

## TABLE OF CONTENTS

<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>2.0 HOW TO INSTALL .....</b>	<b>1</b>
<b>3.0 HOW TO USE VEP.....</b>	<b>2</b>
<b>3.1 DATA INPUT .....</b>	<b>2</b>
<b>3.2 DATABASE TO USE .....</b>	<b>3</b>
<b>3.3 ANALYSIS.....</b>	<b>3</b>
<b>3.4 OUTPUT FILE INTERPRETATION .....</b>	<b>4</b>

### 1.0 Introduction

The Ensembl Variant Effect Predictor (VEP) is a powerful toolset for the analysis, annotation, and prioritization of genomic variants in coding and non-coding regions. It provides access to an extensive collection of genomic annotation, with a variety of interfaces to suit different requirements, and simple options for configuring and extending analysis. It is open source, free to use, and supports full reproducibility of results. The Ensembl Variant Effect Predictor can simplify and accelerate variant interpretation in a wide range of study designs.

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), see variant consequences
- Known variants that match yours, and associated minor allele frequencies from the 1000 Genomes Project
- SIFT and PolyPhen-2 scores for changes to protein sequence
- ... And more!

### 2.0 How to install

VEP has three different interfaces. Of all, web interface is the most user-friendly platform as it doesn't require any bioinformatic skill. To access to web interface, please click:

<https://asia.ensembl.org/Tools/VEP>

VEP interfaces

Web interface	Command line tool	REST API
 <ul style="list-style-type: none"><li>• Point-and-click interface</li><li>• Suits smaller volumes of data</li></ul> <a href="#">Documentation</a>	 <ul style="list-style-type: none"><li>• More options and flexibility</li><li>• For large volumes of data</li></ul> <a href="#">Documentation</a>	 <ul style="list-style-type: none"><li>• Language-independent API</li><li>• Simple URL-based queries</li></ul> <a href="#">Documentation</a>
	<a href="#">Clone from GitHub</a> <a href="#">Download (zip)</a> <a href="#">Pull Docker image from DockerHub</a>	<a href="#">VEP REST API</a>

### 3.0 How to use VEP

When you reach the VEP web interface, you will be presented with a form to enter your data and alter various options.

Species:  Assembly: GRCh38.p13  
[Add/remove species](#)  
If you are looking for VEP for Human GRCh37, please go to [GRCh37 website](#).

Name for this job (optional):

Input data:

Either paste data:

```
rs699  
rs171  
rs665
```

Run instant VEP for current line >

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

Or upload file:  No file chosen

Or provide file URL:

Transcript database to use:

☒ Ensembl/GENCODE transcripts  
☐ Ensembl/GENCODE basic transcripts  
☐ RefSeq transcripts  
☐ Ensembl/GENCODE and RefSeq transcripts

#### 3.1 Data input

First select the correct species for your data. Ensembl hosts many vertebrate genomes; genomes for plants, protists and fungi can be found at <http://ensemblgenomes.org/>. You can optionally choose a name for the data you upload - this can make it easier for you to identify jobs and files that you have uploaded to the VEP at a later point. You have three options for uploading your data:

- File upload - click the "Choose file" button and locate the file on your system

- Paste file - simply copy and paste the contents of your file into the large text box
- File URL - point the VEP to a file hosted on a publically accessible address. This can be either a http:// or ftp:// address.

Once you have uploaded some data, you can select it as the input for future jobs by choosing the data from the drop-down menu. The format of your data is automatically detected.

For pasted data you can get an instant preview of the results of your first variant by clicking the button that appears when you paste your data. This quickly shows you the consequence type, the IDs of any overlapping variants, genes, transcripts, and regulatory features, as well as SIFT and PolyPhen predictions. To see the full results set submit your job as normal.

### 3.2 Database to use

For some species you can select which transcript database to use. The default is to use Ensembl transcripts, which offer the richest annotation through VEP.

GENCODE Basic is a subset of the GENCODE gene set, and is intended to provide a simplified, high-quality subset of the GENCODE transcript annotations that will be useful to most users. GENCODE Basic includes all genes in the GENCODE gene set, with a representative subset of the transcripts (splice variants).

You can also select to use RefSeq transcripts from the other features database; note though that these transcripts are simply aligned to the reference genome and the database is missing much of the annotation found when using the main Ensembl database (e.g. protein domains, CCDS identifiers).

Should you wish to make adjustment on the additional configuration, please refer to <https://asia.ensembl.org/info/docs/tools/vep/online/input.html#ident>

### 3.3 Analysis

Once you have clicked "Run", your input will be checked and submitted to the VEP as a job. All jobs associated with your session or account are shown in the "Recent Tickets" table. You may submit multiple jobs simultaneously.

Show All entries		Show/hide columns (1 hidden)		Filter	
Analysis	Jobs			Submitted at	
Variant Effect Predictor	VEP analysis of pasted data in Bos_taurus	Done	[View results]	13/07/2015, 09:44	[Icons]
Variant Effect Predictor	VEP analysis of pasted data in Ovis_aries	Done	[View results]	08/07/2015, 13:19	[Icons]
Variant Effect Predictor	VEP analysis of pasted data in Ovis_aries	Failed		07/07/2015, 16:51	[Icons]

The "Jobs" column of the table shows the status of the job.

- Queued - your job is waiting to be submitted to the system
- Running - your job is currently running
- Done - your job is finished - click the [View results] link to be taken to the results page
- Failed - there is a problem with your job - click the magnifying glass icon to see more details

You may delete a job by clicking the trash can icon. If you are logged in to Ensembl, you can save the job by clicking the save icon. You may also resubmit a job (for example, to re-run with the same data but change some parameters) by clicking the edit icon. You can see a summary of the options that you selected for your VEP job by clicking on the magnifying glass icon.

### 3.4 Output file interpretation

The VEP presents a summary and a detailed results preview on its results page.

#### 3.4.1 Summary

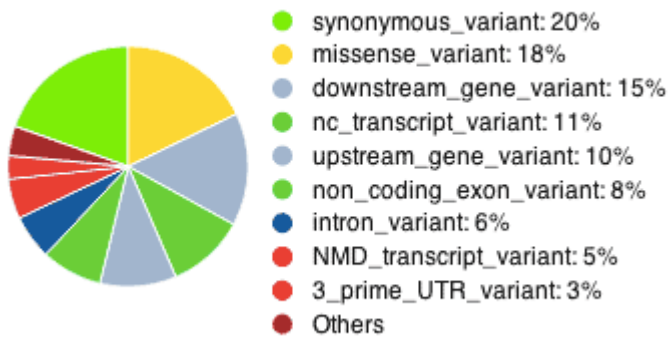
The summary panel on the VEP results page gives a brief overview of the VEP job, along with some basic statistics about the results. Various statistics are listed in a table, including:

- Variants processed - any variants not parsed by the VEP are not included in this count
- Variants remaining after filtering
- Novel / known variants - the number and percentage of novel variants vs existing variants in the input (see input page documentation)
- Number of overlapped genes, transcripts, and regulatory features

Category	Count
Variants processed	498955
Variants remaining after filtering	498955
Novel / known variants	-
Overlapped genes	825
Overlapped transcripts	2888
Overlapped regulatory features	7309

Pie charts are shown detailing the proportion of consequence types called across all variants in the results. The colour scheme of the pie chart matches the colours used to draw variants on the Ensembl region in detail view.

Consequences (all)



### 3.4.2 Results preview table

The results table shows one row per transcript and variant. By default, all the columns are shown; to temporarily hide columns, click the blue "Show/hide columns" button and select or deselect the columns you wish to view. The columns you select will be recalled when viewing other jobs.

Show/hide columns										
Uploaded variation	Location	Feature	Feature type	Consequence	CDS position	Protein position	Amino acids	Codons	SIFT	GMAF
rs116383664	1:1115461	ENSR00000528923	RegulatoryFeature	regulatory_region_variant	-	-	-	-	-	T:0.0137
rs116383664	1:1115461	ENST00000379317	Transcript	upstream_gene_variant	-	-	-	-	-	T:0.0137
rs116383664	1:1115461	ENST00000486379	Transcript	upstream_gene_variant	-	-	-	-	-	T:0.0137
rs116383664	1:1115461	ENST00000379289	Transcript	missense_variant	247	83	R/W	Cgg/Tgg	tolerated(0.06)	T:0.0137
rs116383664	1:1115461	ENST00000460998	Transcript	upstream_gene_variant	-	-	-	-	-	T:0.0137
rs116383664	1:1115461	ENST00000514695	Transcript	upstream_gene_variant	-	-	-	-	-	T:0.0137
rs116383664	1:1115461	ENST00000379290	Transcript	missense_variant	247	83	R/W	Cgg/Tgg	tolerated(0.06)	T:0.0137
rs116383664	1:1115461	ENST00000379288	Transcript	missense_variant	28	10	R/W	Cgg/Tgg	deleterious(0.03)	T:0.0137

The table can be sorted by any column - click the column header to toggle sorting behaviour. The default output format ("VEP" format when downloading from the web interface) is a 14-column tab-delimited file. Empty values are denoted by '-'. The output columns are:

- Uploaded variation - as chromosome\_start\_alleles
- Location - in standard coordinate format (chr:start or chr:start-end)
- Allele - the variant allele used to calculate the consequence
- Gene - Ensembl stable ID of affected gene
- Feature - Ensembl stable ID of feature
- Feature type - type of feature. Currently one of Transcript, RegulatoryFeature, MotifFeature.
- Consequence - consequence type of this variant
- Position in cDNA - relative position of base pair in cDNA sequence
- Position in CDS - relative position of base pair in coding sequence

- Position in protein - relative position of amino acid in protein
- Amino acid change - only given if the variant affects the protein-coding sequence
- Codon change - the alternative codons with the variant base in upper case
- Co-located variation - identifier of any existing variants. Switch on with --check\_existing
- Extra - this column contains extra information as key=value pairs separated by ";" For other field information, please refer to [https://asia.ensembl.org/info/docs/tools/vep/vep\\_formats.html#output](https://asia.ensembl.org/info/docs/tools/vep/vep_formats.html#output)

The VEP allows you to download either your full or filtered results set in a choice of data formats.

- VCF - VCF is a portable format for variant data. Consequence data is encoded as a series of delimited strings under the "CSQ" key in the VCF INFO field.
- VEP - The default VEP output format gives one row per variant and transcript overlap.
- TXT - Text format is a tab-delimited format, equivalent to what can be seen in the results table. Note that the columns you select to be visible in the table do not affect the downloaded file - all columns are outputted. This format is best if you intend to import the results into a spreadsheet program such as Microsoft Excel.

You can also send the genes or known variants in your current preview to BioMart. This allows you to easily retrieve any of BioMart's rich data associated with these genes (other database references, GO terms, orthologues/paralogues) and variants (phenotype annotations, synonyms, citations).

*-End-*