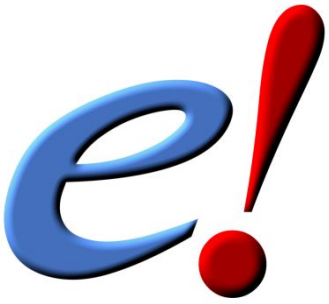
The Variant Effect Predictor



www.ensembl.org
www.ensemblgenomes.org

Coursebook v95

http://training.ensembl.org/events/2019/2019-04-02-VEP_Breda

Variant Effect Prediction Course, Breda
2nd-5th April 2019

CC BY

Introduction to Ensembl

Getting started with Ensembl
www.ensembl.org

Ensembl is a project based at the EBI (European Bioinformatics Institute) that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'Compara' team. Most annotation is updated every two to three months to generate increasing Ensembl versions (84, 85, 86, etc.); however, the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates—namely, bacteria, plants, fungi, metazoa, and protists.

Ensembl provides genes and other **annotation** such as predicted regulatory regions, base pairs conserved across species, and observed sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (**Perl API**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge on the part of the user.

Synopsis — What can I do with Ensembl?

- View genes, with other annotation, along the chromosome.
- View alternative transcripts (such as splice variants) for a given gene.
- For any gene, explore homologues and phylogenetic trees across more than 70 species.
- Compare whole genome alignments and conserved regions across species.
- View microarray sequences matching Ensembl genes.
- View ESTs, clones, mRNAs, and proteins for any chromosomal region.
- Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region.
- View SNPs across strains (rat, mouse), populations (human), or breeds (dog).
- View positions and sequences of mRNAs and proteins that align against Ensembl genes.
- Upload your own data.
- Use BLAST or BLAT against any Ensembl genome.
- Export sequence or create a table of gene information with BioMart.
- Determine how your variants affect genes and transcripts using the Variant Effect Predictor.
- Share Ensembl views with your colleagues and collaborators.

Need more help?

- Check Ensembl documentation
- Watch video tutorials on YouTube
- View the FAQs
- Try some exercises
- Read some publications
- Go to our online course

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list
- **Find us on Facebook or follow us on Twitter**
  - https://www.facebook.com/Ensembl.org/
  - @ensembl
  - @ensemblgenomes

Further reading

Cunningham, F. *et al.*
Ensembl 2019
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gky1113

Kersey, PJ *et al.*
Ensembl Genomes 2018: an integrated omics infrastructure for non-vertebrate species
Nucleic Acids Research (Database Issue)
https://doi.org/10.1093/nar/gkx1011

For a complete list of publications, visit:
http://www.ensembl.org/info/about/publications.html
http://ensemblgenomes.org/info/publications

Demo: The Variant Effect Predictor (VEP)

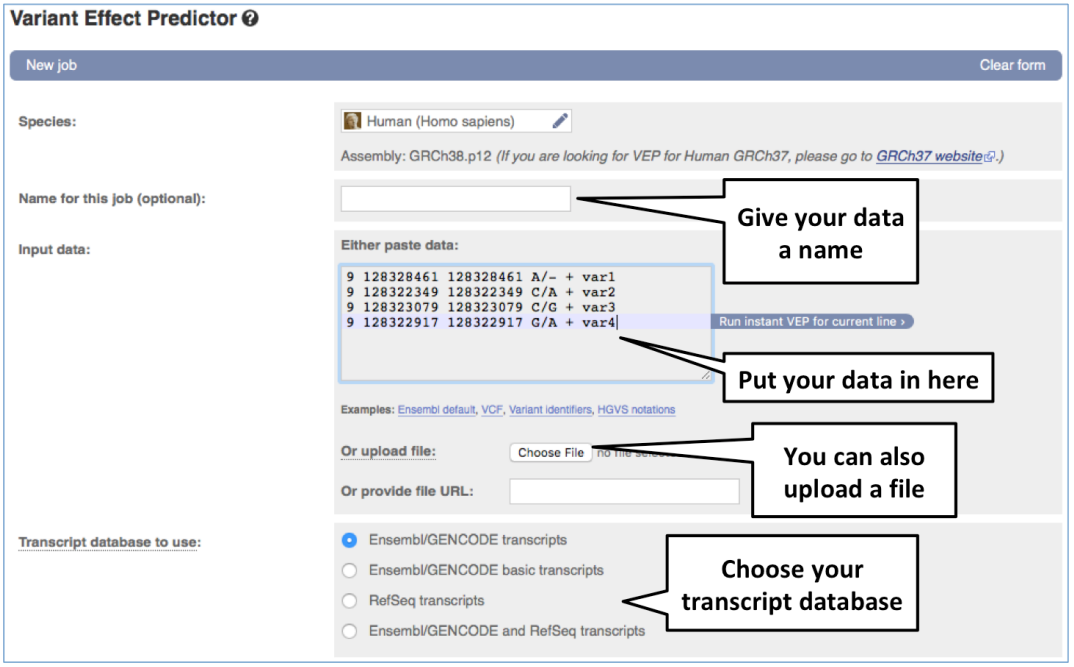
We have identified four variants on human chromosome nine, an A deletion at 128328461, C->A at 128322349, C->G at 128323079, and G->A at 128322917.

We will use the **Ensembl VEP** to determine:

- If the variants have already been annotated in Ensembl
- The genes affected by my variants
- If any of my variants affect gene regulation

Go to the Ensembl homepage and click on the VEP link in the blue navigation bar.

This page contains information about the VEP, including links to download the script version of the tool. Click on Launch VEP to open the input form.

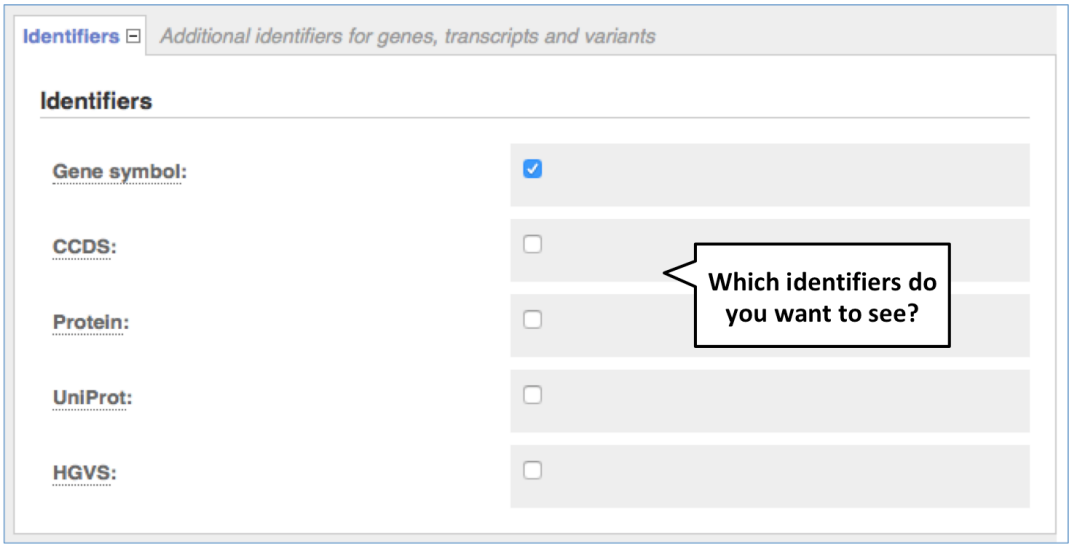


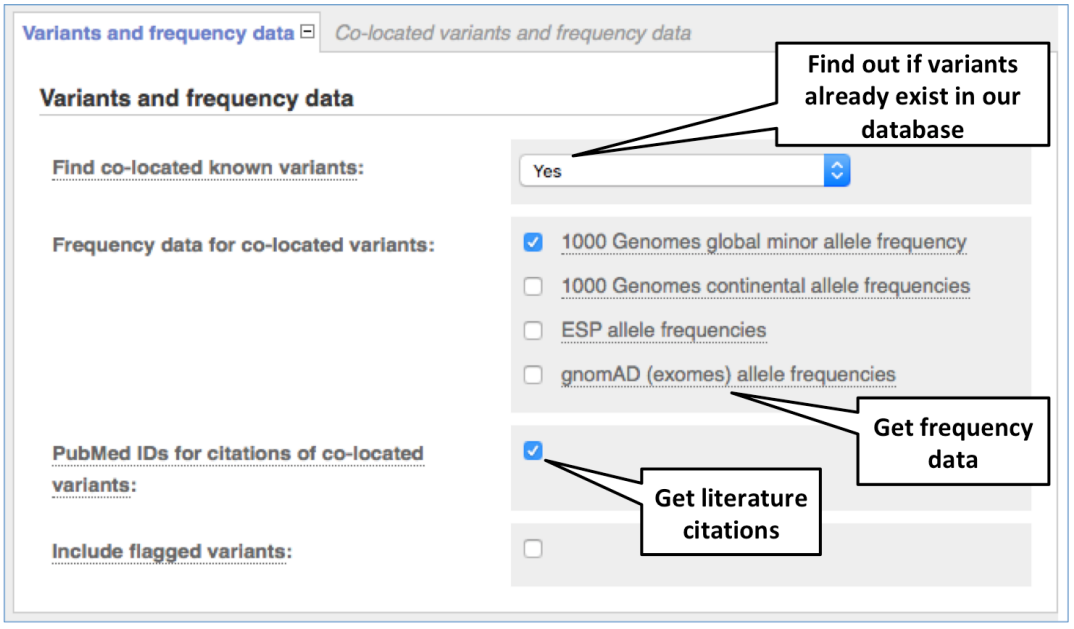
The data are in the format:
Chromosome Start End Alleles (reference/mutation) Strand

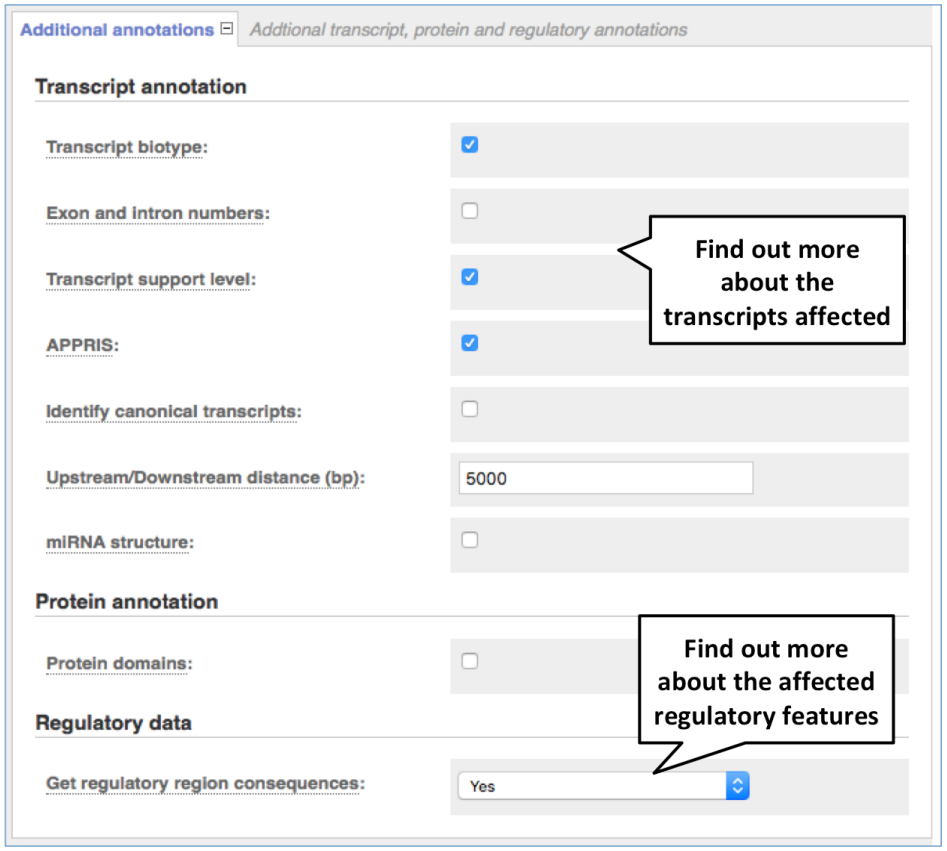
Put the following into the Paste data box:
9 128328461 128328461 A/- + var1
9 128322349 128322349 C/A + var2
9 128323079 128323079 C/G + var3
9 128322917 128322917 G/A + var4

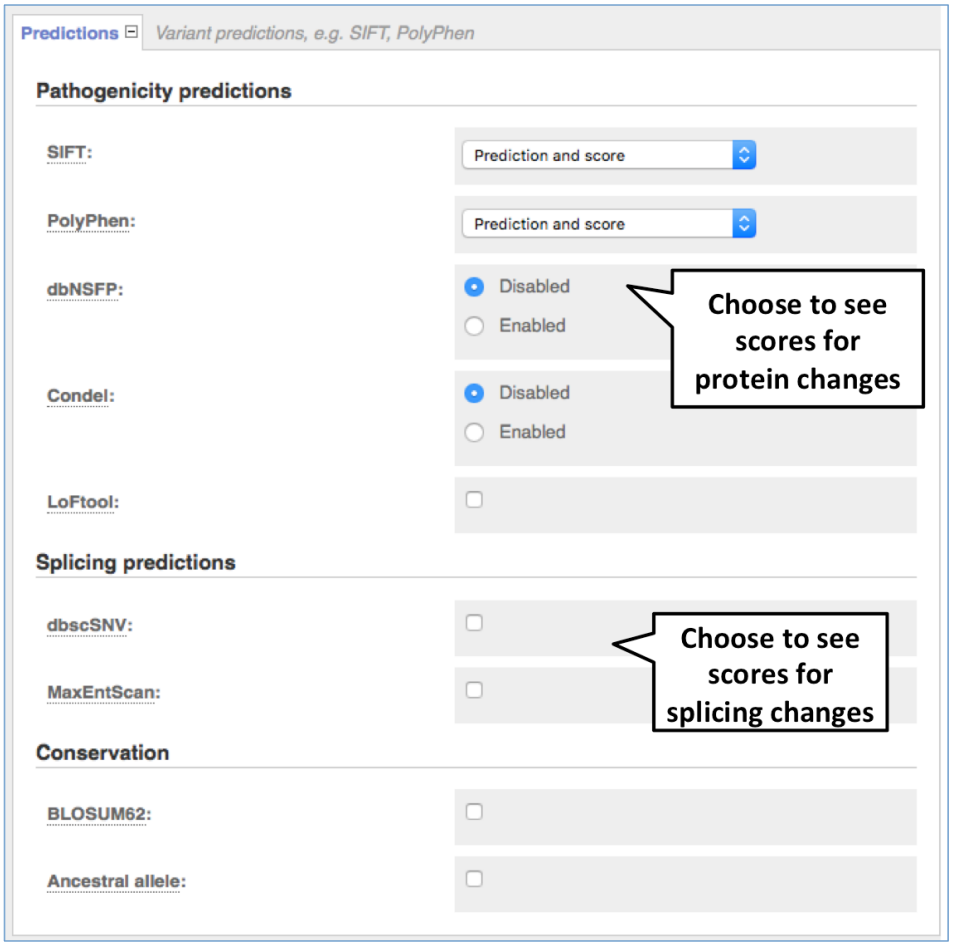
The VEP will automatically detect that the data are in Ensembl default format. (Note that the deletion in the first variant is indicated by "-".)

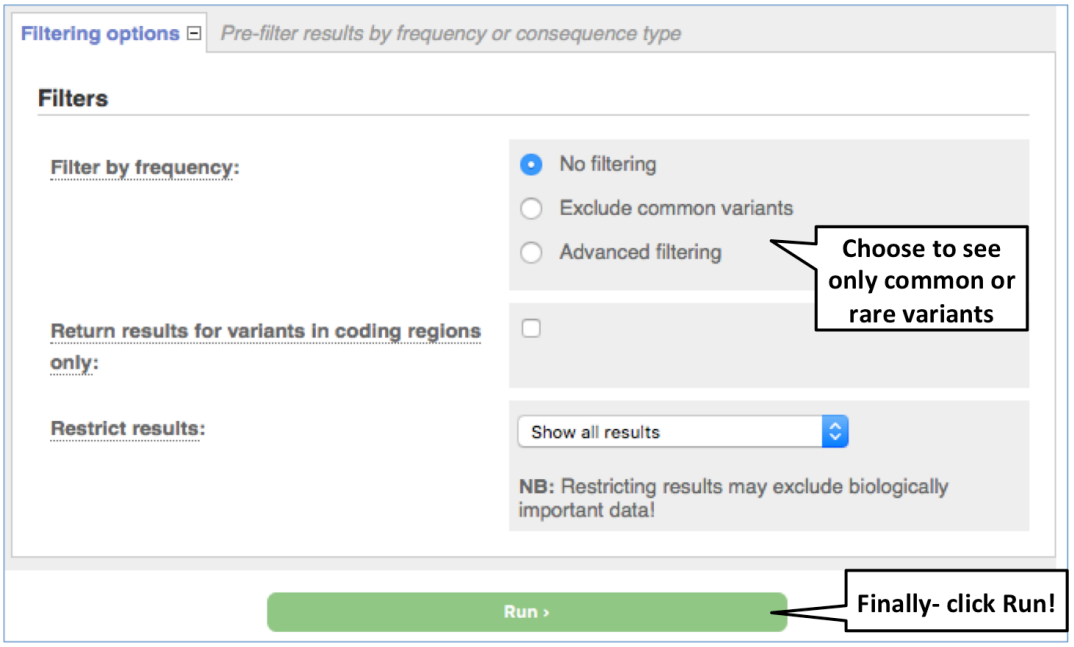
There are further options that you can choose for your output. These are categorised as Identifiers, Variants and frequency data, Additional annotations, Predictions and Filtering options. Let's open all the menus and take a look.






Additional annotations
Addtional 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
Find out more about the transcripts affected
Find out more about the affected regulatory features


Predictions
Variant predictions, e.g. SIFT, PolyPhen
Pathogenicity predictions
SIFT:
Prediction and score
PolyPhen:
Prediction and score
dbNSFP:
Disabled
Enabled
Condel:
Disabled
Enabled
LoFtool:
Splicing predictions
dbscSNV:
MaxEntScan:
Conservation
BLOSUM62:
Ancestral allele:
Choose to see scores for protein changes
Choose to see scores for splicing changes



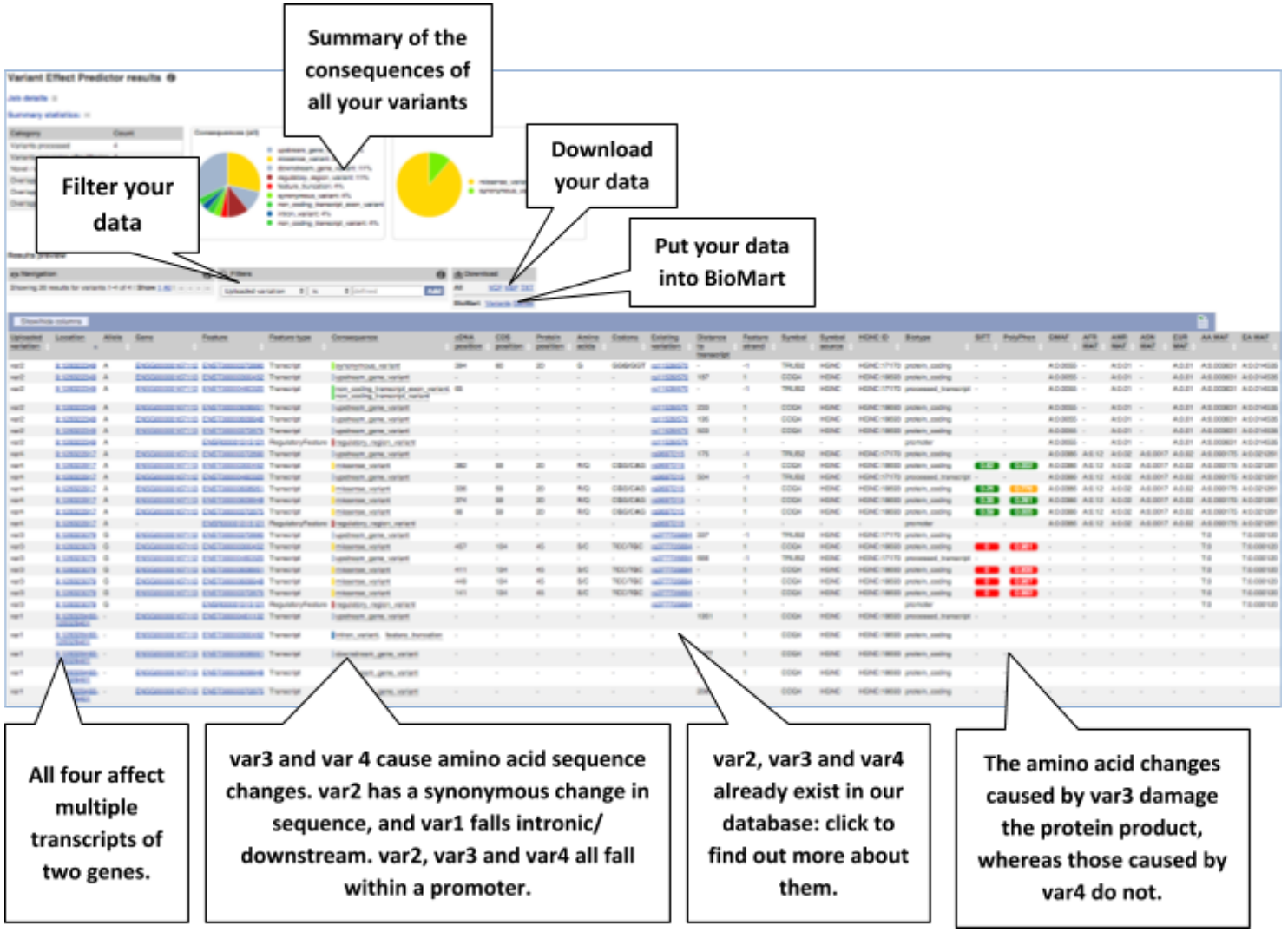
Hover over the options to see definitions.

When you've selected everything you need, scroll right to the bottom and click Run.

The display will show you the status of your job. It will say Queued, then automatically switch to Done when the job is done: you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

Click View Results once your job is done.

In your results you will see a graphical summary of your data as well as a table of your results. (Note that some empty columns in the results table have been hidden in the following screenshot to save space.)


Summary of the consequences of all your variants
Filter your data
Download your data
Put your data into BioMart
All four affect multiple transcripts of two genes.
var3 and var 4 cause amino acid sequence changes. var2 has a synonymous change in sequence, and var1 falls intronic/ downstream. var2, var3 and var4 all fall within a promoter.
var2, var3 and var4 already exist in our database: click to find out more about them.
The amino acid changes caused by var3 damage the protein product, whereas those caused by var4 do not.

Exercises

Exercise 1 – VEP using variant coordinates

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (ENSG00000001626) has revealed the following variants (alleles defined in the forward strand):

- G/A at 7: 117,530,985
- T/C at 7: 117,531,038
- T/C at 7: 117,531,068

Use the VEP tool in Ensembl and choose the options to see SIFT and PolyPhen predictions. Do these variants result in a change in the proteins encoded by any of the Ensembl genes? Which gene? Have the variants already been found?

Exercise 2 – viewing structural variants with the VEP

We have details of a genomic deletion in a breast cancer sample in VCF format:
13 32307062 sv1 . <DEL> . . SVTYPE=DEL;END=32332466

(a) What are the HGNC identifiers of the affected genes?

(b) Does the SV cause deletion of any complete transcripts?

(c) Display your variant in the Ensembl browser.

Exercise 3 – uploading a VCF file for VEP analysis

Sequencing of a patient with early onset Alzheimer disease, followed by alignment to the GRCh38 reference genome sequence and variant calling has produced the VCF located in the following directory:

http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2018/VEPTC_2018/VEPTC_VCF.vcf

(a) How many variants are known? How many are novel?

(b) Of the known variants, which variants are associated with susceptibility to Alzheimer disease?

(c) Do any variants have the ‘Regulatory_region_variant’ consequence? Which gene do the promoter region(s) overlap? Are any TF binding motifs affected?

(d) Which gene(s) are affected by missense variant(s). Are these gene(s) associated with any phenotypes/diseases?

(e) Are there any ‘stop_lost’ consequence predictions? Which gene is affected?

Quick Guide to Databases and Projects

Here is a list of databases and projects you will come across in these exercises. Google any of these to learn more. Projects include many species, unless otherwise noted.

Other help:
The Ensembl Glossary: http://www.ensembl.org/Help/Glossary
Ensembl FAQs: http://www.ensembl.org/Help/Faq

SEQUENCES
EMBL-Bank, NCBI GenBank, DDBJ – Contain nucleic acid sequences deposited by submitters such as wet-lab biologists and gene sequencing projects. These three databases are synchronised with each other every day, so the same sequences should be found in each.

CCDS – coding sequences that are agreed upon by Ensembl, VEGA-Havana, UCSC, and NCBI. *(Human and mouse)*

NCBI Entrez Gene – NCBI's gene collection.

NCBI RefSeq – NCBI's collection of "reference sequences", includes genomic DNA, transcripts, and proteins. NM stands for "Known mRNA" (e.g. NM_005476) and NP (e.g. NP_005467) are "Known proteins".

UniProtKB – the "Protein knowledgebase", a comprehensive set of protein sequences. Divided into two parts: Swiss-Prot and TrEMBL.

UniProt Swiss-Prot – the manually annotated, reviewed protein sequences in the UniProtKB. High quality.

UniProt TrEMBL – the automatically annotated, unreviewed set of proteins (EMBL-Bank translated). Varying quality.

VEGA – Vertebrate Genome Annotation, a selection of manually-curated genes, transcripts, and proteins. *(Human, mouse, zebrafish, gorilla, wallaby, pig, and dog)*

VEGA-HAVANA – The main contributor to the VEGA project, located at the Wellcome Trust Sanger Institute, Hinxton, UK.

GENE NAMES

HGNC – HUGO Gene Nomenclature Committee, a project assigning a unique and meaningful name and symbol to every human gene. *(Human)*

ZFIN – The Zebrafish Model Organism Database. Gene names are only one part of this project. *(Z-fish)*

PROTEIN SIGNATURES

InterPro – A collection of domains, motifs, and other protein signatures. Protein signature records are extensive, and combine information from individual projects such as UniProt, along with other databases such as SMART, PFAM, and PROSITE (explained below).

PFAM – A collection of protein families.

PROSITE – A collection of protein domains, families, and functional sites.

SMART – A collection of evolutionarily conserved protein domains.

OTHER PROJECTS

NCBI dbSNP – A collection of sequence polymorphisms, mainly single nucleotide polymorphisms, along with insertion-deletions.

NCBI OMIM – Online Mendelian Inheritance in Man – a resource showing phenotypes and diseases related to genes. *(Human)*