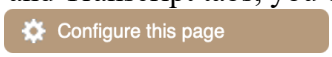
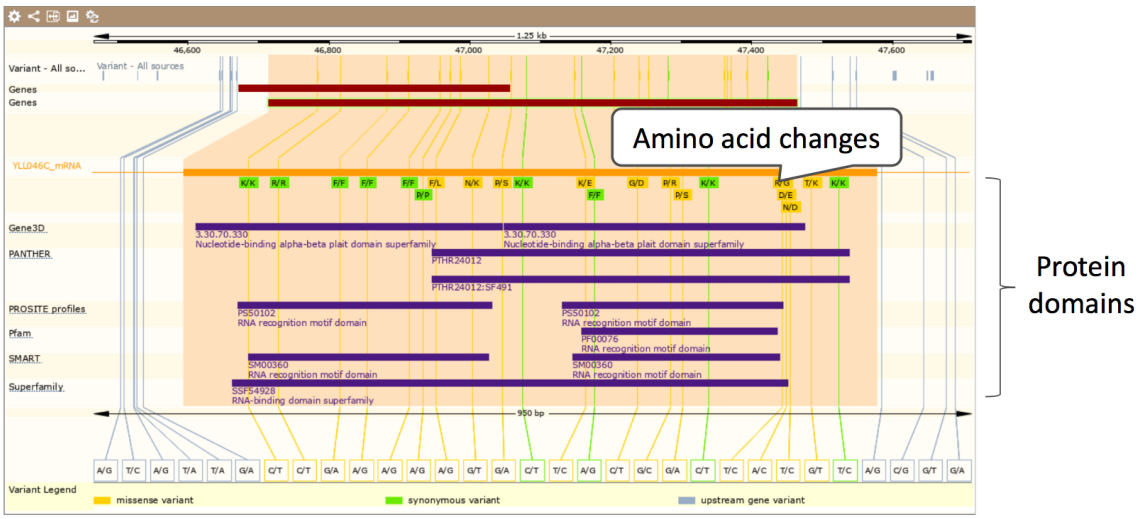
Exercise: Exploring variants in Ensembl Fungi

In any of the sequence views shown in the Gene and Transcript tabs, you can view variants on the sequence. You can do this by clicking on 

from any of these views.

Let's take a look at the Gene sequence view for *RNP1* (Gene Stable ID: YLL046C). This gene is a ribonuclease protein in *Saccharomyces cerevisiae*. Select Saccharomyces cerevisiae on the Ensembl Fungi homepage, search for YLL046C and go to the Variant image view.

This view shows variants mapped to the gene structure and protein domains.



We can examine all variants and filter to see ones we are interested in using the variant table. Click on the Variant table link.

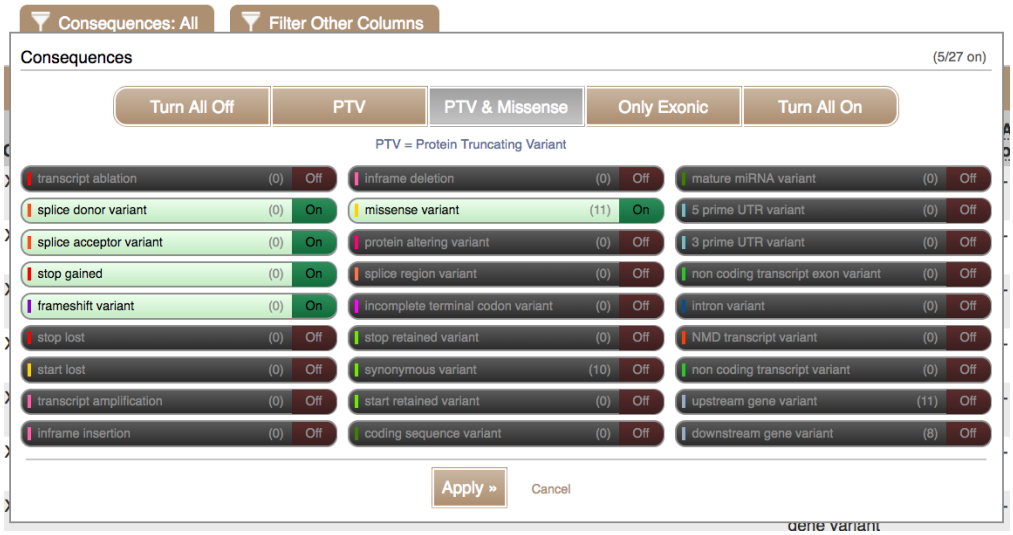
This table shows the variants in order of their occurrence through the genome, and they are reported on the forward strand. The gene *RNP1* is located on the reverse strand, so we are first shown variants downstream of the gene (starting a the 3' downstream region).

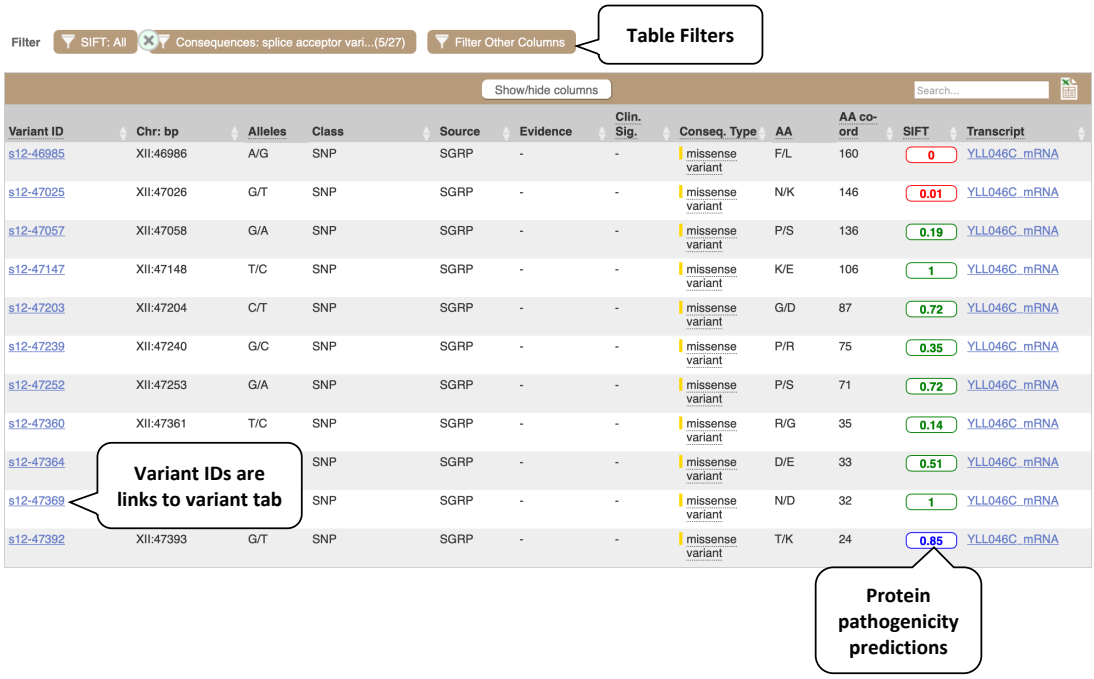
(a) How many variants in this gene are predicted to be missense?

You can filter the table to view variants that alter the protein sequence. Click on the Consequences: All button above the table. Click the option 'PTV and Missense' in the pop-up, then Apply. You can also filter by other columns such as variant Class.

(b) Are there any known variants in this gene predicted to be deleterious?

The SIFT scores predict the consequence of the variant on the function of the protein taking into account chemical changes and conservation of amino acids. Scores <0.05 and coloured red are ‘deleterious’ while scores >0.05 and coloured green are tolerated.





Let’s have a look at a specific variant. Click on the top result in the filtered table, or search for s12-46985. This will open up the variation tab.

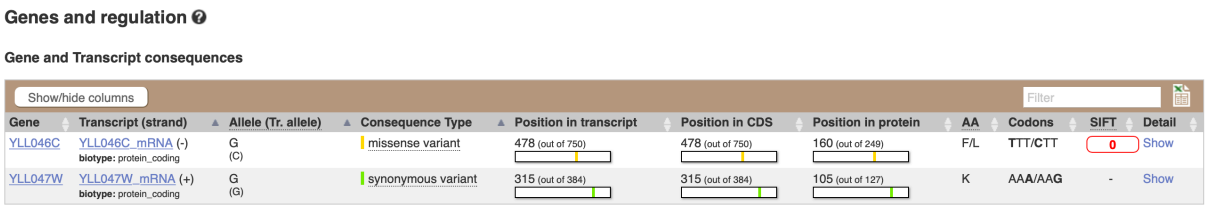
The icons show you what information is available for this variant.

(c) What are the genomic coordinates of this variant?

(d) What is the reference allele? (*Hint: Ensembl always reports alleles on the forward strand. Reference allele is given first.*)

(e) How many genes are affected by this variant? Does it have the same consequence across different transcripts of different genes?

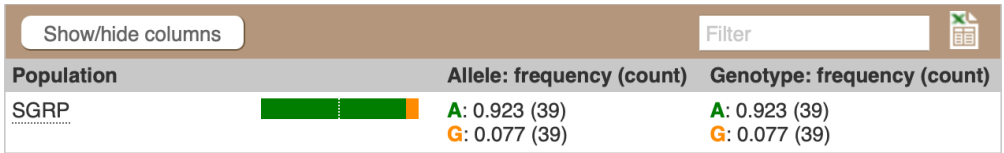
Click on Genes and regulation, or follow the link at the left.



This variant overlaps two genes. It causes a change in the protein sequence (missense variant) in the YLL046C gene we were looking at, but it is a synonymous variant in the other gene, not causing a change in the amino acid. Only missense variants have SIFT scores, hence why this is missing for the other gene.

(f) Which allele is major in the SGRP study?

Click on Genotype frequency in the left-hand menu. Note that the reference allele is more frequent than alternative allele in this case.



Additional Exercise – Variation data in *Fusarium oxysporum*

(a) Select the Fusarium oxysporum genome and search for FOXG_13574T0 gene. One of its upstream variants is SNP tmp_10_6610. What are the possible alleles for this polymorphic position? Which one is on the reference genome?

tmp_10_6610 SNP

Most severe consequence	**upstream gene variant** I See all predicted consequences
Alleles	**C/T** I Highest population MAF: **0.15**
Location	Chromosome **10:6610** (forward strand) I VCF: `10 6610 tmp_10_6610 C T`
HGVS name	10:g.6610C>T
External Links	
Original source	
About this variant	This variant overlaps 4 transcripts and has 10 sample genotypes.

(b) What is the most frequent allele at this position? How many heterozygous individuals were observed in the melonis population?

Show/hide columns Filter

Population	Allele: frequency (count)		Genotype: frequency (count)		
melonis	**C**: 0.850 (17)	**T**: 0.150 (3)	**CIC**: 0.800 (8)	**CIT**: 0.100 (1)	**TIT**: 0.100 (1)

(c) Which individuals have got genotypes C|T and T|T?

Show/hide columns Filter

Sample (Male/Female/Unknown)	Genotype (forward strand)	Population(s)	Father	Mother
886599 (U)	CIC	melonis	-	-
889404 (U)	CIC	melonis	-	-
889405 (U)	CIC	melonis	-	-
889406 (U)	CIC	melonis	-	-
889407 (U)	CIC	melonis	-	-
889408 (U)	CIC	melonis	-	-
889410 (U)	CIC	melonis	-	-
909453 (U)	CIC	melonis	-	-
909454 (U)	CIT	melonis	-	-
909455 (U)	TIT	melonis	-	-

Exercise: The Ensembl Fungi Variant Effect Predictor (VEP)

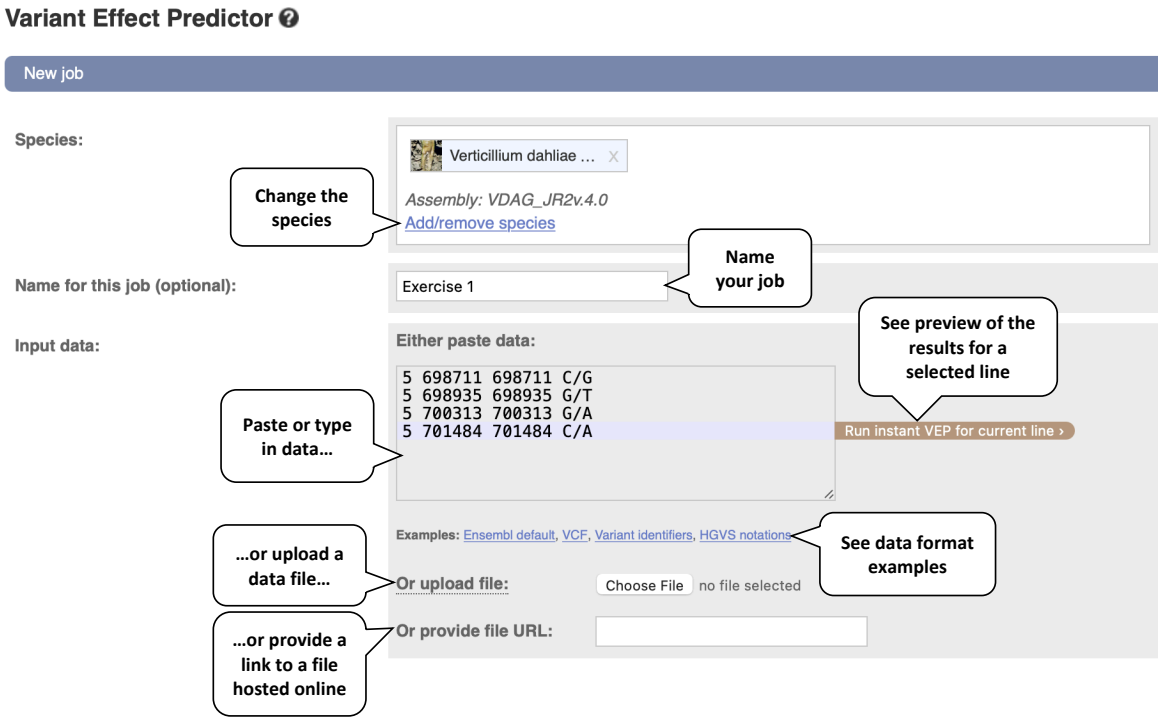
We have identified four variants in *Verticillium dahliae* JR2: chromosome 5, C->G at 698711, G->T at 698935, G->A at 700313 and C->A at 701484.

Use the Ensembl VEP to determine:
(a) Are your variants novel or have they already been annotated in Ensembl?
(b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?

Click on Tools in the top brown bar from any Ensembl Fungi page, then Variant Effect Predictor to open the input form. You will need to change the species to *Verticillium dahliae* JR2 and paste your input data in the provided text box.

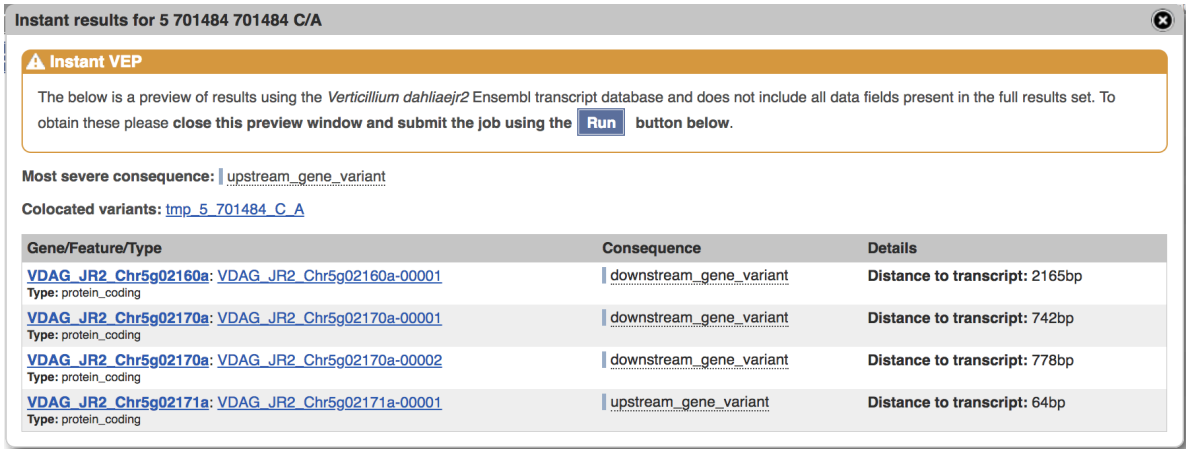
The VEP recognises a number of input formats including the Ensembl default format, VCF, Variant identifiers and HGVS notations.

The Ensembl default format is composed of four compulsory columns and additional 'strand' column: Chromosome, Start Position, End Position, Alleles (reference/alternate), Strand (1 for forward; -1 for reverse), with ne line per variant. Your variants in this format would look like this:
5 698711 698711 C/G
5 698935 698935 G/T
5 700313 700313 G/A
5 701484 701484 C/A

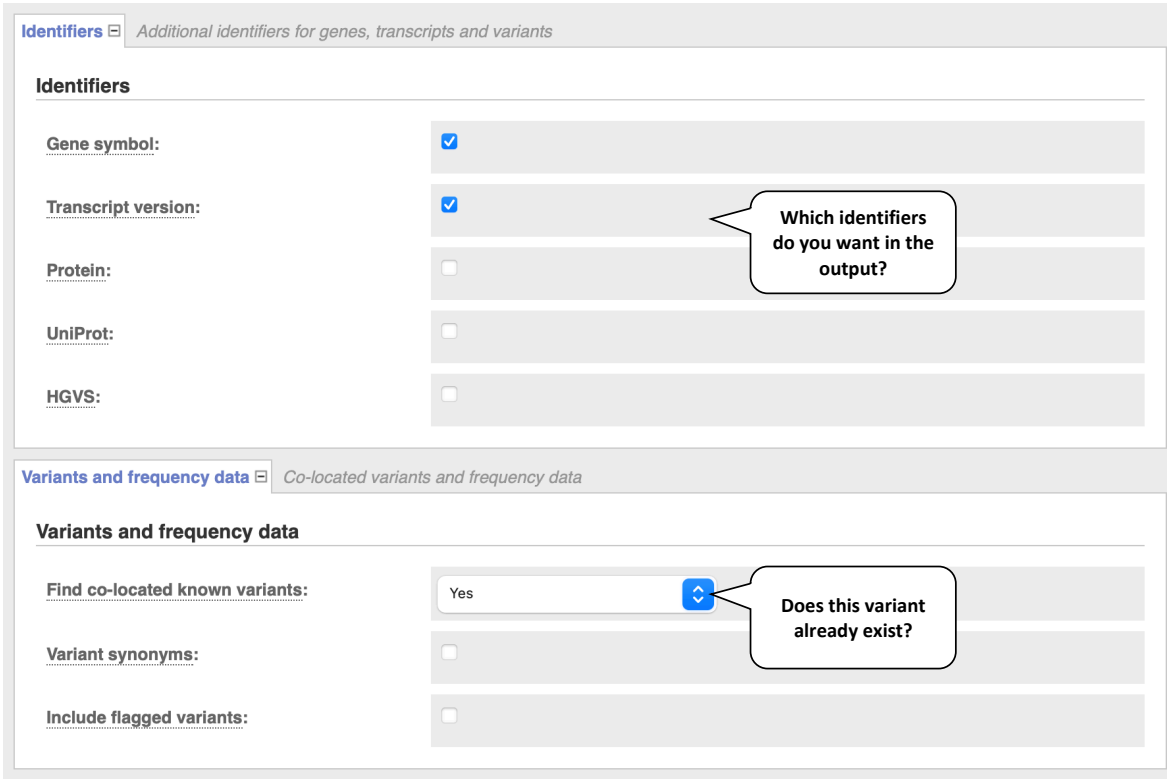


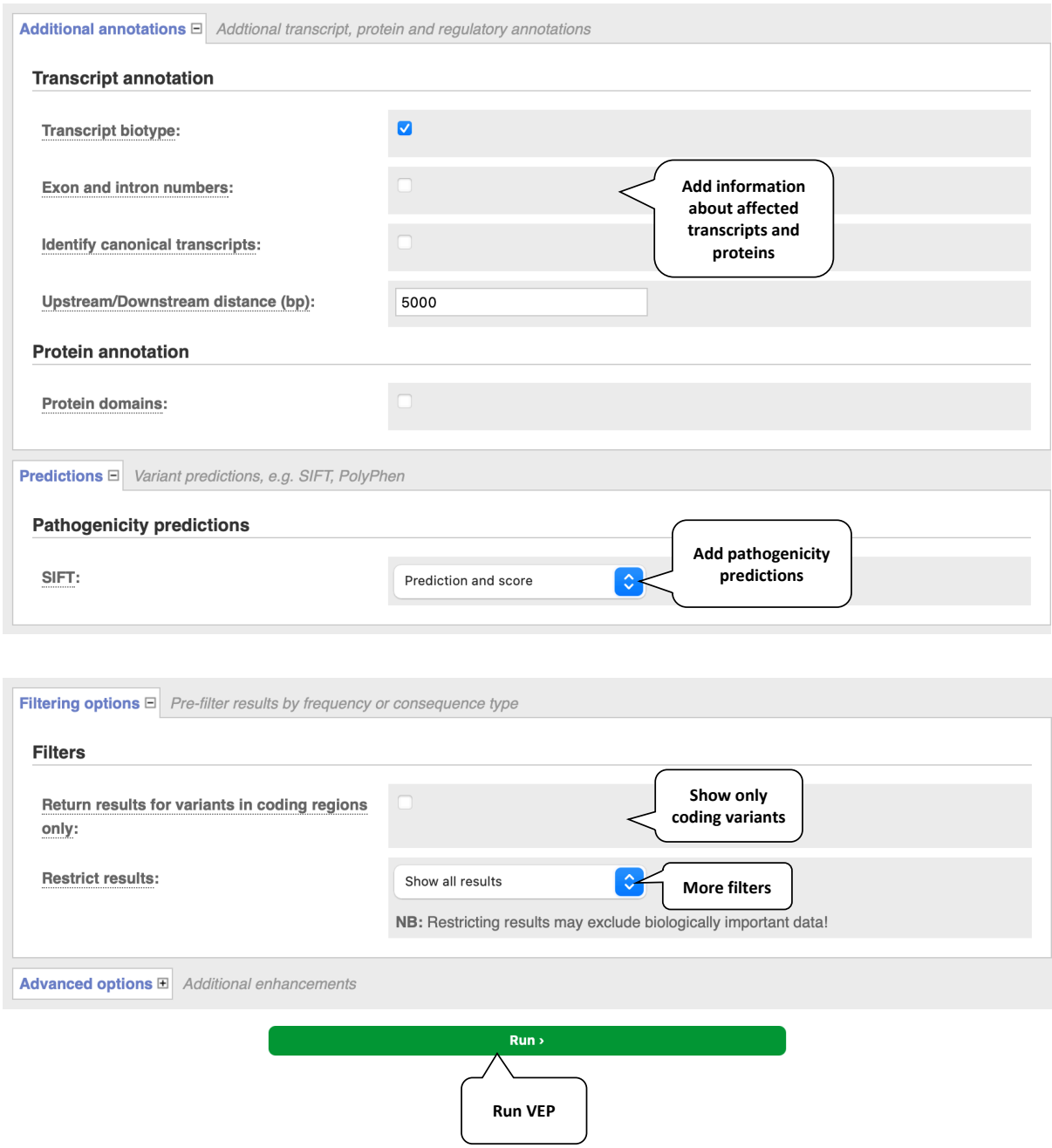
The VEP will automatically detect that the data is in Ensembl default format.

Clicking on the ‘Run instant VEP for current line’ will generate a pop-up with summarised results for that individual variant.

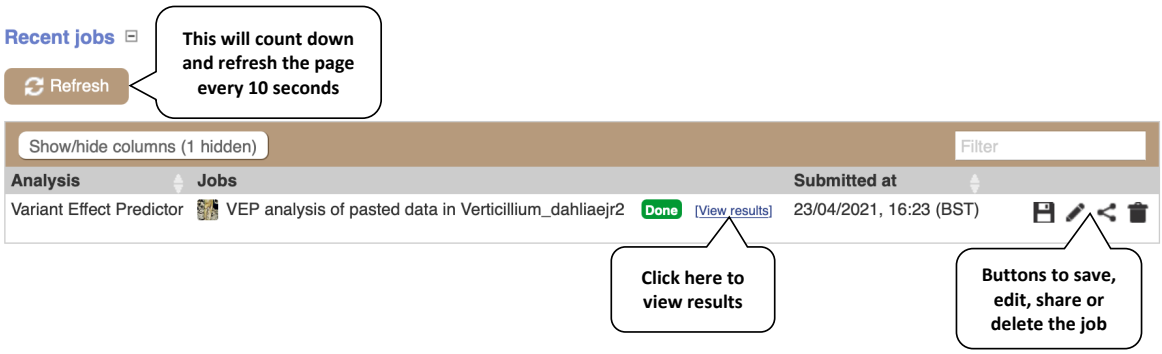


There are further options that you can choose for your output. These are categorised as Identifiers, Variants and frequency data, Additional annotations, Predictions, Filtering options and Advanced options. Let’s open all the menus and take a look.





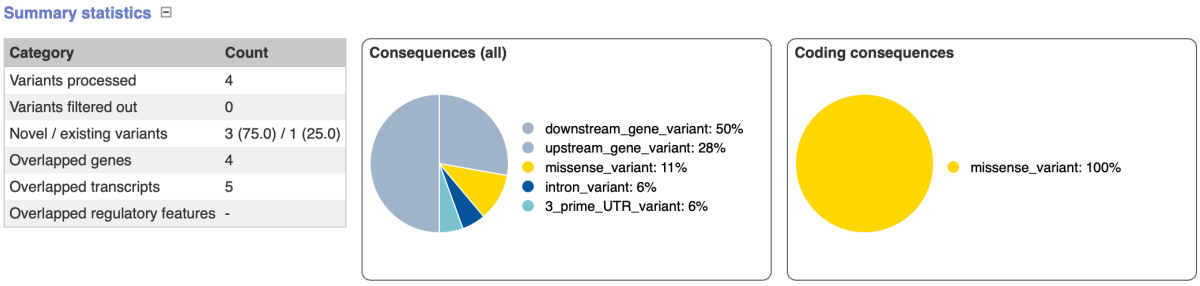
Hover over the options to see definitions. When you've selected everything you need, scroll right to the bottom and click Run.



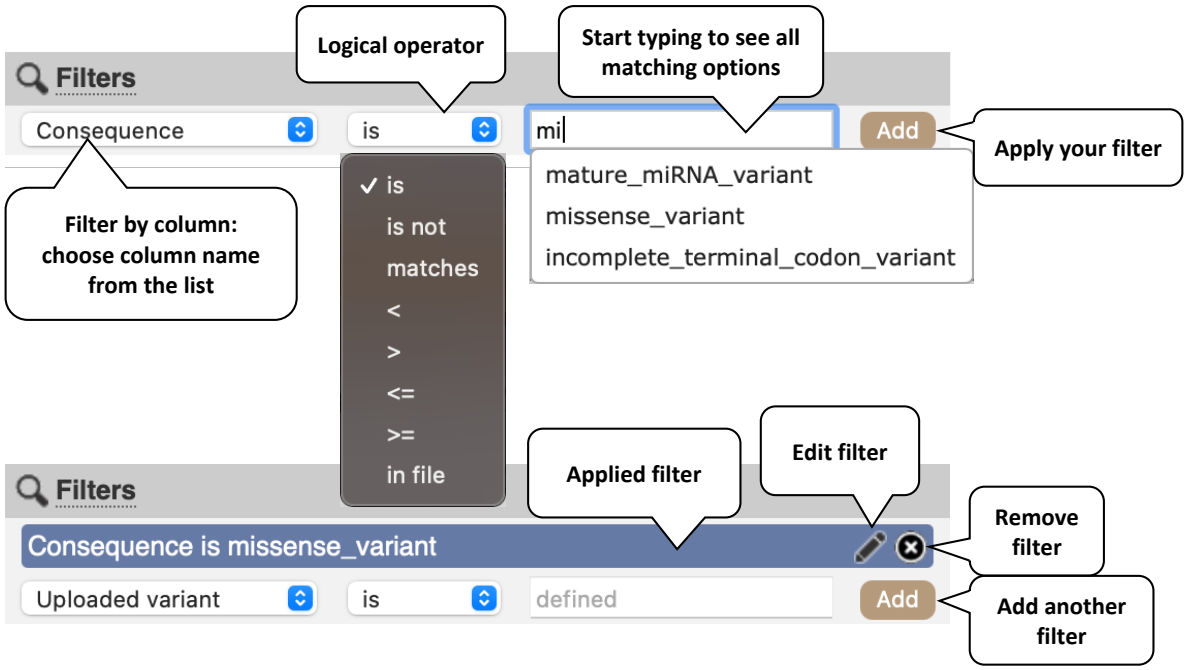
A table display will show you the status of your job. It will say Queued, then automatically switch to Done when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

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.

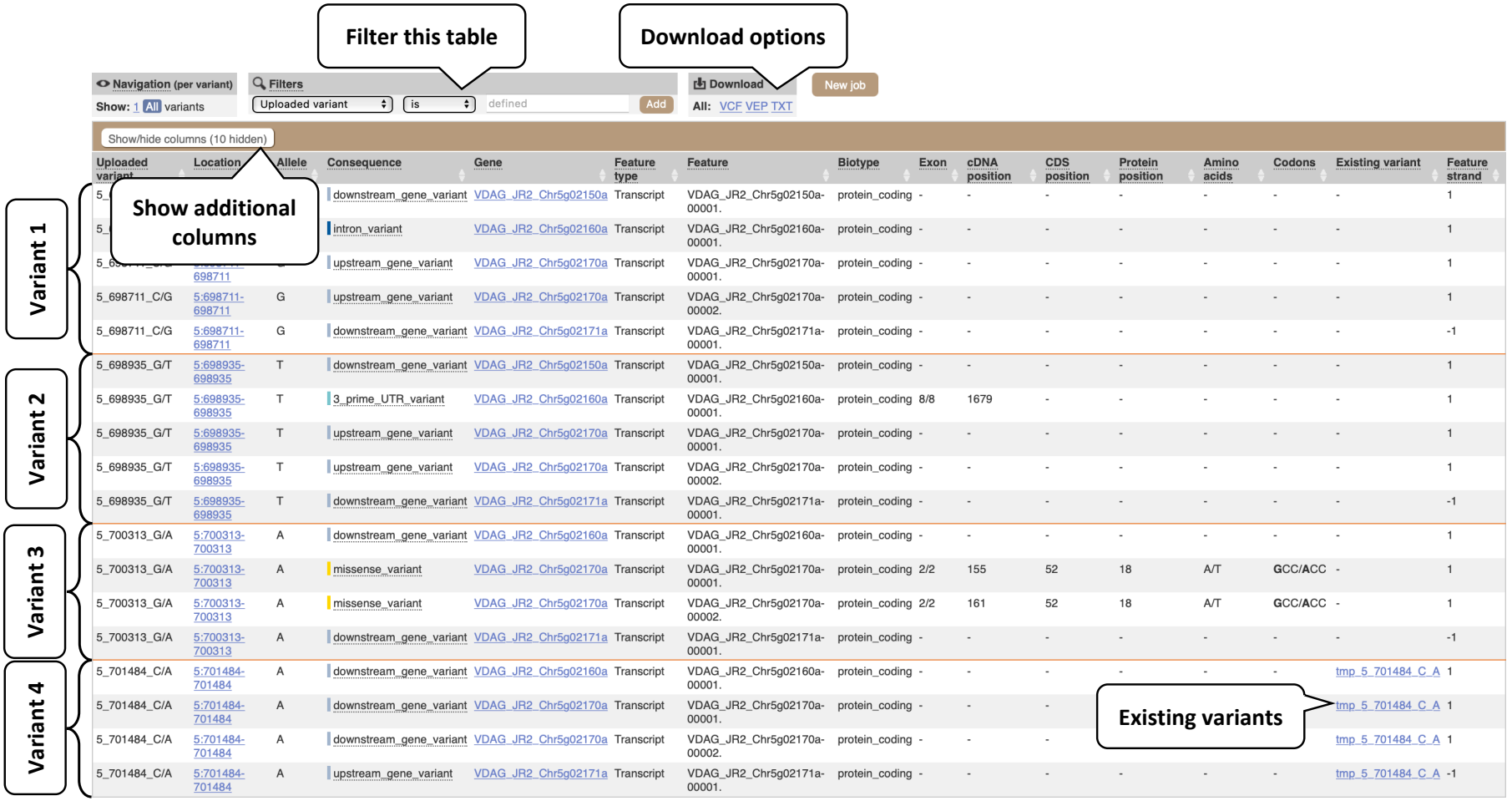
Let's come back to our questions:
(a) Are your variants novel or have they already been annotated in Ensembl?
(b) What genes are affected by your variants?
(c) Do any of your variants affect gene regulation?



The output table reports one variant consequence per row. If your variants have multiple alternate alleles, hit multiple genes or transcripts, you'll find few lines per variant. If the output table is large, you might want to use the filter option to narrow it down. Once you've added a filter, it will appear in the filter box, allowing you to add other filters.



Filter text box is by default set to 'defined', which can be used to filter out empty values, e.g. 'Existing variant' 'is' 'defined' will filter out variants with empty values in the 'Existing variant' column, leaving you with known variants only. Note that you should not type 'define' in the search box, just leave it as it is.



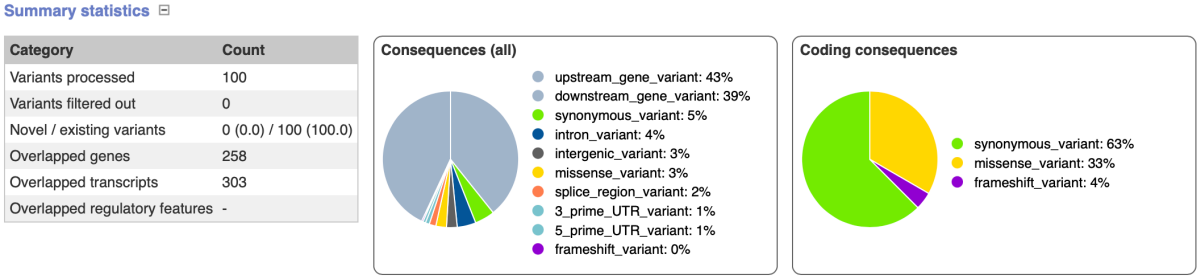
Uploaded variant	Location	Allele	Consequence	Gene	Feature type	Feature	Biotype	Exon	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Feature strand
5_			downstream_gene_variant	VDAG_JR2_Chr5g02150a	Transcript	VDAG_JR2_Chr5g02150a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_			intron_variant	VDAG_JR2_Chr5g02160a	Transcript	VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_	698711		upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00002.	protein_coding	-	-	-	-	-	-	-	1
5_698711_C/G	5:698711-698711	G	downstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript	VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_698935_G/T	5:698935-698935	T	downstream_gene_variant	VDAG_JR2_Chr5g02150a	Transcript	VDAG_JR2_Chr5g02150a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	3_prime_UTR_variant	VDAG_JR2_Chr5g02160a	Transcript	VDAG_JR2_Chr5g02160a-00001.	protein_coding	8/8	1679	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	upstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00002.	protein_coding	-	-	-	-	-	-	-	1
5_698935_G/T	5:698935-698935	T	downstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript	VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_700313_G/A	5:700313-700313	A	downstream_gene_variant	VDAG_JR2_Chr5g02160a	Transcript	VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	-	1
5_700313_G/A	5:700313-700313	A	missense_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00001.	protein_coding	2/2	155	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5:700313-700313	A	missense_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00002.	protein_coding	2/2	161	52	18	A/T	GCC/ACC	-	1
5_700313_G/A	5:700313-700313	A	downstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript	VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	-	-1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02160a	Transcript	VDAG_JR2_Chr5g02160a-00001.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00001.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	downstream_gene_variant	VDAG_JR2_Chr5g02170a	Transcript	VDAG_JR2_Chr5g02170a-00002.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	1
5_701484_C/A	5:701484-701484	A	upstream_gene_variant	VDAG_JR2_Chr5g02171a	Transcript	VDAG_JR2_Chr5g02171a-00001.	protein_coding	-	-	-	-	-	-	tmp_5_701484_C_A	-1

Additional Exercise: The Ensembl Fungi Variant Effect Predictor (VEP)

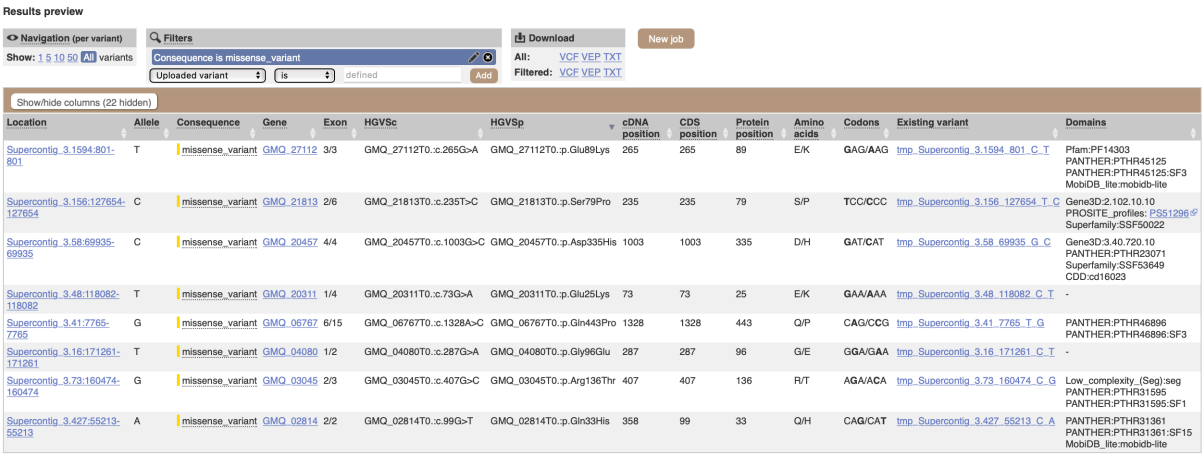
On the course file page, you will find a VCF file labelled VEP_exercise.vcf. This is a small subset of the outcome of *Puccinia graminis Ug99* whole genome sequencing and variant calling experiment. This file can also be found on our FTP site under the following link: http://ftp.ebi.ac.uk/pub/databases/ensembl/training/2021/FungalPathogens/VEP_exercise.vcf

Run the file through the VEP by downloading and uploading it from your computer, or alternatively by attaching it as a remote file hosted online (you will need to provide the FTP file URL).

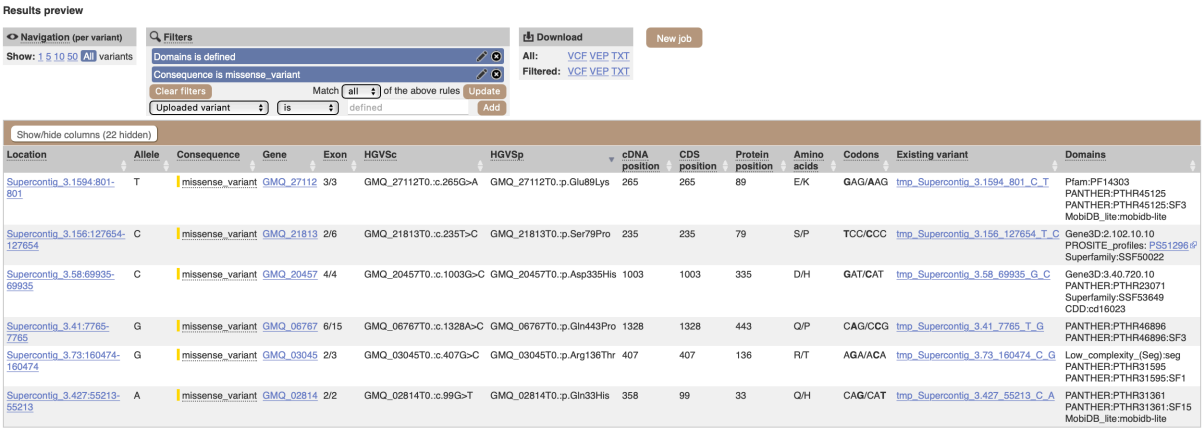
(a) How many variants have been processed?
(b) How many genes and transcripts are overlapped by variants in this file?



(c) Do any of the variants change the amino acid sequences of any proteins? What genes? What is the amino acid change? (*Hint: use the filters above the table to filter by consequences*.)

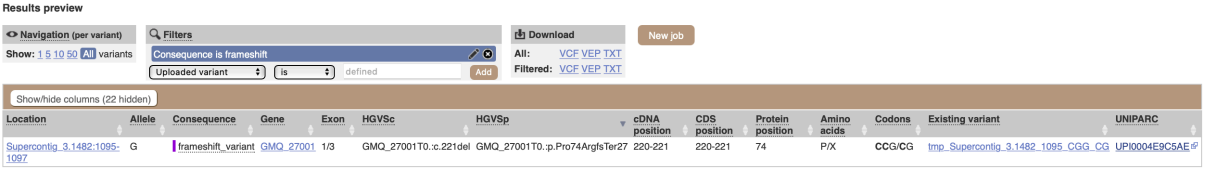


(d) What are the HGVSp notations of missense variants falling in known protein domains?



Location	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	Domains
Supercontig_3.1594:801-801	T	missense_variant	GMQ_27112	3/3	GMQ_27112T0.:c.265G>A	GMQ_27112T0.:p.Glu89Lys	265	265	89	E/K	**G**AG/**A**AG	tmp_Supercontig_3.1594_801_C_T	Pfam:PF14303 PANTHER:PTHR45125 PANTHER:PTHR45125:SF3 MobiDB_lite:mobidb-lite
Supercontig_3.156:127654-127654	C	missense_variant	GMQ_21813	2/6	GMQ_21813T0.:c.235T>C	GMQ_21813T0.:p.Ser79Pro	235	235	79	S/P	**T**CC/**C**CC	tmp_Supercontig_3.156_127654_T_C	Gene3D:2.102.10.10 PROSITE_profiles: PS51296 Superfamily:SSF50022
Supercontig_3.58:69935-69935	C	missense_variant	GMQ_20457	4/4	GMQ_20457T0.:c.1003G>C	GMQ_20457T0.:p.Asp335His	1003	1003	335	D/H	**G**AT/**C**AT	tmp_Supercontig_3.58_69935_G_C	Gene3D:3.40.720.10 PANTHER:PTHR23071 Superfamily:SSF53649 CDD:cd16023
Supercontig_3.41:7765-7765	G	missense_variant	GMQ_06767	6/15	GMQ_06767T0.:c.1328A>C	GMQ_06767T0.:p.Gln443Pro	1328	1328	443	Q/P	CA**G**/CC**G**	tmp_Supercontig_3.41_7765_T_G	PANTHER:PTHR46896 PANTHER:PTHR46896:SF3
Supercontig_3.73:160474-160474	G	missense_variant	GMQ_03045	2/3	GMQ_03045T0.:c.407G>C	GMQ_03045T0.:p.Arg136Thr	407	407	136	R/T	A**G**A/A**C**A	tmp_Supercontig_3.73_160474_C_G	Low_complexity_(Seg):seg PANTHER:PTHR31595 PANTHER:PTHR31595:SF1
Supercontig_3.427:55213-55213	A	missense_variant	GMQ_02814	2/2	GMQ_02814T0.:c.99G>T	GMQ_02814T0.:p.Gln33His	358	99	33	Q/H	CA**G**/CA**T**	tmp_Supercontig_3.427_55213_C_A	PANTHER:PTHR31361 PANTHER:PTHR31361:SF15 MobiDB_lite:mobidb-lite

(e) How many variants are frameshift? Which gene(s) do they fall in and which exons? Can you find a UniParc ID of protein(s) affected by this variant?



Location	Allele	Consequence	Gene	Exon	HGVSc	HGVSp	cDNA position	CDS position	Protein position	Amino acids	Codons	Existing variant	UNIPARC
Supercontig_3.1482:1095-1097	G	frameshift_variant	GMQ_27001	1/3	GMQ_27001T0.:c.221del	GMQ_27001T0.:p.Pro74ArgfsTer27	220-221	220-221	74	P/X	**CC**G/**C**G	tmp_Supercontig_3.1482_1095_CGG_CG	UPI0004E9C5AE