Precision medicine: NGS variant analysis and interpretation for translational research

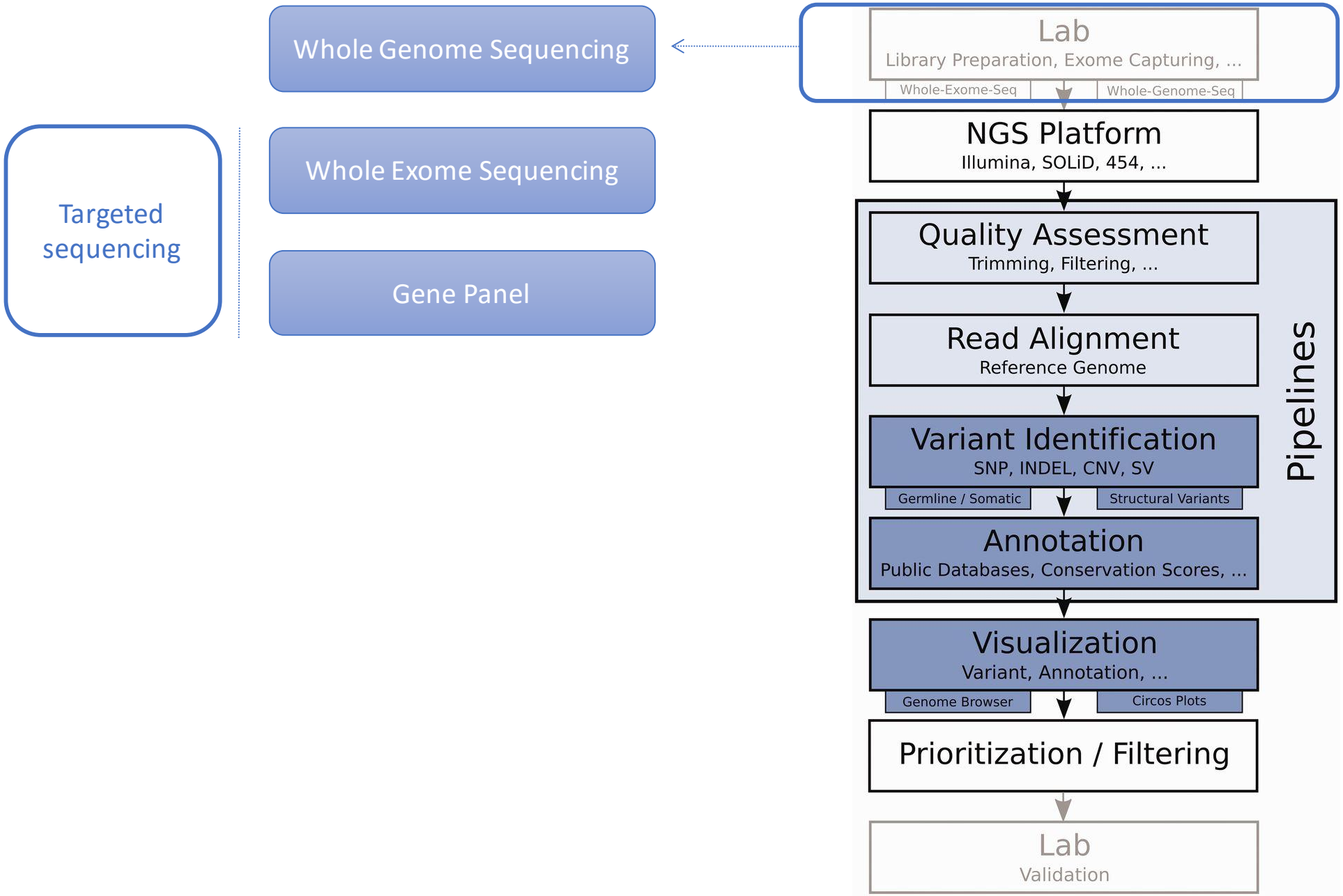
NGSII: Variant annotation

Fátima Al-Shahrour ● Javier Perales ● Elena Piñeiro

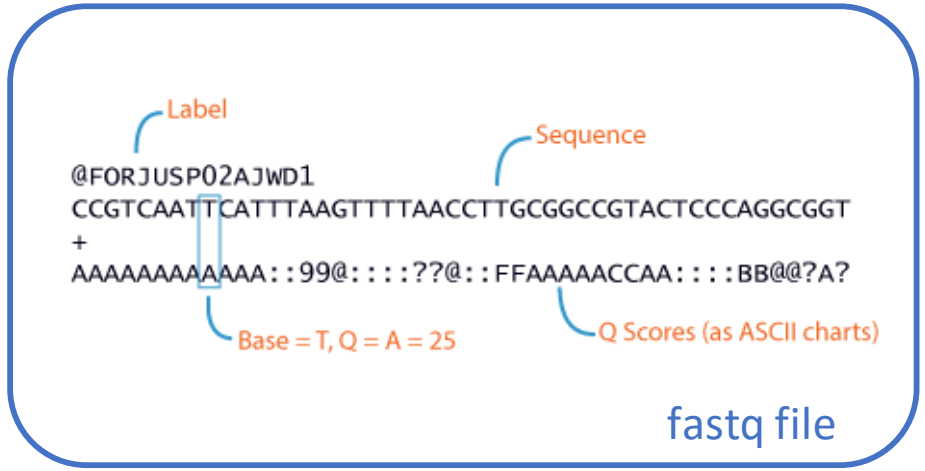
November 15, 2017

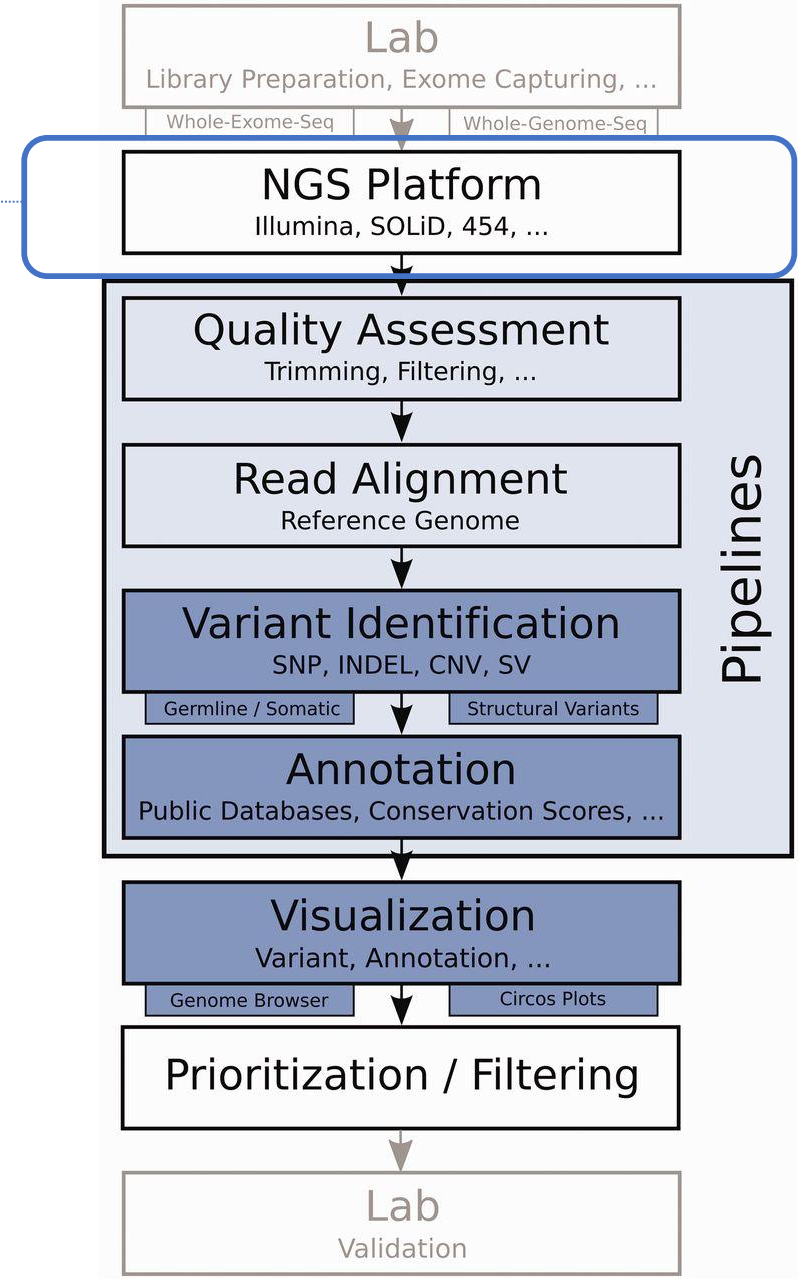
Centro Nacional de Investigaciones Oncológicas

INSTITUTO GULBENKIAN DE CIÊNCIA

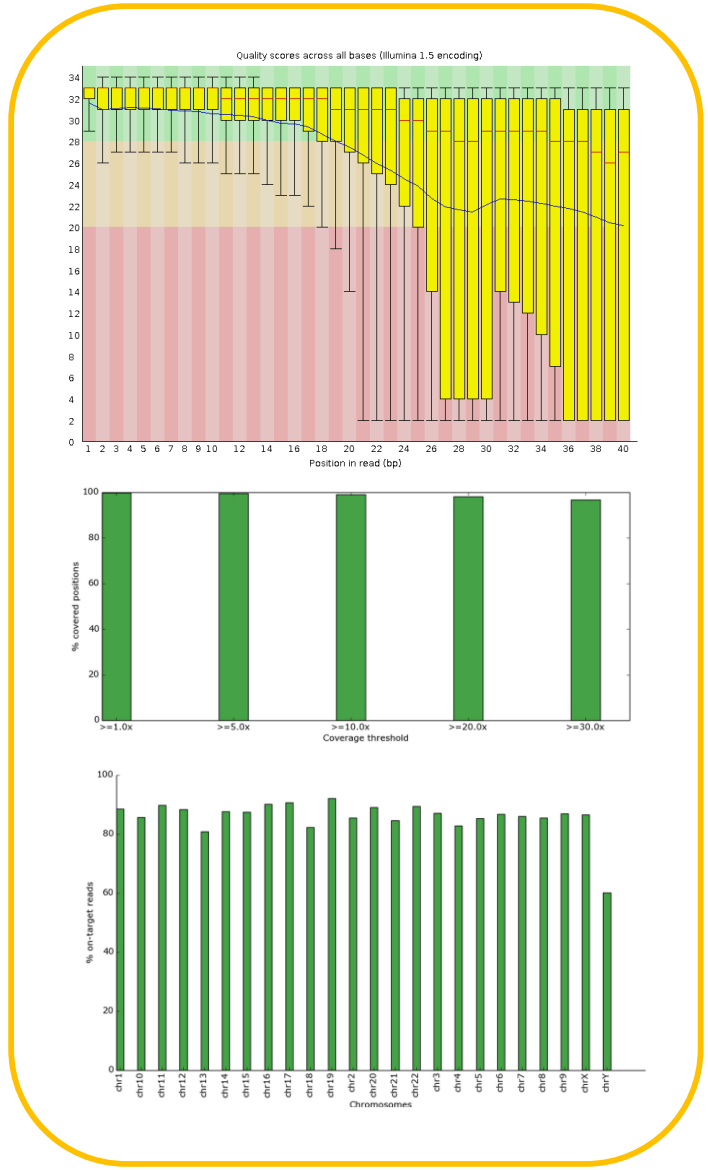


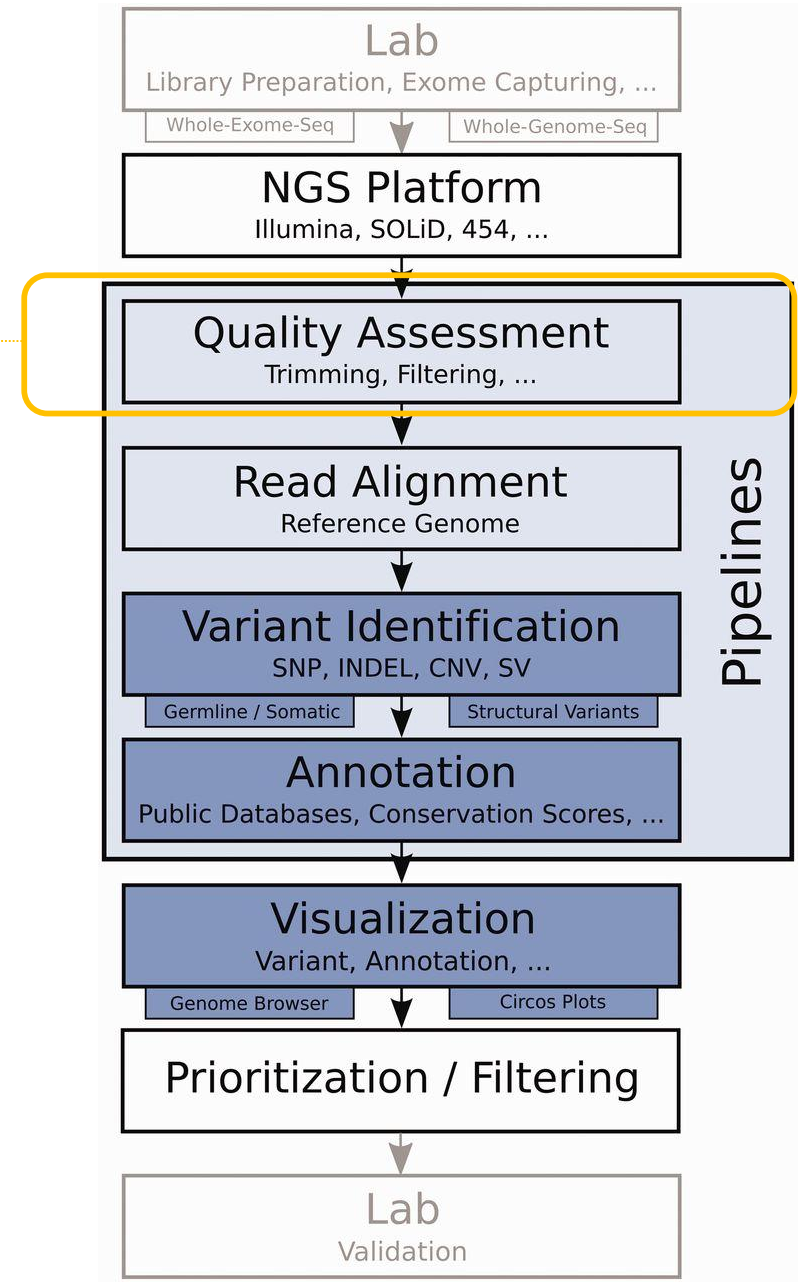
Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086



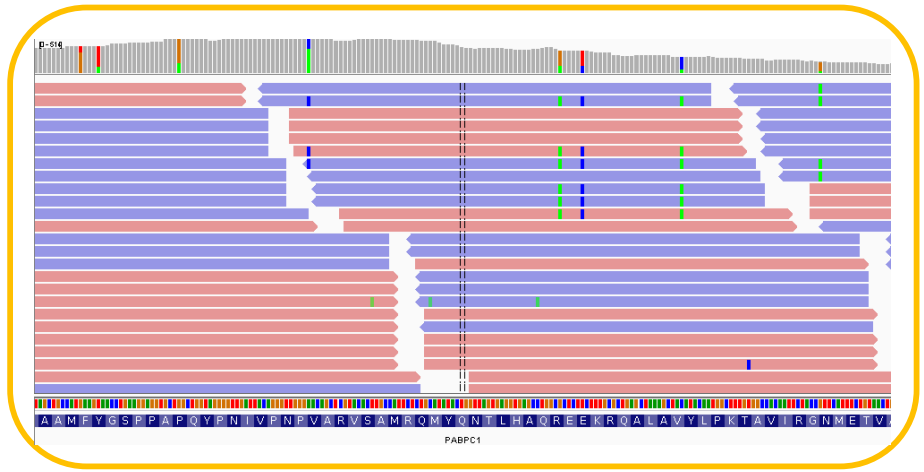


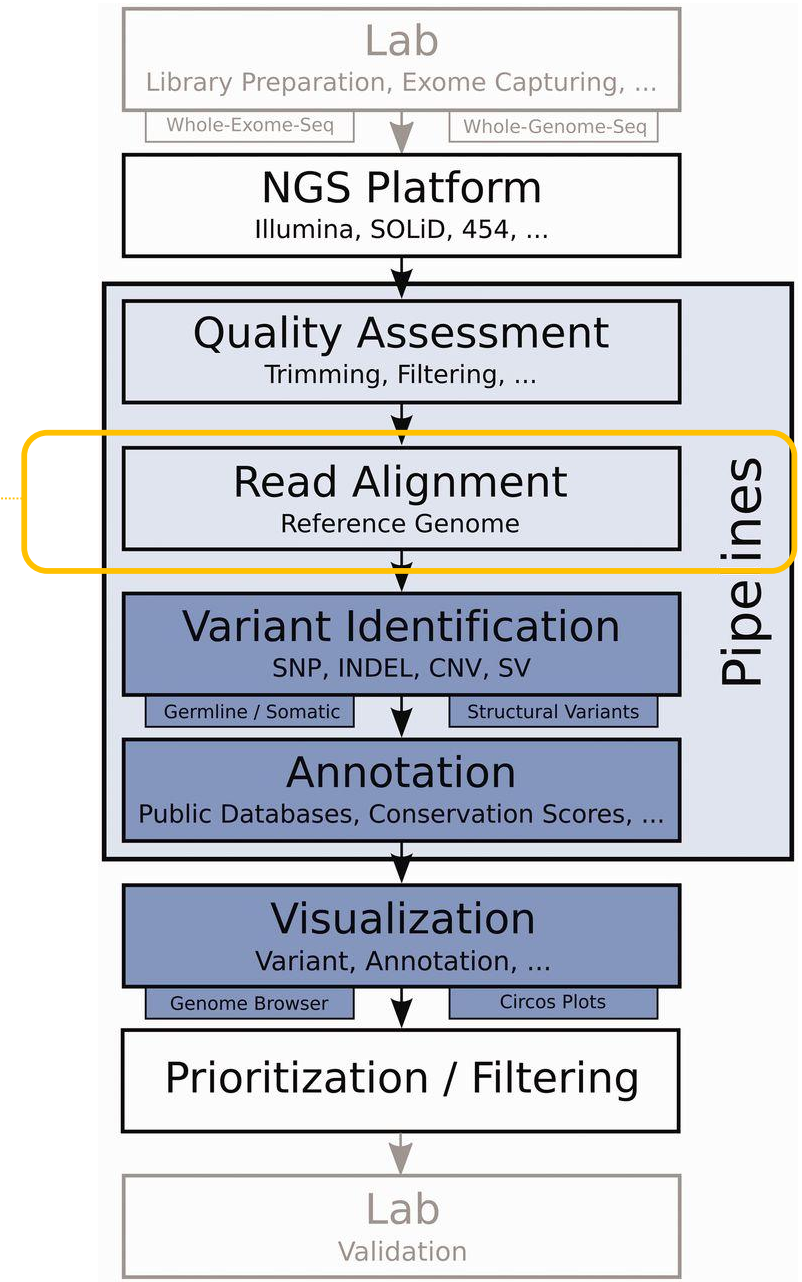
Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086





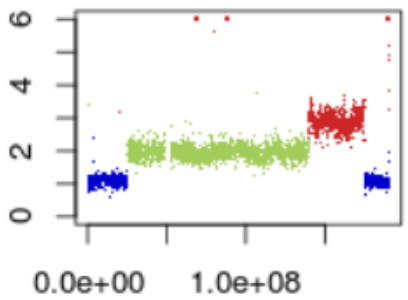
Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086



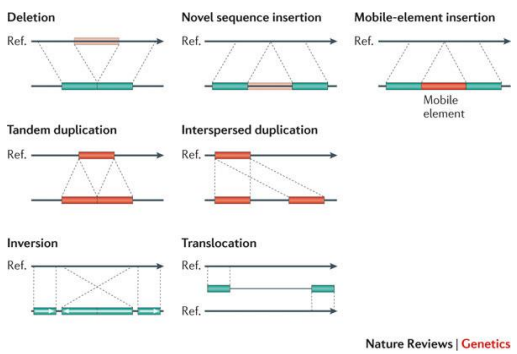


Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Copy Number Variants (CNV)



Structural Variation



Single Nucleotide Variant (SNV)

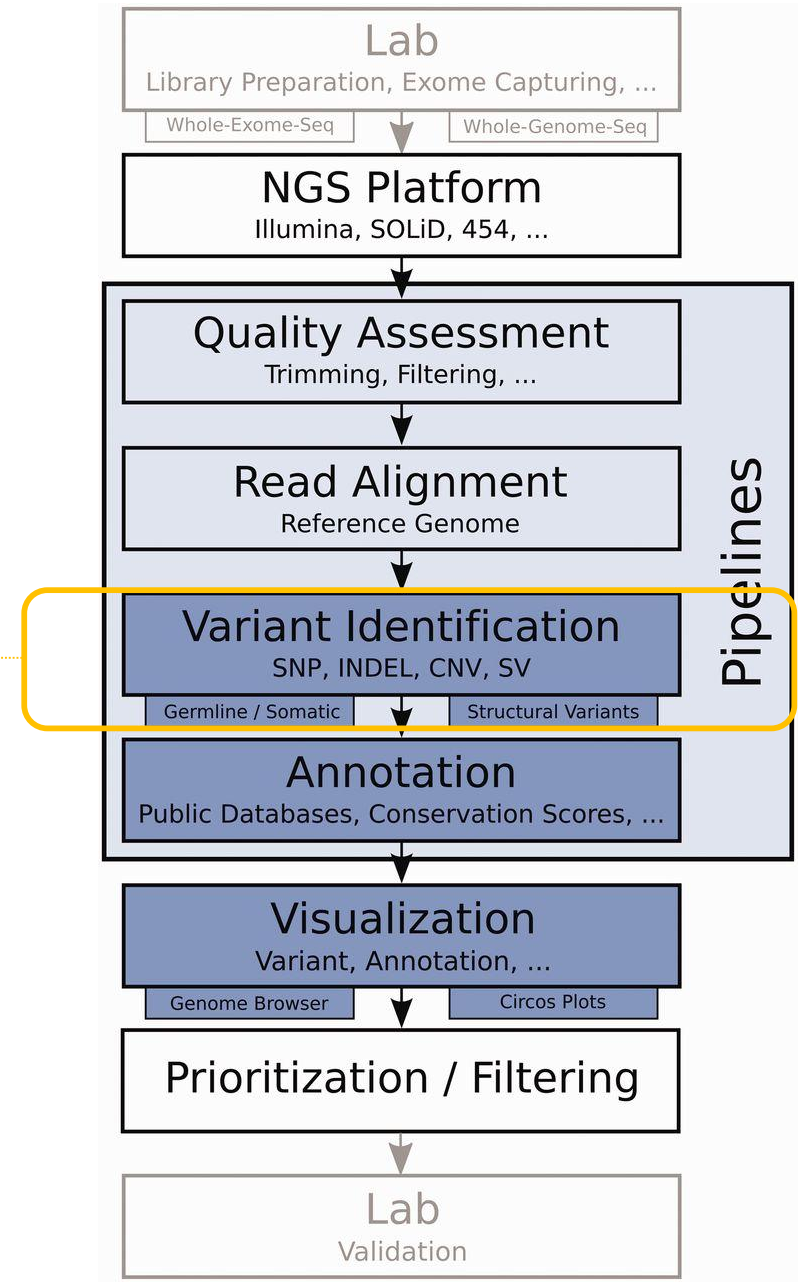
Reference: AACGGCCTGTAAC
Alternative: AACGGCCAGTAAC

Insertion

Reference: AACGGCCTGTAAC
Alternative: AACGGCCAGCTAAC

Deletion

Reference: AACGGCCTGTAAC
Alternative: AACGGCC-GTAAC



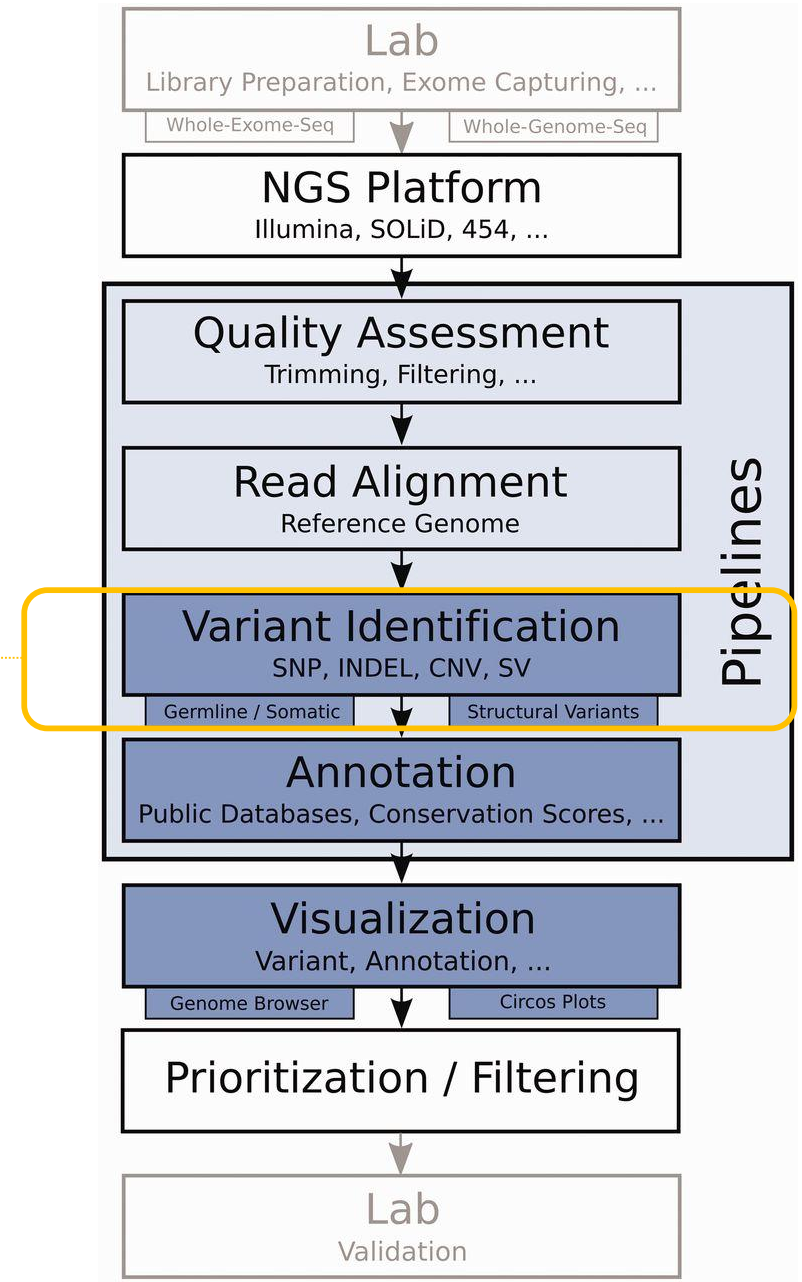
Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

Somatic

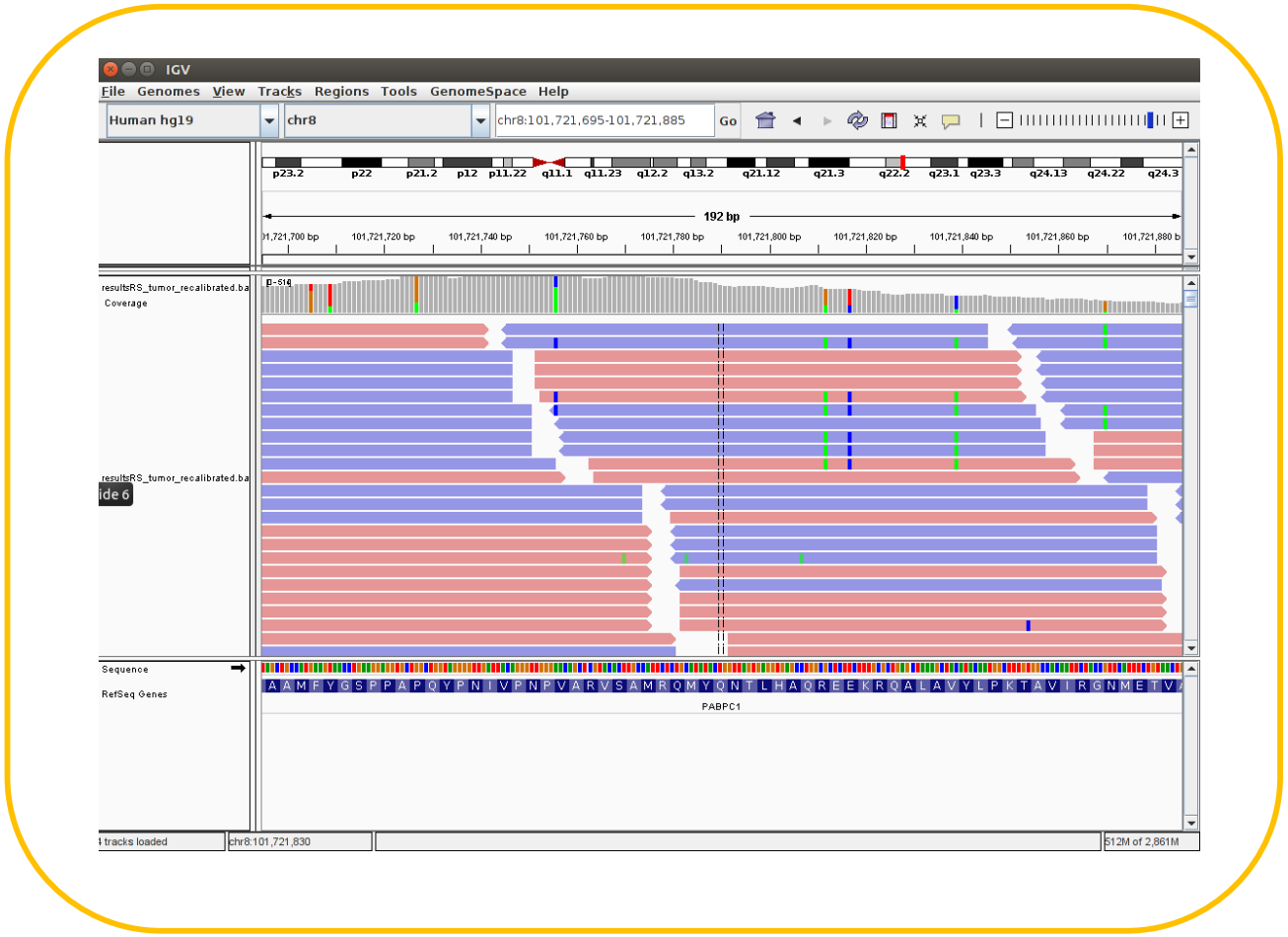
Occur in non-germline tissues
Are not inherited

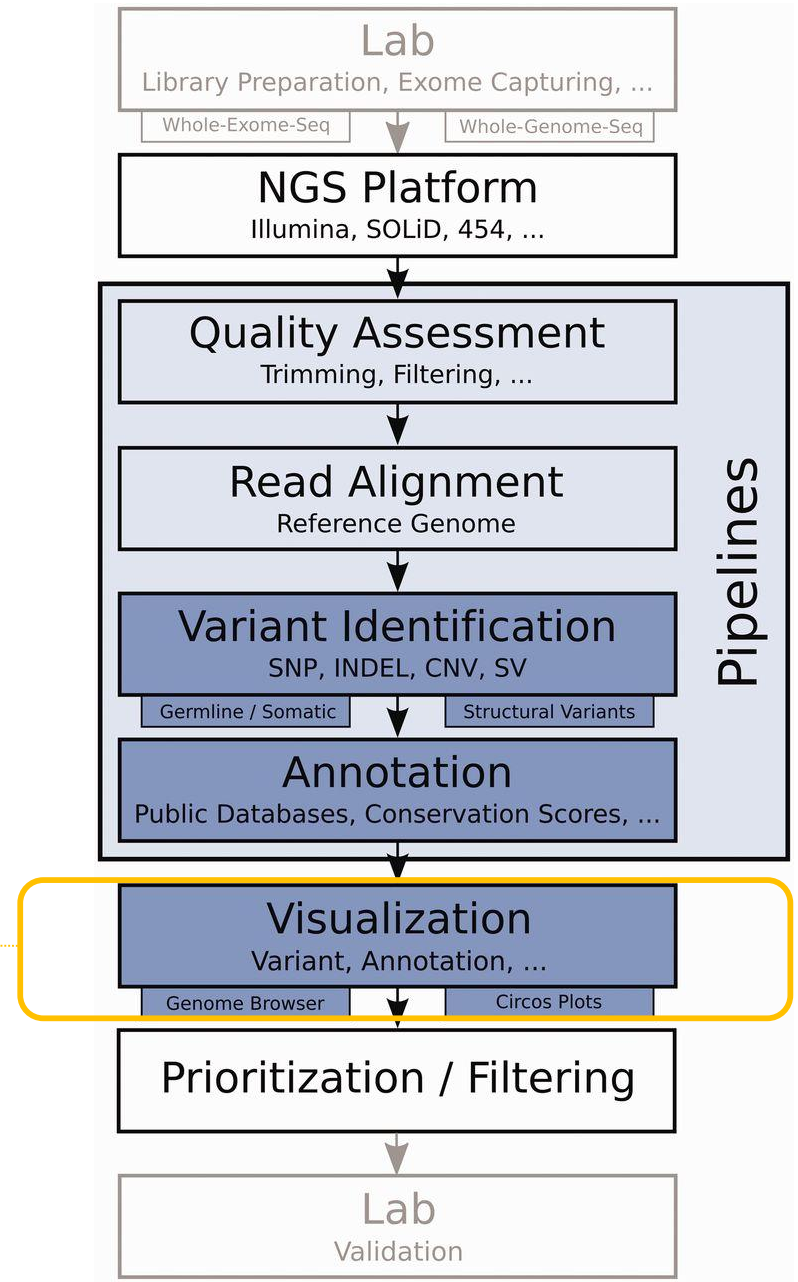
Germline

Present in germline cells
Can be inherited



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

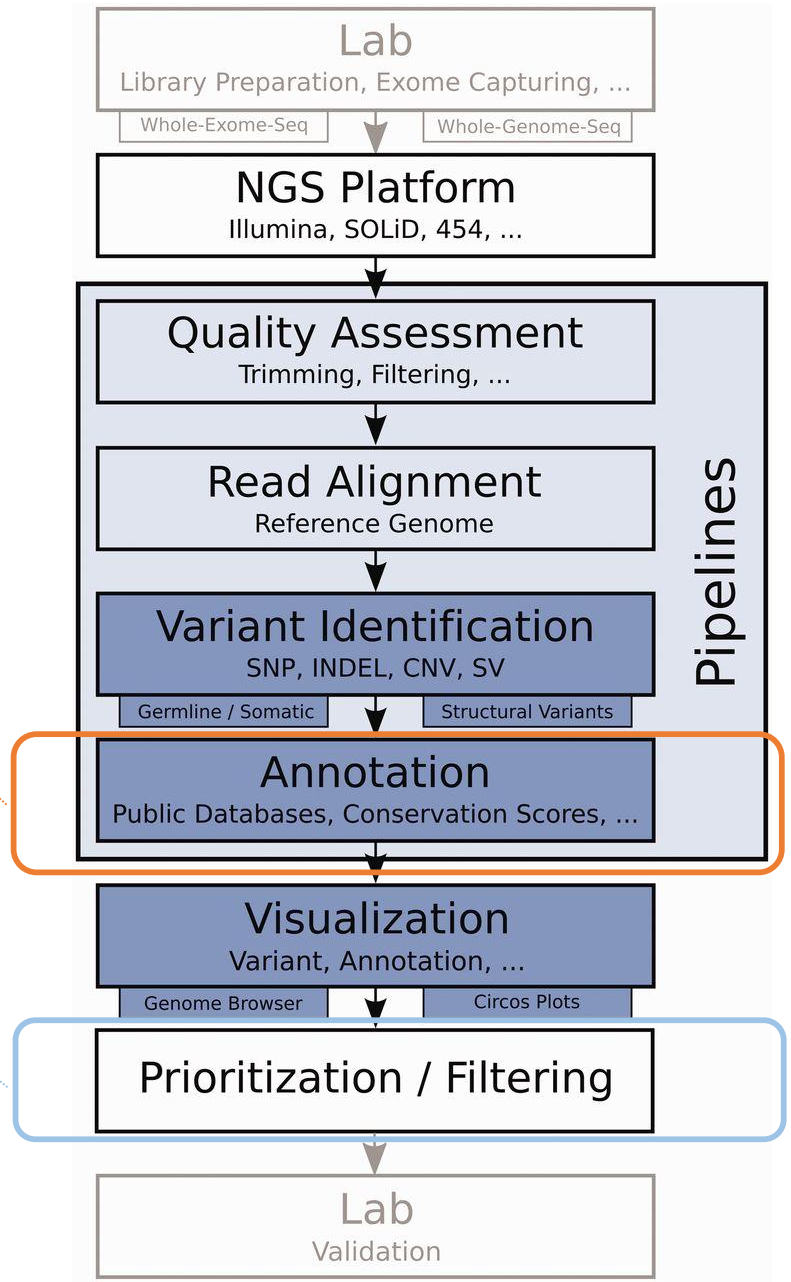




Stephan Pabinger et al. Brief Bioinform 2013;bib.bbs086

NGSII: Variant Annotation

Selecting the most relevant variants: How to filter



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

Variant annotation

- Technical information: quality parameters, filters, …
  - Variant callers provide technical parameters associated to each variant, that allow us to remove sequencing artifacts and select the most reliable variants.
- Descriptive information: nomenclature, genotype, …
  - Variant callers and variant annotators
- Functional annotation: transcriptional consequence, functional prediction, …
  - Variant annotators provide useful information to select the biologically relevant variants

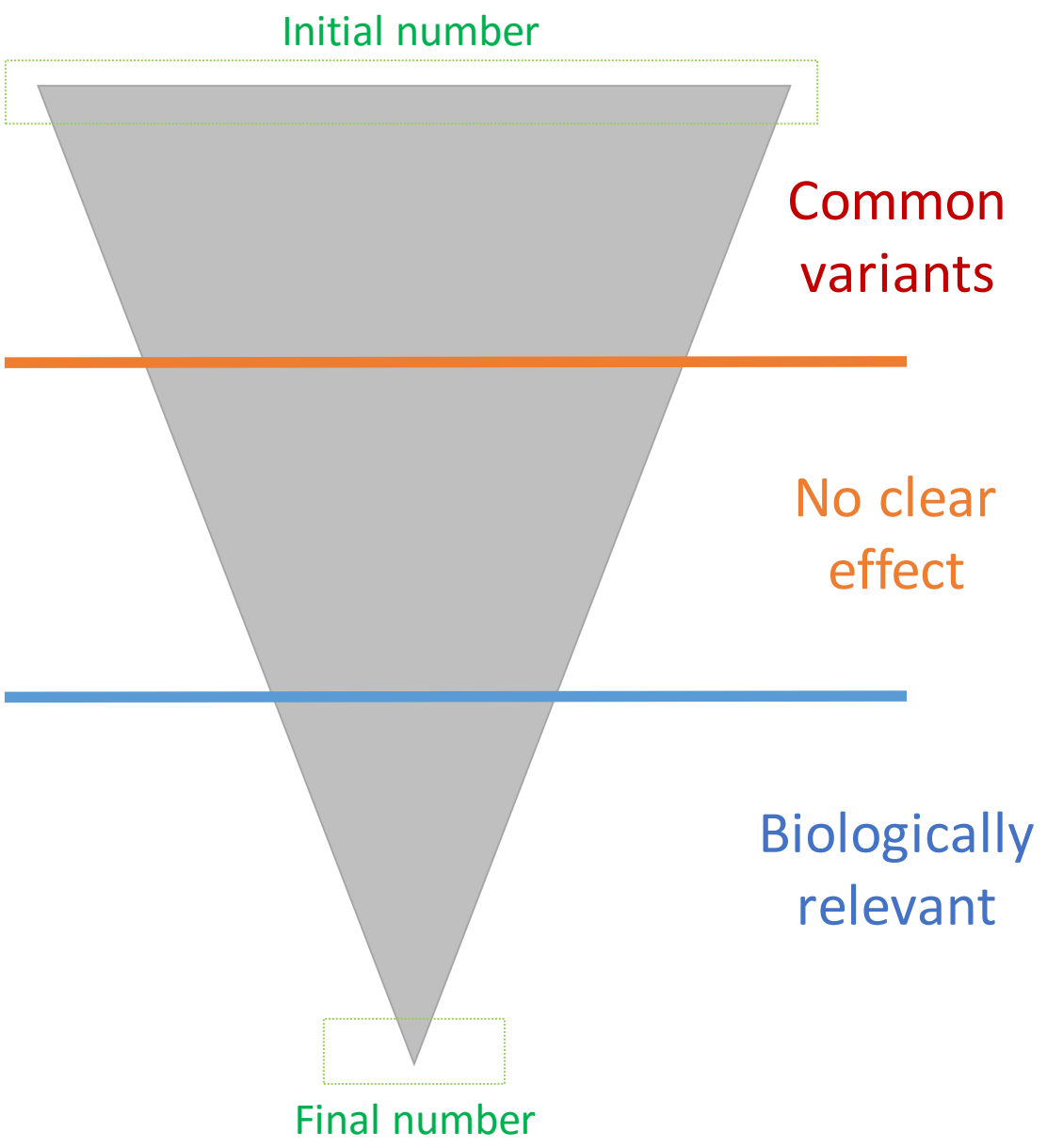
Variant annotation

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

- They are mainly polymorphisms (maybe predisposition or drug response): at least 1% in population

- Variants without a clear effect (synonym variants, benign...)

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;

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

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

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

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 preditction	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	

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

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, MAF, 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;
SnpSIFT + snpEff [159]	VCF, SNPs, insertions, deletions, and MNPs	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

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

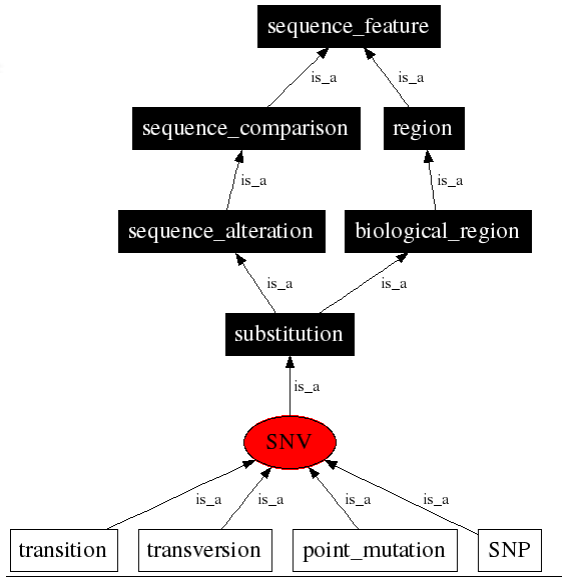
Variant annotations

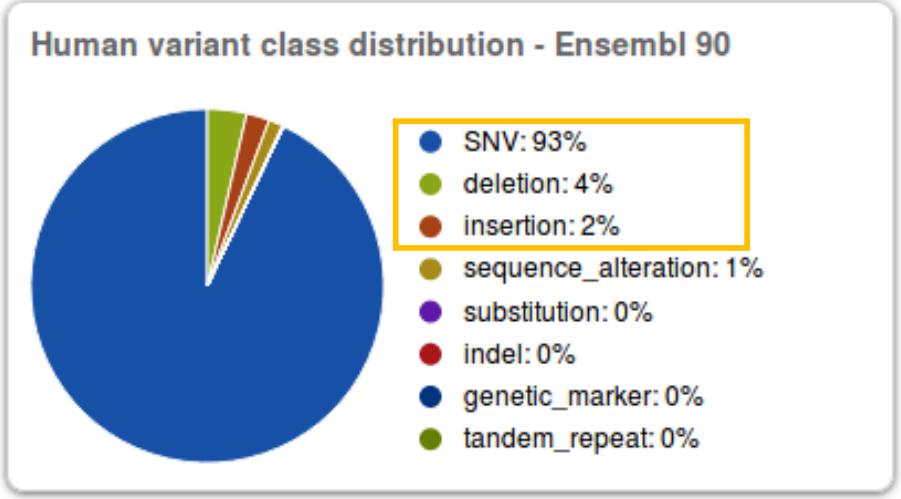
NOMENCLATURE

- Chromosomal coordinates e.g. 1 12854414 G/A (PRAMEF1)
- HGVS nomenclature
  - For coding sequence e.g. c.638G>A
  - For protein sequence e.g. p.Arg213His

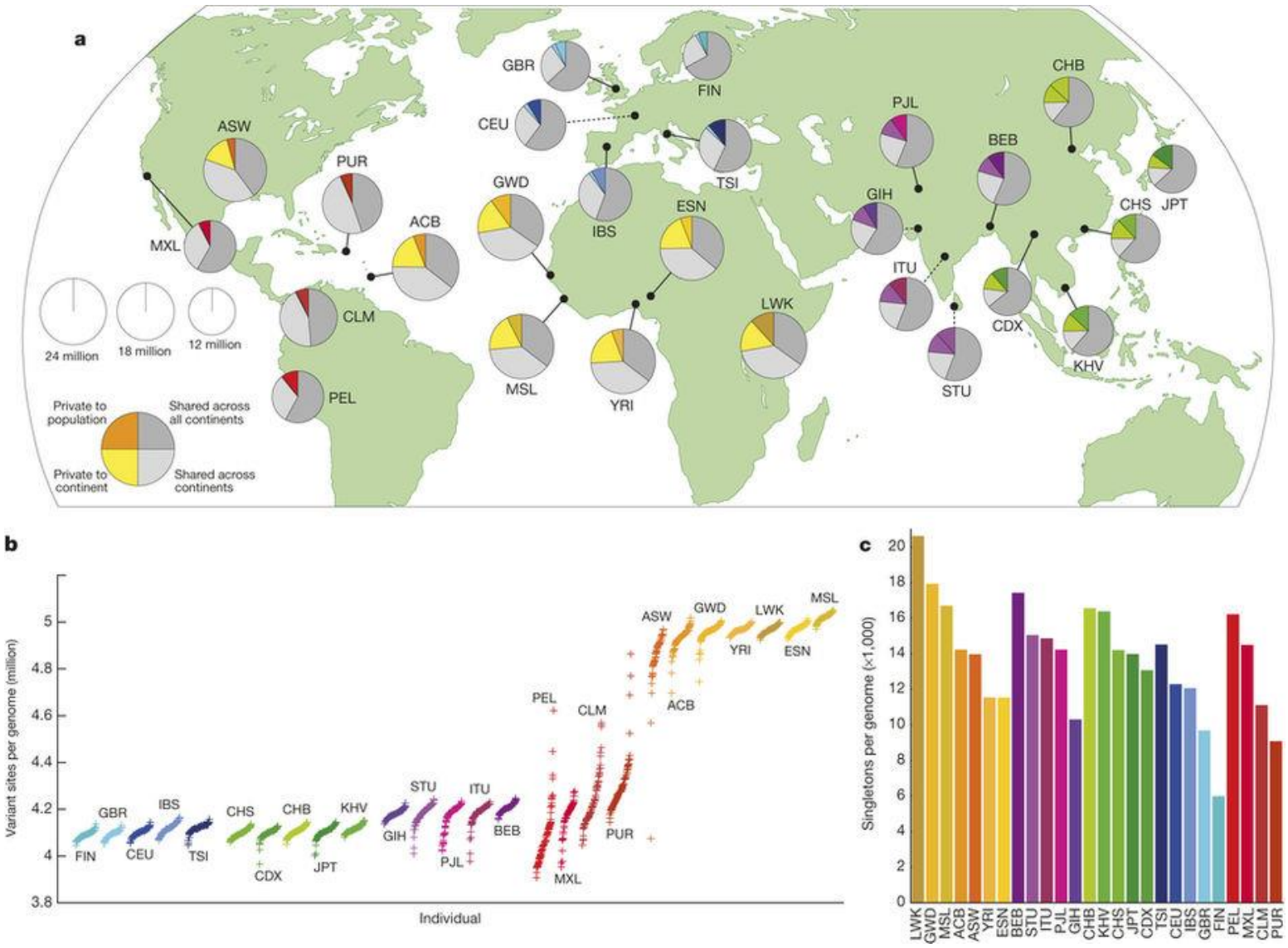
VARIANT CLASS

The Sequence Ontology





POPULATION FREQUENCIES



http://www.internationalgenome.org/

IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

Phase III:
2504 individuals
26 populations

POPULATION FREQUENCIES

http://exac.broadinstitute.org/

ExAC Browser Beta | About | Downloads | Terms | Contact | Jobs | FAQ

Interested in working on the development of this resource? Apply here.

ExAC Browser (Beta) | Exome Aggregation Consortium

Contributing projects

- 1000 Genomes
- Bulgarian Trios
- Finland-United States Investigation of NIDDM Genetics (FUSION)
- GoT2D
- Inflammatory Bowel Disease
- METabolic Syndrome In Men (METSIM)
- Jackson Heart Study
- Myocardial Infarction Genetics Consortium:
  - Italian Atherosclerosis, Thrombosis, and Vascular Biology Working Group
  - Ottawa Genomics Heart Study
  - Pakistan Risk of Myocardial Infarction Study (PROMIS)
  - Precocious Coronary Artery Disease Study (PROCARDIS)
  - Registre Gironi del COR (REGICOR)
- NHLBI-GO Exome Sequencing Project (ESP)
- National Institute of Mental Health (NIMH) Controls
- SIGMA-T2D
- Sequencing in Suomi (SISu)
- Swedish Schizophrenia & Bipolar Studies
- T2D-GENES
- Schizophrenia Trios from Taiwan
- The Cancer Genome Atlas (TCGA)
- Tourette Syndrome Association International Consortium for Genomics (TSAICG)

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

August 8, 2016

- CNV calls are now available on the ExAC browser

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!

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
Filtering AF	0.05274 (Latino)
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

non coding transcript exon

- PPARA - ENST00000493286

intron

- PPARA - ENST00000434345

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**

POPULATION FREQUENCIES

http://gnomad.broadinstitute.org/

gnomAD browser beta

About Downloads Terms Contact Jobs FAQ

Interested in working on the development of this resource? Apply here.

Contributing projects

1000 Genomes
1958 Birth Cohort
ALSGEN
Alzheimer's Disease Sequencing Project (ADSP)
Atrial Fibrillation Genetics Consortium (AFGen)
Estonian Genome Center, University of Tartu (EGCUT)
Bulgarian Trios
Finland-United States Investigation of NIDDM Genetics (FUSION)
Finnish Twin Cohort Study
FINN-ADGEN
FINRISK
Framingham Heart Study
Génome Québec - Genizon Biobank
Genomic Psychiatry Cohort
GoT2D
Genotype-Tissue Expression Project (GTEx)
Health2000
Inflammatory Bowel Disease:
- Helsinki University Hospital Finland
- NIDDK IBD Genetics Consortium
- Quebec IBD Genetics Consortium

Jackson Heart Study
Kuopio Alzheimer Study
LifeLines Cohort
MESTA
METabolic Syndrome In Men (METSIM)
Finnish Migraine Study
Myocardial Infarction Genetics Consortium (MIGen):
- Leicester Exome Seq
- North German MI Study
- Ottawa Genomics Heart Study
- Pakistan Risk of Myocardial Infarction Study (PROMIS)
- Precocious Coronary Artery Disease Study (PROCARDIS)
- Registre Gironi del COR (REGICOR)
- South German MI Study
- Variation in Recovery: Role of Gender on Outcomes of Young AMI Patients (VIRGO)

National Institute of Mental Health (NIMH) Controls
NHLBI-GO Exome Sequencing Project (ESP)
NHLBI TOPMed
Schizophrenia Trios from Taiwan
Sequencing Initiative Suomi (SiSu)
SIGMA-T2D
Swedish Schizophrenia & Bipolar Studies
T2D-GENES
- GoDARTS

T2D-SEARCH
The Cancer Genome Atlas (TCGA)

gnomAD browser beta | genome Aggregation Database

Search for a gene or variant or region

Example - Gene: PCSK9, Variant: 1-55516888-G-GA

About gnomAD

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

The data set provided on this website spans 123,136 exome sequences and 15,496 whole-genome sequences from unrelated individuals sequenced as part of various disease-specific and population genetic studies. The gnomAD Principal Investigators and groups that have contributed data to the current release are listed here.

All data here are released for the benefit of the wider biomedical community, without restriction on use - see the terms of use here.

Sign up for our mailing list for future release announcements here.

Recent News

October 3, 2017

gnomAD r2.0.2 released. Sample composition is identical to the previous release (r2.0.1), however we have made a change to the variant filtering process that you can read about here.

February 27, 2017

Official gnomAD release (version 2.0) with browser updates and data available for download.

October 19, 2016

Public release of gnomAD Browser (beta) at ASHG!

Variant: 22:46615880 T / C

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

- 22-46615880-T-A

	Exomes	Genomes	Total
Filter	Pass	Pass	
Allele Count	2726	140	2866
Allele Number	246166	30970	277136
Allele Frequency	0.01107	0.004521	0.01034
dbSNP	rs1800234		
UCSC	22-46615880-T-C		
ClinVar	Click to search for variant in Clinvar		

Genotype Quality Metrics

Site Quality Metrics

Report this variant

Annotations

This variant falls on 7 transcripts in 1 genes:

missense

- PPARA Transcripts

non coding transcript exon

- PPARA - ENST00000493286

intron

- PPARA - ENST00000434345

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

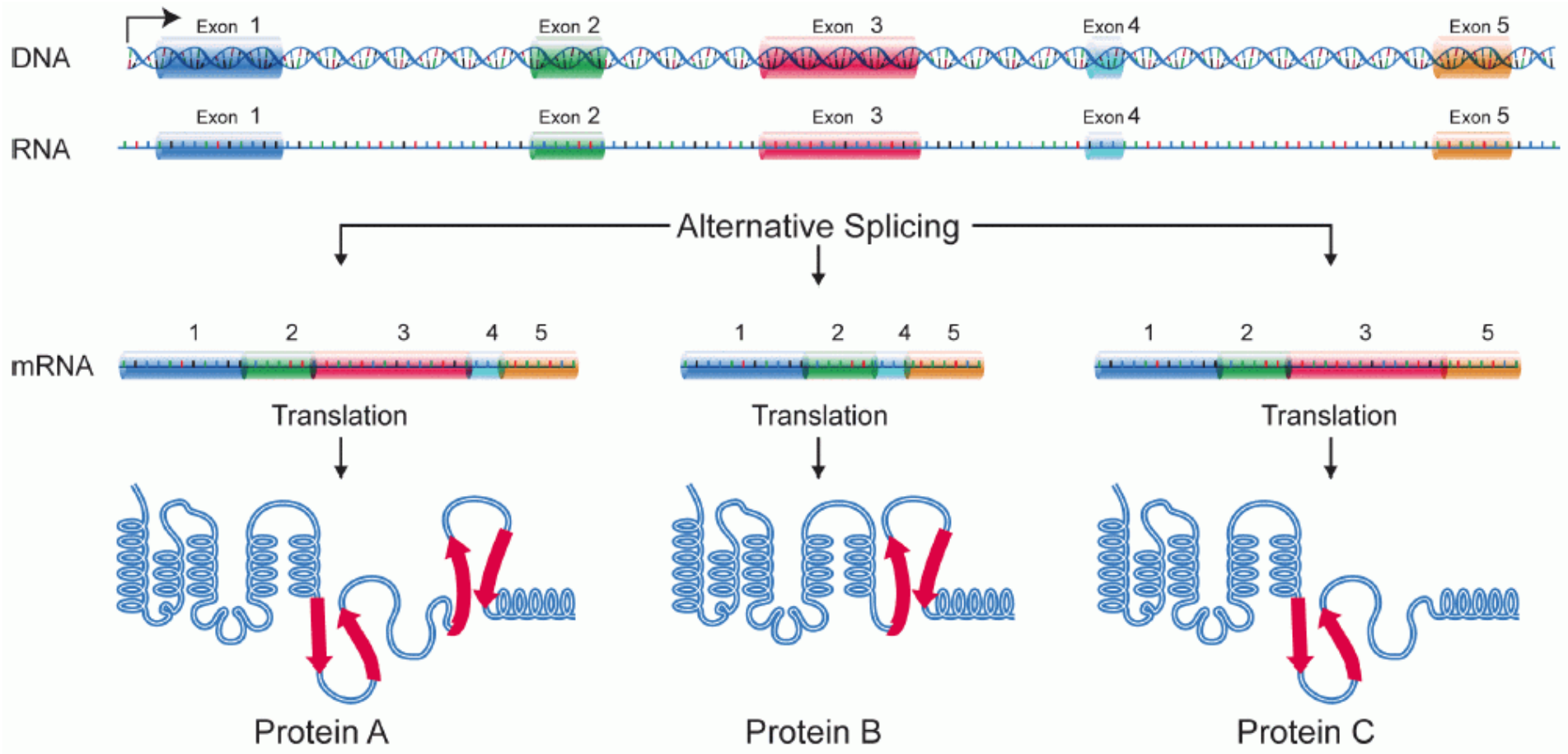
Population Frequencies

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
Latino	1684	34416	55	0.04893
East Asian	818	18866	22	0.04336
Other	70	6464	0	0.01083
Ashkenazi Jewish*	45	10146	0	0.004435
European (Finnish)	90	25730	1	0.003498
South Asian	83	30780	2	0.002697
African	13	24030	0	0.0005410
European (Non-Finnish)	63	126704	0	0.0004972
Total	**2866**	**277136**	**80**	**0.01034**

Include: Exomes Genomes

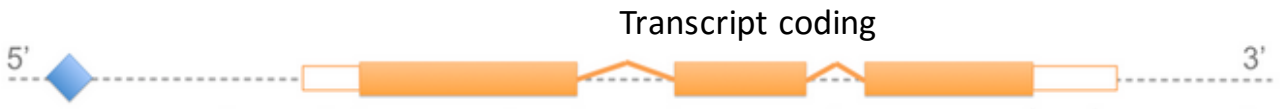
* For detailed analysis of Ashkenazi Jewish frequency see the IBD Exomes Browser.

Variants can affect at different levels



DNA level

DNA REGIONS



Regulatory Region
Transcription Factor Binding site

BIOTYPE Type of transcript or regulatory feature. e.g. protein_coding, processed_pseudogene...

GENE IDENTIFIER

- HUGO Gene Symbol (Standard nomenclature for the human genes) e.g.: MYC
- Specific nomenclature from databases

  e.g. ENSG00000136997 (ensembl)

  4609 (Entrez gene NCBI)

EXON/INTRON NUMBER

ESSENTIALITY

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

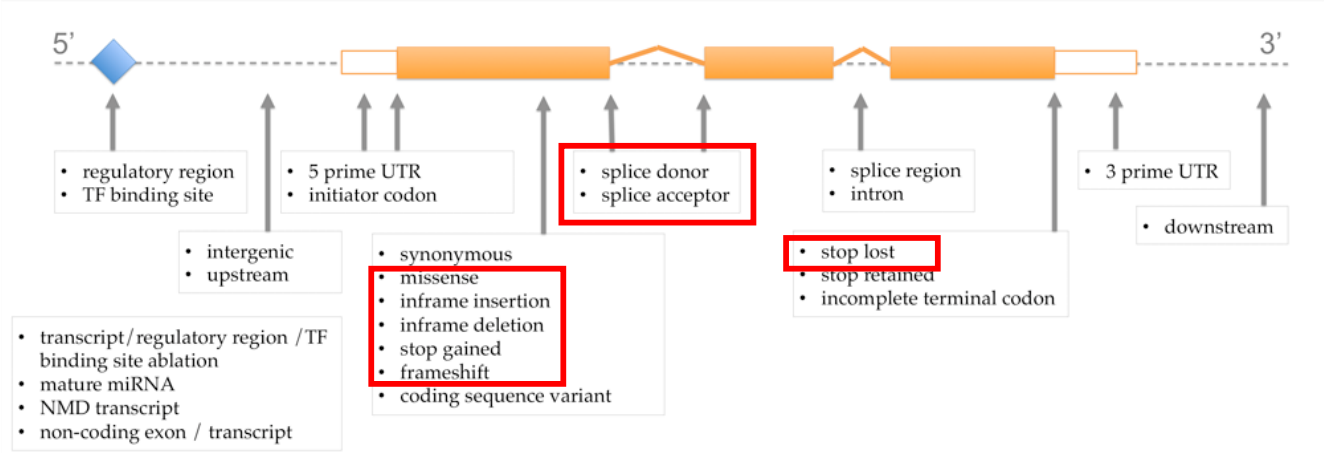
$$\frac{\text{LoF}}{\text{Synonymous}}$$

Transcript level

TRANSCRIPT IDENTIFIERS

- ensembl e.g. ENST00000426406
- RefSeq e.g. NM_001005221.2

CONSEQUENCE



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 codo	SO:0002012	Start lost	HIGH
	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequenc	SO:0001821	Inframe insertion	MODERATE
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequenc	SO:0001822	Inframe deletion	MODERATE
	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE

HIGH

MODERATE

High

Moderate

Low

Modifier

Transcript level

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.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 inteface

APPRIS Database currently houses annotations for vertebrate genomes »

Human

Assemblies: GRCh38
Assemblies: GRCh37

Mouse

Assemblies: GRCm38

Zebrafish

Assemblies: GRCz10
Assemblies: Zv9

Rat

Assemblies: Rnor_6.0
Assemblies: Rnor_5.0

Pig

Assemblies: Sscrofa10.2

Chimpanzee

Assemblies: CHIMP2.1.4

APPRIS Database currently houses annotations for invertebrate genomes »

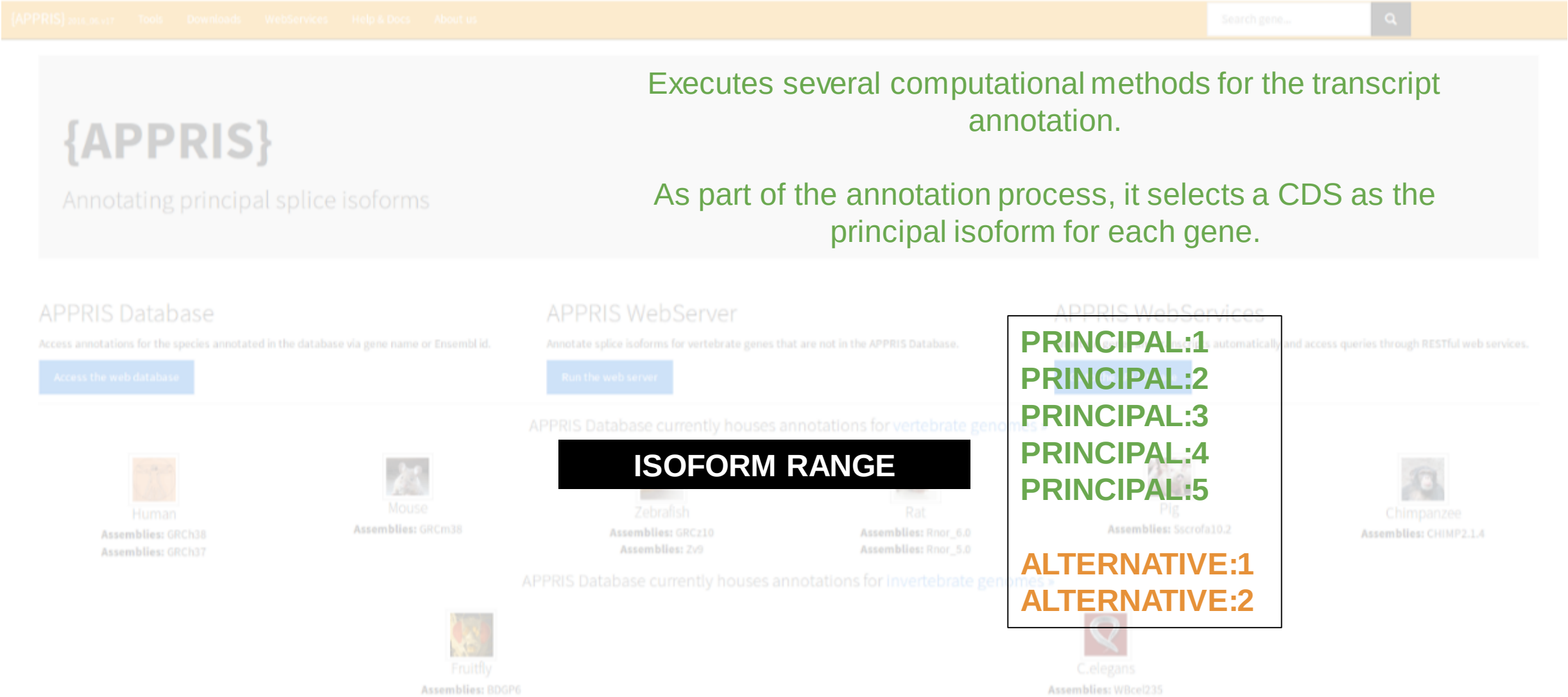
Fruitfly

Assemblies: BDGP6

C.elegans

Assemblies: WBcel235

Principal Isoform - APPRIS



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.

ISOFORM RANGE

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

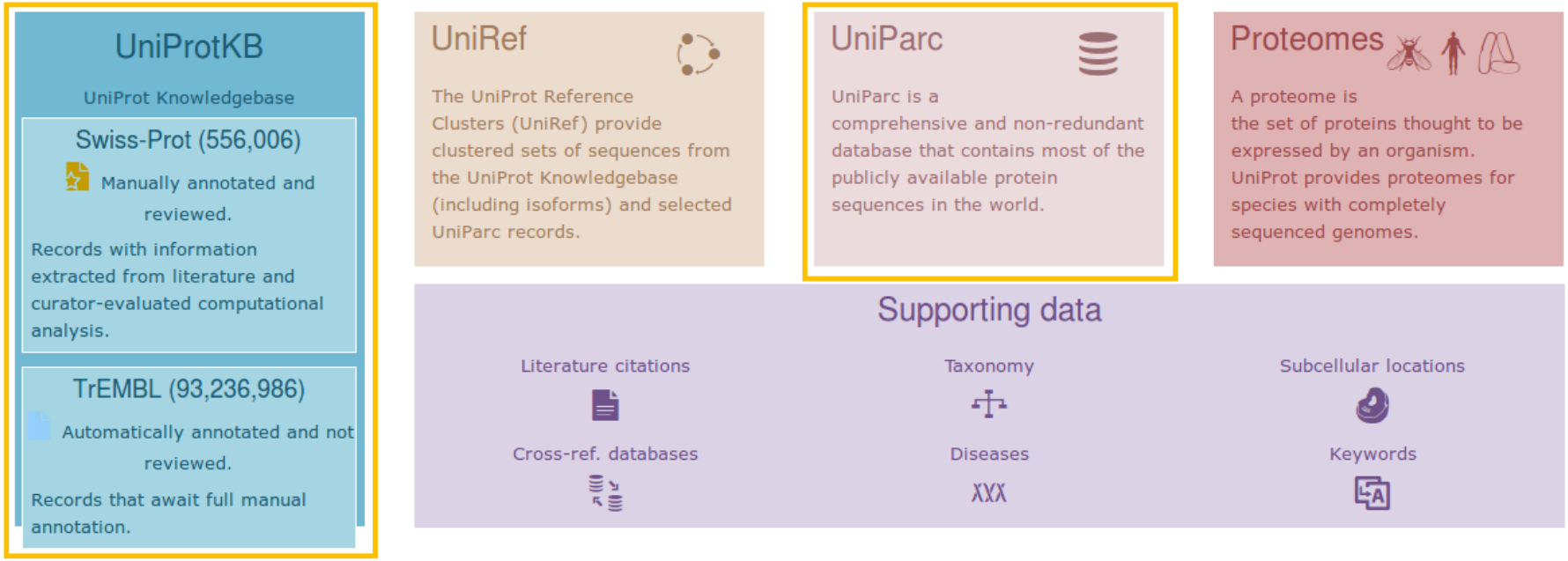
ALTERNATIVE:1
ALTERNATIVE:2

Protein level

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)



Consequence: Functional impact prediction

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

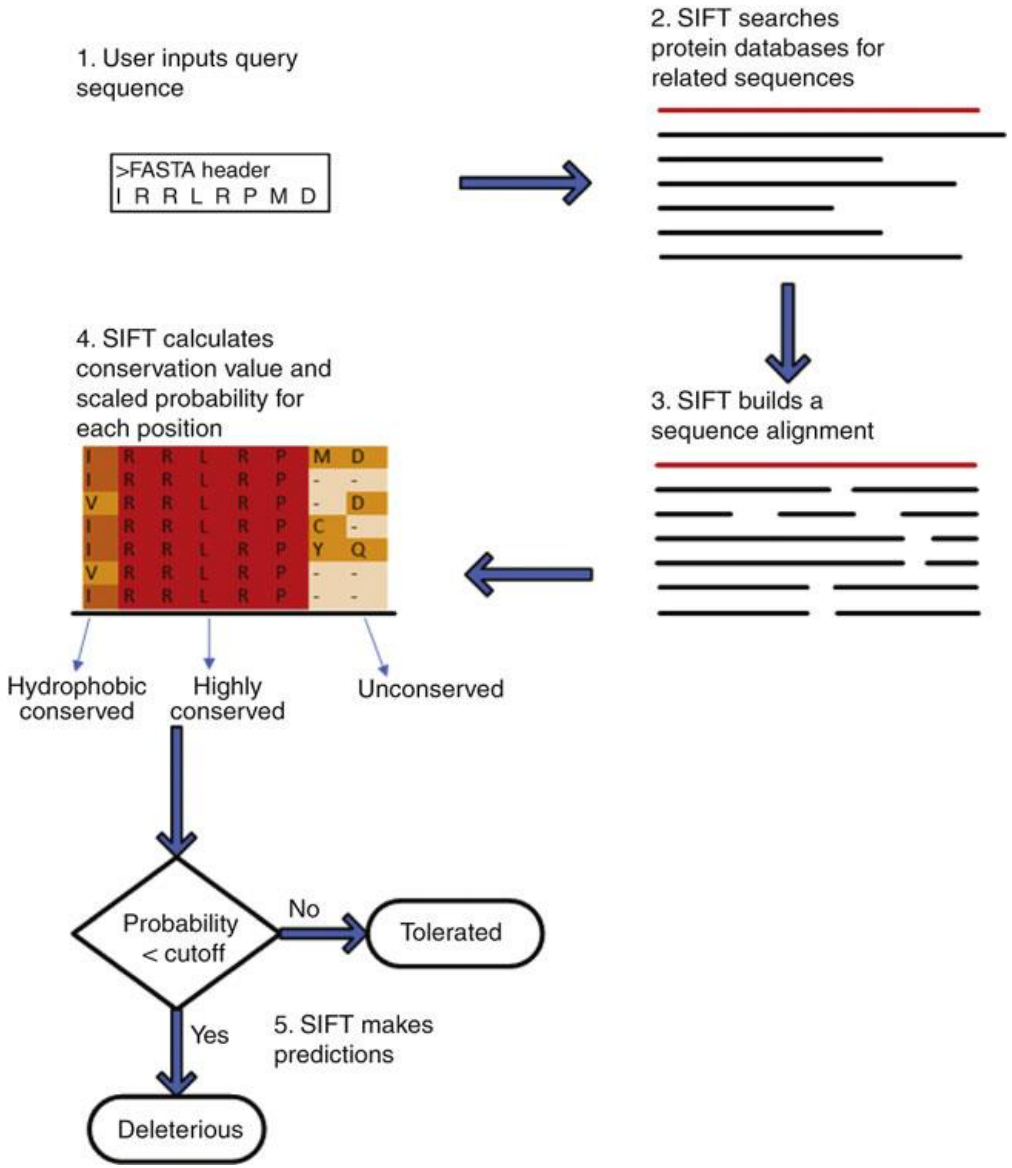
Deleterious – Tolerated

0 < 0.05 Deleterious 1

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

Unknown - Benign - Possibly damaging - Probably damaging

0 Benign < ~0.435 < Damaging 1



Consequence: Functional impact prediction

SIFT

Deleterious – Tolerated

0 < 0.05 Deleterious 1

PolyPhen2

Unknown - Benign - Possibly damaging - Probably damaging

0 Benign < ~0.435 < Damaging 1

CONDEL CONsensus DELeteriousness score computes a consensus score based on:

MutationAssessor
FatHMM

Impact prediction – Other predictors

- **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

Most of predictors allows only the prediction of coding non-synonymous SNV

- **PROVEAN**

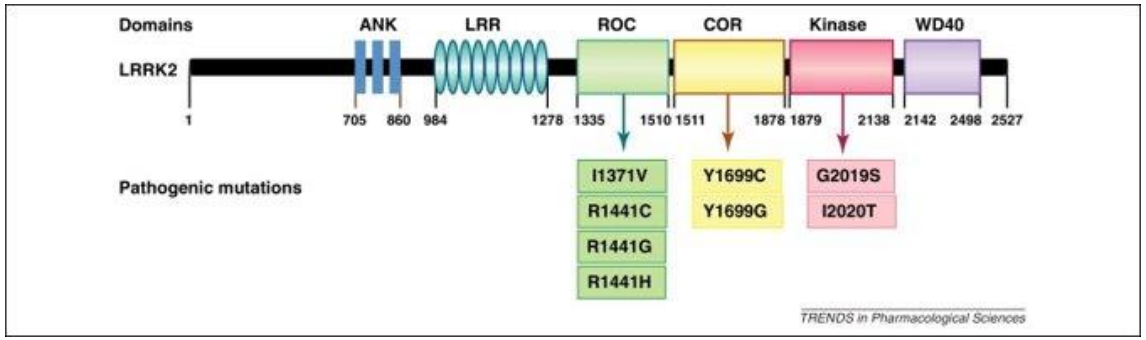
Functional prediction for nonsynonymous mutations or indels

Consequence: Domains

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

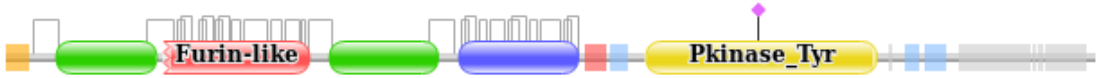
Protein family membership

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

Homologous superfamilies

Domains and repeats

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



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

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

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

Variant Effect Predictor (VEP)

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

Login/Register

Search all species...

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...

Go

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!

Launch VeP

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

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

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

With several ways of execution

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

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

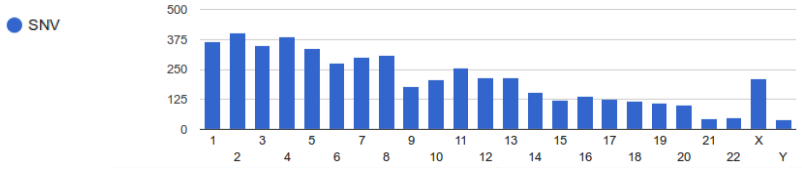
VEP version (API)	85 (85)
Cache/Database	/home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/.vep/homo_sapiens/85_GRCh37
Species	homo_sapiens
Command line options	`--format vcf --sift b --polyphen b --ccds --uniprot --hgvs --symbol --numbers --domains --regulatory --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

Variants by chromosome



VEP standalone script output

```
##VEP=v85 cache=/home/epineiro/analysis/pancancer/vep/ensembl-tools-release-85/scripts/variant_effect_predictor/.vep/homo_sapiens/85_GRCh37 db=. dbSNP=144 gencode=GENCODE 19
ESP=20141103 sift=sift5.2.2 regbuild=13 assembly=GRCh37.p13 polyphen=2.2.2 ClinVar=201507 HGMD-PUBLIC=20152 genebuild=2011-04 COSMIC=71
##Condel=Consensus deleteriousness score for an amino acid substitution based on SIFT and PolyPhen-2
##INFO=<ID=CSQ,Number=.,Type=String,Description="Consequence annotations from Ensembl VEP. Format: Allele|Consequence|IMPACT|SYMBOL|Gene|Feature_type|Feature|BIOTYPE|EXON|INTRON|
HGVSc|HGVSp|cDNA_position|CDS_position|Protein_position|Amino_acids|Codons|Existing_variation|DISTANCE|STRAND|FLAGS|VARIANT_CLASS|SYMBOL_SOURCE|HGNC_ID|CANONICAL|TSL|CCDS|ENSP|
SWISSPROT|TREMBL|UNIPARC|RefSeq|SIFT|PolyPhen|DOMAINS|HGVS_OFFSET|GMAF|AFR_MAF|AMR_MAF|EAS_MAF|EUR_MAF|SAS_MAF|AA_MAF|EA_MAF|ExAC_MAF|ExAC_Adj_MAF|ExAC_AFR_MAF|ExAC_AMR_MAF|
ExAC_EAS_MAF|ExAC_FIN_MAF|ExAC_NFE_MAF|ExAC_OTH_MAF|ExAC_SAS_MAF|CLIN_SIG|SOMATIC|PHENO|MOTIF_NAME|MOTIF_POS|HIGH_INF_POS|MOTIF_SCORE_CHANGE|Condel">
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
1   1244671 .   G   A   255.0   .   Callers=broad,dkfz,muse,sanger;NumCallers=4;dbsnp=rs774706740;VAF=0.1475;t_alt_count=9;t_ref_count=52;CSQ=A|downstream_gene_variant|MODIFIER|
CPSF3L|ENSG00000127054|Transcript|ENST00000323275|retained_intron|||||||||rs774706740|2309|-1||SNV|HGNC|26052||||||||||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|A:0|A:0|A:0|
A:0|||||||,A|upstream_gene_variant|MODIFIER|ACAP3|ENSG00000131584|Transcript|ENST00000353662|protein_coding|||||||||rs774706740|1402|-1||SNV|HGNC|16754||||ENSP00000321139|Q96P50|
Q8N2W2|UPI000012749C|||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|A:0|A:0|A:0|A:0|||||||,A|upstream_gene_variant|MODIFIER|ACAP3|ENSG00000131584|Transcript|ENST00000354700|
protein_coding|||||||||rs774706740|1273|-1||SNV|HGNC|16754|YES||CCDS19.2|ENSP00000346733|Q96P50|Q8WTZ1&Q8N2W2|UPI0000050F41|NM_030649.2||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|
A:0|A:0|A:0|A:0|||||||,A|upstream_gene_variant|MODIFIER|ACAP3|ENSG00000131584|Transcript|ENST00000354980|nonsense_mediated_decay|||||||||rs774706740|1470|-1||SNV|HGNC|16754||||
ENSP00000347075||F8W850|UPI000198C4CE|||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|A:0|A:0|A:0|A:0|||||||,A|intron_variant|MODIFIER|PUSL1|ENSG00000169972|Transcript|ENST00000379031|
protein_coding||3/7|ENST00000379031.5:c.323+18G>A|||||||rs774706740||1||SNV|HGNC|26914|YES||CCDS20.1|ENSP00000368318|Q8N0Z8|J3KTG4|UPI0000051C19|NM_153339.1||||||A:0|||||||A:0|
A:2.644e-05|A:0|A:0|A:0|A:0|A:0|A:0|||||||,A|downstream_gene_variant|MODIFIER|CPSF3L|ENSG00000127054|Transcript|ENST00000411962|protein_coding|||||||||rs774706740|2310|-1||SNV|
HGNC|26052||||ENSP00000400548||J3QRY6&C9IYS7|UPI0000EE7E25|NM_001256462.1||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|A:0|A:0|A:0|A:0|||||||,A|downstream_gene_variant|MODIFIER|CPSF3L|
ENSG00000127054|Transcript|ENST00000419704|protein_coding|||||||||rs774706740|2315|-1||SNV|HGNC|26052|||CCDS57961.1|ENSP00000404886|Q5TA45|J3QRY6|UPI000014103F|
NM_001256463.1||||||A:0|||||||A:0|A:2.644e-05|A:0|A:0|A:0|A:0|A:0|A:0|||||||,A|downstream_gene_variant|MODIFIER|CPSF3L|ENSG00000127054|Transcript|ENST00000421495|
```

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

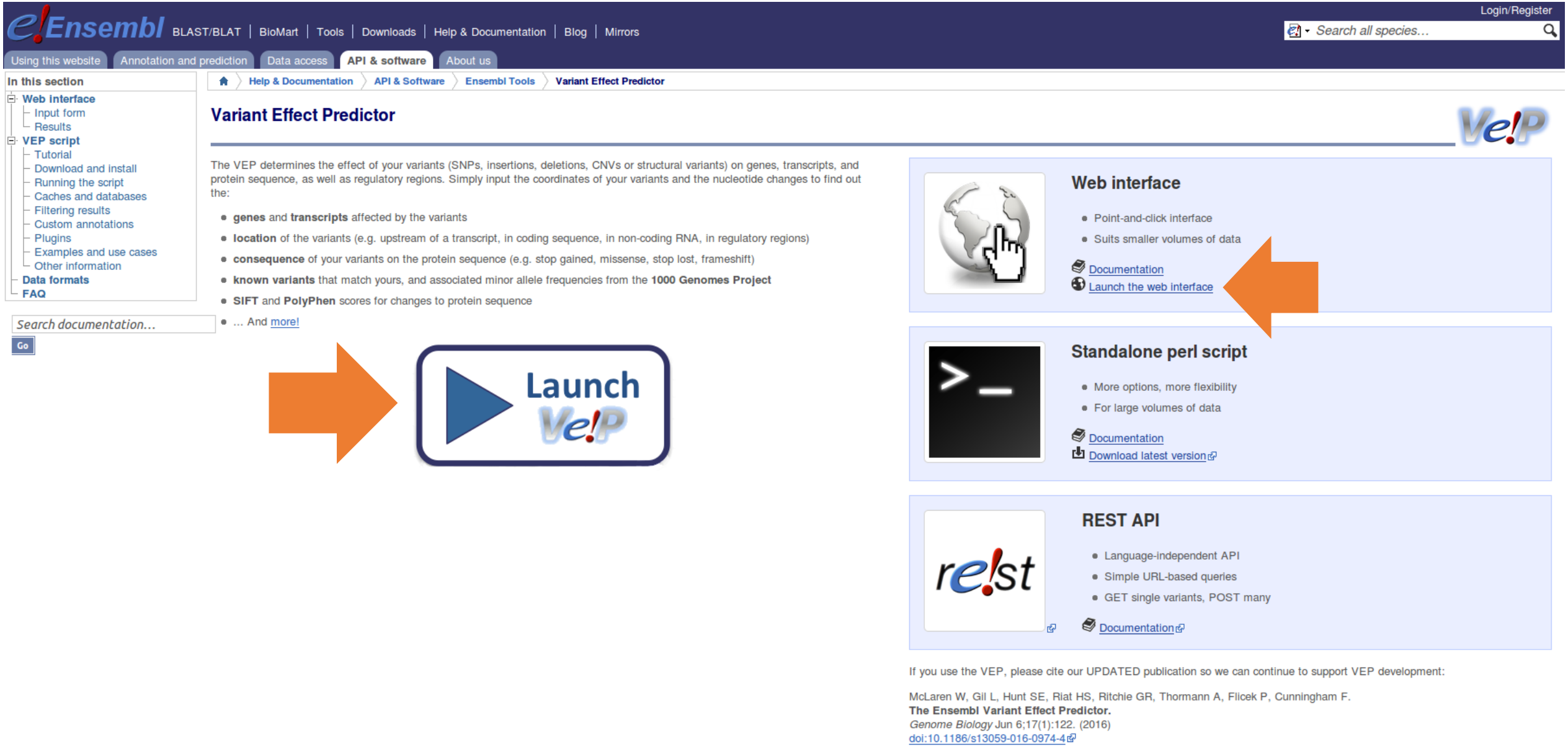
HGVSc (--hgvs)
HGVSp (--hgvs)
cDNA_position
CDS_position
Protein_position
Amino_acids
Codons
Existing_variation
DISTANCE
STRAND

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



web execution

Variant Effect Predictor

VEP for Human GRCh37

If you are looking for VEP for Human GRCh37, please go to GRCh37 website.

hg19

Species: Human (Homo sapiens)

Assembly: GRCh38.p7

hg38

Name for this job (optional):

Either paste data:

Examples: Ensembl default, VCF, Variant identifiers, HGVS notations, Pileup

Or upload file: Browse... No file selected.

Or provide file URL:

Input options

Transcript database to use:

- (●) Ensembl transcripts
- ( ) Gencode basic transcripts
- ( ) RefSeq transcripts
- ( ) Ensembl and RefSeq transcripts

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:	[x] protein coding, pseudogene, processed pseudogene, miRNA, rRNA, scRNA, snoRNA and snRNA
Protein domains:	[ ]
Exon and intron numbers:	[ ]
Transcript support level:	[x]
APPRIS:	[x]
Identify canonical transcripts:	[ ]
Upstream/Downstream distance (bp):	5000
miRNA structure(p):	[ ]

Pathogenicity predictions

SIFT:	Prediction and score
PolyPhen:	Prediction and score
dbNSFP(p):	(•) Disabled ( ) Enabled
Condel(p):	(•) Disabled ( ) Enabled
LoFtool(p):	[ ]

web execution

Regulatory data

Get regulatory region consequences: Yes

Splicing predictions

dbscSNV(p): ☐

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 Close form

web execution

Show/hide columns (1 hidden) Filter

Analysis	Jobs	Submitted at
Variant Effect Predictor	VEP analysis of 0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf in Homo_sapiens Queued	13/03/2017, 14:58 (GMT)

Show/hide columns (1 hidden) Filter

Analysis	Jobs	Submitted at
Variant Effect Predictor	VEP analysis of 0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf in Homo_sapiens Running	13/03/2017, 14:29 (GMT)

Show/hide columns (1 hidden) Filter

Analysis	Jobs	Submitted at
Variant Effect Predictor	VEP analysis of 0a6be23a-d5a0-4e95-ada2-a61b2b5d9485.vcf in Homo_sapiens Done [View results]	13/03/2017, 14:29 (GMT)

web execution

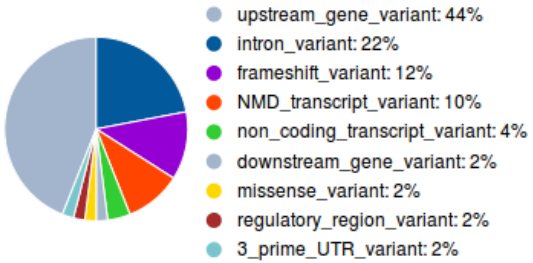
Variant Effect Predictor results

Job details

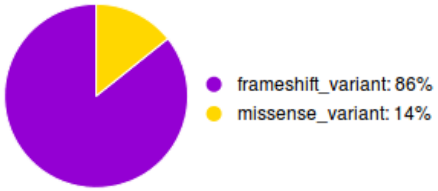
Summary statistics

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 Add

Download

Download

All: VCF VEP TXT

BioMart: Variants Genes

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	Pr po
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	-	-	-	-	-	-	-