GULBENKIAN CIÊNCIA

PO: Precision Oncology Course

Variant Annotation using VEP

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer

Exercise

Annotation of the panel 1 using VEP webpage

Study case

Panel 1

tumor_passlable_filtered.vcf

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

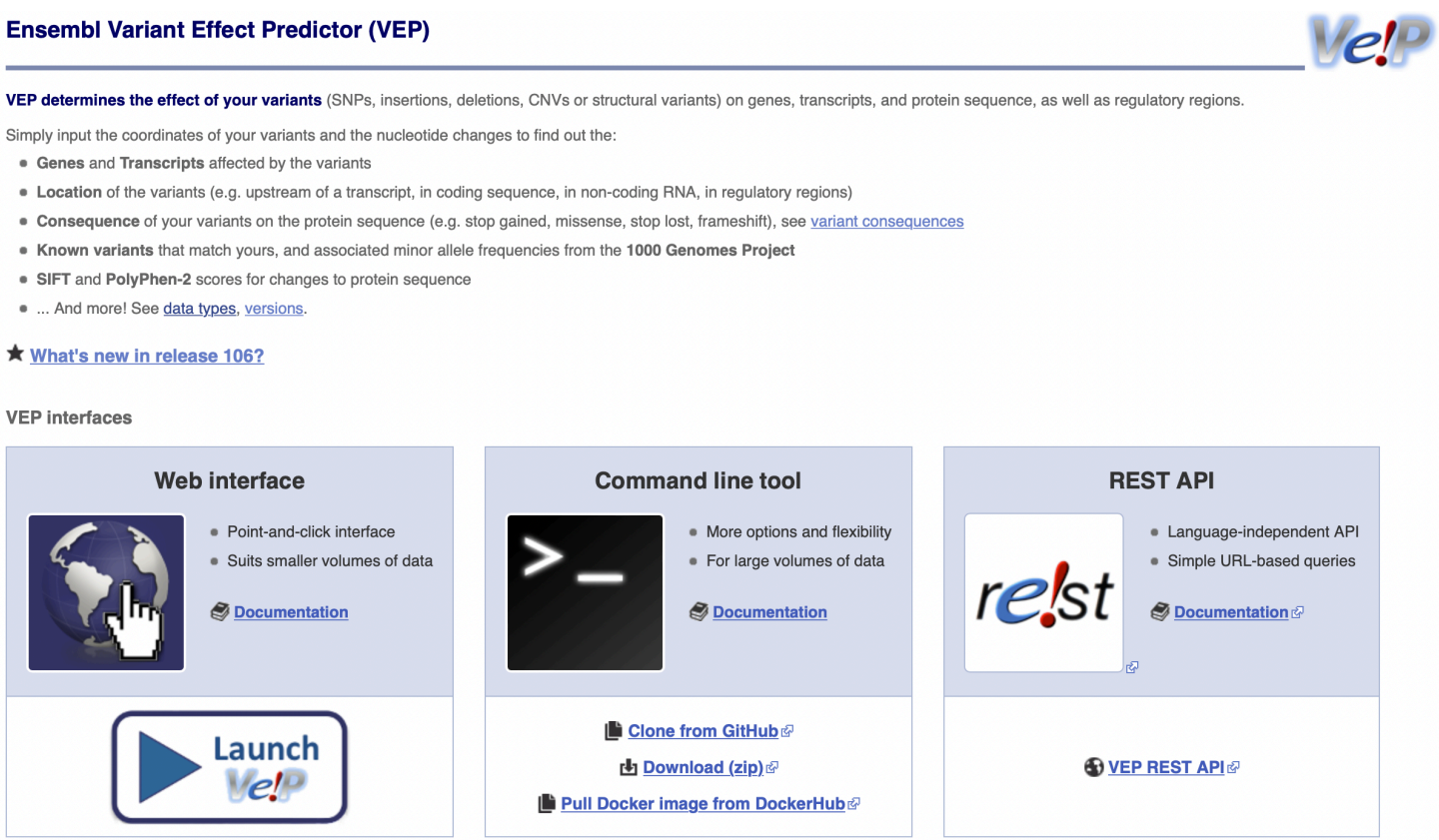
Data: https://drive.google.com/file/d/1BknV7nyQDrUJ6LgAxh4In8qVriUNI8-F/view?usp=sharing

Reference genome: hg19

Steps

Run VEP from the web

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



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 >= 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.

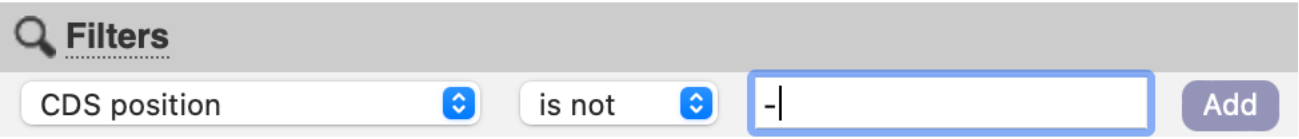
Answers

Variants falling in coding regions

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

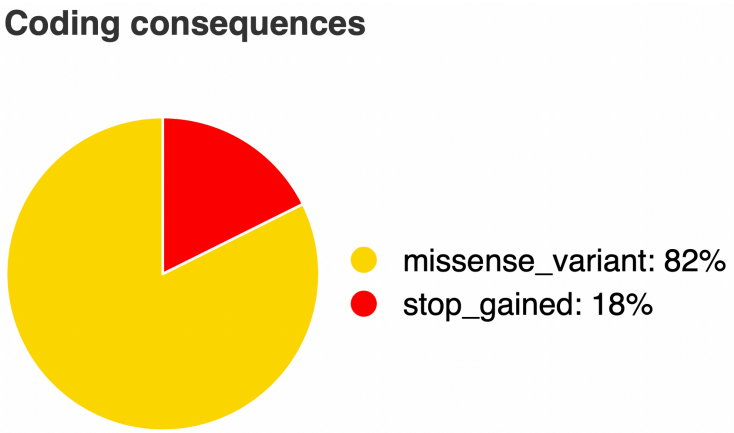
CDS (Coding DNA Sequence): Portion of a gene that codes for a protein. The field “CDS position” in VEP indicates the CDS position in which the variant appears. If the variant doesn’t appear within a CDS, the field will be empty (“-“).

So we can keep just the rows with a CDS position ≠ “-“:



Answers

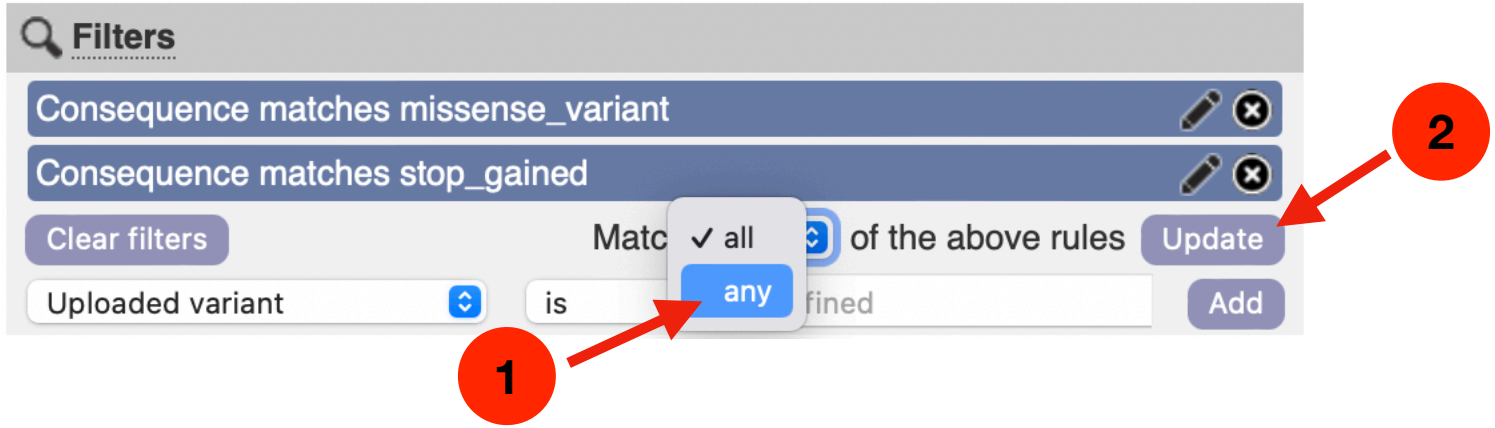
Variants falling in coding regions



Filter by consequence (suggested by Monica)

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
- HGVSp
- cDNA position
- Protein position
- Amino acids
- Codons
- Existing variant
- AF
- gnomAD AF
- gnomAD NFE AF

GULBENKIAN CIÊNCIA

Thanks!

BU CNIO BIOINFORMATICS UNIT

cnio stop cancer