

Precision medicine: NGS variant analysis and interpretation for translational research

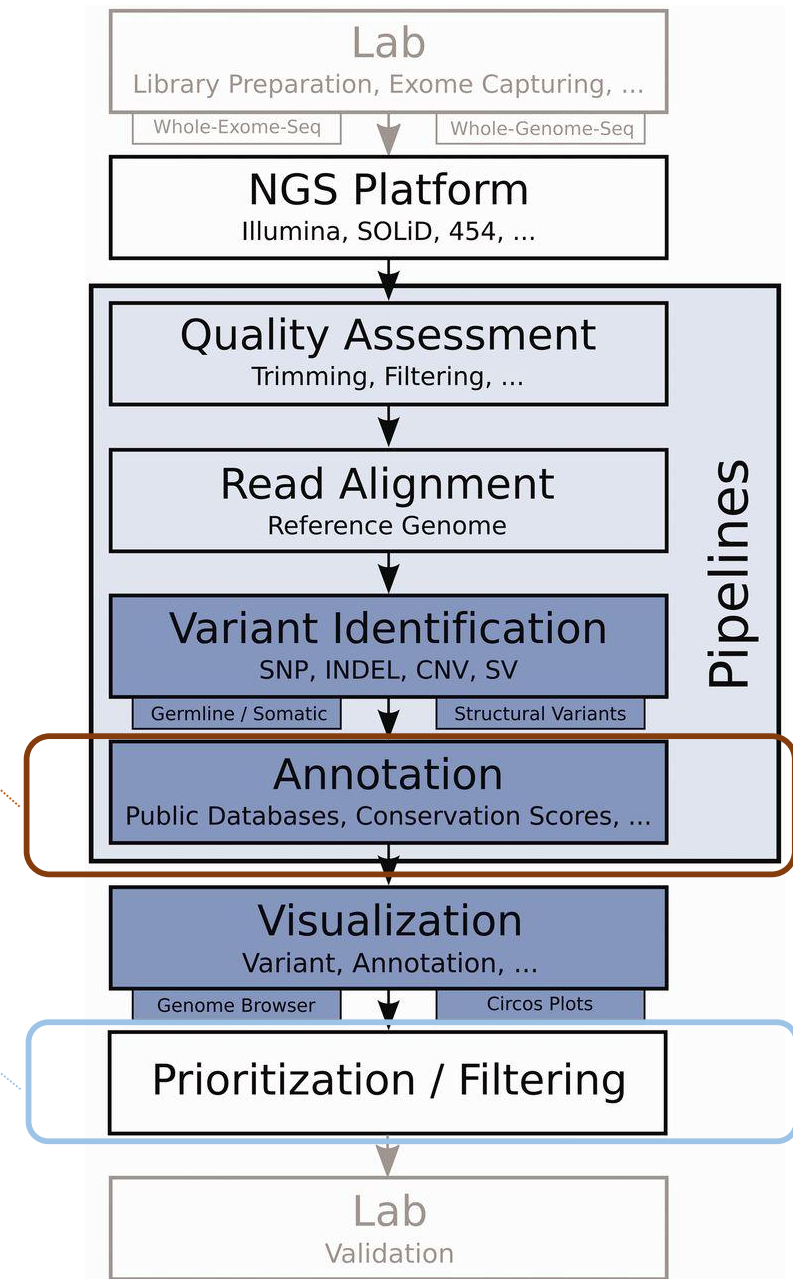
NGSII: Variant annotation

Fátima Al-Shahrour ● Javier Perales ● Elena Piñeiro

September 28, 2016

NGSII: Variant Annotation

Selecting the most relevant variants: How to filter



Variant annotation

- Technical information: quality parameters, filters, ...
- Descriptive information: nomenclature, genotype, ...
- Functional annotation: consequence, functional prediction, ...

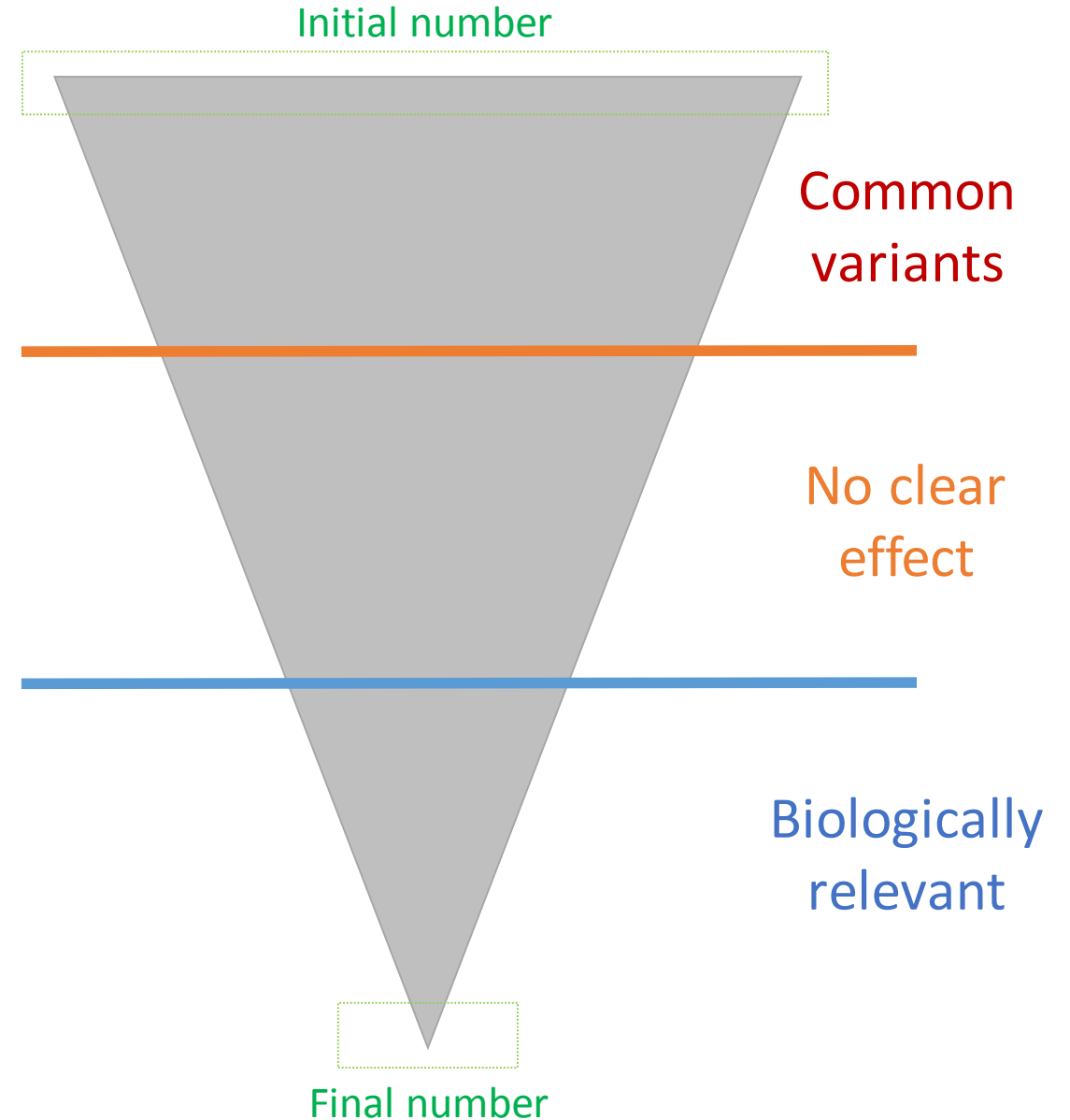
Variant callers provide **technical parameters** associated to each variant, that allow us to remove sequencing artifacts and select the most reliable variants.

Variant annotation

Not all the variants have a relevant impact in the phenotype of study. They are mainly polymorphisms (maybe predisposition) or variants without a clear effect.

To identify and select the most relevant variants we need to add **descriptive and functional annotations**.

There are different software programs that perform variant annotation.



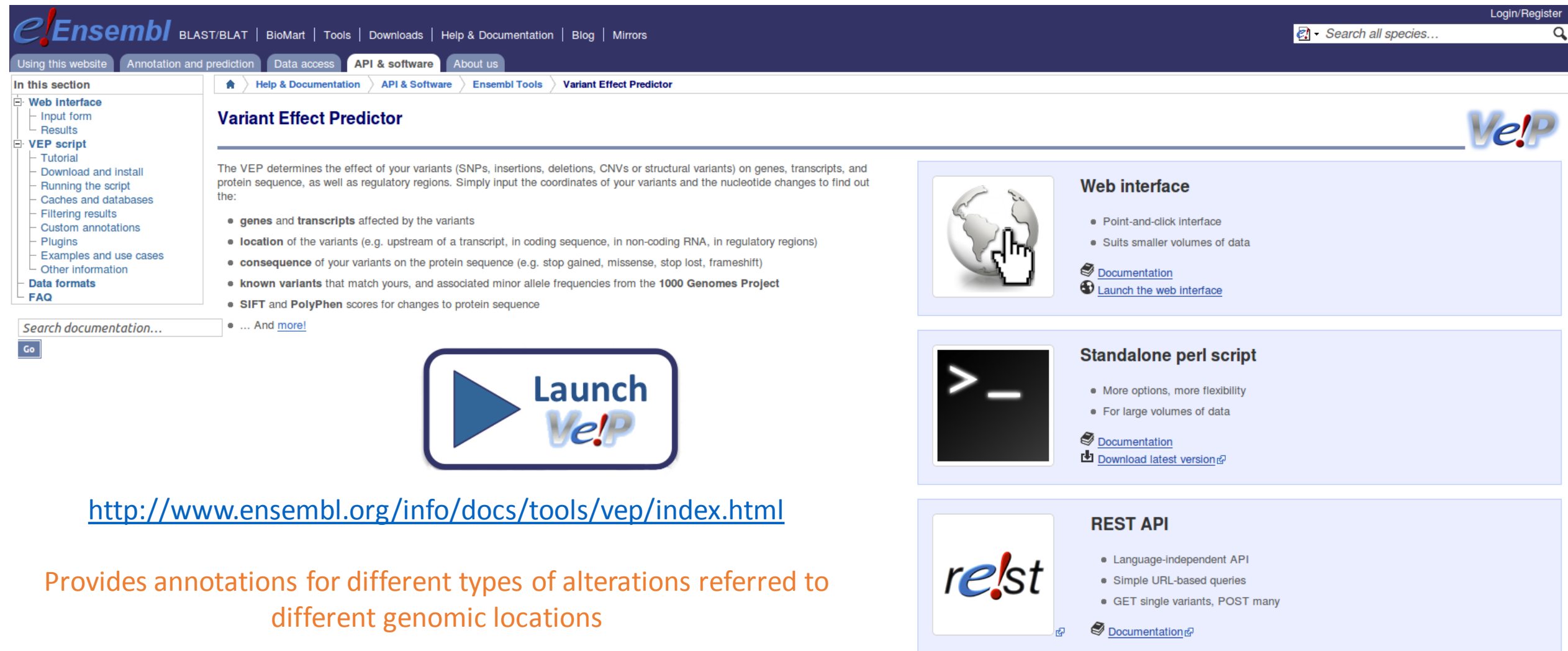
Name	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
ABSOLUTE [91]	HAPSEG output, sample level variance, precomputed models of cancer types, sigma values	Plot showing the Purity/Ploidy, R data file	yes	no	yes	no	yes	no	Comes bundles with HAPSEG;
Align-GVGD [92]	FASTA, substitutions list	Web report	yes	no	no	no	no	yes	Estimates SNP risk;
ANNOVAR [93]	VCF4, Complete Genomics, GFF3-SOLiD, CSV in Annovar format;	Gene-based annotation; Region-based annotations; Filter-based annotation. For all categories	yes	yes	yes	no	yes	no	Integrated tool providing gene annotation, db ids and various scores;
Ann Tools [94]	VCF, pileup, CSV	VCF	yes	yes	yes	no	yes	no	Provides a set of helper tools for custom annotation;
Auto-mute [95]	PDB ID, Chain, Mutation	Web report	yes	no	no	no	no	yes	The tool performs stability and disease potential predictions.
CandiSNPer [96]	dbSNP ID, population	Web report	yes	no	no	no	no	yes	
CHASM and SNVBox [97]	Passenger mutation rates, AA changes	CSV including CHASM score, p-value, and FDR	yes	no	no	no	yes	no	Predicts the functional significance of somatic missense mutations observed in the genomes of cancer cells and features prioritization of mutations;
CUPSAT [98]	PDB ID; PDB file format	Web report	yes	no	no	no	no	yes	Performs protein stability prediction;
dbNSFP [99]			yes	no	no	no	yes	no	Integrated SNP database; provides a simple JAVA CLI tool for searching;
VEP (Ensembl - Variant Effect Predictor) [100]	CSV, VCF, Pileup, HGVS, Variant Identifiers	Web report	yes	no	no	no	yes	yes	
ESEfinder [101]	FASTA	Web report, CSV	-	-	-	no	no	yes	Analyzes sequences for the presence of ESE motifs;
ESRSearch [102]	plain sequence; FASTA	Web report	-	-	-	no	no	yes	Finds ESR sequences;
FANS [103]	FASTA format; or variation information via web interface	Web report, CSV	yes	no	no	no	no	yes	Prioritized variations based on risk levels; divided into: Genome View, Gene View, Transcript View, Variation View;
FastSNP [104]	Gene Symbol, dbSNP ID	Web report	yes	no	no	no	no	yes	Outputs prioritized list of SNPs with risk assessment;
FESD [105]	Gene name	Web report	yes	no	no	no	no	yes	Output includes regions: promoter, CpG, islands, translation start, splice site, translation stop, poly(A) signal, transcript

Name	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
FOLD-X [106]			yes	no	no	no	yes	yes	It performs protein stability analysis.
F-SNP [107]	SNP ID; disease; gene; chromosomal region		yes	no	no				The software integrates information obtained from 16 bioinformatics tools and databases about the functional effects of SNPs.
GERP++ [108]		Web report	yes	no	no	no	yes	yes	It produces evolutionary conservation scores.
GSITIC [109]	Segmentation File, Markers File, FASTA, (Array List File, CNV File)	Lesions, Amplification Genes, Deletion Genes, Gistic Scores, Plots	no	no	yes	no	yes	no	Identifies regions of the genome that are significantly amplified or deleted across a set of samples;
HOPE [110]	FASTA, accession code for protein	Web report on structural differences between wild type and mutations	yes	no	no	no	no	yes	The web-based tool offers a simple web interface for entering protein sequence and amino acid mutation.
Human Splicing Finder (HSF) [111]	Ensembl / RefSeq ID, plain text sequences		yes	no	no	no	no	yes	
I-Mutant2.0 [112]	One letter residue code, sequence residue number	Web report	yes	no	no	no	yes	yes	The tool is based on support vector machines.
LS-SNP [113]	SwissProt ID, dbSNP ID, Kegg Pathway ID, HUGO Gene ID	Web report	yes	no	no	no	no	yes	The tool offers prediction of disease association and confidence of prediction and is based on support vector machines (SVM).
MAPP [114]	FASTA	CSV in MAPP format	yes	no	no	no	yes	no	
MuD [115]		Web report	yes	no	no	no	no	yes	
MutaGeneSys [116]		Web report / CSV	yes	no	no	no	yes	yes	The query Interface is not working.
MutationAssessor [117]	CSV in MutationAssessor format, Uniprot ID, Refseq ID	CSV in MutationAssessor format	yes	no	no	no	no	yes	
MutationTaster [118]	ORF, cDNA sequence, genomic sequence, alteration	Web report	yes	yes	no	no	yes	yes	
MutPred [119]	FASTA sequence, CSV file of mutations	Web report	yes	no	no	no	no	yes	Calculates the impact of mutation on different protein properties; is based on SIFT and offers precomputed dbSNP results;
MutSig [120]	List of mutations, regions to investigate	CSV	yes	yes	no	no	yes	no	Still in beta testing – available upon request
NGS-SNP [121]	VCF, pileup, CSV	VCF	yes	no	no	no	yes	no	
nsSNPAnalyzer [122]	FASTA, substitutions list	Web report	yes	no	no	no	no	yes	The tool outputs various SNP features and predicts the phenotypic class
Oncotator [123]	Oncotator format	CSV	yes	yes	no	no	no	yes	Annotations with data relevant to cancer researcher; collects Genomic Annotations, Protein Annotations, Cancer Annotations

Name	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
PANTHER [124]	Protein sequence and Substitution	subPSEC score	yes	no	no	no	yes	yes	Uses subPSEC score;
Parepro [125]	Protein sequence and Substitution	-	yes	no	no	no	yes	no	It is based on support vector machines (SVM).
PESX [126]	plain sequence; FASTA	Web report	-	-	-	no	no	yes	Finds ESE sequences;
pfSNP [127]	SNP ID; chromosome region; Gene ID;	Web report	yes	no	no	no	no	yes	
PHAST [128]	FASTA, PHYLIP, MPM, MAF, SS	Conservation score	-	-	-	no	yes	no	Phylogenetic analysis toolbox, including phastCons and phyloP;
PhD-SNP [129]	One letter residue code, Swiss-Prot protein code, Sequence file	Effect prediction	yes	no	no	no	yes	no	
PMUT [130]	FASTA sequence/file SWISSProt code	Web report	yes	no	no	no	no	yes	Offers different prediction modes and is able to output detailed mutation analysis reports;
PolyDoms [131]	Gene/protein symbol(s), RefSeqID dbSNP ID	Web report	yes	no	no	no	no	yes	
PolyMAPr [132]	-	-	-	-	-	no	yes	no	No longer available;
PolyPhen-2 [133]	UniProt ID, FASTA, dbSNP ID	CSV in PolyPhen format	yes	no	no	no	yes	yes	
PupaSNP Finder [134]	dbSNP ID, Gene/Transcript ID; PED format	Web report	yes	no	no	no	no	yes	
QuickSNP [135]	genomic position; HUGO gene symbol	Web report	yes	no	no	no	no	yes	
RescueESE [136]	plain text; multi-FASTA	predicts sequences with ESE activity	-	-	-	no	no	yes	
SAPRED [137]	FASTA and mutation file		yes	no	no				The website is offline.
SCAN [138]		Web report	yes	no	yes	no	no	yes	
SCONE [139]	MAF	Conservation score	-	-	-	no	yes	no	
SeattleSeq Annotation [140]	Maq, GFFm CASAVA, VCF, GATK bed	VCF, own format	yes	yes	no	no	no	yes	
SeqAnt [141]	FASTA sequence file	Web report	yes	yes	no	no	no	yes	
SeqProfCod [142]	-	-	yes	no	no	-	-	-	Not available online;
SVA (Sequence Variant Analyser) [143]	VCF of variants, project file (for command line version)	--potential biological function --dbSNP/Kegg/GO/1000 Genomes/DGV annotation --identifies protein-truncating variants --filtering by function	yes	yes	no	yes	yes	no	

Name	Input Format	Output Format	SNP	INDEL	CNV	GUI	CLI	Web	Notes
SIFT [144]	Multiple proteins, dbSNP ID, NCBI GI number, protein sequence, protein sequence alignment, Pileup, VCF4, maq, soap, gff3, casava, cg	XXX in SIFT format	yes	no	no	no	yes	yes	
SIFT Indel [145]			no	yes	no	no	no	yes	
SiPhy [146]	FASTA, MAE, PHYLIP		-	-	-	no	yes	no	
SNAP [147]	AA in FASTA, substitutions format	Web report	yes	no	no	no	no	yes	This tool offers a user friendly web interface.
SNP Function Portal [148]	RefSNP Ids, OMIM Ids	Web report	yes	no	no	no	no	yes	
SNP@Domain [149]			yes	no	no	-	-	-	Not available anymore;
SNPdbe [150]	Gene/protein symbol, FASTA	Web report	yes	no	no	no	no	yes	The protein function is predicted using SNAP and SIFT and entries are augmented with experimental information from public databases.
SNPeffect 4.0 [151]	FASTA, PDB file, PDB ID, UniProt ID	Web report	yes	no	no	no	no	yes	This tool mainly uses protein structure information.
SNPHunter [152]	Gene symbol; dbSNP ID;	Web report	yes	no	no	yes	no	no	
SNPnexus [153]	CSV in SNPnexus input format	CSV in SNPnexus output format	yes	yes	yes	no	no	yes	Outputs CNV, INDELs, inversions;
SNPper [154]	dbSNP ID, TSC ID, position	Web report	yes	no	no	no	no	yes	
SNPs&GO [155]	One letter residue code; Swiss-Prot protein code; Sequence file; GO terms; CSV	Web report	yes	no	no	no	no	yes	Predicts neutral/deleterious; calculates reliability index and disease probability;
SNPs3D [156]	Gene symbol, SNP ID	Web report	yes	no	no	no	no	yes	
SNPseek [157]	-	-	-	-	-	-	-	-	Tool that performs neural network based protein stability prediction which is not available anymore;
SNPselector [158]	-	-	-	-	-	no	no	yes	No longer available;
SnpsSIFT + snpEff [159]	VCF, SNPs, insertions, deletions, and MNP s	CSV	yes	yes	no	no	yes	no	A collection of tools to manipulate VCF files;
SPOT [160]	SNPs and p-values,	Web report	yes	no	no	no	no	yes	Outputs various DB ids and scores;
StSNP [161]	protein sequence; protein name; dbSNP ID; gene symbol	Web report	yes	no	no	no	no	yes	
TAMAL [162]	-	-	-	-	-	-	-	-	No longer available;
TopoSNP [163]	Protein ID, protein sequence	Web report	yes	no	no	no	no	yes	Predicts whether substitution is on surface of the protein structure; conservation score based on Pfam

Variant Effect Predictor (VEP)



The screenshot shows the Ensembl VEP website. At the top is the Ensembl logo and navigation links: BLAST/BLAT, BioMart, Tools, Downloads, Help & Documentation, Blog, and Mirrors. A search bar on the right says "Search all species...". Below the navigation bar are tabs: "Using this website", "Annotation and prediction", "Data access", "API & software", and "About us". The "API & software" tab is active, showing a breadcrumb trail: Home > Help & Documentation > API & Software > Ensembl Tools > Variant Effect Predictor. On the left is a sidebar with a tree view under "In this section" containing links for "Web interface", "VEP script", and "Data formats". The main content area is titled "Variant Effect Predictor" and contains a description of the tool's purpose. Below the description is a list of features. To the right of the main content are three panels: "Web interface" with a globe icon, "Standalone perl script" with a terminal icon, and "REST API" with the "re!st" logo. Each panel lists its features and provides links to documentation and download pages.

Variant Effect Predictor

The 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!](#)

[Launch Ve!P](#)

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

Web interface

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

[Documentation](#)
[Launch the web interface](#)

Standalone perl script

- More options, more flexibility
- For large volumes of data

[Documentation](#)
[Download latest version](#)

REST API

- Language-independent API
- Simple URL-based queries
- GET single variants, POST many

[Documentation](#)

Provides annotations for different types of alterations referred to different genomic locations

With several ways of execution

If you use the VEP, please cite our UPDATED publication so we can continue to support VEP development:

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F.
The Ensembl Variant Effect Predictor.
Genome Biology Jun 6;17(1):122. (2016)
[doi:10.1186/s13059-016-0974-4](https://doi.org/10.1186/s13059-016-0974-4)

Standalone execution

```
perl variant_effect_predictor.pl --format vcf --sift b --polyphen b --ccds --uniprot --hgvs --symbol --numbers --domains --  
regulatory --canonical --protein --biotype --uniprot --tsl --gmaf --variant_class --xref_refseq --maf_1kg --maf_esp --  
maf_exac --dir /home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/.vep -  
i /home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf --config  
/home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/registry.local --  
output_file / /home/epineiro/analysis/pancancer/genotypes/0a6be23a-d5a0-4e95-ada2-  
a61b2b5d9485.vcf_output_VEP.txt --force_overwrite --vcf --no_progress --plugin  
Condel,/home/epineiro/analysis/pancancer/vep/ensembl-tools-release-  
85/scripts/variant_effect_predictor/.vep/Plugins/config/Condel/config,b --fork 8 --offline
```

Configuration options

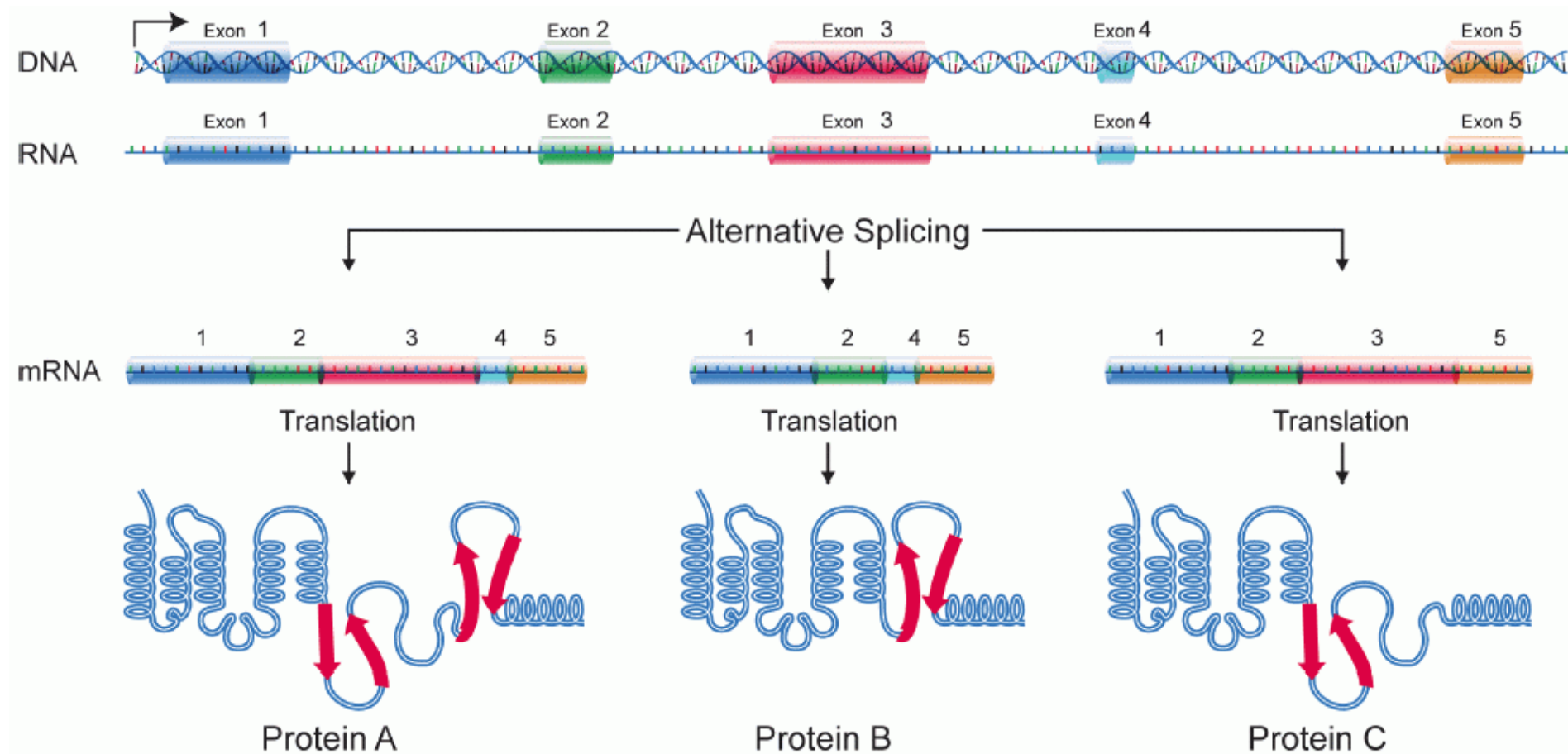
Input, output and format

Annotations

Plugins

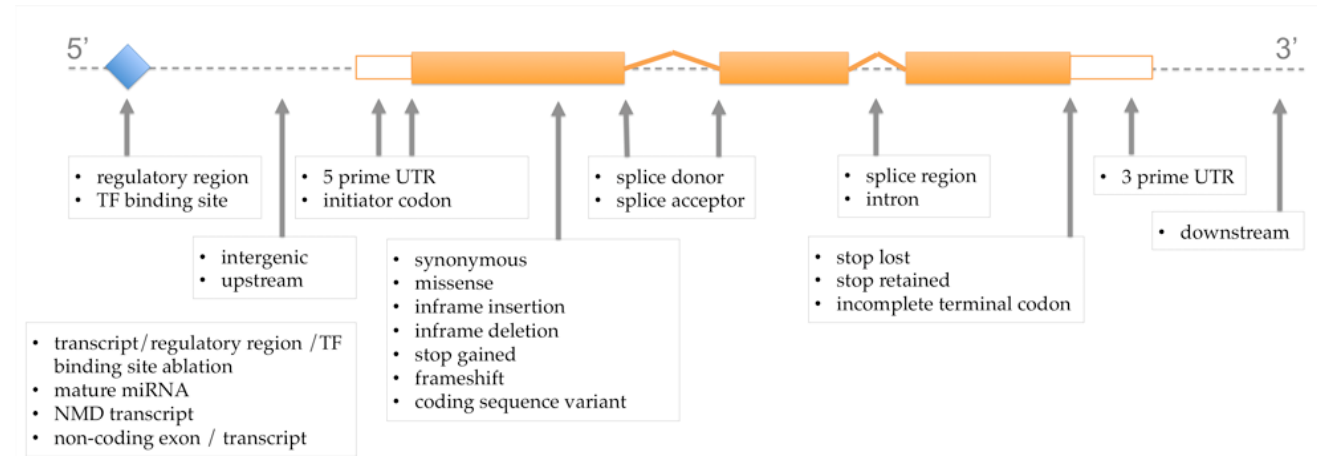
http://www.ensembl.org/info/docs/tools/vep/script/vep_options.html#basic

Annotations provided for each affected transcript



Annotations about consequence: Transcriptional

Sequence Ontology



* SO term	SO description	SO accession	Display term	IMPACT
transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequenc	SO:0001821	Inframe insertion	MODERATE
inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	SO:0001822	Inframe deletion	MODERATE
missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE

H
I
G
H

High

Moderate

Low

Modifier

M
O
D
E
R
A
T
E

Annotations about consequence: Functional impact prediction

SIFT PREDICTION (`--sift [p|s|b]`) predicts whether an amino acid substitution affects protein function based on sequence homology and the physical properties of amino acids. The VEP can output the **p**rediction term, **s**core or **b**oth

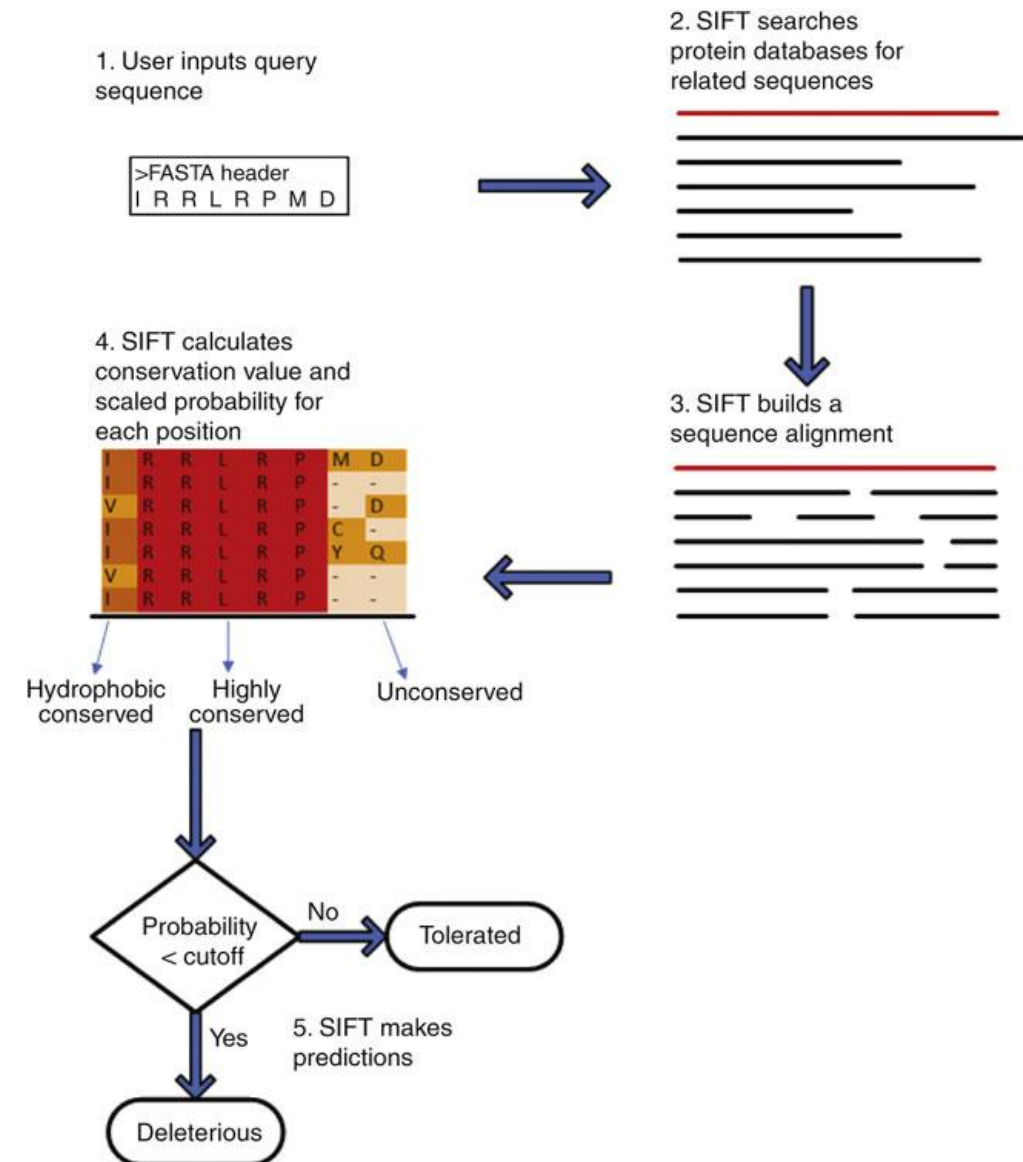
Deleterious – Tolerated

< 0.05 Deleterious

PolyPhen PREDICTION (`--polyphen [p|s|b]`) predicts the possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations. The VEP can output the **p**rediction term, **s**core or **b**oth.

Probably damaging - Possibly damaging - Benign – Unknown

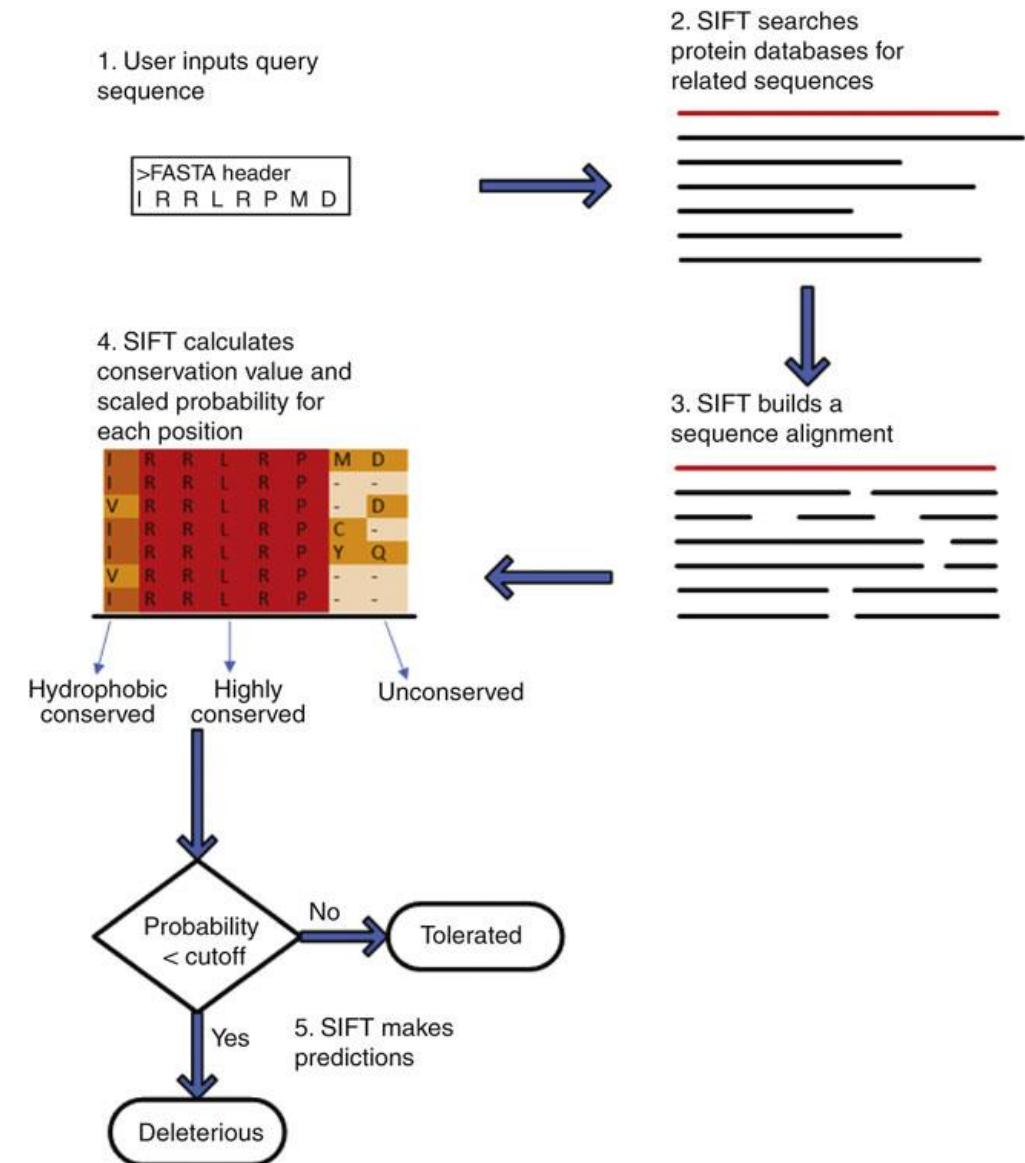
Benign < ~0.435 < Damaging



Annotations about consequence: Functional impact prediction

The built-in functionality in VEP of these predictors allows only the prediction of coding non-synonymous SNV

CONDEL (`--plugin`) Condel **CONsensus DELeteriousness score** runs an adapted version of CONDEL computing a score based on pre-calculated SIFT and PolyPhen-2 scores from the Ensembl API.



Impact prediction – Other predictors

- **dbNSFP**

functional predictions and annotations for human nonsynonymous single-nucleotide variants and splice-site variants

- **LoFtool**

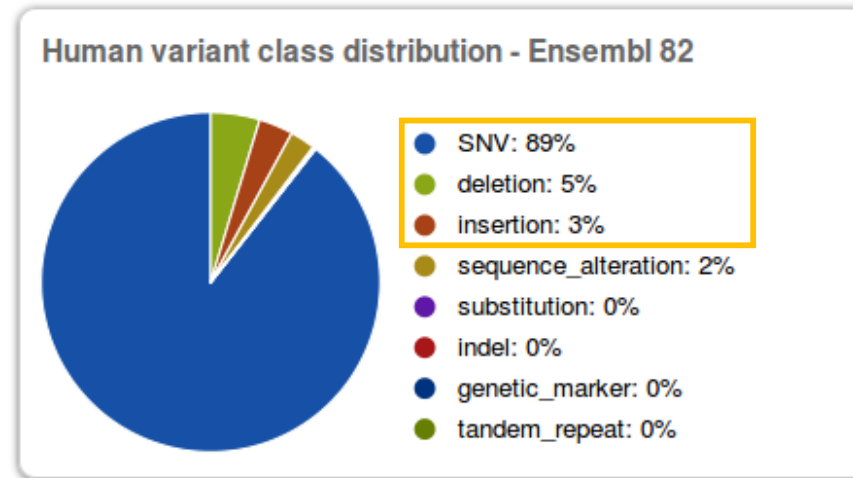
gene intolerance ranking system, based on the ratio of Loss-of-function (LoF) to synonymous mutations for each gene from ExAC data

- **PROVEAN**

Functional prediction for nonsynonymous mutations or indels

Variant level annotations

VARIANT CLASS (`--variant_class`): Sequence Ontology variant class



HGVS NOMENCLATURE (`--hgvs`): HGVS nomenclature additional to Ensembl identifier. Both coding and protein sequence. e.g. ENST00000263967.3:c.3140A>G

GMAF (`--gmaf`): *Global minor allele frequency* (MAF) from 1000 Genomes Phase 1. e.g. G:0.4036

ExAC Browser (Beta) | Exome Aggregation Consortium

Search for a gene or variant or region

Examples - Gene: [PCSK9](#), Transcript: [ENST00000407236](#), Variant: [22-46615880-T-C](#), Multi-allelic variant: [rs1800234](#), Region: [22:46615715-46615880](#)

About ExAC

The [Exome Aggregation Consortium](#) (ExAC) is a coalition of investigators seeking to aggregate and harmonize exome sequencing data from a wide variety of large-scale sequencing projects, and to make summary data available for the wider scientific community.

The data set provided on this website spans 60,706 unrelated individuals sequenced as part of various disease-specific and population genetic studies. The ExAC Principal Investigators and groups that have contributed data to the current release are listed [here](#).

All data here are released under a [Fort Lauderdale Agreement](#) for the benefit of the wider biomedical community - see the terms of use [here](#).

Sign up for our mailing list for future release announcements [here](#).

Recent News

March 14, 2016

- Version 0.3.1 ExAC data and browser (beta) is released! ([Release notes](#))

January 13, 2015

- Version 0.3 ExAC data and browser (beta) is released! ([Release notes](#))

October 29, 2014

- Version 0.2 ExAC data and browser (beta) is released! Sign up for our mailing list for future release announcements [here](#).

October 20, 2014

- Public release of ExAC Browser (beta) at ASHG!

October 15, 2014

- Internal release to consortium now available!

ExAC (--maf_exac)

Variant: 22:46615880 T / C

Note: This variant is multiallelic! The other alt alleles are:

- [22-46615880-T-A](#)

Filter Status PASS
dbSNP [rs1800234](#)
Allele Frequency 0.009613
Allele Count 1163 / 120986
UCSC [22-46615880-T-C](#) [↗](#)
ClinVar [Click to search for variant in Clinvar](#) [↗](#)

Genotype Quality Metrics

Site Quality Metrics

Annotations

This variant falls on 7 transcripts in 1 genes:

missense

- [PPARA](#)

Transcripts ▾

intron

- [PPARA - ENST00000434345](#)

non coding transcript exon

- [PPARA - ENST00000493286](#)

Note: This list may not include additional transcripts in the same gene that the variant does not overlap.

Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Latino	649	11522	28	0.05633
East Asian	361	8618	8	0.04189
Other	10	904	0	0.01106
European (Finnish)	22	6606	1	0.00333
South Asian	42	16440	2	0.002555
European (Non-Finnish)	73	66602	0	0.001096
African	6	10294	0	0.0005829
Total	1163	120986	39	0.009613

Gene and genetic region

GENE SYMBOL (**--symbol**) e.g. MYC

Standard nomenclature for the human genes

EXON/INTRON NUMBER (**--numbers**) The format is Number/Total.

OVERLAPPING REGULATORY REGIONS (**--regulatory**) Script also detect if the variant falls in a **transcription factor binding site**. Output lines have a Feature type of **RegulatoryFeature** or **MotifFeature**.

BIOTYPE (**--biotype**) Adds the biotype of the transcript or regulatory feature. e.g. protein_coding, processed_pseudogene...

Transcript annotation

CCDS (**--ccds**) Adds the CCDS transcript identifier. e.g. CCDS1639.1
CCDS: Consensual CoDing Sequence

TRANSCRIPT SUPPORT LEVEL (**--tsl**)

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

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

PRINCIPAL ISOFORM (**--canonical**) Adds a flag indicating if the transcript is the canonical transcript for the gene.

1. Longest CCDS with no stop codons
2. Longest Ensembl/Havana merged translation with no stop codons
3. Longest translation with no stop codons
4. Longest non-protein-coding transcript (if no translation)

Principal Isoform - APPRIS

<http://appris.bioinfo.cnio.es/#/>

{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 CDS as the principal isoform for each gene.

APPRIS Database

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

[Access the web database](#)

APPRIS WebServer

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

[Run the web server](#)

APPRIS WebServices

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

[Go to the API interface](#)

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



Human

Assemblies: GRCh38

Assemblies: GRCh37



Mouse

Assemblies: GRCm38



Zebrafish

Assemblies: GRCz10

Assemblies: Zv9



Rat

Assemblies: Rnor_6.0

Assemblies: Rnor_5.0



Pig

Assemblies: Sscrofa10.2



Chimpanzee

Assemblies: CHIMP2.1.4

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



Fruitfly

Assemblies: BDGP6



C.elegans

Assemblies: WBcel235

Principal Isoform - APPRIS

{APPRIS}

Annotating principal splice isoforms

Executes several computational methods for the transcript annotation.

As part of the annotation process, it selects a CDS as the principal isoform for each gene.

APPRIS Database

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

[Access the web database](#)

APPRIS WebServer

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

[Run the web server](#)

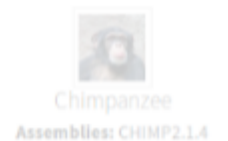
APPRIS WebServices

Access annotations for the species annotated in the database via gene name or Ensembl id. Access queries through RESTful web services.

ISOFORM RANGE

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

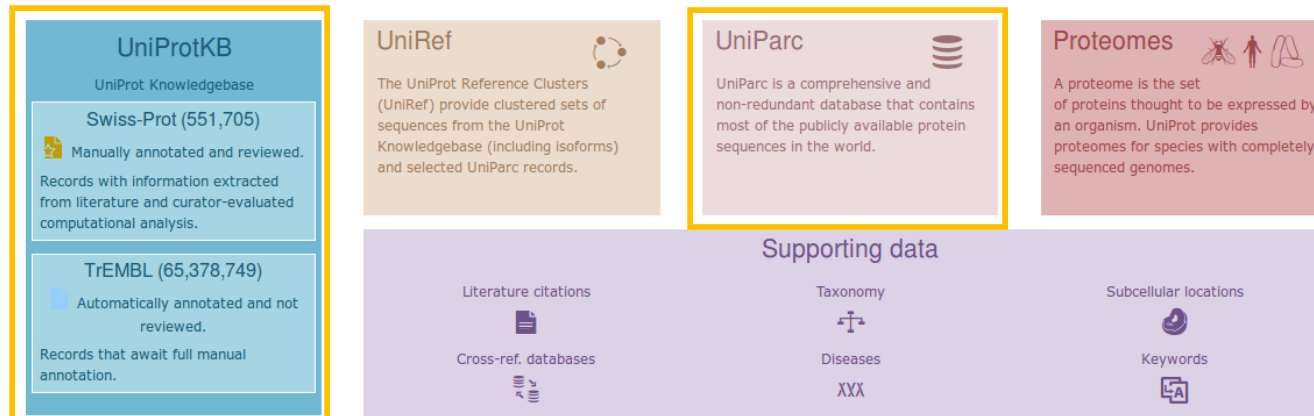
ALTERNATIVE:1
ALTERNATIVE:2



Protein annotation

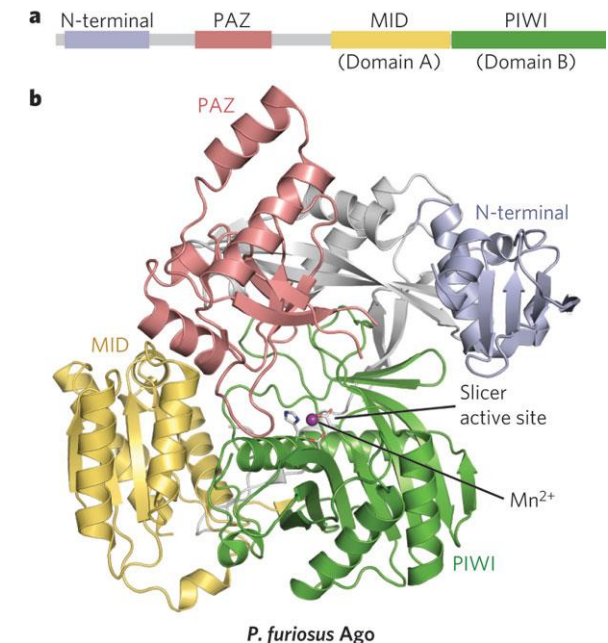
PROTEIN(--protein) Ensembl protein identifier e.g. ENSP00000470877

UNIPROT IDENTIFIERS (--uniprot) Protein identifiers in UniProt databases (SWISSPROT, TREMBL y UniParc) e.g. SH3Y1_HUMAN (SWISSPROT), C9J4Z8_HUMAN (TREMBL), UPI0000208A67 (UniParc)



DOMAINS (--domains) Protein overlapping domains

Pfam, Prosite, InterPro
e.g. Pfam_domain:PF00071



VEP standalone script output



Links

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

VEP run statistics

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

General statistics

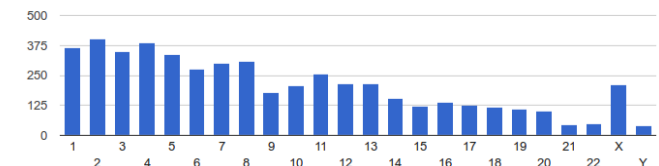
Lines of input read	5025
Variants processed	5013
Variants remaining after filtering	5013
Lines of output written	5013
Novel / existing variants	4273 (85.2%) / 740 (14.8%)
Overlapped genes	2761
Overlapped transcripts	10274
Overlapped regulatory features	490

Variant classes



● SNV

Variants by chromosome



VEP standalone script output

[illegible]

Allele
Consequence
IMPACT
SYMBOL (--symbol)
Gene
Feature_type (--regulatory)
Feature
BIOTYPE (--biotype)
EXON (--numbers)
INTRON (--numbers)


HGVSc (--hgvs)
HGVSp (--hgvs)
cDNA_position
CDS_position
Protein_position
Amino_acids
Codons
Existing_variation
DISTANCE
STRAND

FLAGS
 VARIANT_CLASS (--variant_class)
 SYMBOL_SOURCE
 HGNC_ID
 CANONICAL (--canonical)
 TSL (--tsl)
 CCDS (--ccds)
 ENSP (--protein)
 SWISSPROT (--uniprot)
 TREMBL (--uniprot)

UNIPARC (--uniprot)
RefSeq (--xref_refseq)
SIFT (--sift)
PolyPhen (--polyphen)
DOMAINS (--domains)
HGVS_OFFSET
GMAF (--gmaf)
*_MAF (--maf_1kg)
AA_MAF (--maf_esp)
EA_MAF (--maf_esp)

ExAC_MAF (--maf_exac)
ExAC_*_MAF (--maf_exac)
CLIN_SIG
SOMATIC
PHENO
MOTIF_NAME
MOTIF_POS
HIGH_INF_POS
MOTIF_SCORE_CHANGE
Condel (--condel)

web execution

 BLAST/BLAT | BioMart | Tools | Downloads | Help & Documentation | Blog | Mirrors

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 the script
 - Caches and databases
 - Filtering results
 - Custom annotations
 - Plugins
 - Examples and use cases
 - Other information
- Data formats
- FAQ


Search documentation...


Help & Documentation | API & Software | Ensembl Tools | **Variant Effect Predictor**

Variant Effect Predictor

The 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!](#)







Web interface

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


 [Documentation](#)


 [Launch the web interface](#)




Standalone perl script

- More options, more flexibility
- For large volumes of data


 [Documentation](#)

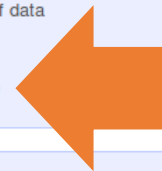
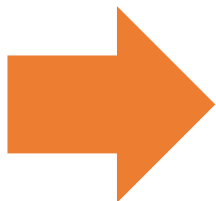
 [Download latest version](#)



REST API

- Language-independent API
- Simple URL-based queries
- GET single variants, POST many

 [Documentation](#)



If you use the VEP, please cite our UPDATED publication so we can continue to support VEP development:

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F.
The Ensembl Variant Effect Predictor.
Genome Biology Jun 6;17(1):122. (2016)
[doi:10.1186/s13059-016-0974-4](https://doi.org/10.1186/s13059-016-0974-4)

web execution

Variant Effect Predictor

 VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Species:

 Human (Homo sapiens) 

Assembly: GRCh38.p7

Name for this job (optional):

Either paste data:

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

Or upload file:

No file selected.

Or provide file URL:

Transcript database to use:

☒ Ensembl transcripts

☐ Gencode basic transcripts

☐ RefSeq transcripts

☐ Ensembl and RefSeq transcripts

hg19

hg38

Input
options

web execution

Identifiers and frequency data  Additional identifiers for genes, transcripts and variants; frequency data

Identifiers

Gene symbol:



CCDS:



Protein:



Uniprot:



HGVS:



CSN^(p):



Unshifted HGVS^(p):



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



ExAC allele frequencies

PubMed IDs for citations of co-located variants:



Include flagged variants:



(p) = functionality from [VEP plugin](#)

web execution

Extra options  e.g. SIFT, PolyPhen and regulatory data

Miscellaneous

Transcript biotype:	<input checked="" type="checkbox"/>	protein coding, pseudogene, processed pseudogene, miRNA, rRNA, scRNA, snoRNA and snRNA
Protein domains:	<input type="checkbox"/>	
Exon and intron numbers:	<input type="checkbox"/>	
Transcript support level:	<input checked="" type="checkbox"/>	
APPRIS:	<input checked="" type="checkbox"/>	
Identify canonical transcripts:	<input type="checkbox"/>	
miRNA structure ^(p) :	<input type="checkbox"/>	
Upstream/Downstream distance ^(p) :	<input checked="" type="radio"/> Disabled <input type="radio"/> Enabled	

Pathogenicity predictions

SIFT:	Prediction and score
PolyPhen:	Prediction and score
Condel ^(p) :	<input checked="" type="radio"/> Disabled <input type="radio"/> Enabled
LoFtool ^(p) :	<input type="checkbox"/>

web execution

Regulatory data

Get regulatory region consequences:

Yes

Splicing predictions

MaxEntScan^(p):


☐

Conservation

BLOSUM62^(p):

☐

(p) = functionality from [VEP plugin](#)

Filtering options  Pre-filter results by frequency or consequence type

Filters

Filter by frequency:

- ☒ No filtering
- ☐ Exclude common variants
- ☐ Advanced filtering

Return results for variants in coding regions only:

☐

Restrict results:

Show all results

NB: Restricting results may exclude biologically important data!

Run >

[Clear](#)



web execution

Show/hide columns (1 hidden)		Filter	
Analysis	Jobs	Submitted at	
Variant Effect Predictor	 VEP analysis of pasted data in Homo_sapiens Queued	09/08/2016, 16:00 (BST)	   

Show/hide columns (1 hidden)		Filter	
Analysis	Jobs	Submitted at	
Variant Effect Predictor	 VEP analysis of pasted data in Homo_sapiens Done View results	09/08/2016, 16:00 (BST)	   

web execution

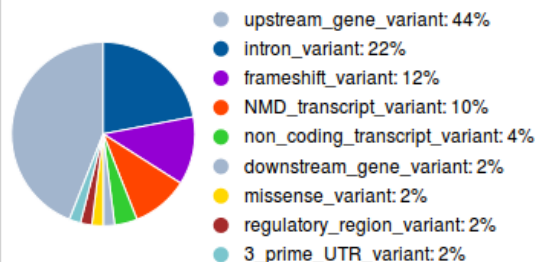
Variant Effect Predictor results

Job details

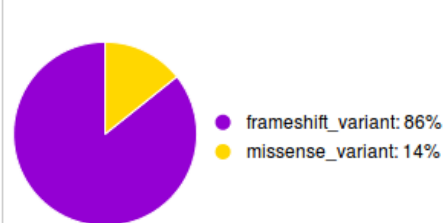
Summary statistics

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

Consequences (all)



Coding consequences



Summary

Results preview

Filtering

Navigation
Page: 1 of 1 | Show: 1 All variants

Filters
Uploaded variant is defined

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	1:818046-818046	C	missense_variant	MODERATE	AL645608.2	ENSG00000269308	Transcript	ENST00000594233	protein_coding	1/3	-	-	-	4	4	2
2_265023_C/A	2:265023-265023	A	intron_variant	MODIFIER	ACP1	ENSG00000143727	Transcript	ENST00000272065	protein_coding	-	1/5	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	intron_variant	MODIFIER	ACP1	ENSG00000143727	Transcript	ENST00000272067	protein_coding	-	1/5	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	upstream_gene_variant	MODIFIER	SH3YL1	ENSG00000035115	Transcript	ENST00000356150	protein_coding	-	-	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	upstream_gene_variant	MODIFIER	SH3YL1	ENSG00000035115	Transcript	ENST00000402632	protein_coding	-	-	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	upstream_gene_variant	MODIFIER	SH3YL1	ENSG00000035115	Transcript	ENST00000403657	protein_coding	-	-	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	upstream_gene_variant	MODIFIER	SH3YL1	ENSG00000035115	Transcript	ENST00000403658	protein_coding	-	-	-	-	-	-	-
2_265023_C/A	2:265023-265023	A	upstream_gene_variant	MODIFIER	SH3YL1	ENSG00000035115	Transcript	ENST00000403712	protein_coding	-	-	-	-	-	-	-

THE END

Fátima Al-Shahrour
falshahrour@cnio.es

Javier Perales-Patón
jperales@cnio.es

Elena Piñeiro-Yáñez
epineiro@cnio.es