

Variant Annotation

Overview of variant databases



Presenter: Nguyen Le Duc Minh

Every genome contains many rare, potentially functional variants

- ~ 2 millions high quality variants in a variant call file (vcf)
- ~ 20,000 variants within coding (exome) regions
- ~ 500 rare missense variants (1/3 of which are predicted damaging by in silico predictors)
- ~ 100 LoF variants: ~20 homozygous, ~20 rare
- ~ 100 rare variants in known disease genes
- ~ 50 reported disease-causing mutations
- 1-2 *de novo* coding mutations
- 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)

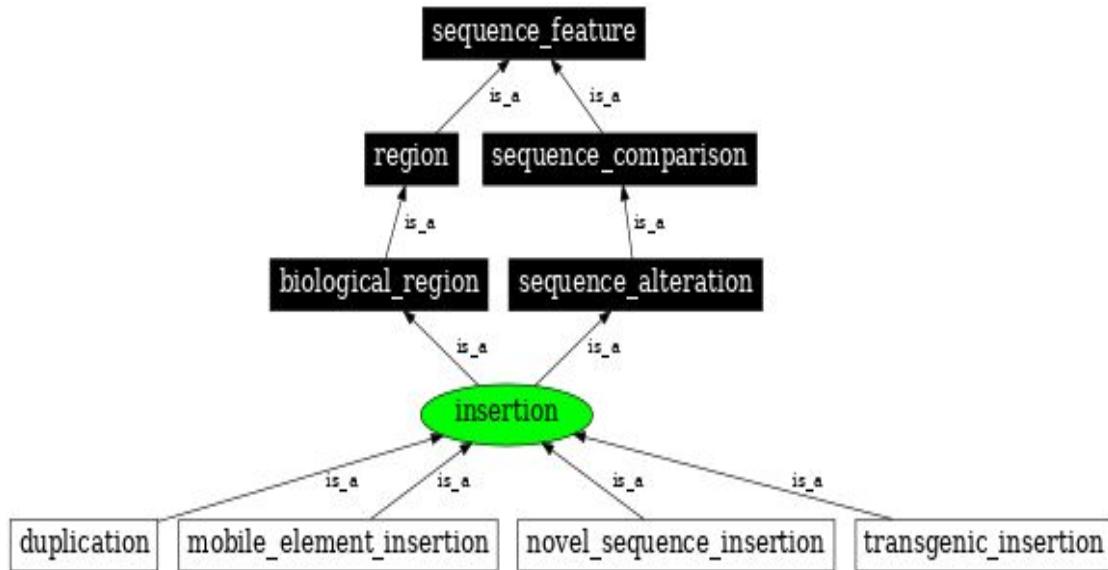
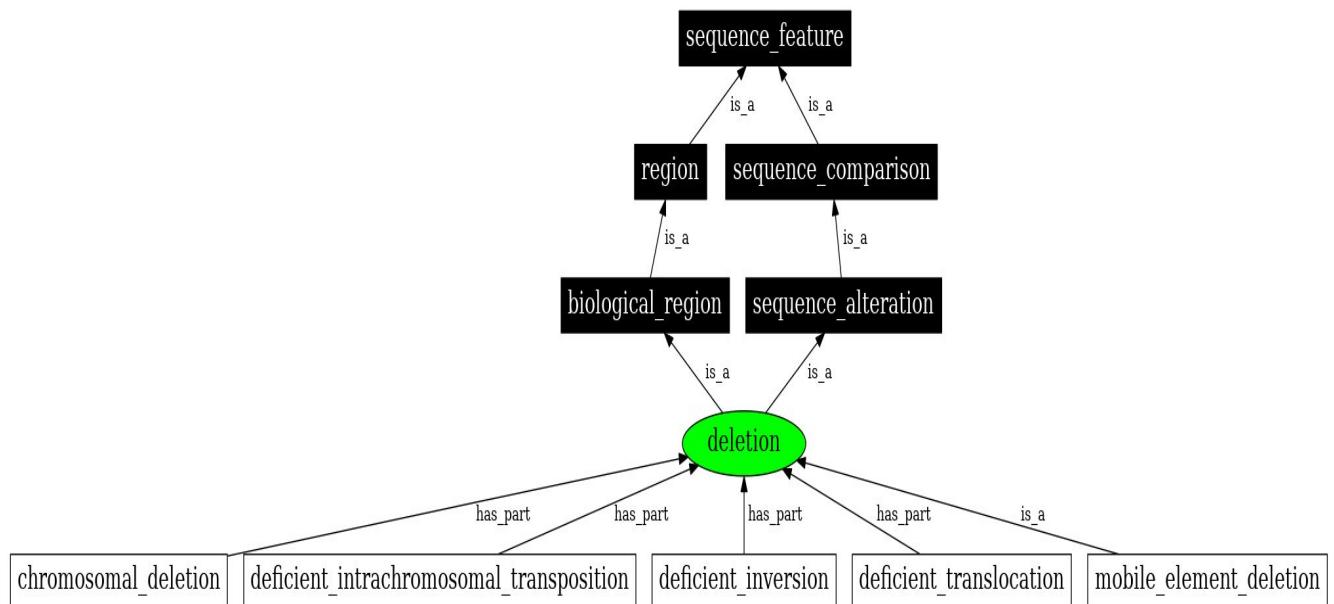
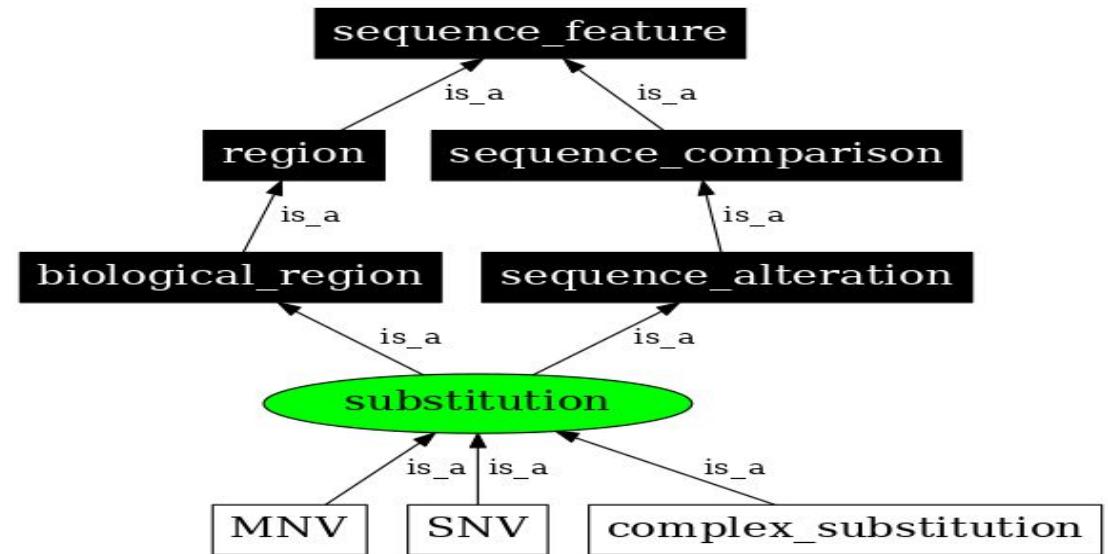
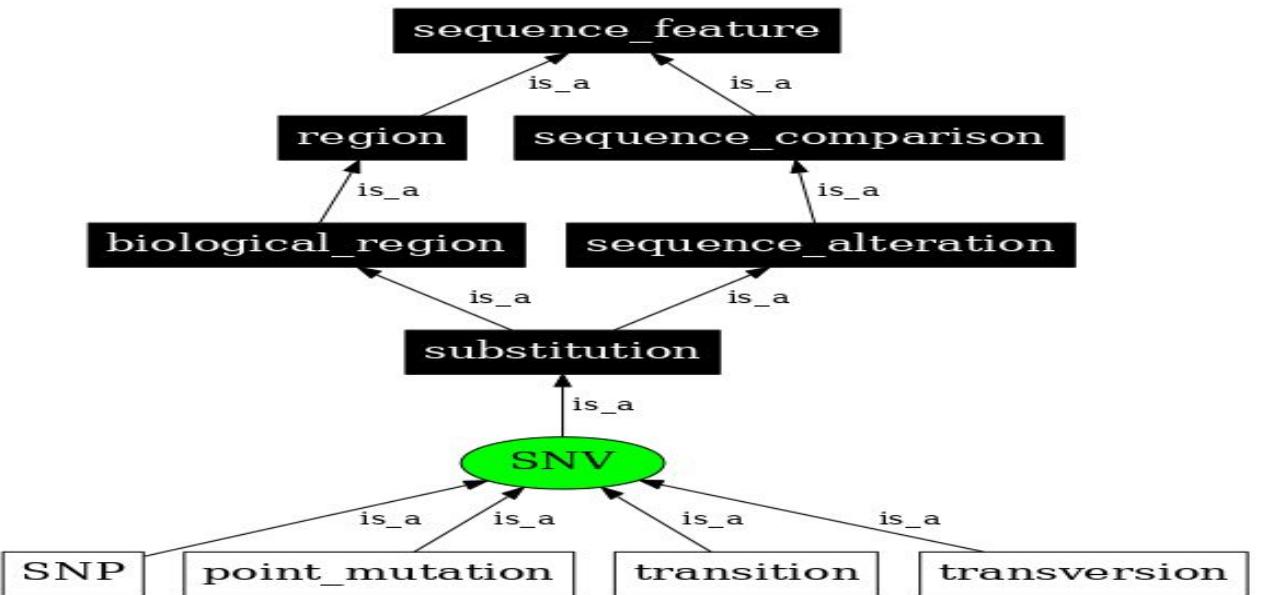
Sequence variants

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 GT...	Alt: ... TTGGT...
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 GT...	Alt: ... TTG TA GT...

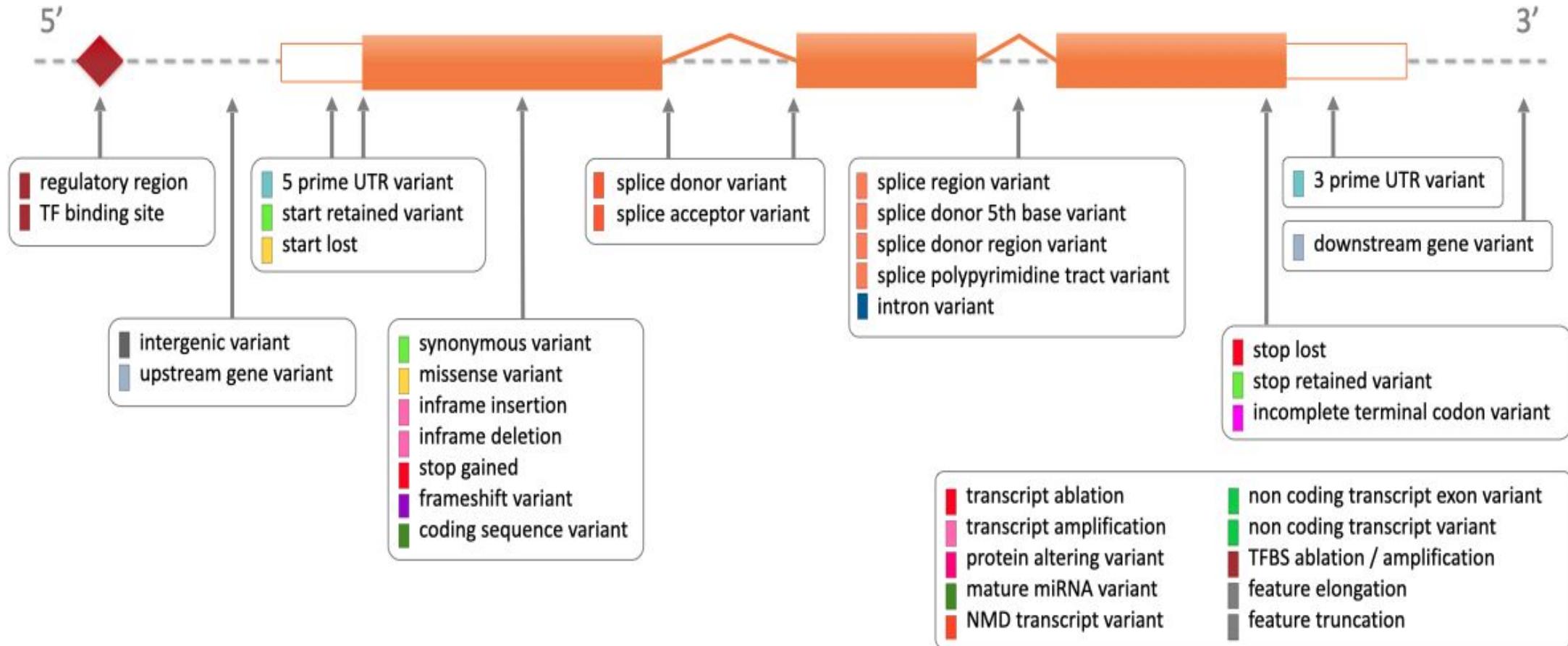
Structural variants

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: 

Ensembl Variation - Variant classification (2)



Ensembl Variation - Calculated variant consequences (1)



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

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

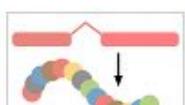
Variant data can be viewed in the browser through pages such as:

- **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"](#).

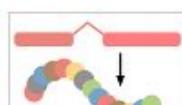
Examples:



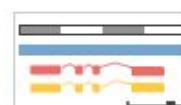
Gene
Variation Table



Transcript
Variation Table



Transcript
Protein Variation



Location
Region in detail



Phenotype
Karyotype view

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

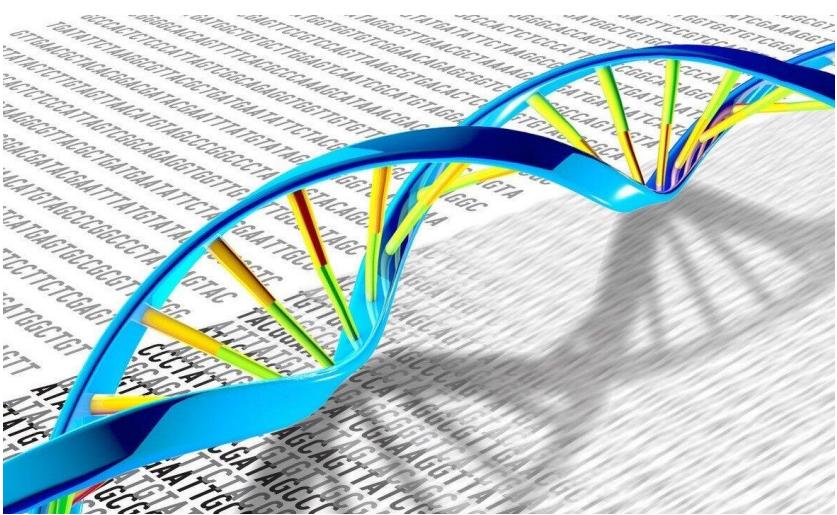
Evidence status

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

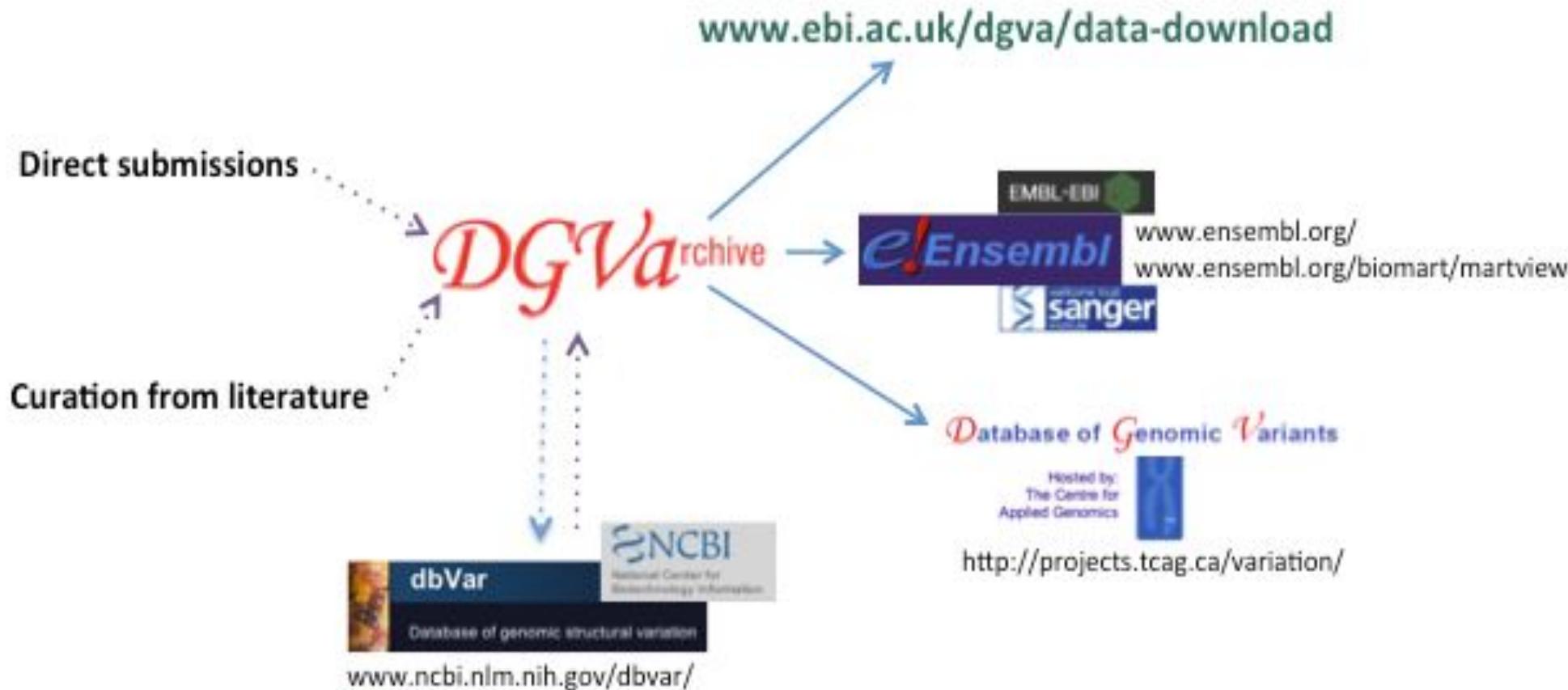
Icon	Name	Description
	Multiple observations	The variant has multiple independent dbSNP submissions, i.e. submissions with a different submitter handles or different discovery samples.
	Frequency	The variant is reported to be polymorphic in at least one sample.
	Cited	The variant is cited in a PubMed 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

Overview of variant databases



The Database of Genomic Variants archive (DGVa)



Frequency of variants databases

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

Functional prediction of variants databases

Whole-genome data	Whole-exome data	Splice variants
Gerp++	dbnsfp	dbscsv
Cadd		Spidex
Dann		
Fathmm		
Eigen		
Gwava		

Disease specific variants related to clinical

Disease specific 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)	

Ensembl database

e! EnsemblGenomes Providing genome data for non-vertebrate species, with tools for the manipulation, analysis and visualisation of that data [Contact us](#)

 [Latest release notes, updates & news from our blog](#)

e! Ensembl COVID-19

 [SARS-CoV-2 Genome sequence & annotation data](#) [Go](#)

e! Ensembl Rapid Release

2-weekly releases of new assemblies with gene & protein feature annotation [Go](#)

e! EnsemblPlants

 [Triticum aestivum](#)
IWGSC

 [Oryza sativa Japonica Group](#)
IRGSP-1.0

 [Arabidopsis thaliana](#)
TAIR10

[Go to Ensembl Plants](#)

e! EnsemblMetazoa

 [Caenorhabditis elegans](#)
WBcel235

 [Drosophila melanogaster](#)
BDGP6.28

 [Bombyx mori](#)
ASM15162v1

[Go to Ensembl Metazoa](#)

e! EnsemblProtists

 [Plasmodium falciparum 3D7](#)
ASM276v2

 [Dictyostelium discoideum](#)
dicty_2.7

 [Phytophthora infestans](#)
ASM14294v1

[Go to Ensembl Protists](#)

e! EnsemblFungi

 [Magnaporthe oryzae](#)
MG8

 [Saccharomyces cerevisiae](#)
R64-1-1

 [Aspergillus nidulans](#)
ASM1142v1

[Go to Ensembl Fungi](#)

e! EnsemblBacteria

[Streptococcus pneumoniae](#)
ASM688v1

[Escherichia coli](#)
ASM584v2

[Bacillus subtilis](#)
ASM73511v1

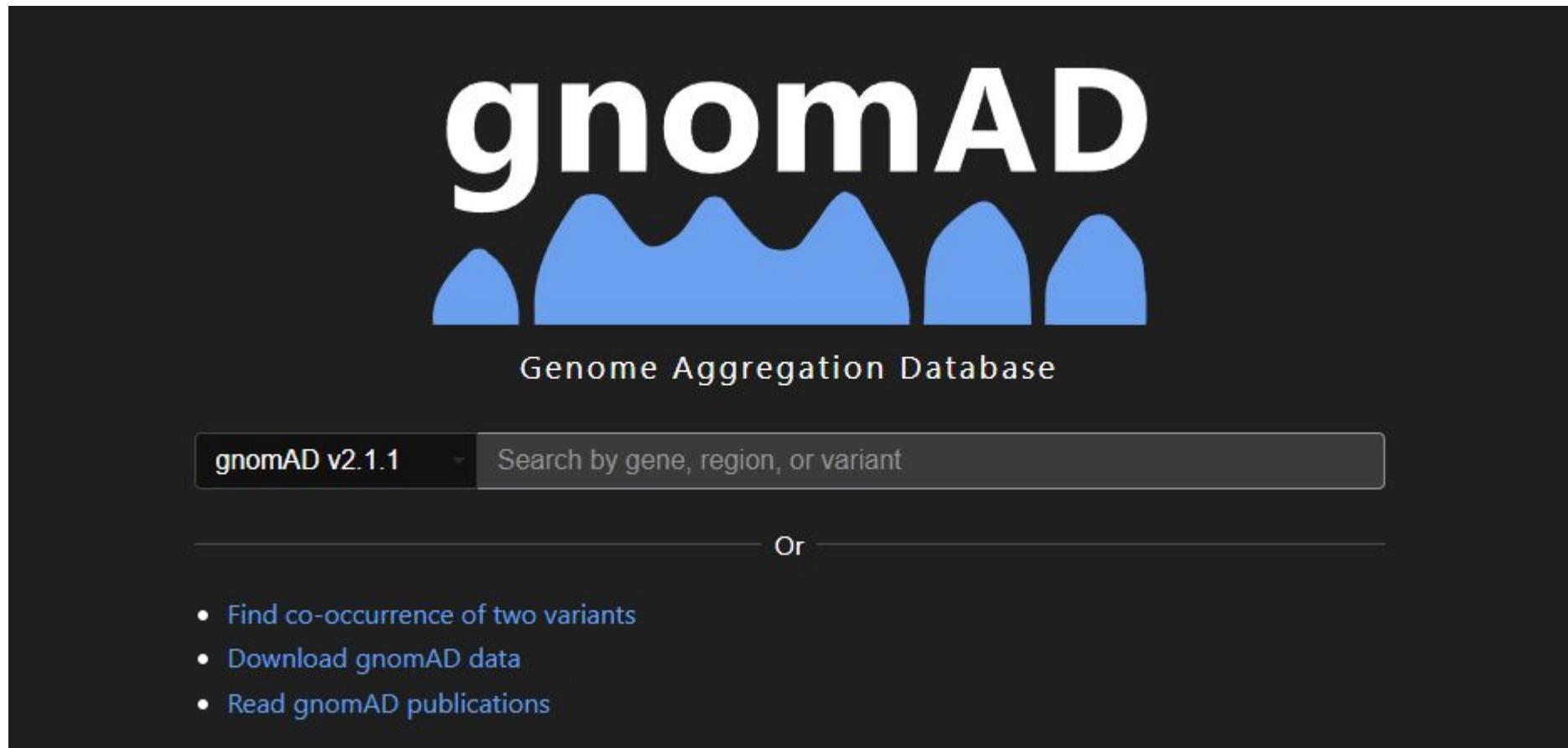
[Go to Ensembl Bacteria](#)

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.



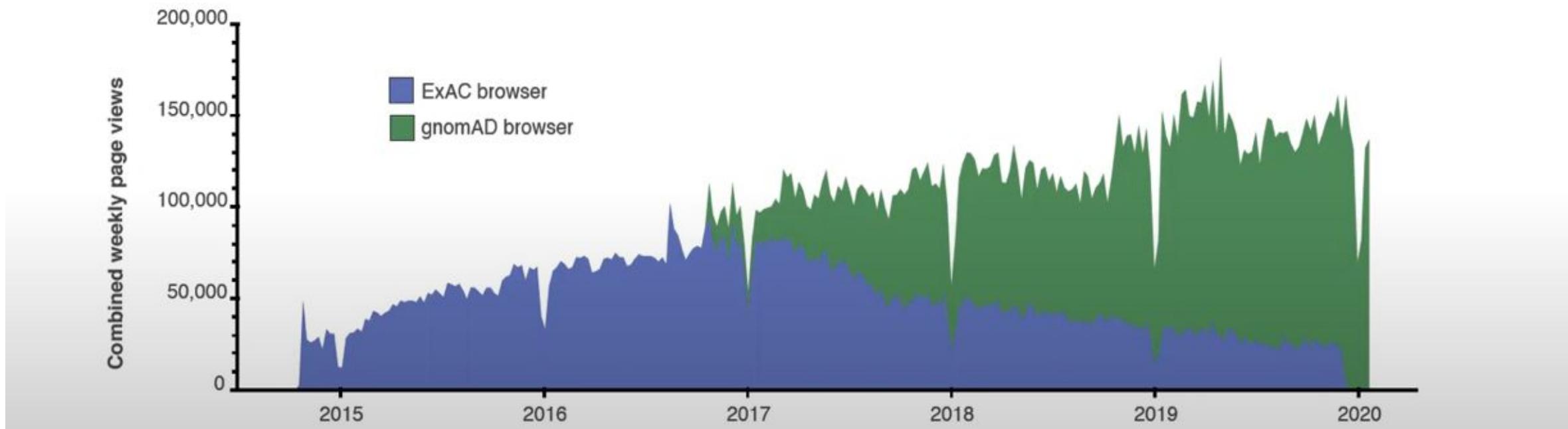
<https://gnomad.broadinstitute.org/>

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

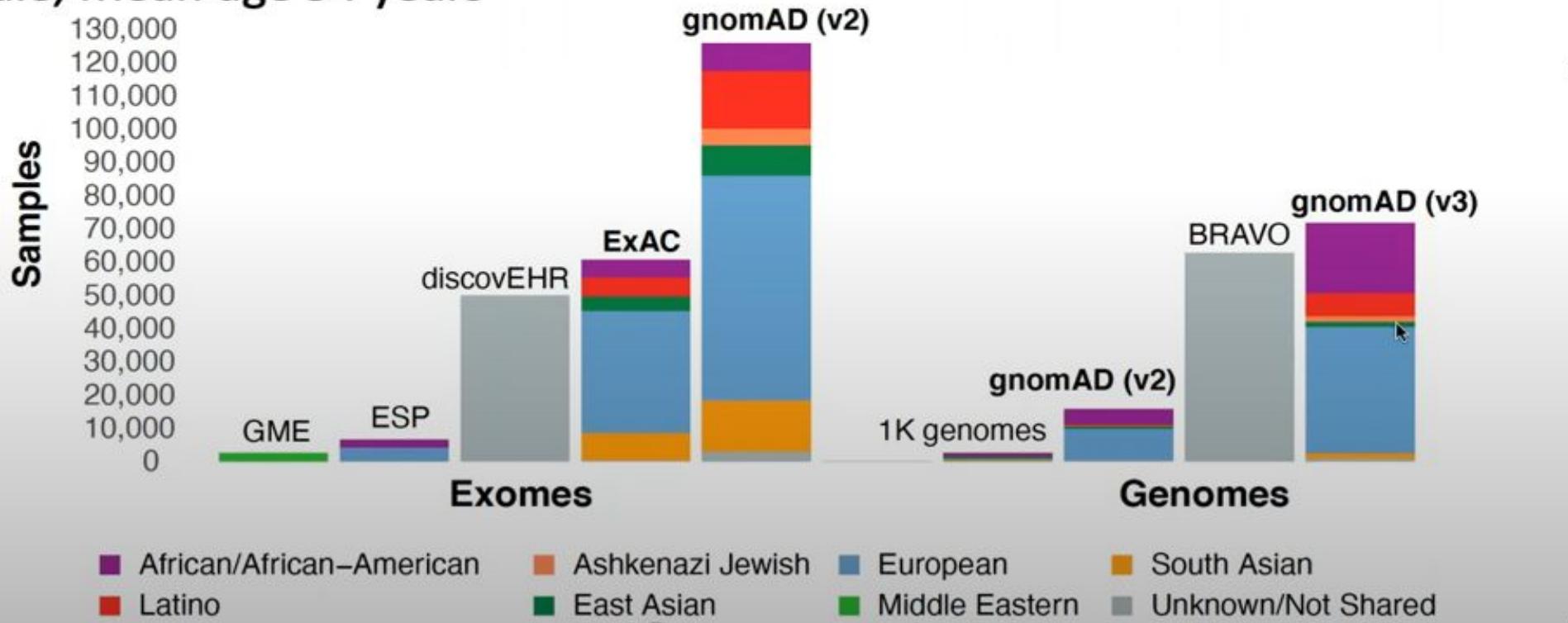
Genome Aggregation Database (gnomAD) – v3
71,702 genomes
Genome Build 38
released October 2019



- 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

gnomAD browser gnomAD v3.1.2 Search About Team News Changelog Downloads Policies Publications Feedback Help

PIK3CA phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha

Dataset gnomAD v3.1.2 ▾ gnomAD SVs v2.1 ▾ ⓘ

Genome build GRCh38 / hg38
Ensembl gene ID ENSG00000121879.6
MANE Select transcript ⓘ ENST00000263967.4 / NM_006218.4
Ensembl canonical transcript ⓘ ENST00000263967.4
Other transcripts ENST00000468036.1, ENST00000675467.1, and 7 more
Region 3:179148114-179240093
External resources Ensembl, UCSC Browser, and more

Viewing full gene. Zoom in

Constraint ⓘ Variant co-occurrence ⓘ
Constraint not yet available for gnomAD v3.

Per-base mean depth of coverage

Include: Coding regions (CDS) Untranslated regions (UTRs) Non-coding transcripts

Show transcripts

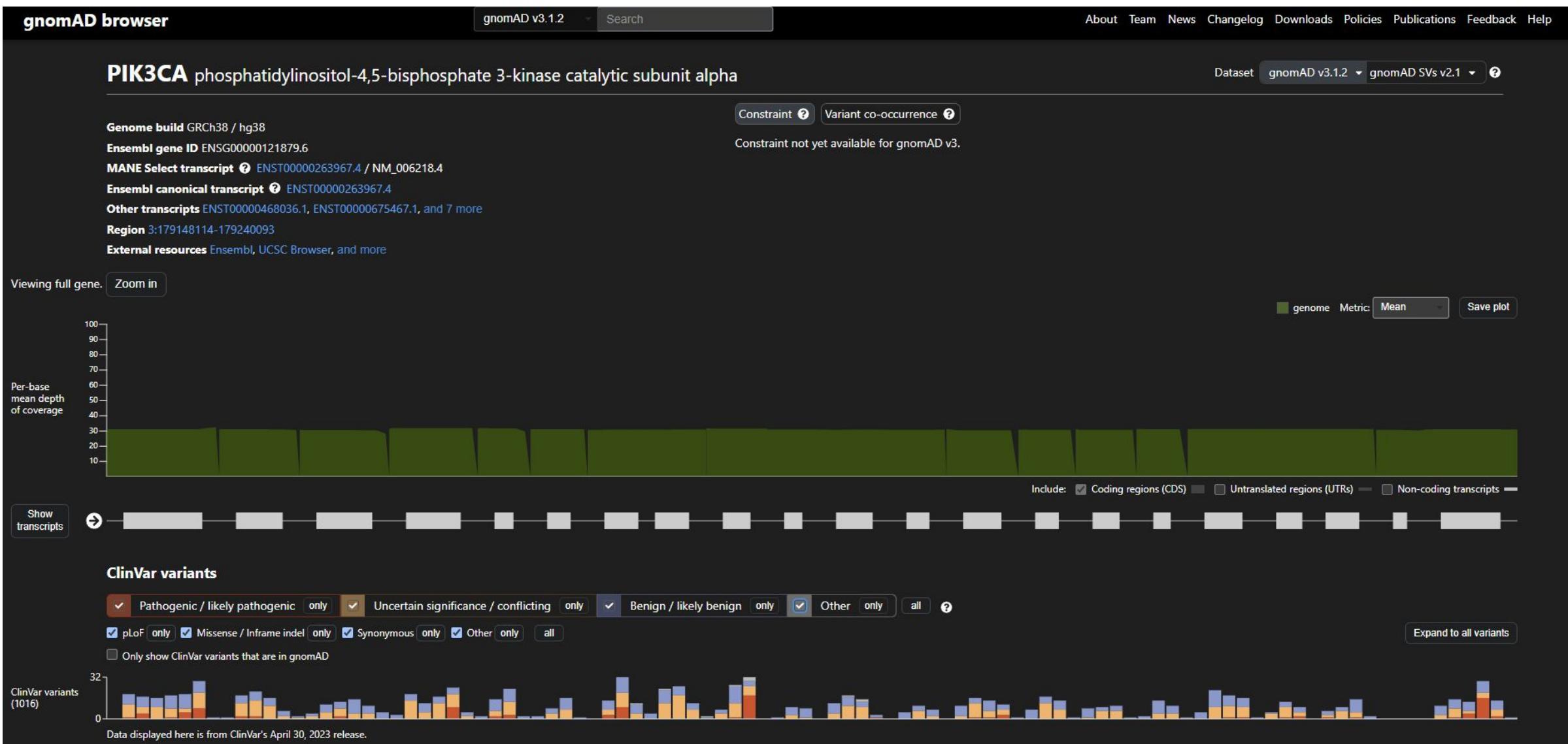
ClinVar variants

Pathogenic / likely pathogenic only Uncertain significance / conflicting only Benign / likely benign only Other only all ⓘ
 pLoF Missense / Inframe indel only Synonymous only Other only all
 Only show ClinVar variants that are in gnomAD

Expand to all variants

ClinVar variants (1016)

Data displayed here is from ClinVar's April 30, 2023 release.



A screenshot of the gnomAD browser interface for the PIK3CA gene. The top navigation bar includes links for About, Team, News, Changelog, Downloads, Policies, Publications, Feedback, and Help. Below the title, the dataset is set to gnomAD v3.1.2 and gnomAD SVs v2.1. The main content area shows the gene name 'PIK3CA' and its description as 'phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha'. It provides genome build information (GRCh38/hg38), Ensembl gene ID (ENSG00000121879.6), and various transcript details. A constraint section indicates that constraint data is not yet available for gnomAD v3. The central feature is a coverage plot showing 'Per-base mean depth of coverage' across the gene region (3:179148114-179240093). Below the plot, a transcript track shows the genomic structure with exons and introns. A 'ClinVar variants' section allows filtering by clinical significance (Pathogenic/Likely Pathogenic, Uncertain, Benign/Likely Benign, Other) and variant types (pLoF, Missense/Inframe indel, Synonymous, Other). A summary bar chart at the bottom shows the count of ClinVar variants (1016) across the gene. A note at the bottom states that the data is from ClinVar's April 30, 2023 release.

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.



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.

PMCID: [PMC7497289](#)

Published online 2020 Jun 28. doi: [10.1007/s00439-020-02199-3](https://doi.org/10.1007/s00439-020-02199-3)

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. Cooper¹

► Author information ► Article notes ► Copyright and License information [Disclaimer](#)

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).

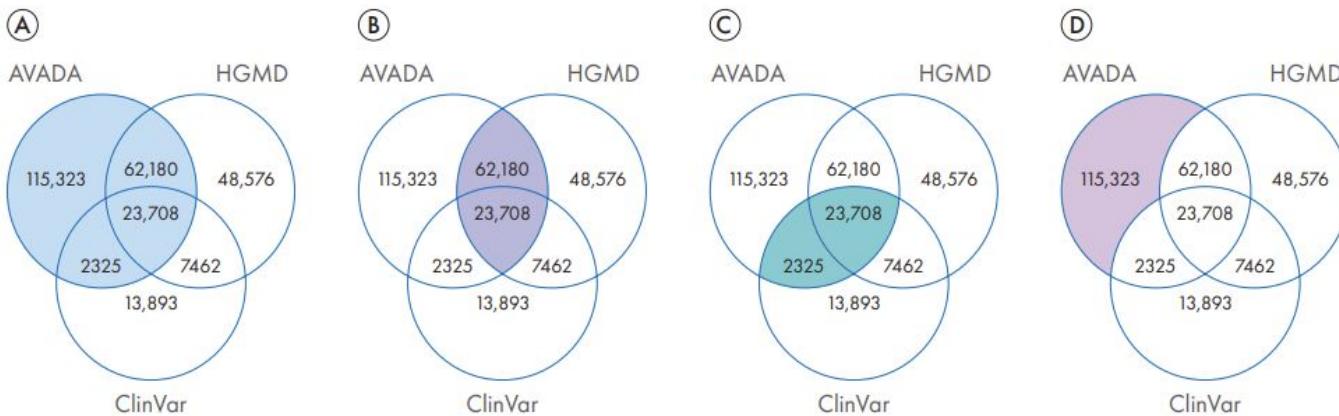


Figure 1. A) 203,536 total variants were automatically extracted by AVADA; B) 61% of variants reported as disease-causing in the expert-curated HGMD were found by AVADA; C) 55% of variants reported as likely/pathogenic in the ClinVar were also found in AVADA; D) 56% of variants retrieved by AVADA were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

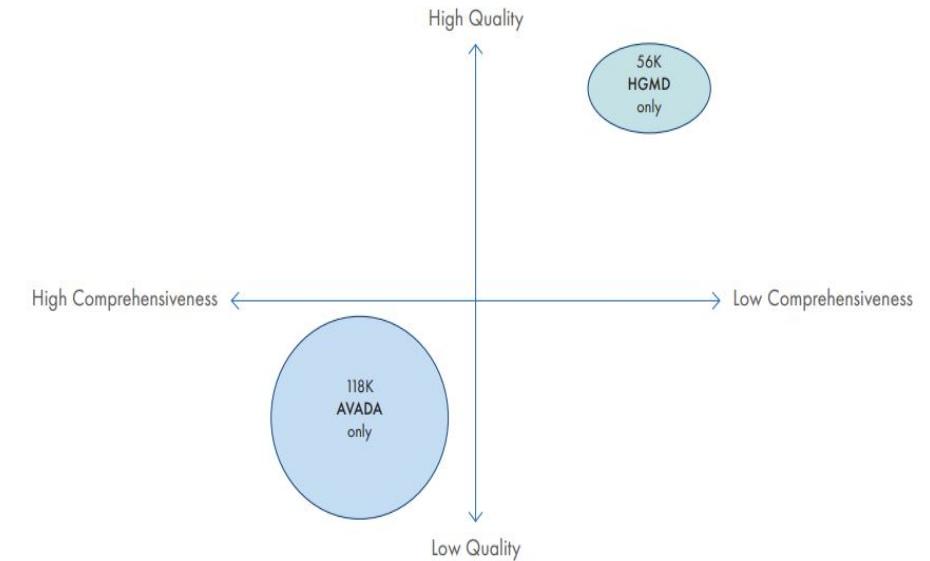


Figure 2. Manual inspection of randomly selected AVADA variants showed lower quality and comprehensiveness compared to expert-curated, high-quality HGMD data even though twice as much data was automatically extracted.

[https://digitalinsights.qiagen.com/products-overview/clinical-i
nsights-portfolio/human-gene-mutation-database/](https://digitalinsights.qiagen.com/products-overview/clinical-insights-portfolio/human-gene-mutation-database/)

Mastermind database

MASTERMIND®

Search Add genes, variants, etc... SEARCH Gene BRCA1 Variant BRCA1:p.E23fs

Genomenon 62.7K ClinVar 14.3K

90,755 Articles

BRCA1
BRCA1 DNA repair associated
protein-coding

36124 Variants 33 Variants All Records 3 Diseases

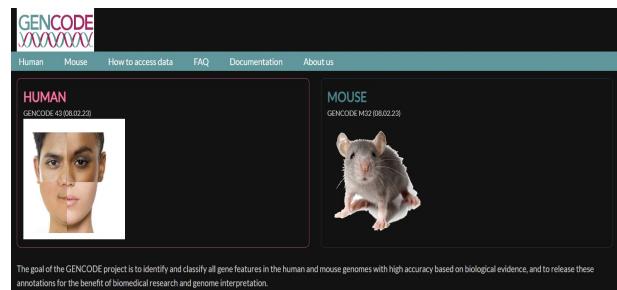
PRO EDITION FEATURE Genomenon has disease classifications for 30.0K gene variants Sign up for full access

Missense 36124
Synonymous 5217
Truncating 9673
In-Frame 1543
Splice 1130
Non-Coding 9049

Variant	Type	cDNA Position	Total Articles
p.M18T +	Missense	c.52, c.53, c.54 ▾	87
p.M18K +	Missense	c.52, c.53, c.54 ▾	29
p.A18V +	Missense	c.52, c.53, c.54 ▾	17
p.M18R +	Missense	c.52, c.53, c.54 ▾	17
p.A18fs +	Frameshift	c.52, c.53, c.54 ▾	14
p.M18I +	Missense	c.52, c.53, c.54 ▾	14
p.M18fs +	Frameshift	c.52, c.53, c.54 ▾	14

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		
- pseudogenes	15	Total No of distinct translations	65519
Immunoglobulin/T-cell receptor gene segments		Genes that have more than one distinct translations	13618
- protein coding segments	410		
- pseudogenes	236		

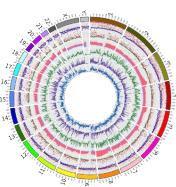
GTF data format

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

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

ClinVar database

```
ACTGATGGTATGGGCCAAGAGATATATCT  
CAGGTACGGCTGTCACTCACTTAGACCTCAC  
CAGGGCTGGGCATAAAAGTCAGGGCAGAGC  
CCATGGTGCATCTGACTCCTGAGGAGAAGT  
GCAGGTTGGTATCAAGGTTACAAGACAGGT  
GGCACTGACTCTCTGCCTATTGGTCTAT
```

ClinVar

- 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

Clinvar publicity

1.3 million
submitted records

Submitted by
>1800 organizations

From
80 countries

ClinVar is a global database



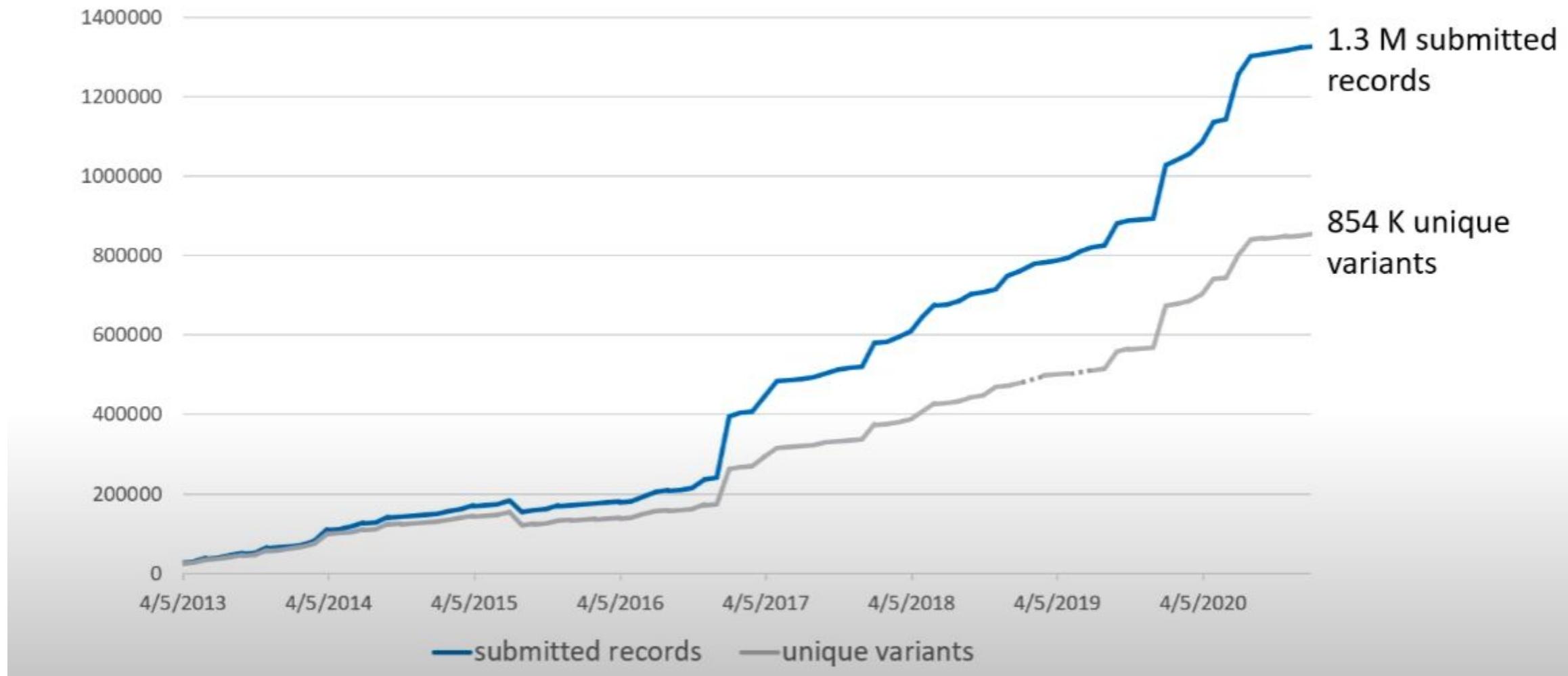
U.S. National Library of Medicine
National Center for Biotechnology Information



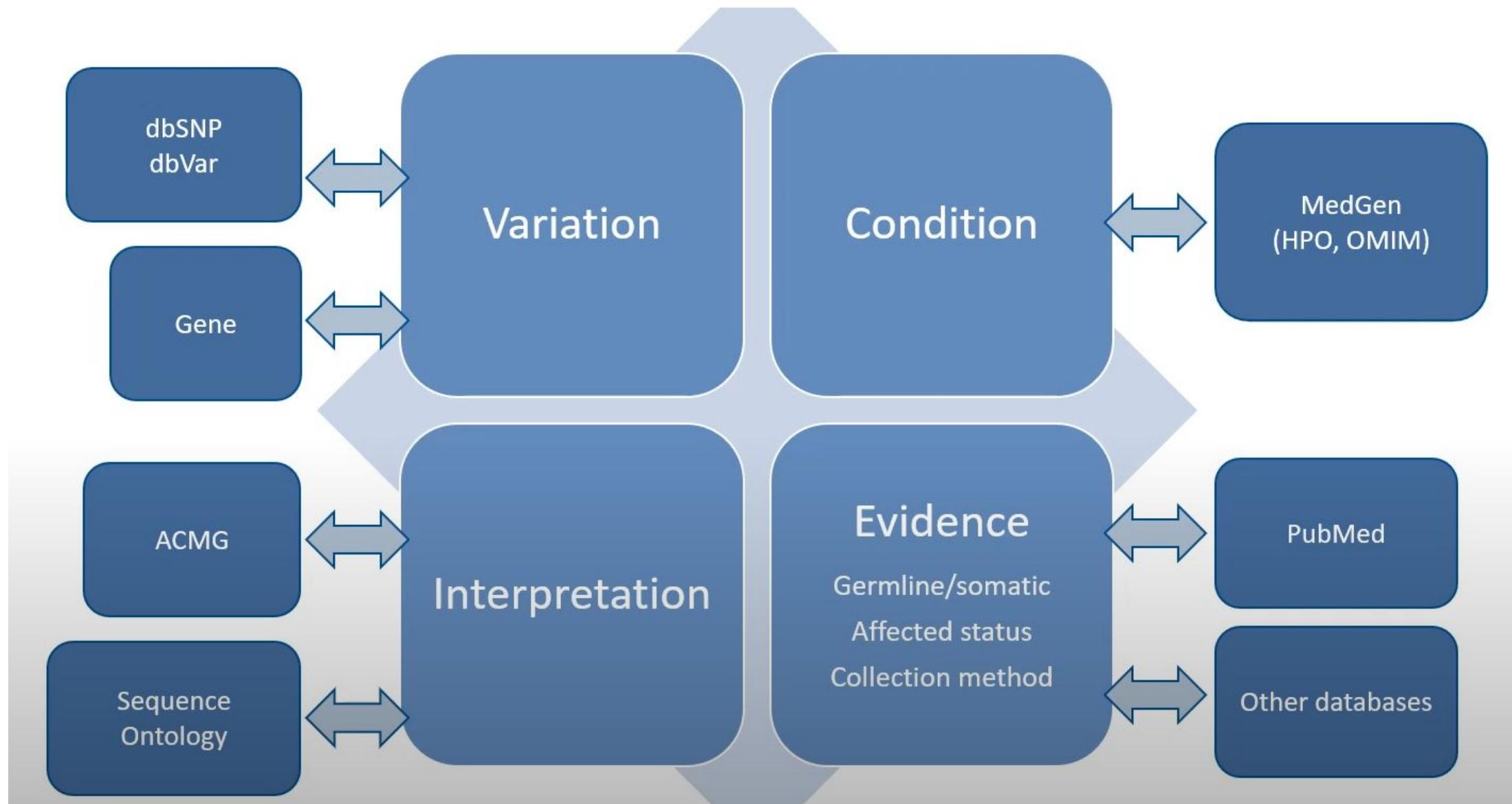
NCBI

ClinVar evolution

Growth of data in ClinVar



ClinVar integrates 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



FTP



E-utilities



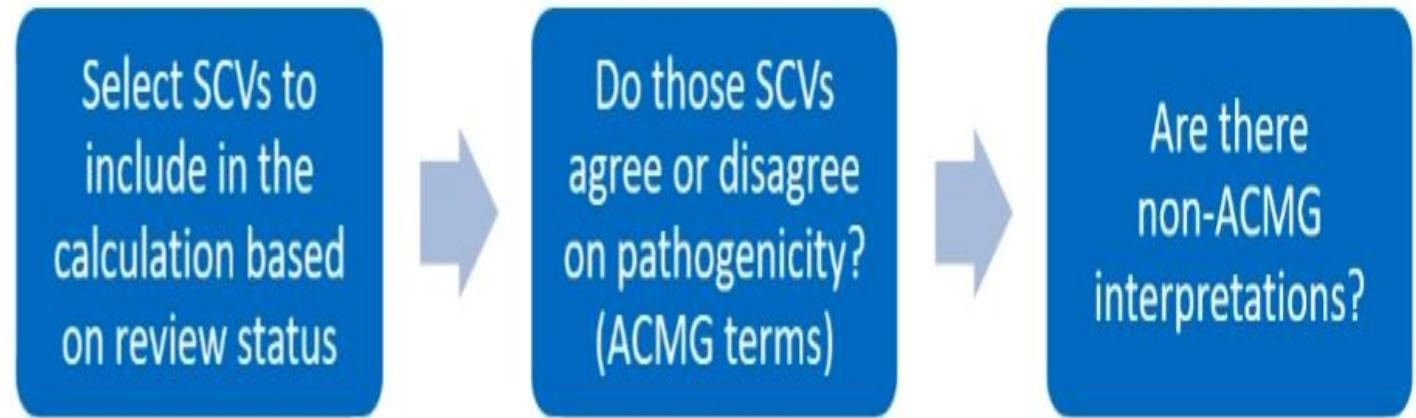
Web site

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

Calculating aggregate interpretation

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



Precedence:

1. Practice guidelines
2. Expert panel
3. criteria provided, single submitter
4. no assertion criteria provided

Examples:

1. Pathogenic
2. Pathogenic/Likely pathogenic
3. Conflicting interpretations of pathogenicity

Examples:

1. Pathogenic, drug response
2. Pathogenic/Likely pathogenic, risk factor

The ClinVar variation report sample

Search ClinVar
Advanced search

FEEDBACK

NM_198056.2(SCN5A):c.4478A>G (p.Lys1493Arg)

Cite this record



NM_198056.2(SCN5A):c.4478A>G (p.Lys1493Arg)

A

Cite this record

Interpretation:

Conflicting interpretations of pathogenicity
Likely pathogenic(1);Uncertain significance(2)

★☆☆☆ criteria provided, conflicting interpretations

4 (Most recent: Jul 30, 2018)

Nov 21, 2017

VCV000067898.1

67898

single nucleotide variant

FEEDBACK

Conflicting interpretations of pathogenicity
Likely pathogenic(1);Uncertain significance(2)

★☆☆☆ criteria provided, conflicting interpretations
4 (Most recent: Jul 30, 2018)

Nov 21, 2017

VCV000067898.1

67898

single nucleotide variant

Variant details

Conditions

Gene(s)

NM_198056.2(SCN5A):c.4478A>G (p.Lys1493Arg)

Allele ID: 78791

Variant type: single nucleotide variant

Variant length: 1 bp

Cytogenetic location: 3p22.2

Genomic location: 3: 38555720 (GRCh38) GRCh38 UCSC

3: 38597211 (GRCh37) GRCh37 UCSC

HGVS:

Nucleotide	Protein	Molecular consequence
NC_000003.11:g.38597211T>C		
NC_000003.12:g.38555720T>C		
LRG_289t1:c.4478A>G	LRG_289p1:p.Lys1493Arg	

... more HGVS

Protein change: K1492R

Other names: -

Functional consequence: -

Global minor allele frequency (GMAF): -

Allele frequency: The Genome Aggregation Database (gnomAD), exomes 0.00001

Exome Aggregation Consortium (ExAC) 0.00002

The Genome Aggregation Database (gnomAD) 0.00003

Trans-Omics for Precision Medicine (TOPMed) 0.00006

Links: UniProtKB: Q14524#VAR_074742

dbSNP: rs199473260

B

C

Variant details

Conditions

Gene(s)

Aggregate interpretations per condition

Interpreted condition	Interpretation	Number of submissions	Review status	Last evaluated	Variation/condition record
Atrial fibrillation	Likely pathogenic	1	criteria provided, single submitter	Jun 24, 2013	RCV000171569.1
Cardiovascular phenotype	Uncertain significance	1	criteria provided, single submitter	Sep 25, 2016	RCV000619395.1
Brugada syndrome	Uncertain significance	1	criteria provided, single submitter	Nov 21, 2017	RCV000638673.1
Congenital long QT syndrome	not provided	1	no assertion provided	-	RCV000058678.3

Print Download

NM_198056.2(SCN5A):c.4478A>G (p.Lys1493Arg)

Cite this record



FEEDBACK

Conflicting interpretations of pathogenicity
Likely pathogenic(1);Uncertain significance(2)

★☆☆☆ criteria provided, conflicting interpretations

4 (Most recent: Jul 30, 2018)

Nov 21, 2017

VCV000067898.1

67898

single nucleotide variant

Variant details

Conditions

Gene(s)

Gene	OMIM	ClinGen Gene Dosage Sensitivity Curation		Variation viewer	Related variants
		HI score	TS score		
SCN5A		Some evidence for dosage pathogenicity	No evidence available	GRCh38 GRCh37	1362 1507

Print Download

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 SCNSA 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)

Review status in ClinVar

ClinVar

Genomic variation as it relates to human health

Search by gene symbols, location, HGVS expressions, c-dot, p-dot, conditions, and more

Search ClinVar



Advanced search

About

Access

Submit

Stats

FTP

Help

Were new search queries using location, c-dot, and p-dot helpful?



Follow



Print

Download

NM_000314.8(PTEN):c.139A>G (p.Arg47Gly)

Cite this record



Interpretation: Pathogenic

Review status: reviewed by expert panel FDA RECOGNIZED DATABASE

Submissions: 3

First in ClinVar: May 28, 2018

Most recent Submission: Oct 1, 2022

Last evaluated: Jun 18, 2020

Accession: VCV000189401.9

Variation ID: 189401

Description: single nucleotide variant

Variation Viewer

Variation Viewer

Share this page

Reset All

More Info



New to Variation Viewer? Read our quick overview! [X](#)

Homo sapiens
(human)

Assembly: GRCh38.p14 (GCF_000001405.40) • Chr 1 (NC_000001.11)

Search assembly

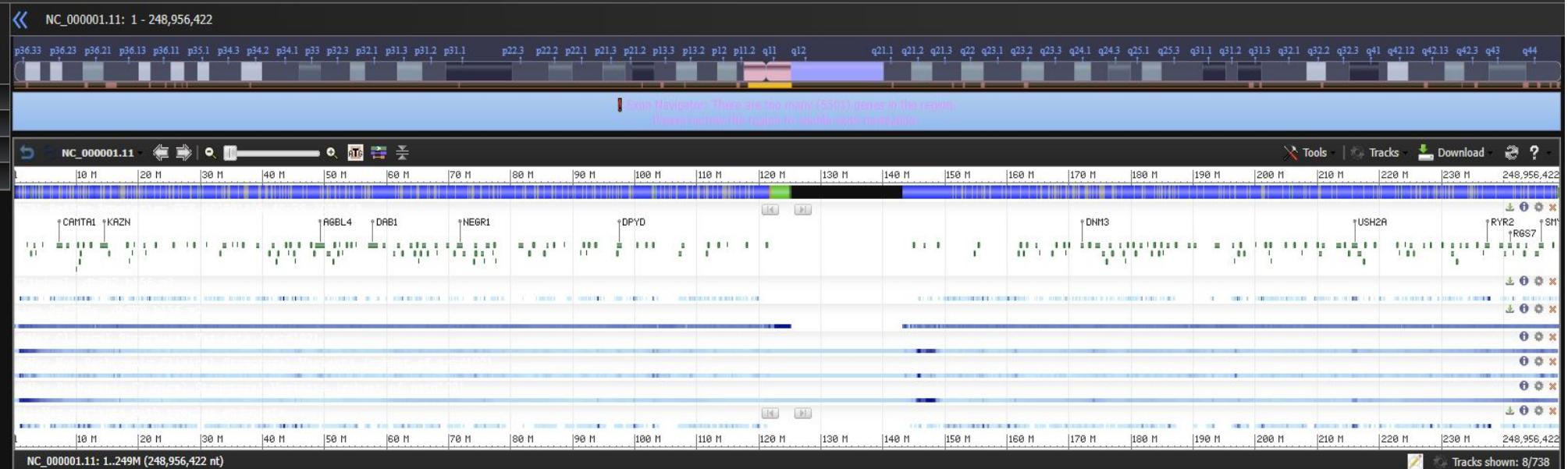
Location, gene or phenotype [Examples](#)

Pick Assembly

Tracks and User Data

History

Assembly Region Details



Variation Data

Filter by	Variant ID	Location	Variant type	Gene	Molecular consequences	Most severe clinical significance	1000G MAF	GO-ESP MAF	ExAC MAF	Publications
Source database	nsv6138160	10,000 - 52,000	copy number variation	WASH7P	and 6 more					1
<input type="checkbox"/> dbSNP (86844566)	nsv945698	10,001 - 22,118	copy number variation	WASH7P	and 2 more					1
<input type="checkbox"/> dbVar (587059)	nsv4398922	10,001 - 127,330	copy number variation	LOC100996442	and 10 more					
In ClinVar	nsv482937	10,001 - 2,368,561	copy number variation	LOC100996442	and 139 more					1
<input type="checkbox"/> Yes (118988)	nsv482957	10,001 - 15,873,505	copy number variation	KAZN	and 470 more					1
<input type="checkbox"/> No (87312637)	nsv6299964	10,001 - 52,000	copy number variation	WASH7P	and 6 more					
	nsv871113	10,001 - 782,105	copy number variation	LOC100996442	and 43 more					1

LitVar database

LitVar²

BRCA1



BRCA2



NIH NLM

YEARS



SECTIONS

<input type="checkbox"/> ABSTRACT	131
<input type="checkbox"/> RESULTS	76
<input type="checkbox"/> DISCUSSION	73
<input type="checkbox"/> TABLE	61
<input type="checkbox"/> INTRODUCTION	56
<input type="checkbox"/> METHODS	54
<input type="checkbox"/> OTHER	50
<input type="checkbox"/> FIGURE LEGEND	9
<input type="checkbox"/> CONCLUSION	4
<input type="checkbox"/> TITLE	4

TOP JOURNALS

<input type="checkbox"/> HERED CANCER CLIN PRACT	29
<input type="checkbox"/> Br J CANCER	18
<input type="checkbox"/> PLoS ONE	16
<input type="checkbox"/> BMC CANCER	15
<input type="checkbox"/> BREAST CANCER Res	10

Showing 1 to 10 of 410 publications.

SORT BY RELEVANCE RECENCY

◀ Page 1 of 41 ▶

PMID37731153 • PMC10673778

Dec 1, 2023

The Landscape of BRCA Mutations among Egyptian Women with Breast Cancer

Azim HA, Loutfy SA ... Kassem L • Oncol Ther

TABLE

No text snippet can be shown as the variant **5382insC** is found in the [Table](#) content.

[Share](#) [Cite](#)

PMID38203374 • PMC10778779

Dec 22, 2023

BRCA Mutations and Fertility Preservation

Dias Nunes J, Demeestere I, Devos M • Int J Mol Sci

OTHER

Notably, those germline mutations were more commonly detected in Ashkenazi Jews population, among whom pathogenic variants are present in approximately 2.5% of the individuals, of which **185delAG** and **5382insC** in BRCA1, and **6174delT** in **BRCA2**.

[Share](#) [Cite](#)

PMID36845707 • PMC9950741

Feb 10, 2023

The cancer-risk variant frequency among Polish population reported by the first national whole-genome sequencing study.

Mroczek M, Liu J ... Dobosz P • Front Oncol

INTRODUCTION

VARIANT

c.5382insC BRCA1

[Download](#) [RSS](#)

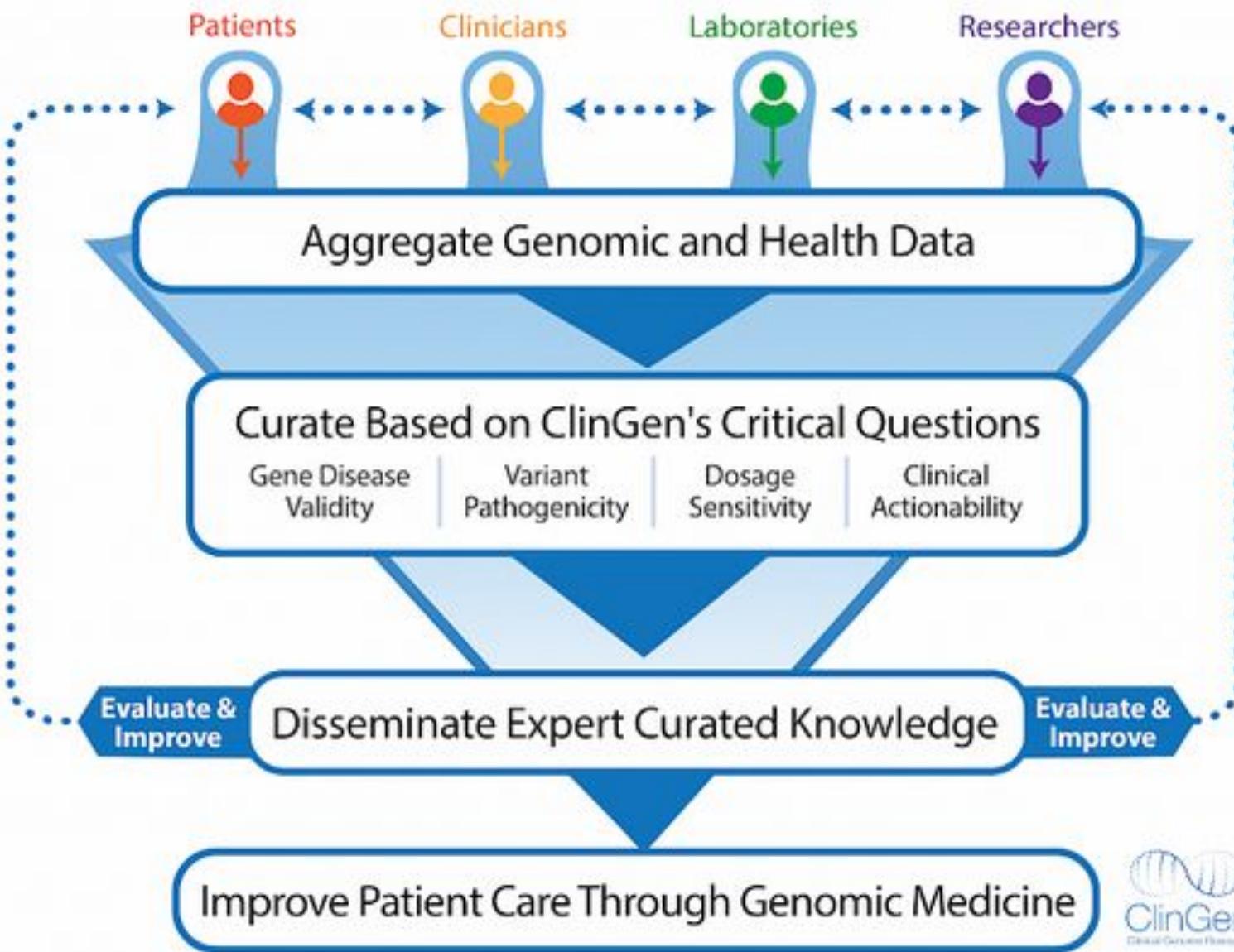
BIOCONCEPTS

GENE DISEASE

CHEMICAL VARIANT

SPECIES CELLINE

Clinical Genome Resource (ClinGen)



ClinGen database

ClinGen Curation Summary Statistics

2756

Unique genes with
at least one curation 

6531

Unique variants with
at least one curation



Gene-Disease Validity

2512

Total reports
(Number of curations
for this activity)

2061

Unique genes
(Total genes with at
least one curation)



Dosage Sensitivity

3879

Total reports
(Number of curations
for this activity)

1561

Unique genes
(Total genes with at
least one curation)



Clinical Actionability

243

Total reports
(Number of reports
for this activity)

286

Unique genes
(Total genes with at
least one report)



Variant Pathogenicity

6533

Total reports
(Number of curations
for this activity)

6531

Unique variants
(Total variants with at
least one curation)



Pharmacogenomics

689

Total reports
(Number of gene-drug pairs
for this activity)

128

Unique genes
(Total genes with at
least one gene-drug pair)

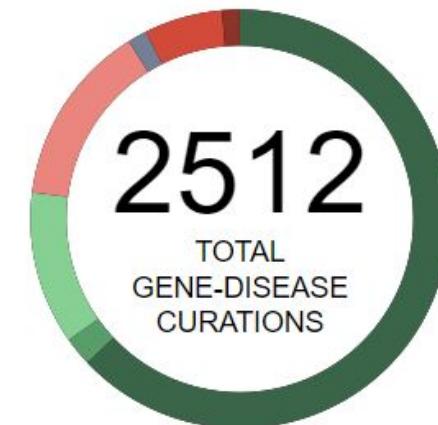
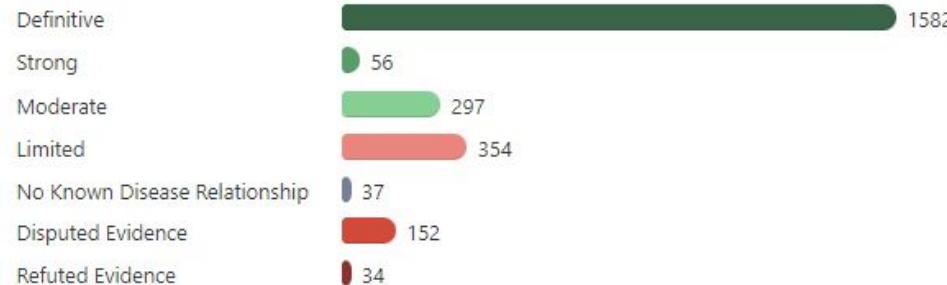


Gene-Disease Clinical Validity Statistics

The ClinGen Gene-Disease Clinical Validity curation process involves evaluating the strength of evidence supporting or refuting a claim that variation in a particular gene causes a particular disease.

Classification Statistics

Gene-Disease Clinical Validity has **2512 curations** encompassing **2061 genes**.



56 ClinGen Gene Curation Expert Panels

[Sort by Name](#) [Sort by Count](#)



Aminoacidopathy



Amyotrophic Lateral Sclerosis Spectrum Disorders



Antibody Deficiencies GCEP



Arrhythmogenic Right Ventricular Cardiomyopathy



Brain Malformations



Breast/Ovarian Cancer



Brugada Syndrome



Catecholaminergic Polymorphic Ventricular Tachycardia

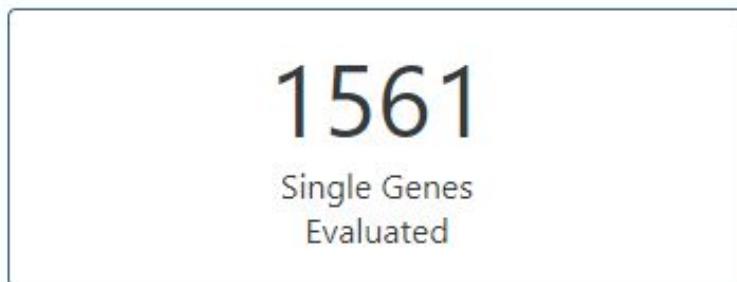
[Load more Gene Curation Expert Panels](#)



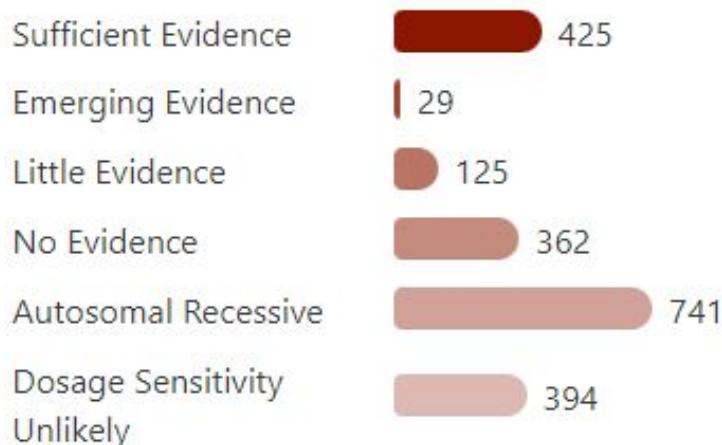
Dosage Sensitivity Statistics

The ClinGen Dosage Sensitivity curation process collects evidence supporting/refuting the haploinsufficiency and triplosensitivity of genes and genomic regions.

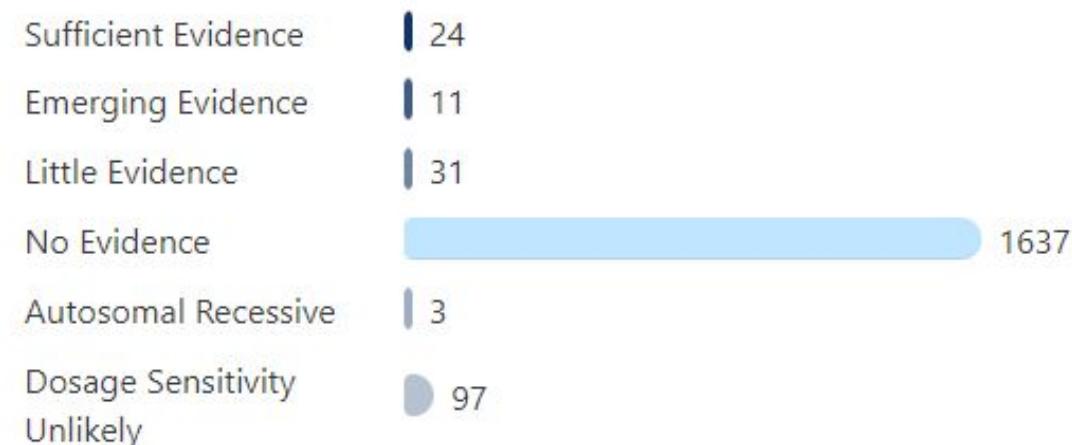
3879 Total Dosage Sensitivity Curations



Haploinsufficiency Classifications Visualized



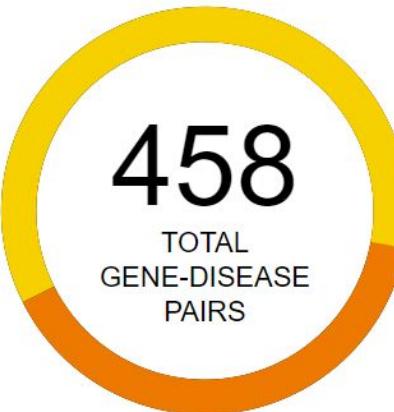
Triplosensitivity Classifications Visualized



A Clinical Actionability

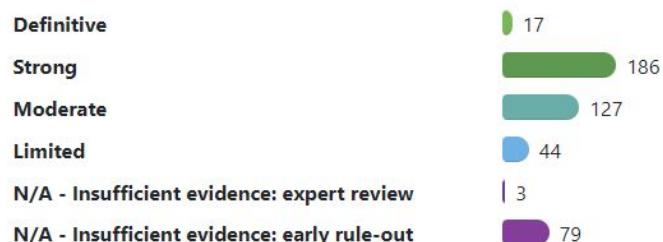
The overarching goal of the Clinical Actionability curation process is to identify those human genes that, when significantly altered, confer a high risk of serious disease that could be prevented or mitigated.

243 Total Clinical Actionability Topics



541 Total Actionability Assertions (by gene)

74 assertions under revision ⓘ

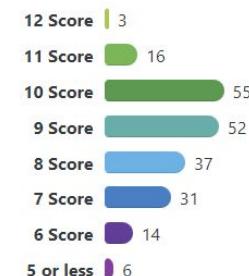


Adult Context

112 Adult Actionability Reports
28 Total Topics Failed Early Rule-out
245 Unique Genes In Adult Actionability Topics

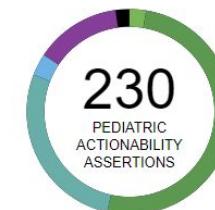


214 Total Adult Outcome-Intervention Scored Pairs



Pediatric Context

86 Pediatric Actionability Reports
17 Total Pediatric Topics Failed Early Rule-out
209 Unique Genes In Pediatric Actionability Topics



144 Total Pediatric Outcome-Intervention Scored Pairs



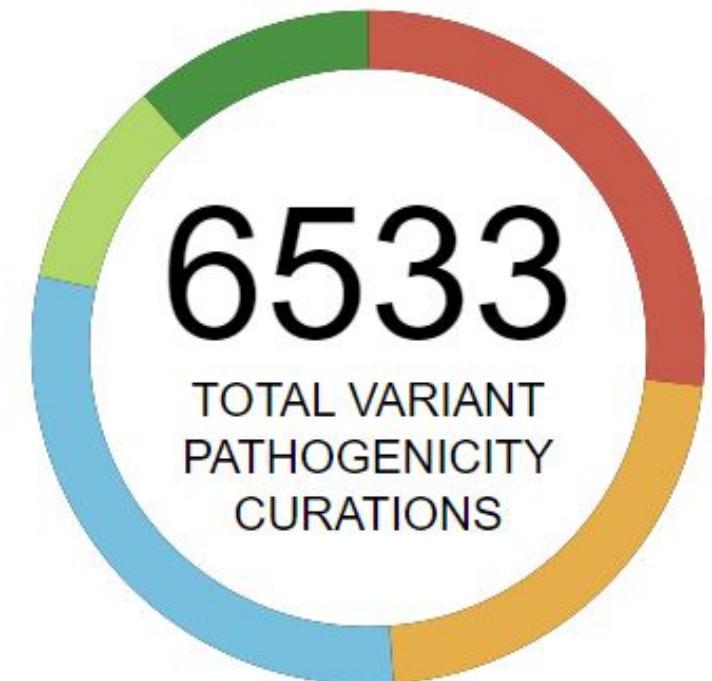


Variant Pathogenicity Statistics

ClinGen's Variant Curation Expert Panels (VCEPs) classify variants using ACMG/AMP sequence variant interpretation guidelines specified for the genes and/or diseases within their scope. These specifications are reviewed and approved as part of the ClinGen VCEP application process.

Classification Statistics

Variant Pathogenicity has **6533 curations**.



27 Approved ClinGen Variant Curation Expert Panels

(For a complete list of VCEPs at different stages of the approval process, click [here](#))

Sort by Name

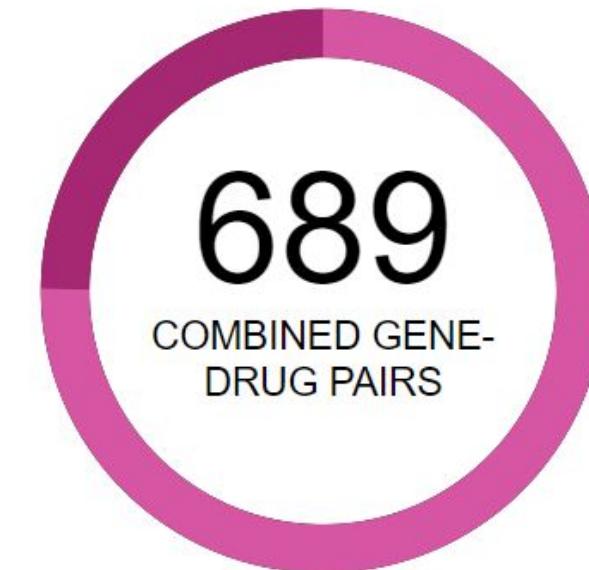
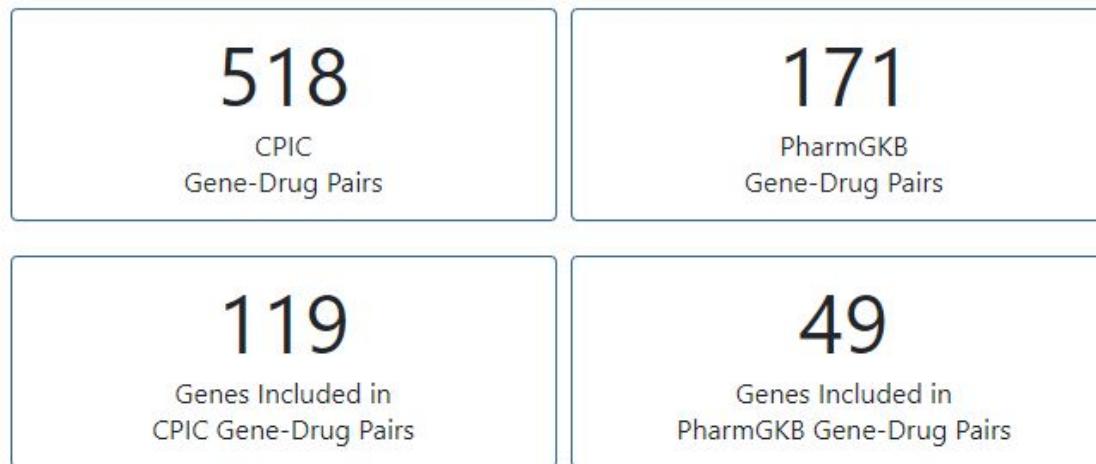
Sort by Count



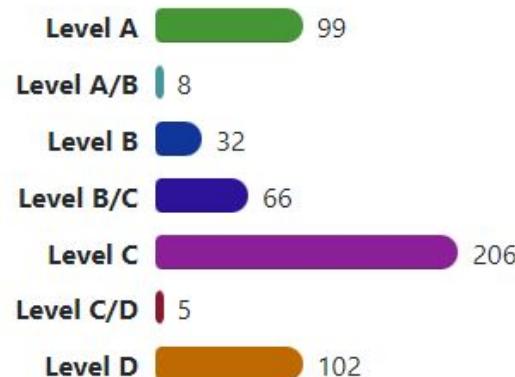
Pharmacogenomics

The overarching goal of the Pharmacogenomics is to study the variances in genes and their effects on drug response.

689 Combined Pharmacogenomics Gene-Drug Pairs



CPIC Gene-Drug Pairs by Highest Level of Actionability **Highest Levels Visualized**



PharmGKB Gene-Drug Pairs by Highest Level of Evidence **Highest Levels Visualized**



ClinGen Evidence Repository



EVIDENCE REPOSITORY

The ClinGen Evidence Repository is an **FDA-recognized human genetic variant database** containing expert-curated assertions regarding variants' pathogenicity and supporting evidence summaries. [\[Disclaimer\]](#)

Gene

[Browse Classifications by Gene](#)

Expert Panel

[Browse Classifications by Expert Panel](#)

Condition

[Browse Classifications by Condition](#)

CA114360, PTEN, Phenylketonuria, c.581T>C

Search

Download ▾

[View API docs](#)

[Browse Criteria Specifications](#)

[Browse Evidence Repository](#)

[Show Advanced Search](#)

**Classification Summary for 5517 Curated Variants
Across 89 Genes**

MedGene database (NCBI)

An official website of the United States government [Here's how you know](#) 

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

MedGen

MedGen

Limits Advanced

Search

Help

 MedGen

Organizes information related to human medical genetics, such as attributes of conditions with a genetic contribution.

Using MedGen

- MedGen Quick Start
- List of Professional Guidelines
- Select condition and phenotype terms for ClinVar and GTR
- Frequently asked questions
- Downloads/FTP
- Follow us on Twitter
-  Check out our video!
- Contact us

Tools

- GTR submission portal
- ClinVar submission portal
- Locate a genetics professional
- Genome Data Viewer
- Variation Viewer
- NHGRI Talking Glossary

Other Resources

- ClinVar
- Gene
- Genetic Testing Registry (GTR®)
- GeneReviews®
- OMIM®
- RefSeqGene
- Medical Genetics Summaries

OMIM database

OMIM®

An Online Catalog of Human Genes and Genetic Disorders

Updated August 24, 2023

Search OMIM for clinical features, phenotypes, genes, and more...



[Advanced Search](#) : OMIM, Clinical Synopses, Gene Map | Search History

[Need help?](#) : Example Searches, OMIM Search Help, OMIM Video Tutorials

[Mirror site](#) : <https://mirror.omim.org>

OMIM is supported by a grant from NHGRI, licensing fees, and generous contributions from people like you.

[Make a donation!](#)

OMIM statistics

OMIM Gene Map Statistics

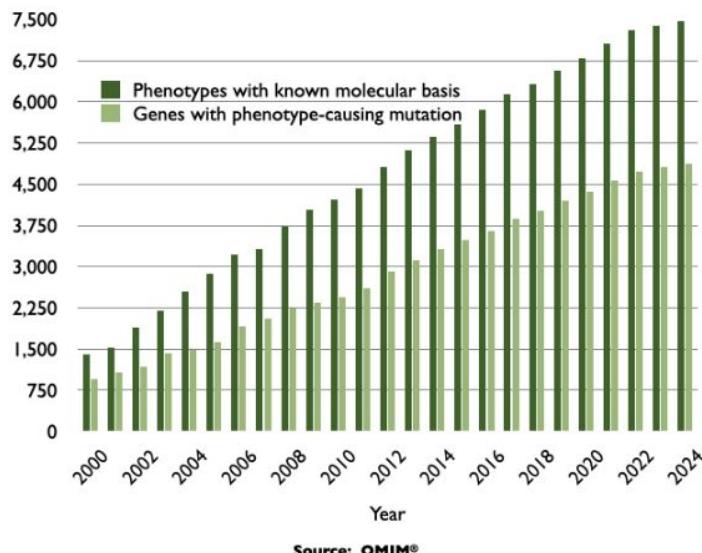
OMIM Morbid Map Scorecard (Updated March 7th, 2024) :

Total number of phenotypes* for which the molecular basis is known	7,502
Total number of genes with phenotype-causing mutation	4,893
* Phenotypes include (1) single-gene mendelian disorders and traits; (2) susceptibilities to cancer and complex disease (e.g., BRCA1 and familial breast-ovarian cancer susceptibility, 113705.0001, and CFH and macular degeneration, 134370.0008); (3) variations that lead to abnormal but benign laboratory test values ("nondiseases") and blood groups (e.g., lactate dehydrogenase B deficiency, 150100.0001 and ABO blood group system, 110300.0001); and (4) select somatic cell genetic disease (e.g., GNAS and McCune-Albright syndrome, 139320.0008 and IDH1 and glioblastoma multiforme, 147700.0001.)	

Distribution of Phenotypes across Genes (Updated March 7th, 2024) :

Number of genes with 1 phenotype	3,429
Number of genes with 2 phenotypes	891
Number of genes with 3 phenotypes	320
Number of genes with 4+ phenotypes	253

Growth of Gene-Phenotype Relationships January 2024



Dissected OMIM Morbid Map Scorecard (Updated March 7th, 2024) :

Class of phenotype	Phenotype	Gene *
Single gene disorders and traits	6,443	4,530
Susceptibility to complex disease or infection	677	500
"Nondiseases"	151	118
Somatic cell genetic disease	238	131

*Some genes may be counted more than once because mutations in a gene may cause more than one phenotype and the phenotypes may be of different classes (e.g., activating somatic BRAF mutation underlying cancer, 164757.0001, and germline BRAF mutation in Noonan syndrome, 164757.0022.)

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/

Variant annotation tools

Funcotator



SnpEff
SnpSift

Pablo Cingolani



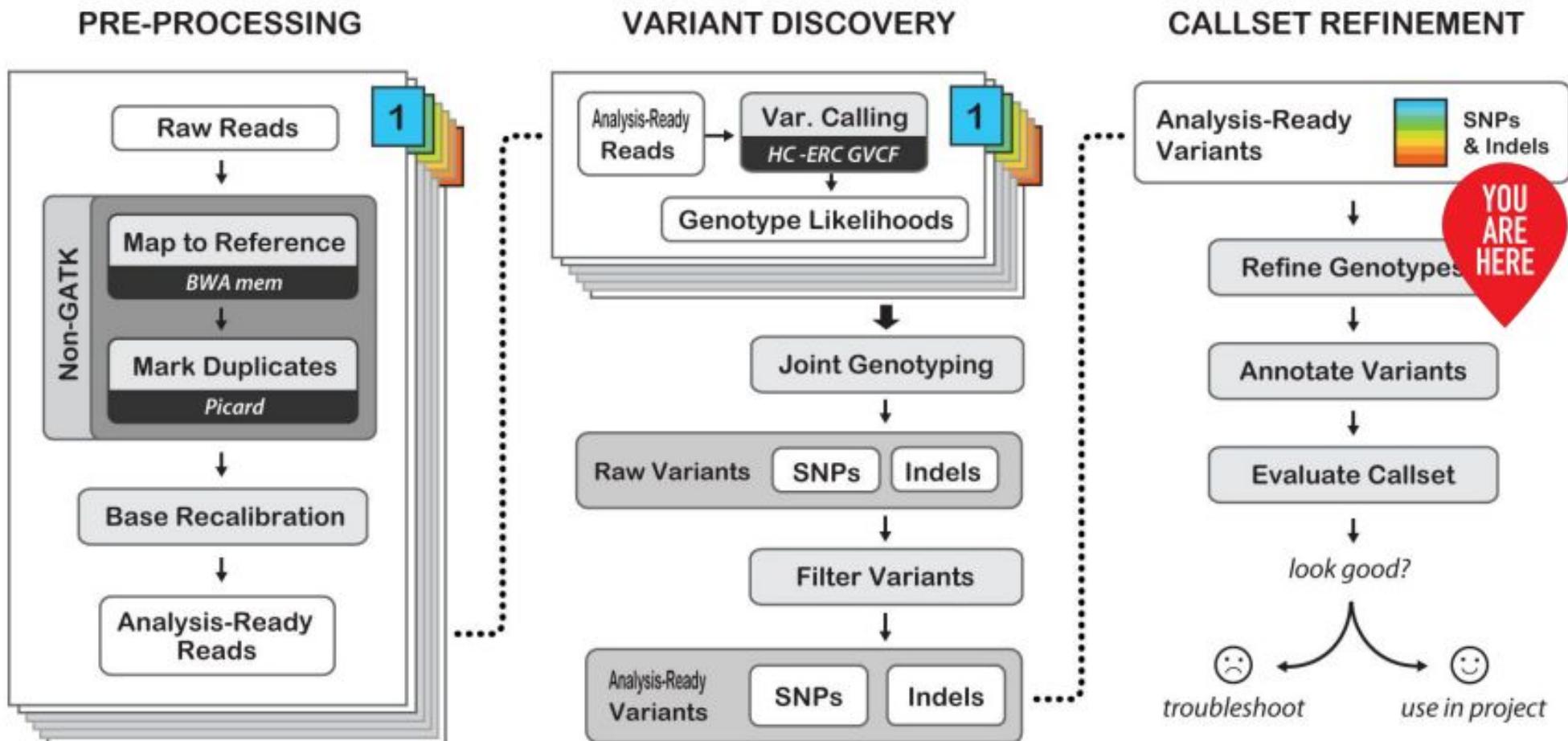
McLaren W, Gil L, Hunt SE, Riat HS,
Ritchie GR, Thormann A, Flicek P,
Cunningham F.

The Ensembl Variant Effect Predictor.
Genome Biology Jun 6;17(1):122. (2016)



Wang K, Li M, Hakonarson H.

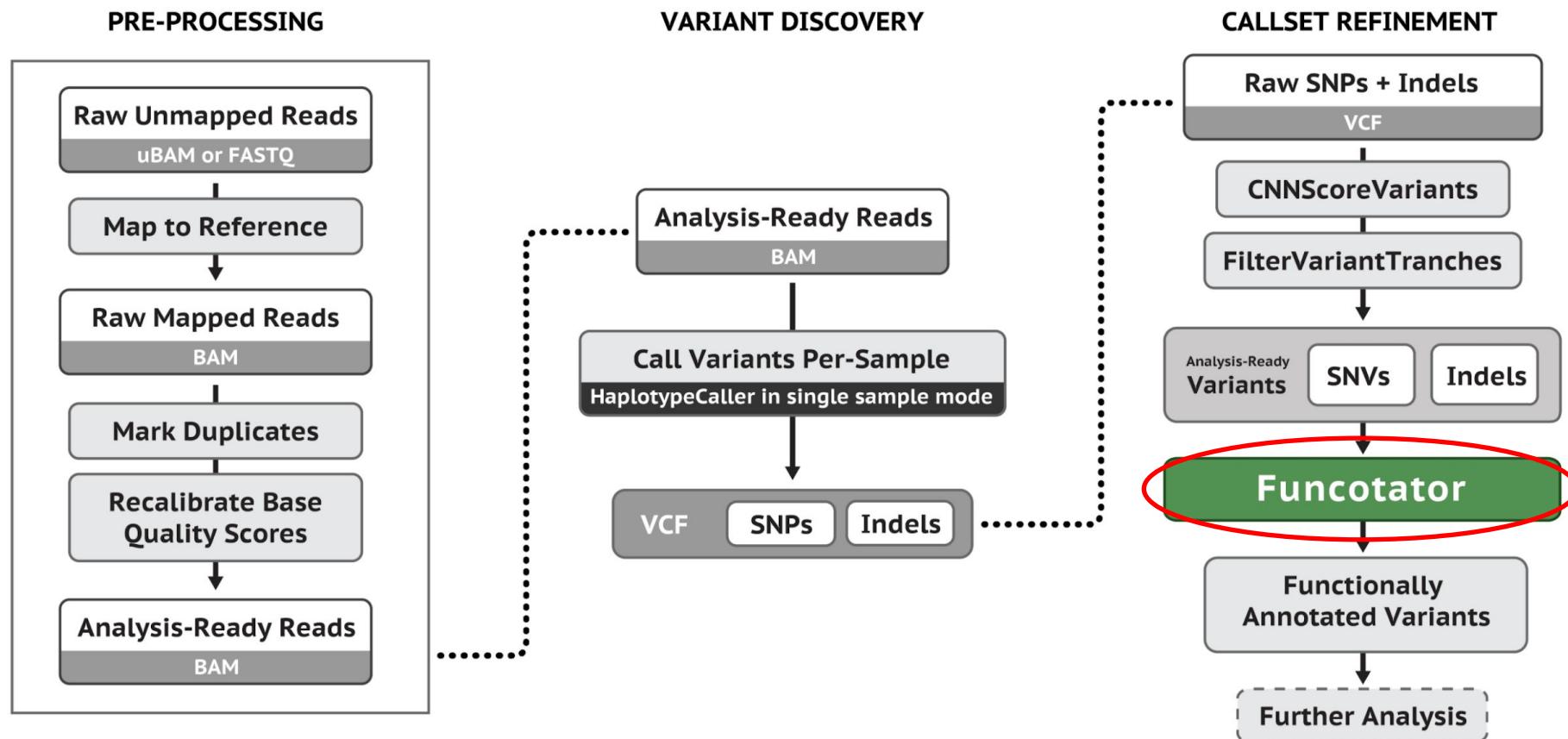
Variants Annotation process



Ready for annotating variants!!

GATK Functional Annotator

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



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/>

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
```

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)

Variant Effect Predictor Web interface

Use the VEP online to analyse your variants through a simple point-and-click interface.

The web interface allows you to access the key features of the VEP without using the command line. Interactively filter your results to find the data you want. Download your results in multiple data formats, easily share your results with others, and integrate your variation data with the powerful Ensembl web browser.

If you use the VEP in your work, please cite McLaren et. al. ([doi:10.1093/bioinformatics/btq330](https://doi.org/10.1093/bioinformatics/btq330) 

Any questions? Send an email to the Ensembl developer's mailing list, dev@ensembl.org or contact the Ensembl Helpdesk at helpdesk@ensembl.org.

Variant Effect Predictor Command line VEP

Use VEP to analyse your variation data locally. No limits, powerful, fast and extendable, command line VEP is the way to get the most out of [VEP](#) and Ensembl.

VEP is a powerful and highly configurable tool - have a browse through the [documentation](#). You might also like to read up on the [data formats](#) that VEP uses, and the different ways you can access [genome data](#). The VEP script can annotate your variants with [custom data](#), be extended with [plugins](#), and use powerful [filtering](#) to find biologically interesting results.

Beginners should have a run through the [tutorial](#), or try the [web interface](#) first.

If you use VEP in your work, please cite our latest publication McLaren et. al. 2016 ([doi:10.1186/s13059-016-0974-4](https://doi.org/10.1186/s13059-016-0974-4) 

Any questions? Send an email to the Ensembl [developers' mailing list](#) or contact the [Ensembl Helpdesk](#).

WGLab/doc- ANNOVAR



Documentation for the ANNOVAR software

83 7

Contributors

● 95

Issues

★ 182

Stars

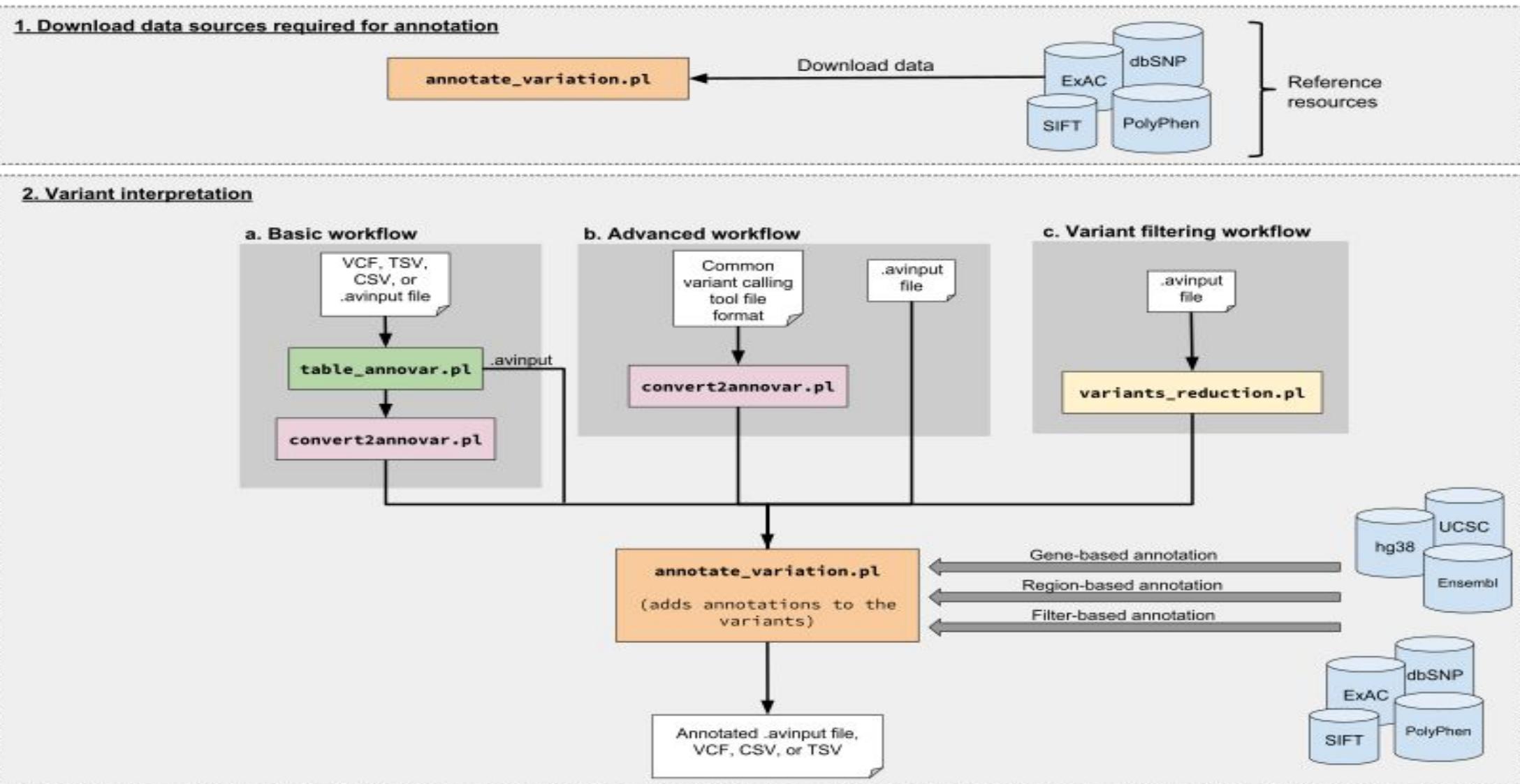
🍴 241

Forks



Annovar

ANNOVAR software package workflow



Overview of main scripts in Annovar program

Script	Purpose	Description	Input	Output	Requirements
<code>annotate_variation.pl</code>	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.
<code>convert2annovar.pl</code>	file converter	Converts various file formats to the custom ANNOVAR input file format.	See "Conversion to the ANNOVAR input file format" section.	.avinput	
<code>table_annovar.pl</code>	automated variant annotator	A wrapper around <code>annotate_variation.pl</code> 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.
<code>variants_reduction.pl</code>	variant reducer	<p>Performs stepwise variant reduction on a large set of input variants to narrow down to a subset of functionally important variants. Filtering procedures include:</p> <ul style="list-style-type: none"> 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.



THANK
YOU