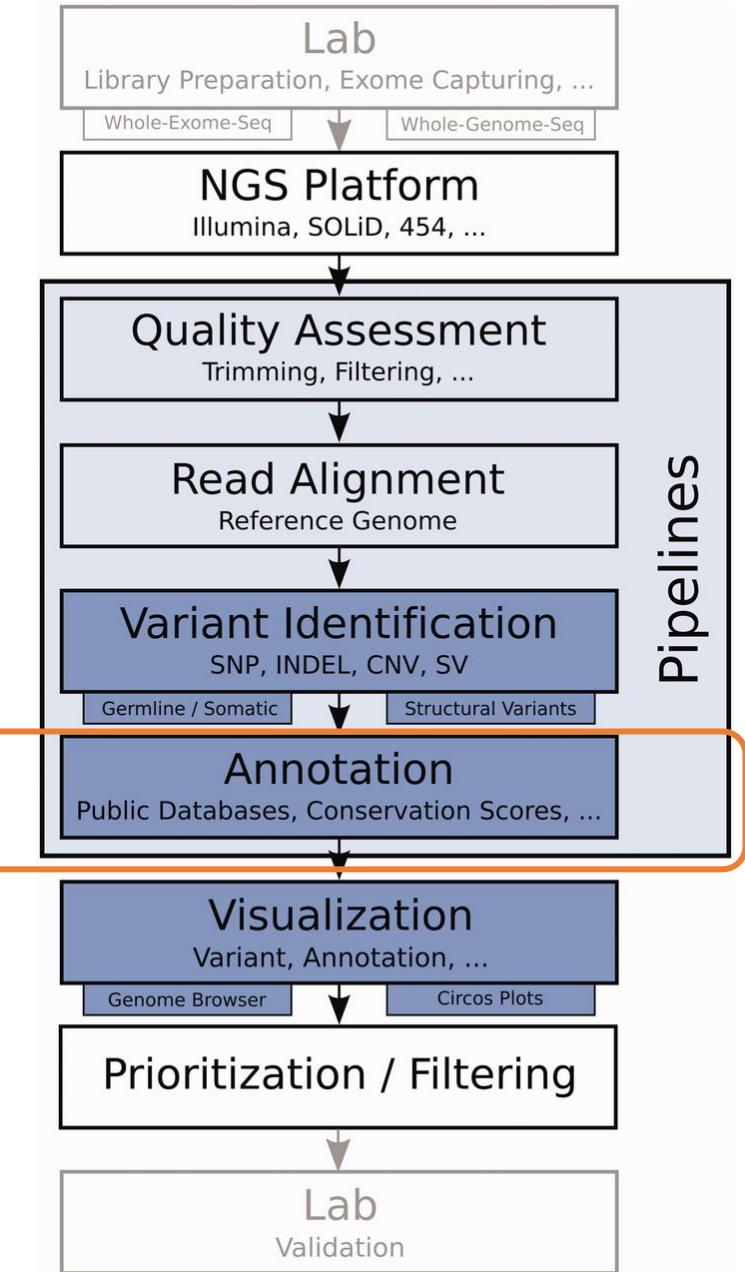


# PO21: Precision Oncology Course

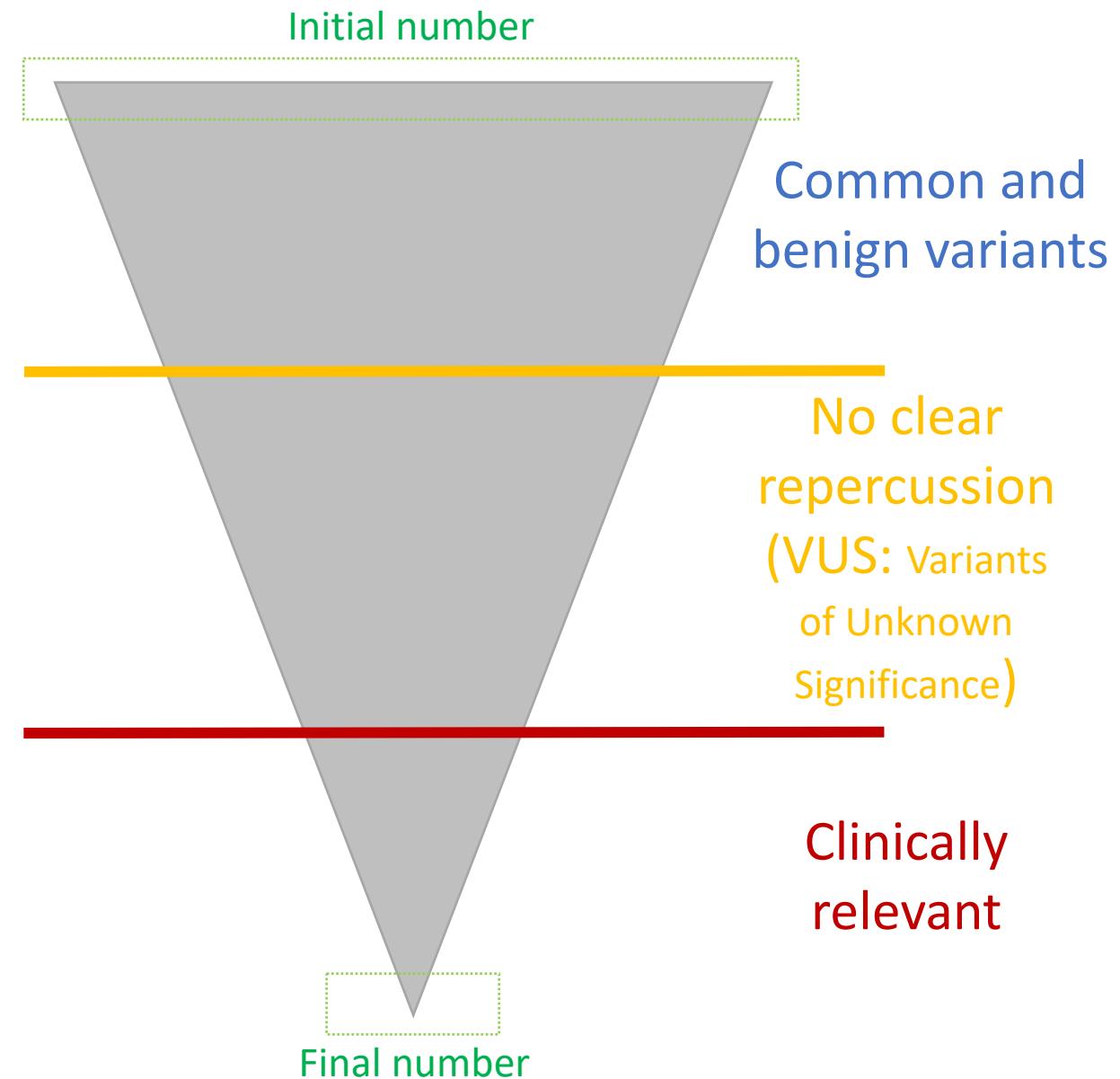
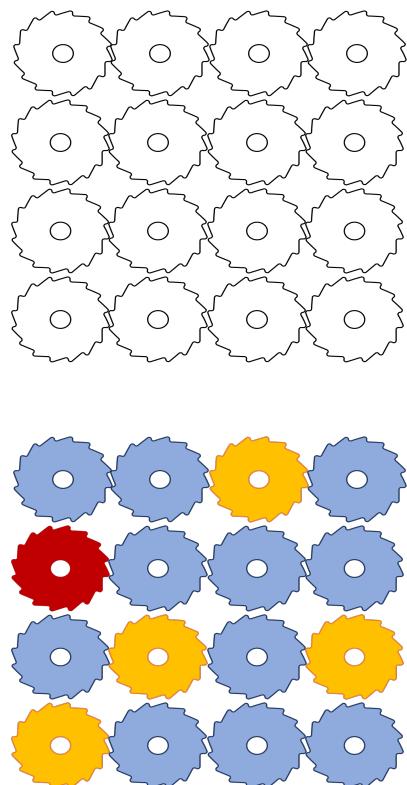
## Variant Annotation

# Annotation



# 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)**

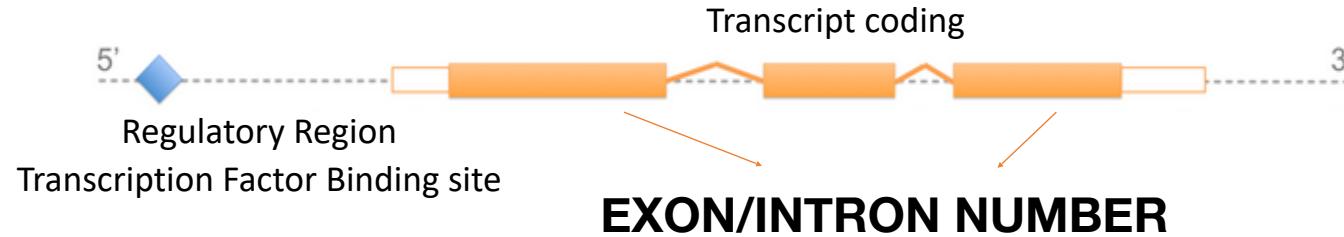
- Coding sequence e.g. c.638G>A
- Protein sequence e.g. p.Arg213His

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

“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

<http://varnomen.hgvs.org>

# Genomic location

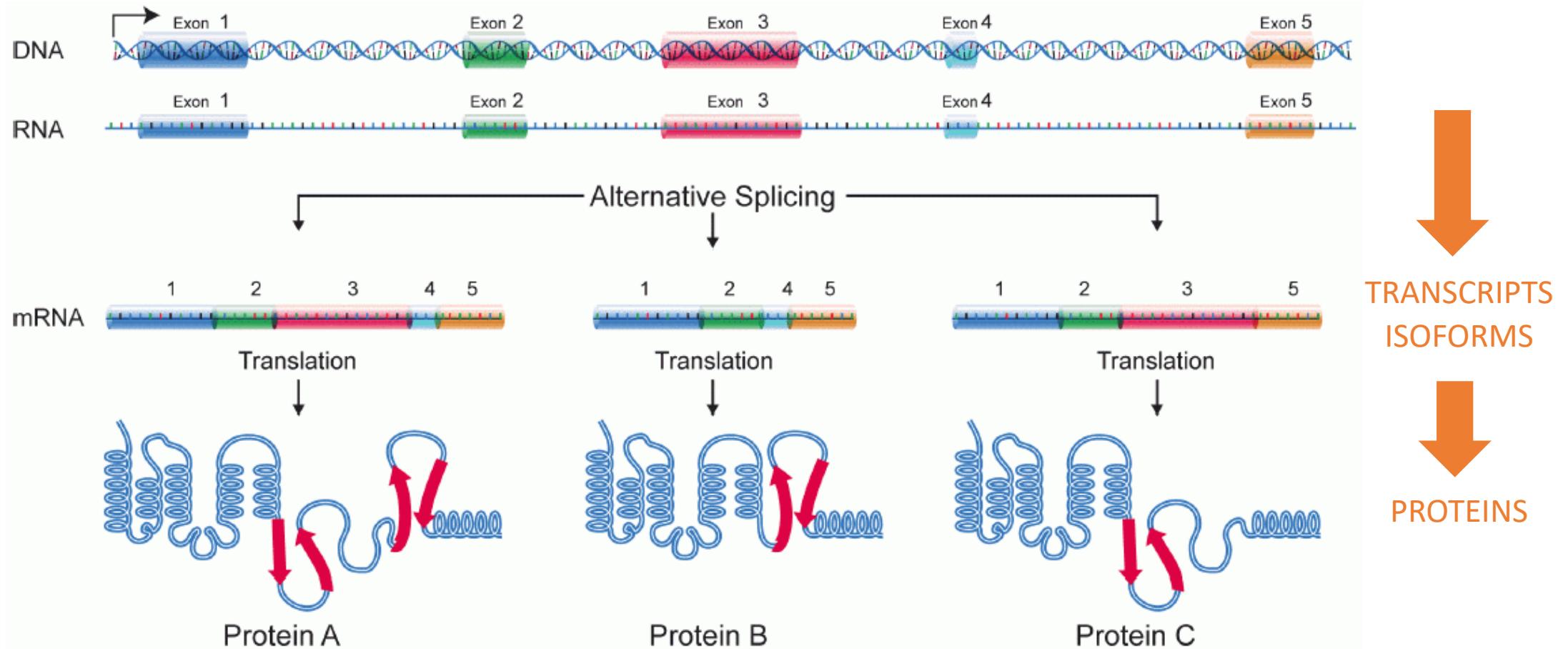


## BIOTYPE

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

# Affected elements



# 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 (Consensus Coding Sequence)** 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](#)



Human

**Assemblies:** GRCh38  
**Assemblies:** GRCh37



Mouse

**Assemblies:** GRCm38

### APPRIS WebServer

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

[Run the web server](#)



Zebrafish

**Assemblies:** GRCz10  
**Assemblies:** Zv9



Rat

**Assemblies:** Rnor\_6.0  
**Assemblies:** Rnor\_5.0



Pig

**Assemblies:** Scrofa10.2



Chimpanzee

**Assemblies:** CHIMP2.1.4

### APPRIS WebServices

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

[Go to the API interface](#)

APPRIS Database currently houses annotations for [vertebrate genomes](#) »

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 i

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 i

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 |
|-------------------------------------|-----------------|-------------|-------------------------|--------------------|---------------|--------------------|---------------------------|-----------------|----------------------|
| <input checked="" type="checkbox"/> | ENST00000369487 | 240         | 0                       | 1.066              | 0             | 1                  | 2                         | -               | 4                    |
| <input checked="" type="checkbox"/> | ENST00000369489 | 250         | 0                       | 1.066              | 0             | 1.5                | 3                         | -               | 4                    |
| <input checked="" type="checkbox"/> | ENST00000457047 | 248         | 0                       | 1.066              | 0             | 1.5                | 3                         | -               | 4                    |
| <input checked="" type="checkbox"/> | ENST00000464088 | 237         | 0                       | 1.066              | 0             | 1                  | 2                         | -               | 4                    |
| <input checked="" type="checkbox"/> | ENST00000526981 | 131         | 0                       | 1                  | 0             | 0                  | 3                         | -               | 2                    |

## ISOFORM RANGE

PRINCIPAL:1

PRINCIPAL:2

PRINCIPAL:3

PRINCIPAL:4

PRINCIPAL:5

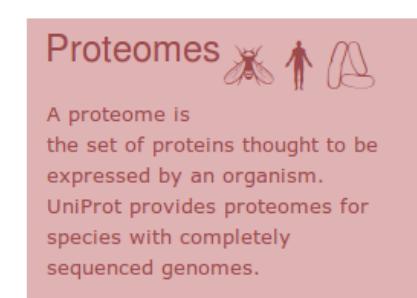
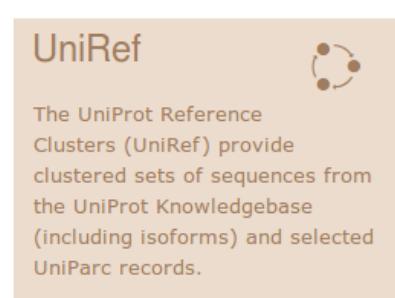
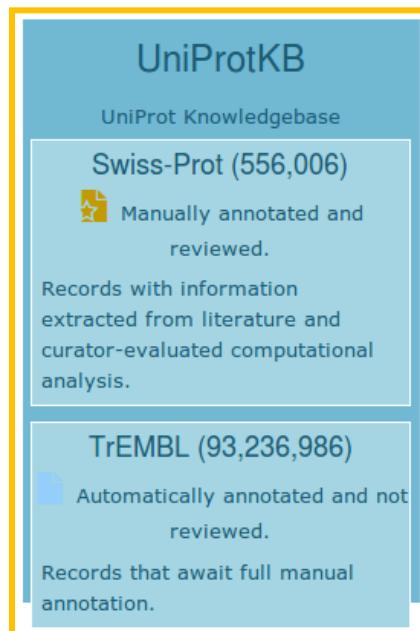
ALTERNATIVE:1

ALTERNATIVE:2

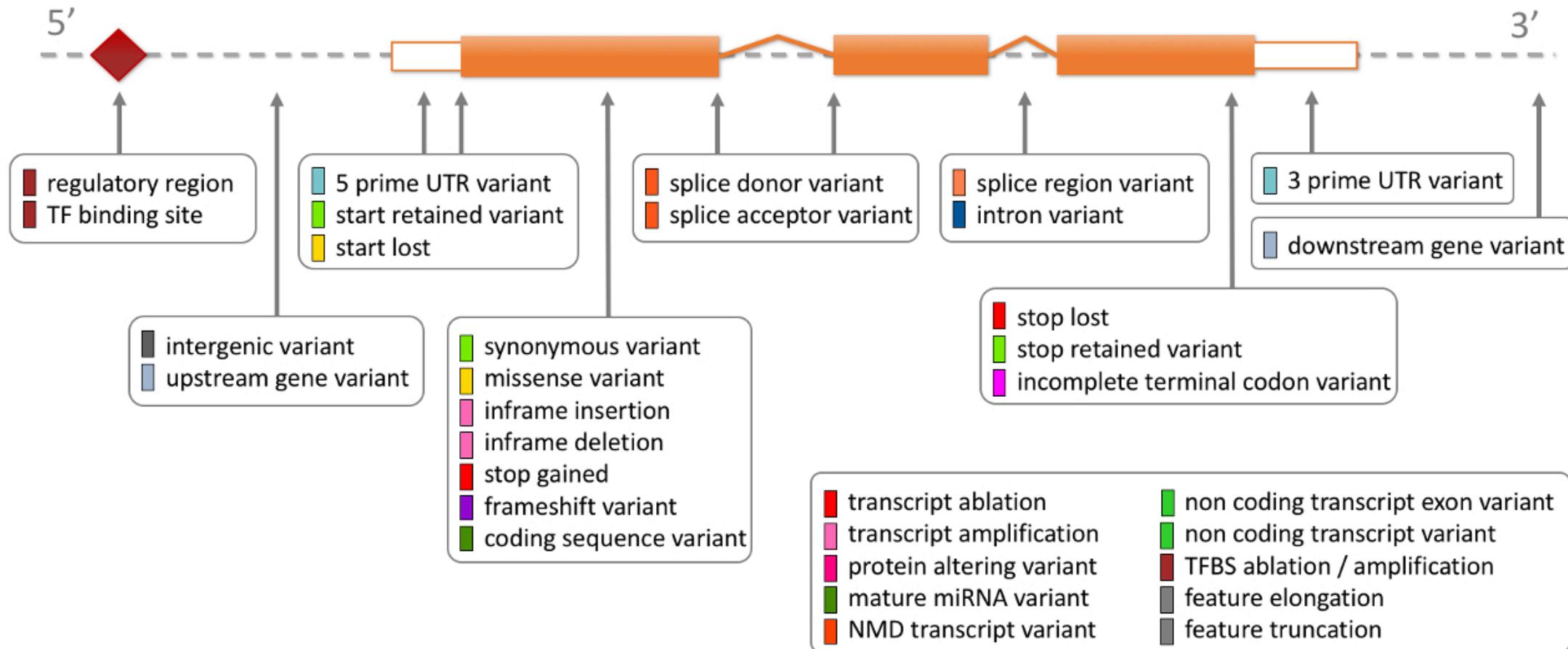
# Affected elements: proteins

## PROTEIN IDENTIFIERS

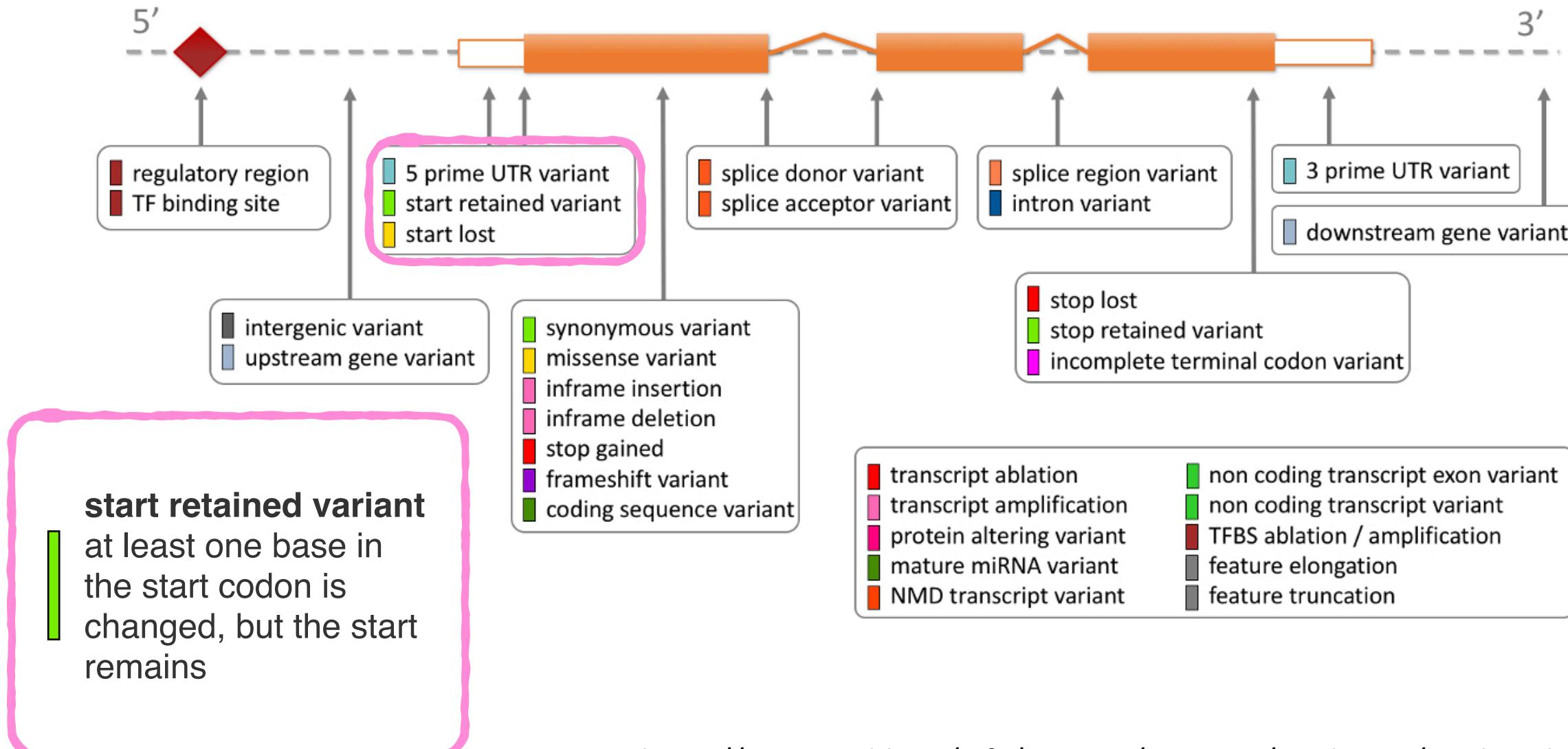
- Ensembl identifier e.g. ENSP00000409316
- RefSeq identifier e.g. NP\_001005221
- Uniprot identifiers (SWISSPROT, TrEMBL y UniParc)  
e.g. Q6IEY1 (SWISSPROT), A0A126GV92 (TREMBL), UPI0000041D3C (UniParc)



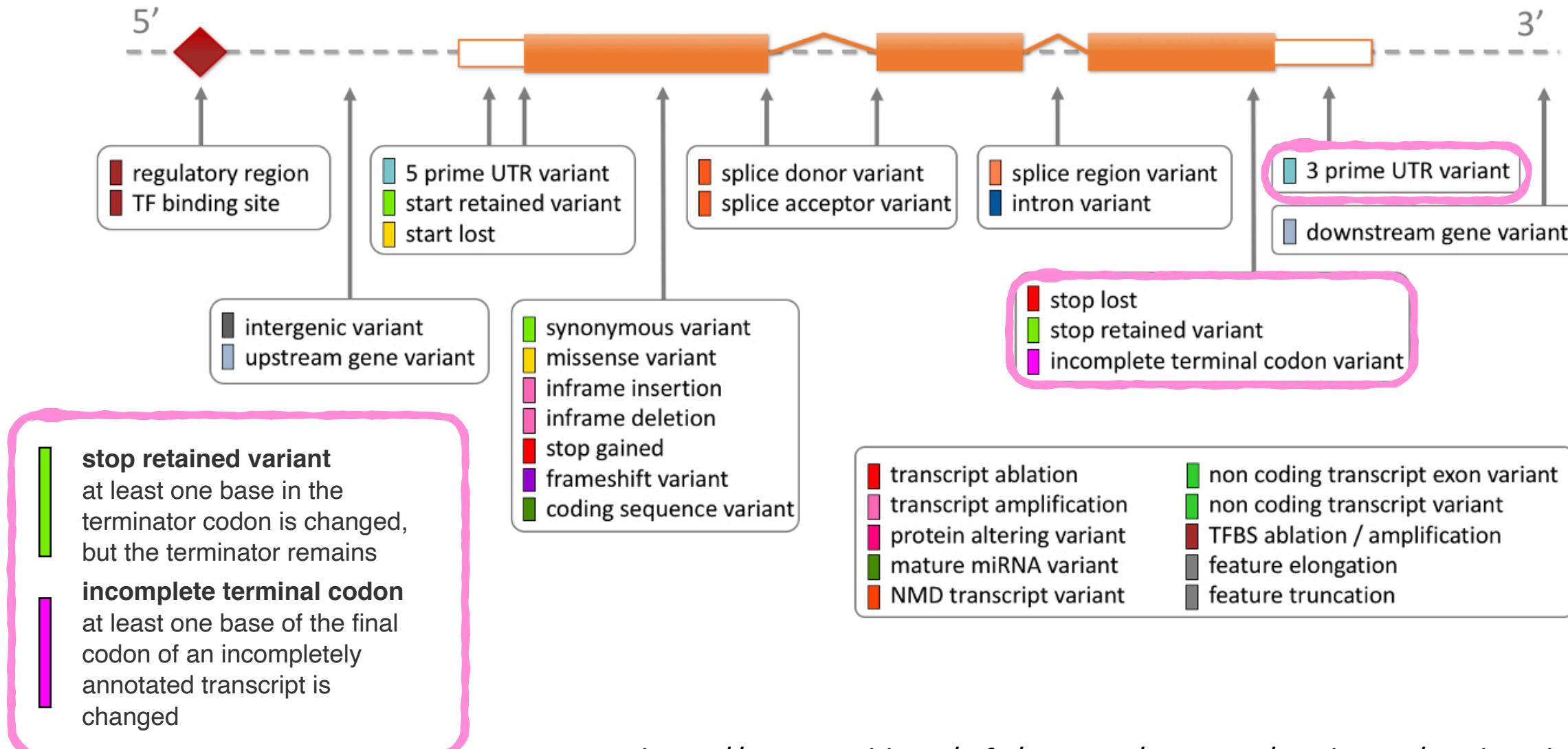
# Sequence consequences



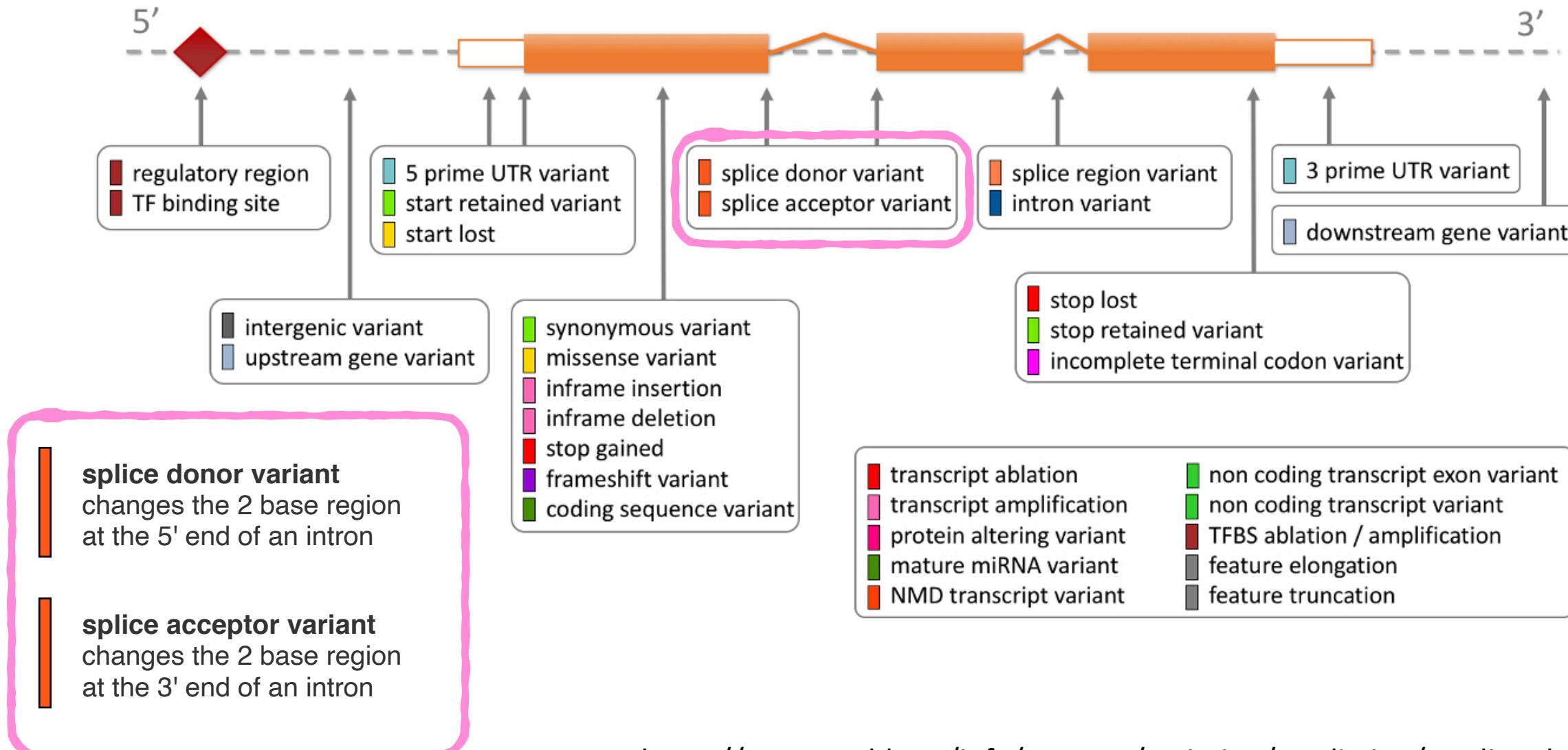
# Sequence consequences



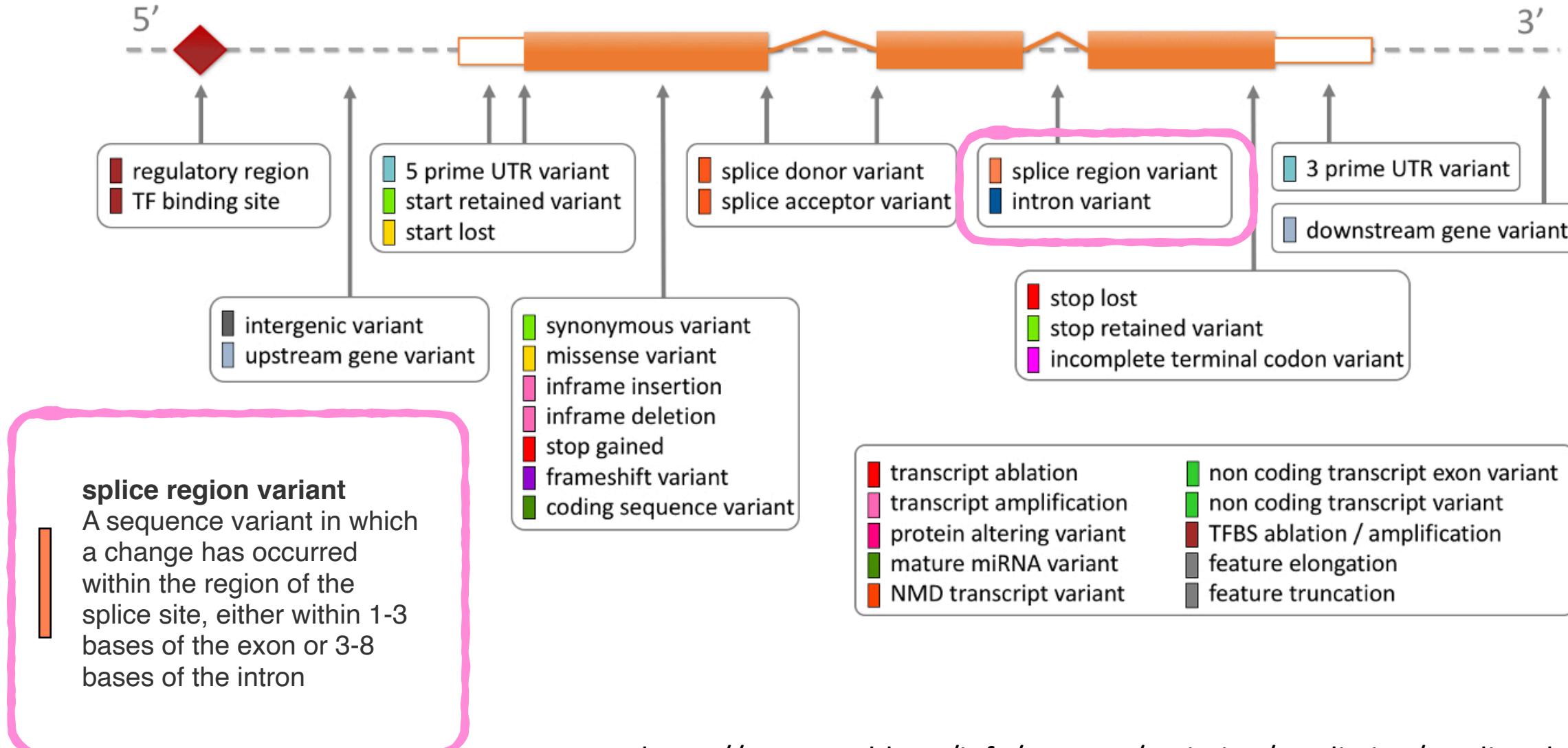
# Sequence consequences



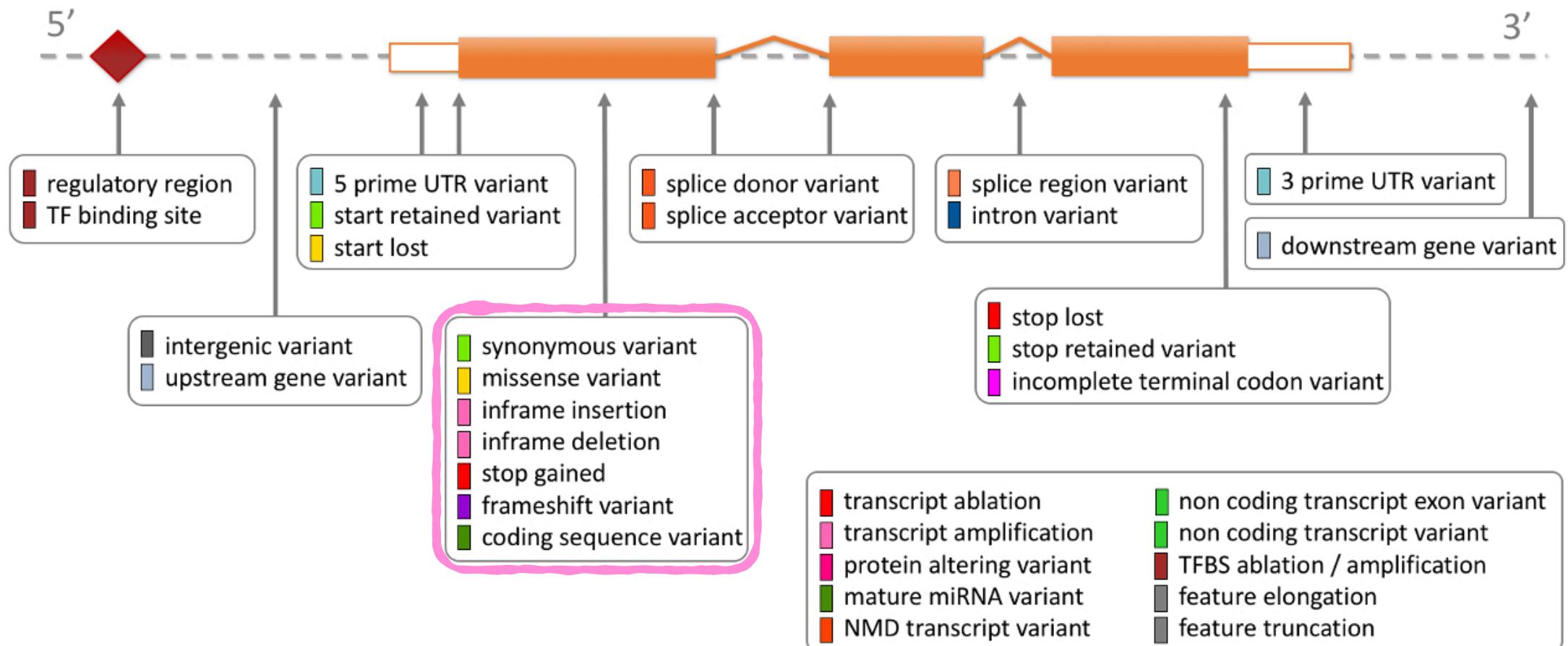
# Sequence consequences



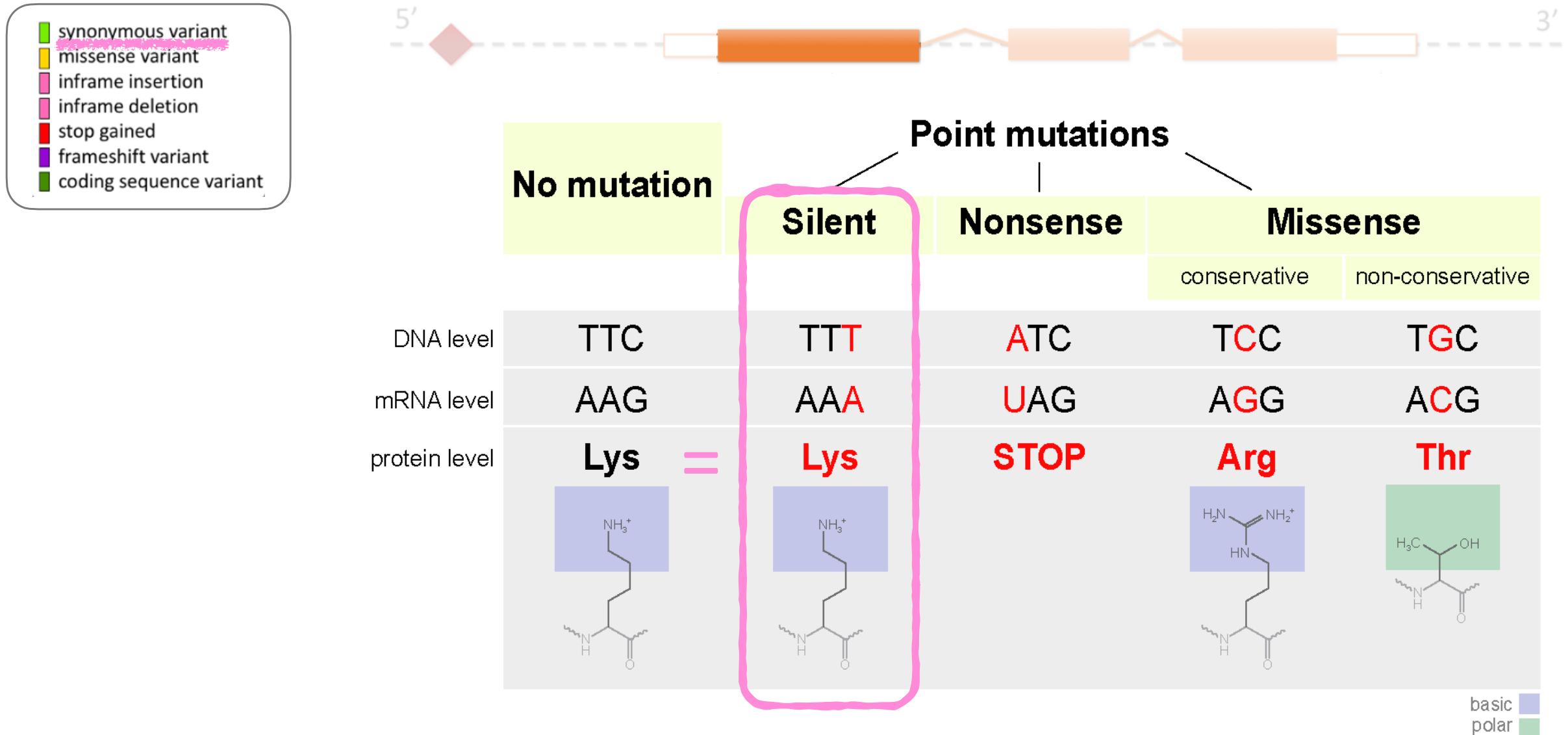
# Sequence consequences



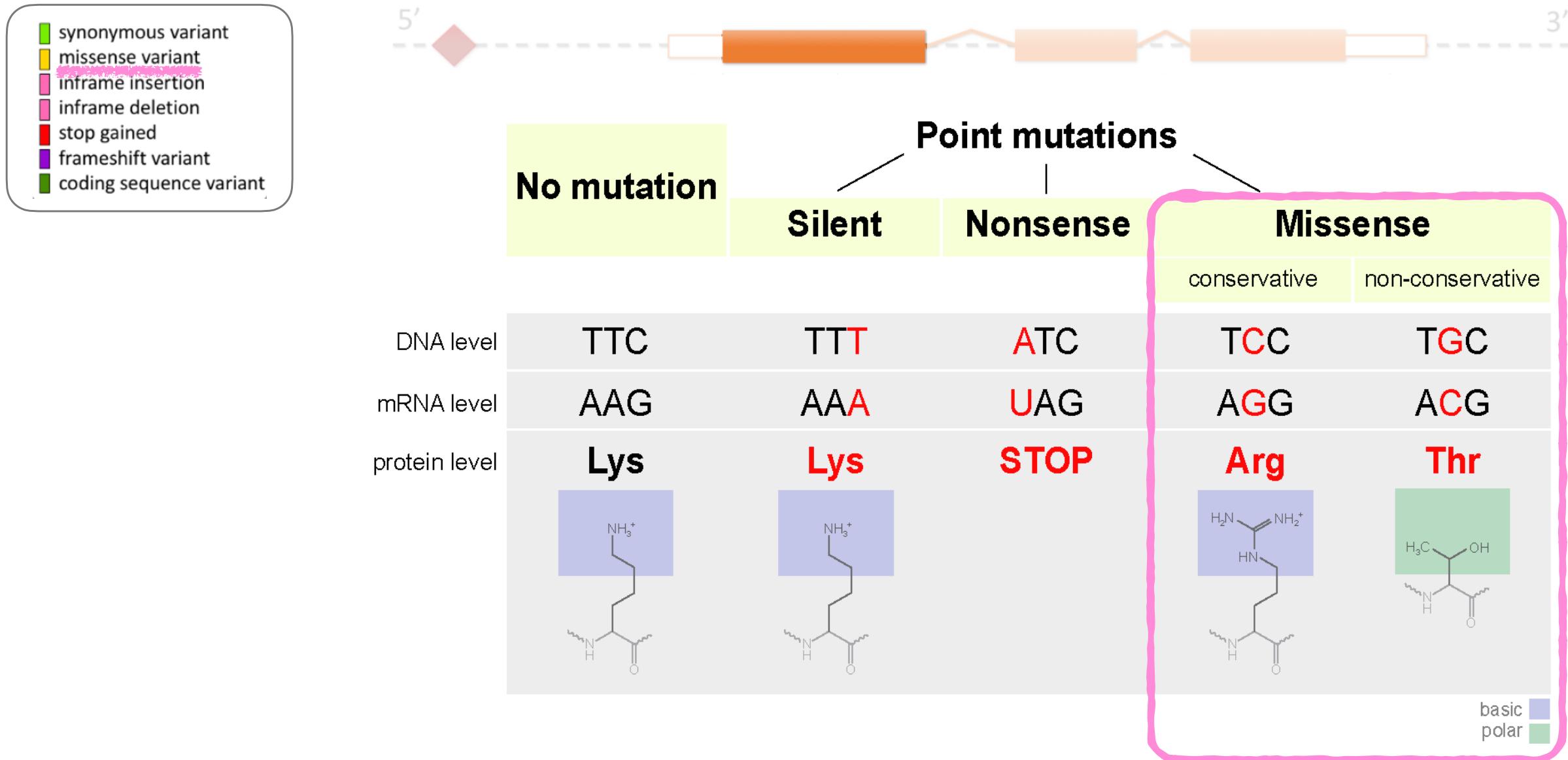
# Sequence consequences



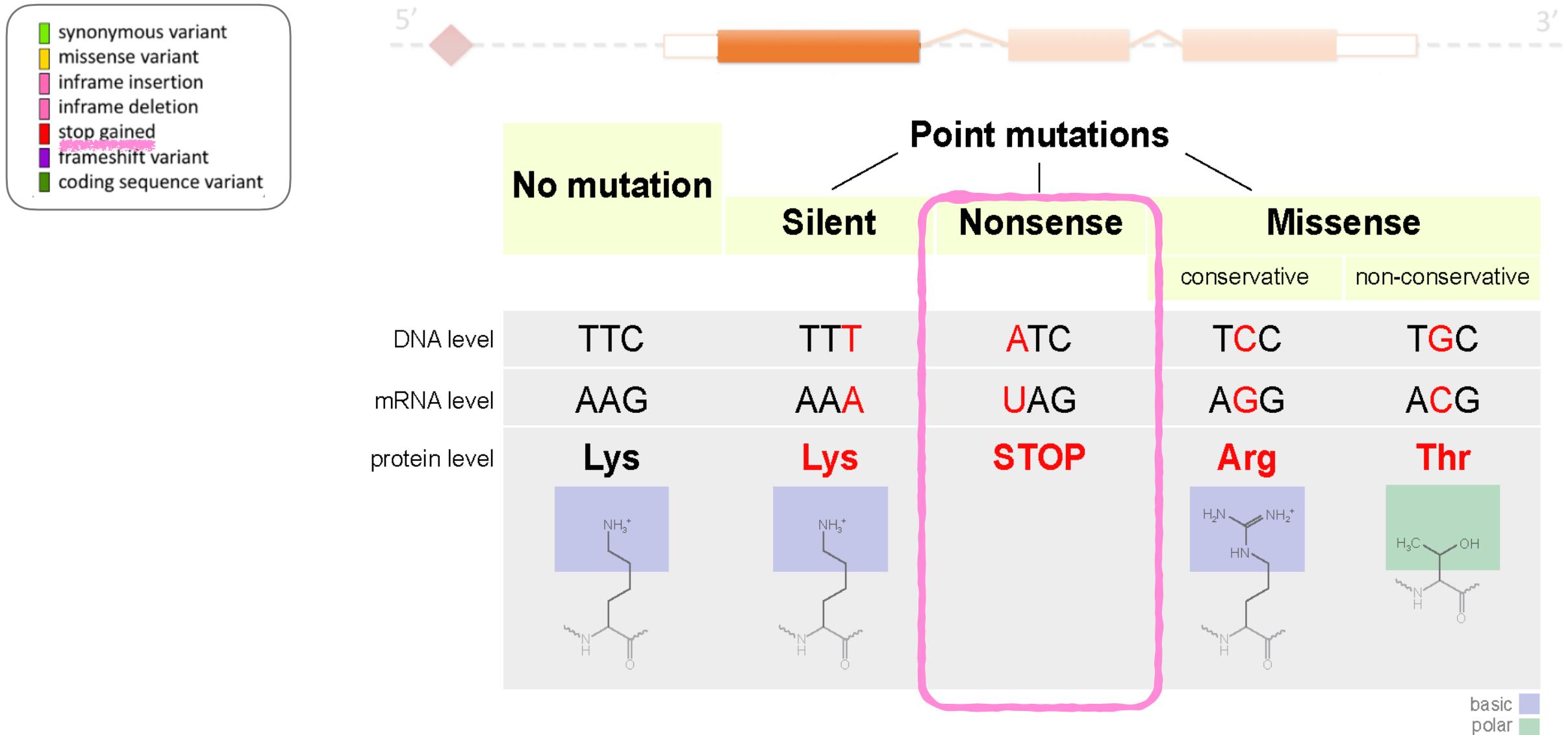
# Sequence consequences



# Sequence consequences

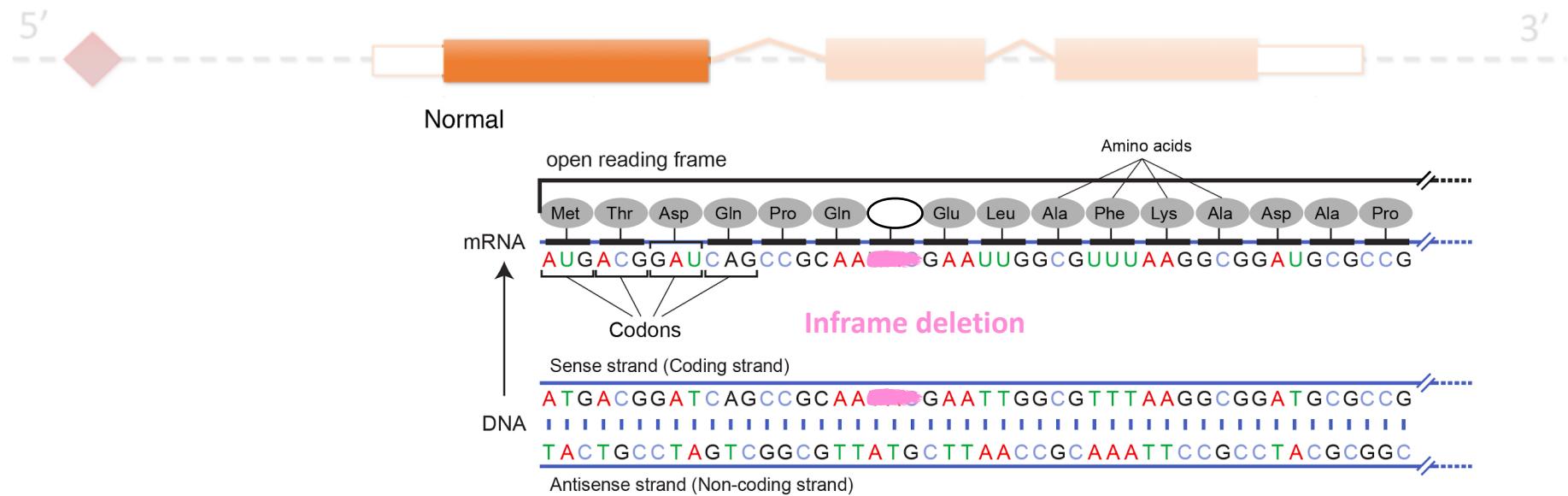


# Sequence consequences

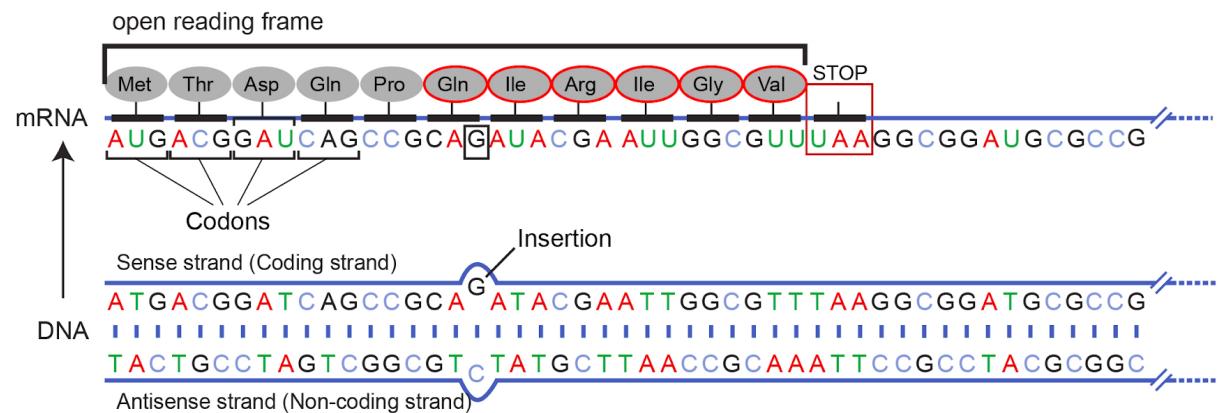


# Sequence consequences

- [green] synonymous variant
- [yellow] missense variant
- [pink] inframe insertion
- [pink] inframe deletion
- [red] stop gained
- [purple] frameshift variant
- [dark green] coding sequence variant



Frameshift mutation - single nucleotide insertion



# 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   | <a href="#">SO:0001893</a> | Transcript ablation      | HIGH     |
| splice_acceptor_variant  | A splice variant that changes the 2 base region at the 3' end of an intron  | <a href="#">SO:0001574</a> | Splice acceptor variant  | HIGH     |
| splice_donor_variant     | A splice variant that changes the 2 base region at the 5' end of an intron  | <a href="#">SO:0001575</a> | 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                        | <a href="#">SO:0001587</a> | 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 | <a href="#">SO:0001589</a> | 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  | <a href="#">SO:0001578</a> | Stop lost                | HIGH     |
| start_lost               | A codon variant that changes at least one base of the canonical start codon   | <a href="#">SO:0002012</a> | Start lost               | HIGH     |
| transcript_amplification | A feature amplification of a region containing a transcript   | <a href="#">SO:0001889</a> | Transcript amplification | HIGH     |
| inframe_insertion        | An inframe non synonymous variant that inserts bases into in the coding sequence  | <a href="#">SO:0001821</a> | Inframe insertion        | MODERATE |
| inframe_deletion         | An inframe non synonymous variant that deletes bases from the coding sequence   | <a href="#">SO:0001822</a> | 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                                | <a href="#">SO:0001583</a> | Missense variant         | MODERATE |
| protein_altering_variant | A sequence variant which is predicted to change the protein encoded in the coding sequence  | <a href="#">SO:0001818</a> | Protein altering variant | MODERATE |



[https://www.ensembl.org/info/genome/variation/prediction/predicted\\_data.html](https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html)

## **Functional impact prediction**

Through algorithms that predict the variant effect in  
the function and structure of the protein

# In silico predictors

They use **several evidences**:

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

They use **several algorithms**:

- Random Forest
- Neural Network

General

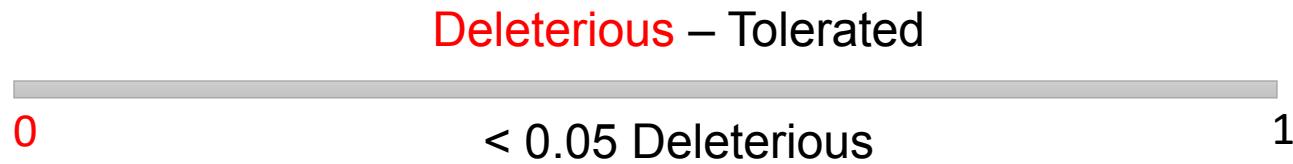
**TABLE 2** | Tools, software, and databases for functional prediction and annotation of variant impact.

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

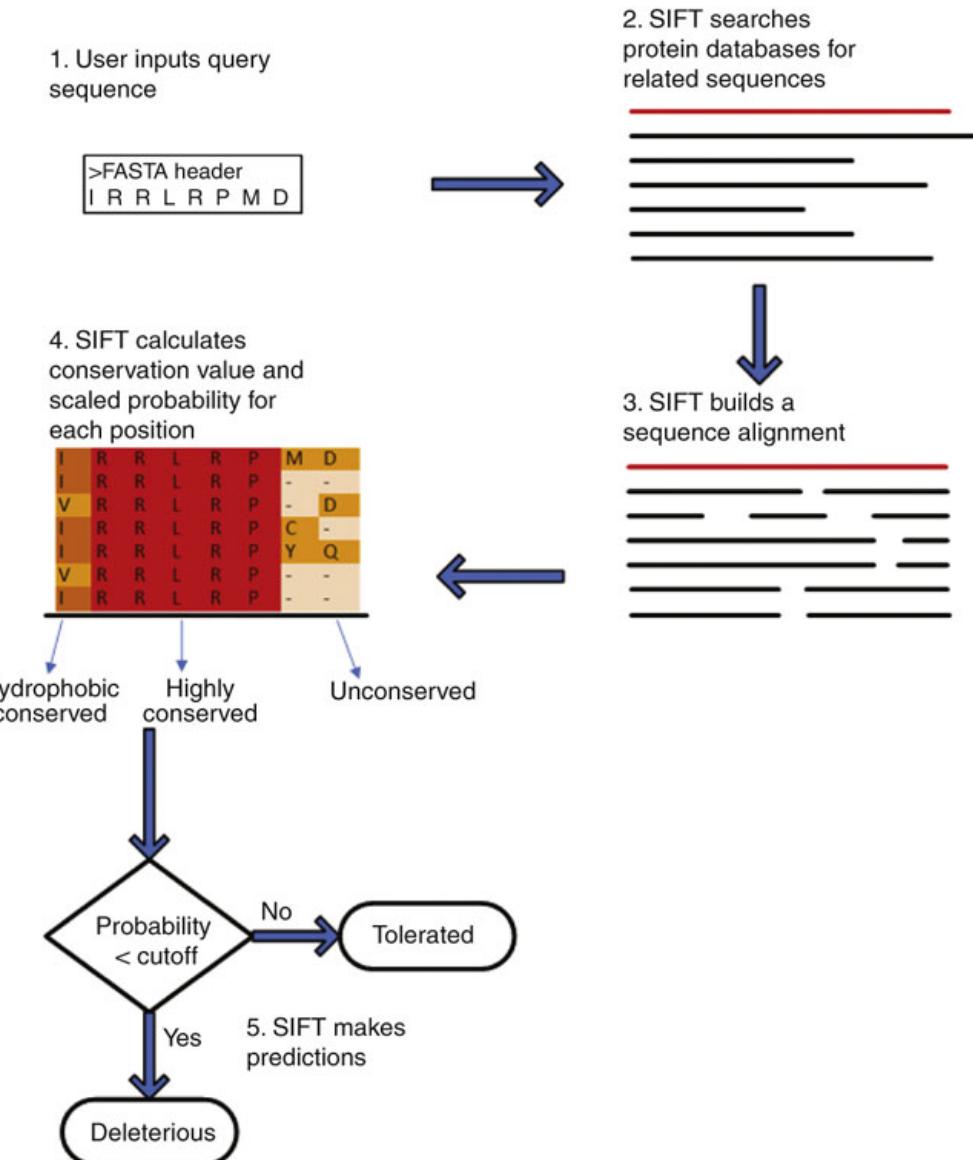
<sup>a</sup>Aggregated databases combine outputs of other databases and algorithms are, therefore, efficient resources to use in annotation pipelines. Adding these resources to observed variants is supported software in **Table 4** including Ensembl VEP software (noted<sup>b</sup> in this table), Annovar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

# Sift and PolyPhen

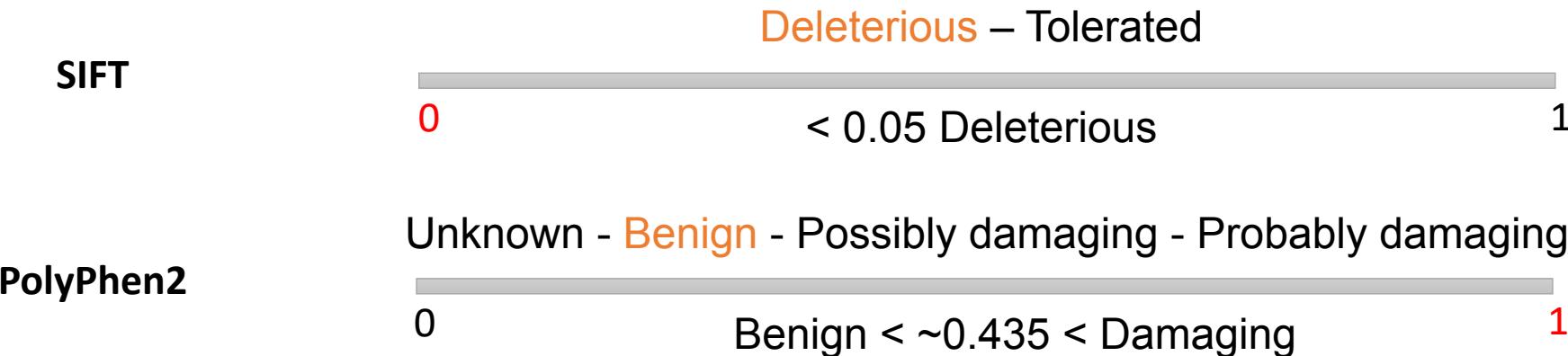
**SIFT PREDICTION** predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids.



**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.



# 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*

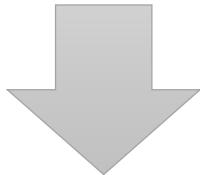
**Table 2** In silico predictive algorithms

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



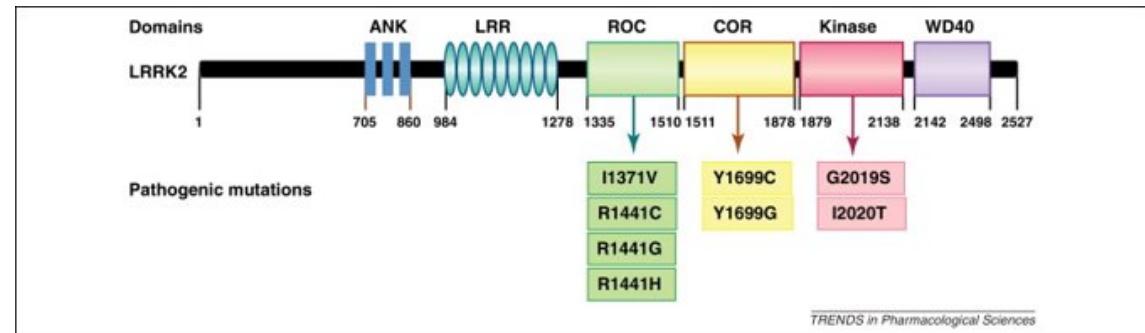
**Predictions should be taken cautiously and not been used as the only evidence for the relevance of the variant.**

# Impact through domain alteration

**DOMAINS** Protein overlapping domains

Pfam, Prosite, InterPro

e.g. Pfam\_domain: PF00071



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

Epidermal growth factor receptor (P00533)

Export FASTA

Accession P00533 (EGFR\_HUMAN)

Species Homo sapiens (Human)

Length 1,210 amino acids (complete)

Source: UniProtKB

Protein family membership

Tyrosine protein kinase, EGF/ERB/XmrK receptor (IPR016245)

Homologous superfamilies



Domains and repeats



<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

$\geq 1\%$  :  
**polymorphism**  
(common variant)

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

## Data Bases

**Table 1** Population, disease-specific, and sequence databases

| Population databases  |  |
|---|--|
| Exome Aggregation Consortium<br><a href="http://exac.broadinstitute.org/">http://exac.broadinstitute.org/</a>                 | 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<br><a href="http://evs.gs.washington.edu/EVS">http://evs.gs.washington.edu/EVS</a>                       | 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<br><a href="http://browser.1000genomes.org">http://browser.1000genomes.org</a>                           | 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<br><a href="http://www.ncbi.nlm.nih.gov/snp">http://www.ncbi.nlm.nih.gov/snp</a>  | Database of short genetic variations (typically $\leq 50$ bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.  |
| dbVar<br><a href="http://www.ncbi.nlm.nih.gov/dbvar">http://www.ncbi.nlm.nih.gov/dbvar</a>                                    | Database of structural variation (typically $>50$ bp) submitted from many sources.   |
| Disease databases   |  |
| ClinVar<br><a href="http://www.ncbi.nlm.nih.gov/clinvar">http://www.ncbi.nlm.nih.gov/clinvar</a>                              | Database of assertions about the clinical significance and phenotype relationship of human variations.   |
| OMIM<br><a href="http://www.omim.org">http://www.omim.org</a>   | Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.  |
| Human Gene Mutation Database<br><a href="http://www.hgmd.org">http://www.hgmd.org</a>   | 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<br><a href="http://www.hgvs.org/dblist/dblist.html">http://www.hgvs.org/dblist/dblist.html</a> | 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<br><a href="http://www.lovd.nl">http://www.lovd.nl</a>   | A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.   |
| DECIPHER<br><a href="http://decipher.sanger.ac.uk">http://decipher.sanger.ac.uk</a>   |  |
| Sequence databases  |  |
| NCBI Genome<br><a href="http://www.ncbi.nlm.nih.gov/genome">http://www.ncbi.nlm.nih.gov/genome</a>                            | Source of full human genome reference sequences.   |
| RefSeqGene<br><a href="http://www.ncbi.nlm.nih.gov/refseq/rsg">http://www.ncbi.nlm.nih.gov/refseq/rsg</a>                     | Medically relevant gene reference sequence resource.   |
| Locus Reference Genomic (LRG)<br><a href="http://www.lrg-sequence.org">http://www.lrg-sequence.org</a>                        |  |
| MitoMap<br><a href="http://www.mitomap.org/MITOMAP/HumanMitoSeq">http://www.mitomap.org/MITOMAP/HumanMitoSeq</a>              | Revised Cambridge reference sequence for human mitochondrial DNA.  |

# 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 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.



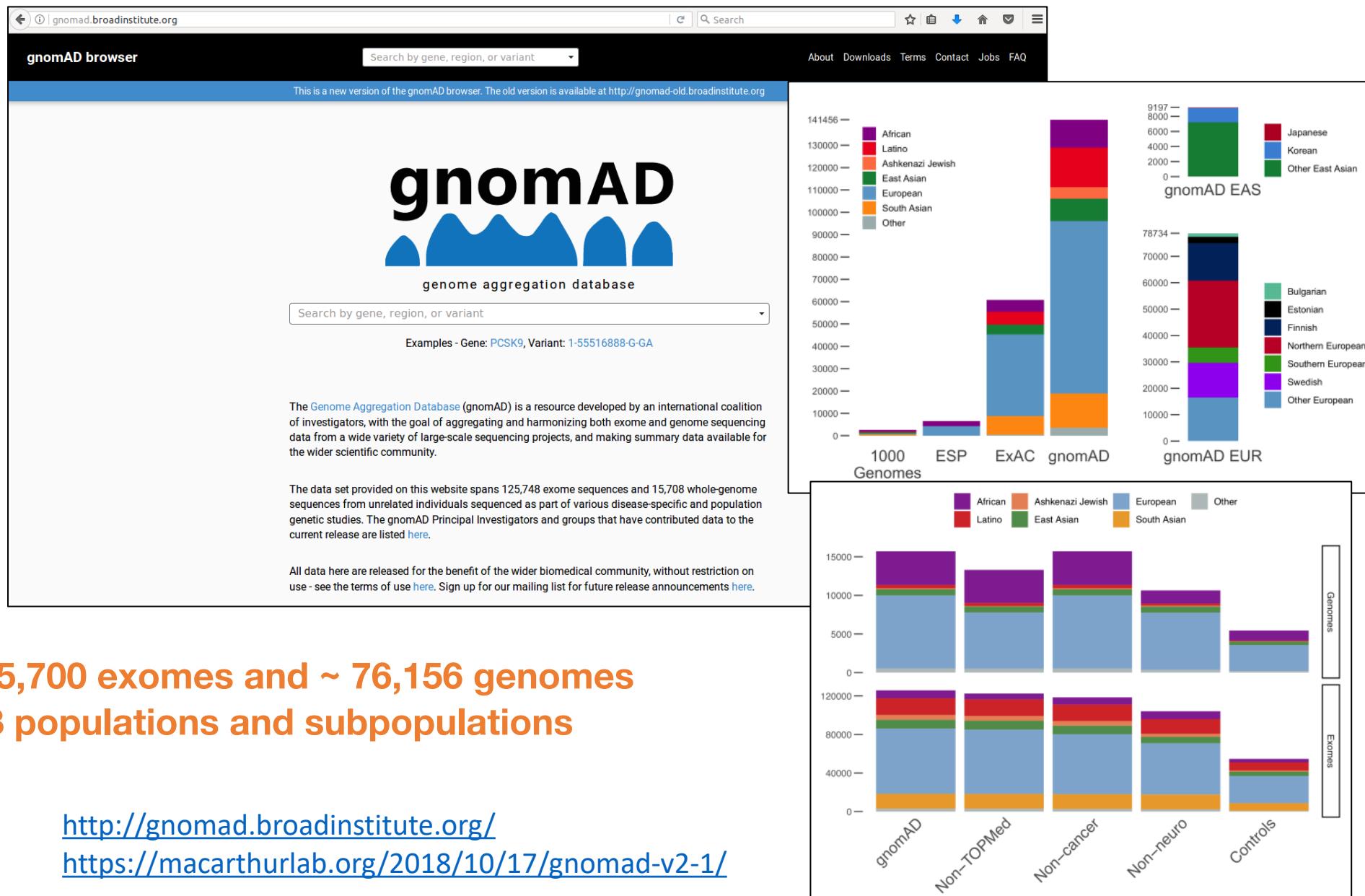
Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;



| 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)



# Association with pathologies

## Data Bases

**Table 1** Population, disease-specific, and sequence databases

| Population databases  |  |
|---|--|
| Exome Aggregation Consortium<br><a href="http://exac.broadinstitute.org/">http://exac.broadinstitute.org/</a>                 | 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<br><a href="http://evs.gs.washington.edu/EVS">http://evs.gs.washington.edu/EVS</a>                       | 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<br><a href="http://browser.1000genomes.org">http://browser.1000genomes.org</a>                           | 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<br><a href="http://www.ncbi.nlm.nih.gov/snp">http://www.ncbi.nlm.nih.gov/snp</a>  | Database of short genetic variations (typically $\leq 50$ bp) submitted from many sources. May lack details of the originating study and may contain pathogenic variants.  |
| dbVar<br><a href="http://www.ncbi.nlm.nih.gov/dbvar">http://www.ncbi.nlm.nih.gov/dbvar</a>                                    | Database of structural variation (typically $> 50$ bp) submitted from many sources.  |
| Disease databases   |  |
| ClinVar<br><a href="http://www.ncbi.nlm.nih.gov/clinvar">http://www.ncbi.nlm.nih.gov/clinvar</a>                              | Database of assertions about the clinical significance and phenotype relationship of human variations.   |
| OMIM<br><a href="http://www.omim.org">http://www.omim.org</a>   | Database of human genes and genetic conditions that also contains a representative sampling of disease-associated genetic variants.  |
| Human Gene Mutation Database<br><a href="http://www.hgmd.org">http://www.hgmd.org</a>   | 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<br><a href="http://www.hgvs.org/dblist/dblist.html">http://www.hgvs.org/dblist/dblist.html</a> | 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<br><a href="http://www.lovd.nl">http://www.lovd.nl</a>   | A molecular cytogenetic database for clinicians and researchers linking genomic microarray data with phenotype using the Ensembl genome browser.   |
| DECIPHER<br><a href="http://decipher.sanger.ac.uk">http://decipher.sanger.ac.uk</a>   |  |
| Sequence databases  |  |
| NCBI Genome<br><a href="http://www.ncbi.nlm.nih.gov/genome">http://www.ncbi.nlm.nih.gov/genome</a>                            | Source of full human genome reference sequences.   |
| RefSeqGene<br><a href="http://www.ncbi.nlm.nih.gov/refseq/rsg">http://www.ncbi.nlm.nih.gov/refseq/rsg</a>                     | Medically relevant gene reference sequence resource.   |
| Locus Reference Genomic (LRG)<br><a href="http://www.lrg-sequence.org">http://www.lrg-sequence.org</a>                        |  |
| MitoMap<br><a href="http://www.mitomap.org/MITOMAP/HumanMitoSeq">http://www.mitomap.org/MITOMAP/HumanMitoSeq</a>              | Revised Cambridge reference sequence for human mitochondrial DNA.  |

# OMIM

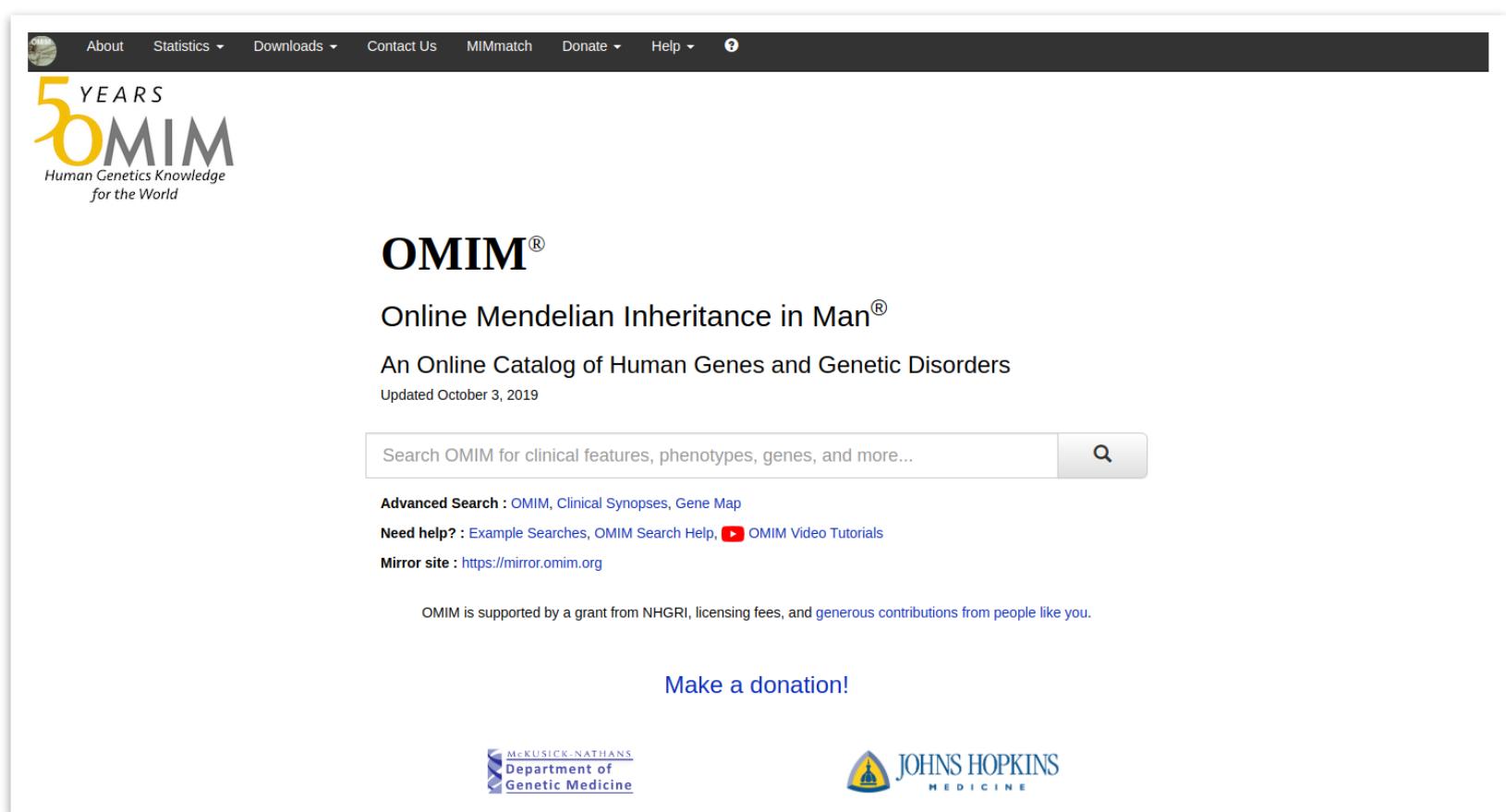
~ 15.000 human genes



Mendelian disorders

Susceptibility to cancer and other complex diseases

<https://omim.org/>



50 YEARS  
**OMIM**  
Human Genetics Knowledge  
for the World

**OMIM®**  
Online Mendelian Inheritance in Man®  
An Online Catalog of Human Genes and Genetic Disorders  
Updated October 3, 2019

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

Advanced Search : OMIM, Clinical Synopses, Gene Map  
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!](#)

 McKUSICK-NATHANS  
Department of  
Genetic Medicine

 JOHNS HOPKINS  
MEDICINE

# LOVD

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.

<http://www.lovd.nl>

The screenshot shows the LOVD v.3.0 homepage with a blue header containing the LOVD logo, version information, and navigation links for Home, News, FAQ, Documentation, Download, Contact, Developers, and social media links. Below the header is a search bar and a link to the public list of LOVD installations. The main content area is titled "List of public LOVD installations" and includes a note about checking the list of LSDBs. It features a search bar for gene symbols and a link to download the full list of genes. A table lists 1,026,284,303 variants across 68 installations, with columns for URL, LOVD version, number of genes, and total variants. The table includes entries for BIPMed SNP Array (HG38, HG19), BIPMed WES (HG19), and Whole genome datasets.

| In total: 1,026,284,303 variants (6,624,805 unique) in 1,274,946 individuals in 68 LOVD installations. |  |                |                    |
|--|--|----------------|--------------------|
| <a href="http://bipmed.iqm.unicamp.br/snparray_296/">http://bipmed.iqm.unicamp.br/snparray_296/</a>    | 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  |                    |
| <a href="http://bipmed.iqm.unicamp.br/snparray_hg19/">http://bipmed.iqm.unicamp.br/snparray_hg19/</a>  | LOVD 3.0-21  | 15440 genes    | 267807776 variants |
| BIPMed SNP Array - HG19  | A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GNT,AAAC,AA... | 893557 unique  |                    |
| <a href="http://bipmed.iqm.unicamp.br/snparray/">http://bipmed.iqm.unicamp.br/snparray/</a>            | LOVD 3.0-20  | 17391 genes    | 222715116 variants |
| BIPMed SNP Array   | A1BG,AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT,A4GN... | 902273 unique  |                    |
| <a href="http://bipmed.iqm.unicamp.br/wes_hg19/">http://bipmed.iqm.unicamp.br/wes_hg19/</a>            | LOVD 3.0-21  | 18203 genes    | 197345407 variants |
| BIPMed WES - HG19  | A1BG,A1CF,A2M,A2M-AS1,A2ML1,A3GALT2,A4GALT,A4GNT,AAAS,AAC... | 824599 unique  |                    |
| <a href="http://bipmed.iqm.unicamp.br/">http://bipmed.iqm.unicamp.br/</a>                              | LOVD 3.0-20  | 20930 genes    | 66158522 variants  |
| BIPMed WES   | A1BG,A1BG-AS1,A1CF,A2M,A2M-AS1,A2ML1,A2MP1,A3GALT2,A4GALT... | 622610 unique  |                    |
| <a href="http://databases.lovd.nl/whole_genome/">http://databases.lovd.nl/whole_genome/</a>            | 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/>

BIPMed - Neurofibromatosis

LOVD 3.0-21

NF1,NF2

2 genes

66 variants

66 unique

# COSMIC

Expert Curation Data  
from peer reviewed  
publications

Genome-wide screen  
data from publications or  
consortiums

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

The screenshot shows the COSMIC website homepage. At the top, there's a navigation bar with links for Projects, Data, Tools, News, Help, About, Genome Version, a search bar, and a login link. Below the header, a banner announces 'COSMIC v83, released 07-NOV-17'. It highlights the resource as the world's largest and most comprehensive for somatic mutations in cancer. A search bar below the banner contains the placeholder 'eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell' with a 'SEARCH' button. The main content area is divided into sections: 'Projects' (listing COSMIC, Cell lines project, COSMIC-3D, and Cancer Gene Census), 'Data curation' (listing Gene curation, Gene fusion curation, Genome Annotation, and Drug Resistance), and 'Tools' (listing Cancer browser, Genome browser, CONAN, GA4GH Beacon, and COSMIC In BigQuery). On the right side, a 'COSMIC News' box features a circular logo and text about the November release (v83) which includes three new fully curated genes: TGFBR2, ERBB4 and BCL9L; a substantial update to VHL; and the new fusion pair ETV6-ABL1. It also mentions the switch between GRCh37 and GRCh38 and changes to the Cancer Gene Census.

# Catalog of somatic mutations in cancer

# Clinical Implications of variants: ClinVar

## Variation

|     | Variation<br>Location  | Gene(s)               |
|-----|--|-----------------------|
| 19. | <input type="checkbox"/> NM_001005862.2(ERBB2):c.1376C>T ( p.Pro459Leu)<br><i>GRCh37:</i> Chr17:37872145<br><i>GRCh38:</i> Chr17:39715892                                | <a href="#">ERBB2</a> |
| 20. | <input type="checkbox"/> NM_001005862.2(ERBB2):c.1703C>A ( p.Ala568Asp)<br><i>GRCh37:</i> Chr17:37873628<br><i>GRCh38:</i> Chr17:39717375                                | <a href="#">ERBB2</a> |
| 21. | <input type="checkbox"/> NM_001005862.2(ERBB2):c.1870A>G ( p.Ile624Val)<br><i>GRCh37:</i> Chr17:37879585<br><i>GRCh38:</i> Chr17:39723332                                | <a href="#">ERBB2</a> |
| 22. | <input type="checkbox"/> NM_001005862.2(ERBB2):c.1873A>G ( p.Ile625Val)<br><i>GRCh37:</i> Chr17:37879588<br><i>GRCh38:</i> Chr17:39723335                                | <a href="#">ERBB2</a> |
| 23. | <input type="checkbox"/> NM_001005862.2(ERBB2):c.2173_2174 delTTinsCC ( p.Leu725Pro)<br><i>GRCh37:</i> Chr17:37880219-37880220<br><i>GRCh38:</i> Chr17:39723966-39723967 | <a href="#">ERBB2</a> |

## Condition Significance

| Condition(s)                                | Frequency                            | Clinical significance<br>(Last reviewed) | Review status                  |
|---|--------------------------------------|--|--------------------------------|
| not specified                               | GMAF:0.00040(T)                      | not provided<br>(Sep 19, 2013)           | no assertion provided          |
| not specified                               |                                      | not provided<br>(Sep 19, 2013)           | no assertion provided          |
| <b>ERBB2</b><br>POLYMORPHISM, not specified | GO-ESP:0.00707(G)<br>GMAF:0.00260(G) | Benign<br>(Feb 1, 1993)                  | no assertion criteria provided |
| <b>ERBB2</b><br>POLYMORPHISM, not specified | GO-ESP:0.16854(G)<br>GMAF:0.12140(G) | Benign<br>(Feb 1, 1993)                  | no assertion criteria provided |
| Adenocarcinoma of lung                      |                                      | Pathogenic<br>(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



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

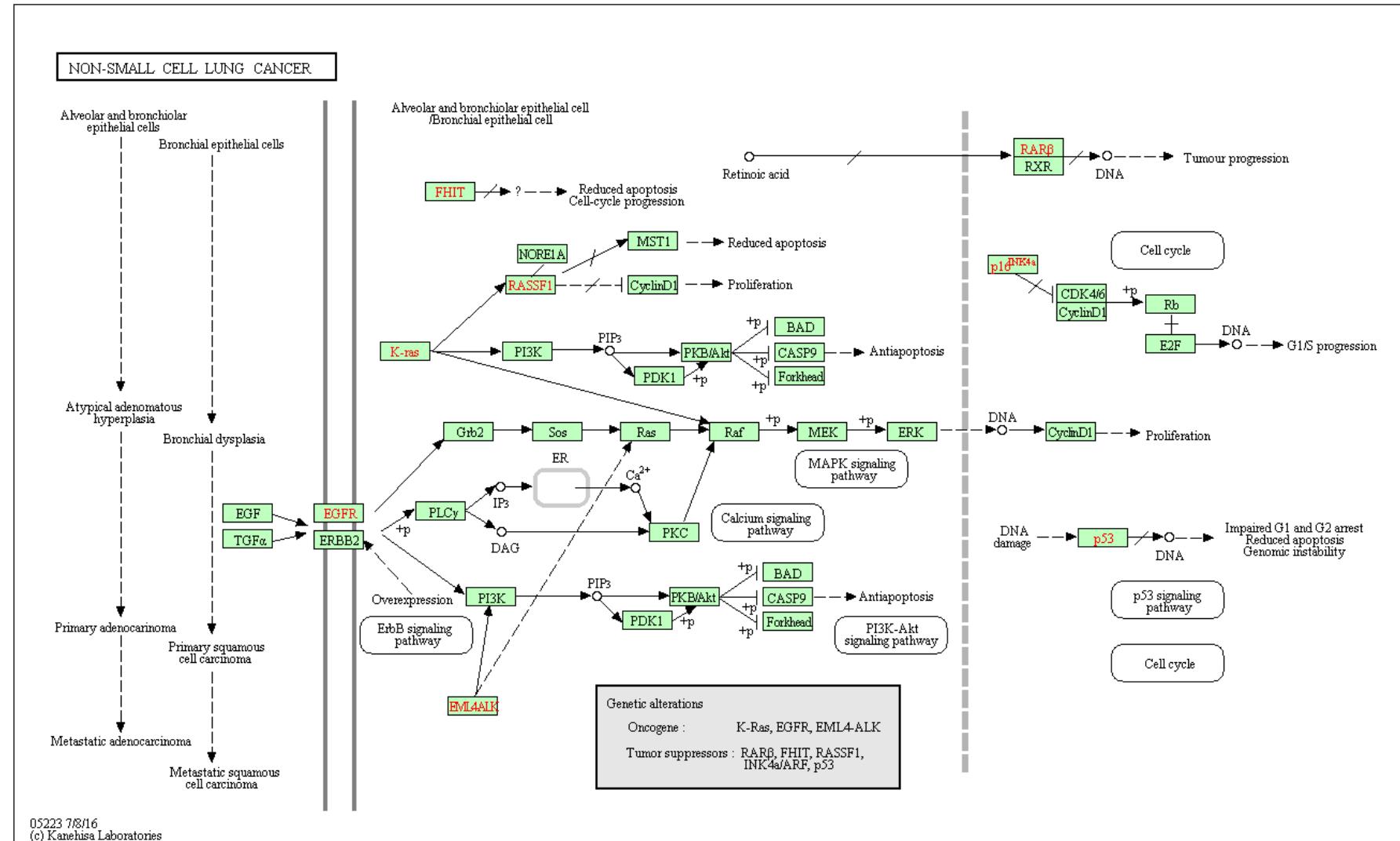
### 3. Environmental Information Processing

### 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

[Switch to classic site](#)

**rs121913529**

Current Build 152  
Released October 2, 2018

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

**FEEDBACK**

Variant Details      **50 citations for rs121913529**

Filter:

| PMID   | Title   | Author               | Year | Journal                              |
|--|---|----------------------|------|--------------------------------------|
| <a href="#">263727</a><br><a href="#">03</a> | Prognostic value of the KRAS G12V mutation in 841 surgically resected Caucasian lung adenocarcinoma cases.  | Renaud S et al.      | 2015 | British journal of cancer            |
| <a href="#">251579</a><br><a href="#">68</a> | Prospective enterprise-level molecular genotyping of a cohort of cancer patients.   | MacConaill LE et al. | 2014 | The Journal of molecular diagnostics |
| <a href="#">250441</a><br><a href="#">03</a> | 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. | Chaff JE et al.      | 2014 | Clinical lung cancer                 |
| <a href="#">234060</a><br><a href="#">27</a> | Selumetinib-enhanced radioiodine uptake in advanced thyroid cancer.   | Ho AL et al.         | 2013 | The New England journal of medicine  |

# 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<sup>i</sup>

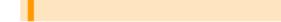
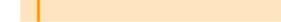
<http://www.uniprot.org>

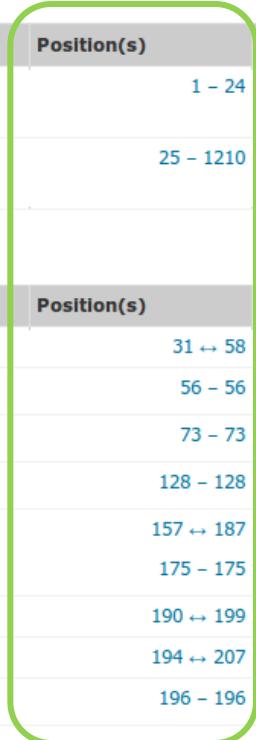
### PTM / Processing<sup>i</sup>

#### Molecule processing

| Feature key                 | Position(s) | Length | Description   | Graphical view  | Feature identifier | Actions  |
|-----------------------------|-------------|--------|---|---|--------------------|--|
| Signal peptide <sup>i</sup> | 1 – 24      | 24     |  2 Publications  |  |                    |  Add<br> BLAST |
| Chain <sup>i</sup>          | 25 – 1210   | 1186   | Epidermal growth factor receptor  |  | PRO_0000016665     |  Add<br> BLAST |

#### Amino acid modifications

| Feature key                 | Position(s) | Length | Description   | Graphical view  | Feature identifier | Actions  |
|-----------------------------|-------------|--------|---|---|--------------------|--|
| Disulfide bond <sup>i</sup> | 31 ↔ 58     |        |   |    |                    |  |
| Glycosylation <sup>i</sup>  | 56 – 56     |        | 1 N-linked (GlcNAc...) (complex); atypical; partial |    | CAR_000227         |  Add<br> BLAST    |
| Glycosylation <sup>i</sup>  | 73 – 73     |        | 1 N-linked (GlcNAc...) (complex); atypical          |  |                    |  Add<br> BLAST  |
| Glycosylation <sup>i</sup>  | 128 – 128   |        | 1 N-linked (GlcNAc...) (complex); atypical          |  |                    |  Add<br> BLAST |
| Disulfide bond <sup>i</sup> | 157 ↔ 187   |        |   |  |                    |  |
| Glycosylation <sup>i</sup>  | 175 – 175   |        | 1 N-linked (GlcNAc...) (complex); atypical          |  |                    |  Add<br> BLAST |
| Disulfide bond <sup>i</sup> | 190 ↔ 199   |        |   |  |                    |  |
| Disulfide bond <sup>i</sup> | 194 ↔ 207   |        |   |  |                    |  |
| Glycosylation <sup>i</sup>  | 196 – 196   |        | 1 N-linked (GlcNAc...) (complex); atypical          |  |                    |  Add<br> BLAST |



# Annotation Tools

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

| ANNOVAR                  | Web report        |
|--------------------------|-------------------|
| Variant Effect Predictor | NCBI Rep          |
| VarAFT                   | predicts ESE acti |
| Snpeff                   |                   |

| 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  | -   | -   | -   | -   |     |     |     |  |  |  |
| pfSNP [127]                           | SNP ID; chromosome region; Gene ID;                        | Web report  | yes | no  | no  | no  |     |     |     |  |  |  |
| PHAST [128]                           | FASTA, PHYLIP, MPM, MAF, SS                                | Conservation score  | -   | -   | -   | -   |     |     |     |  |  |  |
| PhD-SNP [129]                         | DNA sequence; code; Swiss-Prot protein code; Sequence file | Effect prediction   | yes | no  | no  | no  |     |     |     |  |  |  |
| PMUT [130]                            | FASTA, PDB ID, UniProt ID, SWISSProt code                  | Web report  | -   | -   | -   | -   |     |     |     |  |  |  |
| PolyDomms [131]                       | Gene/protein symbol(s), RefSeqID dbSNP ID                  | Web report  | yes |     |     |     |     |     |     |  |  |  |
| PolyMAPr [132]                        | -  | -   | -   | -   | -   | -   |     |     |     |  |  |  |
| PolyPhen-2 [133]                      | UniProt ID, FASTA, dbSNP ID                                | CSV in PolyPhen format  | yes |     |     |     |     |     |     |  |  |  |
| PupaSNP Finder [134]                  | dbSNP ID; GenBank transcript ID; PED format                | Web report  | yes |     |     |     |     |     |     |  |  |  |
| QuickSNP [135]                        | gene symbol  | Web report  | yes |     |     |     |     |     |     |  |  |  |
| RescueESE [136]                       | plain text; mRNA sequence                                  | predicts sequences with ESE activity                                | -   | -   | -   | -   |     |     |     |  |  |  |
| SAPRED [137]                          | FASTA and VCF  | -   | yes |     |     |     |     |     |     |  |  |  |
| SCAN [138]                            | -  | Web report  | yes |     |     |     |     |     |     |  |  |  |
| SCONE [139]                           | MAF  | Conservation score  | -   | -   | -   | -   |     |     |     |  |  |  |
| SeattleSeq Annotation [140]           | Maq, GFFm CASAVA, VCF, GATK bed                            | VCF, own format   | yes | yes | yes | yes |     |     |     |  |  |  |
| SeqAnt [141]                          | FASTA sequence file  | Web report  | yes | yes | no  | 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<br>--dbSNP/Kegg/GO/1000 Genomes/DGV | yes | yes | no  | yes | yes | yes | no  |  |  |  |

# Variant Effect Predictor (VEP)

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

The screenshot shows the VEP page on the Ensembl website. The header includes the Ensembl logo, navigation links for BLAST/BLAT, BioMart, VEP, Tools, Downloads, Help & Docs, and Blog, and a search bar. The main content area has a sidebar titled 'In this section' with links for Web interface, VEP script (including Tutorial, Download and install, Running VEP, Annotation sources, Filtering results, Custom annotations, Plugins, Examples and use cases, Other information, Data formats, and FAQ), and a 'Search documentation...' button. The main content area features the 'Variant Effect Predictor' logo and a brief description of its function: determining the effect of variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions. It also lists what can be found by inputting variant coordinates and nucleotide changes. Below this is a list of execution methods: Web interface, Standalone Perl script, and REST API.

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

With **several ways**  
**of execution**

The image compares three ways to execute VEP:

- Web interface:** Point-and-click interface, suits smaller volumes of data. Includes a 'Launch Ve!P' button.
- Standalone Perl script:** More options and flexibility, for large volumes of data. Includes 'Clone from GitHub' and 'Download (zip)' buttons.
- REST API:** Language-independent API, simple URL-based queries. Includes a 'VEP REST API' button.

# 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

**CONDEL**

**Sift and PolyPhen  
consensus**

[http://www.ensembl.org/info/docs/tools/vep/script/vep\\_options.html#basic](http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html#basic)

# VEP standalone script output: html

**e!Ensembl** **Ve!P**

**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**

|                             |   |
|-----------------------------|---|
| <b>VEP version (API)</b>    | 85 (85)   |
| <b>Cache/Database</b>       | /home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/vep/homo_sapiens/85_GRCh37  |
| <b>Species</b>              | homo_sapiens  |
| <b>Command line options</b> | --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 |
| <b>Start time</b>           | 2016-09-03 09:01:35   |
| <b>End time</b>             | 2016-09-03 09:43:38   |
| <b>Run time</b>             | 2523 seconds  |
| <b>Input file (format)</b>  | /home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf (VCF)  |
| <b>Output file</b>          | /home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf_output_VEP.txt [ <a href="#">text</a> ]  |

**General statistics**

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

**Variant classes**

A donut chart with a single blue segment labeled "100%" in white text, representing the proportion of SNVs.

**Variants by chromosome**

A bar chart showing the number of variants per chromosome. The x-axis labels are 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, X, and Y. The y-axis ranges from 0 to 500. The highest counts are on chromosomes 1, 2, and 4.

| Chromosome | Number of Variants |
|------------|--------------------|
| 1          | ~380               |
| 2          | ~380               |
| 4          | ~380               |
| 5          | ~350               |
| 6          | ~320               |
| 7          | ~320               |
| 8          | ~320               |
| 9          | ~200               |
| 10         | ~180               |
| 11         | ~220               |
| 12         | ~180               |
| 13         | ~180               |
| 14         | ~150               |
| 15         | ~120               |
| 16         | ~120               |
| 17         | ~100               |
| 18         | ~100               |
| 19         | ~100               |
| 20         | ~100               |
| 21         | ~50                |
| X          | ~150               |
| Y          | ~50                |

**SNV**

# VEP standalone script output: VCF

Annotations per transcript are separated by comma

## One set of annotations per transcript

## Order of annotations in INFO column

# Annotations starts with **CSQ=** tag

Each individual annotation is separated by |

# web execution

**e!Ensembl** BLAST/BLAT | BioMart | VEP | Tools | Downloads | Help & Docs | Blog

Login/Register

Using this website Annotation and prediction Data access API & software About us

In this section

- Web interface
  - Input form
  - Results
- VEP script
  - Tutorial
  - Download and install
  - Running VEP
  - Annotation sources
  - Filtering results
  - Custom annotations
  - Plugins
  - Examples and use cases
  - Other information
- Data formats
- FAQ

Search documentation... Go

Help & Documentation > API & Software > Ensembl Tools > Variant Effect Predictor

## Variant Effect Predictor

**VEP determines the effect of your variants** (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

Simply input the coordinates of your variants and the nucleotide changes to find out the:

- Genes and Transcripts affected by the variants
- Location of the variants (e.g. upstream of a transcript, in coding sequence, in non-coding RNA, in regulatory regions)
- Consequence of your variants on the protein sequence (e.g. stop gained, missense, stop lost, frameshift)
- Known variants that match yours, and associated minor allele frequencies from the **1000 Genomes Project**
- SIFT and PolyPhen scores for changes to protein sequence
- ... And more! See [data types](#), [versions](#).

**Ve!P**

**Web interface**



- Point-and-click interface
- Suits smaller volumes of data

[Documentation](#)

**Launch Ve!P**

**Standalone Perl script**



- More options and flexibility
- For large volumes of data

[Documentation](#)

[Clone from GitHub](#)

[Download \(zip\)](#)

**REST API**



- Language-independent API
- Simple URL-based queries

[Documentation](#)

**VEP REST API**

# web execution: options

Variant Effect Predictor ?

<http://grch37.ensembl.org/info/docs/tools/vep/index.html>

hg19

New job

Clear form

Species:



Homo\_sapiens

X

hg38

Assembly: GRCh38.p13  
[Add/remove species](#)

Name for this job (optional):

Input data:

Either paste data:

Examples: [Ensembl default](#), [VCF](#), [Variant identifiers](#), [HGVS notations](#)

Or upload file:  Choose File no file selected

Or provide file URL:

Input  
options

Transcript database to use:

- Ensembl/Gencode transcripts
- Ensembl/Gencode basic transcripts
- RefSeq transcripts
- Ensembl/Gencode and RefSeq transcripts

## Identifiers

### Identifiers

Gene symbol:

CCDS:

Protein:

UniProt:

HGVS:

## Variants and frequency data

Variants and frequency data

Find co-located known variants:

Yes

Frequency data for co-located variants:

- 1000 Genomes global minor allele frequency
- 1000 Genomes continental allele frequencies
- ESP allele frequencies
- gnomAD (exomes) allele frequencies

PubMed IDs for citations of co-located variants:

Include flagged variants:

## Additional annotations

Additional transcript, protein and regulatory annotations

### Transcript annotation

Transcript biotype:

Exon and intron numbers:

Transcript support level:

APPRIS:

Identify canonical transcripts:

Upstream/Downstream distance (bp):

5000

miRNA structure:

### Protein annotation

Protein domains:

### Regulatory data

Get regulatory region consequences:

Yes

## Predictions

Variant predictions, e.g. SIFT, PolyPhen

### Pathogenicity predictions

SIFT:

Prediction and score

PolyPhen:

Prediction and score

dbNSFP:

Disabled

Enabled

Condel:

Disabled

Enabled

LoFTool:

Disabled

### Splicing predictions

dbSCNV:

Disabled

MaxEntScan:

Disabled

### Conservation

BLOSUM62:

Disabled

Ancestral allele:

Disabled

Filtering options

Pre-filter results by frequency or consequence type

Run >

Calculates the Consensus Deleteriousness (Condel) score for a missense mutation based on the pre-calculated SIFT and PolyPhen scores

web execution

# web execution



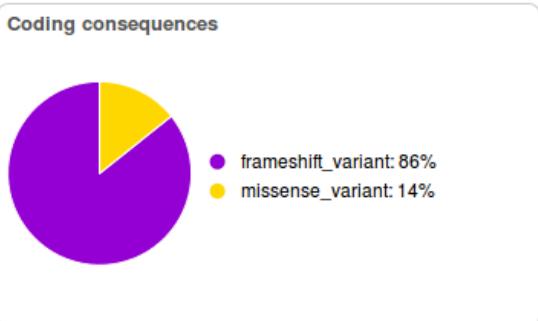
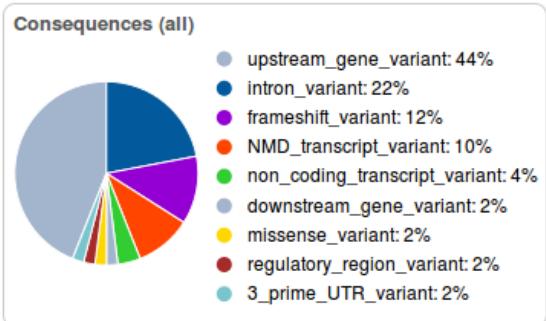
# web execution: output

## Variant Effect Predictor results

Job details 

Summary statistics 

| Category                           | Count |
|------------------------------------|-------|
| Variants processed                 | 3     |
| Variants remaining after filtering | 3     |
| Novel / existing variants          | -     |
| Overlapped genes                   | 4     |
| Overlapped transcripts             | 42    |
| Overlapped regulatory features     | 1     |



Summary

Results preview

Filtering

Navigation  
Page: < > 1 of 1 >> | Show: [1 All](#) variants

Filters

Uploaded variant is defined [Add](#)

Download  
All: [VCF VEP TXT](#)  
BioMart: Variants [Genes](#)

Download

Results table

| Show/hide columns |                                 |        |                       |          |            |                                 |              |                                 |                |      |        |       |       |               |              |                  |
|-------------------|---------------------------------|--------|-----------------------|----------|------------|---------------------------------|--------------|---------------------------------|----------------|------|--------|-------|-------|---------------|--------------|------------------|
| Uploaded variant  | Location                        | Allele | Consequence           | Impact   | Symbol     | Gene                            | Feature type | Feature                         | Biotype        | Exon | Intron | HGVSc | HGVSp | cDNA position | CDS position | Protein position |
| 1_818046_T/C      | <a href="#">1:818046-818046</a> | C      | missense_variant      | MODERATE | AL645608.2 | <a href="#">ENSG00000269308</a> | Transcript   | <a href="#">ENST00000594233</a> | protein_coding | 1/3  | -      | -     | -     | 4             | 4            | 2                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | intron_variant        | MODIFIER | ACP1       | <a href="#">ENSG00000143727</a> | Transcript   | <a href="#">ENST00000272065</a> | protein_coding | -    | 1/5    | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | intron_variant        | MODIFIER | ACP1       | <a href="#">ENSG00000143727</a> | Transcript   | <a href="#">ENST00000272067</a> | protein_coding | -    | 1/5    | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | upstream_gene_variant | MODIFIER | SH3YL1     | <a href="#">ENSG00000035115</a> | Transcript   | <a href="#">ENST00000356150</a> | protein_coding | -    | -      | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | upstream_gene_variant | MODIFIER | SH3YL1     | <a href="#">ENSG00000035115</a> | Transcript   | <a href="#">ENST00000402632</a> | protein_coding | -    | -      | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | upstream_gene_variant | MODIFIER | SH3YL1     | <a href="#">ENSG00000035115</a> | Transcript   | <a href="#">ENST00000403657</a> | protein_coding | -    | -      | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | upstream_gene_variant | MODIFIER | SH3YL1     | <a href="#">ENSG00000035115</a> | Transcript   | <a href="#">ENST00000403658</a> | protein_coding | -    | -      | -     | -     | -             | -            | -                |
| 2_265023_C/A      | <a href="#">2:265023-265023</a> | A      | upstream_gene_variant | MODIFIER | SH3YL1     | <a href="#">ENSG00000035115</a> | Transcript   | <a href="#">ENST00000403712</a> | protein_coding | -    | -      | -     | -     | -             | -            | -                |



Specific cancer  
information



---

Disease databases

ClinVar

<http://www.ncbi.nlm.nih.gov/cinvar>

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.



## NIH NATIONAL CANCER INSTITUTE GDC Data Portal

## 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

## Data Portal Summary

Data Release 30.0 - September 23, 2021

## PROJECTS

70

## PRIMARY SITES

67

## CASES

85,414

## FILES

619,488

## GENES

23,621

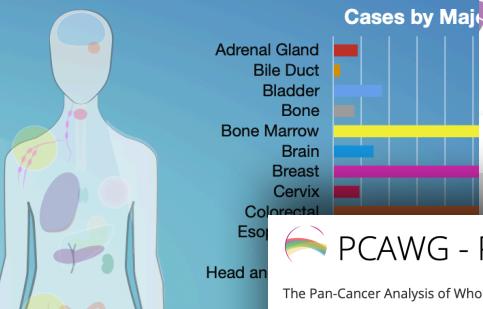
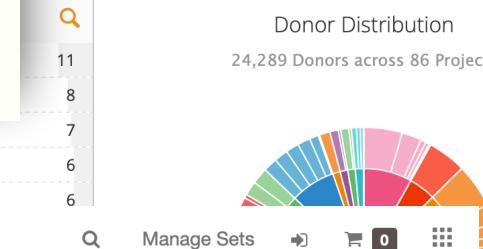
## MUTATIONS

3,599,319



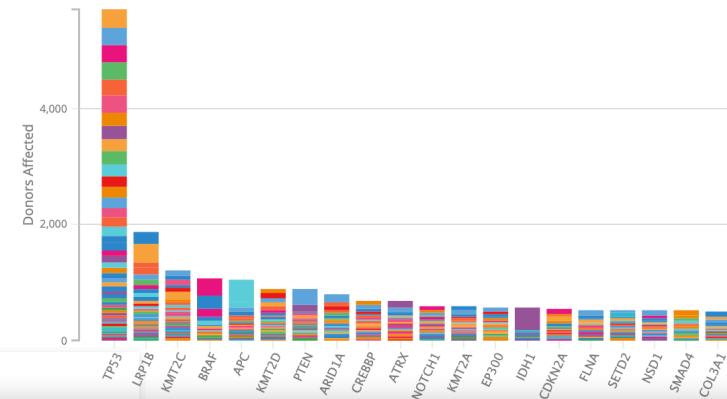
## NCI PROJECTS

Summary      Details      History



Top 20 Mutated Cancer Genes with High Functional Impact SSMs

19,729 Unique SSM-Tested Donors

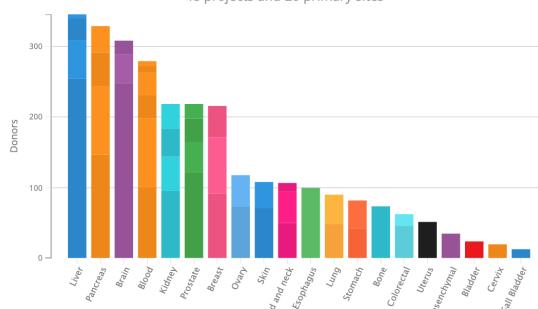


## PCAWG - PANCANCER ANALYSIS OF WHOLE GENOMES

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



# 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/>

The screenshot shows the TumorPortal homepage with a dark blue background featuring a green and blue stylized human figure logo. At the top right are links for 'Sign In', 'Go to Gene', and a search bar. The main header reads 'Welcome to TumorPortal' with subtext 'Genes, Cancers, DNA Mutations & Annotations'. Below this is a section titled 'Explore the pan-cancer dataset:' with buttons for 'by Tumor Types', 'by Genes', and 'by Figures'. A link 'Questions or comments? Please contact us.' is also present. The bottom half of the page is divided into two sections: 'Explore dataset by tumor types' (listing various tumor types like AML, BLCA, BRCA, etc.) and 'Explore dataset by Genes' (listing genes like TP53, PIK3CA, PTEN, KRAS, etc.). Each item in these lists includes a percentage value indicating its prevalence.

| Tumor Type                          | Number of Patients |
|-------------------------------------|--------------------|
| Acute myeloid leukemia AML          | 196 patients       |
| Bladder BLCA                        | 99 patients        |
| Breast BRCA                         | 892 patients       |
| Carcinoid CARC                      | 54 patients        |
| Chronic lymphocytic leukemia CLL    | 159 patients       |
| Colorectal CRC                      | 233 patients       |
| Diffuse large B-cell lymphoma DLBCL | 58 patients        |
| Esophageal adenocarcinoma ESO       | 141 patients       |
| Glioblastoma multiforme GBM         | 291 patients       |
| Head and neck HNSC                  | 384 patients       |
| Kidney clear cell KIRC              | 417 patients       |
| Lung adenocarcinoma LUAD            | 405 patients       |
| Lung squamous cell carcinoma LUSC   | 178 patients       |
| Medulloblastoma MED                 | 92 patients        |
| Melanoma MEL                        | 118 patients       |
| Multiple myeloma MM                 | 207 patients       |
| Neuroblastoma NB                    | 81 patients        |
| Ovarian OV                          | 316 patients       |
| combined cohort PanCan              | 4742 patients      |
| Prostate PRAD                       | 138 patients       |
| Rhabdoid tumor RHAB                 | 35 patients        |
| Endometrial UCEC                    | 248 patients       |

| Gene       | Percentage of all patients |
|------------|----------------------------|
| TP53       | 36% of all patients        |
| PIK3CA     | 14% of all patients        |
| PTEN       | 7% of all patients         |
| KRAS       | 7% of all patients         |
| APC        | 6% of all patients         |
| MLL3       | 6% of all patients         |
| FAT1       | 6% of all patients         |
| MLL2       | 5% of all patients         |
| ARID1A     | 5% of all patients         |
| VHL        | 4% of all patients         |
| PBRM1      | 4% of all patients         |
| NF1        | 4% of all patients         |
| EGFR       | 4% of all patients         |
| ATM        | 4% of all patients         |
| PIK3R1     | 3% of all patients         |
| BRAF       | 3% of all patients         |
| CDKN2A     | 3% of all patients         |
| SETD2      | 3% of all patients         |
| CREBBP     | 3% of all patients         |
| FBXW7      | 3% of all patients         |
| SPEN       | 3% of all patients         |
| MTOR       | 3% of all patients         |
| RB1        | 2% of all patients         |
| SMARCA4    | 2% of all patients         |
| NOTCH1     | 2% of all patients         |
| Other Gene | Gene Symbol                |

**ICGC-TCGA Consortium**



## COSMIC v83, released 07-NOV-17

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell

**SEARCH**

## COSMIC News

 Follow @cosmic\_sanger

### COSMIC Release v83

The November release (v83) is now live! New in this release we have 3 fully curated genes: TGFBR2, ERBB4 and BCL9L; a substantial update to VHL; and the new fusion pair ETV6-ABL1. [More...](#)

### Switch between GRCh37 and GRCh38

You still have access to GRCh37 and it is now possible to switch between GRCh37 and GRCh38 from any page within the main site. [More...](#)

### Cancer Gene Census changes

As we mentioned at the last release, as part of integrating the Hallmarks of Cancer feature, the Cancer Gene Census (CGC) has had a thorough re-evaluation. [More...](#)

## Tools

- ➊ [Cancer browser](#) — browse COSMIC data by tissue type and histology
- ➋ [Genome browser](#) — browse the human genome with COSMIC annotations
- ➌ [CONAN](#) — the COSMIC copy number analysis tool
- ➍ [GA4GH Beacon](#) — access COSMIC data through the [GA4GH Beacon Project](#)
- ➎ [COSMIC in BigQuery](#) — search COSMIC via the [ISB Cancer Genomics Cloud](#)

## Help

- ➊ [Downloads](#) — data that you can download from our SFTP site
- ➋ [Documentation](#) — view our help documentation
- ➌ [FAQ](#) — a compilation of our Frequently Asked Questions
- ➍ [Release notes](#) — Information about the latest COSMIC release
- ➎ [Licensing](#) — information about our licensing policy

variant/gene frequency  
in database





## 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 |
|----------------------|---|----------------|------------------|---------|------------------|
| <a href="#">PTEN</a> | <a href="#">PTEN</a> , <a href="#">ENST00000371953</a> , <a href="#">PTEN.html...</a> | 69987          | 3975             | 0       | 3975             |

Showing 1 to 1 of 1 entries

First Previous **1** Next Last

[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 |
|----------------------|----------------------------------|--|------------|
| <a href="#">PTEN</a> | <a href="#">c.388C&gt;G</a>      | <a href="#">PTEN c.388C&gt;G...</a>      | 162        |
| <a href="#">PTEN</a> | <a href="#">c.389G&gt;A</a>      | <a href="#">PTEN c.389G&gt;A...</a>      | 134        |
| <a href="#">PTEN</a> | <a href="#">c.697C&gt;T</a>      | <a href="#">PTEN c.697C&gt;T...</a>      | 124        |
| <a href="#">PTEN</a> | <a href="#">c.388C&gt;T</a>      | <a href="#">PTEN c.388C&gt;T...</a>      | 94         |
| <a href="#">PTEN</a> | <a href="#">c.?_?del?</a>        | <a href="#">PTEN c.?_?del?...</a>        | 91         |
| <a href="#">PTEN</a> | <a href="#">c.?</a>              | <a href="#">PTEN c.?</a>                 | 88         |
| <a href="#">PTEN</a> | <a href="#">c.1_1212del1212</a>  | <a href="#">PTEN c.1_1212del1212...</a>  | 88         |
| <a href="#">PTEN</a> | <a href="#">c.800delA</a>        | <a href="#">PTEN c.800delA...</a>        | 56         |
| <a href="#">PTEN</a> | <a href="#">c.517C&gt;T</a>      | <a href="#">PTEN c.517C&gt;T...</a>      | 47         |
| <a href="#">PTEN</a> | <a href="#">c.950_953delTACT</a> | <a href="#">PTEN c.950_953delTACT...</a> | 44         |

Showing 1 to 10 of 1,968 entries

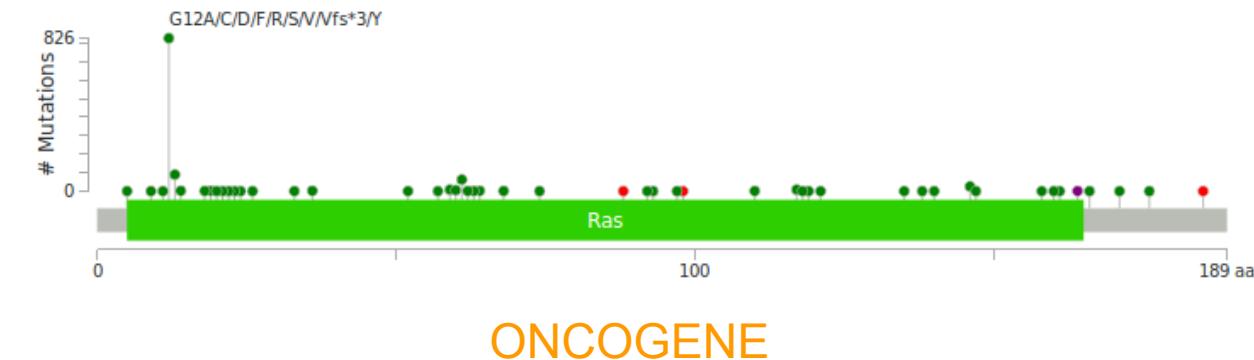
First Previous **1** 2 3 4 5 ... 197 Next Last

**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

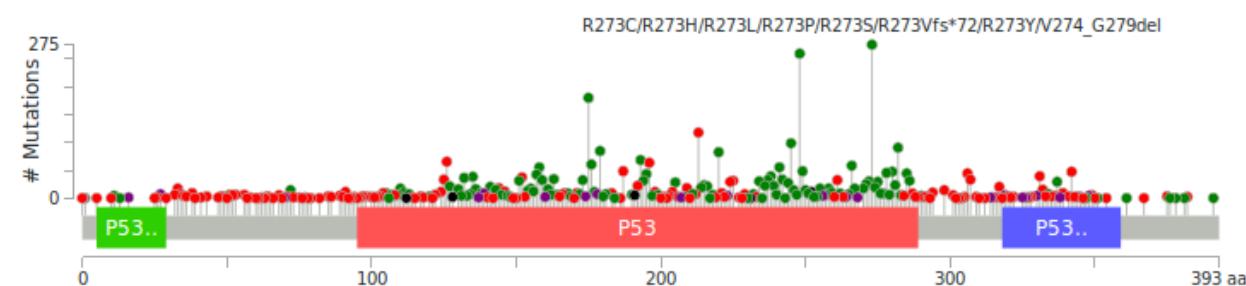
**KRAS:**  
RASK\_HUMAN PDF SVG Customize Color Codes



ONCOGENE

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

**TP53:**  
P53\_HUMAN PDF SVG Customize Color Codes



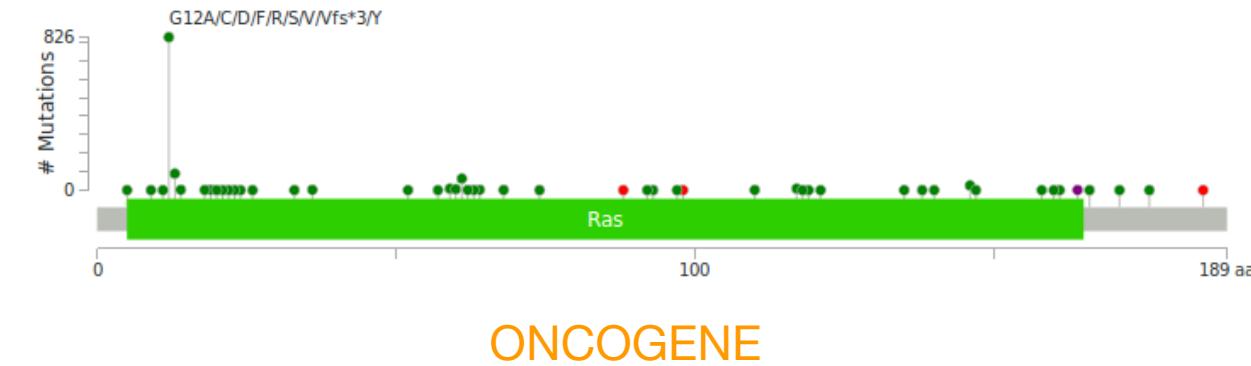
TUMOR SUPPRESSOR GENE

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

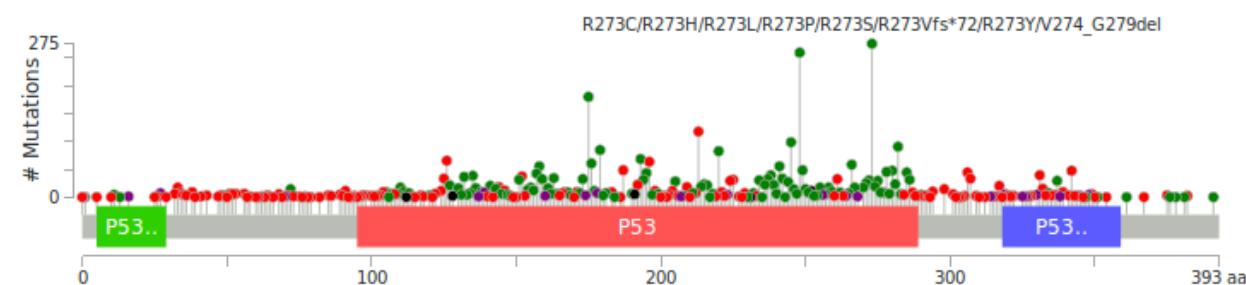
**KRAS:**  
RASK\_HUMAN [PDF](#) [SVG](#) [Customize](#) [Color Codes](#)



ONCOGENE

**High frequency** of  
variant expected in a  
cohort  
Direct inhibition

**TP53:**  
P53\_HUMAN [PDF](#) [SVG](#) [Customize](#) [Color Codes](#)



TUMOR SUPPRESSOR GENE

**Lower frequency** of  
variant expected in a  
cohort  
Downstream inhibition



## COSMIC v83, released 07-NOV-17

COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer.

Start using COSMIC by searching for a gene, cancer type, mutation, etc. below.

eg Braf, COLO-829, Carcinoma, V600E, BRCA-UK, Campbell

SEARCH

### Projects

COSMIC is divided into several distinct projects, each presenting a separate dataset or view of our data:



#### COSMIC

The core of COSMIC, an expert-curated database of somatic mutations



#### Cell lines project

Mutation profiles of over 1,000 cell lines used in cancer research



#### COSMIC-3D

An interactive view of cancer mutations in the context of 3D structures



#### Cancer Gene Census

A catalogue of genes with mutations that are causally implicated in cancer

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

### Data curation

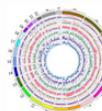
- ❶ [Gene curation](#) — details of our manual curation process
- ❶ [Gene fusion curation](#) — details of our curation process for gene fusions
- ❶ [Genome Annotation](#) — information on the annotation of genomes
- ❶ [Drug Resistance](#) — curation of mutations conferring drug resistance

### COSMIC News

[Follow @cosmic\\_sanger](#)

#### COSMIC Release v83

The November release (v83) is now live! New in this release we have 3 fully curated genes: TGFBR2, ERBB4 and BCL9L; a substantial update to VHL; and the new fusion pair ETV6-ABL1. [More...](#)



#### Switch between GRCh37 and GRCh38

You still have access to GRCh37 and it is now possible to switch between GRCh37 and GRCh38 from any page within the main site. [More...](#)



#### Cancer Gene Census changes

As we mentioned at the last release, as part of integrating the Hallmarks of Cancer feature, the Cancer Gene Census (CGC) has had a thorough re-evaluation. [More...](#)

### Tools

- ❶ [Cancer browser](#) — browse COSMIC data by tissue type and histology
- ❶ [Genome browser](#) — browse the human genome with COSMIC annotations
- ❶ [CONAN](#) — the COSMIC copy number analysis tool
- ❶ [GA4GH Beacon](#) — access COSMIC data through the [GA4GH Beacon Project](#)
- ❶ [COSMIC in BigQuery](#) — search COSMIC via the [ISB Cancer Genomics Cloud](#)

### Help

- ❶ [Downloads](#) — data that you can download from our SFTP site
- ❶ [Documentation](#) — view our help documentation
- ❶ [FAQ](#) — a compilation of our Frequently Asked Questions
- ❶ [Release notes](#) — Information about the latest COSMIC release
- ❶ [Licensing](#) — information about our licensing policy

<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 Genelid         | Genome Location            | Tier | Hallmark | Chr Band | Somatic | Germline | Tumour Types(Somatic)                     | Tumour Types(Germline) | Cancer Syndrome | Tissue Type |
|--------------------------|--|------------------------|----------------------------|------|----------|----------|---------|----------|---|------------------------|-----------------|-------------|
| <a href="#">ARHGAP5</a>  | Rho GTPase activating protein 5                        | <a href="#">394</a>    | 14:32090670-32154948<br>   | 2    |          | 14q12    | yes     |          | colon cancer; glioma                      |                        | E; O            |             |
| <a href="#">ARHGEF12</a> | RHO guanine nucleotide exchange factor (GEF) 12 (LARG) | <a href="#">23365</a>  | 11:120337244-120485077<br> | 1    |          | 11q23.3  | yes     |          | AML                                       |                        | L               | D           |
| <a href="#">ARID1A</a>   | AT rich interactive domain 1A (SWI-like)               | <a href="#">8289</a>   | 1:26696404-26780756<br>    | 1    |          | 1p35.3   | yes     |          | clear cell ovarian carcinoma; RCC; breast |                        | E               | R           |
| <a href="#">ARID1B</a>   | AT rich interactive domain 1B                          | <a href="#">57492</a>  | 6:156778104-157207891<br>  | 1    |          | 6q25.1   | yes     |          | breast; hepatocellular carcinoma          |                        | E               | R           |
| <a href="#">ARID2</a>    | AT rich interactive domain 2                           | <a href="#">196528</a> | 12:45729837-45905078<br>   | 1    |          | 12q12    | yes     |          | hepatocellular carcinoma                  |                        | E               | R           |



# 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.

Loss of function cutoff: 0.3

Activating cutoff: 0.7

Cancer driver list

Cancer5000

HCD

[Download classification](#)

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)

Elena Piñeiro-Yáñez

[epineiro@cnio.es](mailto:epineiro@cnio.es)

Coral Fustero-Torre

[cfustero@cnio.es](mailto:cfustero@cnio.es)

María José Jiménez-Santos

[mjjimenez@cnio.es](mailto:mjjimenez@cnio.es)

Fátima Al-Shahrour

[falshahrour@cnio.es](mailto:falshahrour@cnio.es)