Hereditary **hemochromatosis**, **Hemochromatosis** type 3	Pathogenic/Likely pathogenic (Jan 17, 2022)	**criteria provided, multiple submitters, no conflicts**
7.	NM_213653.4(HJV):c.399del (p.Ala134fs) *GRCh37:* Chr1:145415577 *GRCh38:* Chr1:146019433	HJV	A134fs, A21fs	**Hemochromatosis** type 2A, not provided	Pathogenic/Likely pathogenic (Feb 17, 2022)	**criteria provided, multiple submitters, no conflicts**
8.	NM_014585.6(SLC40A1):c.533G>A (p.Arg178Gln) *GRCh37:* Chr2:190430307	SLC40A1	R178Q	**Hemochromatosis** type 4	Pathogenic/Likely pathogenic (Sep 2, 2021)	**criteria provided, multiple submitters, no conflicts**

NM_213653.4(HJV):c.187C>T (p.Arg63Ter)

Cite | Follow | Print | Download

Germline

Classification ★★☆☆

Pathogenic/Likely pathogenic
2 out of 2 submissions contributed to this classification

Somatic

No data submitted for somatic clinical impact

Somatic

No data submitted for oncogenicity

Variant Details

Identifiers: NM_213653.4(HJV):c.187C>T (p.Arg63Ter)
Variation ID: 1327996 Accession: VCV001327996.3

Type and length: single nucleotide variant, 1 bp

Location: Cytogenetic: 1q21.1 1: 146019645 (GRCh38) [NCBI UCSC

Timeline in ClinVar:

	First in ClinVar	Last submission	La
Germline	Dec 18, 2021	Dec 31, 2022	

HGVS:

Nucleotide	Protein
NM_213653.4:c.187C>T MANE SELECT	NP_998818.1
NM_001316767.2:c.-22+53C>T	
NM_001379352.1:c.187C>T	NP_0013662

... more HGVS

Protein change: R63*
Other names: -
Canonical SPDI: NC_000001.11:146019644:G:A
Functional consequence : -
Global minor allele frequency (GMAF) : -
Allele frequency : Trans-Omics for Precision Medicine (TOPMed) 0.00000
Links : dbSNP: rs1652592444
ClinGen: CA342143559
VarSome

Submissions - Germline

Classification (Last evaluated)	Review status (Assertion criteria)	Condition	Submitter	More information
Pathogenic (Jul 01, 2021) C Contributing to aggregate classification	★☆☆☆ (ACMG Guidelines, 2015) Method: research	Hemochromatosis type 2A Affected status: yes Allele origin: germline	BloodGenetics Accession: SCV002033800.1 First in ClinVar: Dec 18, 2021 Last updated: Dec 18, 2021	Publications: PubMed (1)
Likely pathogenic (Oct 19, 2021) C Contributing to aggregate classification	★☆☆☆ (ACMG Guidelines, 2015) Method: clinical testing	Hemochromatosis type 2A Affected status: unknown Allele origin: unknown	Fulgent Genetics, Fulgent Genetics Accession: SCV002813262.1 First in ClinVar: Dec 31, 2022 Last updated: Dec 31, 2022	

Genes

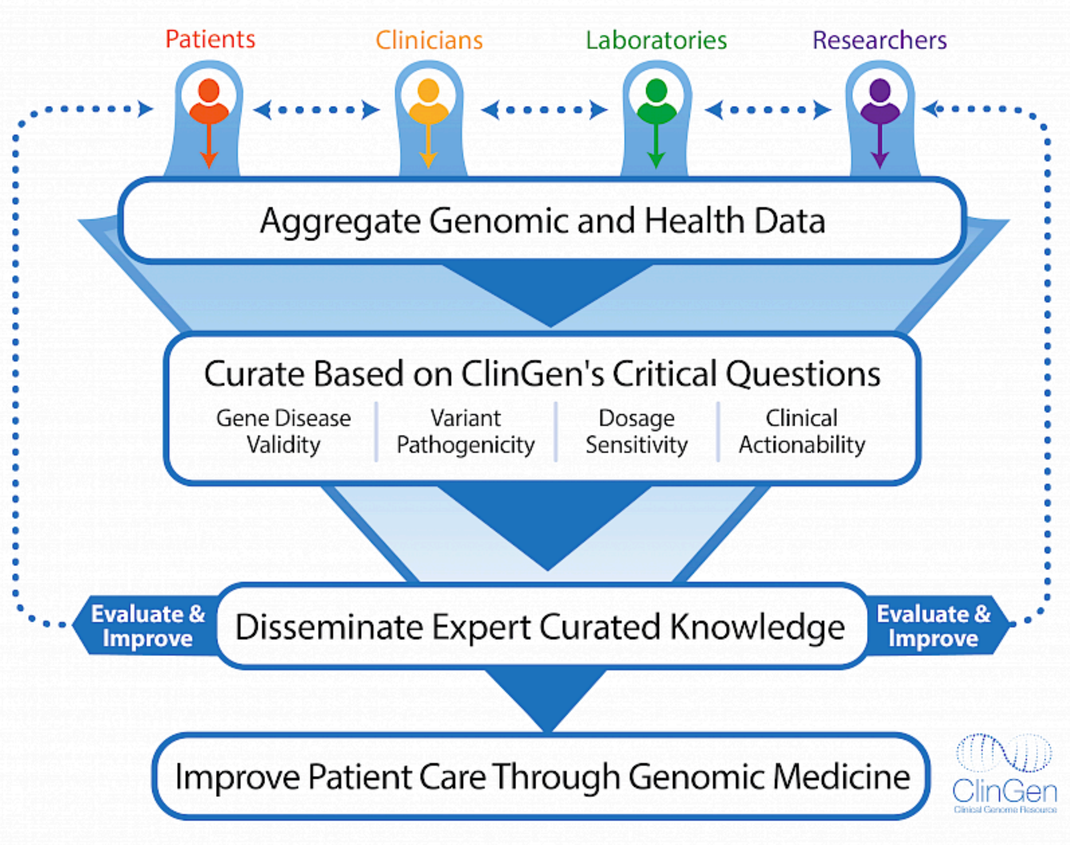
Gene	OMIM	ClinGen Gene Dosage Sensitivity Curation		Variation Viewer	Related variants	
		HI score	TS score		Within gene	All
HJV		-	-	GRCh38 GRCh37	411	603

Conditions - Germline

Condition	Classification (# of submissions)	Review status	Last evaluated	Variation/condition record
Hemochromatosis type 2A	Pathogenic/Likely pathogenic (2)	★★☆☆	Oct 19, 2021	RCV001794942.3

ClinGen

an authoritative central resource that defines the clinical relevance of genes and variants for use in precision medicine and research.



Curation activities

- Gene Disease Validity: Can variation in this gene cause disease?

Curators review genetic and experimental data in the scientific literature to identify genes in which pathogenic variants cause disease.

- Dosage Sensitivity: Does loss or gain of a copy of this gene or genomic region result in disease?

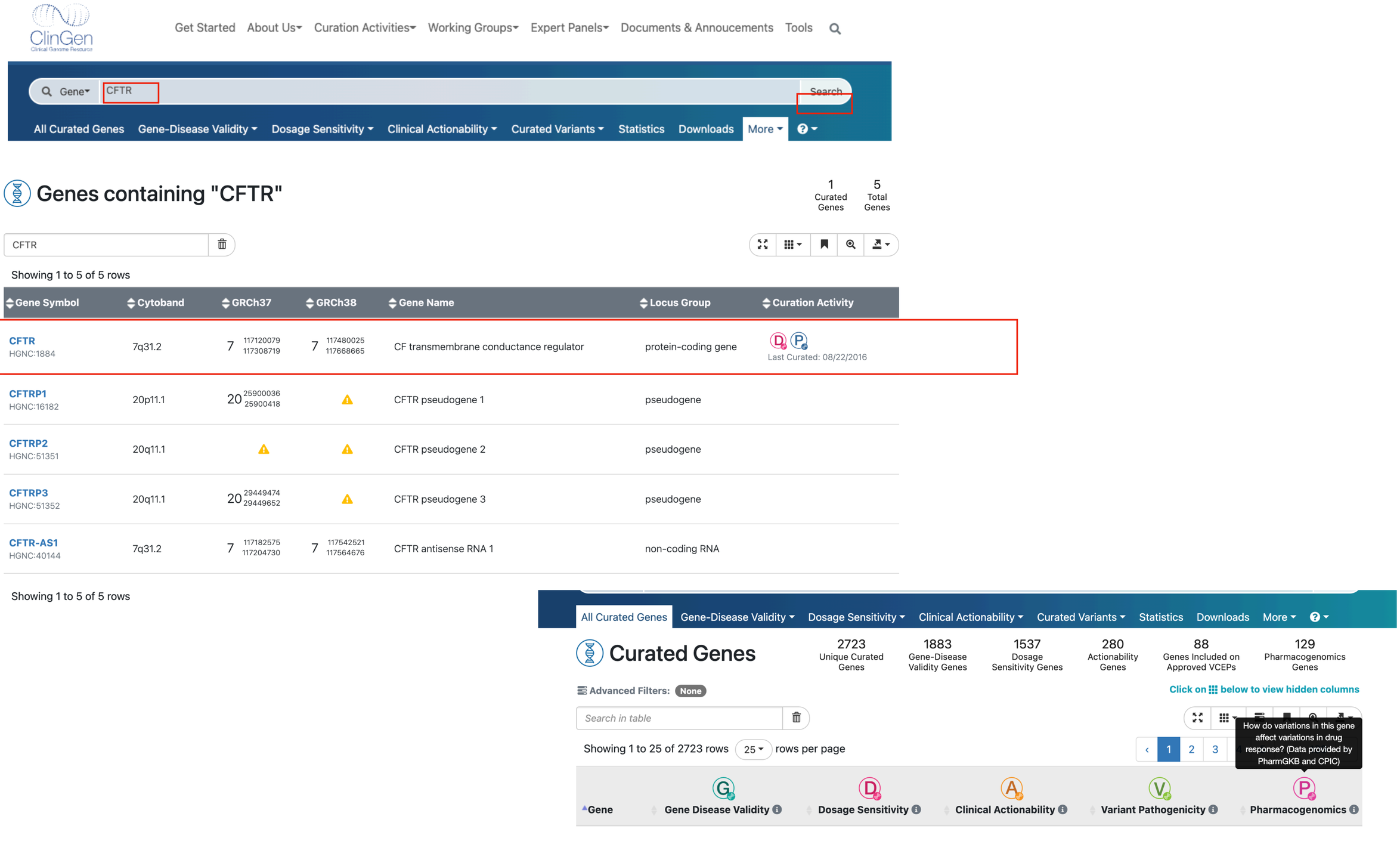
The dosage sensitivity curation process collects evidence supporting or refuting haploinsufficiency (loss) and triplosensitivity (gain) as mechanisms for disease for genes and larger genomic regions.

- Variant Pathogenicity: Which changes in this gene cause disease?

The variant curation process combines clinical, genetic, population, and functional evidence with expert review to classify variants according to ACMG/AMP guidelines.

- Clinical Actionability: Are there actions that could be taken to improve outcomes for patients with this genetic risk?

The actionability curation process evaluates availability of effective medical interventions, accounting for the chance the outcome will happen, the severity of the condition to be avoided, and the risks associated with the intervention.


ClinGen
Clinical Genome Resource
Get Started
About Us
Curation Activities
Working Groups
Expert Panels
Documents & Annoucements
Tools
Gene
CFTR
Search
All Curated Genes
Gene-Disease Validity
Dosage Sensitivity
Clinical Actionability
Curated Variants
Statistics
Downloads
More
Genes containing "CFTR"
1 Curated Genes
5 Total Genes
CFTR
Showing 1 to 5 of 5 rows
Gene Symbol
Cytoband
GRCh37
GRCh38
Gene Name
Locus Group
Curation Activity
CFTR
HGNC:1884
7q31.2
7 117120079 117308719
7 117480025 117668665
CF transmembrane conductance regulator
protein-coding gene
Last Curated: 08/22/2016
CFTRP1
HGNC:16182
20p11.1
20 25900036 25900418
CFTR pseudogene 1
pseudogene
CFTRP2
HGNC:51351
20q11.1
CFTR pseudogene 2
pseudogene
CFTRP3
HGNC:51352
20q11.1
20 29449474 29449652
CFTR pseudogene 3
pseudogene
CFTR-AS1
HGNC:40144
7q31.2
7 117182575 117204730
7 117542521 117564676
CFTR antisense RNA 1
non-coding RNA
Showing 1 to 5 of 5 rows
All Curated Genes
Gene-Disease Validity
Dosage Sensitivity
Clinical Actionability
Curated Variants
Statistics
Downloads
More
Curated Genes
2723 Unique Curated Genes
1883 Gene-Disease Validity Genes
1537 Dosage Sensitivity Genes
280 Actionability Genes
88 Genes Included on Approved VCEPs
129 Pharmacogenomics Genes
Advanced Filters: None
Click on below to view hidden columns
Search in table
Showing 1 to 25 of 2723 rows
25
rows per page
1
2
3
How do variations in this gene affect variations in drug response? (Data provided by PharmGKB and CPIC)
Gene
Gene Disease Validity
Dosage Sensitivity
Clinical Actionability
Variant Pathogenicity
Pharmacogenomics

CFTR

View Gene Facts

0	1	0	0	2 / 3	Follow Gene
Gene-Disease Validity Classifications	Dosage Sensitivity Classifications	Clinical Actionability Assertions	Variant Pathogenicity Assertions	CPIC / PharmGKB High Level Records	

Curation Summaries | Status and Future Work (0) | External Genomic Resources | ClinVar Variants

Dosage Sensitivity

Group By Activity | Group By Gene-Disease Pair

Gene	Disease	Working Group	HI Score & TS Score	Report & Date
CFTR	cystic fibrosis MONDO:0009061	Dosage Sensitivity WG	30 (Gene Associated with Autosomal Recessive Phenotype)	08/22/2016

Pharmacogenomics - CPIC

Gene	Drug	CPIC Level	Date Accessed	CPIC Clinical Guidelines
CFTR	ivacaftor	Level A	09/19/2022	Guideline
CFTR	ataluren	Level C	09/19/2022	Provisional

Pharmacogenomics - PharmGKB

Gene	Drug	Highest Level of Evidence	Last Curated	Information
CFTR	ivacaftor	Level 1A	03/24/2021	View
	ivacaftor / lumacaftor	Level 1A	03/24/2021	View
	ivacaftor / tezacaftor	Level 1A	03/24/2021	View

CFTR

View Gene Facts

0	1	0	0	2 / 3	Follow Gene
Gene-Disease Validity Classifications	Dosage Sensitivity Classifications	Clinical Actionability Assertions	Variant Pathogenicity Assertions	CPIC / PharmGKB High Level Records	

Gene Facts

External Data Attribution

HGNC Symbol	CFTR (HGNC:1884) HGNC Entrez Ensembl OMIM UCSC Uniprot GeneReviews LOVD LSDB ClinVar
HGNC Name	CF transmembrane conductance regulator
Gene type	protein-coding gene
Locus type	gene with protein product
Previous symbols	CF, ABCC7
Alias symbols	MRP7, ABC35, TNR-CFTR, dJ760C5.1, CFTR/MRP
GenCC Classifications	Limited 1 Definitive 2 Supportive 2 (Read more about GenCC Classifications)
%HI	2.33 (Read more about the DECIPHER Haploinsufficiency Index)
pLI	0 (Read more about gnomAD pLI score)
LOEUF	1.31 (Read more about gnomAD LOEUF score)
Cytoband	7q31.2
Genomic Coordinates	GRCh37/hg19: chr7:117120079-117308719 NCBI Ensembl UCSC GRCh38/hg38: chr7:117480025-117668665 NCBI Ensembl UCSC
MANE Select Transcript	NM_000492.4 ENST00000003084.11 (Read more about MANE Select)
Function	Epithelial ion channel that plays an important role in the regulation of epithelial ion and water transport and fluid homeostasis (PubMed:26823428). Mediates the transport of chloride ions across the cell membrane (PubMed:10792060, PubMed:11524016, PubMed:11707463, PubMed:12519745, PubMed:15010471, PubMed:12588899, PubMed:17036051, PubMed:19398555, PubMed:19621064, PubMed:22178883, PubMed:25330774, PubMed:1712898, PubMed:8910473, PubMed:9804160, PubMed:12529365, PubMed:17182731, PubMed:26846474,... *(Source: Uniprot)*

All Curated Genes | Gene-Disease Validity | Dosage Sensitivity | Clinical Actionability | Curated Variants | Statistics | Downloads | More

CFTR

Gene Facts

30	-5
Haplo Score	Triplo Score

Dosage Sensitivity Summary (Gene)

Dosage ID:	ISCA-30165 View legacy report...
Curation Status:	Complete
Issue Type:	Dosage Curation - Gene
Haploinsufficiency:	Gene Associated with Autosomal Recessive Phenotype (30) Read full report...
Triplosensitivity:	Not Yet Evaluated Read full report...
Last Evaluated:	08/22/2016

Haploinsufficiency (HI) Score Details

HI Score:	30
HI Evidence Strength:	Gene Associated with Autosomal Recessive Phenotype (Disclaimer)
HI Disease:	cystic fibrosis Monarch

DISCLAIMER

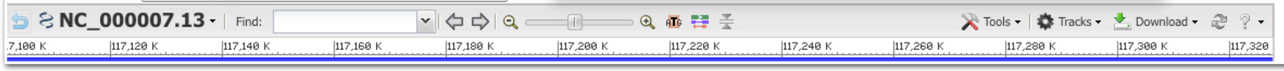
The loss of function score should be used to evaluate deletions, and the triplosensitivity score should be used to evaluated duplications. CNVs encompassing more than one gene must be evaluated in their totality (e.g. overall size, gain vs. loss, presence of other genes, etc). The rating of a single gene within the CNV should not necessarily be the only criteria by which one defines a clinical interpretation. Individual interpretations must take into account the phenotype described for the patient as well as issues of penetrance and expressivity of the disorder. ACMG has published guidelines for the characterization of postnatal CNVs, and these recommendations should be utilized (Genet Med (2011)13: 680-685). Exceptions to these interpretive correlations will occur, and clinical judgment should always be exercised.

Triplosensitivity (TS) Score Details

TS Evidence Strength:	Not Yet Evaluated (Discla

Genomic View

Select assembly: GRCh37/hg19 chr7:117120079-117308719 (NC_



Explore ClinVar

Task:

1.Learn HGVS nomenclature (https://hgvs-nomenclature.org/stable/background/simple/)

2.Create a relational list (**graded**, ddl 04.15 09:00am)

rename the template to “[name]_var_gene_disease_relation.md”

Disease	Disease description	Gene	Variants（HGVS)
Hemochromatosis	a disorder that causes the body to absorb too much iron from the diet	HJV	NM_213653.4:c.187C>T
Thalassemia			
Haemophilia			
Cystic Fibrosis			
Tay sachs disease			
Fragile X syndrome			
Huntington’s disease			

(in 2025.03 I'd like to choose NM_003227.4:c.2093_2096del in LOC113687175,TFR2 gene)

Explore ClinGen

Task: Create a relational list (**graded**, ddl 04.15 09:00am)

Gene	Gene name	Chromosomal location	Gene product	Disease	Disease description
CFTR	CF transmembrane conductance regulator	7q31.2	epithelial ion channel, transport of chloride ions across the cell membrane	Cystic fibrosis	a genetic disorder characterized by the production of sweat with a high salt content and mucus secretions with an abnormal viscosity
CYBB					
HJV					
CDKN2A					
KRAS					
TP53					
				Fragile X syndrome	a genetic disorder characterized by mild-to-moderate intellectual disability

BIO392: Introduction to BLAST

Hangjia Zhao
hangjia.zhao@uzh.ch
2025.04.11

BLAST is an algorithm used for comparing biological sequences, either amino-acids for comparing proteins or nucleotides for DNA and RNA. It can help find similar sequences.

Why do we use sequence similarity search tools

- Find the function of an unknown protein by comparing with very similar proteins
- Check the specificity of primers and probes *in-silico*
- Select data for phylogenic tree construction as well as define a related but different sequence as outgroup
- Identify host contamination in metagenomic data
- Describe the taxonomic profile of viral metagenomes

Where can we find sequence information?

Nucleic acid database

INSDC (International Nucleotide Sequence Database Collaboration)

These databases are synchronised meaning that they share the same information after synchronisation.

Protein database

Uniprot

· **Unreviewed entries:** Presents the sequence in the way it was submitted, plus automatic annotation. There are many errors or missing features, and these data are poorly updated.

· **Manually reviewed entries**: The sequence annotation has been curated by reviewers, with addition of biological knowledge. These data are updated.

Databases	Unreviewed data	Manually reviewed data
Nucleotides	INSDC (GenBank, EMBL-EBI, DDBJ)	NCBI Reference Sequences (RefSeq)
Proteins	UniprotKB/TrEMBL	UniprotKB/Swiss-Prot

Similarity & Homology

Similarity

· It refers to the "likeness" or percentage of identity between 2 sequences

· It can be quantified by calculating a shared statistically significant number of bases or amino acids

Score	Expect	Method	Identities	Positives	Gaps
107 bits(267)	3e-25	Compositional matrix adjust.	50/60(83%)	55/60(91%)	0/60(0%)

```
Query  1   MANSKEVKSFLWTQALRRELGQYCSTVKSSIIKDAQSLLHSLDFSEVSNIQRLMRKDKRN  60
           M+NSKEVKSFLWTQALRREL  YC+ VK  +IKDAQSLL+SLDFSEVSN+QRLMRKDKRN
Sbjct  1   MSNSKEVKSFLWTQALRRELSPYCTNVKLQVIKDAQSLLNSLDFSEVSNVQRLMRKDKRN  60
```

The figure above shows an alignment of two protein sequences

Amino acids represent identical amino acids between both sequences. '+' represents two amino acids with similar chemical properties.

50 identical amino acids out of 60 amino acids mean that these sequences are 83% identical.

Homology

· Most of the time, users will perform sequence searches on databases to identify genes that have an evolutionary relationship with the input sequence.

· This is **homology** : two sequences are said to be homologous if they are derived from a common ancestor. So either they are homologous or not.

· Homology usually implies similarity and cannot be quantified

Search algorithms

Exhaustive vs Heuristic Search Strategies

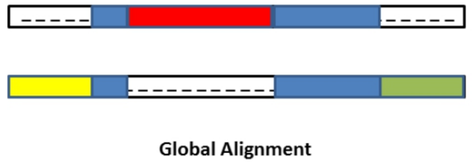
- An **exhaustive search** is a search process enumerating all possible candidates for the solution and checking whether each candidate provides a possible best match.
  - It becomes problematic since the number of comparisons required grow exponentially with the database size.
  - Such as Needleman–Wunsch algorithm (global alignment) and Smith-Waterman algorithm (local alignment)

- A **heuristic search** is to solve a problem in a faster and more efficient fashion, but not necessarily optimal for a difficult optimisation problem.
  - Such as BLAST

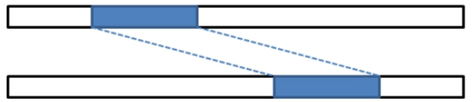
Local vs Global Alignment

- **Global alignment** algorithms considers the entire sequence, adding gaps when necessary.

- **Local alignment** algorithms find the region (or regions) of highest similarity between two sequences regardless of the other lengths of sequences. BLAST is based on local alignment.



Global Alignment



Local Alignment

Scoring system

Nucleotide

Identity matrix is used to examine the alignment between query and database hit sequence. Each nucleotide identity or mismatch corresponds to a score. The score for each nucleotide is added, resulting in the alignment raw score.

The value itself is meaningless, but allows the comparison of sequence similarity with regards to the query. Therefore the scoring system is not fixed and the user can decide the values for a match or a mismatch.

In this example Match= +1 Mismatch= -3 Gap= -3

```
CAGGTAGCAAGCTTGCATGTCA
||  ||||||||||||  |||||
CACGTAGCAAGCTTG-GTGTCA
```

	A	G	C	T
A	1	-3	-3	-3
G	-3	1	-3	-3
C	-3	-3	1	-3
T	-3	-3	-3	1

The raw score is the sum: 19 (*1) matches - 2 (*3) mismatches and -1 (*3) Gap => 19-6-3= score of 10

Scoring system

Protein

- Unlike nucleotides, mutations in proteins do not all have the same weight in term of functionality.

  For example, an alanine could be replaced by a valine without major consequence, but replacing it with a proline could be disastrous.

- An ideal scoring matrix should reflect the biological phenomena that the alignment seeks to expose.

  People use all-purpose matrices called PAM and BLOSUM

BLOSUM (**BLO**cks **SU**bstitution **M**atrix) matrix is a scoring matrix used for sequence alignment of proteins. BLOSUM matrices are used to score alignments between evolutionarily divergent protein sequences.

- BLOSUM 30, 62, 80, 100

The choice of the matrix used depends on the similarity of the proteins you are considering. To compare closely related sequences, BLOSUM matrices with higher numbers are created, e.g. BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.

BLOSUM62 is BLAST default matrix.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

Scoring system

Gap score

- Gaps indicate an absence of alignment and therefore cannot be scored in terms of similarity. Still, the presence of gaps must be considered when scoring alignments.
- The method used the most in BLAST is called affine gap-penalty. The penalty is composed of two parts: a penalty for the existence of a gap (gap open), and a further length-dependent penalty (gap extension).O+E* (L-1)

Final score of an alignment

The quality of the alignment is represented by the score, which is the sum of scores for each position, minus gap penalties. It should be noted that different matrices produce different scores.

BLAST= Basic Local Alignment Search Tool

It is a heuristic algorithm based on local alignment

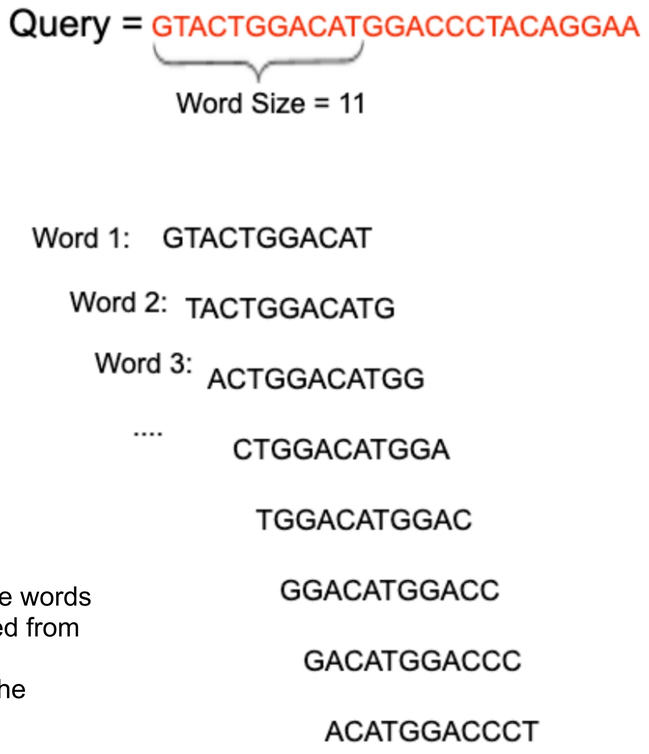
BLAST finds similar sequences by: 1) searching for matching “words” rather than individual residues. 2) using statistics to determine if a match might have occurred by chance

Steps

I. The query sequence is divided into small units, called words
II. Words are matched with database sequences
III. Pairwise alignments are created between matching and query sequences
IV. Each pairwise alignment is scored and the result is sorted on the basis of these scores

Words in BLAST

Nucleotide words



Representative words were generated from the query and compared to the database.

Protein words

The word size is adjustable

- In BLAST nucleotide, it can be reduced from the default value of 11 to a minimum of 7
- In BLAST protein, it can be reduced from the default value of 3 to a minimum of 2
- The use of short words will increase sensitivity but the task will take longer in that there are more words to compare.

Words in BLAST

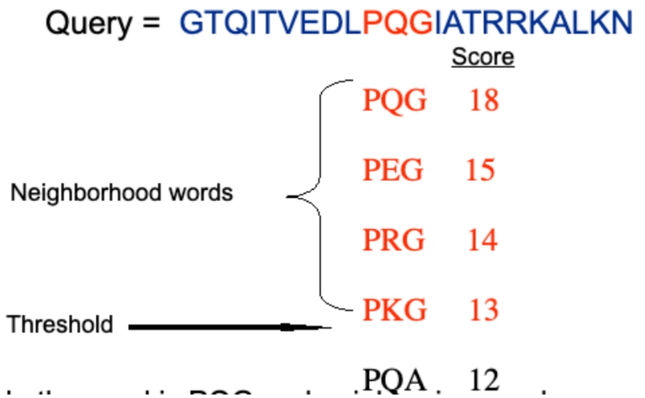
Neighborhood words

When comparing two sequences, BLAST searches for exact word matches called word *Hits*. Some alignments do not contain identical words.The neighborhood of a word contains the word itself and all the words whose score is significant when compared to a scoring matrix.

Minimum requirements for a Hit

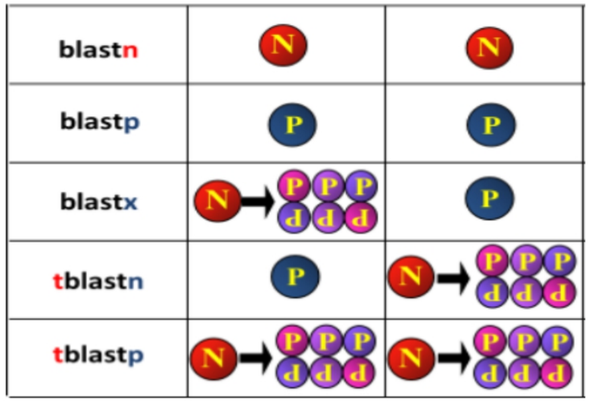
Nucleotide BLAST requires one exact word match

Protein BLAST requires two neighboring matches within 40 residues



PKG is a neighboring word, PQA is not.

Type of BLAST



BLASTn: search nucleotide sequences against nucleotide data
Used to find nucleotide similar sequences

BLASTp: search protein sequences against protein data
Find similar protein sequences and information about protein function

BLASTx: search nucleotide sequence to protein data
Used to identify coding regions in a nucleotide sequence

tBLASTn: search protein sequence to nucleotide data
Used to compare a protein sequence on nucleotide data, to find similar proteins even if they have not been annotated

tBLASTx: search translated nucleotide to translated nucleotide data
Used as a gene prediction tool used for unannotated sets of genomes

BLASTn and BLASTp are the most widely used

BLAST output interpretation

Expected value (E-value)

- Some amino acids are more common than others and so similarity among them can occur just by statistical chance. The significance of an alignment is given by the **Expected value (E-value).**

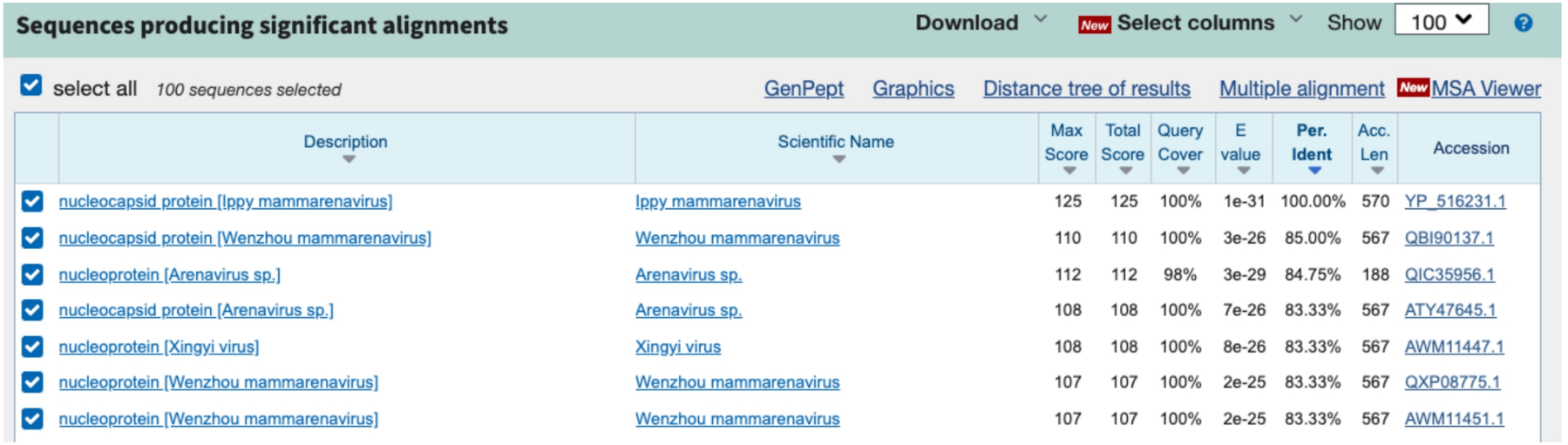
- The definition of the E – value is: the number of expected hits of similar quality (score) that could be found just by chance.

  e.g. E-value of 10 means that up to 10 hits can be expected to be found just by chance, given the same size of a random database.

- The typical threshold for a good E-value from a BLAST search is 10^{-5} or lower.

- Database size is taken into account during the E-value calculation. The same search done at different times may therefore give two E-values, if the size of the database has changed between the two searches.

BLAST output interpretation



Sequences producing significant alignments Download New Select columns Show 100

select all *100 sequences selected* GenPept Graphics Distance tree of results Multiple alignment New MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
nucleocapsid protein [Ippy mammarenavirus]	Ippy mammarenavirus	125	125	100%	1e-31	100.00%	570	YP_516231.1
nucleocapsid protein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	110	110	100%	3e-26	85.00%	567	QBI90137.1
nucleoprotein [Arenavirus sp.]	Arenavirus sp.	112	112	98%	3e-29	84.75%	188	QIC35956.1
nucleocapsid protein [Arenavirus sp.]	Arenavirus sp.	108	108	100%	7e-26	83.33%	567	ATY47645.1
nucleoprotein [Xingyi virus]	Xingyi virus	108	108	100%	8e-26	83.33%	567	AWM11447.1
nucleoprotein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	107	107	100%	2e-25	83.33%	567	QXP08775.1
nucleoprotein [Wenzhou mammarenavirus]	Wenzhou mammarenavirus	107	107	100%	2e-25	83.33%	567	AWM11451.1

Keep an eye on query coverage. A partial similarity may score better than a true protein homolog. Therefore:

- Do not trust the first hit alone.
- Be careful of homology between pathogens and host. For example, viruses and their host are very different organisms and often a protein can have acquired a very different function when moving from one to another.

Question 1

Where can you find nucleotide sequences?

i. NCBI GenBank

ii. Uniprot

Question 2

If you want to find local regions with the highest level of similarity between sequences, which alignment strategy is preferred?

i. Local alignment

ii. Global alignment

Question 3

What is the default matrix chosen by BLAST?

i. BLOSUM-80
ii. BLOSUM-62
iii. BLOSUM-45

Question 4

What does the BLAST algorithm search for?

I. Individual nucleotides/amino acids
II. Words

Question 5

The higher the E-value, the more significant the alignment?

I. Yes
II. No

Reference & Useful Links

SIB e-learning resource (https://viralzone.expasy.org/e_learning/alignments/1/start.html)

Blast in NCBI tutorial (https://www.youtube.com/watch?v=RzC-V67z5LA) 2:35-5:24

Blast in Uniprot tutorial (https://www.youtube.com/watch?v=ethkJ2_zCBY) 10:03-12:43

Exercise

1. Use blast in NCBI to search the unknown nucleotide sequence
    - Which organism does this sequence belong to?
    - Pick one blast result. What is the accession number, max score, query cover and E value?
    - Which region does this sequence cover the subject sequence? (The answer could be different which depends on the accession that you choose)
    - Is it DNA or RNA sequence?
    - Does it encode a (part of) protein? If yes, which protein? (Hint: use different blast type)

Exercise

2. Use blast in Uniprot to search the unknown protein sequence

- Select the most possible one among manually reviewed entries. What is its Uniprot ID?
- What protein does this sequence come from?
- Which organism does this sequence belong to?
- What is the function of this protein?
- What is the variant associated with acute myeloid leukemia (AML) in this protein?

3. If you have more time, play around to feel the difference of blast service from different databases

For example,

- Use Blast in NCBI to query the protein sequence
- Use Blast in Uniprot to query the nucleotide sequence