GULBENKIAN SCIENCE

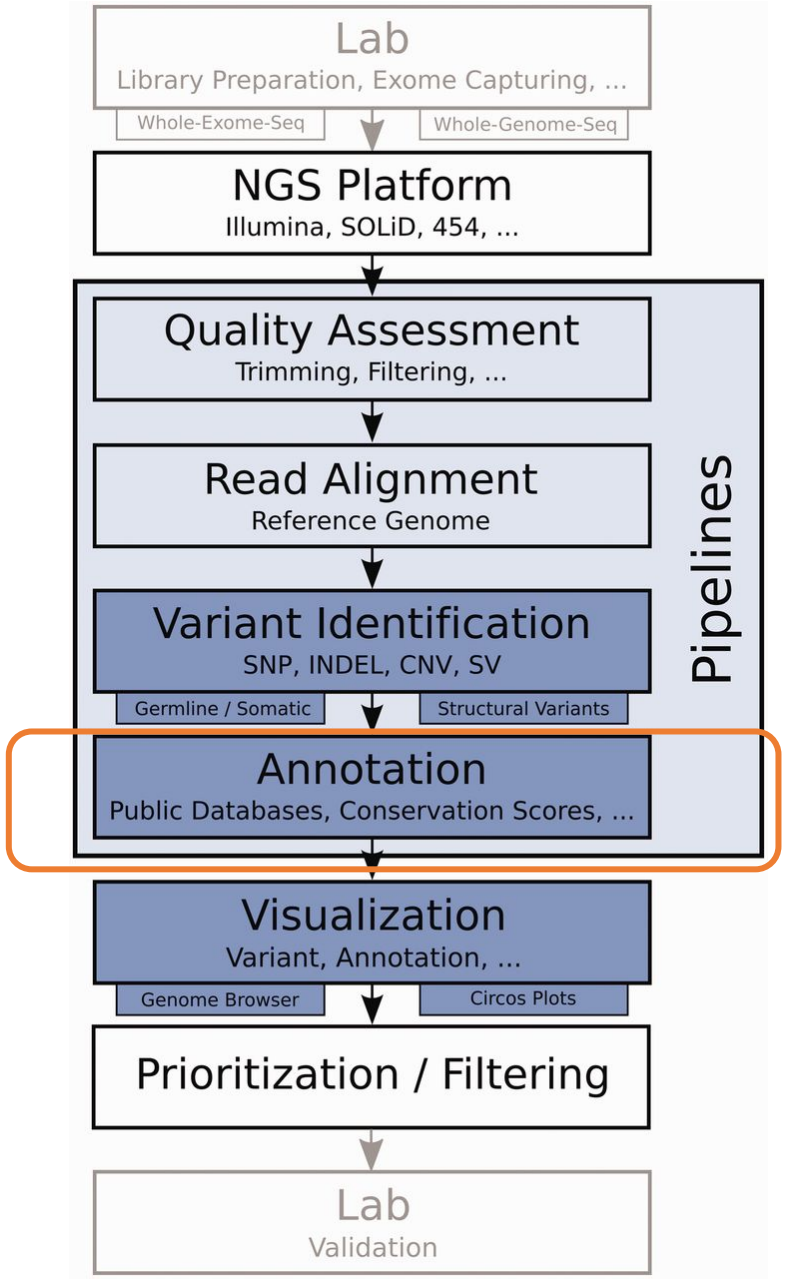
PO: Precision Oncology Course

Variant Annotation

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer

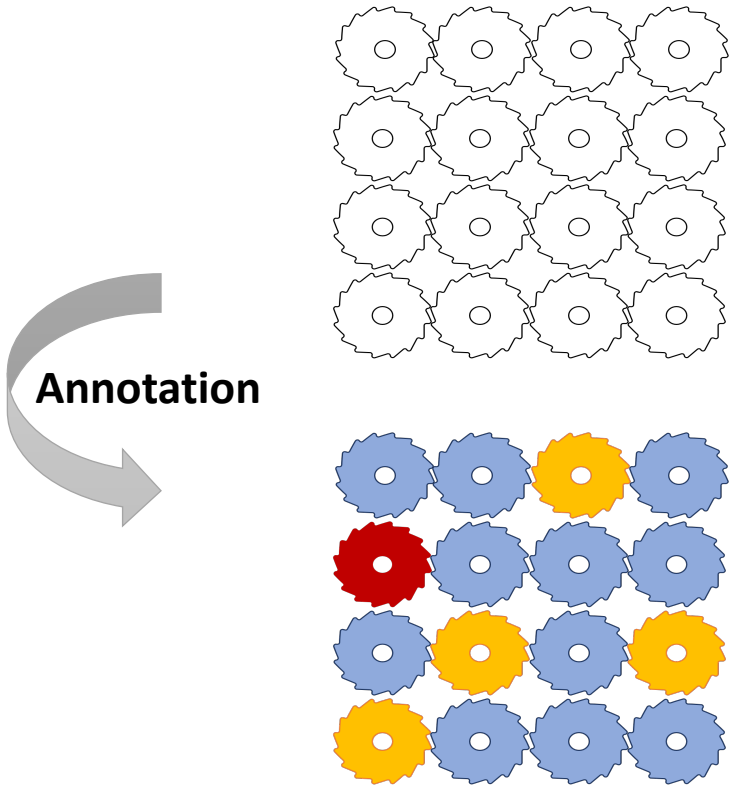
Annotation

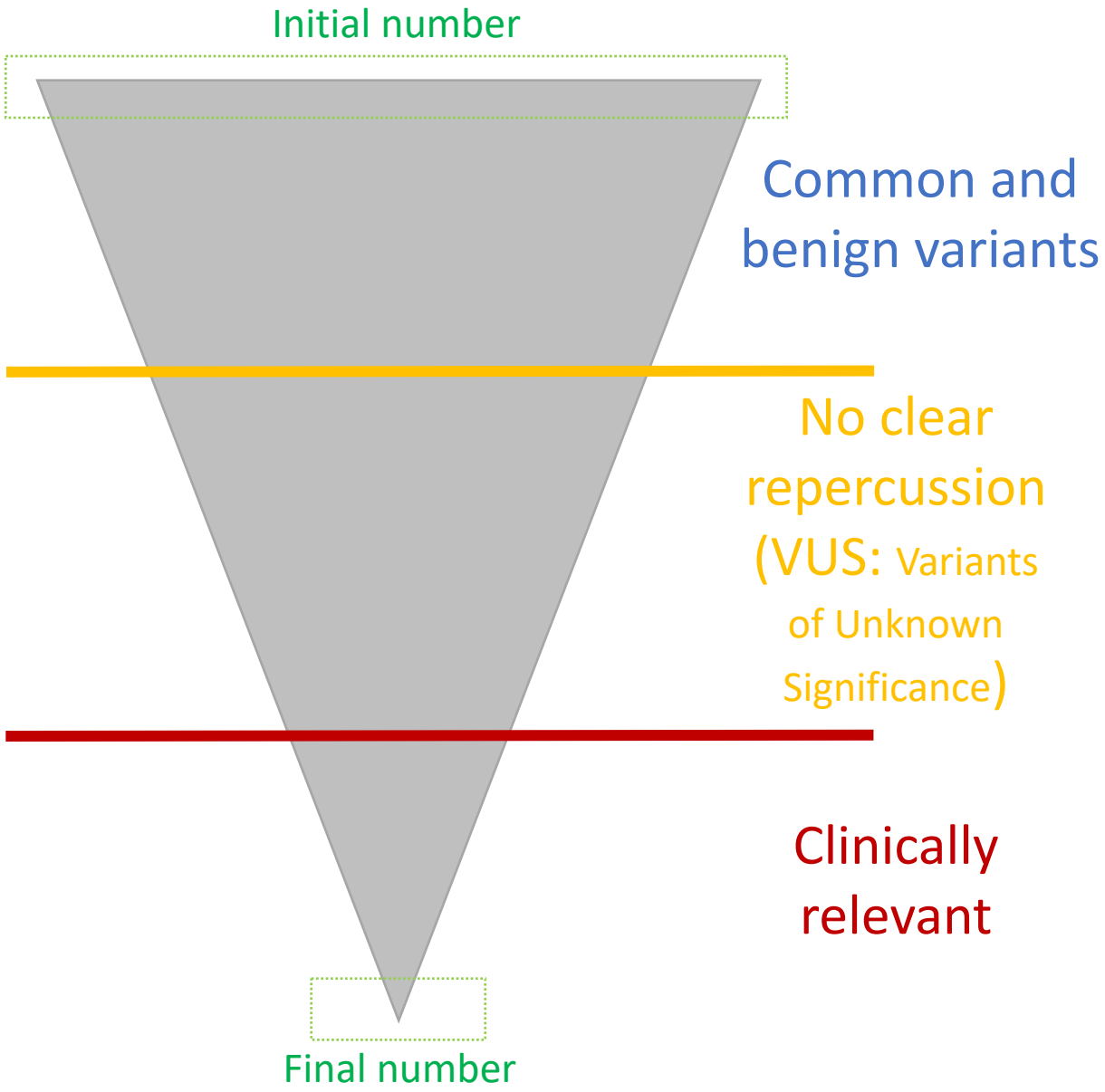


Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Why to annotate variants?

Not all the variants have a relevant impact in the phenotype of study





Variant annotations

Variant nomenclature

Type of variant
Indicates if it is a SNV or an indel (insertion or deletion)

Affected molecules

Genomic location

Sequence consequences

Functional impact prediction

Population frequencies

Association with pathologies

Nomenclature of variants

- **Chromosome coordinates** (They **change** depending on the reference genome version)

1 12794077 T/C (PRAMEF1) <=> 1 12734114 T/C (PRAMEF1)

GRCh37 GRCh38

1 12794076 TC/T <=> 1 12794077 C/-

VCF

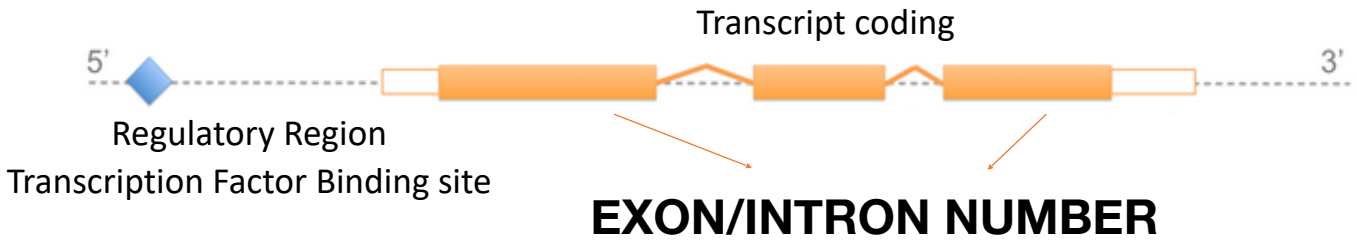
- **HGVS nomenclature (standard)**
  - Coding sequence e.g. c.638G>A
  - Protein sequence e.g. p.Arg213His

Coding sequence nomenclature uses
A of ATG codon as number 1 position

http://varnomen.hgvs.org

"**g.**" for a genomic reference sequence
"**c.**" for a coding DNA reference sequence
"**m.**" for a mitochondrial DNA reference sequence
"**n.**" for a non-coding DNA reference sequence
"**r.**" for an RNA reference sequence (transcript)
"**p.**" for a protein reference sequence

Genomic location

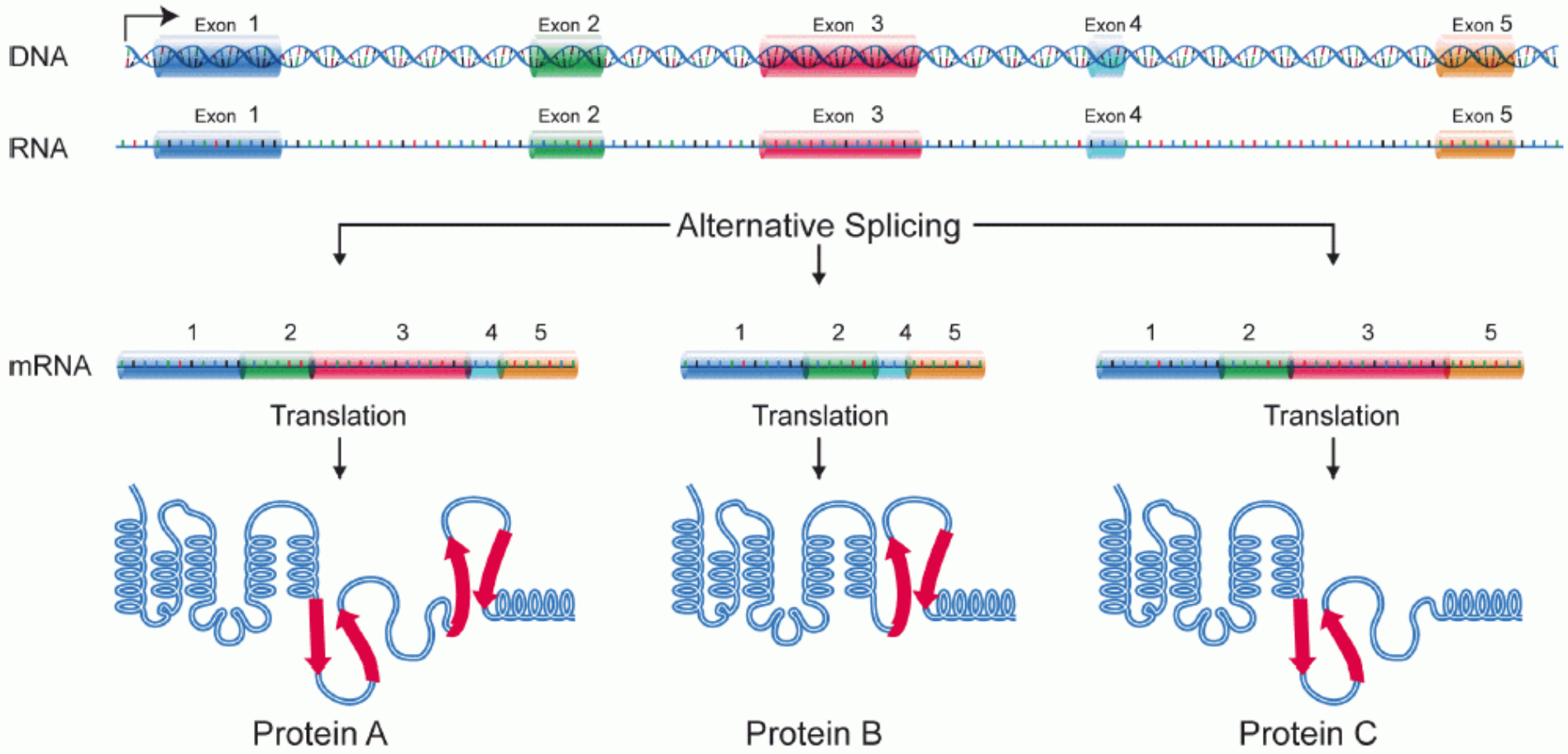


BIOTYPE

Type of transcript or regulatory feature. e.g. promoter, protein_coding, processed_pseudogene, miRNA, rRNA, scRNA, snoRNA, snRNA...

http://vega.archive.ensembl.org/info/about/gene_and_transcript_types.html

Affected elements



TRANSCRIPTS
ISOFORMS

PROTEINS

Affected elements: genes and transcripts

GENE IDENTIFIERS

- HUGO Gene Symbol (Standard nomenclature for the human genes) e.g.: MYC
  https://www.genenames.org/

- Identifiers from databases
  e.g. ENSG00000136997 (ensembl)
  4609 (Entrez gene NCBI)

TRANSCRIPT IDENTIFIERS

Ensembl e.g. ENST00000426406

RefSeq e.g. NM_001005221.2

Transcript support and relevance

CCDS (**C**onsensus **Cod**ing **S**equence) e.g. CCDS1639.1

Coding regions consistently annotated and with high quality

TRANSCRIPT SUPPORT LEVEL (Ensembl)

The Transcript Support Level (TSL) indicates if the transcript model is well or poorly supported

tsl1 > tsl2 > tsl3 > tsl4 > tsl5 > tslNA (the transcript was not analyzed)

PRINCIPAL ISOFORM

- Traditionally based on length criteria: Principal isoform = longest transcript
- Based on functional evidences: protein structure, conservation among species, functional features

Principal Isoform - APPRIS

http://appris-tools.org/#/

{APPRIS} 2016_06.v17 Tools Downloads WebServices Help & Docs About us Search gene...

{APPRIS}

Annotating principal splice isoforms

Executes several computational methods for the transcript annotation.

As part of the annotation process, it selects a transcript(s) as the principal isoform for each gene.

APPRIS Database

Access annotations for the species annotated in the database via gene name or Ensembl id.

Access the web database

APPRIS WebServer

Annotate splice isoforms for vertebrate genes that are not in the APPRIS Database.

Run the web server

APPRIS WebServices

Annotate genes and transcripts automatically and access queries through RESTful web services.

Go to the API inteface

APPRIS Database currently houses annotations for vertebrate genomes »

Human

Assemblies: GRCh38

Assemblies: GRCh37

Mouse

Assemblies: GRCm38

Zebrafish

Assemblies: GRCz10

Assemblies: Zv9

Rat

Assemblies: Rnor_6.0

Assemblies: Rnor_5.0

Pig

Assemblies: Sscrofa10.2

Chimpanzee

Assemblies: CHIMP2.1.4

APPRIS Database currently houses annotations for invertebrate genomes »

Fruitfly

Assemblies: BDGP6

C.elegans

Assemblies: WBcel235

Principal Isoform - APPRIS

Id	Name	Biotype	Species	Assembly	Location
ENSG00000116815	CD58	protein_coding	Homo sapiens	GRCh38	1:116514535-11657103...

Principal Isoforms

Hide panel

Seq. id	Seq. name	Length (aa)	Biotype	CCDS	Flags	Principal Isoform
ENST00000369487	CD58-201	240	protein_coding	-	TSL1	ALTERNATIVE:2
ENST00000369489	CD58-202	250	protein_coding	CCDS888	TSL1	PRINCIPAL:3
ENST00000457047	CD58-203	248	protein_coding	CCDS44199	TSL1	ALTERNATIVE:2
ENST00000464088	CD58-204	237	nonsense_mediated_decay	-	TSL1	MINOR
ENST00000526981	CD58-205	131	protein_coding	-	TSL1	ALTERNATIVE:2

APPRIS annotations

Hide panel

All / None	Seq. id	Length (aa)	No. Functional Residues	3D Structure Score	Domains score	Conservation score	No. Transmembrane Helices	Signal Sequence	No. Mapping Peptides
☑	ENST00000369487	240	0	1.066	0	1	2	-	4
☑	ENST00000369489	250	0	1.066	0	1.5	3	-	4
☑	ENST00000457047	248	0	1.066	0	1.5	3	-	4
☑	ENST00000464088	237	0	1.066	0	1	2	-	4
☑	ENST00000526981	131	0	1	0	0	3	-	2

ISOFORM RANGE

PRINCIPAL:1
PRINCIPAL:2
PRINCIPAL:3
PRINCIPAL:4
PRINCIPAL:5

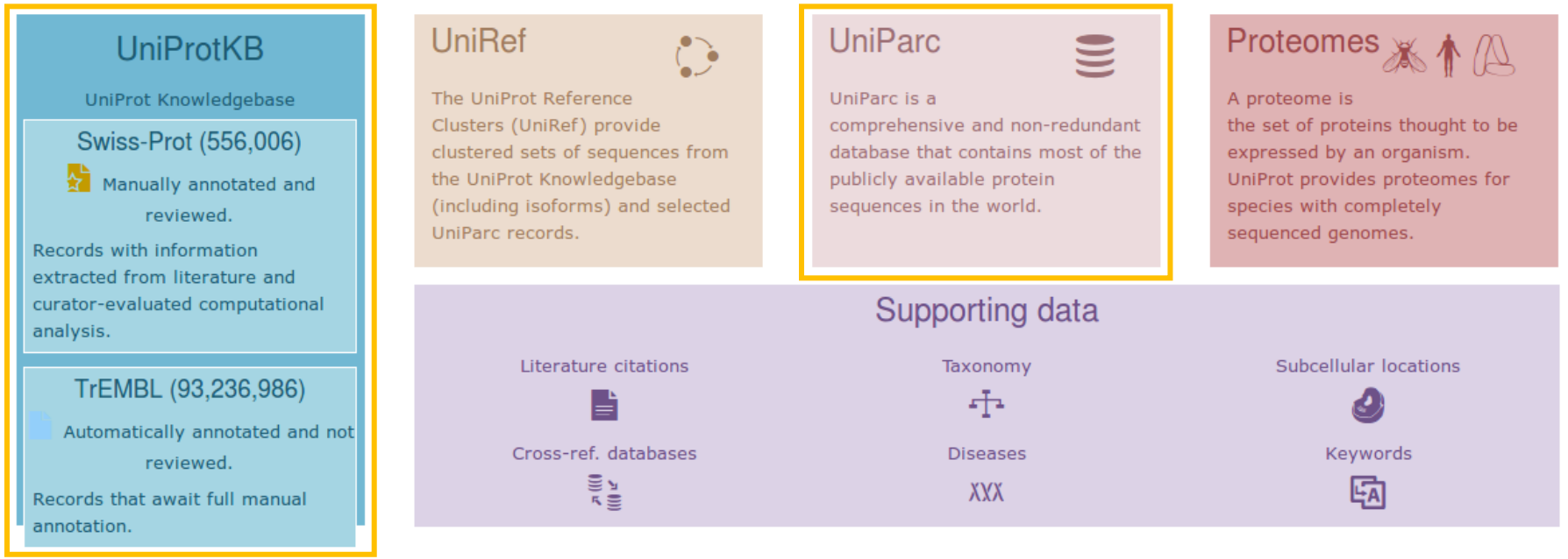
ALTERNATIVE:1
ALTERNATIVE:2

Affected elements: proteins

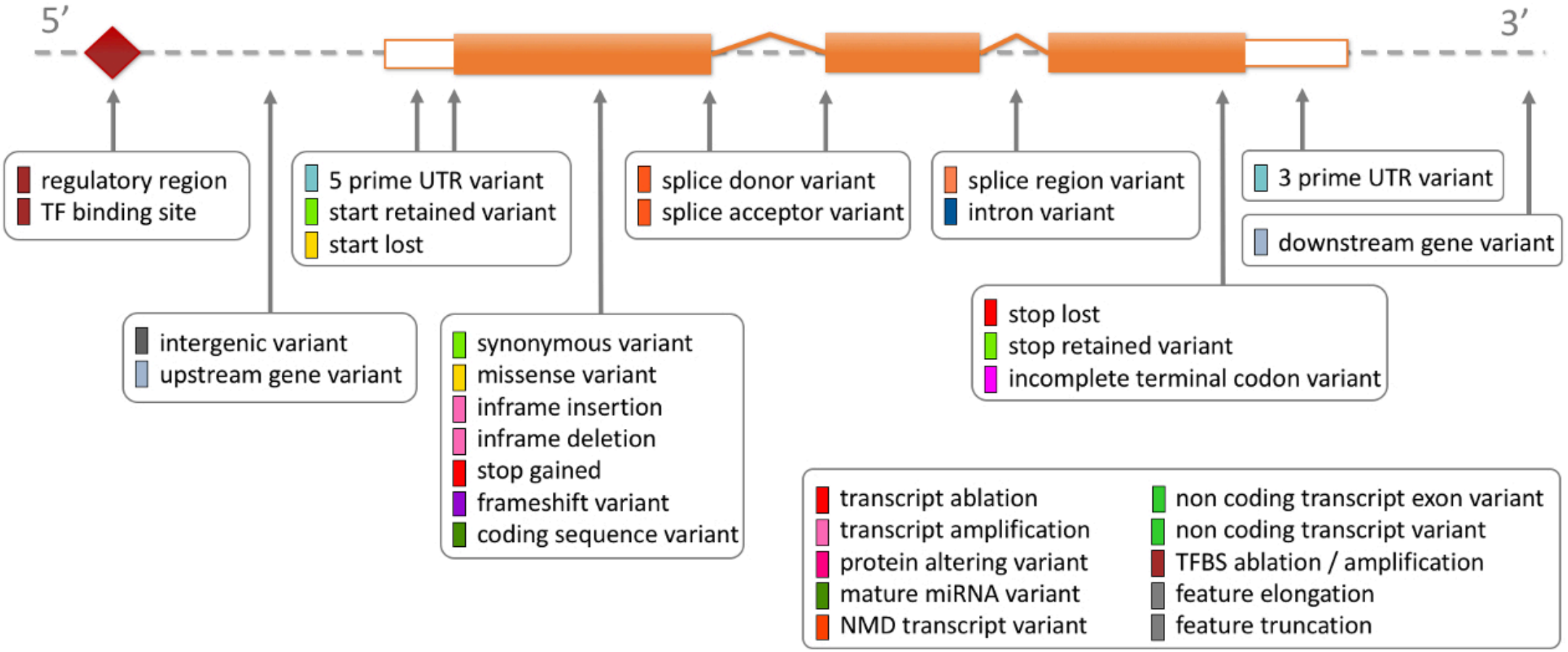
PROTEIN IDENTIFIERS

- Ensembl identifier e.g. ENSP00000409316
- RefSeq identifier e.g. NP_001005221

- Uniprot identifiers (SWISSPROT, TrEMBL y UniParc)
  e.g. Q6IEY1 (SWISSPROT), A0A126GV92 (TREMBL), UPI0000041D3C (UniParc)

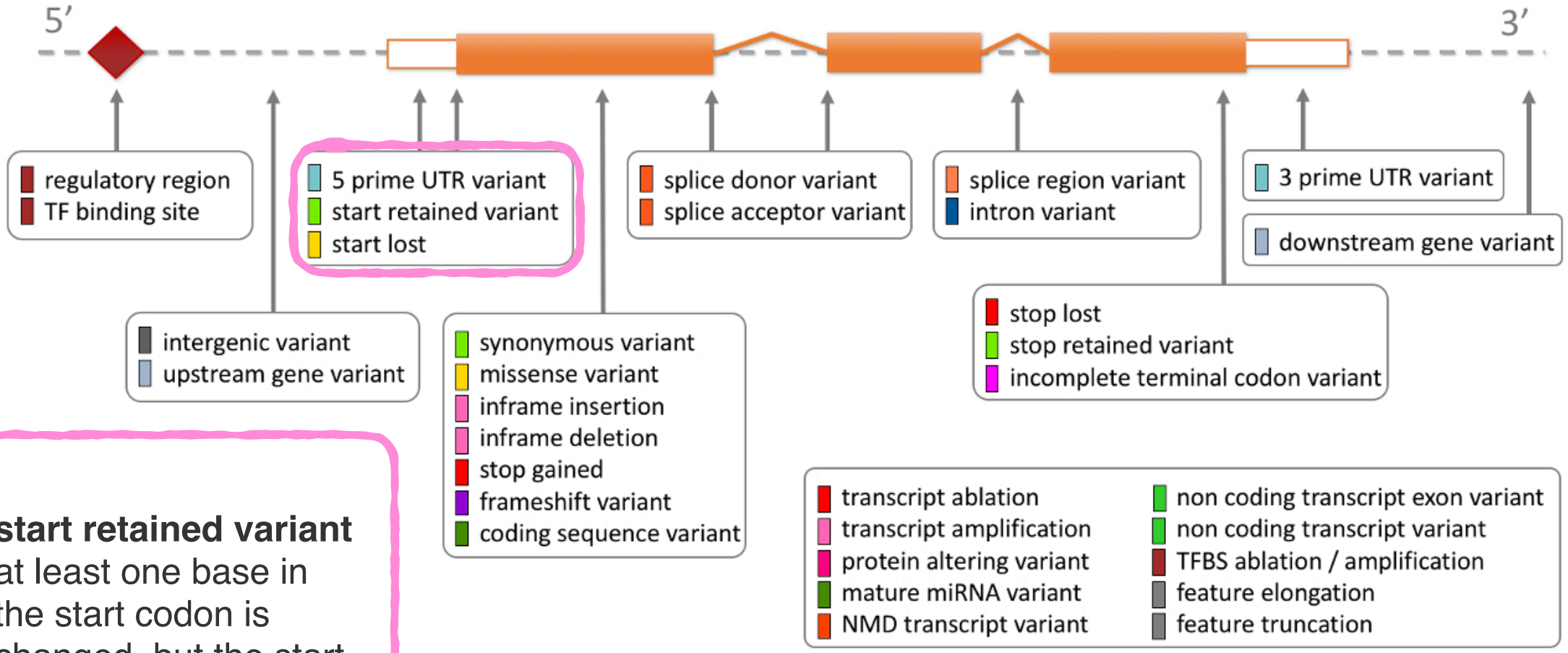


Sequence consequences



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

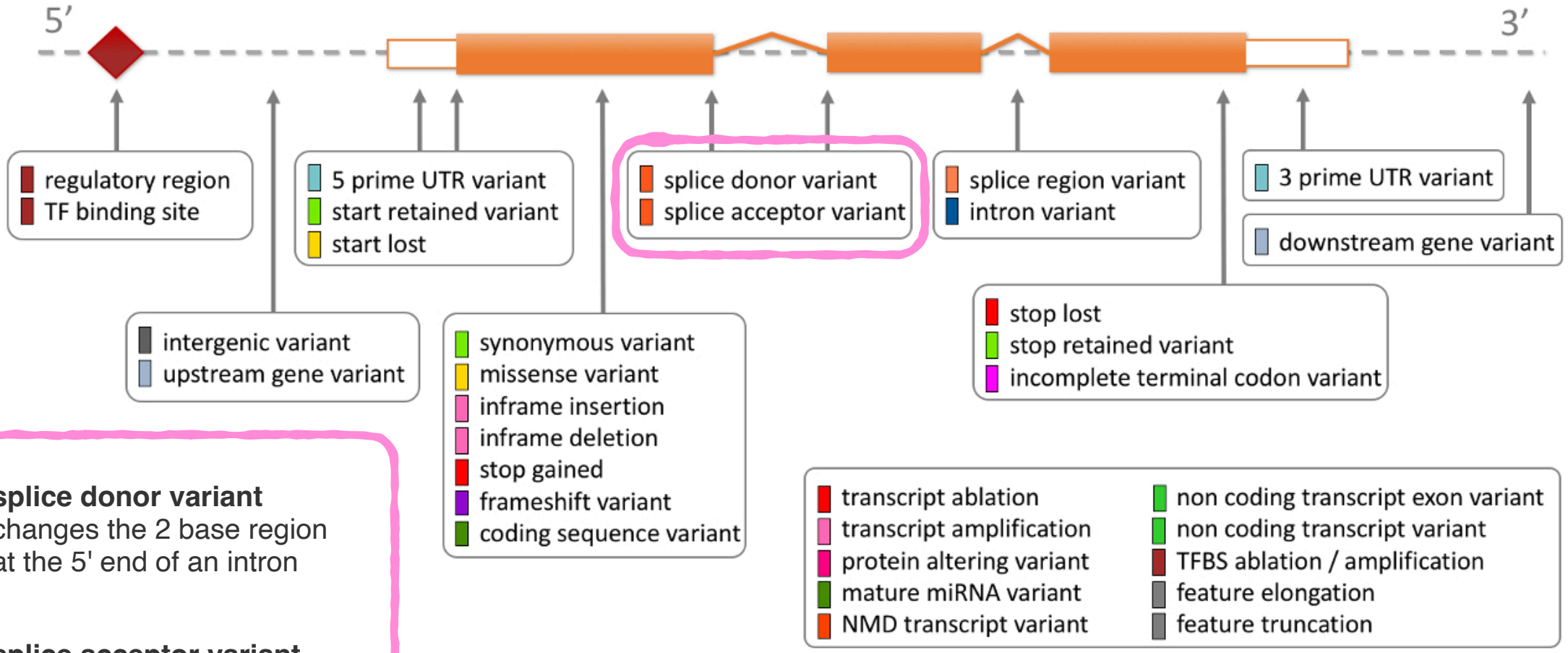
Sequence consequences



start retained variant
at least one base in the start codon is changed, but the start remains

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

Sequence consequences

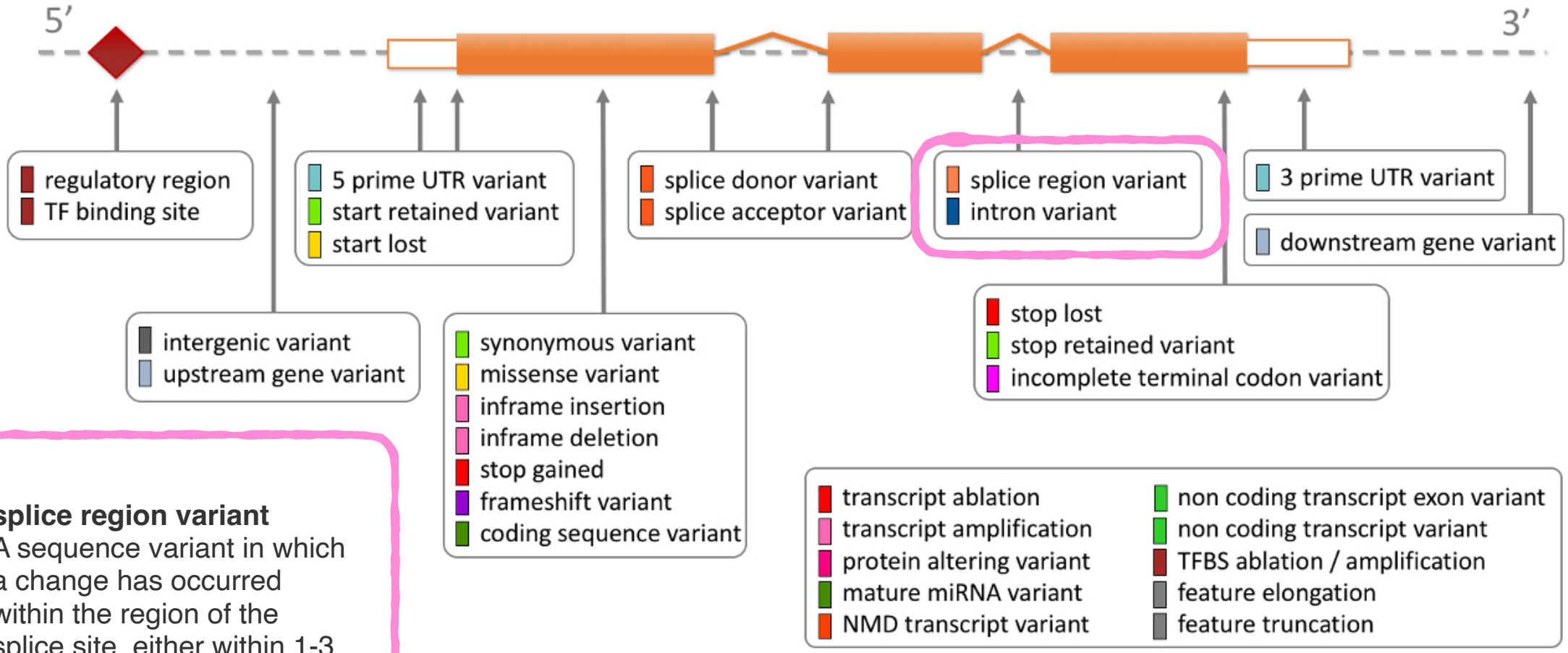


splice donor variant
changes the 2 base region at the 5' end of an intron

splice acceptor variant
changes the 2 base region at the 3' end of an intron

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

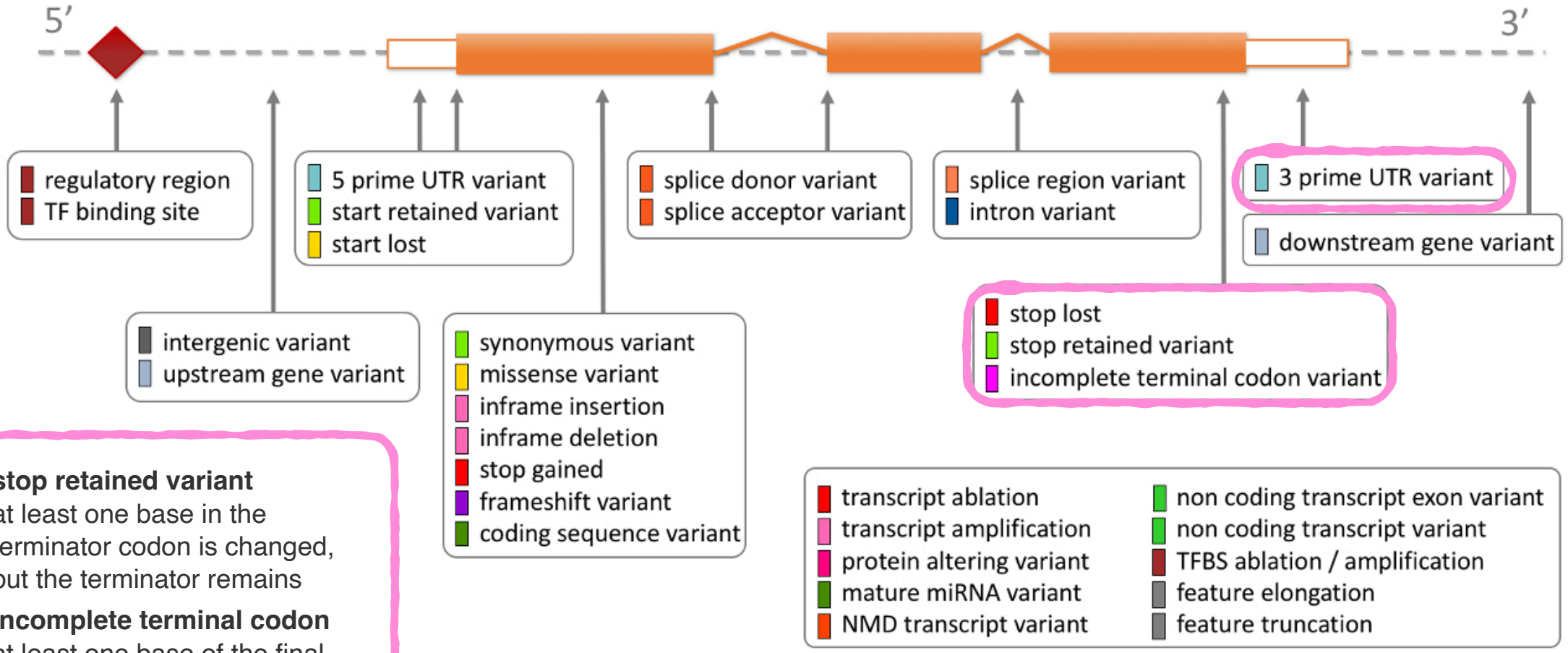
Sequence consequences



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

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

Sequence consequences

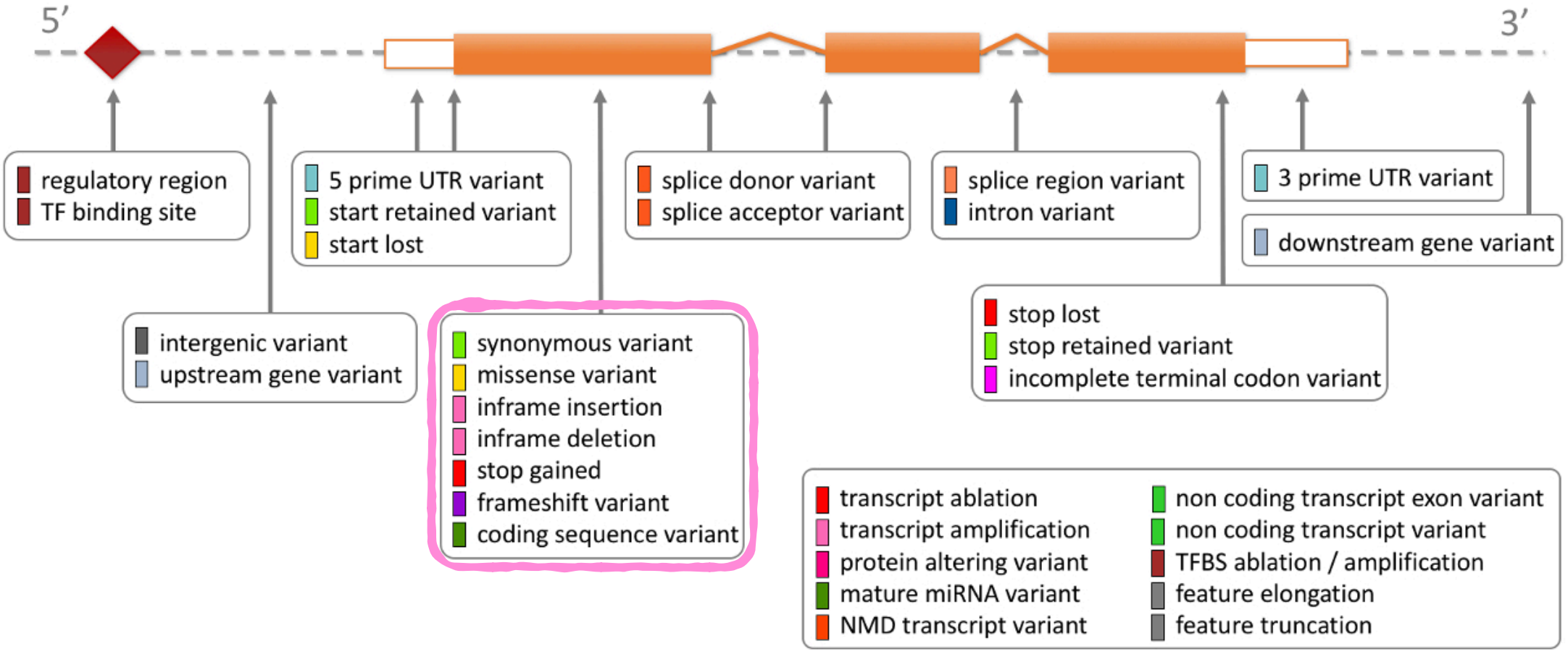


stop retained variant
at least one base in the terminator codon is changed, but the terminator remains

incomplete terminal codon
at least one base of the final codon of an incompletely annotated transcript is changed

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

Sequence consequences



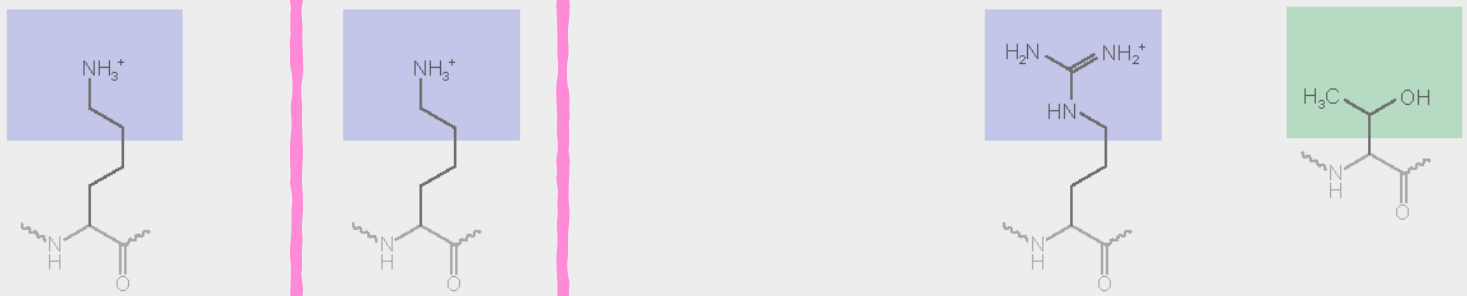
https://m.ensembl.org/info/genome/variation/prediction/predicted_data.html

Sequence consequences

- synonymous variant
- missense variant
- inframe insertion
- inframe deletion
- stop gained
- frameshift variant
- coding sequence variant

Point mutations

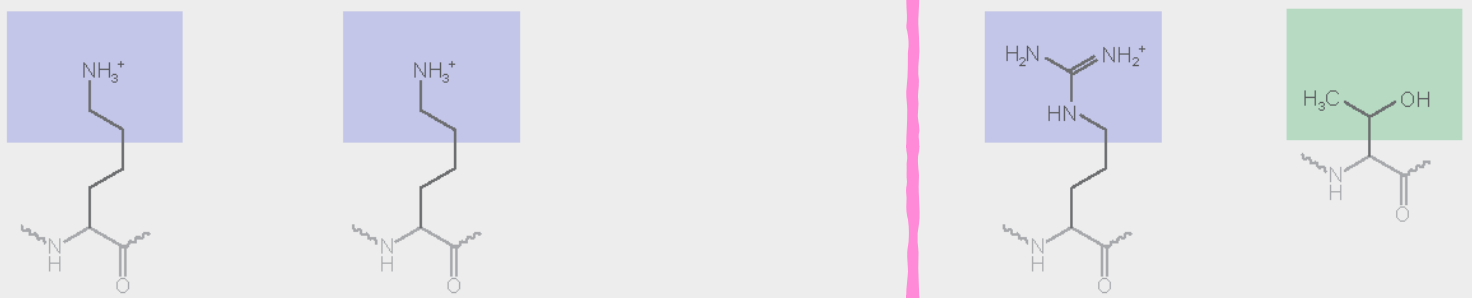
	No mutation	Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



Sequence consequences

- synonymous variant
- missense variant
- inframe insertion
- inframe deletion
- stop gained
- frameshift variant
- coding sequence variant

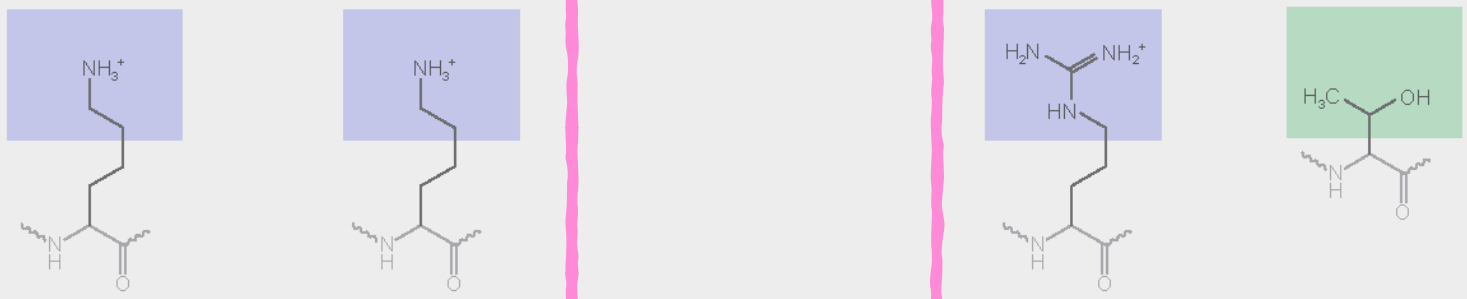
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



Sequence consequences

- synonymous variant
- missense variant
- inframe insertion
- inframe deletion
- stop gained
- frameshift variant
- coding sequence variant

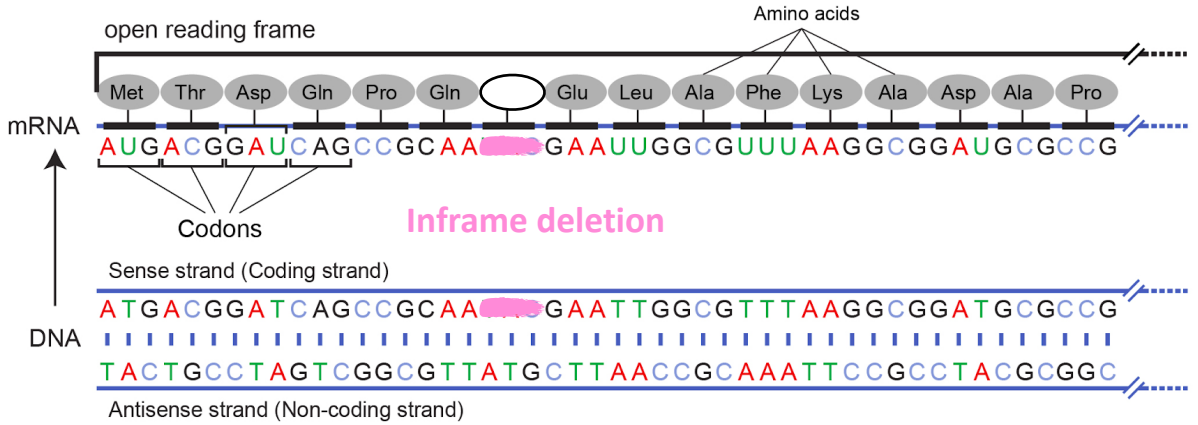
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



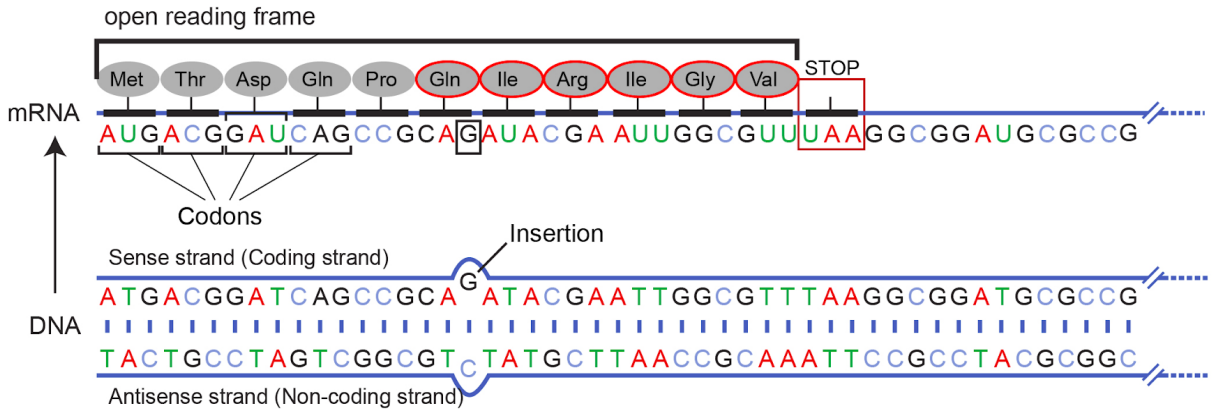
Sequence consequences

- synonymous variant
- missense variant
- inframe insertion
- inframe deletion
- stop gained
- frameshift variant
- coding sequence variant

Normal



Frameshift mutation - single nucleotide insertion



Functional impact prediction

Impact based on sequence change

*	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 codo	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 sequenc	SO:0001821	Inframe insertion	MODERATE
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	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

HIGH

MODERATE

High

Moderate

Low

Modifier

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

Impact based on *in silico* predictors

They use **several evidences**:

- Conservation degree
- Sequence similarity
- Protein structure
- Protein stability

They use **several algorithms**:

- Random Forest
- Neural Network

General

TABLE 2 | Tools, software, and databases for functional prediction and annotation of variant impact.

Resource	URL	Reference	Notes
Integrated predictive methods and aggregated databases			
dbNSFP[a,b,c,d]	https://sites.google.com/site/jpopgen/dbNSFP	(45)	Aggregated database of variant information
myvariant.info[a]	http://myvariant.info/	(46)	Aggregated database of variant information
Functional effect prediction software and algorithms			
PolyPhen-2[b]	http://genetics.bwh.harvard.edu/pph2	(47)	Bayesian classification
SIFT[b]	http://sift.jcvi.org	(48)	Alignment scores
MutationAssessor	http://mutationassessor.org	(27)	Conservation, naive Bayes classifier
MutationTaster	http://www.mutationtaster.org	(49)	
PROVEAN	http://provean.jcvi.org/index.php	(50)	
CADD[b,c]	http://cadd.gs.washington.edu	(51)	
GERP++[c]	http://mendel.stanford.edu/SidowLab/downloads/gerp/index.html	(52)	
PhyloP and PhastCons	http://compgen.cshl.edu/phast/index.php	(53, 54)	
nsSNPAnalyzer	http://snpanalyzer.uthsc.edu/	(55)	Random Forest
SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go/	(56)	SVM
SNAP2	https://rostlab.org/services/snap2web/	(57)	Neural Networks
SNPs3D	http://www.snps3d.org/	(58)	Structure and sequence analysis
MutPred2	http://mutpred.mutdb.org/	(59)	Random Forest
AUTO-MUTE	http://binf2.gmu.edu/automute/	(60)	Topology and statistical contact potential
Panther	http://www.pantherdb.org/tools/csnpScoreForm.jsp	(61)	Hidden Markov Model
stSNP	http://ilyinlab.org/StSNP/	(62)	Comparative modeling of protein structure
Condel[b]	http://bg.upf.edu/fannsdb/	(63)	A weighted average of multiple methods
CoVEC	https://sourceforge.net/projects/covec/files		
CAROL[b]	http://www.sanger.ac.uk/science/tools/carol	(64)	Combines PolyPhen-2 and SIFT
Cancer-specific prediction tools			
CHASM	http://wiki.chasmsoftware.org/index.php/Main_Page	(65)	Random Forest
CanDrA	http://bioinformatics.mdanderson.org/main/CanDrA#CanDrA	(66)	96 structural, evolutionary and gene features

[a]Aggregated databases combine outputs of other databases and algorithms are, therefore, efficient resources to use in annotation pipelines. Adding these resources to observed variants is supported software in ***Table 4*** *including Ensembl VEP software (noted[b] in this table), Annovar (noted[c]), and snpEff (noted[d]).*

Frontiers in Oncology, 2017

Sift and PolyPhen

SIFT PREDICTION predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.

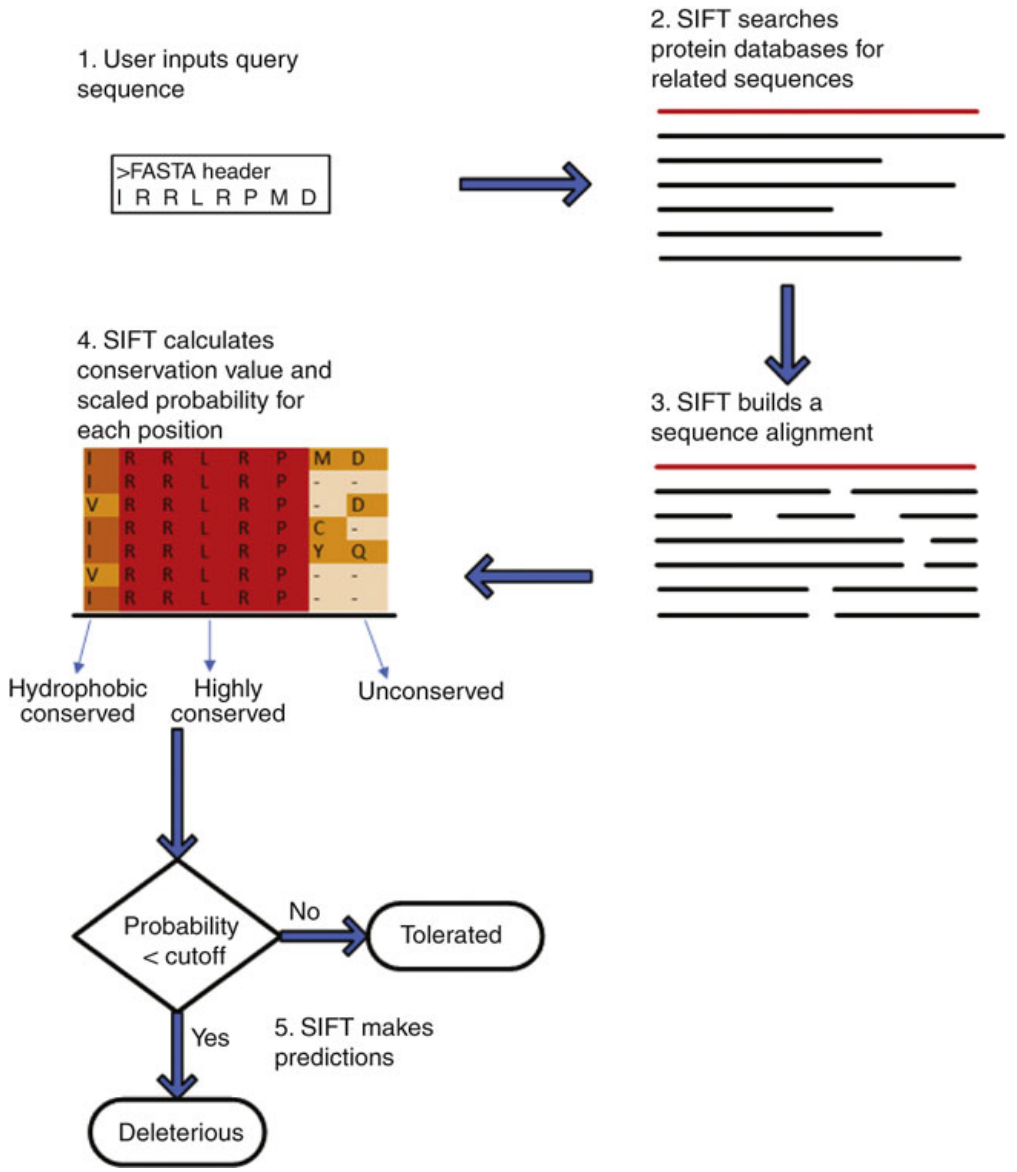
Deleterious – Tolerated

0 < 0.05 Deleterious 1

PolyPhen-2 PREDICTION predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

Unknown - Benign - Possibly damaging - Probably damaging

0 Benign < ~0.435 < Damaging 1



No consensus among predictors

CONDEL CONsensus DELeteriousness score computes a consensus score based on:

MutationAssessor
FatHMM

dbNSFP functional predictions and annotations for human nonsynonymous single-nucleotide variants and splice-site variants from various tools:

MetaSVM, MetaLR, CADD, VEST3, PROVEAN, 4×fitCons, fathmm-MKL, DANN

In silico predictors

Specific of event type:

Missense variants
Splice-site variants

Genetics in Medicine, 2015

Table 2 In silico predictive algorithms

Category	Name	Website	Basis
Missense prediction	ConSurf	http://consurftest.tau.ac.il	Evolutionary conservation
	FATHMM	http://fathmm.biocompute.org.uk	Evolutionary conservation
	MutationAssessor	http://mutationassessor.org	Evolutionary conservation
	PANTHER	http://www.pantherdb.org/tools/csnpScoreForm.jsp	Evolutionary conservation
	PhD-SNP	http://snps.biofold.org/phd-snp/phd-snp.html	Evolutionary conservation
	SIFT	http://sift.jcvi.org	Evolutionary conservation
	SNPs&GO	http://snps-and-go.biocomp.unibo.it/snps-and-go	Protein structure/function
	Align GVGD	http://agvgd.iarc.fr/agvgd_input.php	Protein structure/function and evolutionary conservation
	MAPP	http://mendel.stanford.edu/SidowLab/downloads/MAPP/index.html	Protein structure/function and evolutionary conservation
	MutationTaster	http://www.mutationtaster.org	Protein structure/function and evolutionary conservation
	MutPred	http://mutpred.mutdb.org	Protein structure/function and evolutionary conservation
	PolyPhen-2	http://genetics.bwh.harvard.edu/pph2	Protein structure/function and evolutionary conservation
	PROVEAN	http://provean.jcvi.org/index.php	Alignment and measurement of similarity between variant sequence and protein sequence homolog
	nsSNPAnalyzer	http://snpanalyzer.uthsc.edu	Multiple sequence alignment and protein structure analysis
	Condel	http://bg.upf.edu/fannsdb/	Combines SIFT, PolyPhen-2, and MutationAssessor
	CADD	http://cadd.gs.washington.edu	Contrasts annotations of fixed/nearly fixed derived alleles in humans with simulated variants
Splice site prediction	GeneSplicer	http://www.cbcb.umd.edu/software/GeneSplicer/gene_spl.shtml	Markov models
	Human Splicing Finder	http://www.umd.be/HSF/	Position-dependent logic
	MaxEntScan	http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html	Maximum entropy principle
	NetGene2	http://www.cbs.dtu.dk/services/NetGene2	Neural networks
	NNSplice	http://www.fruitfly.org/seq_tools/splice.html	Neural networks
	FSPLICE	http://www.softberry.com/berry.phtml?topic=fsplice&group=programs&subgroup=gfind	Species-specific predictor for splice sites based on weight matrices model
Nucleotide conservation prediction	GERP	http://mendel.stanford.edu/sidowlab/downloads/gerp/index.html	Genomic evolutionary rate profiling
	PhastCons	http://compgen.bscb.cornell.edu/phast/	Conservation scoring and identification of conserved elements
	PhyloP	http://compgen.bscb.cornell.edu/phast/	
		http://compgen.bscb.cornell.edu/phast/help-pages/phyloP.txt	Alignment and phylogenetic trees: Computation of *P* values for conservation or acceleration, either lineage-specific or across all branches

In silico tools/software prediction programs used for sequence variant interpretation.

In silico predictors – Take into account

- Only valid for **specific type of alterations**.
- They have a **moderate precision and specificity**.
- **Predictions can change**:
  - Depending on the predictor
  - Depending on the gene or protein sequence

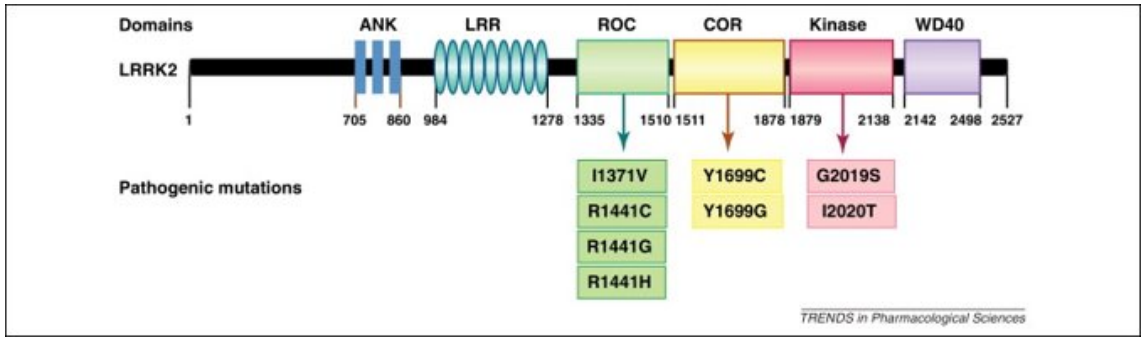
Predictions should be taken cautiously and not been used as the only evidence for the relevance of the variant.

Impact through domain alteration

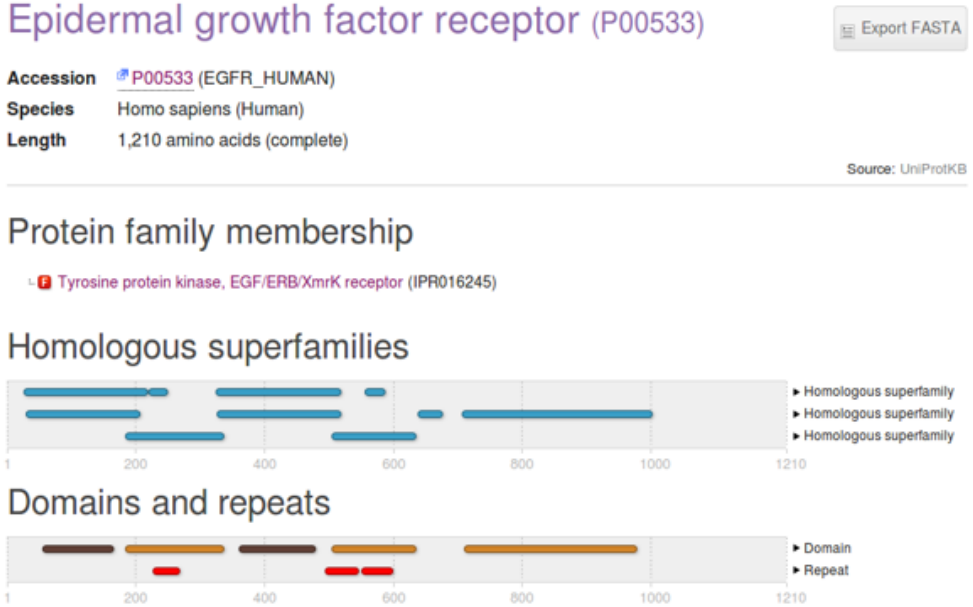
DOMAINS Protein overlapping domains

Pfam, Prosite, InterPro

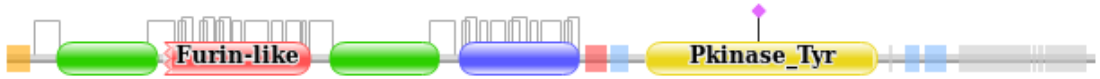
e.g. Pfam_domain: PF00071



https://www.ebi.ac.uk/interpro/



EMBL-EBI Pfam http://pfam.xfam.org



Download the data used to generate the domain graphic in JSON format.

Source	Domain	Start	End
sig_p	n/a	1	24
low_complexity	n/a	6	24
Pfam	Recep_L_domain	57	168
Pfam	Furin-like	177	338
Pfam	Recep_L_domain	361	481
Pfam	GF_recep_IV	505	637
transmembrane	n/a	646	667
low_complexity	n/a	650	665
low_complexity	n/a	674	691
Pfam	Pkinase_Tyr	712	968

Description:	Epidermal growth factor receptor EC=2.7.10.1
Source organism:	Homo sapiens (Human) (NCBI taxonomy ID 9606) View Pfam proteome data.
Length:	1210 amino acids
Reference Proteome:	✓

Population frequencies

Frequency with which a variant appears in a population

≥ 1% : **polymorphism** (common variant)

Common variants can be **involved in predisposition and drug response**

Table 1 Population, disease-specific, and sequence databases

Data Bases	
Population databases	
Exome Aggregation Consortium http://exac.broadinstitute.org/	Database of variants found during exome sequencing of 61,486 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Pediatric disease subjects as well as related individuals were excluded.
Exome Variant Server http://evs.gs.washington.edu/EVS	Database of variants found during exome sequencing of several large cohorts of individuals of European and African American ancestry. Includes coverage data to inform the absence of variation.
1000 Genomes Project http://browser.1000genomes.org	Database of variants found during low-coverage and high-coverage genomic and targeted sequencing from 26 populations. Provides more diversity compared to the Exome Variant Server but also contains lower-quality data, and some cohorts contain related individuals.
dbSNP http://www.ncbi.nlm.nih.gov/snp	Database of short genetic variations (typically ≤50 bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	Database of structural variation (typically >50 bp) submitted from many sources.
Disease databases	
ClinVar http://www.ncbi.nlm.nih.gov/clinvar	Database of assertions about the clinical significance and phenotype relationship of human variations.
OMIM http://www.omim.org	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database http://www.hgmd.org	Database of variant annotations published in the literature. Requires fee-based subscription to access much of the content.
Locus/disease/ethnic/other-specific databases	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html Leiden Open Variation Database http://www.lovd.nl	The Human Genome Variation Society site developed a list of thousands of databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the Leiden Open Variation Database system.
DECIPHER http://decipher.sanger.ac.uk	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.
Sequence databases	
NCBI Genome http://www.ncbi.nlm.nih.gov/genome	Source of full human genome reference sequences.
RefSeqGene http://www.ncbi.nlm.nih.gov/refseq/rsg Locus Reference Genomic (LRG) http://www.lrg-sequence.org	Medically relevant gene reference sequence resource.
MitoMap http://www.mitomap.org/MITOMAP/HumanMitoSeq	Revised Cambridge reference sequence for human mitochondrial DNA.

Genetics in Medicine, 2015

1000 Genomes Project

ARTICLE

OPEN

doi:10.1038/nature15393

A global reference for human genetic variation

The 1000 Genomes Project Consortium*

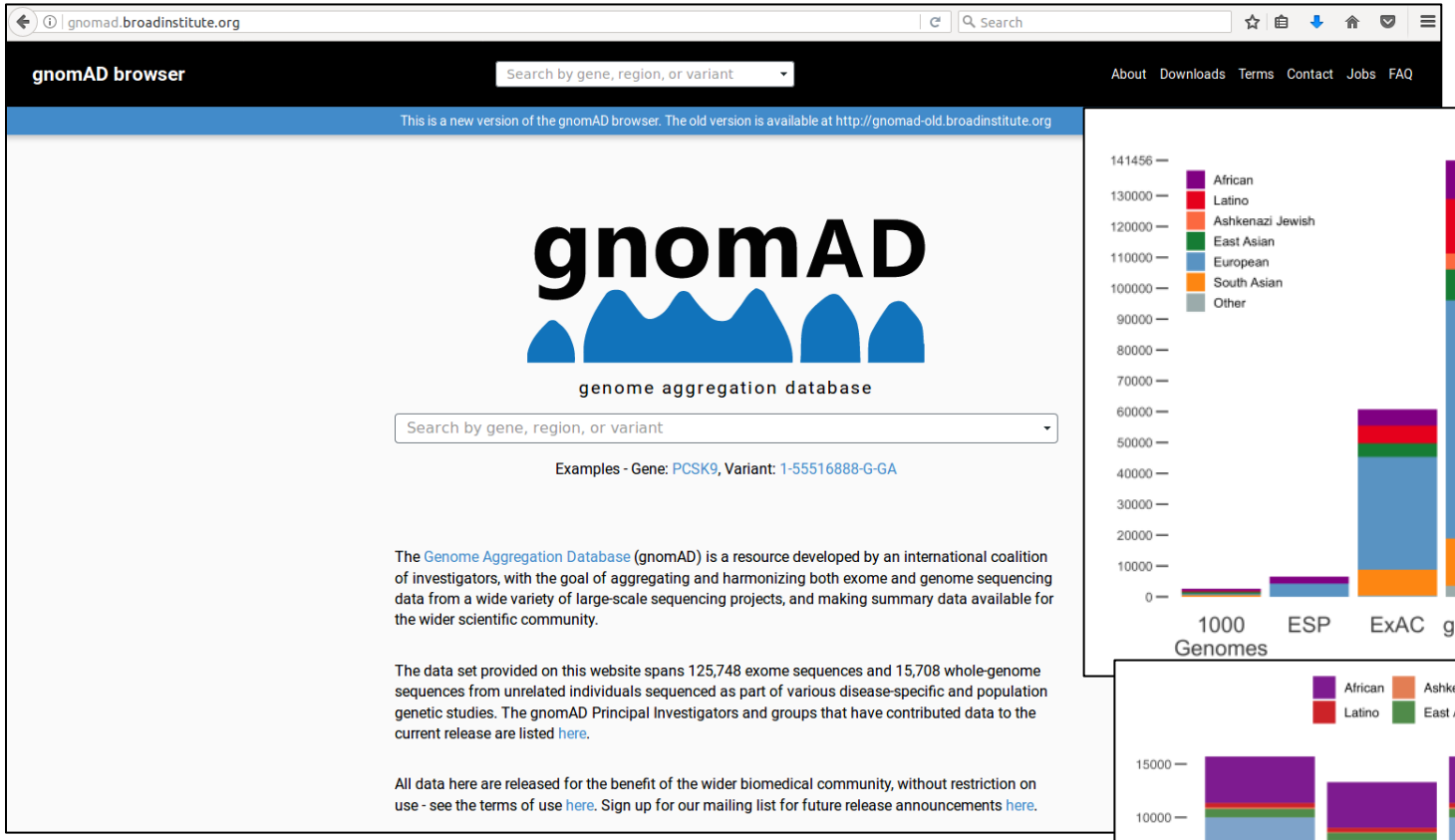
The 1000 Genomes Project set out to provide a comprehensive description of common human genetic variation by applying whole-genome sequencing to a diverse set of individuals from multiple populations. Here we report completion of the project, having reconstructed the genomes of 2,504 individuals from 26 populations using a combination of low-coverage whole-genome sequencing, deep exome sequencing, and dense microarray genotyping. We characterized a broad spectrum of genetic variation, in total over 88 million variants (84.7 million single nucleotide polymorphisms (SNPs), 3.6 million short insertions/deletions (indels), and 60,000 structural variants), all phased onto high-quality haplotypes. This resource includes >99% of SNP variants with a frequency of >1% for a variety of ancestries. We describe the distribution of genetic variation across the global sample, and discuss the implications for common disease studies.

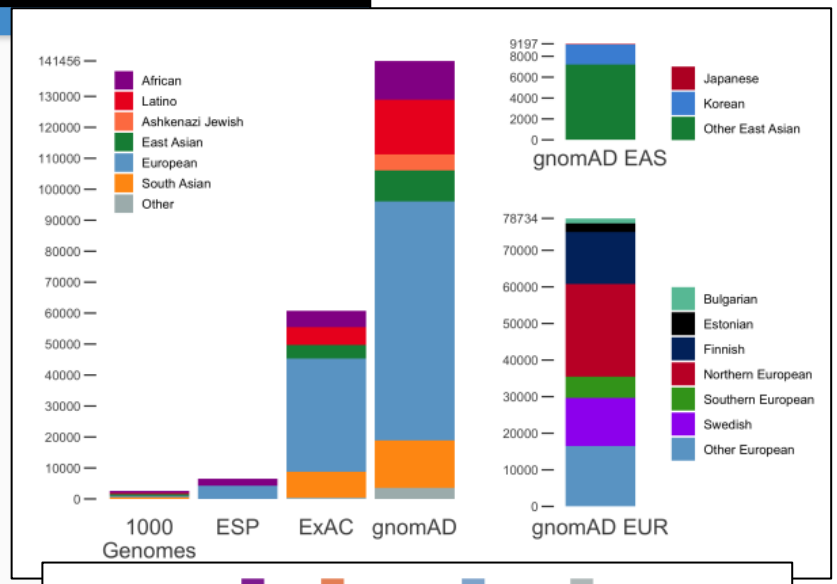
Populations: - African; - American; - East Asian; - European; - South Asian;

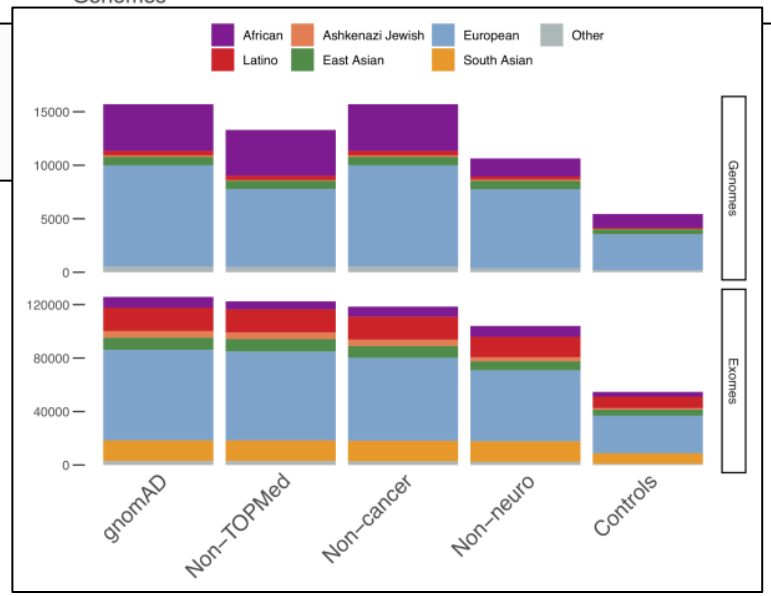
Phase 3	WGS	WExS
Raw bases	89 Tb	18 Tb
Samples	2,504	2,504
Region	Genome	Exome
Mean Depth	8.45x	75x
SNPs	85M	1.5M
Indels	3.6M	22K
Structural Variants	60K	6.5K
Het. Concordance (SNPs)	99.4%	99.8%

http://www.1000genomes.org/about#ProjectSamples ; **Phase 1 n=1092 → Phase 3 n=2504 (26 populations)**

Genome Aggregation Database (gnomAD)







~ 125,700 exomes and ~ 76,156 genomes
8 populations and subpopulations

http://gnomad.broadinstitute.org/
https://macarthurlab.org/2018/10/17/gnomad-v2-1/

Association with pathologies

Table 1 Population, disease-specific, and sequence databases

Population databases	
Exome Aggregation Consortium http://exac.broadinstitute.org/	Database of variants found during exome sequencing of 61,486 unrelated individuals sequenced as part of various disease-specific and population genetic studies. Pediatric disease subjects as well as related individuals were excluded.
Exome Variant Server http://evs.gs.washington.edu/EVS	Database of variants found during exome sequencing of several large cohorts of individuals of European and African American ancestry. Includes coverage data to inform the absence of variation.
1000 Genomes Project http://browser.1000genomes.org	Database of variants found during low-coverage and high-coverage genomic and targeted sequencing from 26 populations. Provides more diversity compared to the Exome Variant Server but also contains lower-quality data, and some cohorts contain related individuals.
dbSNP http://www.ncbi.nlm.nih.gov/snp	Database of short genetic variations (typically ≤50 bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.
dbVar http://www.ncbi.nlm.nih.gov/dbvar	Database of structural variation (typically >50 bp) submitted from many sources.
Disease databases	
ClinVar http://www.ncbi.nlm.nih.gov/clinvar	Database of assertions about the clinical significance and phenotype relationship of human variations.
OMIM http://www.omim.org	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database http://www.hgmd.org	Database of variant annotations published in the literature. Requires fee-based subscription to access much of the content.
Locus/disease/ethnic/other-specific databases	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html Leiden Open Variation Database http://www.lovd.nl	The Human Genome Variation Society site developed a list of thousands of databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the Leiden Open Variation Database system.
DECIPHER http://decipher.sanger.ac.uk	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.
Sequence databases	
NCBI Genome http://www.ncbi.nlm.nih.gov/genome	Source of full human genome reference sequences.
RefSeqGene http://www.ncbi.nlm.nih.gov/refseq/rsg	Medically relevant gene reference sequence resource.
Locus Reference Genomic (LRG) http://www.lrg-sequence.org	
MitoMap http://www.mitomap.org/MITOMAP/HumanMitoSeq	Revised Cambridge reference sequence for human mitochondrial DNA.

Data Bases

Genetics in Medicine, 2015

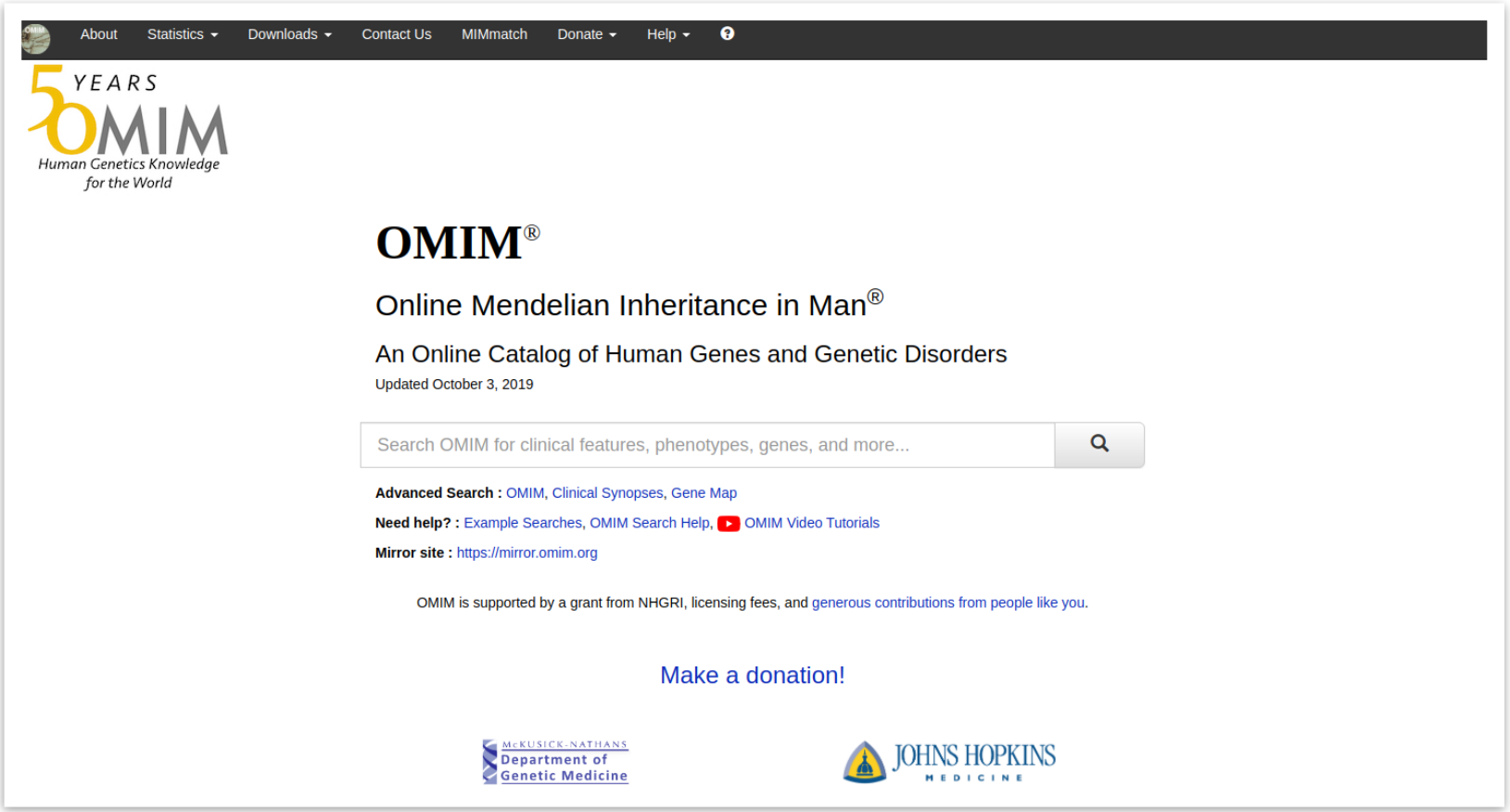
OMIM

https://omim.org/

~ 15.000 **human genes**

Mendelian disorders

Susceptibility to cancer and other **complex diseases**



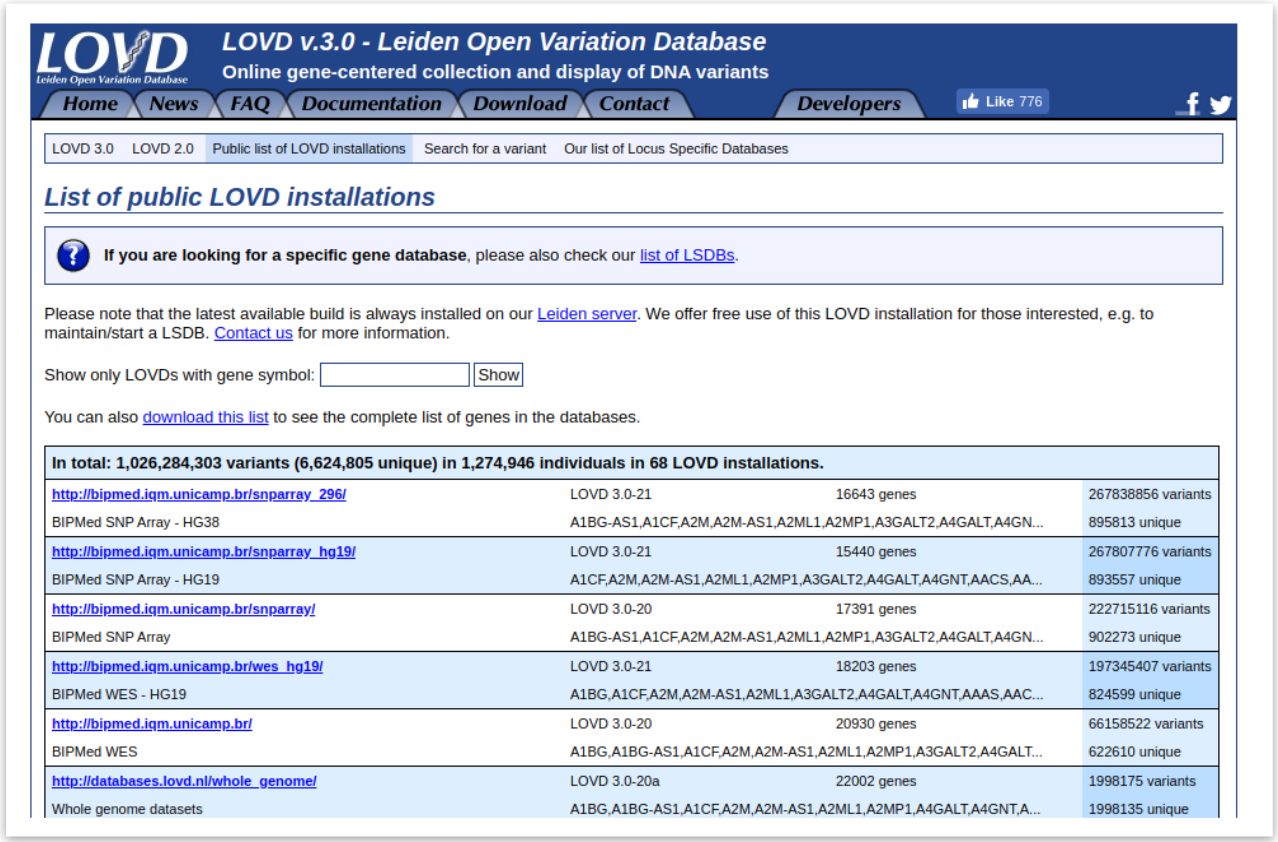
LOVD

http://www.lovd.nl

Collection of databases with variants.

There is **one database for each gene** with its variant list.

Several genes group together in **"installations"**, some of them concerning diseases.



LOVD v.3.0 - Leiden Open Variation Database

Online gene-centered collection and display of DNA variants

Home | News | FAQ | Documentation | Download | Contact | Developers | Like 776

LOVD 3.0 | LOVD 2.0 | Public list of LOVD installations | Search for a variant | Our list of Locus Specific Databases

List of public LOVD installations

If you are looking for a specific gene database, please also check our list of LSDBs.

Please note that the latest available build is always installed on our Leiden server. We offer free use of this LOVD installation for those interested, e.g. to maintain/start a LSDB. Contact us for more information.

Show only LOVDs with gene symbol: [ ] Show

You can also download this list to see the complete list of genes in the databases.

In total: 1,026,284,303 variants (6,624,805 unique) in 1,274,946 individuals in 68 LOVD installations.			
http://bipmed.iqm.unicamp.br/snparray_296/	LOVD 3.0-21	16643 genes	267838856 variants
BIPMed SNP Array - HG38	A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GN...		895813 unique
http://bipmed.iqm.unicamp.br/snparray_hg19/	LOVD 3.0-21	15440 genes	267807776 variants
BIPMed SNP Array - HG19	A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GNT,AACS,AA...		893557 unique
http://bipmed.iqm.unicamp.br/snparray/	LOVD 3.0-20	17391 genes	222715116 variants
BIPMed SNP Array	A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GN...		902273 unique
http://bipmed.iqm.unicamp.br/wes_hg19/	LOVD 3.0-21	18203 genes	197345407 variants
BIPMed WES - HG19	A1BG,A1CF,A2M,A2M-AS1,A2ML1,A3GALT2,A4GALT,A4GNT,AAAS,AAC...		824599 unique
http://bipmed.iqm.unicamp.br/	LOVD 3.0-20	20930 genes	66158522 variants
BIPMed WES	A1BG,A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT...		622610 unique
http://databases.lovd.nl/whole_genome/	LOVD 3.0-20a	22002 genes	1998175 variants
Whole genome datasets	A1BG,A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A4GALT,A4GNT,A...		1998135 unique

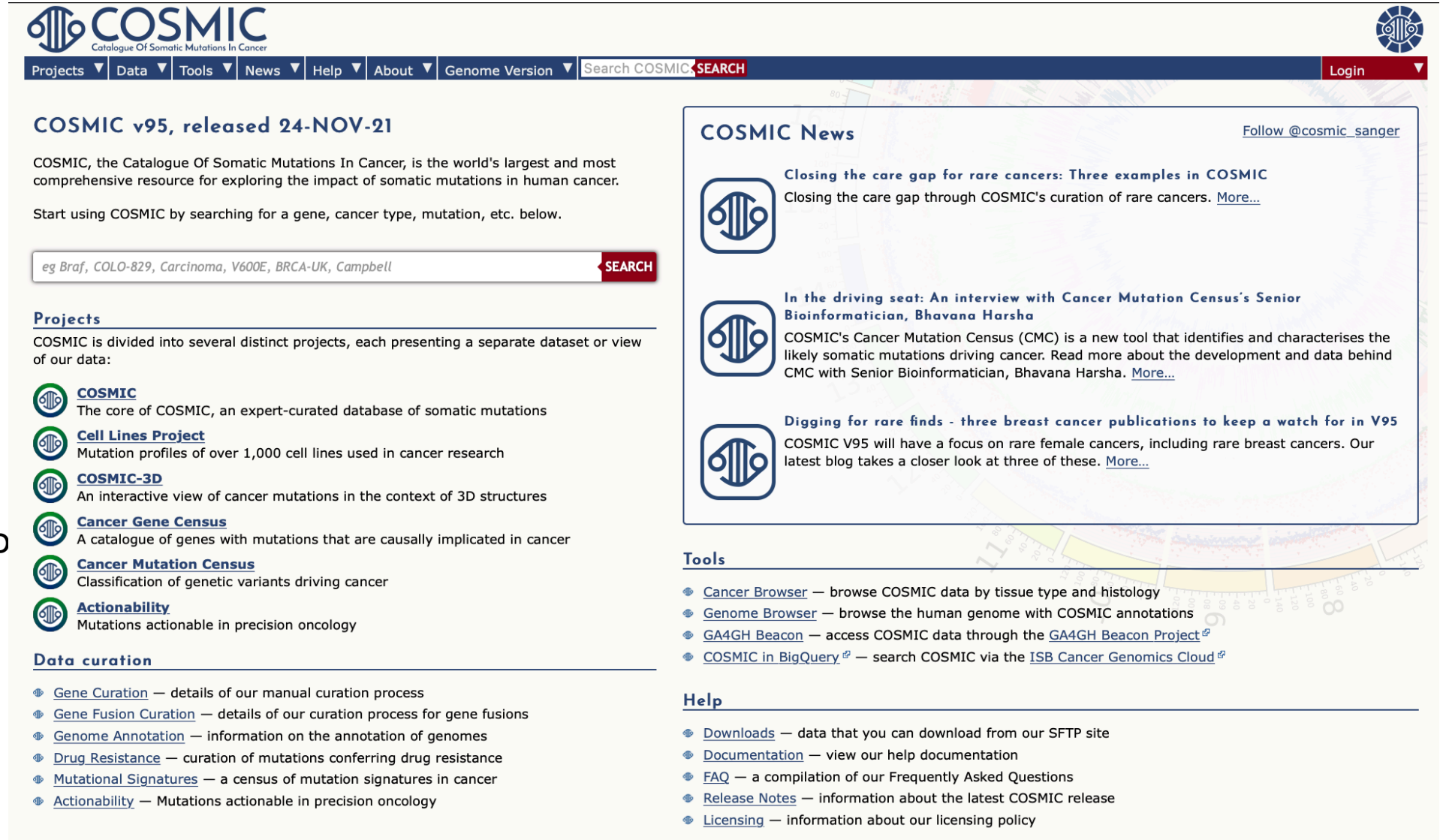
http://bipmed.iqm.unicamp.br/neurofibromatosis/	LOVD 3.0-21	2 genes	66 variants
BIPMed - Neurofibromatosis	NF1,NF2		66 unique

COSMIC

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

Expert Curation Data from peer reviewed publications

Genome-wide screen data from publications o consortiums



Catalog of somatic mutations in cancer

Clinical Implications of variants: ClinVar

Variation **Condition** **Significance**

	Variation *Location*	Gene(s)	Condition(s)	Frequency	Clinical significance (Last reviewed)	Review status
19.	NM_001005862.2(**ERBB2**):c.1376C>T (p.Pro459Leu) *GRCh37:* Chr17:37872145 *GRCh38:* Chr17:39715892	ERBB2	not specified	GMAF:0.00040(T)	not provided (Sep 19, 2013)	no assertion provided
20.	NM_001005862.2(**ERBB2**):c.1703C>A (p.Ala568Asp) *GRCh37:* Chr17:37873628 *GRCh38:* Chr17:39717375	ERBB2	not specified		not provided (Sep 19, 2013)	no assertion provided
21.	NM_001005862.2(**ERBB2**):c.1870A>G (p.Ile624Val) *GRCh37:* Chr17:37879585 *GRCh38:* Chr17:39723332	ERBB2	**ERBB2** POLYMORPHISM, not specified	GO-ESP:0.00707(G) GMAF:0.00260(G)	Benign (Feb 1, 1993)	no assertion criteria provided
22.	NM_001005862.2(**ERBB2**):c.1873A>G (p.Ile625Val) *GRCh37:* Chr17:37879588 *GRCh38:* Chr17:39723335	ERBB2	**ERBB2** POLYMORPHISM, not specified	GO-ESP:0.16854(G) GMAF:0.12140(G)	Benign (Feb 1, 1993)	no assertion criteria provided
23.	NM_001005862.2(**ERBB2**):c.2173_2174 delTTinsCC (p.Leu725Pro) *GRCh37:* Chr17:37880219-37880220 *GRCh38:* Chr17:39723966-39723967	ERBB2	Adenocarcinoma of lung		Pathogenic (Sep 30, 2004)	no assertion criteria provided

- Benign
- Likely benign
- Uncertain significance
- Likely pathogenic
- Pathogenic
- drug response
- Association
- risk factor
- Protective
- Affects
- conflicting data from submitters
- Other
- not provided
- … and more

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

Annotation Tools

Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

ANNOVAR

Variant Effect Predictor

VarAFT

SnpEff

	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
	Multiple proteins, dbSNP ID, NCBI GI number, protein sequence, protein sequence alignment, Pileup, VCF4, maq, soap, gff3, casava, cg	XXX in SIFT format	yes	no	no	no	yes	yes	
del [145]			no	yes	no	no	no	yes	
[146]	FASTA, MAF, PHYLIP		-	-	-	no	yes	no	
[147]	AA in FASTA, substitutions format	Web report	yes	no	no	no	no	yes	This too interfac
Function [148]	RefSNP Ids, OMIM Ids	Web report	yes	no	no	no	no	yes	
Domain			yes	no	no	-	-	-	Not ava
be [150]	Gene/protein symbol, FASTA	Web report	yes	no	no	no	no	yes	The pro SNAP a augmer informa
ect 4.0 [151]	FASTA, PDB file, PDB ID, UniProt ID	Web report	yes	no	no	no	no	yes	This too informa
nter [152]	Gene symbol; dbSNP ID;	Web report	yes	no	no	yes	no	no	
s [153]	CSV in SNPnexus input format	CSV in SNPnexus output format	yes	yes	yes	no	no	yes	Outputs
54]	dbSNP ID, TSC ID, position	Web report	yes	no	no	no	no	yes	
55]	One letter residue code; Swiss-Prot protein code; Sequence file; GO terms; CSV	Web report	yes	no	no	no	no	yes	Predicts reliabilit
	Gene symbol, SNP ID	Web report	yes	no	no	no	no	yes	
	-	-	-	-	-	-	-	-	Tool tha based p is not a
		-	-	-	-	no	no	yes	No long
	SNPs, insertions, s, and MNPs	CSV	yes	yes	no	no	yes	no	A collec files;
	n-values,	Web report	yes	no	no	no	no	yes	Ou
	nce; protein gene	Web report	yes	no	no	no	no	yes	
		-	-	-	-	-	-	-	
		Web report	yes	no	no	no			

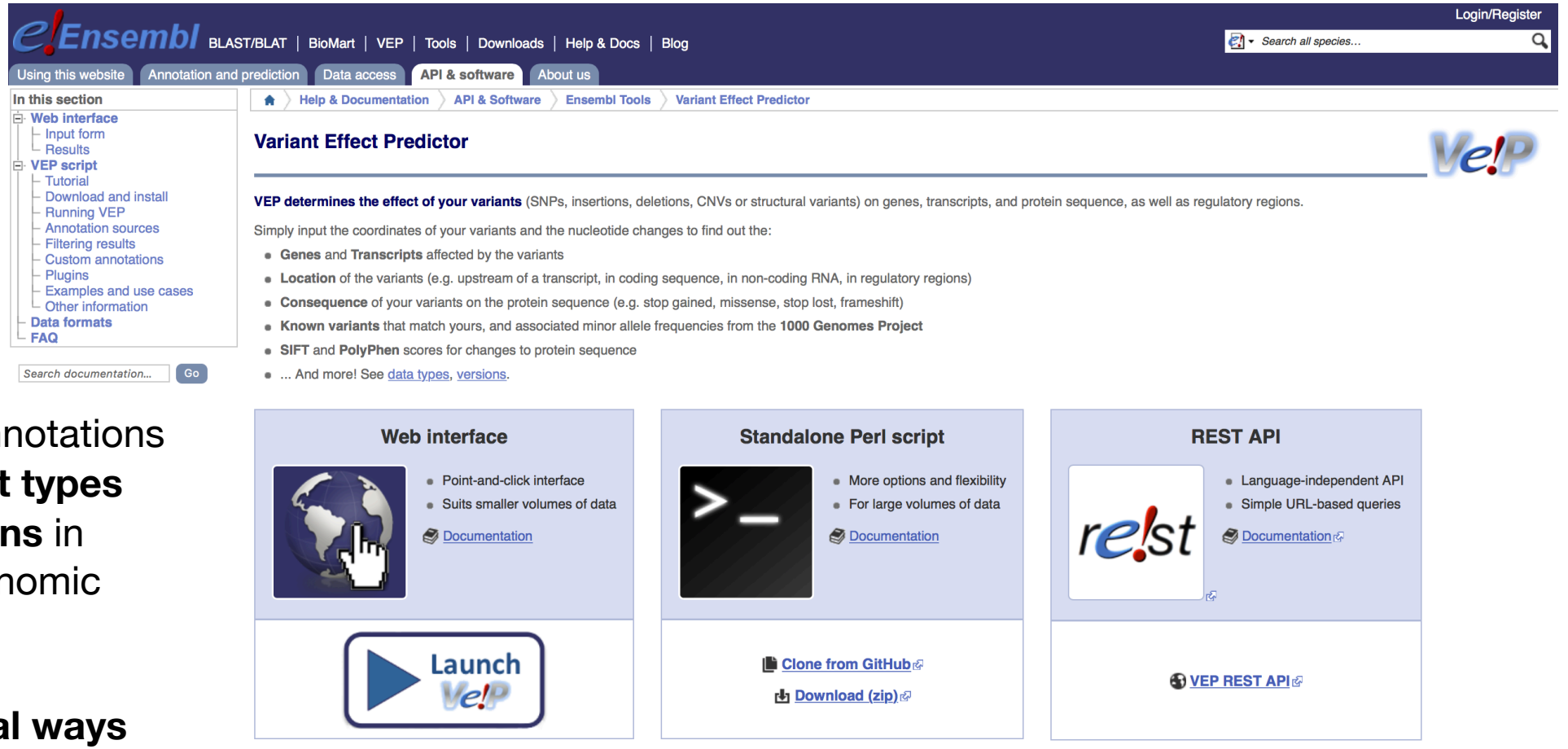
	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
PANTHER [124]	Protein sequence and Substitution	subPSEC score	yes	no	no	no			
Parepro [125]	Protein sequence and Substitution	-	yes	no	no	no			
PESX [126]	plain sequence; FASTA	Web report	-	-	-	no			
pfSNP [127]	SNP ID; chromosome region; Gene ID;	Web report	yes	no	no	no			
PHAST [128]	FASTA, PHYLIP, MPM, MAF, SS	Conservation score	-	-	-	no			
PhD-SNP [129]	Protein sequence; Swiss-Prot protein code; Sequence file	Effect prediction	yes	no	no	no			
PMUT [130]	FASTA sequence file; SWISSProt code	Web report	yes	no	no	no			
PolyDoms [131]	Gene/protein symbol(s), RefSeqID dbSNP ID	Web report	yes	no	no	no			
PolyMAPr [132]	-	-	-	-	-	-			
PolyPhen-2 [133]	UniProt ID, FASTA, dbSNP ID	CSV in PolyPhen format	yes	no	no	no			
PupaSNP Finder [134]	dbSNP ID, GenBank/Ensembl ID; PED format	Web report	yes	no	no	no			
QuickSNP [135]	gene/protein ID, GO term, gene symbol	Web report	yes	no	no	no			
RescueESE [136]	plain text; multiple FASTA	predicts sequences with ESE activity	-	-	-	no			
SAPRED [137]	FASTA and		yes	no	no	no			
SCAN [138]		Web report	yes	no	no	no			
SCONE [139]	MAF	Conservation score	-	-	-	no			
SeattleSeq Annotation [140]	Maq, GFFm CASAVA, VCF, GATK bed	VCF, own format	yes	yes	no	yes			
SeqAnt [141]	FASTA sequence file	Web report	yes	yes	no	no	no	yes	
SeqProfCod [142]	-	-	yes	no	no	-	-	-	
SVA (Sequence Variant Analyser) [143]	VCF of variants, project file (for command line version)	--potential biological function --dbSNP/Kegg/GO/1000 Genomes/DGV	yes	yes	no	yes	yes	no	

	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
		Plot showing the Purity/Ploidy, R data file	yes	no	yes				
	stitutions list	Web report	yes	no	no	no			
	mplete Genomics, OLiD, CSV in ar format;	Gene-based annotation; Region-based annotations; Filter-based annotation. For all categories	yes	yes	yes	no			
	F, pileup, CSV	VCF	yes	yes	yes	no	yes	n	
	DB ID, Chain, Mutation	Web report	yes	no	no	no	no	ye	
	dbSNP ID, population	Web report	yes	no	no	no	no	yes	
	Passenger mutation rates, AA changes	CSV including CHASM score, p-value, and FDR	yes	no	no	no	yes	no	
	PDB ID; PDB file format	Web report	yes	no	no	no	no	yes	
			yes	no	no	no	yes	no	
	SV, VCF, Pileup, HGVS, riant Identifiers	Web report	yes	no	no	no	yes	ye	
	STA	Web report, CSV	-	-	-	no	no		
	equence; FASTA	Web report	-	-	-	no	no		
	ormat; or variation n via web interface	Web report, CSV	yes	no	no	no	no		
	hSNP ID	Web report	yes	no	no	no			
		Web report	yes	no	no				

	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
		Plot showing the Purity/Ploidy, R data file	yes	no	yes	no			
	stitutions list	Web report	yes	no	no	no	no	yes	
	mplete Genomics, OLiD, CSV in var format;	Gene-based annotation; Region-based annotations; Filter-based annotation. For all categories	yes	yes	yes	no	yes	no	
	VCF, pileup, CSV	VCF	yes	yes	yes	no	yes	no	
	PDB ID, Chain, Mutation	Web report	yes	no	no	no	no	yes	
	dbSNP ID, population	Web report	yes	no	no	no	no	yes	
	Passenger mutation rates, AA changes	CSV including CHASM score, p-value, and FDR	yes	no	no	no	yes	no	
	PDB ID; PDB file format	Web report	yes	no	no	no	no	yes	
			yes	no	no	no	yes	no	
	CSV, VCF, Pileup, HGVS, Variant Identifiers	Web report	yes	no	no	no	yes	yes	
	FASTA	Web report, CSV	-	-	-	no	no	yes	
	plain sequence; FASTA	Web report	-	-	-	no	no	yes	
	FASTA format; or variation information via web interface	Web report, CSV	yes	no	no	no	no	yes	
	Gene Symbol, dbSNP ID	Web report	yes	no	no	no	no	yes	
	Gene name	Web report	yes	no	no	no	no	yes	

Variant Effect Predictor (VEP)

http://www.ensembl.org/info/docs/tools/vep/index.html



Provides annotations for **different types of alterations** in different genomic locations

With **several ways of execution**

Standalone execution

```
vep --af --af_1kg --af_esp --af_gnomad --appris --assembly GRCh37 --biotype --cache --canonical --ccds --
compress_output gzip --dir_cache /home/epineiro/Programs/VEP/VEP104/.vep/ --domains --fasta /home/
epineiro/Programs/VEP/VEP104/.vep/homo_sapiens/104_GRCh37/
Homo_sapiens.GRCh37.75.dna.primary_assembly.fa.gz --force_overwrite --fork 4 --hgvs --input_file out/
filtered/all.vcf.gz --numbers --offline --output_file out/annotated/all.vep.vcf.gz --polyphen b --port
3337 --protein --regulatory --sift b --symbol --tsl --uniprot --variant_class --vcf –xref_refseq
```

Configuration options

Input, output and format

Annotations

Plugins

http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html#basic

VEP standalone script output: html

Links

- Top of page
- VEP run statistics
- General statistics
- Variant classes
- Consequences (most severe)
- Consequences (all)
- Coding consequences
- SIFT summary
- PolyPhen summary
- Variants by chromosome
- Position in protein

VEP run statistics

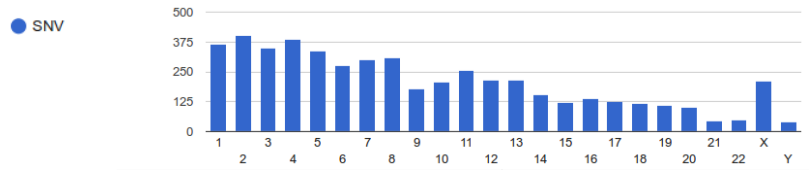
VEP version (API)	85 (85)
Cache/Database	/home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/.vep/homo_sapiens/85_GRCh37
Species	homo_sapiens
Command line options	`--format vcf --sift b --polyphen b --ccds --uniprot --hgvs --symbol --numbers --domains --regulatory --canonical --protein --biotype --uniprot --tsl --gmaf --variant_class --xref_refseq --maf_1`
Start time	2016-09-03 09:01:35
End time	2016-09-03 09:43:38
Run time	2523 seconds
Input file (format)	/home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf (VCF)
Output file	/home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf_output_VEP.txt [text]

General statistics

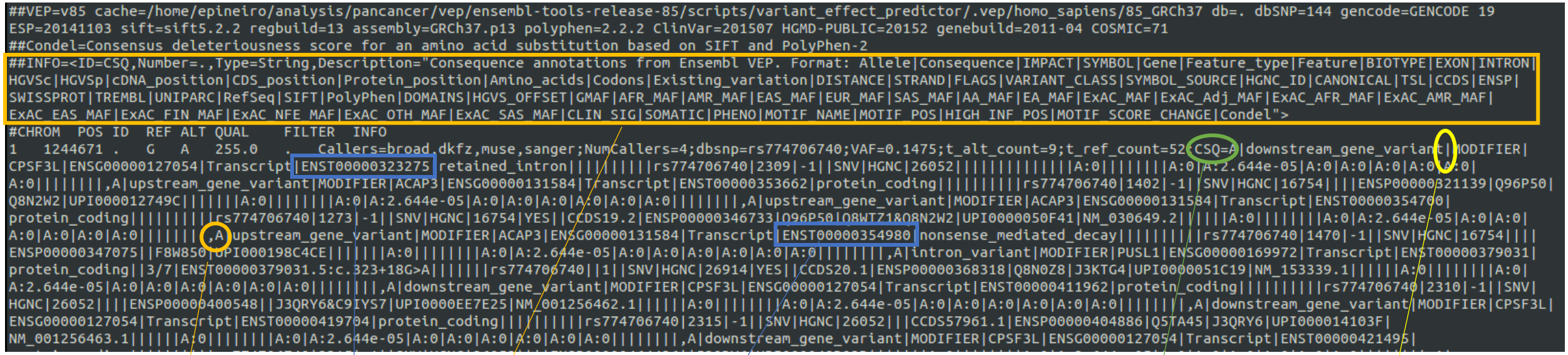
Lines of input read	5025
Variants processed	5013
Variants remaining after filtering	5013
Lines of output written	5013
Novel / existing variants	4273 (85.2%) / 740 (14.8%)
Overlapped genes	2761
Overlapped transcripts	10274
Overlapped regulatory features	490

Variant classes

Variants by chromosome



VEP standalone script output: VCF



Annotations per transcript are separated by comma

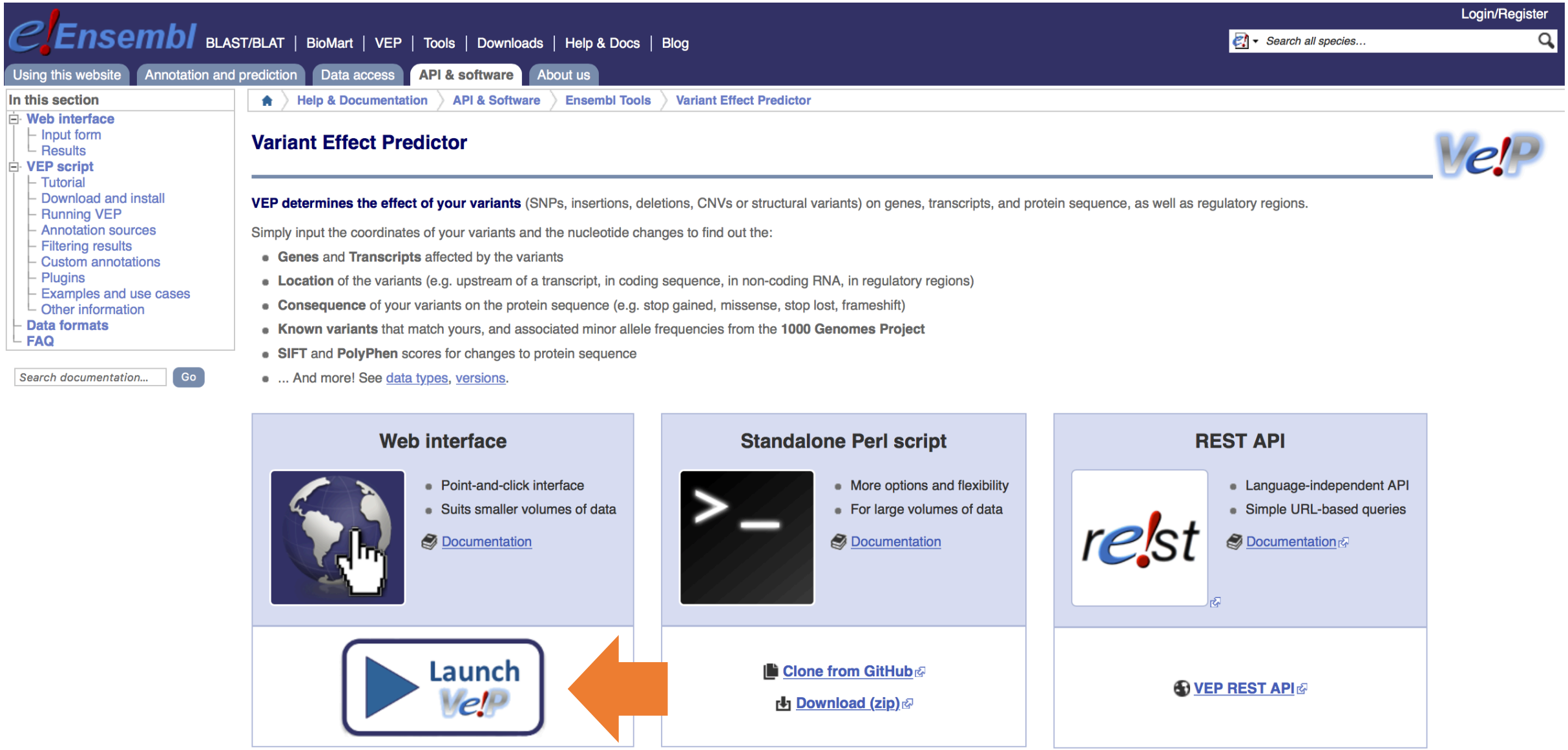
Order of annotations in INFO column

Annotations starts with **CSQ=** tag

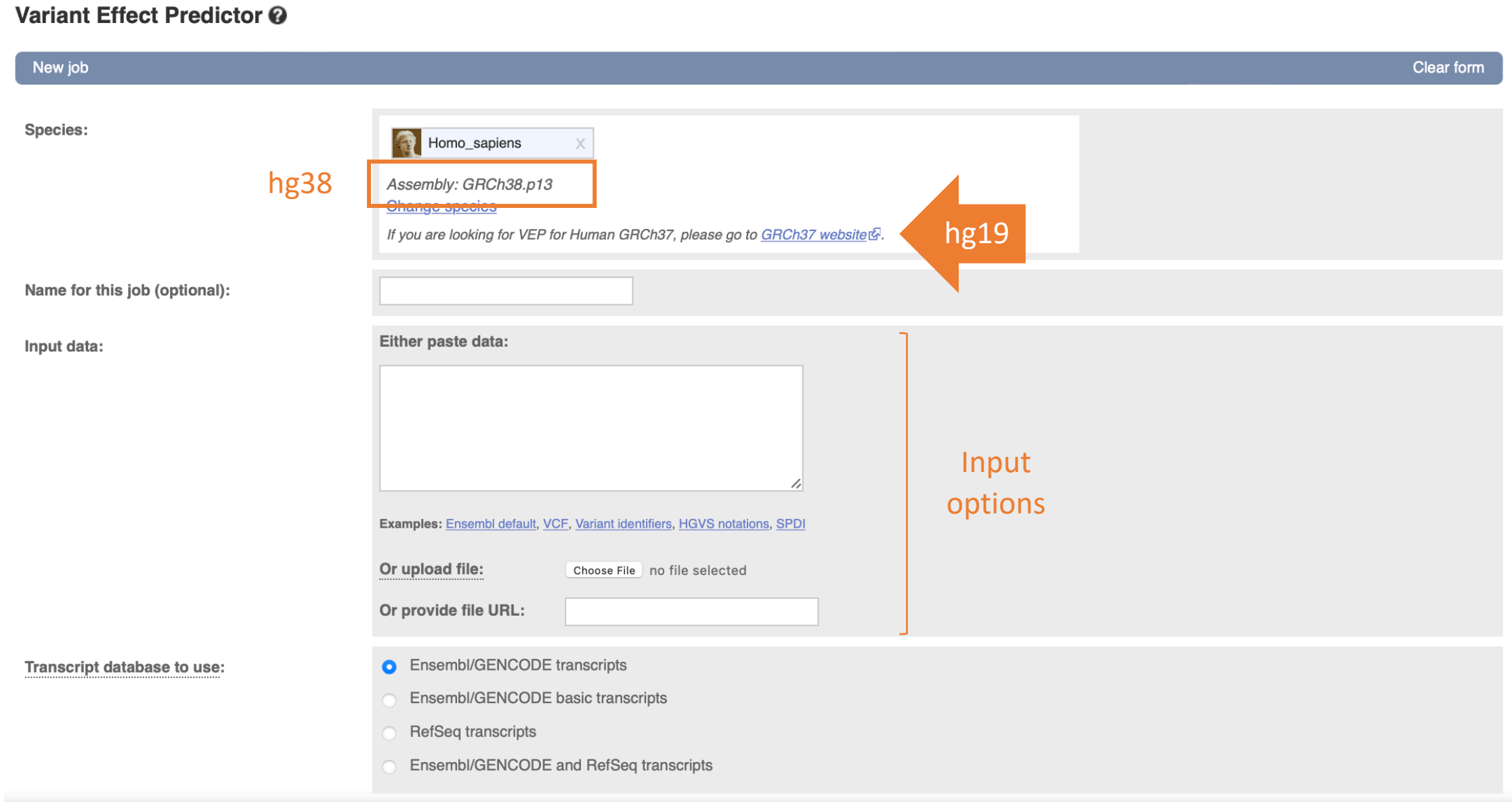
Each individual annotation is separated by |

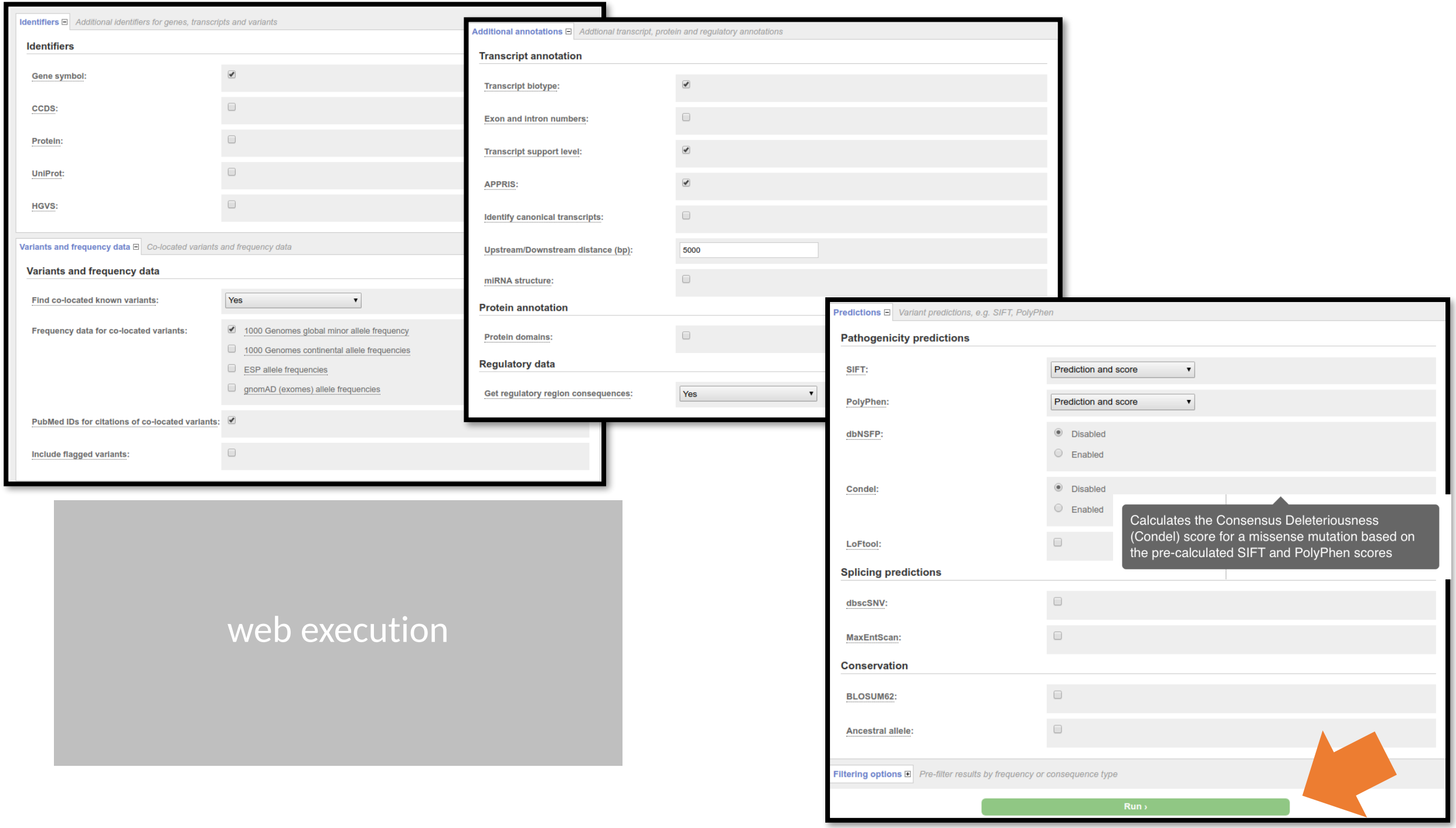
One set of annotations per transcript

web execution



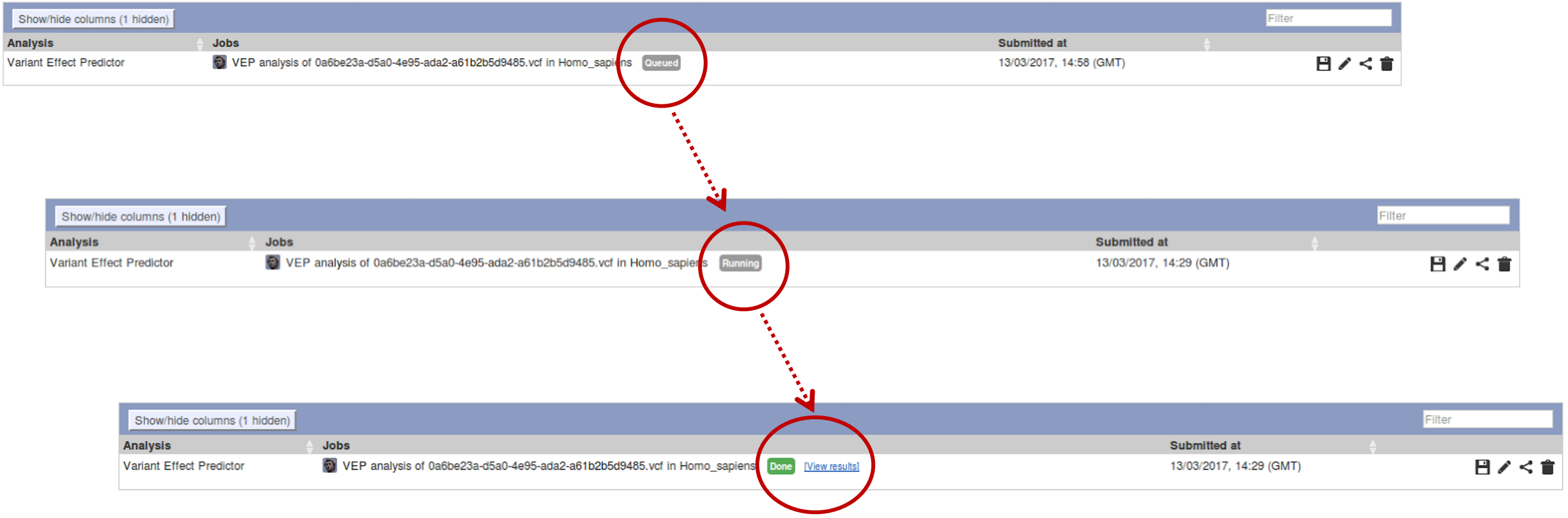
web execution: options



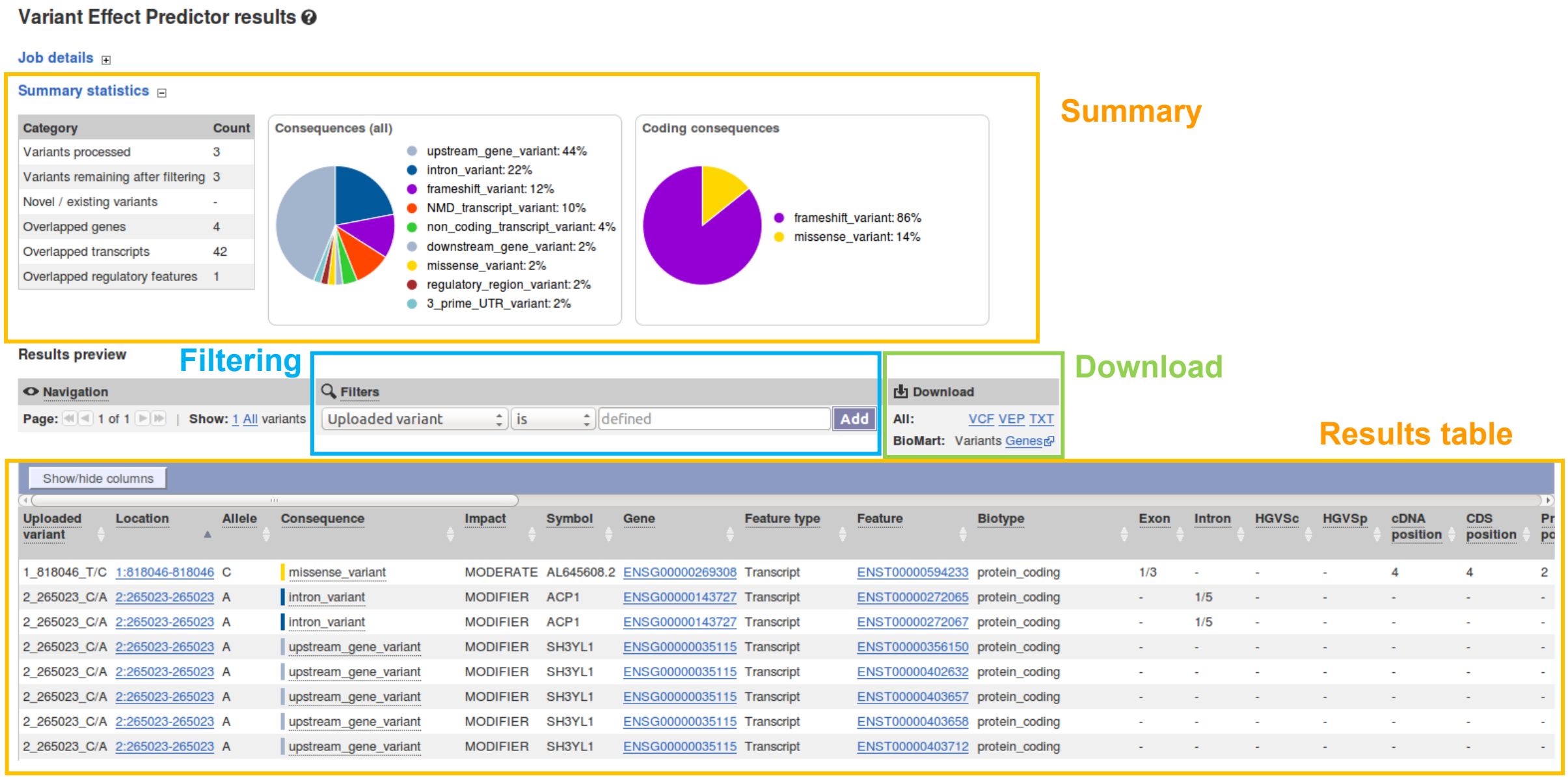


web execution

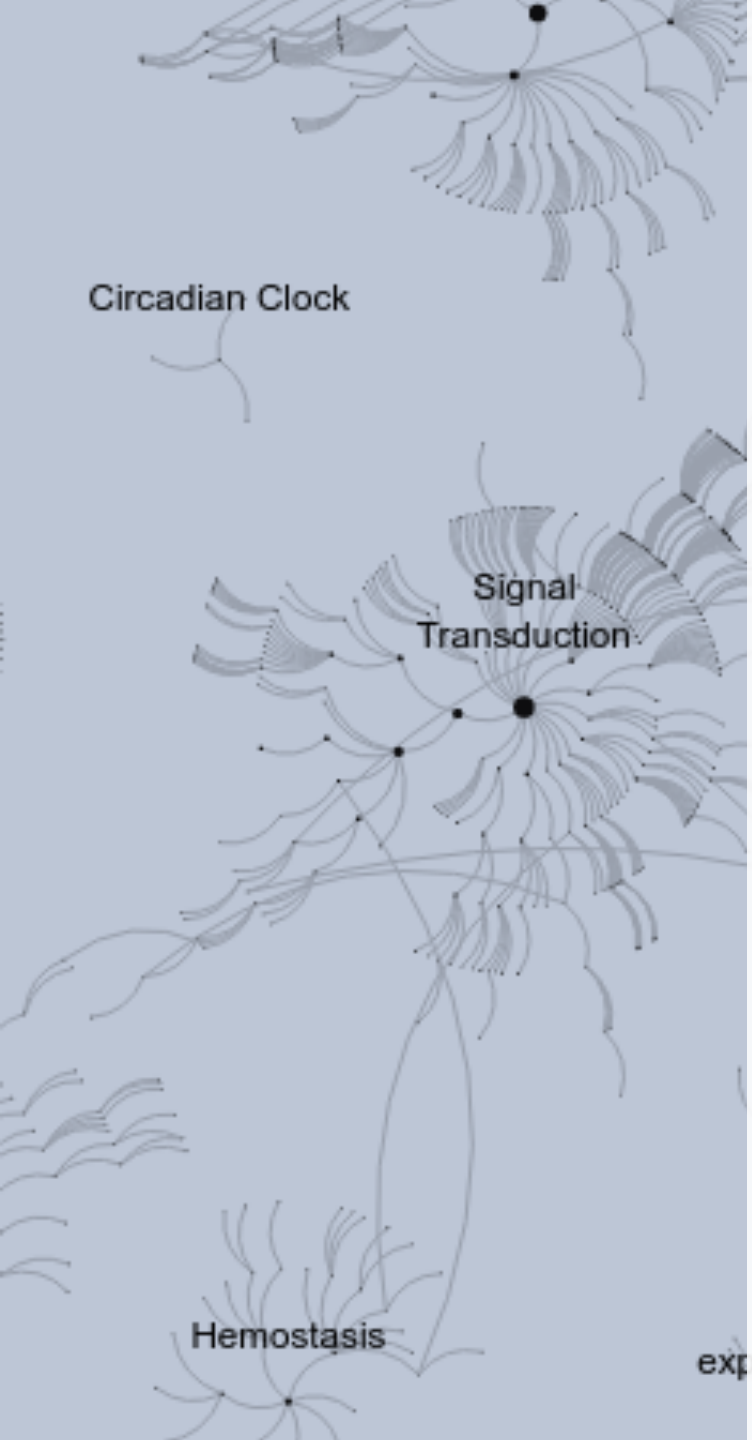
web execution



web execution: output



Extended information



Pathways

- Implication of genes in biological process
- Interaction among proteins -> Therapeutical actions

- **Resources**
  - Reactome
  - KEGG (Kyoto Encyclopedia of Genes and Genomes)

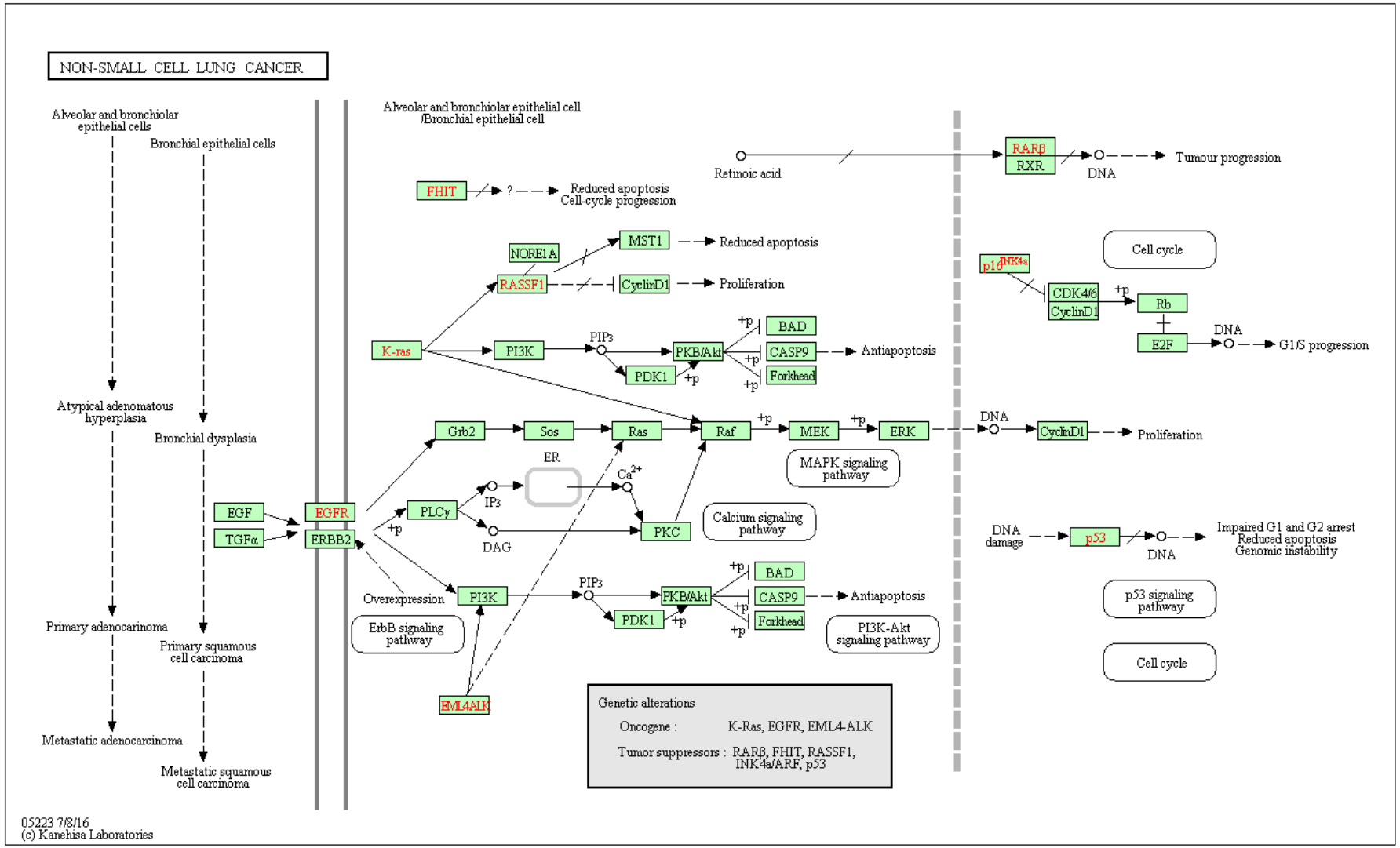
KEGG pathways

https://www.genome.jp/kegg/pathway.html

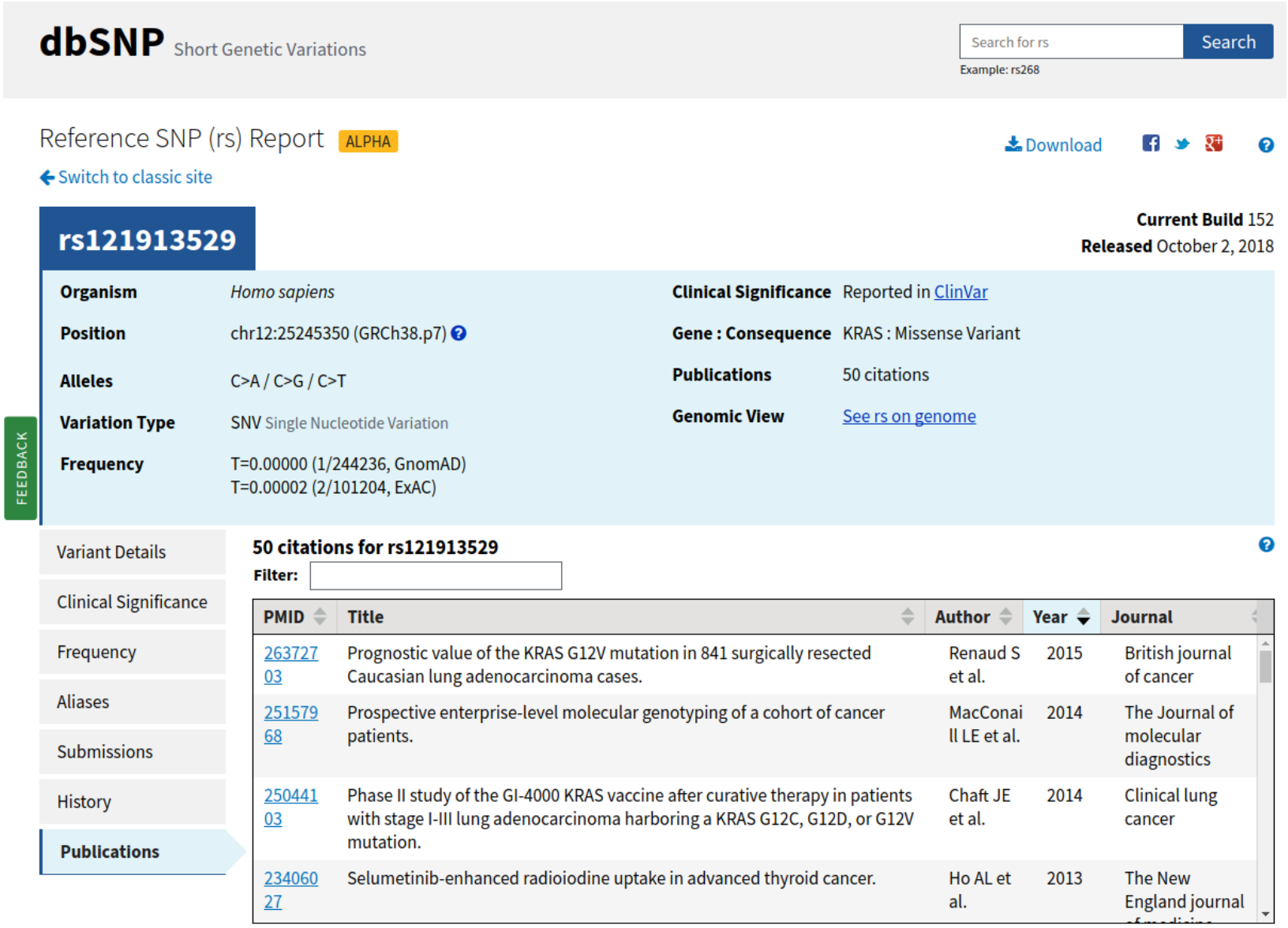
KEGG PATHWAY Database

Wiring diagrams of molecular interactions, reactions, and relations

1. **Metabolism**
2. **Genetic Information Processing**
3. **Environmental Information Processing**
4. **Cellular Processes**
5. **Organismal Systems**
6. **Human diseases**
7. **Drug development (structural relations between compounds)**



Bibliography



dbSNP Short Genetic Variations

Search for rs — Search

Example: rs268

Reference SNP (rs) Report ALPHA

Download

Switch to classic site

rs121913529

Current Build 152
Released October 2, 2018

Organism	*Homo sapiens*	**Clinical Significance**	Reported in ClinVar
Position	chr12:25245350 (GRCh38.p7)	**Gene : Consequence**	KRAS : Missense Variant
Alleles	C>A / C>G / C>T	**Publications**	50 citations
Variation Type	SNV Single Nucleotide Variation	**Genomic View**	See rs on genome
Frequency	T=0.00000 (1/244236, GnomAD) T=0.00002 (2/101204, ExAC)		

FEEDBACK

Variant Details
Clinical Significance
Frequency
Aliases
Submissions
History
Publications

50 citations for rs121913529

Filter:

PMID	Title	Author	Year	Journal
26372703	Prognostic value of the KRAS G12V mutation in 841 surgically resected Caucasian lung adenocarcinoma cases.	Renaud S et al.	2015	British journal of cancer
25157968	Prospective enterprise-level molecular genotyping of a cohort of cancer patients.	MacConaill LE et al.	2014	The Journal of molecular diagnostics
25044103	Phase II study of the GI-4000 KRAS vaccine after curative therapy in patients with stage I-III lung adenocarcinoma harboring a KRAS G12C, G12D, or G12V mutation.	Chaft JE et al.	2014	Clinical lung cancer
23406027	Selumetinib-enhanced radioiodine uptake in advanced thyroid cancer.	Ho AL et al.	2013	The New England journal

Bibliography

UniProtKB - P00533 (EGFR_HUMAN)

Protein | **Epidermal growth factor receptor**

Gene | **EGFR**

Organism | *Homo sapiens (Human)*

Status | Reviewed - Annotation score: ●●●●● - Experimental evidence at protein level[i]

http://www.uniprot.org

PTM / Processing[i]

Molecule processing

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Signal peptide[i]	1 – 24	24	2 Publications			Add BLAST
Chain[i]	25 – 1210	1186	Epidermal growth factor receptor		PRO_0000016665	Add BLAST

Amino acid modifications

Feature key	Position(s)	Length	Description	Graphical view	Feature identifier	Actions
Disulfide bond[i]	31 ↔ 58					
Glycosylation[i]	56 – 56	1	N-linked (GlcNAc...) (complex); atypical; partia 5 Publications		CAR_000227	
Glycosylation[i]	73 – 73	1	N-linked (GlcNAc...); atypica 1 Publication			
Glycosylation[i]	128 – 128	1	N-linked (GlcNAc..) 3 Publications			
Disulfide bond[i]	157 ↔ 187					
Glycosylation[i]	175 – 175	1	N-linked (GlcNAc..) 5 Publications			
Disulfide bond[i]	190 ↔ 199					
Disulfide bond[i]	194 ↔ 207					
Glycosylation[i]	196 – 196	1	N-linked (GlcNAc..) 4 Publications			

Specific cancer information

Disease databases	
ClinVar http://www.ncbi.nlm.nih.gov/clinvar	Database of assertions about the clinical significance and phenotype relationship of human variations.
OMIM http://www.omim.org	Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.
Human Gene Mutation Database http://www.hgmd.org	Database of variant annotations published in the literature. Requires fee-based subscription to access much of the content.
Locus/disease/ethnic/other-specific databases	
Human Genome Variation Society http://www.hgvs.org/dblist/dblist.html	The Human Genome Variation Society site developed a list of thousands of databases that provide variant annotations on specific subsets of human variation. A large percentage of databases are built in the Leiden Open Variation Database system.
Leiden Open Variation Database http://www.lovd.nl	
DECIPHER http://decipher.sanger.ac.uk	A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.

Genet Med. 2015 May;17(5):405-24

Frequency of the variant/alteration of the gene in the disease

Association of the alteration with the disease

TCGA

- Program History +
- TCGA Cancers Selected for Study
- Publications by TCGA
- Using TCGA +
- Contact

The Cancer Genome Atlas Program

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33 cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Over the next dozen years, TCGA generated over 2.5 petabytes of genomic, epigenomic, transcriptomic, and proteomic data. The data, which has already led to improvements in our ability to diagnose, treat, and prevent cancer, will remain publicly available for anyone in the research community to use.

ICGC Data Portal

Quick Search

NCER PROJECTS

Summary | Details | History

Liver 7
Brain 6
Pancreas 6

Donor Distribution

24,289 Donors across 86 Projects

Top 20 Mutated Cancer Genes with High Functional Impact SSMs

19,729 Unique SSM-Tested Donors

Donors Affected

TP53 LRP1B KMT2C BRAF APC KMT2D PTEN ARID1A CREBBP ATRX NOTCH1 KMT2A EP300 IDH1 CDKN2A FLNA SETD2 NSD1 SMAD4 COL3A1

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Manage Sets 0

Harmonized Cancer Datasets

Genomic Data Commons Data Portal

Get Started by Exploring:

Projects | Exploration | Analysis | Repository

e.g. BRAF, Breast, TCGA-BLCA, TCGA-A5-A0G2

Cases by Maj

Adrenal Gland
Bile Duct
Bladder
Bone
Bone Marrow
Brain
Breast
Cervix

Data Portal Summary

Data Release 30.0 - September 23, 2021

PROJECTS	PRIMARY SITES	CASES
70	67	85,414

FILES	GENES	MUTATIONS
619,488	23,621	3,599,319

PCAWG - PANCANCER ANALYSIS OF WHOLE GENOMES

The Pan-Cancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,600 cancer whole genomes from the International Cancer Genome Consortium. Building upon previous work which examined cancer coding regions (Cancer Genome Atlas Research Network, The Cancer Genome Atlas Pan-Cancer analysis project, Nat. Genet. 2013 45:1113, Cell. 2018 Apr 5;173(2):283-285), this project explored the nature and consequences of somatic and germline variations in both coding and non-coding regions, with specific emphasis on cis-regulatory sites, non-coding RNAs, and large-scale structural alterations.

In order to facilitate the comparison among diverse tumor types, all tumor and matched normal genomes have been subjected to a uniform set of alignment and variant calling algorithms, and must pass a rigorous set of quality control tests. The research activities have been coordinated by a series of working groups comprising more than 700 scientists.

Donor Distribution by Primary Site

48 projects and 20 primary sites

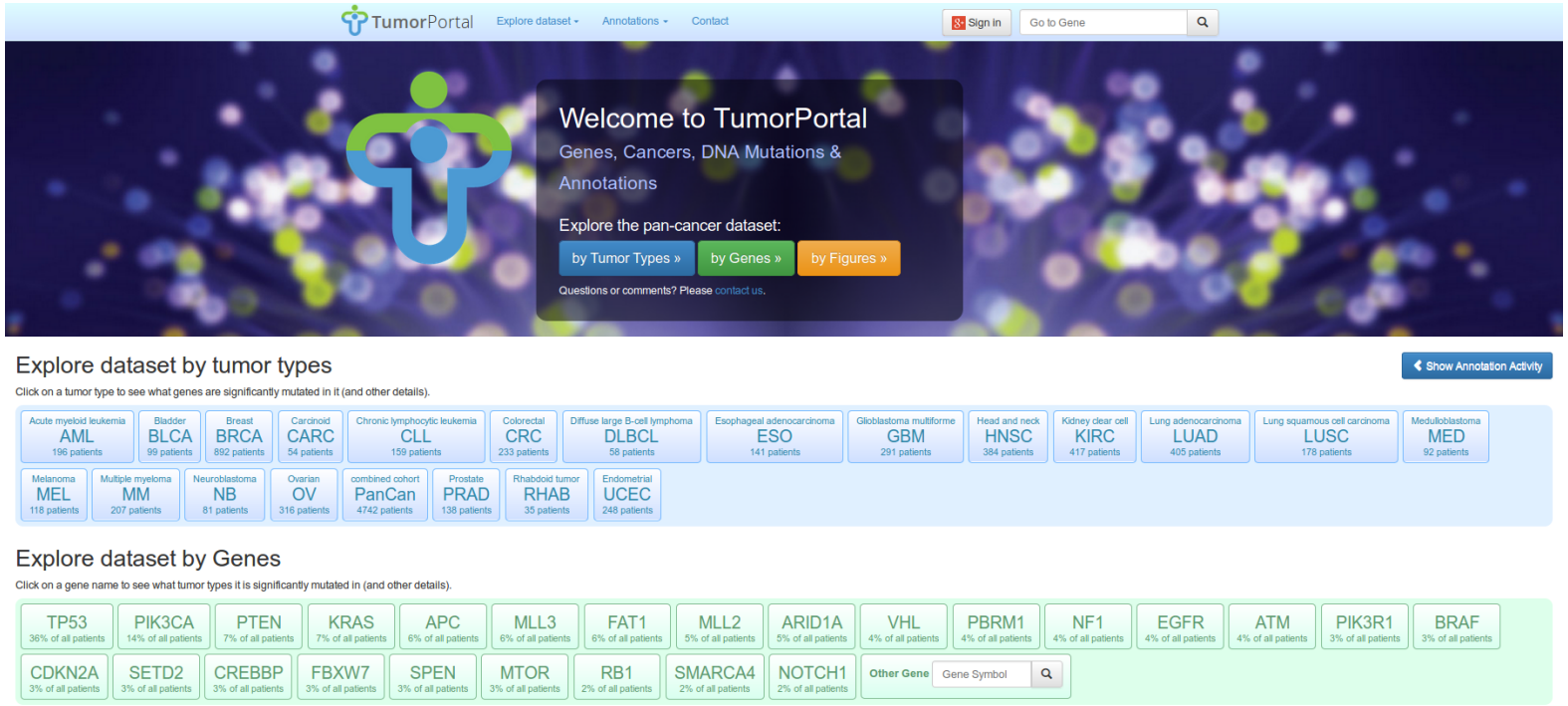
Donors

Liver Pancreas Brain Blood Kidney Prostate Breast Ovary Skin Head and neck Esophagus Lung Stomach Bone Colorectal Uterus Mesenchymal Bladder Cervix Gall Bladder

TumorPortal

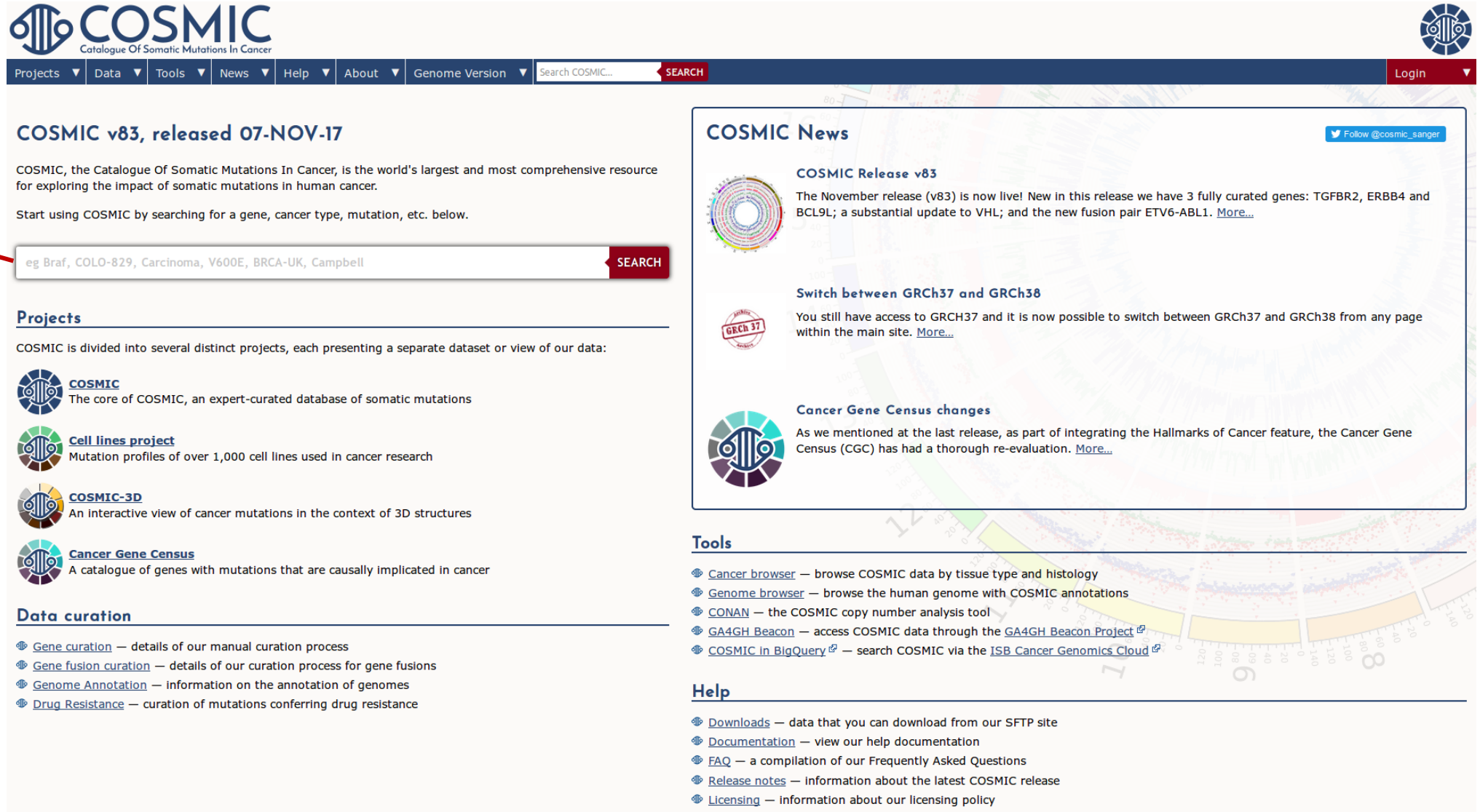
- Somatic mutations in exome sequencing of 4742 human cancer across 21 different tumor types
- Known cancer related genes and genes not previously involved in cancer (apoptosis, proliferation, ...) detected
- Gene classification **in each tumor type** in:
  - **Highly significantly mutated**
  - **Significantly mutated**
  - **Near significance**

http://www.tumorportal.org/



ICGC-TCGA Consortium

variant/gene frequency in database



COSMIC
Catalogue Of Somatic Mutations In Cancer

Projects | Data | Tools | News | Help | About | Genome Version | Search COSMIC... SEARCH | Login

COSMIC search results

Your search term "**pten**" was an exact match for the COSMIC gene **PTEN**.

A search of the whole COSMIC database returned results in **3** sections of the database. More...

Genes (1 hit) | Mutations (1968) | SNPs (0) | Cancer (0) | Tumour Site (0) | Samples (0) | Pubmed (366) | Studies (0)

Show 10 entries

Gene	Alternate IDs	Tested samples	Simple Mutations	Fusions	Coding Mutations
PTEN	PTEN,ENST00000371953,PTEN.html...	69987	3975	0	3975

Showing 1 to 1 of 1 entries

First Previous 1 Next Last

Login

Genes (1 hit) | Mutations (1968) | SNPs (0) | Cancer (0) | Tumour Site (0) | Samples (0) | Pubmed (366) | Studies (0)

Show 10 entries

Gene	Syntax	Alternate IDs	Recurrence
PTEN	c.388C>G	PTEN c.388C>G...	162
PTEN	c.389G>A	PTEN c.389G>A...	134
PTEN	c.697C>T	PTEN c.697C>T...	124
PTEN	c.388C>T	PTEN c.388C>T...	94
PTEN	c.?_?del?	PTEN c.?_?del?...	91
PTEN	c.?	PTEN c.?...	88
PTEN	c.1_1212del1212	PTEN c.1_1212del1212...	88
PTEN	c.800delA	PTEN c.800delA...	56
PTEN	c.517C>T	PTEN c.517C>T...	47
PTEN	c.950_953delTACT	PTEN c.950_953delTACT...	44

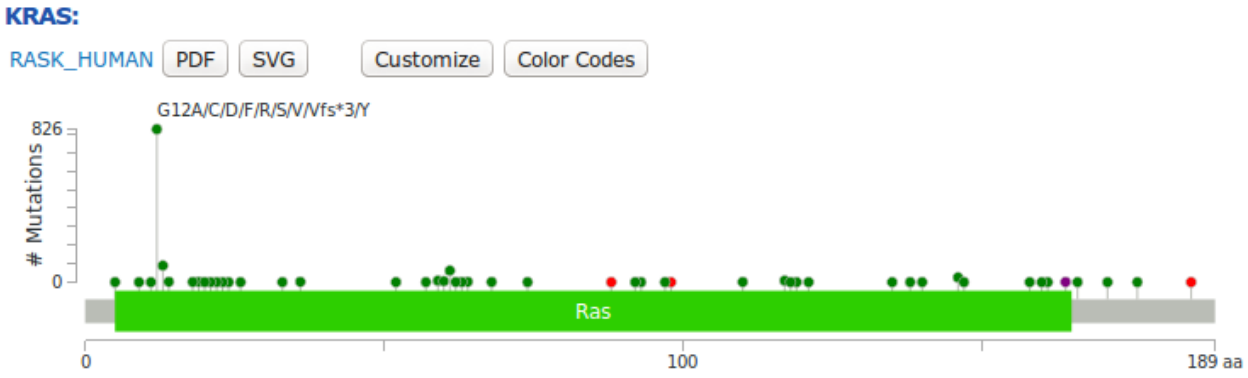
Showing 1 to 10 of 1,968 entries

First Previous 1 2 3 4 5 ... 197 Next Last

Role of the gene: oncogenic role by activation (Oncogene) or inactivation (Tumor Suppressor Gene)

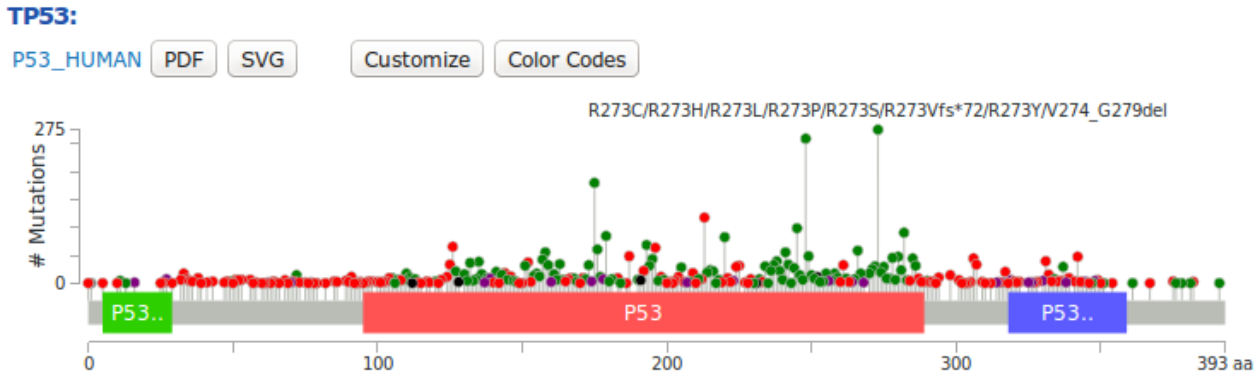
Interpretation of variant frequencies
Therapeutical action

Role of the gene as ONC or TSG



ONCOGENE

Mutated form of a gene involved in normal cell growth
Pro-tumoral activity by **activation**



TUMOR SUPPRESSOR GENE

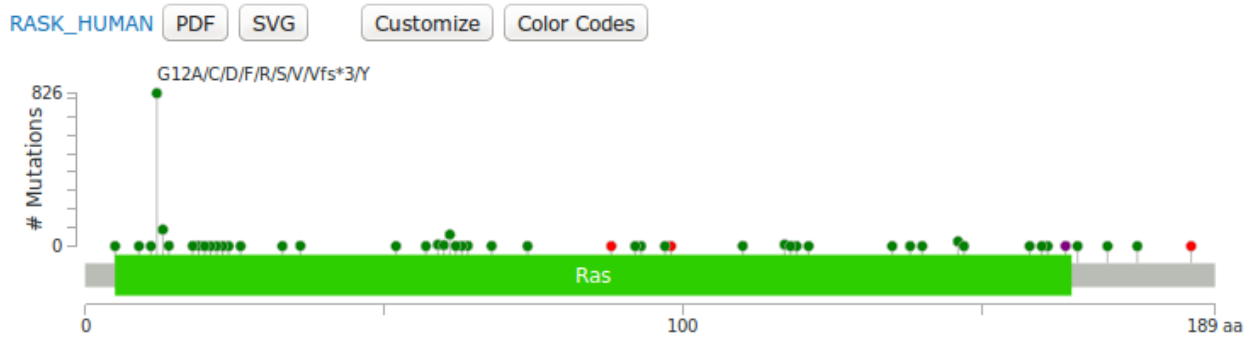
Gene that is involved in the control of cell growth
Pro-tumoral activity by **inactivation**

How variant frequencies in a tumor cohort can be interpreted?

How genes can be actioned for therapy?

Role of the gene as ONC or TSG

KRAS:

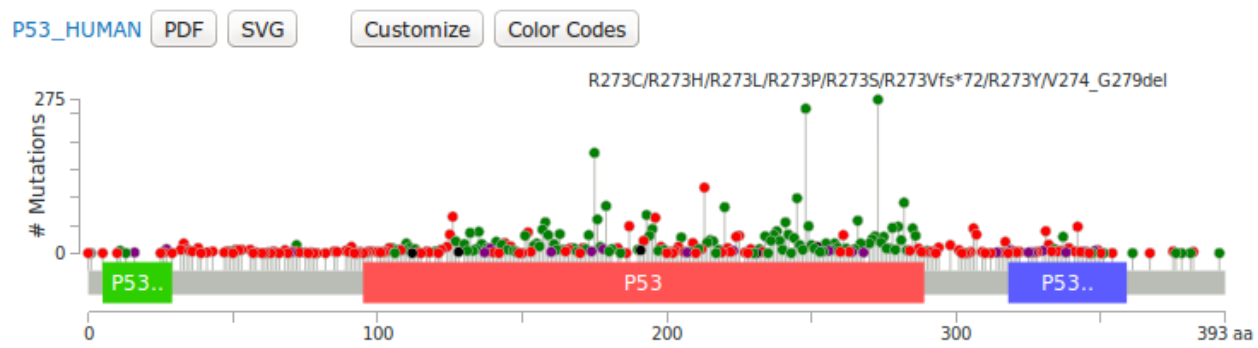


ONCOGENE

High frequency of variant expected in a cohort

Direct inhibition

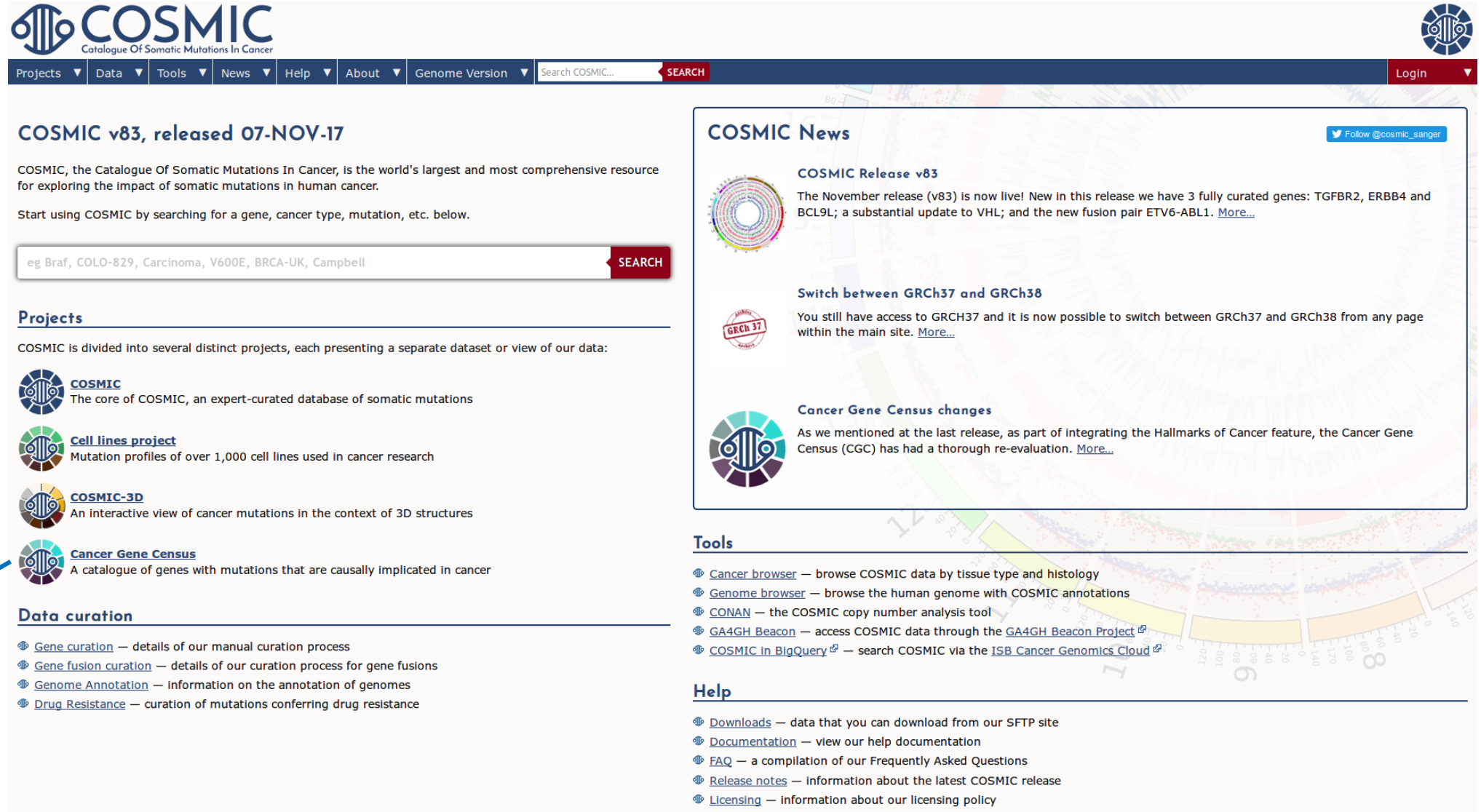
TP53:



TUMOR SUPPRESSOR GENE

Lower frequency of variant expected in a cohort

Downstream inhibition



cancer genes information (e.g. role of the gene)

COSMIC
Catalogue of somatic mutations in cancer

http://cancer.sanger.ac.uk/census/

Cancer Gene Census

Showing both tiers | Show tier 1 | Show tier 2

Show 25 entries Export: CSV TSV Search:

Gene Symbol	Name	Entrez GeneId	Genome Location	Tier	Hallmark	Chr Band	Somatic	Germline	Tumour Types(Somatic)	Tumour Types(Germline)	Cancer Syndrome	Tissue Type	Role in Cancer
ARHGAP5	Rho GTPase activating protein 5	394	14:32090670-32154948	2		14q12	yes		colon cancer; glioma			E; O	oncogene
ARHGEF12	RHO guanine nucleotide exchange factor (GEF) 12 (LARG)	23365	11:120337244-120485077	1		11q23.3	yes		AML			L	TSG; fusion
ARID1A	AT rich interactive domain 1A (SWI-like)	8289	1:26696404-26780756	1		1p35.3	yes		clear cell ovarian carcinoma; RCC; breast			E	TSG
ARID1B	AT rich interactive domain 1B	57492	6:156778104-157207891	1		6q25.1	yes		breast; hepatocellular carcinoma			E	TSG
ARID2	AT rich interactive domain 2	196528	12:45729837-45905078	1		12q12	yes		hepatocellular carcinoma			E	TSG; fusion

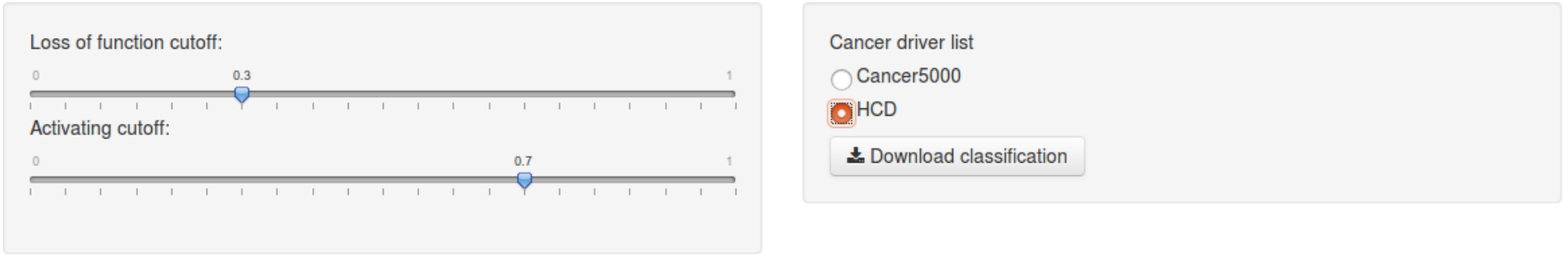
Role of the gene (ONC or TSG)

OncodriveROLE

http://bg.upf.edu/oncodrive-role/

Classifying cancer driver genes into Loss of Function and Activating roles.

We developed the machine-learning based approach OncodriveROLE to classify cancer driver genes into to Activating or Loss of Function roles for cancer gene development. Here you can download the code of the method, and browse the results of applying OncodriveROLE to two recently published list of driver genes (HCDs and Cancer5000) in the respective tabs Plots, Gene classification and performance. You may adjust the cut-offs with the sliders to the left, download the results according to the selected cut-offs or directly download the classifier to use with your own data. For further information please refer to the manuscript.



Download & Usage | Performance and plots | Gene classification | Validation

25 records per page — Search:

ENSG	SYM	oncodriveROLE	Value
ENSG00000000971	CFH	Activating	0.9180
ENSG00000009307	CSDE1	Loss of function	0.2515
ENSG0000			
ENSG0000			

Role of the gene (ONC or TSG)