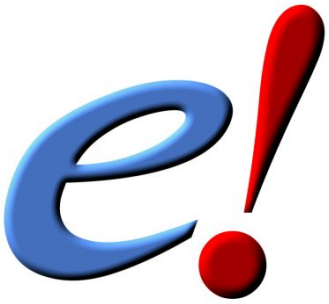
Interpreting Genetic Variation in Clinical Research using Ensembl:

Exploring Ensembl/GENCODE Annotation, MANE Transcripts and the Ensembl Variant Effect Predictor (VEP)



www.ensembl.org
www.ensemblgenomes.org

Variant Effect Predictor (VEP)

ASHG, 27th-30th October 2020

TABLE OF CONTENTS

Introduction to Ensembl

Getting started with Ensembl
www.ensembl.org

Ensembl is a project from 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 months, leading to increasing Ensembl versions (such as version 101), 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 regulatory regions, conserved base pairs across species, and 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 Interfaces (Perl APIs). 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 by the user.

Synopsis – What can I do with Ensembl?

- View genes with other annotation along the chromosome.
- View alternative transcripts (i.e. splice variants) for a given gene.
- Explore homologues and phylogenetic trees across more than 100 species for any gene.
- Compare whole genome alignments and conserved regions across species.
- View microarray sequences that match to Ensembl genes.
- View ESTs, clones, mRNA 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 sequence of mRNAs and proteins that align with 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

Recommend us to a friend

Ensembl do not charge any fees for delivering these workshops. If you know someone at another institute or move onto a new institute who you think would benefit from a course, let them know about us. Find out more about our training.

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list

Further reading

Yates, A. *et al*
Ensembl 2020
https://europepmc.org/abstract/MED/31691826

Howe, KL. *et al*
Ensembl Genomes 2020-enabling non-vertebrate genomic research
https://europepmc.org/abstract/MED/31598706

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

The Variant Effect Predictor (VEP)

Demo

We have identified six variants in patient presenting with neurological and cardiac symptoms:

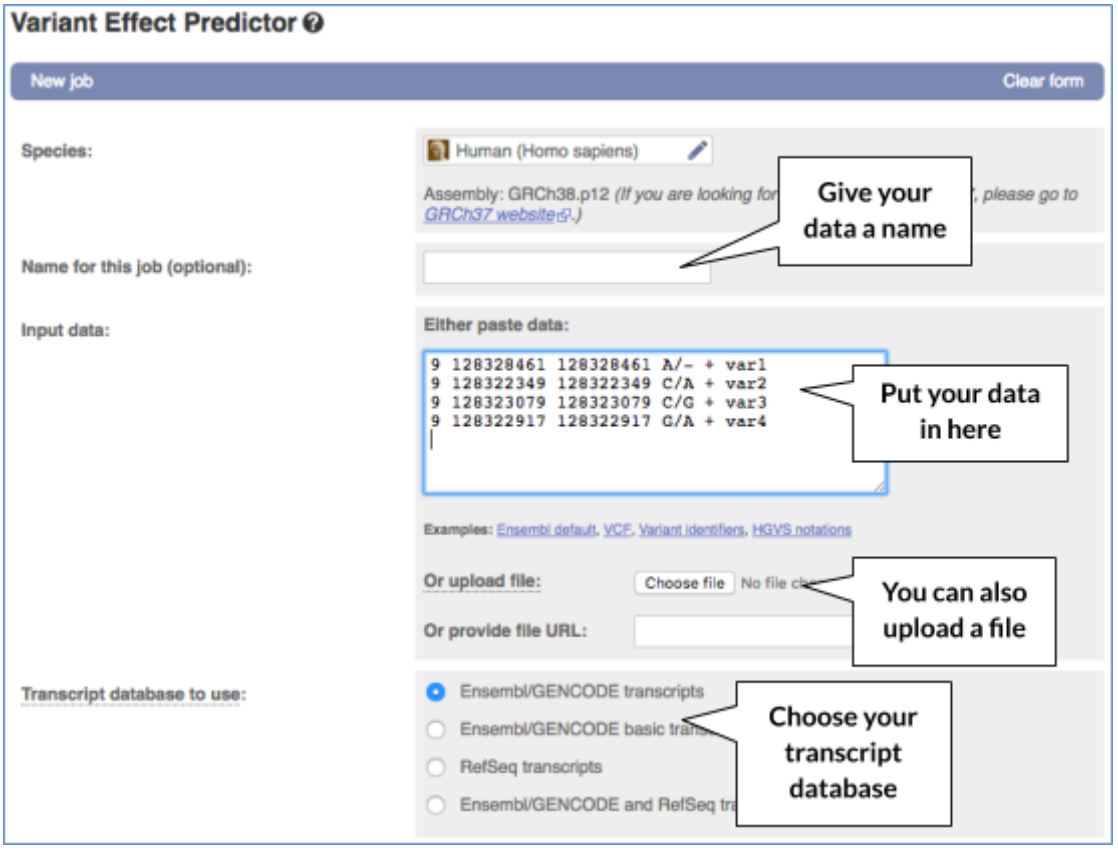
9	128328460	var1	TA	A
9	128322349	var2	C	A
9	128323079	var3	C	G
9	128322917	var4	G	A
14	73953549	var5	G	A
22	20995765	var6	C	T

We will use the Ensembl VEP to determine:

- Have my variants already been annotated in Ensembl?
- What genes are affected by my variants?
- Do any of my variants affect gene regulation?
- Can we prioritise them considering their consequence, allele frequency, clinical significance, pathogenicity predictions and associated phenotypes?

Go to the front page of Ensembl and click on the Variant Effect Predictor.

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.



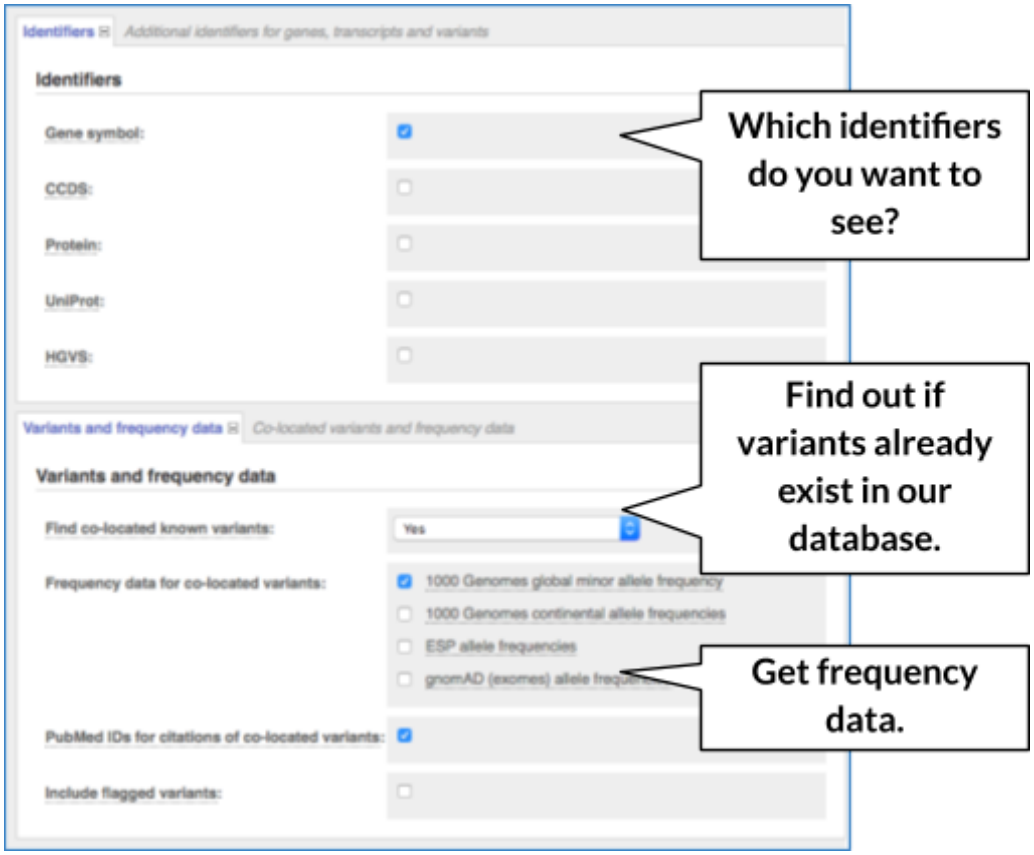
The data is in the VCF format:
`Chromosome Position Variant_ID Reference_allele Alternate_allele`

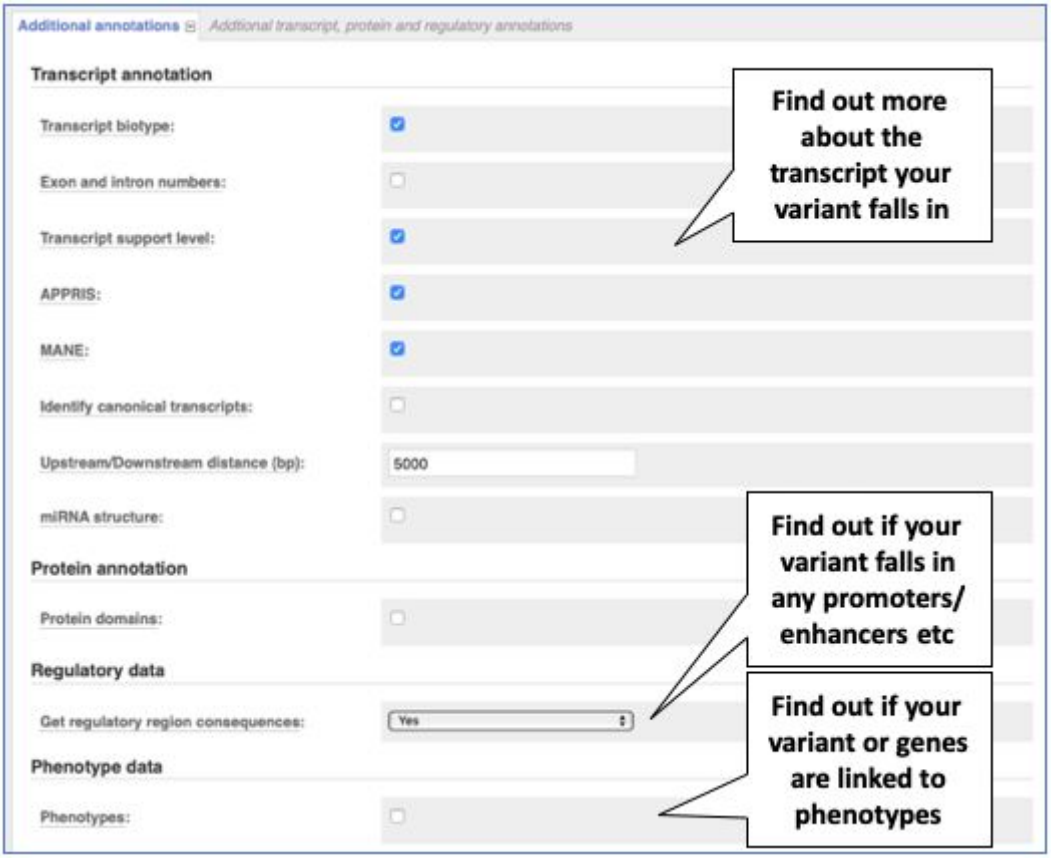
Put the following into the Paste data box:

```
9 128328460 var1 TA A
9 128322349 var2 C A
9 128323079 var3 C G
9 128322917 var4 G A
14 73953549 var5 G A
22 20995765 var6 C T
```

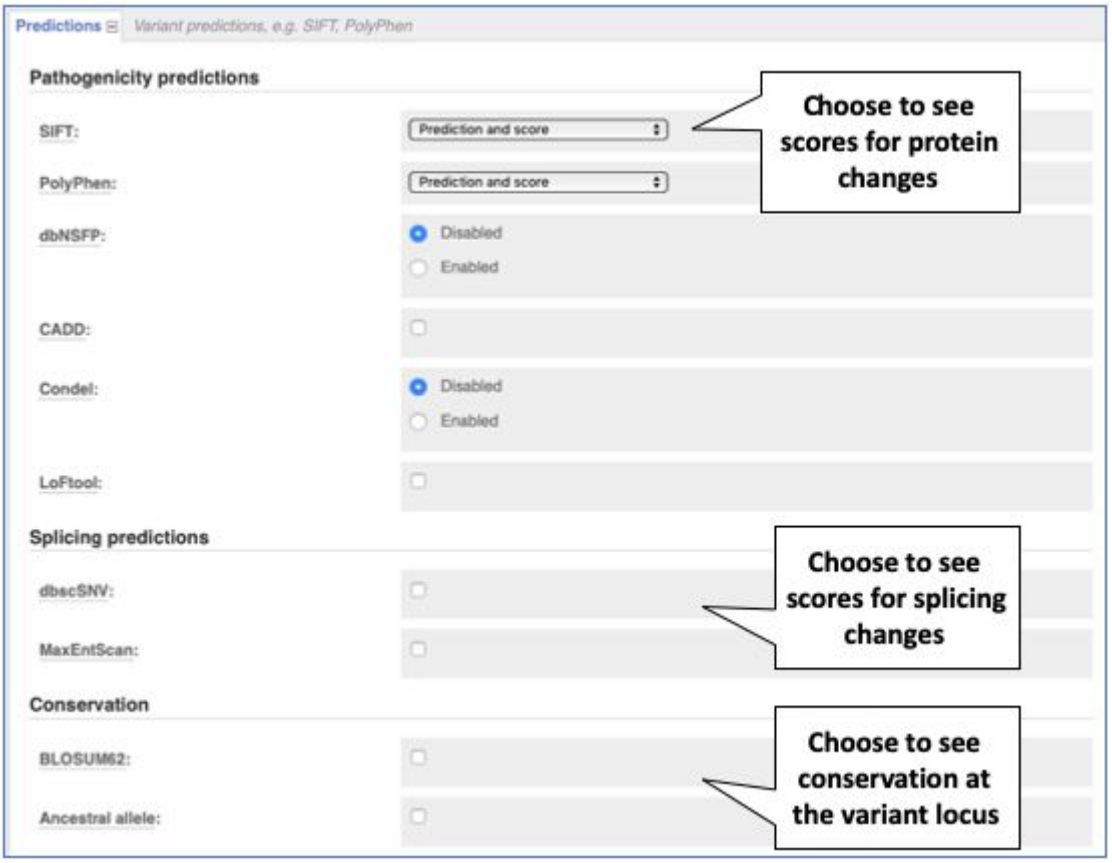
The VEP will automatically detect that the data is in VCF format.

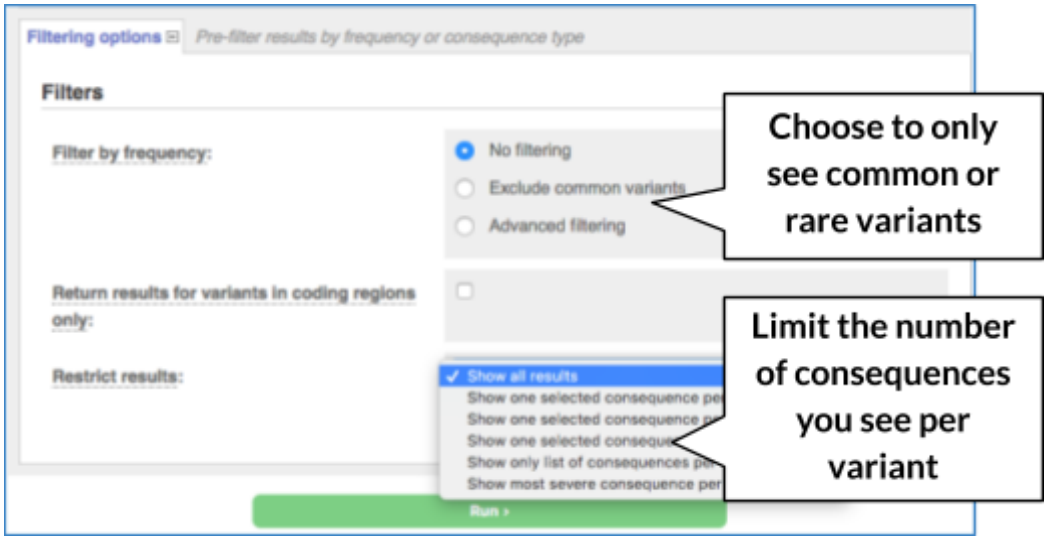
There are further options that you can choose for your output. These are categorised as Identifiers, Variants and frequency data, Additional annotations, Predictions, Filtering options and Advanced options. Let's open all the menus and take a look.

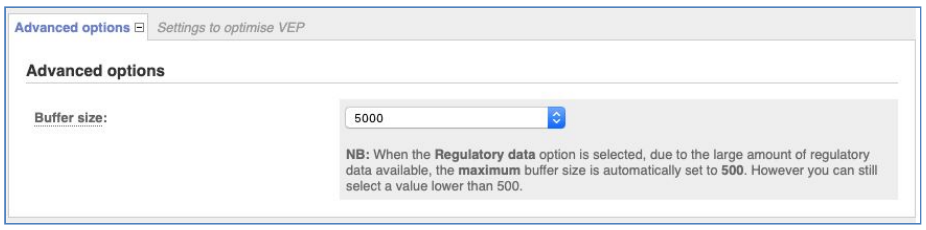




Select Phenotypes.

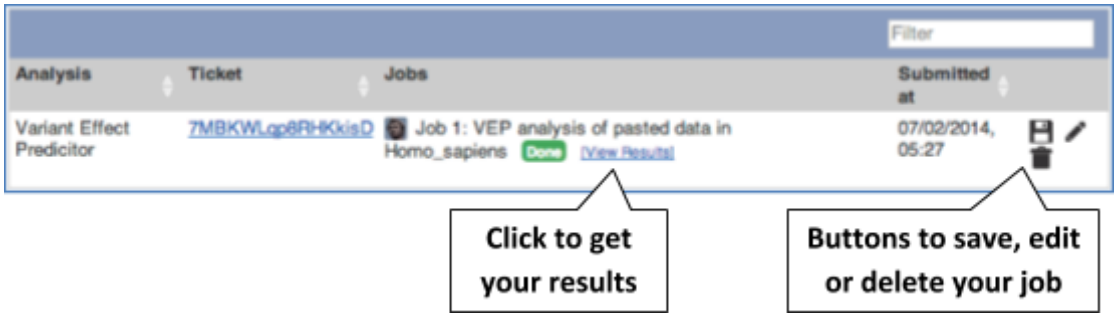






Hover over the options to see definitions.

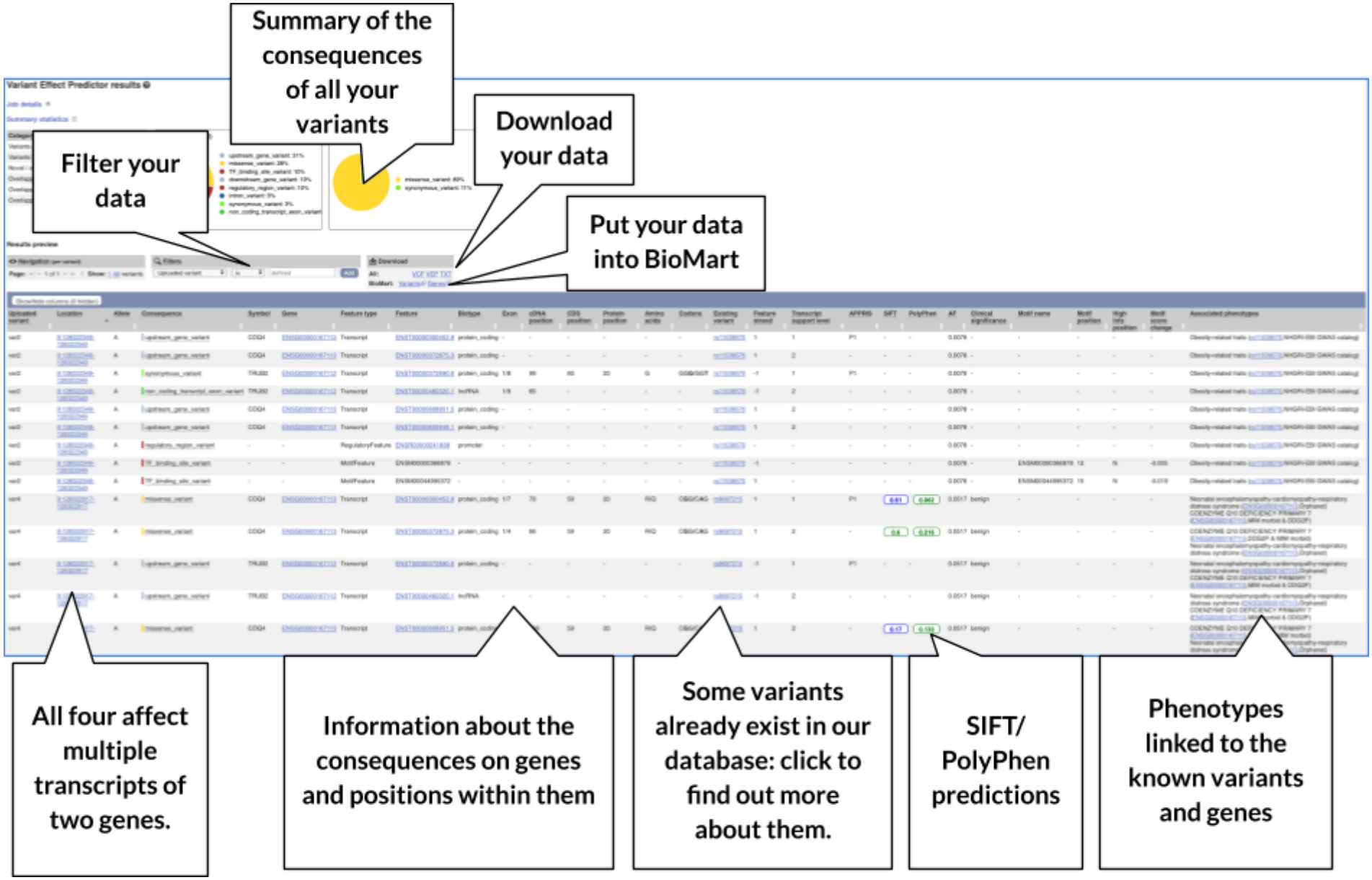
When you've selected everything you need, scroll right to the bottom and click Run.



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.


Summary of the consequences of all your variants
Download your data
Filter your data
Put your data into BioMart
All four affect multiple transcripts of two genes.
Information about the consequences on genes and positions within them
Some variants already exist in our database: click to find out more about them.
SIFT/ PolyPhen predictions
Phenotypes linked to the known variants and genes

Exercises

Exercise 1

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=32908738
```

(a) How many genes have been affected?

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

(c) Display your variant in the Ensembl browser.

Answers

Exercise 1

Go to www.ensembl.org and click on the link tools at the top of the page. Currently there are nine tools listed in that page. Click on Variant Effect Predictor and enter the three variants as below:

```
7     117530985  117530985  G/A
7     117531038  117531038  T/C
7     117531068  117531068  T/C
```

Note: Variation data input can be done in a variety of formats. See more details here http://www.ensembl.org/info/docs/variation/vep/vep_formats.html

Click Run.

When your job is listed as Done, click View Results.

You will get a table with the consequence terms from the Sequence Ontology project (http://www.sequenceontology.org/) (i.e. synonymous, missense, downstream, intronic, 5'

UTR, 3' UTR, etc) provided by VEP for the listed SNPs. You can also upload the VEP results as a track and view them on Location pages in Ensembl. SIFT and PolyPhen are available for missense SNPs only. For two of the entered positions, the variations have been predicted to have missense consequences of various pathogenicity (coordinate 117531038 and 117531068), both affecting *CFTR*. All the three variants have been already described and are known as rs1800078, rs1800077 and rs35516286 in dbSNP and other sources (databases, literature, etc).

Exercise 2 – viewing structural variants with the VEP

(a) Give your data a name, such as Patient deletion.
Paste 13 32307062 sv1 . <DEL> . . SVTYPE=DEL;END=32908738
into the Paste data field then hit Run.

13 genes have been affected.

(b) Use the Filters, selecting Consequence is transcript_ablation, Add.

Yes, there is deletion of complete transcripts of *PDS5B*, *N4BP2L1*, *BRCA2*, *RNY1P4*, *IFIT1P1*, *ATP8A2P2*, *N4BP2L2*, *N4BP2L2-IT2*, *AL137247.1*, *AL137247.2* and *AL138820.1*.

(c) To view your variant in the browser click on the location link in the results table 13: 32307062-32908738.
The link will open the Region in detail view in a new tab. If you have given your data a name it will appear automatically in red. If not, you may need to Configure this page and add it under Personal Data.