Variant Annotation

Overview of variant databases

Presenter: Nguyễn Lê Đức Minh

Every genome contains many rare, potentially functional variants

- ~ 2 millions high quality variants in a variant call file (vcf)
- ~
- ~
- ~
- ~
- ~
- 1
- Unknown number of sequencing errors

In Mendelian disease analysis, how can we identify the pathogenic genetic variant(s) in the sea of benign variation

The power of over 8 billion people

Given known mutation rates, it is almost certain that **every possible single base change compatible with life exists in a living human**

Overall review of different variant classifications and consequences

Based on 

https://asia.ensembl.org/index.html

Ensembl Variation - Variant classification (1)

Sequence variants

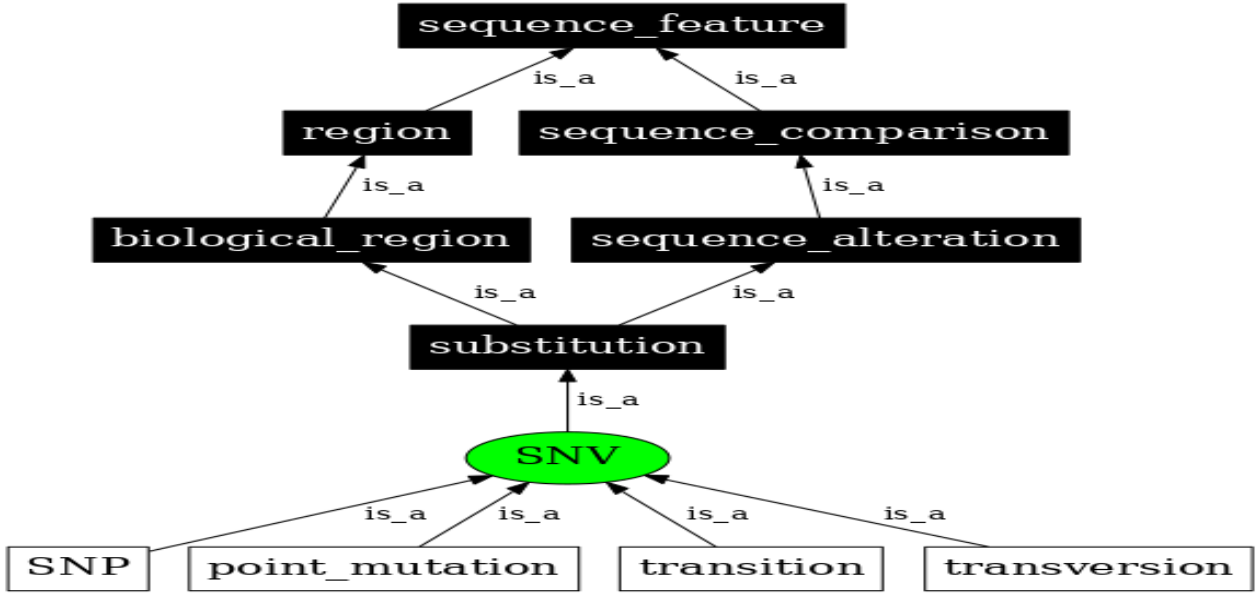
Type	Description	Example (Reference / Alternative)	
SNP	Single Nucleotide Polymorphism	Ref: ...TTG**A**CGTA...	Alt: ...TTG**G**CGTA...
Insertion	Insertion of one or several nucleotides	Ref: ...TTGACGTA...	Alt: ...TTGA**TG**CGTA...
Deletion	Deletion of one or several nucleotides	Ref: ...TTG**AC**GTA...	Alt: ...TTGGTA...
Indel	An insertion and a deletion, affecting 2 or more nucleotides	Ref: ...TTG**A**CGTA...	Alt: ...TTG**GCT**CGTA...
Substitution	A sequence alteration where the length of the change in the variant is the same as that of the reference.	Ref: ...TTG**AC**GTA...	Alt: ...TTG**TA**GTA...

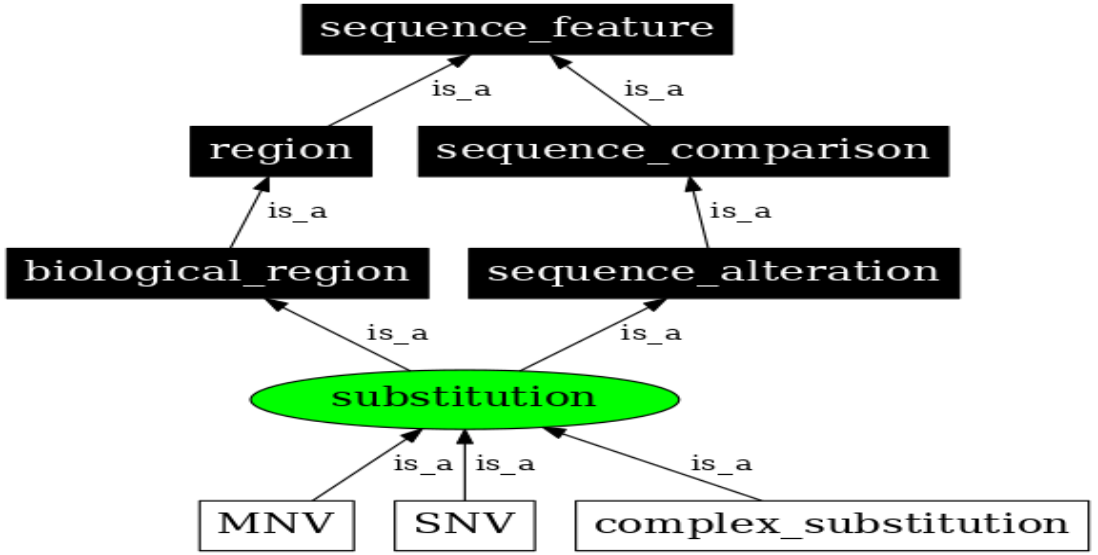
Structural variants

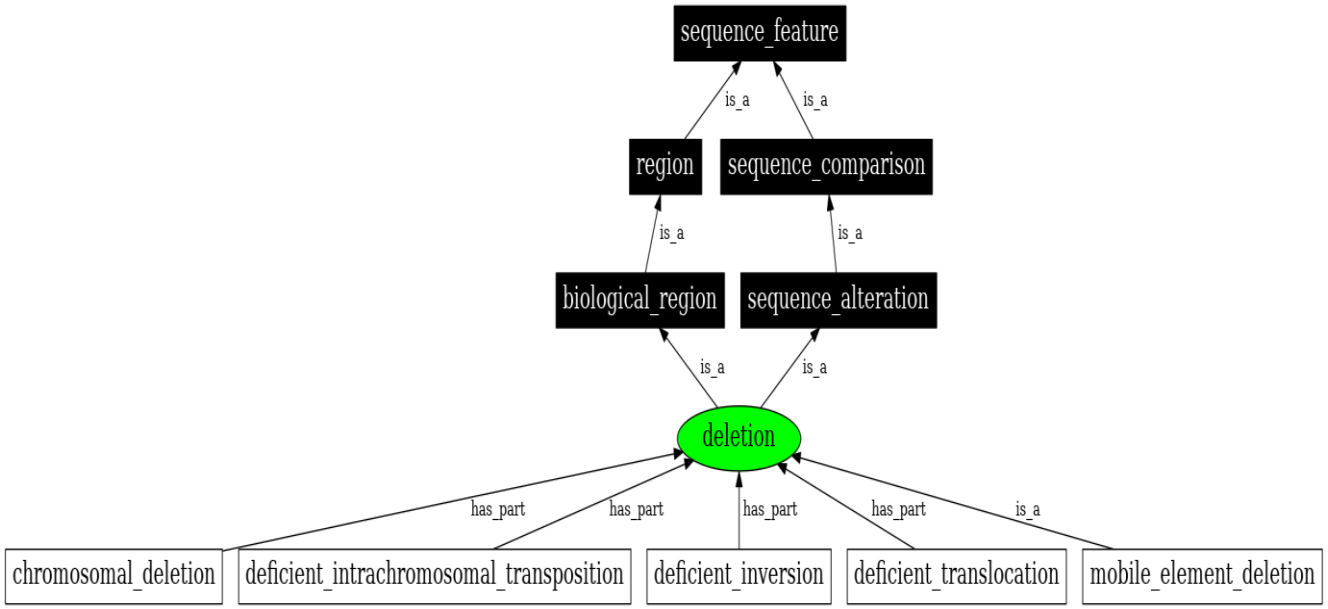
Type	Description	Example (Reference / Alternative)	
CNV	Copy Number Variation: increases or decreases the copy number of a given region	Reference: 	"Gain" of one copy:  "Loss" of one copy: 
Inversion	A continuous nucleotide sequence is inverted in the same position	Reference: 	Alternative: 
Translocation	A region of nucleotide sequence that has translocated to a new position	Reference: 	Alternative: 

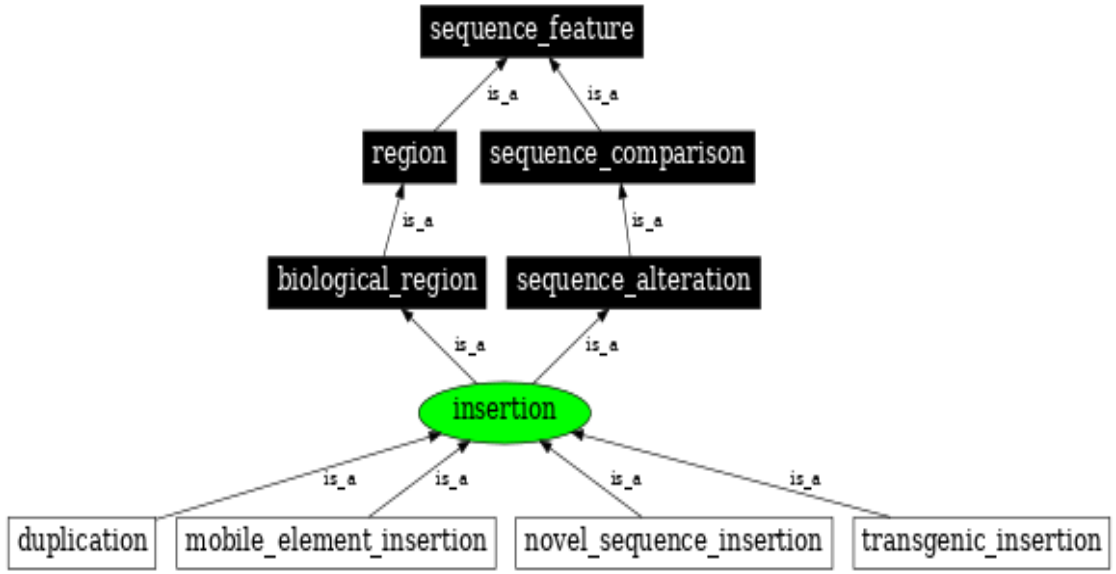
https://asia.ensembl.org/info/genome/variation/prediction/classification.html

Ensembl Variation - Variant classification (2)



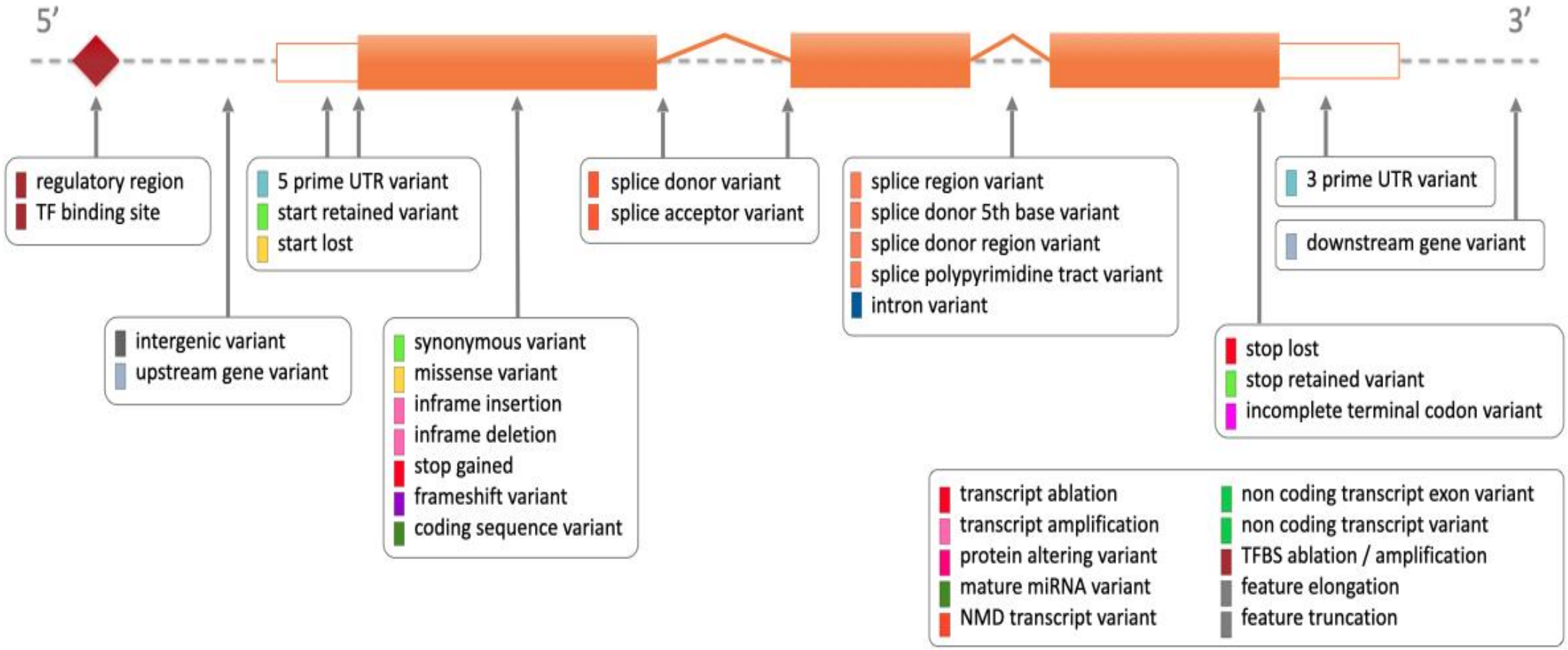






http://www.sequenceontology.org/

Ensembl Variation - Calculated variant consequences (1)



https://asia.ensembl.org/info/genome/variation/prediction/predicted_data.html

Ensembl Variation - Calculated variant consequences (2)

* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
splice_donor_5th_base_variant	A sequence variant that causes a change at the 5th base pair after the start of the intron in the orientation of the transcript	SO:0001787	Splice donor 5th base variant	LOW
splice_donor_region_variant	A sequence variant that falls in the region between the 3rd and 6th base after splice junction (5' end of intron)	SO:0002170	Splice donor region variant	LOW
splice_polypyrimidine_tract_variant	A sequence variant that falls in the polypyrimidine tract at 3' end of intron between 17 and 3 bases from the end (acceptor -3 to acceptor -17)	SO:0002169	Splice polypyrimidine tract variant	LOW
incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW
synonymous_variant	A sequence variant where there is no resulting change to the encoded amino acid	SO:0001819	Synonymous variant	LOW

https://asia.ensembl.org/info/genome/variation/prediction/predicted_data.html

Ensembl Variation - Data access

Ensembl Variation - Data access

Ensembl variation data can be accessed through:

- The website
- BioMart
- The FTP site
- The Perl API
- The REST API
- The MySQL database

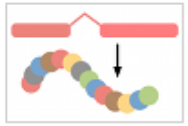
Website

Variant data can be viewed in the browser through pages such as:

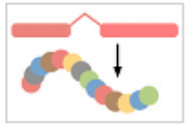
- **Gene:** Variation Table and Variation Image (for all variants in a gene) e.g. for all variants in KCNE2. Structural Variation to see all structural variants overlapping the gene.
- **Transcript:** Population comparison, Comparison image (for comparing variants in a transcript across different individual or strain sequences) e.g. compare Tmco4 in different mouse strains
- **Transcript:** Sequence, protein: list of the coding variants in protein coordinates.
- **Location:** Region in Detail (Variants can be drawn using "Configure this page" at the left. The menu allows display of information in Ensembl databases along with external sources in DAS format such as DGV loci.)
- **Phenotype:** A karyotype view to display the variants associated with a certain phenotype, e.g. phenotype "Glaucoma".

Examples:

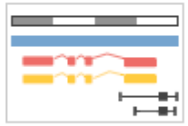
Gene
Variation Table



Transcript
Variation Table



Transcript
Protein Variation



Location
Region in detail

Phenotype
Karyotype view

https://asia.ensembl.org/info/genome/variation/tools/data_access.html

Ensembl Variation - Pathogenicity predictions

Genome wide scores

1. **GERP** (Genomic Evolutionary Rate Profiling): identifies constrained loci in multiple sequence alignments by comparing the level of substitution observed to that expected if there was no functional constraint.
    - Positive scores represent highly-conserved positions while negative scores represent highly-variable positions.
2. **CADD** (Combined Annotation Dependent Depletion): scores the predicted deleteriousness of single nucleotide variants and insertion/deletions variants in the human genome by integrating multiple annotations including conservation and functional information into one metric

Ensembl Variation - Pathogenicity predictions

Evaluate missense variants

1. **SIFT:** predicts whether an amino acid substitution is likely to affect protein function based on sequence homology and the physico-chemical similarity between the alternate amino acids.

SIFT value	Qualitative prediction	Website display example
Less than 0.05	"Deleterious"	0.01
	"Deleterious - low confidence"	0.01
Greater than or equal to 0.05	"Tolerated"	0.8
	"Tolerated - low confidence"	0.8

2. **PolyPhen:** predicts the effect of an amino acid substitution on the structure and function of a protein using sequence homology, Pfam annotations, 3D structures from PDB where available, and a number of other databases and tools.

Polyphen value	Qualitative prediction	Website display example
greater than 0.908	"Probably Damaging"	0.95
greater than 0.446 and less than or equal to 0.908	"Possibly Damaging"	0.5
less than or equal to 0.446	"Benign"	0.25
unknown	"Unknown"	unknown

Ensembl Variation - Variant quality

Evidence status

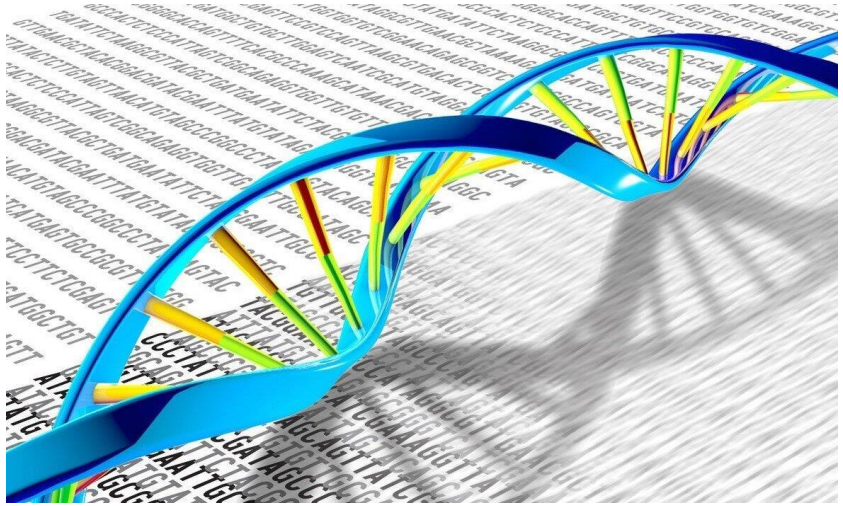
We provide a simple summary of the evidence supporting a variant as a guide to its potential reliability.

Icon	Name	Description
	Multiple observations	The variant has multiple independent dbSNP submissions, i.e. submissions with a different submitter handles or different discovery samples.
	Frequency	The variant is reported to be polymorphic in at least one sample.
	Cited	The variant is cited in a PubMed article.
	Phenotype or Disease	The variant is associated with at least one phenotype or disease.
	1000 Genomes	The variant was discovered in the 1000 Genomes Project (human only).
	gnomAD	The variant was discovered in the genome Aggregation Database (human only).
	TOPMed	The variant was discovered in the Trans-Omics for Precision Medicine program (human only).

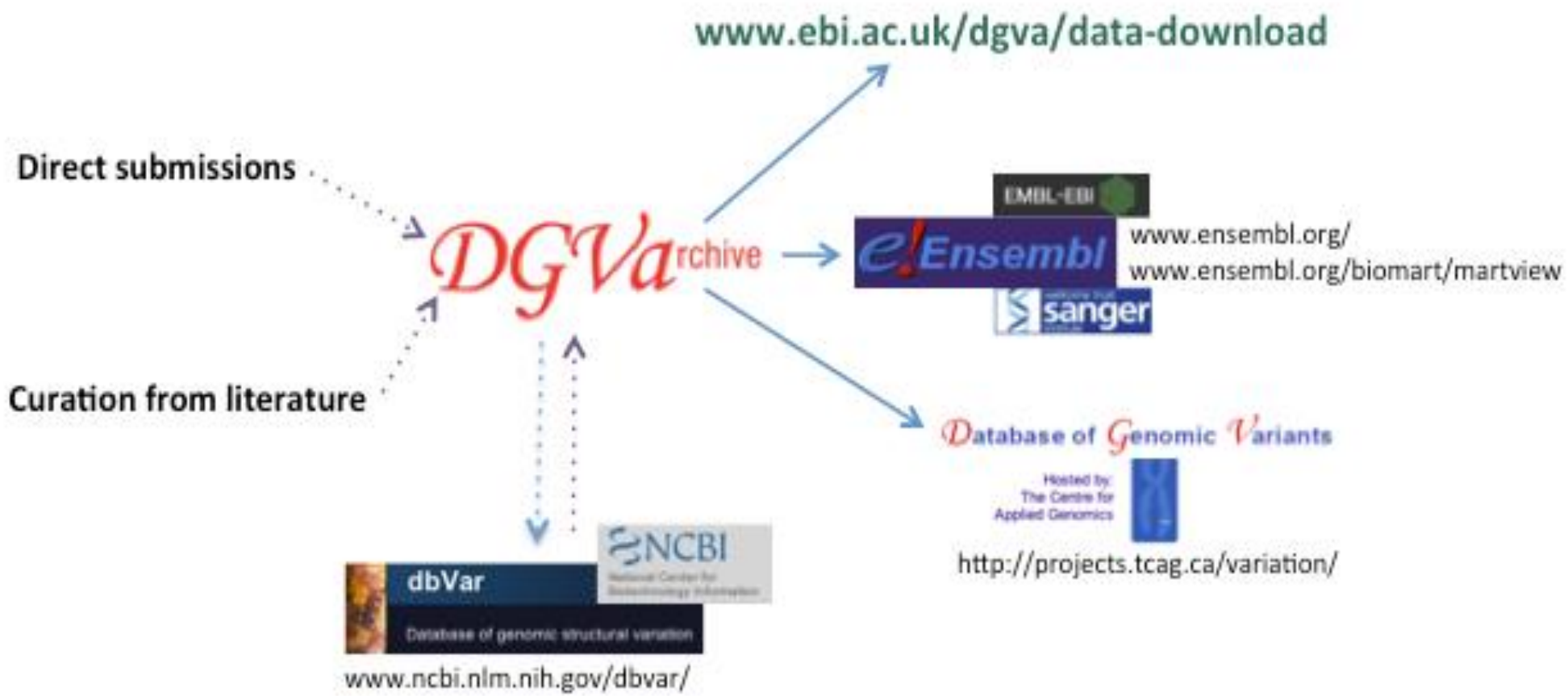
QC Type	Reported failure reason
Mapping checks	Variant does not map to the genome
	Variant maps to more than 1 location
	Mapped position is not compatible with reported alleles
	None of the variant alleles match the reference allele
Checks on the alleles of refSNPs	Loci with no observed variant alleles in dbSNP
	Alleles contain ambiguity codes
	Alleles contain non-nucleotide characters
Checks on the alleles in dbSNP submissions	Additional submitted allele data from dbSNP does not agree with the dbSNP refSNP alleles
External failure classification	Flagged as suspect by dbSNP
New assembly	Variant can not be re-mapped to the current assembly

https://asia.ensembl.org/info/genome/variation/prediction/variant_quality.html

Overview of variant databases



The Database of Genomic Variants archive (DGVa)



Frequency of variants databases

Whole-geome data	Whole-exome data	Isolated or less represented populations
1000g2015aug	Exac03	ajews
Kaviar_20150923	Esp6500siv2	TMC-SNPDB
Hrcr1	Gnomad_exome	gme
Cg69		
Gnomad_genome		

https://annovar.openbioinformatics.org/en/latest/user-guide/filter/

https://www.internationalgenome.org/

Functional prediction of variants databases

Whole-geome data	Whole-exome data	Splice variants
Gerp++	dbnsfp	dbscsnv
Cadd		Spidex
Dann		
Fathmm		
Eigen		
Gwava		

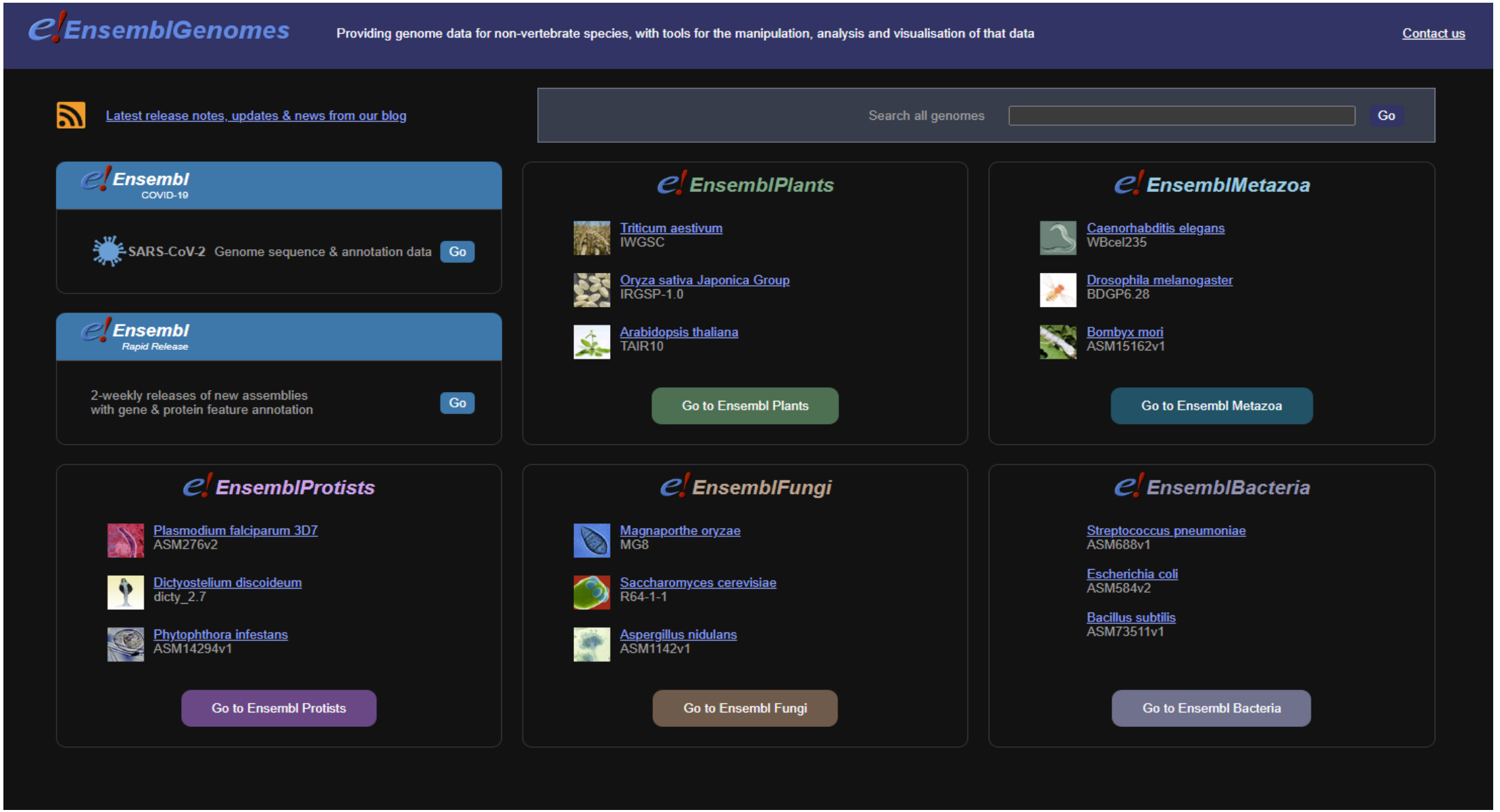
https://annovar.openbioinformatics.org/en/latest/

Disease specific variants related to clinical

Disease specifc variants	Variant identifiers
ClinVar	dbSNP
Cosmic	Avsnp (an abbreviated version of dbSNP with left-normalization by ANNOVAR)
ICGC (International Cancer Genome Consortium)	
NCI (human tumor cell line panel exome sequencing AF)	

https://annovar.openbioinformatics.org/en/latest/

Ensembl database



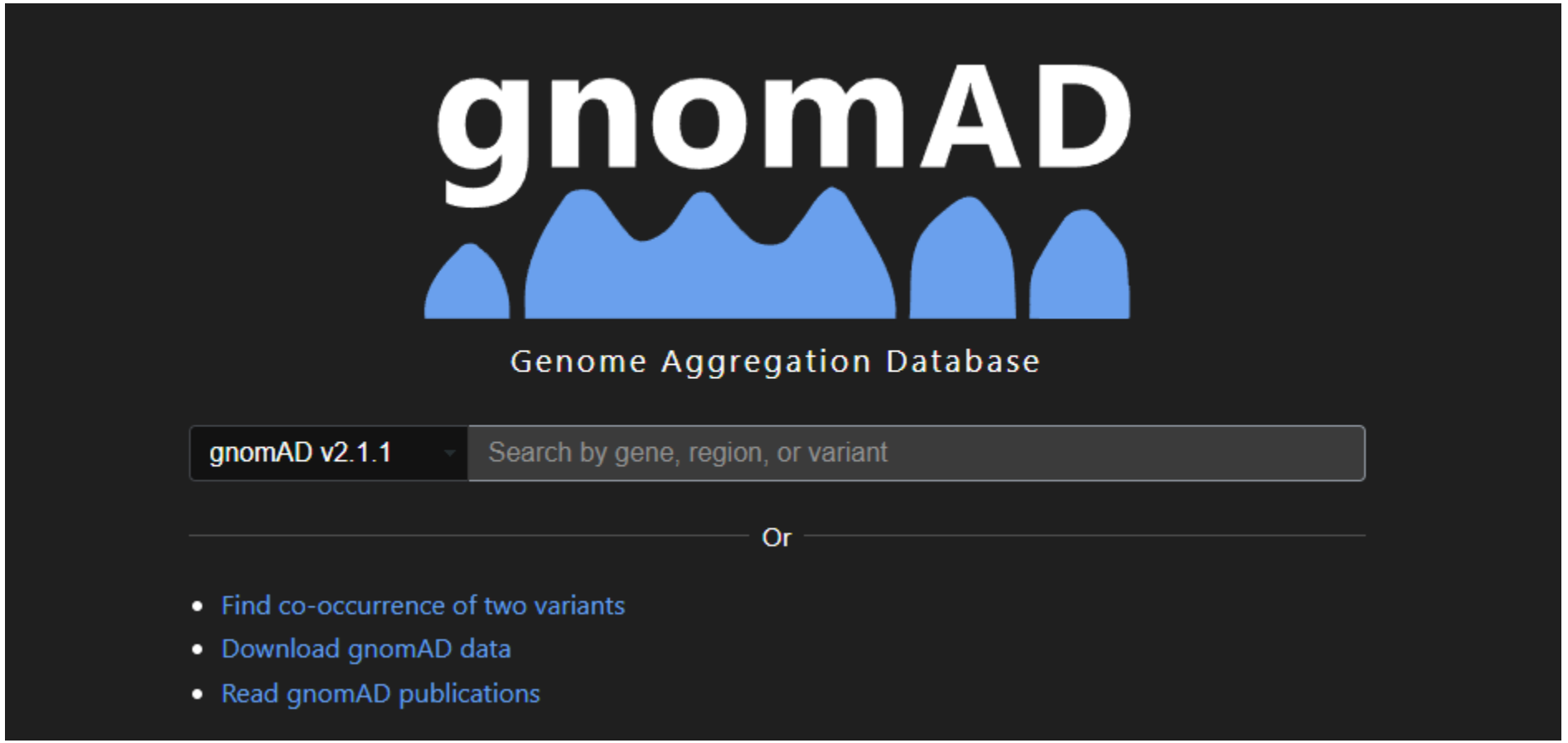
http://asia.ensembl.org/info/genome/variation/tools/data_access.html

Ensembl plays a crucial role in advancing genomics research

- Ensembl is a comprehensive and widely used biological database that provides a centralized resource for genome annotation and analysis.
- It was created to facilitate the understanding and exploration of genomic data from a wide range of organisms, including humans, animals, plants, and microorganisms.
- Ensembl integrates data from multiple sources, including genome sequencing projects, gene expression studies, and protein databases, to offer a unified and up-to-date view of the genome and its associated biological features.

Genome Aggregation Database (gnomAD)

The gnomAD database is composed of exome and genome sequences from around the world.



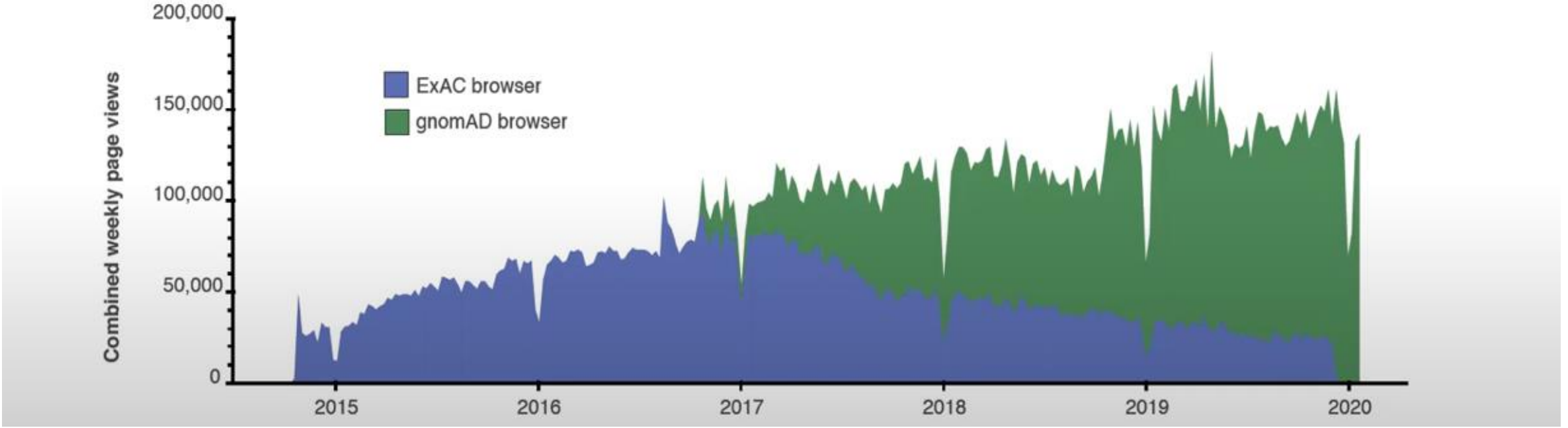
https://gnomad.broadinstitute.org/

Variant aggregation at Broad

Exome Aggregation Consortium (ExAC) – v1
60,076 exomes
Genome Build 37
released October 2014

Genome Aggregation Database (gnomAD) – v2
125,748 exomes + 15,708 genomes
Genome Build 37
released October 2016

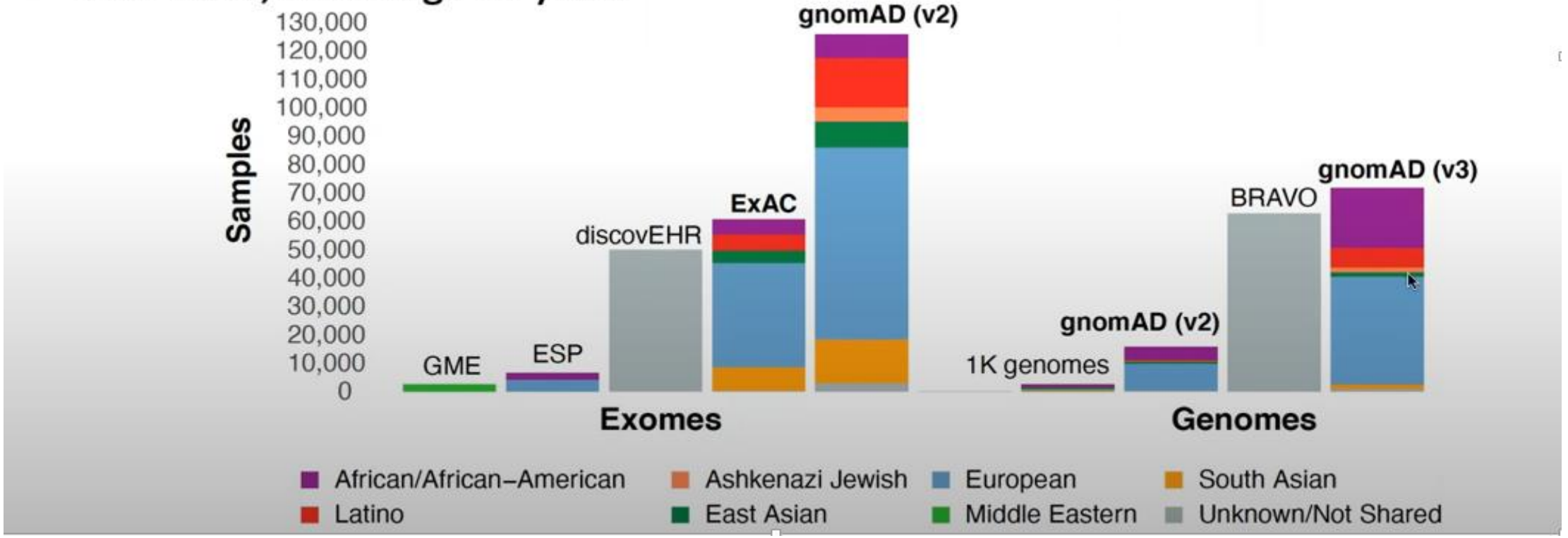
Genome Aggregation Database (gnomAD) – v3
71,702 genomes
Genome Build 38
released October 2019



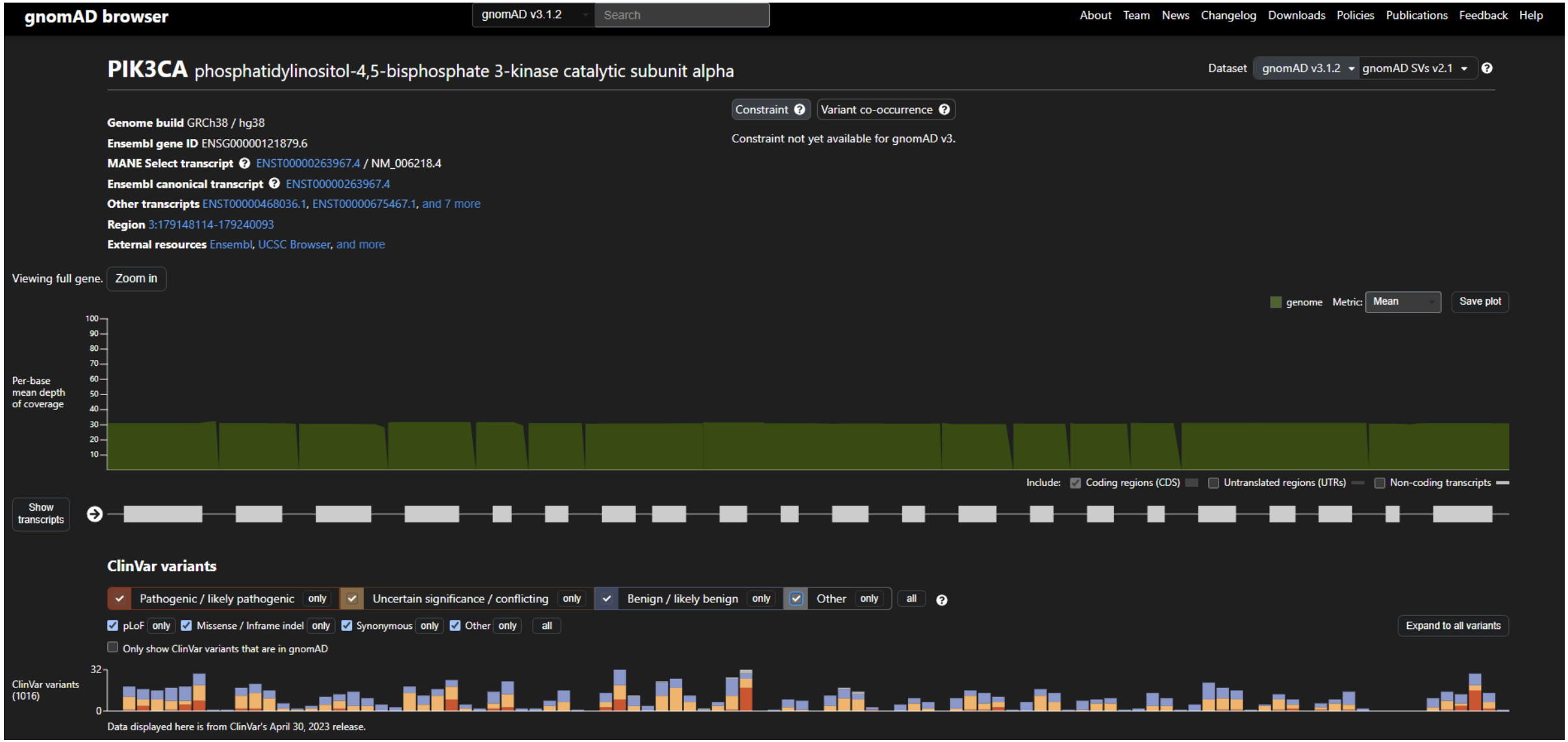
- 20+ M pageviews from 184 countries
- Aided in the diagnosis of over 200,000 patients with rare disease

Who's in the gnomAD database ?

- Case-control studies of complex adult-onset diseases (e.g. type 2 diabetes, heart attack, migraine, bipolar)
- Depleted as much as possible of people **known** to have severe pediatric disease, their first-degree relatives
- 55% Male, Mean age 54 years



gnomAD browser



https://gnomad.broadinstitute.org/gene/ENSG00000121879?dataset=gnomad_r3

The Human Gene Mutation Database (HGMD)

The Human Gene Mutation Database (HGMD) represents an attempt to collate known (published) gene lesions responsible for human inherited disease.

Commercial partnership with Celera Genomics | 2000-2005
HGMD data made available as part of the Celera Discovery System.

Commercial partnership with BIOBASE GmbH | 2006-2015
HGMD Professional stand-alone web application plus local installation.

Commercial partnership with QIAGEN Bioinformatics

2016-present

HGMD Professional, data download plus integration into IVA and QCI.

Human Gene Mutation Database

https://www.hgmd.cf.ac.uk/ac/index.php

HGMD variant classes

DM = Pathological mutation reported to be disease-causing in the corresponding literature report (majority of HGMD data).

DM? = Likely pathological mutation reported to be disease-causing in the corresponding report, but where the author has indicated that there may be some degree of doubt, or subsequent evidence has come to light in the literature, calling the deleterious nature of the variant into question.

DP = A polymorphism reported to be in significant association with a disease/phenotype ($p<0.05$) that is assumed to be functional (e.g. as a consequence of location, evolutionary conservation, replication studies etc), although there may as yet be no direct evidence (e.g. from an expression study) of a functional effect.

DFP = A polymorphism reported to be in significant association with disease ($p<0.05$) that has evidence of being of direct functional importance (e.g. as a consequence of altered gene expression, mRNA studies etc).

FP = A polymorphism reported to affect the structure, function or expression of the gene (or gene product), but with no disease association reported as yet.

R = A variant entry retired from HGMD due to being found to have been erroneously included *ab initio*, or a variant that has been subjected to correction in the literature resulting in the record becoming obsolete, merged or otherwise invalid.

Important role of HMGD

Hum Genet. 2020; 139(10): 1197–1207.
Published online 2020 Jun 28. doi: 10.1007/s00439-020-02199-3
PMCID: PMC7497289
PMID: 32596782

The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting

Peter D. Stenson,[✉1] Matthew Mort,[1] Edward V. Ball,[1] Molly Chapman,[1] Katy Evans,[1] Luisa Azevedo,[1,2] Matthew Hayden,[1] Sally Heywood,[1] David S. Millar,[1] Andrew D. Phillips,[1] and David N. Cooper[1]

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7497289/

A controversial topic

HUMAN EXPERTISE vs. ARTIFICIAL INTELLIGENCE

Can AI replace human curation of genetic variants?

Together with deep learning (DL) and machine learning (ML), AI is currently a buzzword across almost all scientific disciplines and has the potential to revolutionize diagnostic approaches in inherited diseases. But when it comes to variant curation, is AI capable of replacing human expertise? In a new study, Stanford University compares data quality from their Automatic VAriant evidence DAtabase (AVADA) to the Human Gene Mutation Database (HGMD).

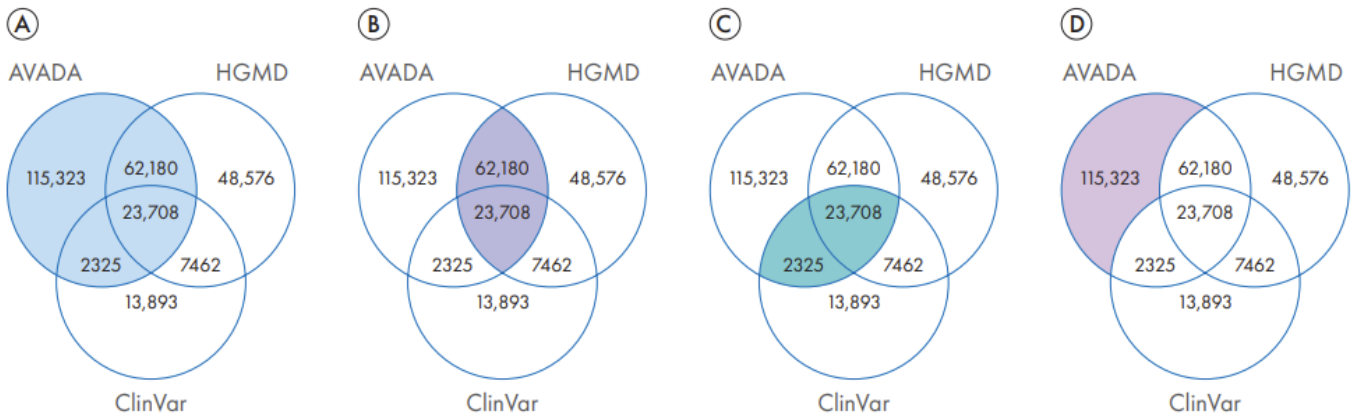


Figure 1. A) 203,536 total variants were automatically extracted by AVADA; B) 61% of variants reported as disease-causing in the expert-curated HGMD were found by AVADA; C) 55% of variants reported as likely/pathogenic in the ClinVar were also found in AVADA; D) 56% of variants retrieved by ADAVA were neither in HGMD nor in ClinVar. These variants are of questionable quality and clinical utility.

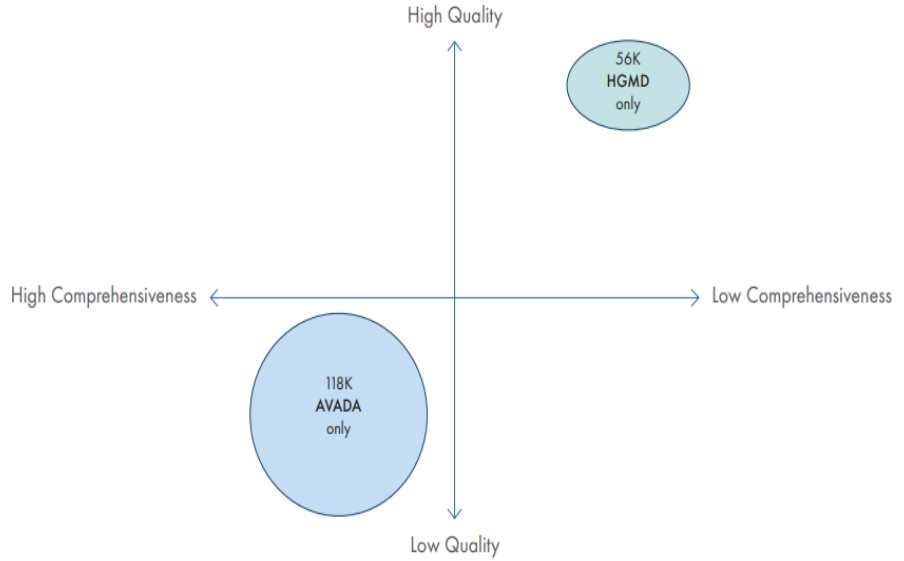
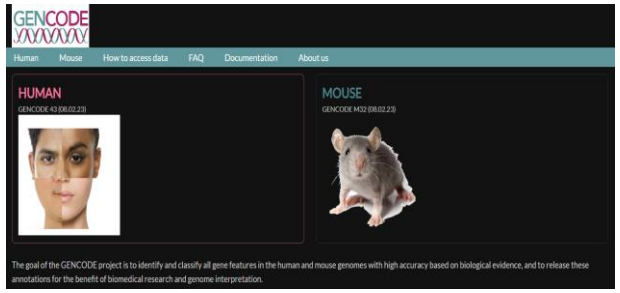


Figure 2. Manual inspection of randomly selected AVADA variants showed lower quality and comprehensiveness compared to expert-curated, high-quality HGMD data even though twice as much data was automatically extracted.

https://digitalinsights.qiagen.com/products-overview/clinical-insights-portfolio/human-gene-mutation-database/

GENCODE

- Originated as a public research consortium named ENCODE from the National Human Genome Research Institute.in September 2003.
- Its goal is to identify all functional elements in the human genome sequence
- The Wellcome Sanger Institute was awarded a grant to carry out a scale-up of the GENCODE project.



- The GENCODE gene sets are used by the entire ENCODE consortium and by many other projects (eg. Genotype-Tissue Expression (GTEx), The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), NIH Roadmap Epigenomics Mapping Consortium, Blueprint Epigenome Project, Exome Aggregation Consortium (EXAC), Genome Aggregation Database (gnomAD), 1000 Genomes Project and the Human Cell Atlas (HCA)) as reference gene sets.

https://www.gencodegenes.org/

Current GENCODE Release version 43

Human

Statistics about the current GENCODE Release (version 43)

The statistics derive from the gtf file that contains only the annotation of the main chromosomes.

For details about the calculation of these statistics please see the README_stats.txt file.

General stats

Total No of Genes	62703	Total No of Transcripts	252913
Protein-coding genes	19393	Protein-coding transcripts	89411
- readthrough genes (not included)	649	- full length protein-coding	64004
Long non-coding RNA genes	19928	- partial length protein-coding	25407
Small non-coding RNA genes	7566	Nonsense mediated decay transcripts	21354
Pseudogenes	14737	Long non-coding RNA loci transcripts	58023
- processed pseudogenes	10662		
- unprocessed pseudogenes	3570		
- unitary pseudogenes	254	Total No of distinct translations	65519
- pseudogenes	15	Genes that have more than one distinct translations	13618
Immunoglobulin/T-cell receptor gene segments			
- protein coding segments	410		
- pseudogenes	236		

https://www.gencodegenes.org/human/stats.html

GTF data format

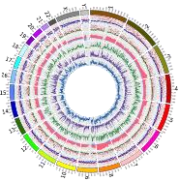
Column-number	Content	Values/format
1	chromosome name	chr{1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20,21,22,X,Y,M} or GRC accession [a]
2	annotation source	{ENSEMBL,HAVANA}
3	feature type	{gene,transcript,exon,CDS,UTR,start_codon,stop_codon,Selenocysteine}
4	genomic start location	integer-value (1-based)
5	genomic end location	integer-value
6	score(not used)	.
7	genomic strand	{+,-}
8	genomic phase (for CDS features)	{0,1,2,.}
9	additional information as key-value pairs	see below

https://www.gencodegenes.org/pages/data_format.html

GTF FTP site

- FTP site of GENCODE public databases

https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/



Catalogue of Somatic Mutations in Cancer (COSMIC)

High Precision Data, Manually Curated by Experts:

- Targeted gene-screening panels
- Over 27,000 peer reviewed papers
- Metadata (environmental factors and patient history)
- Focused on known and suspected cancer genes and mutations
- Objective frequency data as a result of mutation negative samples
- Full details of the curation process and data captured

Genome-wide Screen Data:

- Over 37,000 genomes, consisting of:
  - peer reviewed large scale genome screening data
  - other databases such as TCGA and ICGC
- Provides unbiased, genome-level profiling of diseases
- Objective frequency data, by interpreting non-mutant genes across each genome
- Can be used to discover novel driver genes

COSMIC

COSMIC v97, released 29-NOV-22

https://cancer.sanger.ac.uk/cosmic

ClinVar database

- Archive of interpretations of variants relative to conditions
- Variant-level interpretations
  - Assertion/Clinical significance/Interpretation/Classification/
- Fully public and freely available
- Submission-driven database
- Curation support from NCBI staff

https://www.ncbi.nlm.nih.gov/clinvar/

Clinvar publicity

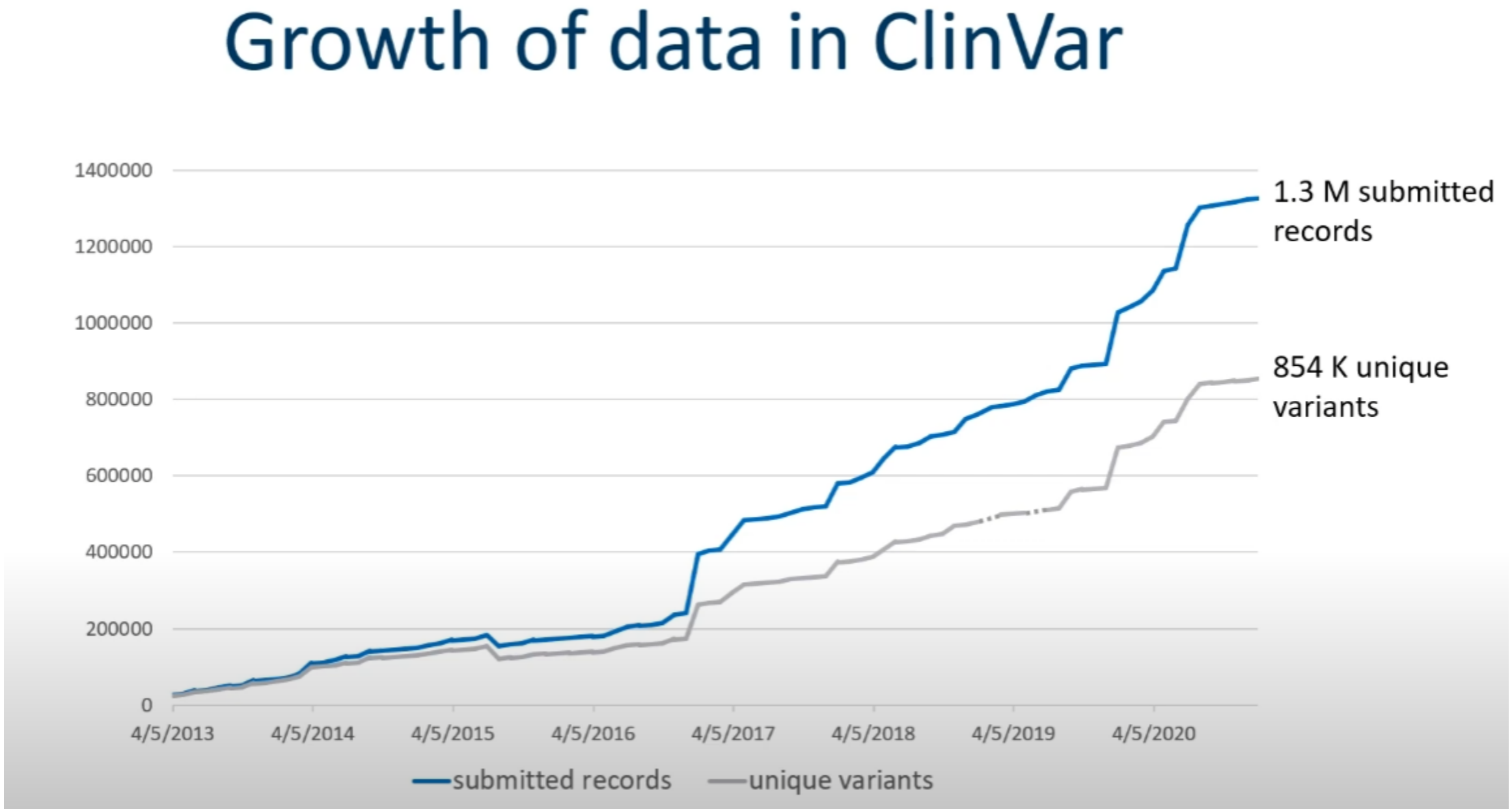
1.3 million
submitted records

Submitted by
>1800 organizations

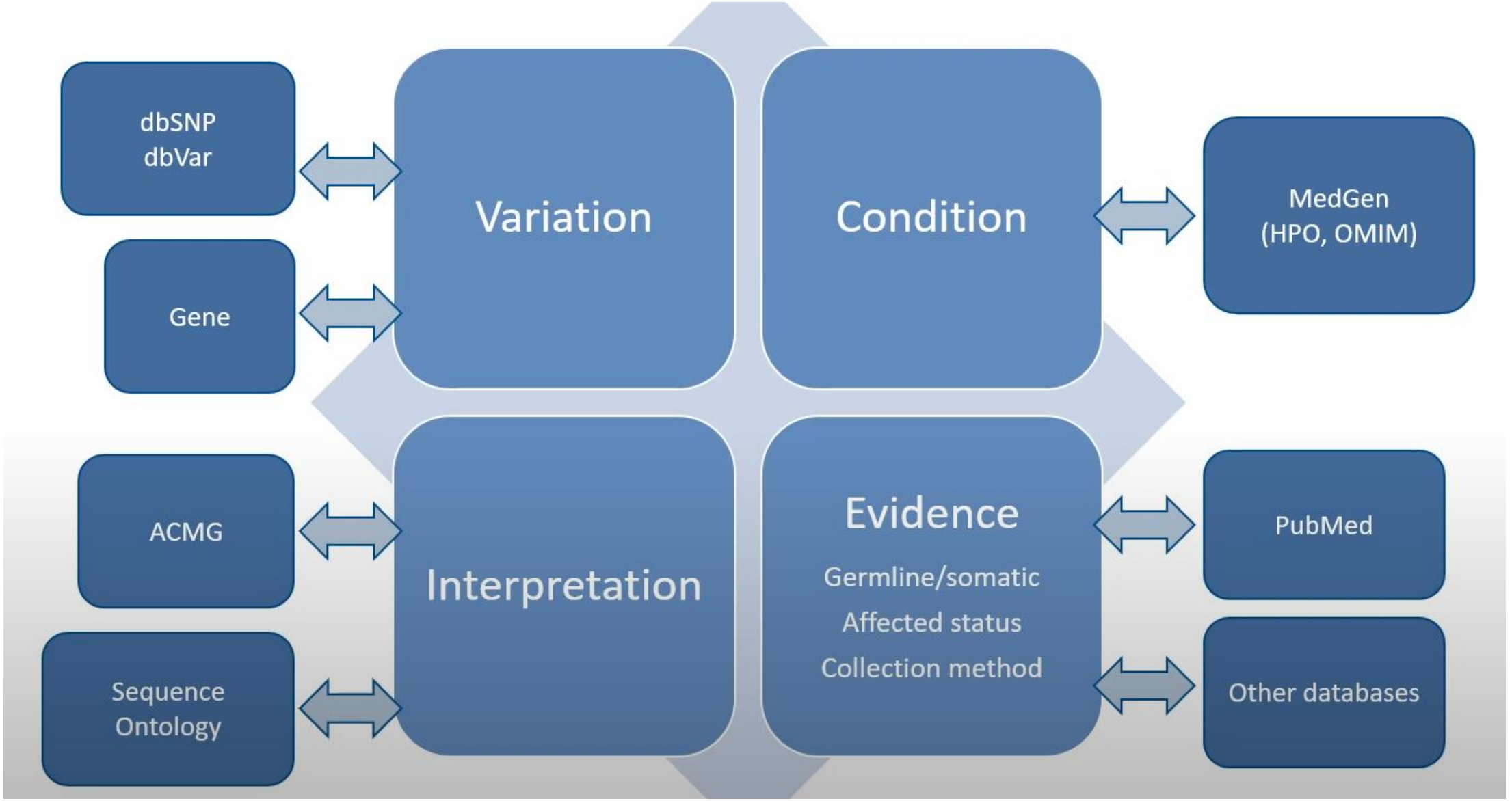
From
80 countries

https://www.ncbi.nlm.nih.gov/clinvar/

ClinVar evolution



ClinVar intergrates four domains of information



ClinVar - Scope of data

Variant

- Anywhere in the genome
- Any type/size of variation
- Single variant or combinations – haplotypes, genotypes

Condition

- Disease, phenotype, or drug response
- Single condition or combination

Interpretation

- Pathogenicity for Mendelian diseases
- Relationship to cancer
- Pharmacogenomics

Accessing ClinVar data

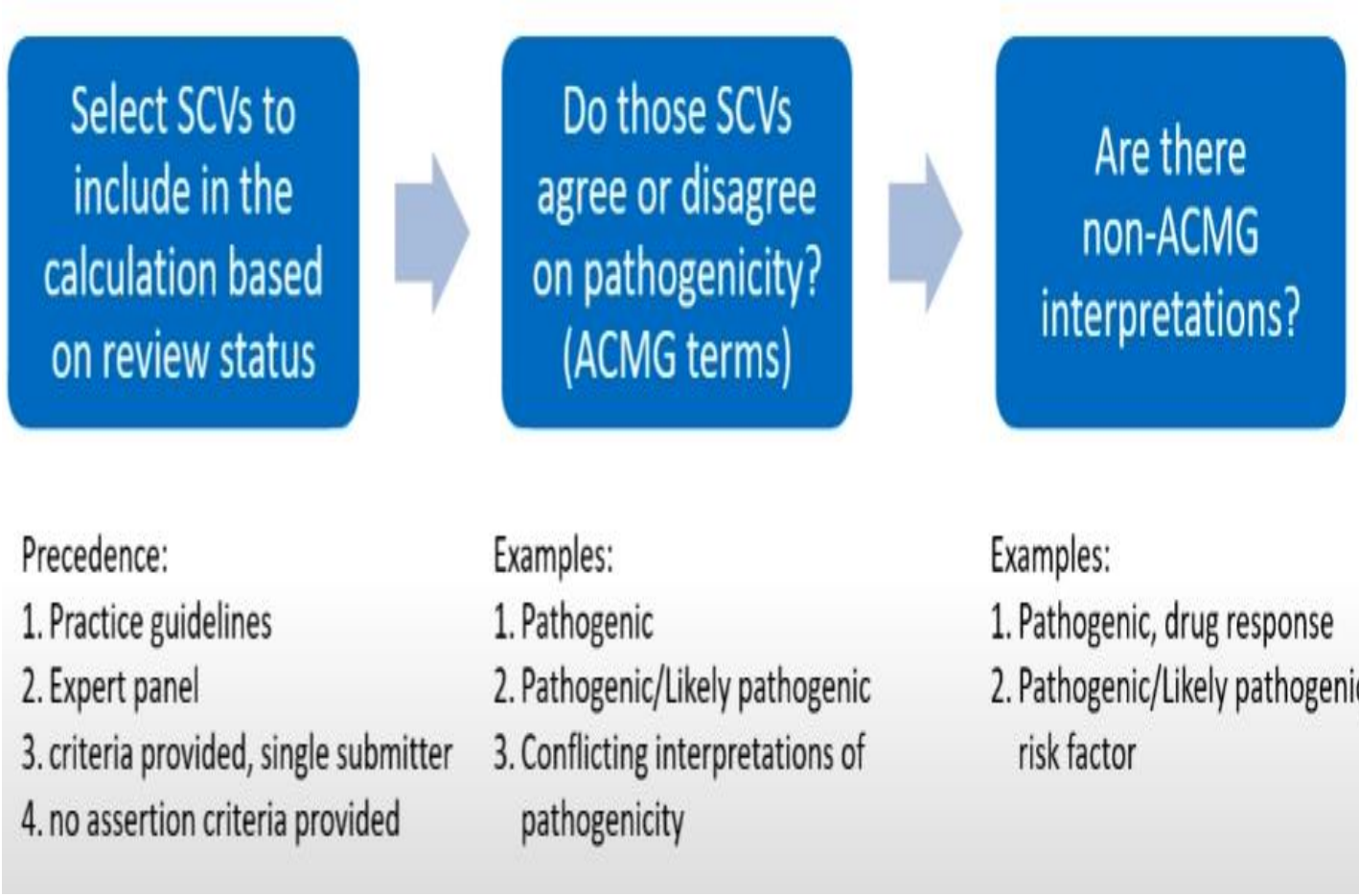
Searching ClinVar

- HGVS expressions
  - NM_000016.5:c.296G>T
  - C.296G>T
  - P.Gly99Val
- rs numbers
- Protein changes
  - V600E
- Gene symbols
  - Official gene symbols from HGNC
- Disease and phenotypes

https://www.ncbi.nlm.nih.gov/clinvar/docs/help/

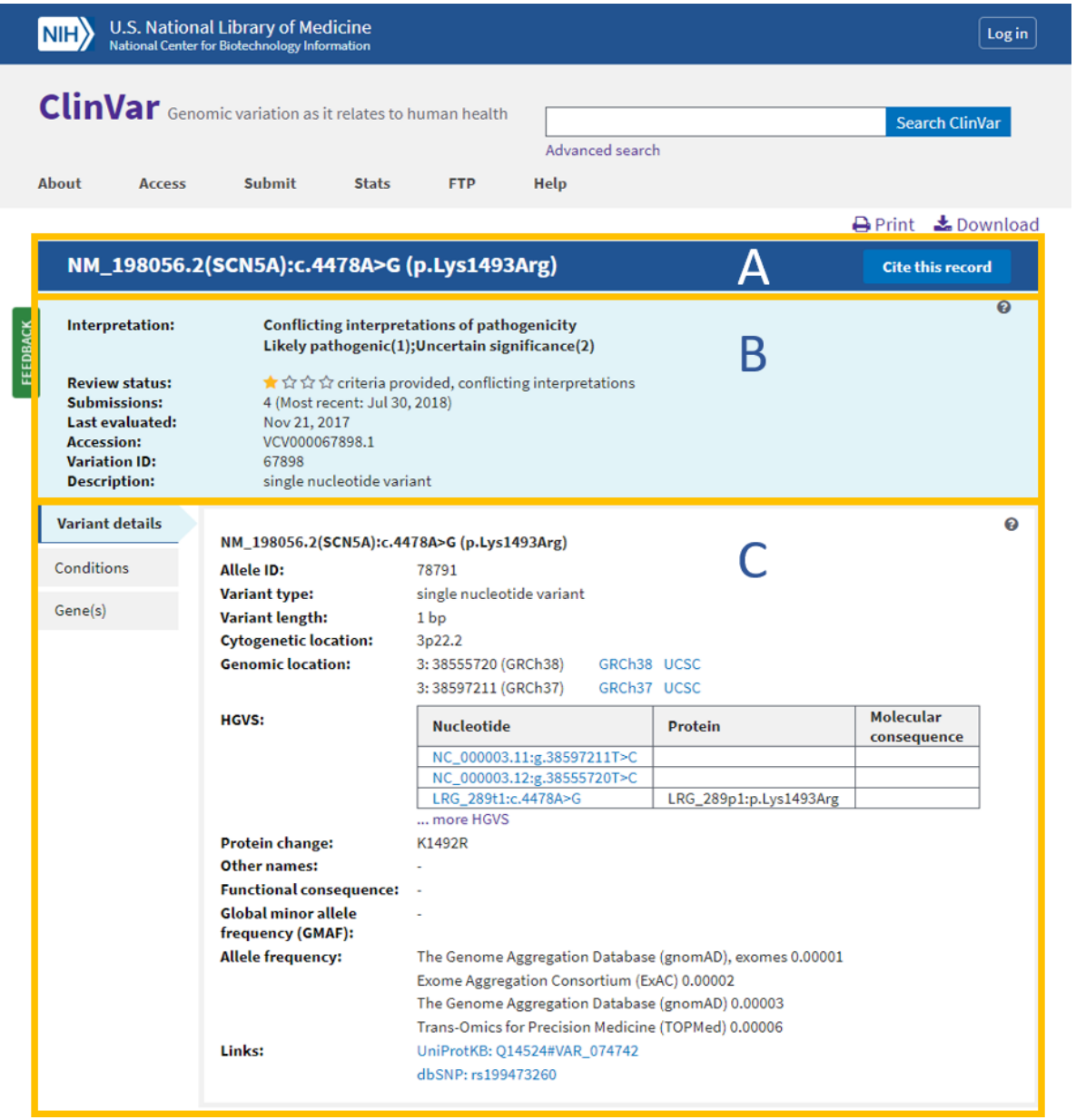
Calculating aggregate interpretation

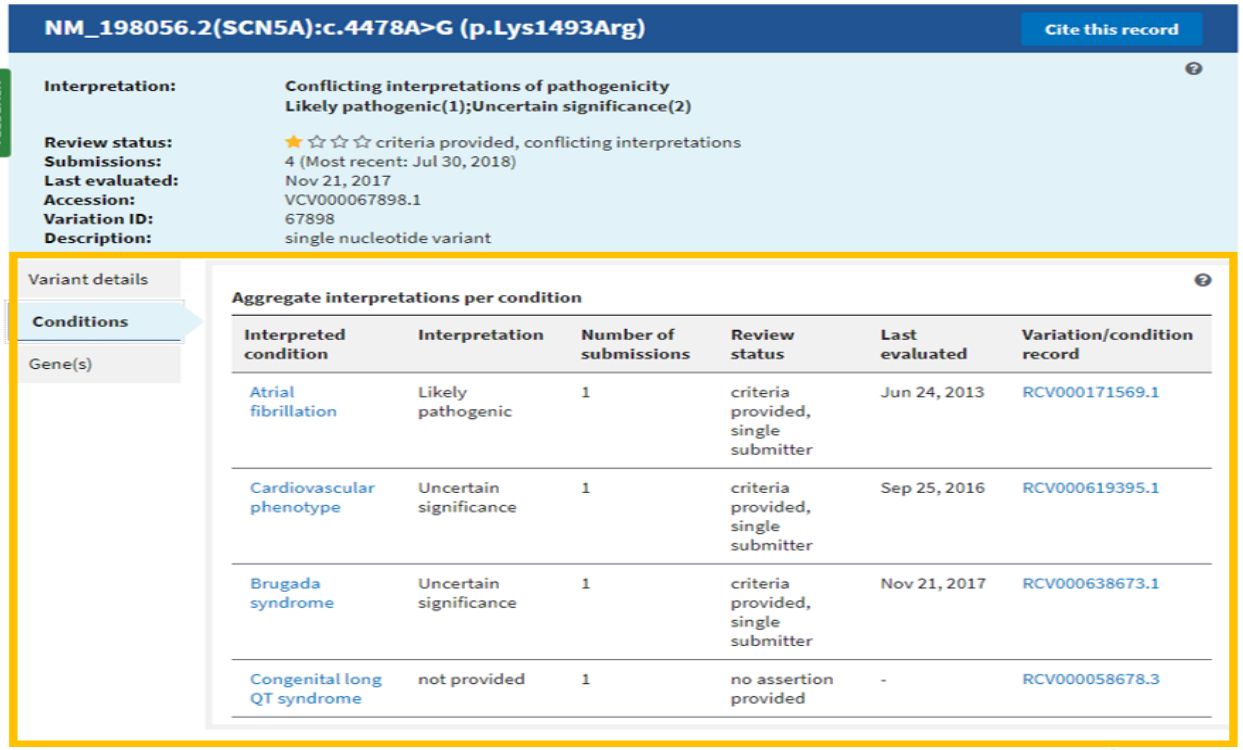
- A submitted interpretation is on each submitted record (SCV).
- An aggregate interpretation is calculated for each VCV record.

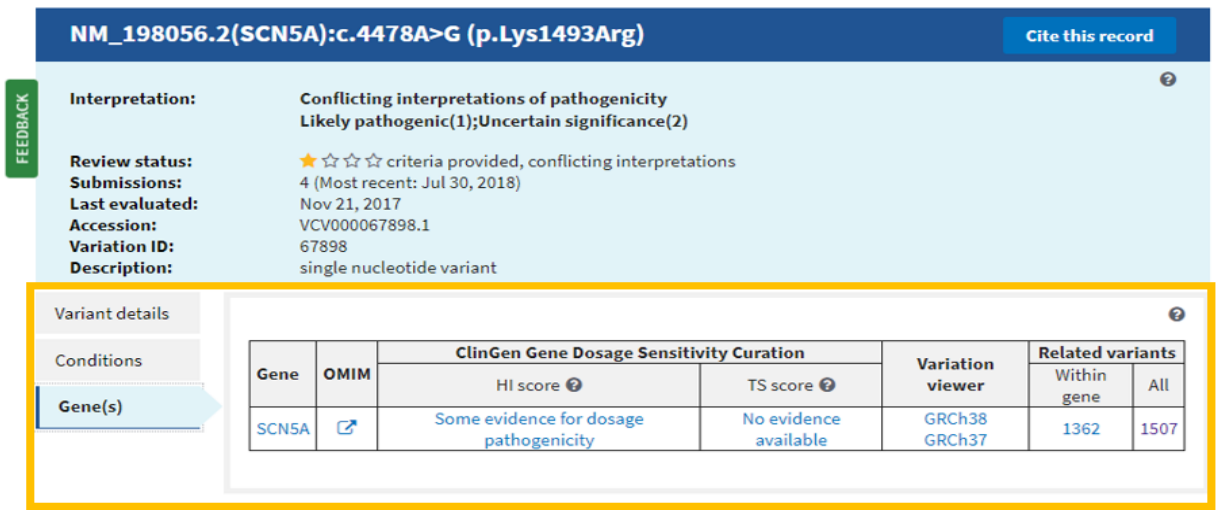


https://www.ncbi.nlm.nih.gov/clinvar/docs/clinsig/#clinsig_scv

The ClinVar variation report sample







https://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report/

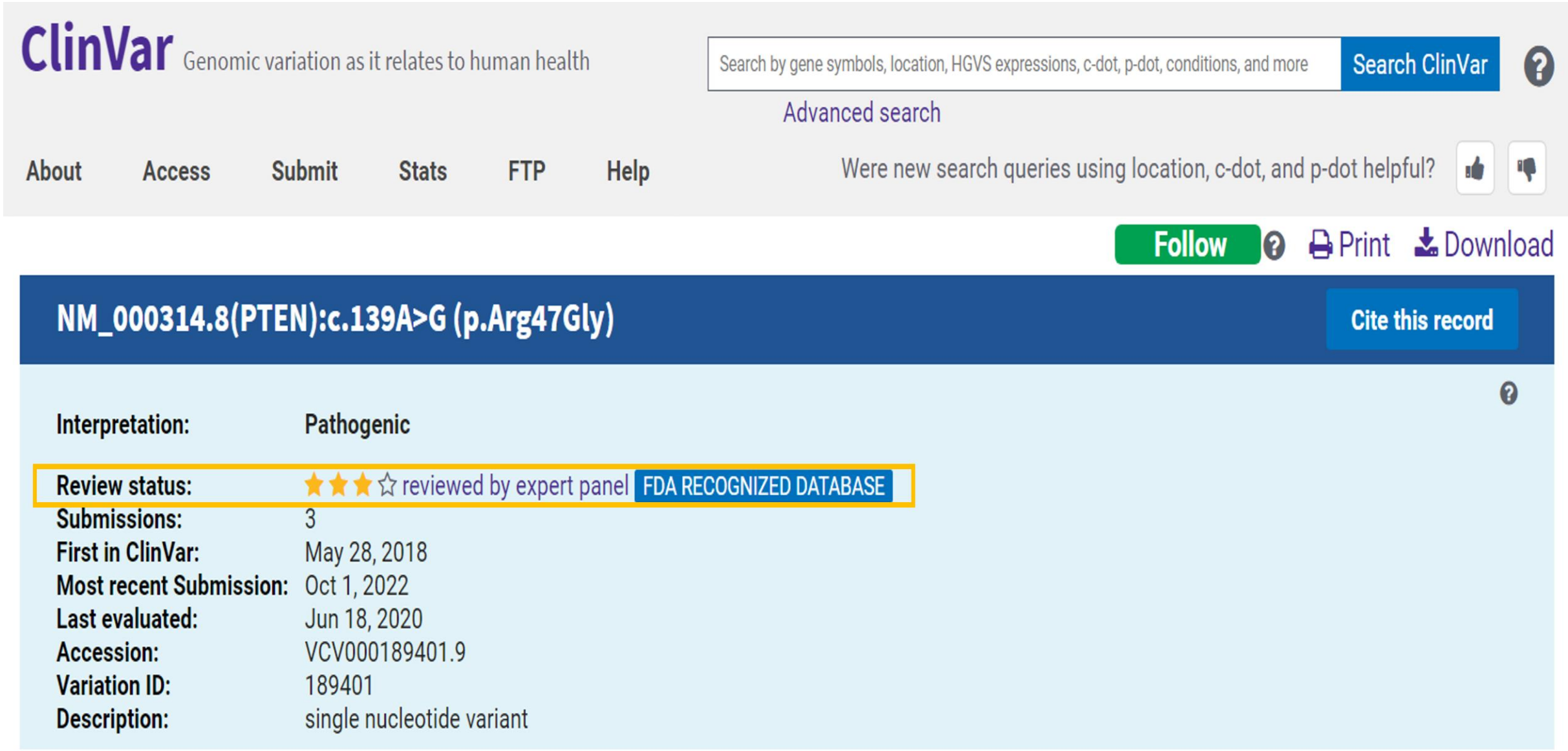
The ClinVar variant interpretations and supported evidence

Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	Supporting information (See all)
Likely pathogenic (Jun 24, 2013)	criteria provided, single submitter (Submitter's publication) Method: research	Atrial fibrillation Allele origin: unknown	Biesecker Lab/Human Development Section,National Institutes of Health Study: ClinSeq Accession: SCV000055299.1 Submitted: (Mar 10, 2015) Comments (2): The study set was not selected for affection status in relation to any cancer. Pathogenicity categories were based on literature curation. See Pubmed ID:23861362 for ... (more) Medical sequencing	Evidence details
Uncertain significance (Nov 21, 2017)	criteria provided, single submitter (Nykamp K et al. (Genet Med 2017)) Method: clinical testing	Brugada syndrome Allele origin: germline	Invitae Accession: SCV000760212.1 Submitted: (Apr 02, 2018)	Evidence details Publications PubMed (3) Comment: This sequence change replaces lysine with arginine at codon 1493 of the SCN5A protein (p.Lys1493Arg). The lysine residue is highly conserved and there is a ... (more)
Uncertain significance (Sep 25, 2016)	criteria provided, single submitter (Ambry Autosomal Dominant and X-Linked criteria (10/2015)) Method: clinical testing	cardiovascular phenotype Allele origin: germline	Ambry Genetics Accession: SCV000737722.2 Submitted: (Jul 30, 2018)	Evidence details Publications PubMed (3) Comment: Lines of evidence used in support of classification: Insufficient evidence
not provided (-)	no assertion provided Method: literature only	Congenital long QT syndrome Allele origin: germline	Cardiovascular Biomedical Research Unit,Royal Brompton & Harefield NHS Foundation Trust Accession: SCV000090198.3 Submitted: (Sep 22, 2016)	Evidence details Publications PubMed (3) Comment: This variant has been reported as associated with Long QT syndrome in the following publications (PMID:19167345;PMID:19716085). This is a literature report, and does not necessarily ... (more)

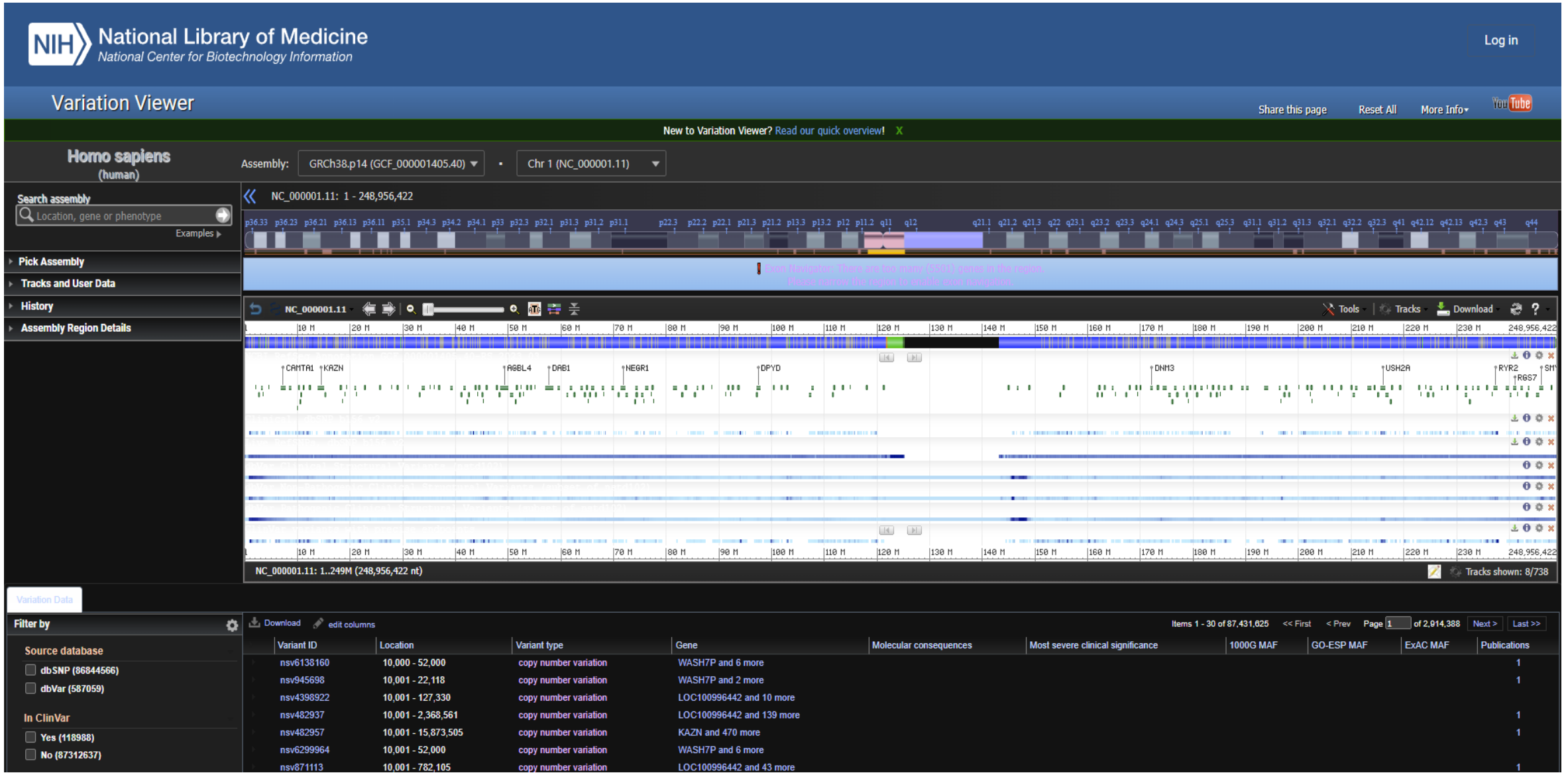
https://www.ncbi.nlm.nih.gov/clinvar/docs/variation_report/

Review status in ClinVar



https://www.ncbi.nlm.nih.gov/clinvar/docs/review_status/

Variation Viewer



https://www.ncbi.nlm.nih.gov/variation/view

Database of Short Genetic Variations (dbSNP)

dbSNP: Database of Short Genetic Variations

An expansive catalog of short nucleotide changes for human
https://www.ncbi.nlm.nih.gov/snp
National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope and Access

The NCBI Short Genetic Variation database (dbSNP) [1], commonly known as dbSNP, catalogs short variations in nucleotide sequences for human. These variations include single nucleotide variations, as well as insertions, deletions, and short tandem repeats less than 50 nucleotides in length. Short genetic variations may be common, thus representing true polymorphisms, or they may be rare. Some rare human entries have additional information associated with them, including disease associations from ClinVar [2], genotype information and allele origin, as some variations arises in somatic rather than from germline.

Short nucleotide variation data can be accessed through the dbSNP homepage and EUtils API:
www.ncbi.nlm.nih.gov/snp and www.ncbi.nlm.nih.gov/books/NBK25501
VCF files JSON files are available for download through FTP:
ftp.ncbi.nlm.nih.gov/snp/latest_release/
API services based on the SPDI notation system [3] is available at:
api.ncbi.nlm.nih.gov/variation/v0/
dbSNP data can also be examined under the genomic context through the Variation Viewer:
www.ncbi.nlm.nih.gov/variation/view/

https://ftp.ncbi.nlm.nih.gov/pub/factsheets/Factsheet_SNP.pdf

https://www.ncbi.nlm.nih.gov/snp/

Variant annotation tools

Funcotator

BROAD INSTITUTE

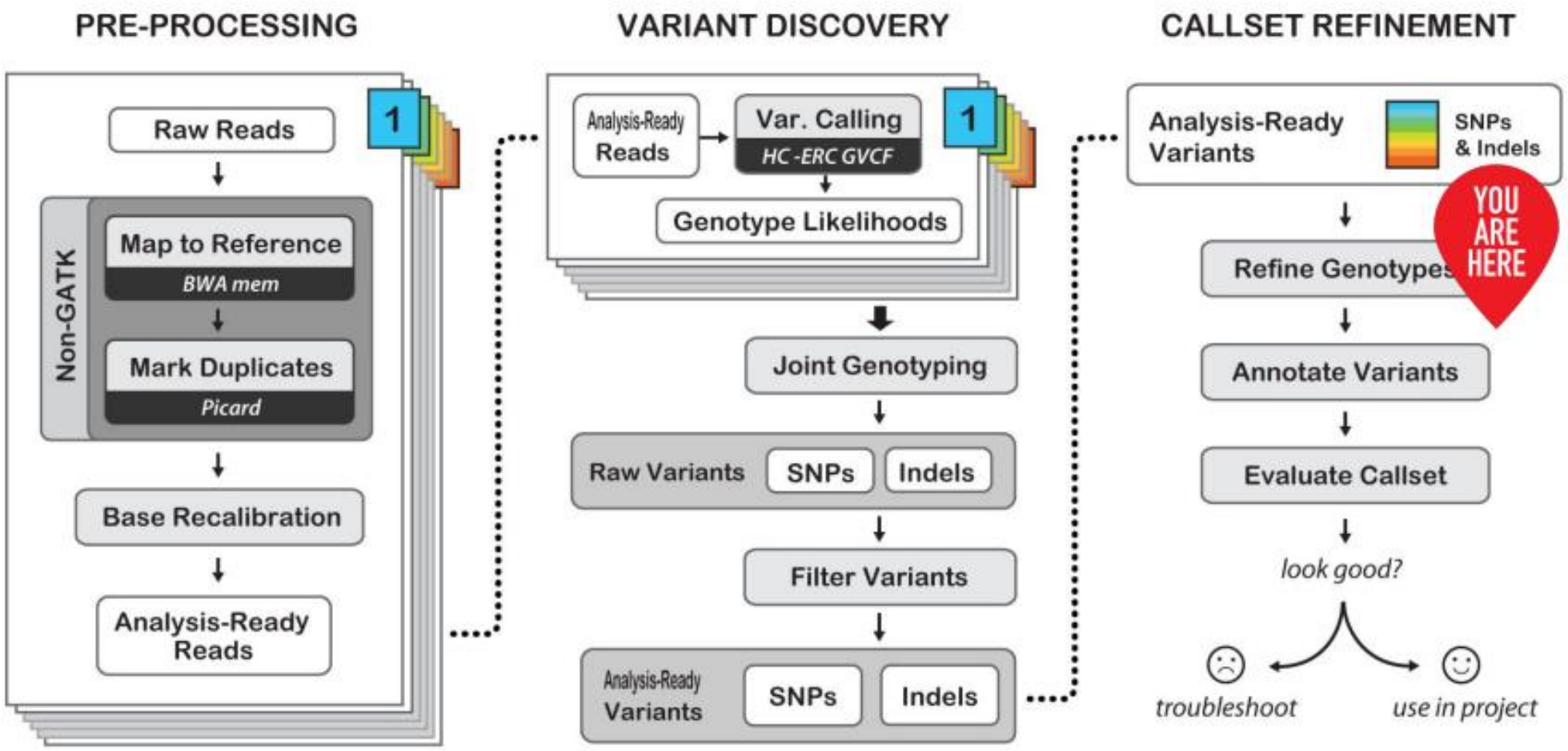
SnpEff
SnpSift

Pablo Cingolani

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F.
The Ensembl Variant Effect Predictor.
Genome Biology Jun 6;17(1):122. (**2016**)

Wang K, Li M, Hakonarson H.

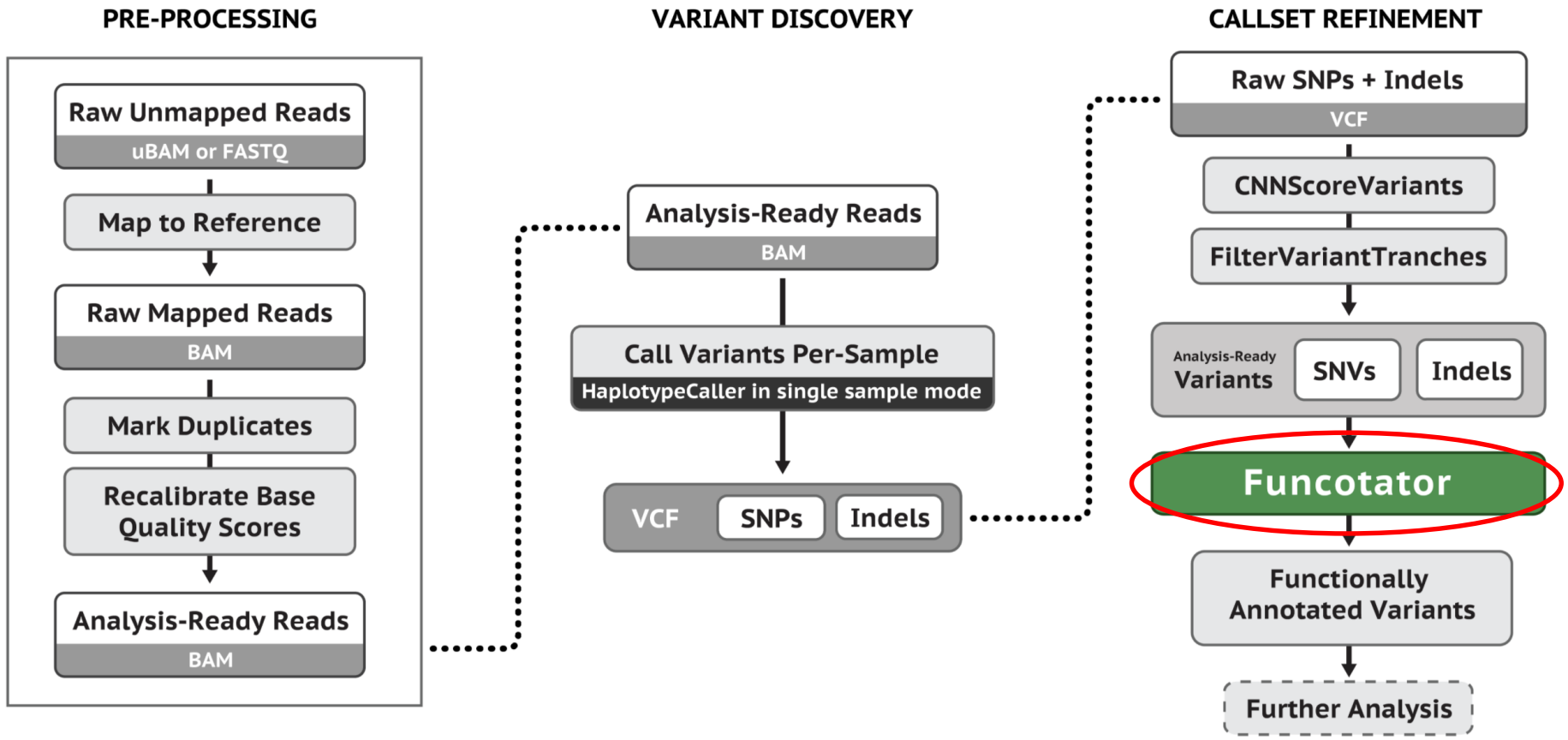
Variants Annotation process



Ready for annotating variants!!

GATK Functional Annotator

- Functional annotator analyzes give variants for their function (as retrieved from a set of data sources) and produces the analysis in a specified output file.



https://gatk.broadinstitute.org/hc/en-us/articles/360035889931-Funcotator-Information-and-Tutorial

Pre-Packaged Data Sources

Versioned gzip archives of data source files are provided here:

- FTP: ftp://gsapubftp-anonymous@ftp.broadinstitute.org/bundle/funcotator/
- Google Cloud Bucket: gs://broad-public-datasets/funcotator/

https://console.cloud.google.com/storage/browser/broad-public-datasets/funcotator

Data Source Downloader Tool

To download and extract the data sources, one can invoke {@link FuncotatorDataSourceDownloader} in the following ways:

- For germline data sources:

```
{@code ./gatk FuncotatorDataSourceDownloader --germline --validate-integrity --extract-after-download}
```

```
dataSourcesFolder/
    Data_Source_1/
        hg19
            data_source_1.config
            data_source_1.data.file.one
            data_source_1.data.file.two
            data_source_1.data.file.three
            ...
        hg38
            data_source_1.config
            data_source_1.data.file.one
            data_source_1.data.file.two
            data_source_1.data.file.three
            ...
    Data_Source_2/
        hg19
            data_source_2.config
            data_source_2.data.file.one
            data_source_2.data.file.two
            data_source_2.data.file.three
            ...
        hg38
            data_source_2.config
            data_source_2.data.file.one
            data_source_2.data.file.two
            data_source_2.data.file.three
            ...
     ...
```

```
./gatk Funcotator \
    --variant variants.vcf \
    --reference Homo_sapiens_assembly19.fasta \
    --ref-version hg19 \
    --data-sources-path funcotator_dataSources.v1.2.20180329 \
    --output variants.funcotated.vcf \
    --output-file-format VCF
```

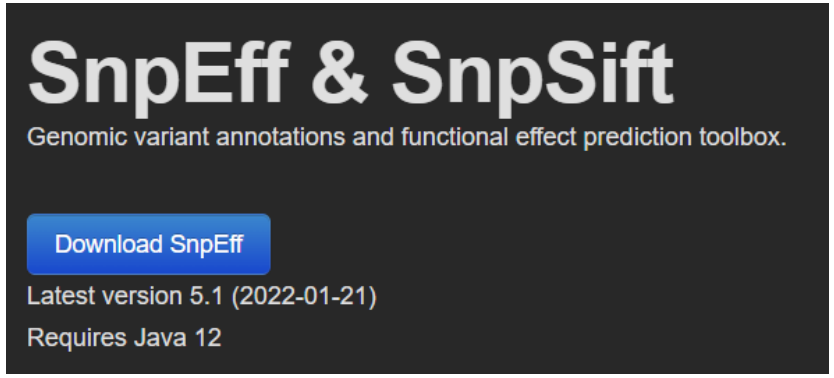
https://gatk.broadinstitute.org/hc/en-us/articles/13832755737115-FuncotatorDataSourceDownloader

SnpEff

1. **SnpEff** is a variant annotation and effect prediction tool. It annotates and predicts the effects of genetic variants (such as amino acid changes).
2. **SnpEff** Summary

   A typical SnpEff use case would be:

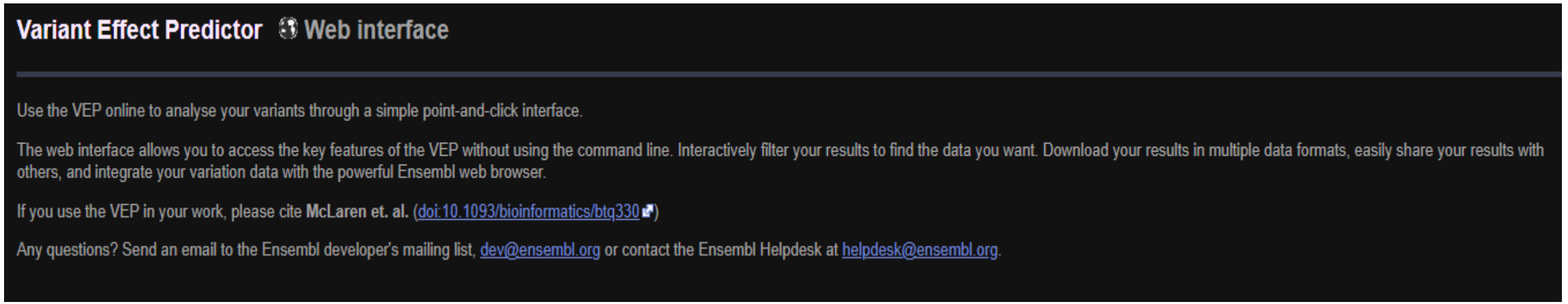
   - Input: The inputs are predicted variants (SNPs, insertions, deletions and MNPs). The input file is usually obtained as a result of a sequencing experiment, and it is usually in variant call format (VCF).
   - Output: SnpEff analyzes the input variants. It annotates the variants and calculates the effects they produce on known genes (e.g. amino acid changes).

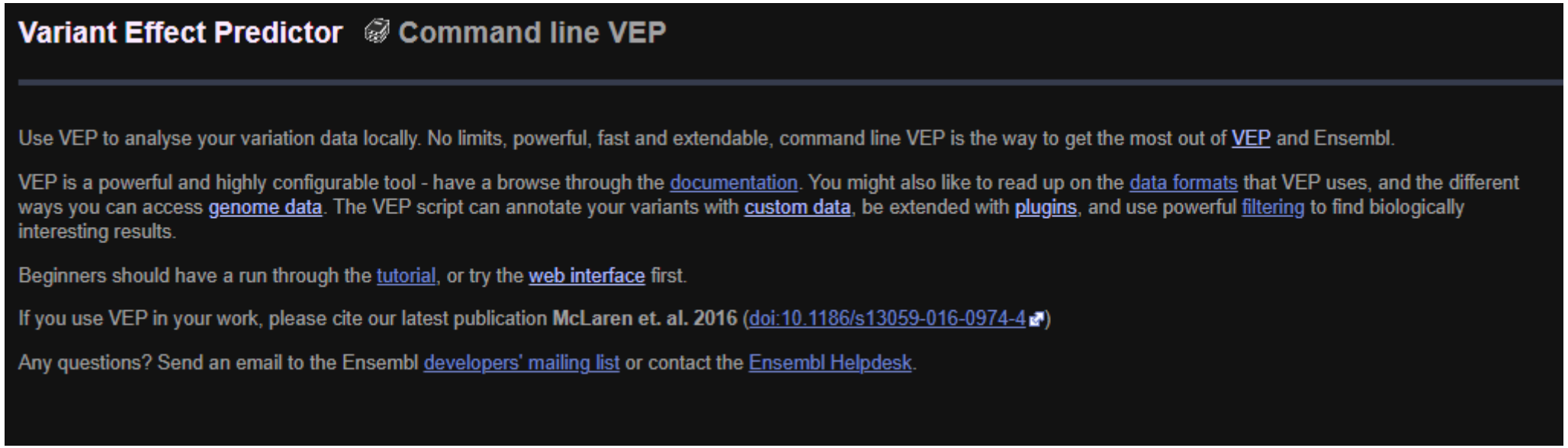


https://pcingola.github.io/SnpEff/

https://pcingola.github.io/SnpEff/se_introduction/

Variant Effect Predictor (VEP)





https://asia.ensembl.org/info/docs/tools/vep/index.html

WGLab/**doc-ANNOVAR**

Documentation for the ANNOVAR software

7 Contributors

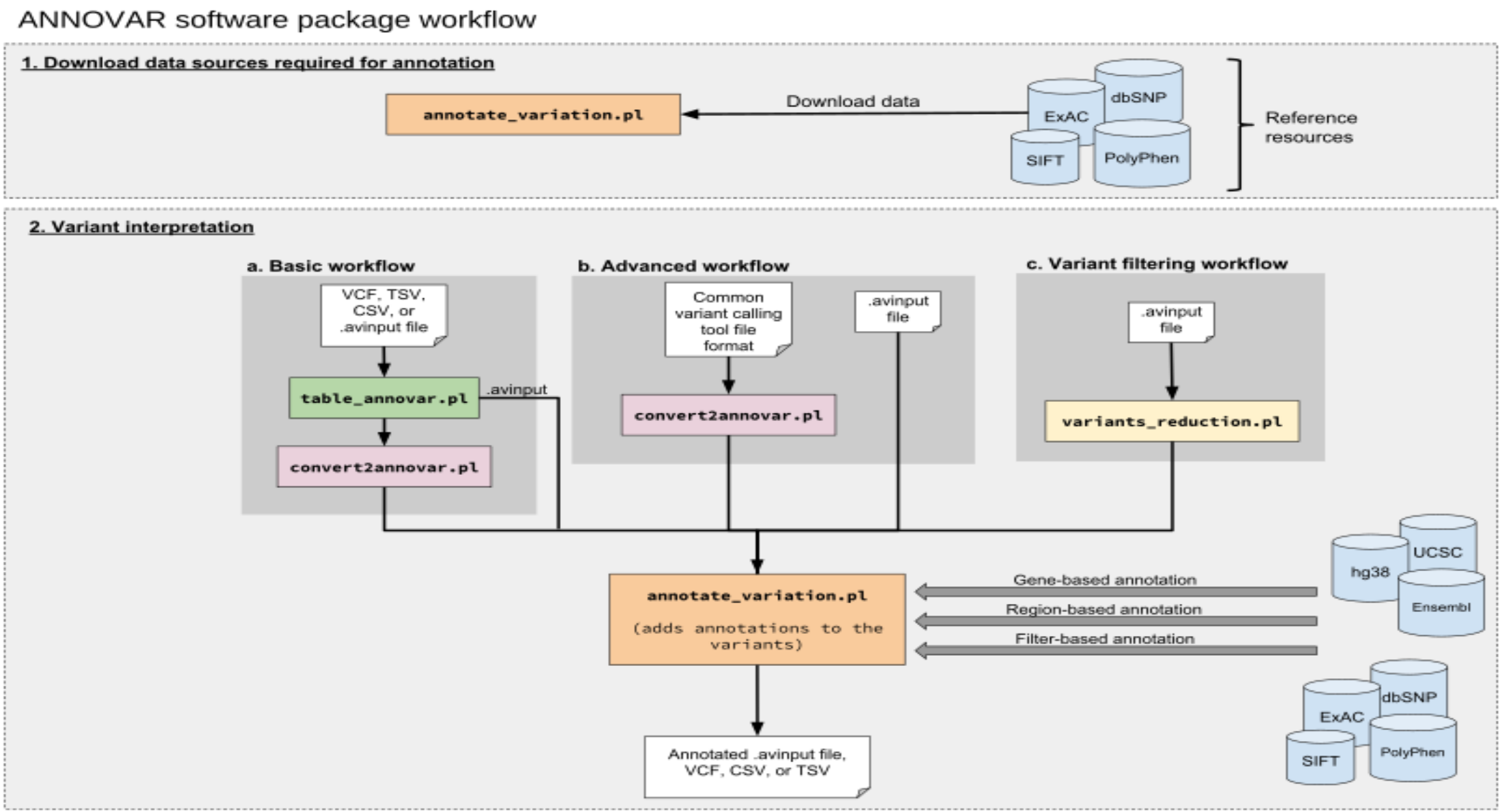
95 Issues

182 Stars

241 Forks

https://annovar.openbioinformatics.org/en/latest/

Annovar



https://en.wikipedia.org/wiki/ANNOVAR

Overview of main scripts in Annovar program

Script	Purpose	Description	Input	Output	Requirements
`annotate_variation.pl`	variant annotator	The core script, which functionally annotates the genetic variants via (1) gene-based, (2) region-based, and/or (3) filter-based annotation.	.avinput	.avinput	Data sources are downloaded for annotation, e.g. hg38, UCSC, 1000 Genomes Project.
`convert2annovar.pl`	file converter	Converts various file formats to the custom ANNOVAR input file format.	See "Conversion to the ANNOVAR input file format" section.	.avinput	
`table_annovar.pl`	automated variant annotator	A wrapper around `annotate_variation.pl` that can take VCF format along with the ANNOVAR format, performs annotation and outputs an Excel-compatible file. Ideal for beginners.	.avinput, CSV, TSV, VCF	CSV, TSV, VCF, TXT	Data sources are downloaded for annotation, e.g. hg38, UCSC, 1000 Genomes Project.
`variants_reduction.pl`	variant reducer	Performs stepwise variant reduction on a large set of input variants to narrow down to a subset of functionally important variants. Filtering procedures include: Applies a stepwise procedure of filtering to identify subsets of variants that are likely to be related to a disease.[2] Such filtering procedures include:[2] • identifying non-synonymous and splicing variants • removing variants in segmental duplication regions • identifying conserved genomic regions • removing variants from 1000 Genomes Project, ESP6500 and dbSNP	.avinput	.avinput	Gene-based annotation data sources and various filter-based annotation data sources are downloaded.

https://en.wikipedia.org/wiki/ANNOVAR

Interpreting results

Variant visulization from sample data: chr21_tumor


NCBI
The National Center for Biotechnology Information

GATK4 Functional annotator output table

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	AC	AN	DP	AF	FUNCOTATION												
2	1	2	48	0.5	[SAMSN1|hg38|chr21|14521246|14521246|INTRON||SNP|A|A|T|g.chr21:14521246A>T|ENST00000400566.6|-|||c.e2+25T>A												
3	2	2	84	1	[SAMSN1|hg38|chr21|14582272|14582272|MISSENSE||SNP|G|G|A|g.chr21:14582272G>A|ENST00000647101.1|-|9|859|c.8												
4	2	2	41	1	[SAMSN1|hg38|chr21|14582339|14582339|MISSENSE||SNP|T|T|C|g.chr21:14582339T>C|ENST00000647101.1|-|9|792|c.74												
5	1	2	57	0.5	[USP25|hg38|chr21|15826970|15826970|INTRON||SNP|G|G|A|g.chr21:15826970G>A|ENST00000285679.10|+|||c.e14-7G>												
6	2	2	36	1	[LINC01549|hg38|chr21|17444014|17444014|RNA||SNP|A|A|G|g.chr21:17444014A>G|ENST00000440664.6|+|||c.e3-69A>G												
7	1	2	33	0.5	[CXADR|hg38|chr21|17513228|17513228|INTRON||SNP|C|C|G|g.chr21:17513228C>G|ENST00000284878.12|+|||c.e1+56C>												
8	1	2	100	0.5	[C21orf91|hg38|chr21|17796838|17796838|MISSENSE||SNP|G|G|C|g.chr21:17796838G>C|ENST00000400558.7|-|3|499|c.4												
9	1	2	78	0.5	[CHODL|hg38|chr21|18134334|18134334|INTRON||SNP|G|G|A|g.chr21:18134334G>A|ENST00000400128.5|+|||c.e3-12217												
10	1	2	63	0.5	[CHODL|hg38|chr21|18134377|18134377|INTRON||SNP|T|T|G|g.chr21:18134377T>G|ENST00000400128.5|+|||c.e3-122132												
11	1	2	56	0.5	[CHODL|hg38|chr21|18134398|18134398|INTRON||SNP|C|C|T|g.chr21:18134398C>T|ENST00000400128.5|+|||c.e3-122111												
12	1	2	142	0.5	[TMPRSS15|hg38|chr21|18269968|18269968|THREE_PRIME_UTR||SNP|G|G|T|g.chr21:18269968G>T|ENST00000284885.8|-												
13	1	2	57	0.5	[TMPRSS15|hg38|chr21|18275155|18275155|INTRON||SNP|G|G|A|g.chr21:18275155G>A|ENST00000284885.8|-|||c.e24-42												
14	1	2	70	0.5	[TMPRSS15|hg38|chr21|18279040|18279040|SILENT||SNP|A|A|G|g.chr21:18279040A>G|ENST00000284885.8|-|23|2851|c.2												
15	1	2	56	0.5	[TMPRSS15|hg38|chr21|18281026|18281026|INTRON||DEL|T|T|-|g.chr21:18281026delT|ENST00000284885.8|-|||c.e22-15AT												
16	1	2	113	0.5	[TMPRSS15|hg38|chr21|18294213|18294213|INTRON||SNP|T|T|C|g.chr21:18294213T>C|ENST00000284885.8|-|||c.e21-57A												
17	1	2	44	0.5	[TMPRSS15|hg38|chr21|18294584|18294584|INTRON||SNP|T|T|A|g.chr21:18294584T>A|ENST00000284885.8|-|||c.e20-19A												
18	1	2	97	0.5	[TMPRSS15|hg38|chr21|18297801|18297801|MISSENSE||SNP|G|G|A|g.chr21:18297801G>A|ENST00000284885.8|-|19|235												
19	1	2	82	0.5	[TMPRSS15|hg38|chr21|18341504|18341504|SILENT||SNP|C|C|T|g.chr21:18341504C>T|ENST00000284885.8|-|13|1636|c.1												
20	1	2	96	0.5	[TMPRSS15|hg38|chr21|18398245|18398245|MISSENSE||SNP|T|T|C|g.chr21:18398245T>C|ENST00000284885.8|-|2|393|c.2												
21	1	2	30	0.5	[TMPRSS15|hg38|chr21|18398417|18398417|INTRON||SNP|G|G|C|g.chr21:18398417G>C|ENST00000284885.8|-|||c.e2+88												
22	1	2	64	0.5	[PPIAP22|hg38|chr21|18857784|18857784|RNA||SNP|C|C|T|g.chr21:18857784C>T|ENST00000358455.4|+|||c.e1+5C>T|||0												
23	2	2	51	1	[NCAM2|hg38|chr21|21286225|21286231|INTRON||DEL|TTTGTGA|TTTGTGA|-|g.chr21:21286225_21286231delTTTGTGA|ENS												
24	2	2	40	1	[NCAM2|hg38|chr21|21324517|21324517|INTRON||SNP|C|C|G|g.chr21:21324517C>G|ENST00000400546.6|+|||c.e6+17C>												

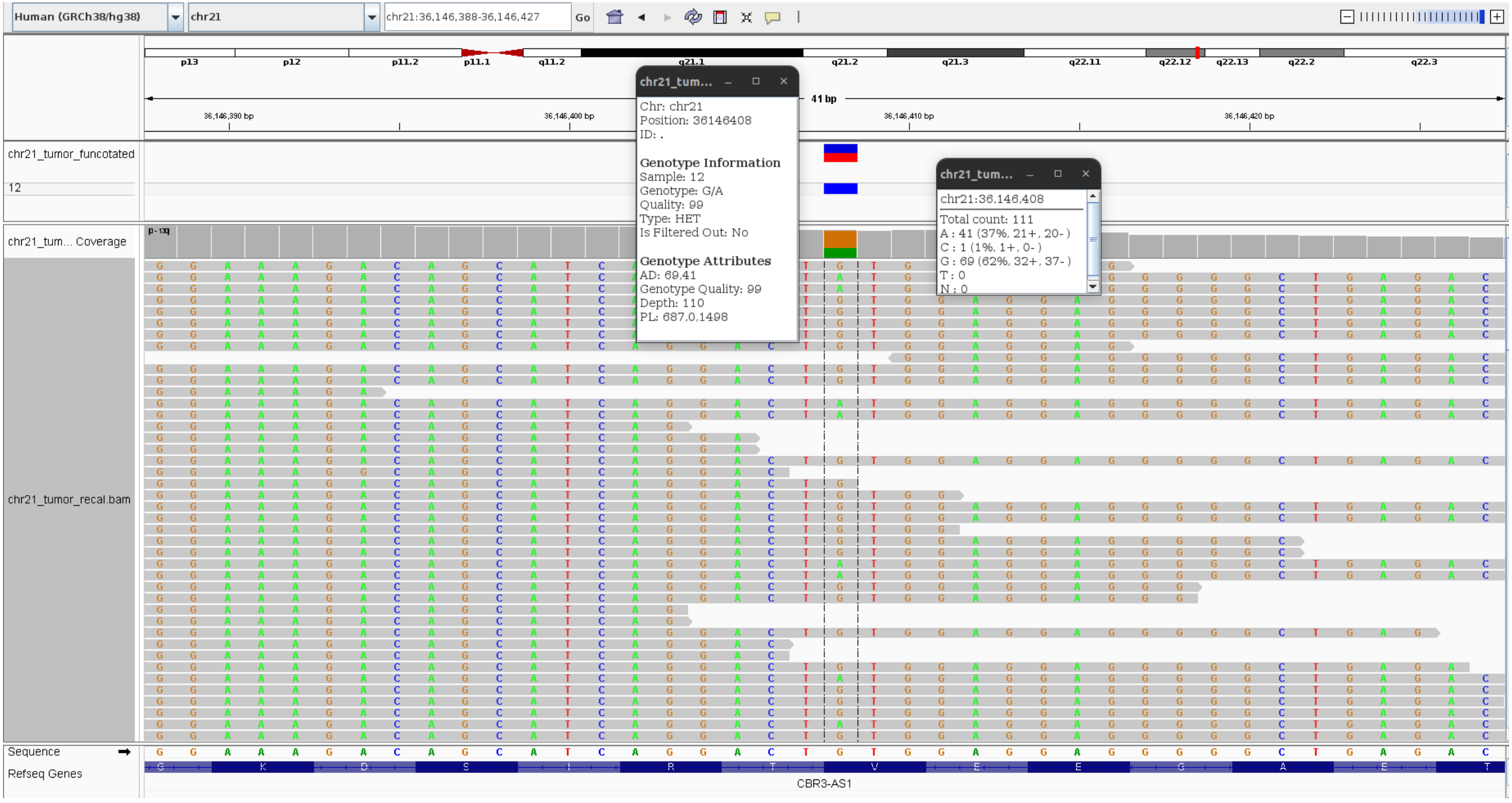
Parsing annotated variants into tabular format for interpretation

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode	Gencode
2	SAMSN1	hg38	chr21	1E+07	1E+07	INTRON		SNP	A	A	T	g.chr21:1	ENST000	-
3	SAMSN1	hg38	chr21	1E+07	1E+07	MISSENSE		SNP	G	G	A	g.chr21:1	ENST000	-
4	SAMSN1	hg38	chr21	1E+07	1E+07	MISSENSE		SNP	T	T	C	g.chr21:1	ENST000	-
5	USP25	hg38	chr21	2E+07	2E+07	INTRON		SNP	G	G	A	g.chr21:1	ENST000	+
6	LINC015	hg38	chr21	2E+07	2E+07	RNA		SNP	A	A	G	g.chr21:1	ENST000	+
7	CXADR	hg38	chr21	2E+07	2E+07	INTRON		SNP	C	C	G	g.chr21:1	ENST000	+
8	C21orf91	hg38	chr21	2E+07	2E+07	MISSENSE		SNP	G	G	C	g.chr21:1	ENST000	-
9	CHODL	hg38	chr21	2E+07	2E+07	INTRON		SNP	G	G	A	g.chr21:1	ENST000	+
10	CHODL	hg38	chr21	2E+07	2E+07	INTRON		SNP	T	T	G	g.chr21:1	ENST000	+
11	CHODL	hg38	chr21	2E+07	2E+07	INTRON		SNP	C	C	T	g.chr21:1	ENST000	+
12	TMPRSS1	hg38	chr21	2E+07	2E+07	THREE_PRIME_UTR		SNP	G	G	T	g.chr21:1	ENST000	-
13	TMPRSS1	hg38	chr21	2E+07	2E+07	INTRON		SNP	G	G	A	g.chr21:1	ENST000	-
14	TMPRSS1	hg38	chr21	2E+07	2E+07	SILENT		SNP	A	A	G	g.chr21:1	ENST000	-
15	TMPRSS1	hg38	chr21	2E+07	2E+07	INTRON		DEL	T	T	-	g.chr21:1	ENST000	-
16	TMPRSS1	hg38	chr21	2E+07	2E+07	INTRON		SNP	T	T	C	g.chr21:1	ENST000	-
17	TMPRSS1	hg38	chr21	2E+07	2E+07	INTRON		SNP	T	T	A	g.chr21:1	ENST000	-
18	TMPRSS1	hg38	chr21	2E+07	2E+07	MISSENSE		SNP	G	G	A	g.chr21:1	ENST000	-
19	TMPRSS1	hg38	chr21	2E+07	2E+07	SILENT		SNP	C	C	T	g.chr21:1	ENST000	-
20	TMPRSS1	hg38	chr21	2E+07	2E+07	MISSENSE		SNP	T	T	C	g.chr21:1	ENST000	-
21	TMPRSS1	hg38	chr21	2E+07	2E+07	INTRON		SNP	G	G	C	g.chr21:1	ENST000	-
22	PPIAP22	hg38	chr21	2E+07	2E+07	RNA		SNP	C	C	T	g.chr21:1	ENST000	+
23	NCAM2	hg38	chr21	2E+07	2E+07	INTRON		DEL	TTTGTGA	TTTGTGA	-	g.chr21:2	ENST000	+
24	NCAM2	hg38	chr21	2E+07	2E+07	INTRON		SNP	C	C	G	g.chr21:2	ENST000	+

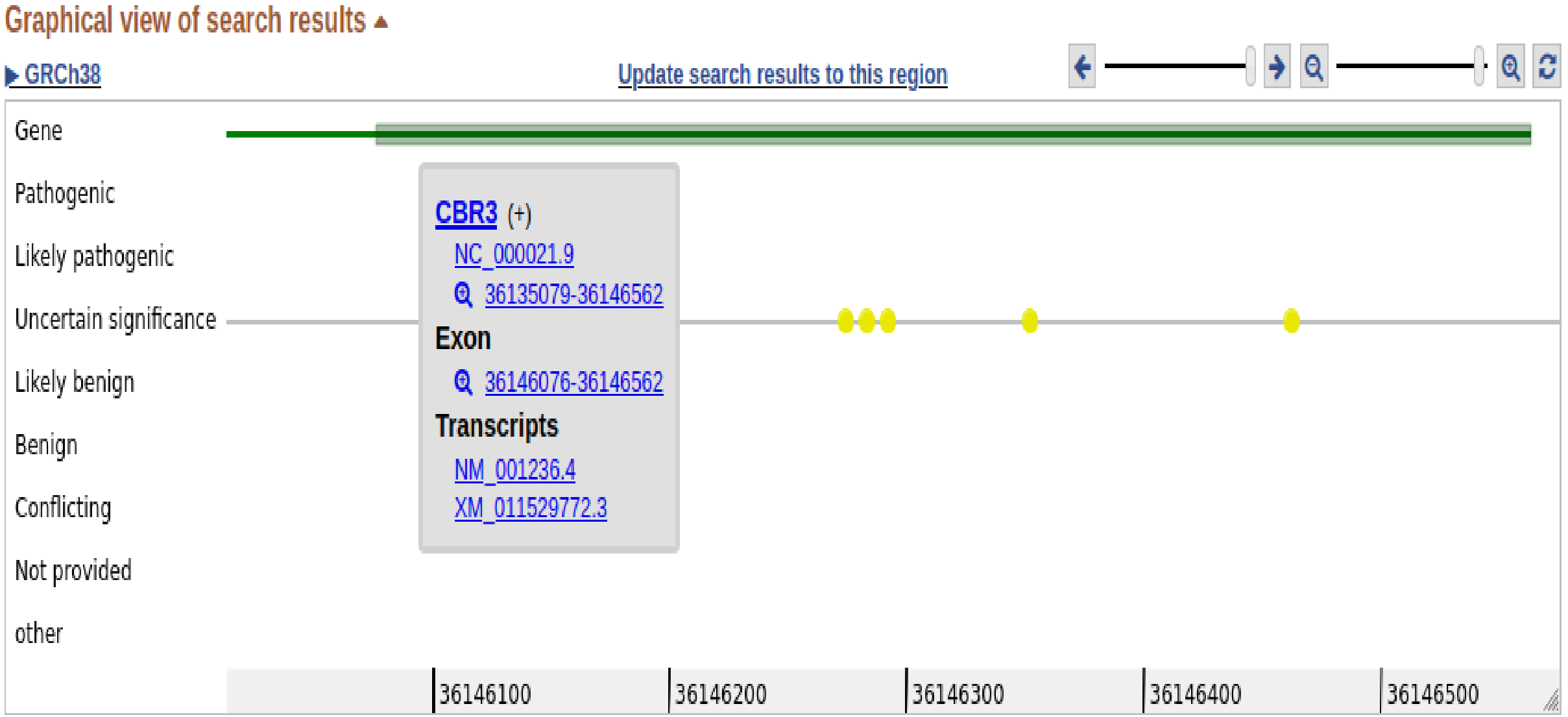
High-risk variants filtration

Gene	Consequence	Type	cDNA	Protein	ClinSig
CBR3	MISSENSE	SNV	c.730G>A	p.V244M	Drug respone
SLC19A1	MISSENSE	SNV	c.80A>G	p.H27R	Uncertain significance
FTCD	FRAMESHIFT INS	INS	c.990dup	p.P331fs	Conflicting interpretations of pathogenicity

CBR3


Human (GRCh38/hg38)
chr21
chr21:36,146,388-36,146,427
Go
p13
p12
p11.2
p11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
41 bp
36,146,390 bp
36,146,400 bp
36,146,410 bp
36,146,420 bp
chr21_tumor_funcotated
12
chr21_tum... Coverage
chr21_tumor_recal.bam
Sequence
Refseq Genes
CBR3-AS1
chr21_tum...
Chr: chr21
Position: 36146408
ID: .
Genotype Information
Sample: 12
Genotype: G/A
Quality: 99
Type: HET
Is Filtered Out: No
Genotype Attributes
AD: 69,41
Genotype Quality: 99
Depth: 110
PL: 687,0,1498
chr21_tum...
chr21:36,146,408
Total count: 111
A : 41 (37%, 21+, 20-)
C : 1 (1%, 1+, 0-)
G : 69 (62%, 32+, 37-)
T : 0
N : 0

CBR3: ClinVar graphical view



CBR3-chr21:36146408 (rs1056892)

Reference SNP (rs) Report

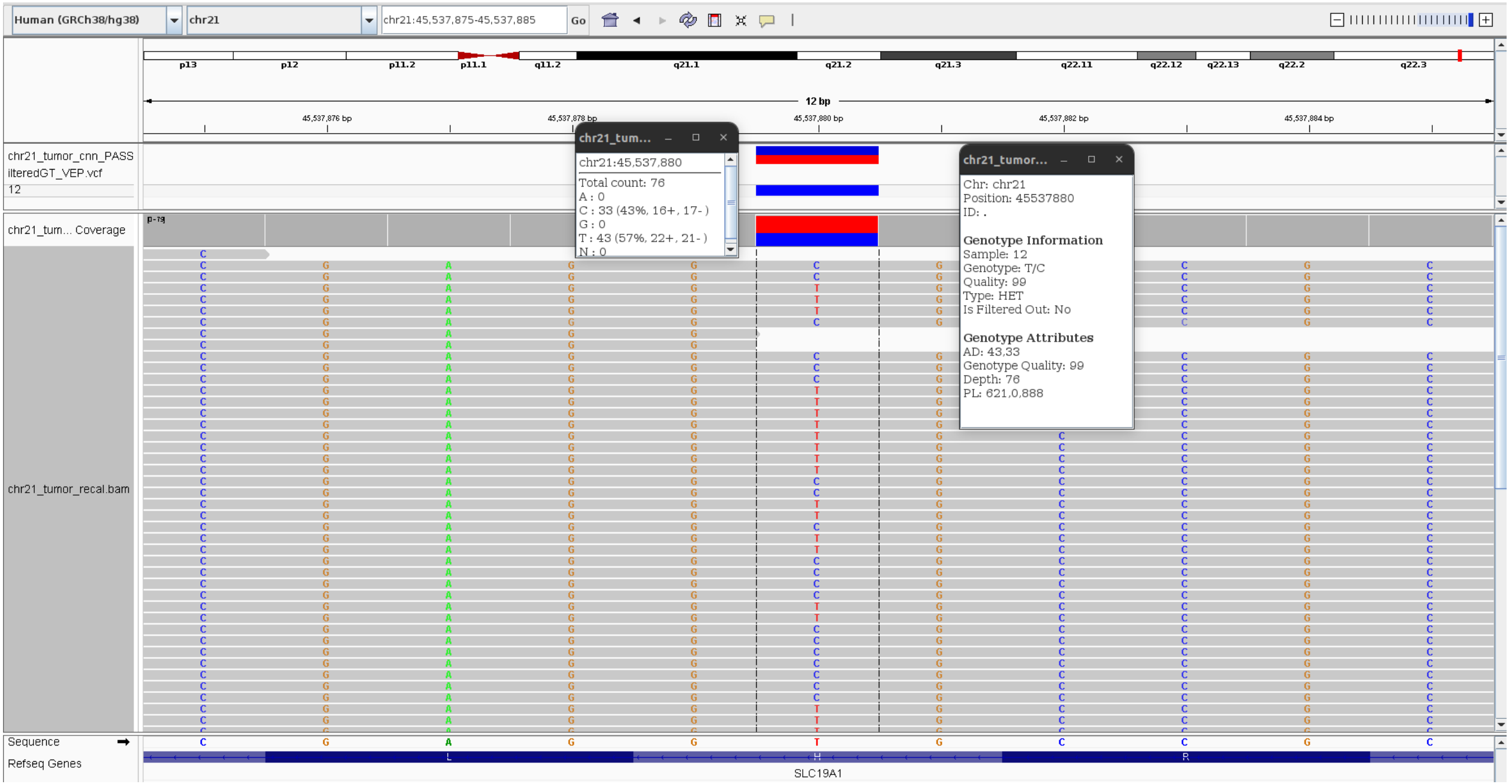
Download

rs1056892

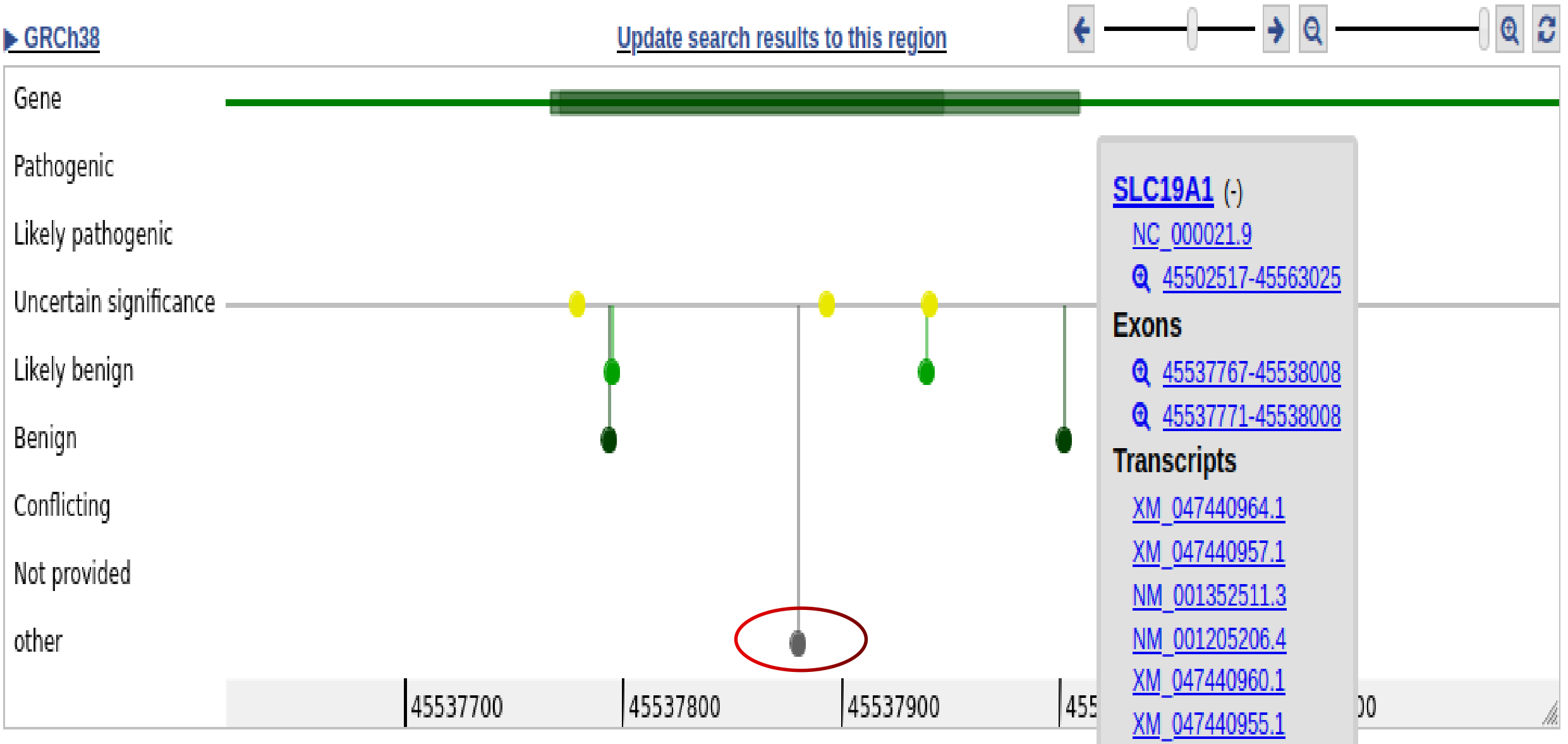
Current Build 156
Released September 21, 2022

Organism	*Homo sapiens*
Position	chr21:36146408 (GRCh38.p14)
Alleles	G>A
Variation Type	SNV Single Nucleotide Variation
Frequency	A=0.362416 (136846/377594, ALFA) A=0.386769 (102374/264690, TOPMED) A=0.367362 (92349/251384, GnomAD_exome) (+ 28 more)
Clinical Significance	Not Reported in ClinVar
Gene : Consequence	CBR3 : Missense Variant CBR3-AS1 : Intron Variant
Publications	15 citations
Genomic View	See rs on genome

SLC19A1


Human (GRCh38/hg38)
chr21
chr21:45,537,875-45,537,885
Go
p13
p12
p11.2
p11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
12 bp
45,537,876 bp
45,537,878 bp
45,537,880 bp
45,537,882 bp
45,537,884 bp
chr21_tumor_cnn_PASSilteredGT_VEP.vcf
12
chr21_tum... Coverage
chr21_tumor_recal.bam
chr21_tum...
chr21:45,537,880
Total count: 76
A : 0
C : 33 (43%, 16+, 17-)
G : 0
T : 43 (57%, 22+, 21-)
N : 0
chr21_tumor...
Chr: chr21
Position: 45537880
ID: .
Genotype Information
Sample: 12
Genotype: T/C
Quality: 99
Type: HET
Is Filtered Out: No
Genotype Attributes
AD: 43,33
Genotype Quality: 99
Depth: 76
PL: 621,0,888
Sequence
Refseq Genes
SLC19A1

SLC19A1 ClinVar graphical view



SLC19A1-chr21:45537880 (rs1051266)

Reference SNP (rs) Report

Download

rs1051266

Current Build 156
Released September 21, 2022

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr21:45537880 (GRCh38.p14)	**Gene : Consequence**	SLC19A1 : Missense Variant
Alleles	T>C / T>G	**Publications**	132 citations
Variation Type	SNV Single Nucleotide Variation		
Frequency	T=0.439253 (120805/275024, ALFA) T=0.487544 (129048/264690, TOPMED) T=0.485867 (68000/139956, GnomAD) (+ 25 more)	**Genomic View**	See rs on genome

SLC19A1 - Variant details

NM_194255.4(SLC19A1):c.80A>G (p.His27Arg)

Interpretation:	**drug response**
Review status:	★★★☆ reviewed by expert panel
Submissions:	3
First in ClinVar:	Nov 13, 2014
Most recent Submission:	Feb 7, 2023
Last evaluated:	Mar 24, 2021
Accession:	VCV000157588.5
Variation ID:	157588
Description:	single nucleotide variant

Variant details

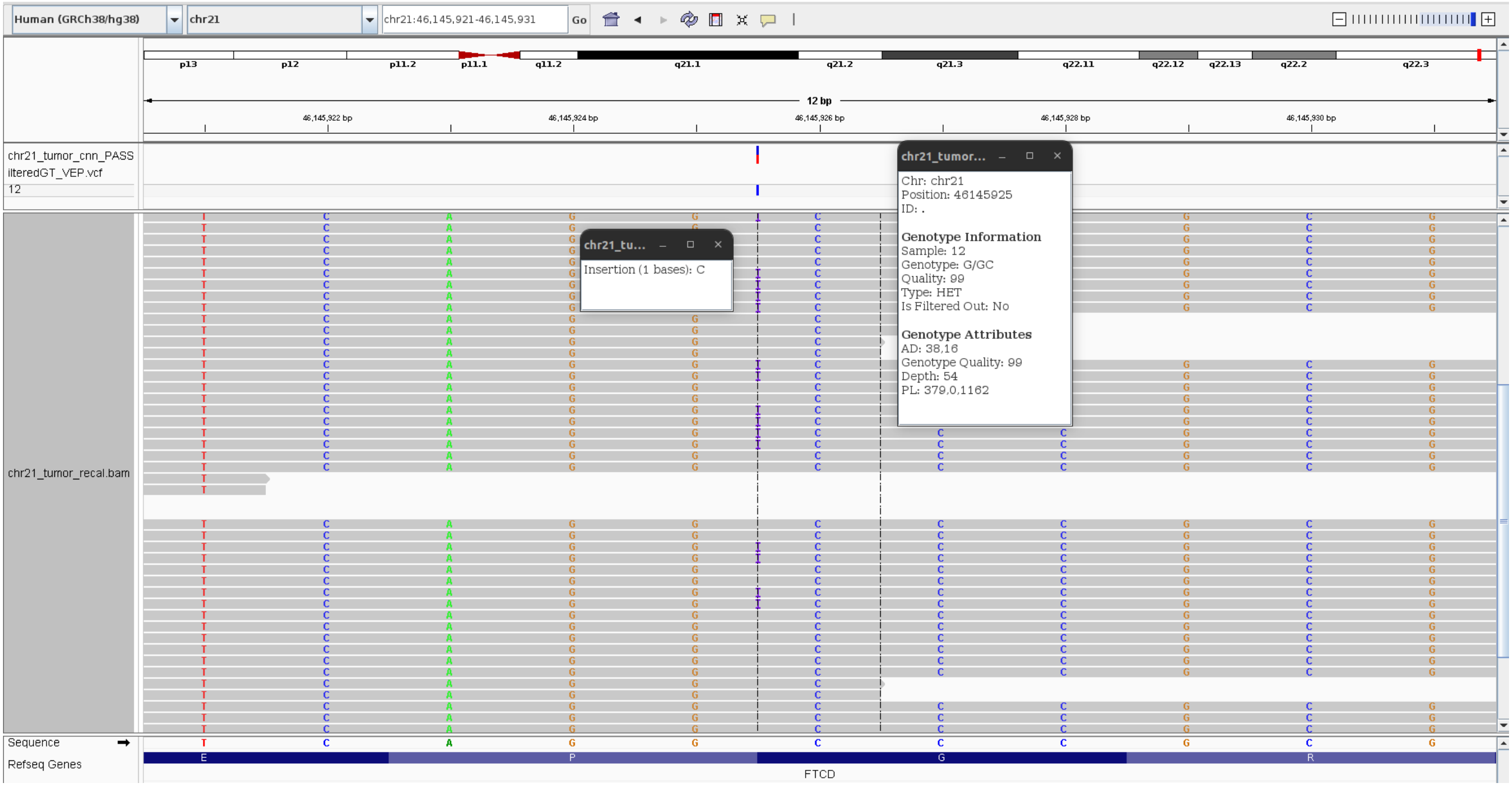
Conditions

Gene(s)

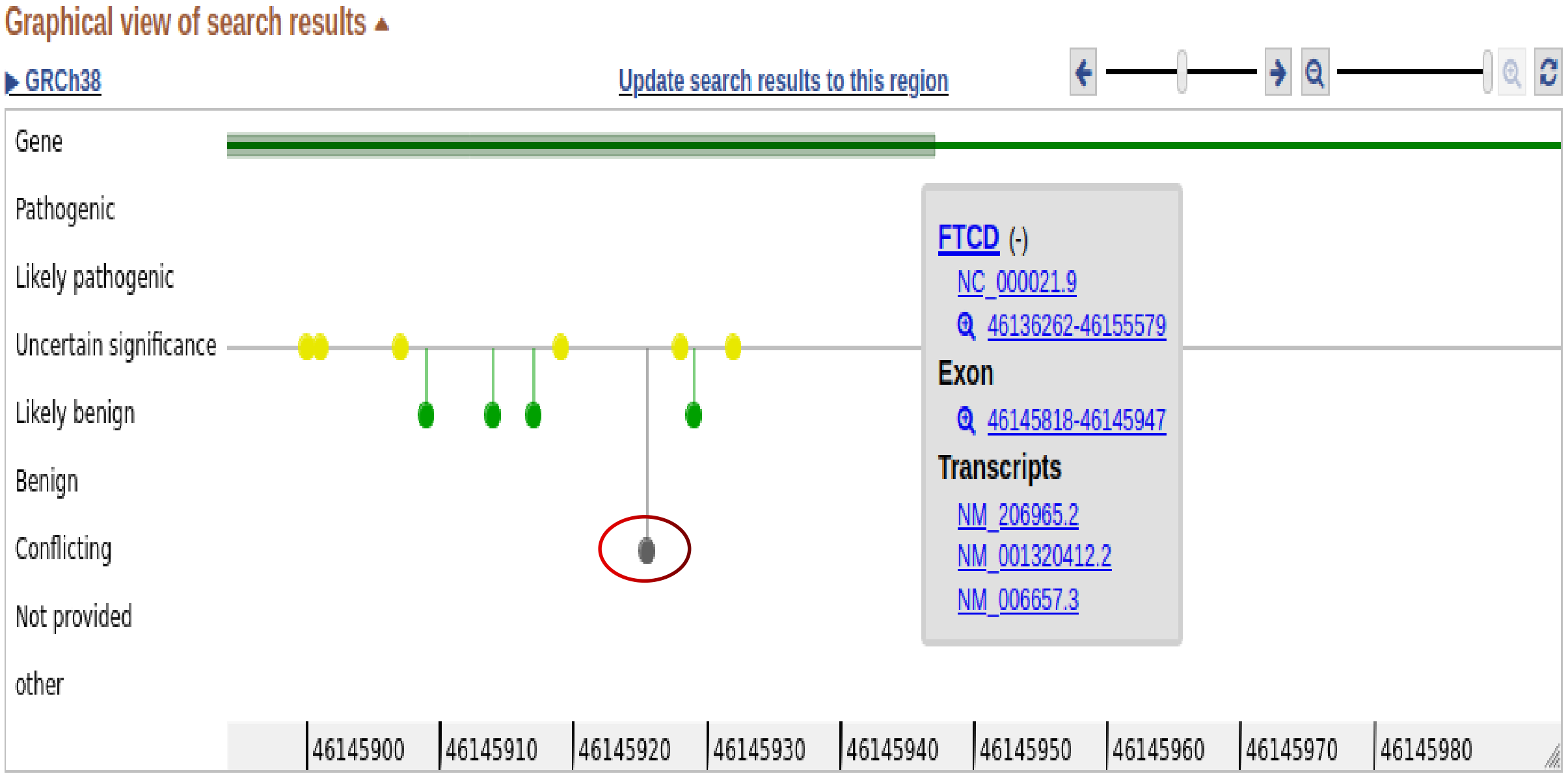
NM_194255.4(SLC19A1):c.80A>G (p.His27Arg)

Allele ID:	167450	
Variant type:	single nucleotide variant	
Variant length:	1 bp	
Cytogenetic location:	21q22.3	
Genomic location:	21: 45537880 (GRCh38)	GRCh38 UCSC
	21: 46957794 (GRCh37)	GRCh37 UCSC

FTCD


Human (GRCh38/hg38)
chr21
chr21:46,145,921-46,145,931
Go
p13
p12
p11.2
p11.1
q11.2
q21.1
q21.2
q21.3
q22.11
q22.12
q22.13
q22.2
q22.3
12 bp
46,145,922 bp
46,145,924 bp
46,145,926 bp
46,145,928 bp
46,145,930 bp
chr21_tumor_cnn_PASS
ilteredGT_VEP.vcf
12
chr21_tu...
Insertion (1 bases): C
chr21_tumor...
Chr: chr21
Position: 46145925
ID: .
Genotype Information
Sample: 12
Genotype: G/GC
Quality: 99
Type: HET
Is Filtered Out: No
Genotype Attributes
AD: 38,16
Genotype Quality: 99
Depth: 54
PL: 379,0,1162
chr21_tumor_recal.bam
Sequence
Refseq Genes
E
P
G
R
FTCD

FTCD ClinVar graphical view



FTCD - chr21: 46145925 - 46145926 (rs398124234)

Reference SNP (rs) Report

Download

rs398124234

Current Build 156
Released September 21, 2022

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr21:46145926-46145928 (GRCh38.p14)	**Gene : Consequence**	FTCD : Frameshift Variant
Alleles	dupC	**Publications**	5 citations
Variation Type	Indel Insertion and Deletion		
Frequency	dupC=0.003355 (888/264690, TOPMED) dupC=0.003893 (536/137694, GnomAD) dupC=0.002909 (326/112050, GnomAD_exome) (+ 5 more)	**Genomic View**	See rs on genome

FTCD - Variant details

NM_206965.2(FTCD):c.990dup (p.Pro331fs)

Interpretation:	**Conflicting interpretations of pathogenicity** **Pathogenic(7); Likely pathogenic(2); Uncertain significance(1)**
Review status:	★ ☆ ☆ ☆ criteria provided, conflicting interpretations
Submissions:	11
First in ClinVar:	Mar 24, 2015
Most recent Submission:	Apr 23, 2023
Last evaluated:	Sep 1, 2022
Accession:	VCV000004019.25
Variation ID:	4019
Description:	1bp duplication

- **Variant details**
- Conditions
- Gene(s)

NM_206965.2(FTCD):c.990dup (p.Pro331fs)

Allele ID:	19058		
Variant type:	Duplication		
Variant length:	1 bp		
Cytogenetic location:	21q22.3		
Genomic location:	21: 46145925-46145926 (GRCh38)	GRCh38	UCSC
	21: 47565839-47565840 (GRCh37)	GRCh37	UCSC


THANK
YOU