

# Precision medicine: NGS variant analysis and interpretation for translational research

## Exercise 3: Prioritization and variant filtering

Fátima Al-Shahrour ● Javier Perales ● Elena Piñeiro

September 28, 2016

# Additional annotations and prioritization

## **VEP\_parser**

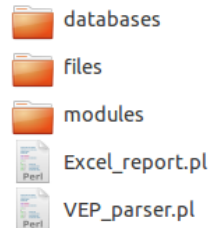
- Adds annotations from additional data sources.
  - Changes vcf format into a tabular format.
  - Formatting and filtering annotations in which we are interested.
- Computes a score for each variant. Cancer oriented.

} Parsing

# Execution: Command line and files

## Command line execution

### Directory structure

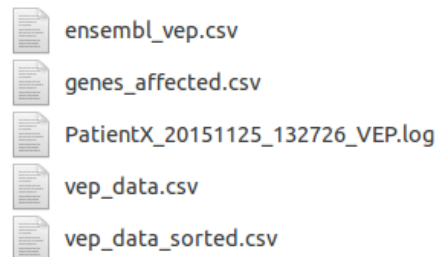


```
epineiro@epineiro:~/ExerciseDay2/VEP_parser$ perl VEP_parser.pl

--vepfile=filename or -f=filename      Input file containing results of VEP from Ensembl analysis. Mandatory.
--output=directory or -o=directory     Execution output dir. Default ./output.
--sample=code or -sp=code              Tumor code. Mandatory.
--root=rootname or -r=rootname         Root name for the PCDA output files. Default none.
--cfile=boolean or -v=boolean          Use of filter consequences file VEP_conseq.csv or not. Default TRUE.
--databases=directory or -d=directory  Absolute path to databases directory. Mandatory.

i.e. VEP_parser.pl -f=file.vcf -o=/home/user/z13-222 -sp=z13-222 -r=analysis -v=TRUE -d=/home/user/VEP_parser/databases
```

### Output files



### vep\_data\_sorted.csv

chr	loc	mut	gene	feature	feature_type	consequence	principal	poly_effect	poly_score
3	178921553	T/A	ENS G00000121879	ENS T00000468036	Transcript	downstream_gene_variant			
3	178921553	T/A	ENS G00000121879	ENS T00000477735	Transcript	downstream_gene_variant			
3	178921553	T/A	ENS G00000121879	ENS T00000263967	Transcript	missense_variant	PRINCIPAL:1	probably_damaging	1
4	1808898	-/C	ENS G00000068078	ENS T00000469068	Transcript	downstream_gene_variant			
4	1808898	-/C	ENS G00000068078	ENS T00000474521	Transcript	downstream_gene_variant			
4	1808898	-/C	ENS G00000068078	ENS T00000507588	Transcript	downstream_gene_variant			
4	1808898	-/C	ENS G00000068078	ENS T00000481110	Transcript	frameshift_variant		inferred	1
4	1808898	-/C	ENS G00000068078	ENS T00000352904	Transcript	frameshift_variant		inferred	1
4	1808898	-/C	ENS G00000068078	ENS T00000412135	Transcript	frameshift_variant		inferred	1
4	1808898	-/C	ENS G00000068078	ENS T00000260795	Transcript	frameshift_variant	PRINCIPAL:3	inferred	1
4	1808898	-/C	ENS G00000068078	ENS T00000440486	Transcript	frameshift_variant	PRINCIPAL:3	inferred	1
4	1808898	-/C	ENS G00000068078	ENS T00000340107	Transcript	frameshift_variant	ALTERNATIVE:2	inferred	1

# Excel generator



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))

Flag possible artifacts:

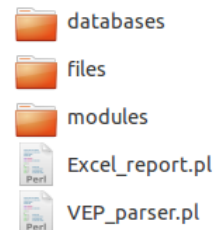
Germline variants defined as somatic - Filtered in normal sample

Very low coverage variants

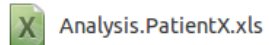
# Execution Excel generator

## Command line execution

### Directory structure



### Output file



```
epineiro@epineiro:~/ExerciseDay2/VEP_parser$ perl Excel_report.pl

--output=directory      Execution output dir. Default ./output.
--case=name             Tumor case name. Mandatory.
--tsample=name          Tumor sample name. Mandatory.
--nsample=name          Normal sample name. Optional.
--vepfile1=path         VEP file absolute path. Mandatory.
--vepfile2=path         VEP file for bothTC calling absolute path. Optional.
--vepfile3=path         VEP file for MuTect calling absolute path. Optional.
--rubioseq=path         RUBioSeq output path. Mandatory.
--mutect=path           MuTect output path. Mandatory if --vepfile3.

i.e. perl IonTorrent_report.pl --output=/home/user/z13-222 --case=z13-222 --tsample=z13-222 --nsample=c13-222 --vepfile1=/home/user/z13-222/z13-222_20130826_104703_VEP/vep_data_sorted.csv --vepfile2=/home/user/z13-222/z13-222-both_20130826_104703_VEP/vep_data_sorted.csv --vepfile3=/home/user/z13-222/z13-222-MuTect_20130826_104703_VEP/vep_data_sorted.csv --rubioseq=/home/user/z13-222/results35 --mutect=/home/user/z13-222/MuTect
```

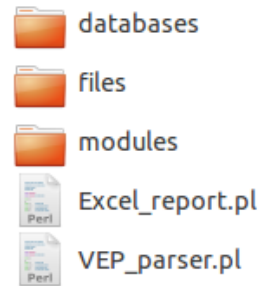
## Output file

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	chr	loc	mut	gene_hgr	tumorport	role_drive	zygosity	var_freq	coverage	gene	feature	consequence	functional	cosmic_id
2	3	17892155	T/A	PIK3CA	PRAD:Ne	CGC:onco	Heterozyg	0.11	625	ENSG000	ENST000	missense	probably	COSM75
3	4	1808898	-/C	FGFR3	NB:Near	CGC:onco	Heterozyg	0.16	924	ENSG000	ENST000	frameshift	inferred/inferred/inferre	
4	4	1808898	-/C	LETM1			Heterozyg	0.16	924	ENSG000	ENST000	downstream_gene_variant		
5	5	14943359	T/G	HMGXB3			Heterozyg	0.94	158	ENSG000	ENST000	downstream_gene_variant		
6	5	14943359	T/G	CSF1R			Heterozyg	0.94	158	ENSG000	ENST000	3_prime_UTR_variant		
7	5	14943359	G/A	HMGXB3			Heterozyg	0.82	184	ENSG000	ENST000	downstream_gene_variant		
8	5	14943359	G/A	CSF1R			Heterozyg	0.82	184	ENSG000	ENST000	3_prime_UTR_variant		
9	12	25378562	C/T	KRAS	OV:Near	CGC:onco	Heterozyg	0.25	581	ENSG000	ENST000	missense	possibly	COSM19
10	12	25378562	C/T	AC087239.1			Heterozyg	0.25	581	ENSG000	ENST000	upstream_gene_variant		

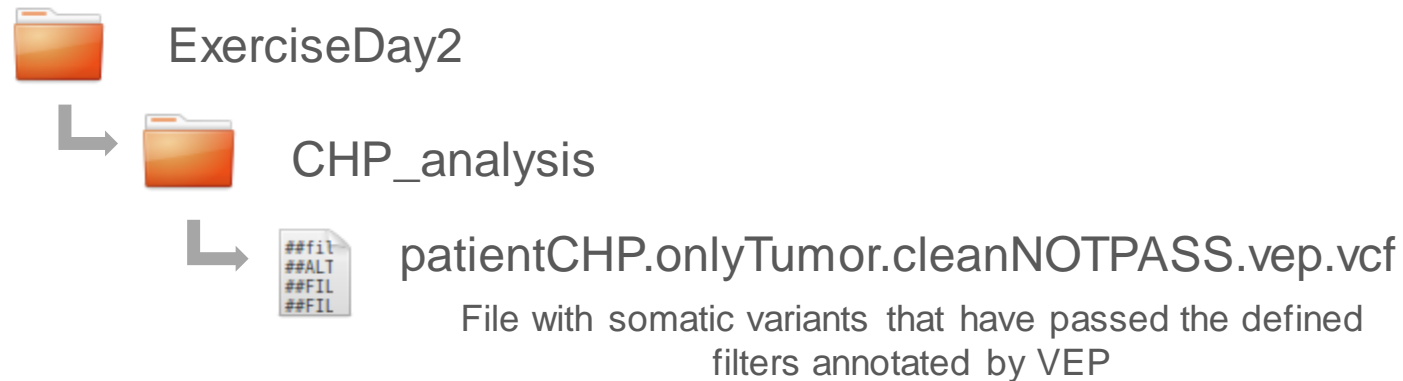


# Things to do

Programs are in ExerciceDay2 folder (VEP\_parser y el Excel\_report)



Our input file is the vcf file annotated in the previous practice with the VEP.



# Execute VEP\_parser

- Without consequence filtering.
- Sample code and rootname = patientCHP
- Output folder = CHP\_analysis
- Database directory = databases folder inside PracticeDay2 folder

```
perl VEP_parser.pl -  
f=/home/participant/ExerciseDay2/CHP_analysis/patientCHP.onlyTumor.cleanNOTPASS.vep.vcf -  
o=/home/participant/ExerciseDay2/CHP_analysis/ -sp=patientCHP -r=patientCHP -v=FALSE -  
d=/home/participant/ExerciseDay2/VEP_parser/databases/
```

Check that all output files have been generated in the new folder in CHP\_analysis.  
Open the file vep\_data\_sorted.csv to check that it has been generated correctly.

# Databases versions

- Cosmic Release v76 - hg19
- Pfam 29.0 (Nov 2015)
- UniProt release 2016\_03 (12/04/2016)
- InterPro 56.0 (13/04/2016)
- Clinvar 1.36 (01/09/2016)
- CGC (Cosmic v76) → The corresponding assembly is GRCH38 (but we search at gene level)
- APPRIS (gen19.ensembl74 13/04/2016)
- ExAC 0.3 (Uses HG19 coordinates)
- KEGG (12/04/2016)



# Execute Excel\_report

- Output folder= CHP\_analysis
- Tumor case name = patientCHP
- Tumor sample name = patientCHP-tumor
- Normal sample name = patientCHP-normal
- vepfile1 = vep\_data\_sorted.csv file from VEP\_parser
- rubioseq = path to output file in RUBioSeq including the folder name

```
perl Excel_report.pl --output=/home/participant/ ExerciseDay2/ CHP_analysis/ --case=patientCHP --  
tsample=patientCHP-tumor --nsample=patientCHP-normal --vepfile1=/home/participant/ExerciseDay2/  
CHP_analysis/patientCHP_20160228_175748_VEP/vep_data_sorted.csv --rubioseq=/home/participant/  
ExerciseDay2/CHP_analysis/RSresults/
```

Check that the output file has been generated

# Play with results

Open the Analysis.patientCHP.xls file and check the results

Have indels the same nomenclature in vcf file? What happens with the coordinates?

Which variants seem false positives? What annotations support this assumption?

Which variants seem more relevant in the pathology? Which annotations support this assumption?

Which threshold could be established for the vscore according to this data?