



Annotation and filtering: hands on

Precision Oncology Course



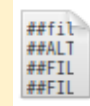
CNIO **BIOINFORMATICS UNIT**

Coral Fustero Torre
Bioinformatics Unit,
Structural Biology Programme.
cfustero@cnio.es | bioinformatics.cnio.es

Exercise 1: Annotations with VEP

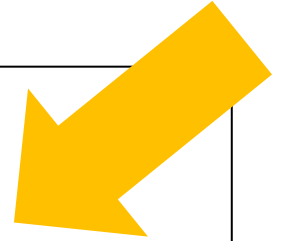
Tumor type: hepatic met – colon adenocarcinoma

- Sequencing: Illumina
- Panel: HiSeq 2500
- Tumor – control paired sample
- File with somatic variants: variants detected in tumor sample but not in the corresponding control
- Assembly version: hg19



tumor_passable_filtered.vcf

File with somatic variants that have passed the defined filters



Link to exercise directory: <https://drive.google.com/drive/folders/1T4PLkl4wgzxdjf-USfvv7gl8W5gW4Ypx?usp=sharing>

Exercise 1: Annotations with VEP

Execution of VEP through the web page

1. Go to:

<http://www.ensembl.org/info/docs/tools/vep/index.html>

2. We want to obtain the annotations using the ensembl transcript set

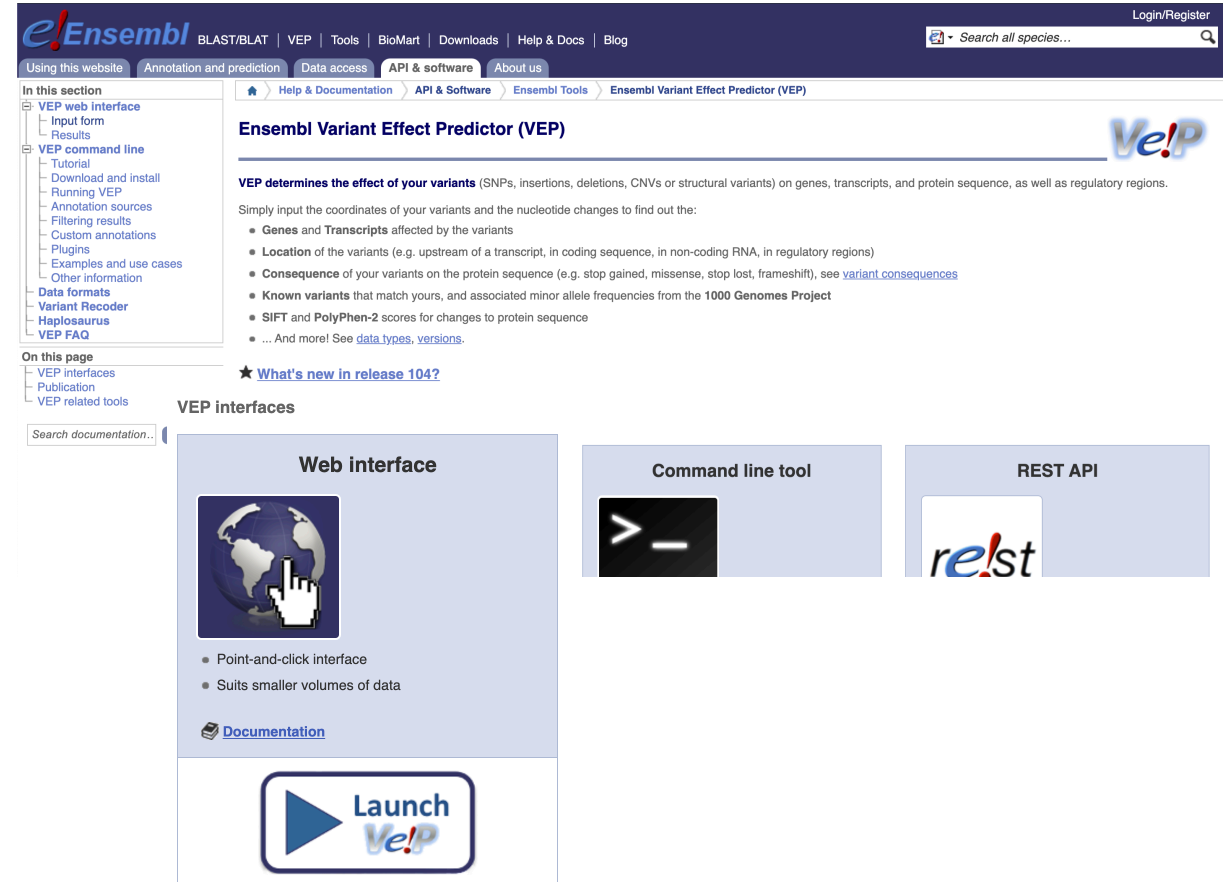
3. We want in the output the following annotations:

- HUGO gene symbol
- The HGVS identifiers for coding DNA and protein
- The Global Minor Allele Frequency of 1000 genomes project
- gnomAD frequencies
- SIFT and PolyPhen prediction and score
- Condel prediction and score

4. You can choose any other parameters to explore the results

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>



The screenshot displays the Ensembl Variant Effect Predictor (VEP) website. The top navigation bar includes links for BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. A search bar is located on the right. The main content area is titled 'Ensembl Variant Effect Predictor (VEP)' and features a 'Ve!P' logo. Below the title, a brief description states: '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:'. A bulleted list follows, detailing the types of annotations provided: Genes and Transcripts affected, Location of variants, Consequence of variants on the protein sequence, Known variants from the 1000 Genomes Project, SIFT and PolyPhen-2 scores, and more. A 'What's new in release 104?' link is also present. The 'VEP interfaces' section is divided into three columns: 'Web interface' (featuring a globe icon, a list of features, and a 'Launch VEP' button), 'Command line tool' (featuring a terminal icon), and 'REST API' (featuring the 're!st' logo).

Exercise 1: Annotations with VEP

Answer the following questions

- How many variants were in the input 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?
- Which is the most represented coding sequence consequence?
- How many variants fall in a coding region in some gene?
- What do the HGVS identifiers mean in each case?
- Does the prediction tool agree in the prediction of functional impact?
- Is there any clear polymorphism within the data?

... and explore the results

Exercise 1: Annotations with VEP

Download the file

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

- Consequence
- Existing_variation
- Feature
- PolyPhen
- Condel
- SIFT
- SYMBOL
- Protein_position
- Amino_acids
- HGVSc
- HGVSp
- AF
- CDS_position
- Allele
- Gene
- Feature_type
- cDNA_position
- Codons
- gnomAD_AF
- gnomAD_NFE_AF
- ExAC

Exercise 2: Understanding the annotations

Tumor type: hepatic met – colon adenocarcinoma

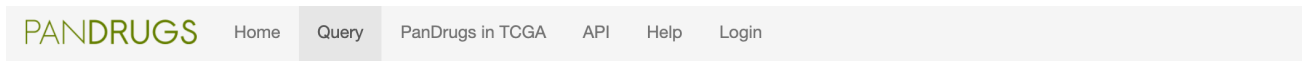
- Sequencing: Illumina
- Panel: HiSeq 2500
- Tumor – control paired sample
- File with somatic variants: variants detected in tumor sample but not in the corresponding control
- Assembly version: hg19

Link to exercise file:

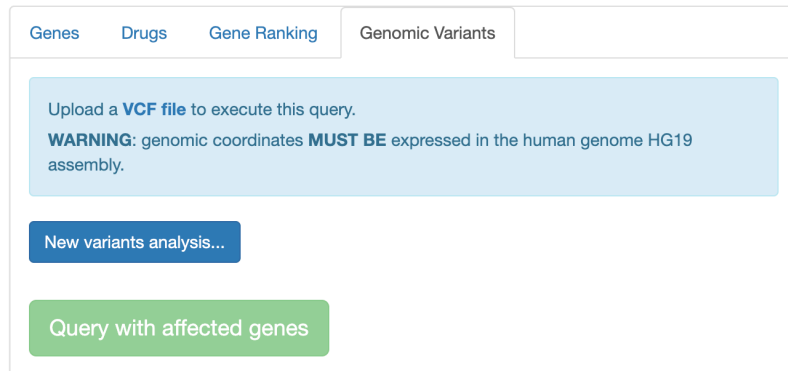
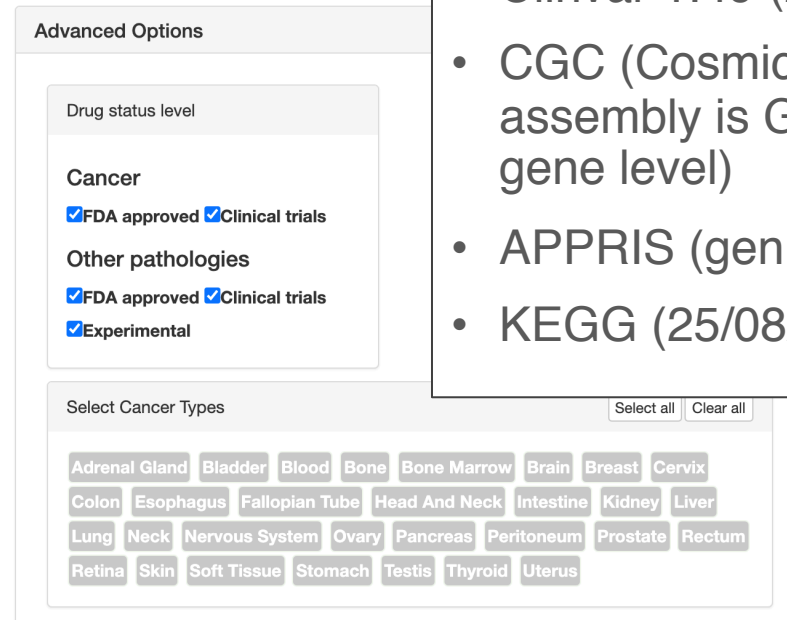
<https://drive.google.com/drive/folders/1T4PLkl4wgzxdjf-USfvv7gl8W5gW4Ypx?usp=sharing>

Exercise 2: Understanding the annotations

1. Go to: pandrugs.org/
 2. Upload the vcf file @ Genomic variants section
 3. Download the results
- This might take some time...***



Query PanDrugs

The image shows the 'Query PanDrugs' form. It has a tabbed interface with 'Genes', 'Drugs', 'Gene Ranking', and 'Genomic Variants'. The 'Genomic Variants' tab is selected. Inside the form, there is a light blue box with instructions: 'Upload a VCF file to execute this query. WARNING: genomic coordinates MUST BE expressed in the human genome HG19 assembly.' Below this is a blue button labeled 'New variants analysis...'. At the bottom is a green button labeled 'Query with affected genes'.The image shows the 'Advanced Options' form. It has a section for 'Drug status level' with 'Cancer' and 'Other pathologies' sub-sections. Under 'Cancer', there are checkboxes for 'FDA approved' (checked), 'Clinical trials' (checked), and 'Experimental' (checked). Under 'Other pathologies', there are checkboxes for 'FDA approved' (checked), 'Clinical trials' (checked), and 'Experimental' (checked). Below this is a 'Select Cancer Types' section with a grid of buttons for various cancer types: Adrenal Gland, Bladder, Blood, Bone, Bone Marrow, Brain, Breast, Cervix, Colon, Esophagus, Fallopian Tube, Head And Neck, Intestine, Kidney, Liver, Lung, Neck, Nervous System, Ovary, Pancreas, Peritoneum, Prostate, Rectum, Retina, Skin, Soft Tissue, Stomach, Testis, Thyroid, and Uterus. There are 'Select all' and 'Clear all' buttons at the top right of the grid.

Databases versions

- Cosmic Release v82 - hg19
- Pfam 31.0 (Mar 2017)
- UniProt release 2017_07 (28/08/2017)
- InterPro 64.0 (28/08/2017)
- Clinvar 1.49 (26/08/2017)
- CGC (Cosmic v82) → The corresponding assembly is GRCH38 (but we search at gene level)
- APPRIS (gen19.ensembl74 29/08/2017)
- KEGG (25/08/2017)

Exercise 2: Understanding the annotations



Selection of principal isoform:

PRINCIPAL:1 - Transcript(s) expected to code for the main functional isoform based solely on the core modules in the APPRIS database

PRINCIPAL:2 - Where the APPRIS core modules are unable to choose a clear principal variant (approximately 25% of human protein coding genes), the database chooses two or more of the CDS variants as "candidates" to be the principal variant

PRINCIPAL:3 - Where the APPRIS core modules are unable to choose a clear principal variant and more than one of the variants have distinct CCDS identifiers, APPRIS selects the variant with lowest CCDS identifier as the principal variant

PRINCIPAL:4 - Where the APPRIS core modules are unable to choose a clear principal CDS and there is more than one variant with a distinct (but consecutive) CCDS identifiers, APPRIS selects the longest CCDS isoform as the principal variant

PRINCIPAL:5 - Where the APPRIS core modules are unable to choose a clear principal variant and none of the candidate variants are annotated by CCDS, APPRIS selects the longest of the candidate isoforms as the principal variant

REST (ALTERNATIVE:1 (Candidate transcript(s) models that are conserved in at least three tested non-primate species), ALTERNATIVE:2 (Candidate transcript(s) models that appear to be conserved in fewer than three tested non-primate species), NO LABEL (Non-candidate transcripts are not flagged and are considered as "MINOR" transcripts))

Reduced to relevant isoforms if PRINCIPAL

Exercise 2: Understanding the annotations

Open the results file and check the results

Try to answer the following questions:

- Which fields indicate polymorphisms?
- Which fields have information about the effect in the sequence?
- Which fields have information about the effect in the protein?
- Which fields give specific information about the pathology under study?
- In which processes are involved APC and FBXW7 gene?
- Is the gene KRAS frequently mutated in the same tumor type?
- Which variant has been reported more times in tumors?
- Should ATM gene be inhibited?
- Name 3 candidates as relevant variants in the disease



Thanks!

Credits for many class materials to:

Héctor Tejero: htejero@cnio.es

Elena Piñeiro: epineiro@cnio.es

Javier Perales-Patón: jperales@cnio.es



CNIO BIOINFORMATICS UNIT