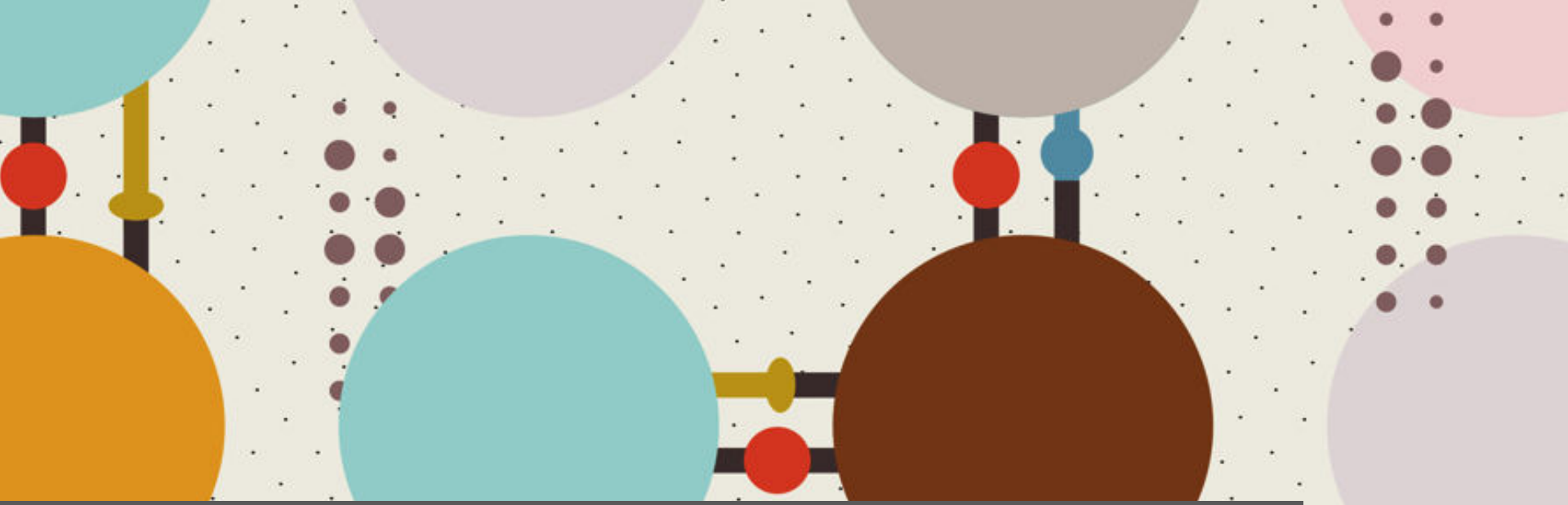


IO17 | Large Scale Bioinformatics for Immuno-Oncology

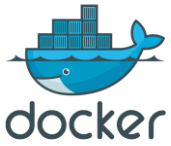
Neoantigens: exercise 2 - Solution

Francesca Finotello, Federica Eduati, and Pedro L. Fernandes

GTPB | The Gulbenkian Training Programme in Bioinformatics
Instituto Gulbenkian de Ciência, Oeiras, Portugal | Sept 19th-22nd, 2017

INSTITUTO GULBENKIAN DE CIÊNCIA

Exercise 2: variant annotation



To use TIminer, Docker must be running on your computer!

We have access to VCF file reporting the somatic DNA mutations predicted for Patient_1 using the GRCh37/hg19 human genome annotation:

Patient_1_mutations.vcf

Predict the proteins affected by the mutations with TIminer (function TIminerAPI.**executeVep**).

Note: use „Patient_1“ as subject ID and /ader/databases/vep as cache directory.

Once you get the results, answer the questions related to this exercise at:

https://b.socrative.com/login/student/

Variant annotation: Python code

```
from TIminer import TIminerAPI

TIminerAPI.executeVep(inputFile="../Input/Patient_1_mutations.vcf",
    subject="Patient_1",
    outputFile="../Output/Patient_1_VEP_37_mutations.txt",
    mutatedSeqOutputFile="../Output/Patient_1_VEP_37_proteins.txt",
    cacheDir="/ader/databases/vep",
    genomeVersion=37)
```

Variant annotation: questions

1) How many mutations from Patient 1 were annotated by VEP (Exercise 2)?

127

Variant annotation: questions

2) According to the variant annotation performed by VEP on the mutational data from Patient 1 (Exercise 2), what is the **HGNC symbol of the gene** affected by the mutation at position 205,902,150 in chromosome 1?

#Uploaded_variation	Patient_1
Location	**1:205902150**
Allele	A
Gene	ENSG00000174502
Feature	ENST00000367134
Feature_type	Transcript
Consequence	missense_variant
cDNA_position	302
CDS_position	188
Protein_position	63
Amino_acids	P/L
Codons	cCc/cTc
Existing_variation	-
Extra	IMPACT=MODERATE;STRAND=-1;SYMBOL=**SLC26A9**;SYMBOL_SOURCE=HGNC;HGNC_ID=14469;ENSP=ENSP00000356102;TREMBL=B1AVM8_HUMAN;UNIPARC=UPI000013DF98

Variant annotation: questions

3) According to the variant annotation performed by VEP on the mutational data from Patient 1 (Exercise 2), what is the **Ensemble ID of the protein** affected by the mutation at position 205,902,150 in chromosome 1?

#Uploaded_variation	Patient_1
Location	**1:205902150**
Allele	A
Gene	ENSG00000174502
Feature	ENST00000367134
Feature_type	Transcript
Consequence	missense_variant
cDNA_position	302
CDS_position	188
Protein_position	63
Amino_acids	P/L
Codons	cCc/cTc
Existing_variation	-
Extra	IMPACT=MODERATE;STRAND=-1;SYMBOL=SLC26A9;SYMBOL_SOURCE=HGNC;HGNC_ID=14469;ENSP=**ENSP00000356102**;TREMBL=B1AVM8_HUMAN;UNIPARC=UPI000013DF98

Variant annotation: questions

4) According to the variant annotation performed by VEP on the mutational data from Patient 1 (Exercise 2), what are the first 80 bases of the protein affected by the mutation at position 205,902,150 in chromosome 1?

Hint: in R, you can measure the length of a string with the **nchar** function

>ENSP00000356102.2:p.Pro63Leu
MSQPRPRYVVDRAAYSLTLFDDEFEKKDRTYPVGEKLRNAFRCSSAKIKAVVFGLLPVLSWLLKYKIKDYIIPDLLGGLS
GGSIQVPQGMAFALLANLPAVNGLYSSFFPLLTYFFLGGVHQMVPGTFAVISILVGNICLQLAPESKFQVFNNATNESYV
DTAAMEAERLHVSATLACLTAIIQMGLGFMQFGFVAIYLSESFIRGFMTAAGLQILISVLKYIFGLTIPSYTGPGSIVFT
FIDICKNLPHTNIASLIFALISGAFLVLVKELNARYMHKIRFPIPTEMIVVVVATAISGGCKMPKKYHMQIVGEIQRGFP
TPVSPVVSQWKDMIGTAFSLAIVSYVINLAMGRTLANKHGYDVDSNQEMIALGCSNFFGSFFKIHVICCALSVTLAVDGA
GGKSQVASLCVSLVVMITMLVLGIYLYPLPKSVLGALIAVNLKNSLKQLTDPYYLWRKSKLDCCIWVVSFLSSFFLSLPY
GVAVGVAFSVLVVVFQTQFRNGYALAQVMDTDIYVNPKTYNRAQDIQGIKIITYCSPLYFANSEIFRQKVIAKTGMDPQK
VLLAKQKYLKKQEKRRMRPTQQRRSLFMKTKTVSLQELQQDFENAPPTDPNNNQTPANGTSVSYITFSPDSSSPAQSEPP
ASAEAPGEPSDMLASVPPFVTFHTLILDMSGVSFVDLMGIKALAKLSSTYGKIGVKVFLVNIHAQVYNDISHGGVFEDGS
LECKHVFPSIHDAVLFAQANARDVTPGHNFQGAPGDAELSLYDSEEDIRSYWDLEQEMFGSMFHAETLTALESLSAAGGC
YPYRSESLVSPLFTRQALAAMDKPPAHSTPPTSALSLAAEGHLDFQLLRVSQKQKDKYNCAGLLYKLQKVSQSPHGSVSD
GVRLSRT

Variant annotation: questions

5) What are instead the first 80 bases of the **wild type** (i.e. non-mutated) protein?

#Uploaded_variation	Patient_1
Location	**1:205902150**
…	…
Protein_position	**63**
Amino_acids	**P/L**
…	…
Extra	…;ENSP=**ENSP00000356102**;…

>ENSP00000356102.2

MSQPRPRYVVDRAAYSLTLF**DDEFEKKDRT**YPVGEKLRNA**FRCSSAKIKA**VVFGLLPVLS**WLPKYKIKDY**IIPDLLGGLS

…

Variant annotation: questions

6) Re-run variant annotation for Patient 1, but using the GRCh38/hg20 genome annotation. How many mutations are annotated by VEP?

Only **2**!

```
from TIminer import TIminerAPI

TIminerAPI.executeVep(inputFile="../Input/Patient_1_mutations.vcf",
    subject="Patient_1",
    outputFile="../Output/Patient_1_VEP_38_mutations.txt",
    mutatedSeqOutputFile="../Output/Patient_1_VEP_38_proteins.txt",
    cacheDir="/ader/databases/vep",
    genomeVersion=38)
```