



# PO: Precision Oncology Course

## Variant Annotation using VEP



# Exercise

Annotation of the panel 1 using VEP webpage

# Study case

## Panel 1

**Tumor type:** Patient with Colon Adenocarcinoma

**Sequencing platform:** Illumina HiSeq2500

**Type of data:** Sequencing panel (paired). Ion Ampliseq Cancer Hotspot Panel v2 (46 genes)

**Samples:** Tumor with matched healthy tissue

**File with somatic variants from Mutect2:** Variants detected in tumor sample but not in the corresponding control

**Data:** <https://drive.google.com/file/d/1BknV7nyQDrUJ6LgAxx4ln8qVriUNI8-F/view?usp=sharing>

**Reference genome:** hg19



tumor\_passable\_filtered.vcf

# Steps

Run VEP from the web

1. Go to: <http://www.ensembl.org/info/docs/tools/vep/index.html>
2. Click on “Web interface”

## Ensembl Variant Effect Predictor (VEP)



**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! See [data types](#), [versions](#).

★ [What's new in release 106?](#)

### VEP interfaces

#### Web interface



- Point-and-click interface
- Suits smaller volumes of data

[Documentation](#)



#### Command line tool



- More options and flexibility
- For large volumes of data

[Documentation](#)

[Clone from GitHub](#)

[Download \(zip\)](#)

[Pull Docker image from DockerHub](#)

#### REST API



- Language-independent API
- Simple URL-based queries

[Documentation](#)

[VEP REST API](#)

# Steps

Run VEP from the web

3. Fill in a new job. We want the following annotations:

- HUGO gene symbol.
- The HGVS identifiers for coding DNA and protein.
- The Global Minor Allele Frequency of 1000 genomes project.
- gnomAD frequencies.

4. You can add any other annotation you want.

**HINTS:** Remember to use the same assembly used in the variant detection. **Further info:** <http://www.ensembl.org/info/docs/tools/vep/online/input.html>

# Questions

30 min

- **How many variants were in the VCF file?**
- **How many of them are not known in the database?**
- **How many genes and transcripts are affected by the variants?**
- **Is there any regulatory region overlapping some variant?**
- **Which is the most represented consequence category?**

# Answers

30 min

- How many variants were in the VCF file? **5**
- How many of them are not known in the database? **0**
- How many genes and transcripts are affected by the variants? **8 genes and 30 transcripts**
- Is there any regulatory region overlapping variants? **No**
- Which is the most represented consequence category?  
**Missense (44%)**

# Questions

30 min

- **Which is the most represented coding sequence consequence?**
- **How many variants fall in a coding region in some gene?**
- **What do the HGVS identifiers represent in each case?**
- **Is there any clear polymorphism within the data?**



# Answers

30 min

- Which is the most represented coding sequence consequence? **Missense (82%)**
- How many variants fall in gene coding regions? **5**
- What do the HGVS identifiers represent in each case?  
**The coding (HGVSc) and protein (HGVSp) changes**
- Is there any clear polymorphism within the data? **No**  
**(gnomAD or AF  $\geq$  0.01)**

# Answers

Variants falling in coding regions

Possible answers:

- Identify the variants that fall in a **CDS** position.
- Filter by **consequence** (suggested by Monica).

Keep in mind that VEP can return **several rows for the same variant**, so you have to **count the unique location + allele** combinations.

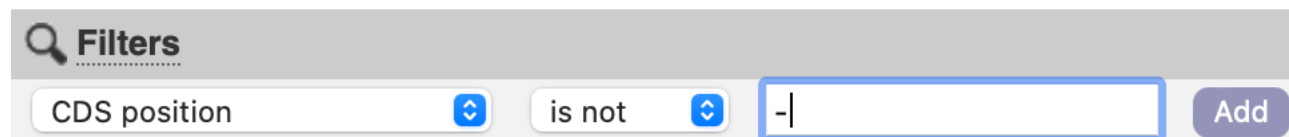
# Answers

Variants falling in coding regions

Identify the variants that fall in a **CDS** position

**CDS (Coding DNA Sequence):** Portion of a gene that codes for a protein. The field “CDS position” in VEP indicates the CDS position in which the variant appears. If the variant doesn’t appear within a CDS, the field will be empty (“-”).

So we can keep just the rows with a CDS position  $\neq$  “-“:



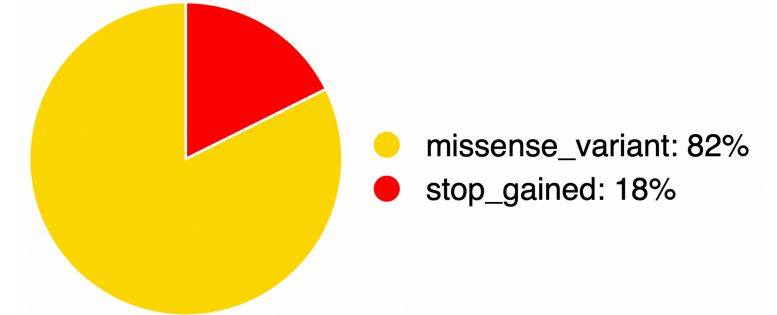
The screenshot shows a 'Filters' interface with a search icon and the title 'Filters'. Below the title, there is a filter rule being constructed. It consists of a dropdown menu with 'CDS position' selected, followed by a blue double-headed arrow icon. Then, there is a text input field containing 'is not', followed by another blue double-headed arrow icon. To the right of this is a text input field containing a hyphen '-' character. Finally, there is a blue button labeled 'Add'.

# Answers

Variants falling in coding regions

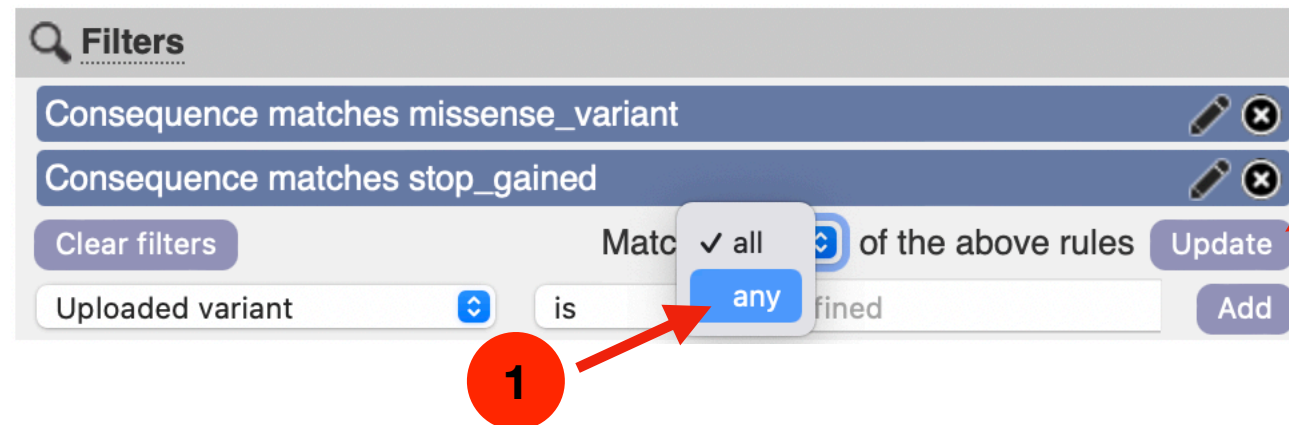
Filter by consequence (suggested by Monica)

Coding consequences



**We have 2 coding consequences: missense and stop gained**

So we can keep just the rows with consequence = “missense\_variant” **OR** consequence = “stop\_gained”:



# Steps

Download the file

1. Save the file in VCF format.
2. Check that the following annotations have been added to the INFO field:

- Allele
- Consequence
- Symbol
- Gene
- Feature type
- Feature
- HGVSc
- HGVSg
- cDNA position
- Protein position
- Amino acids
- Codons
- Existing variant
- AF
- gnomAD AF
- gnomAD NFE AF



# Thanks!

