IO17 | Large Scale Bioinformatics for Immuno-Oncology

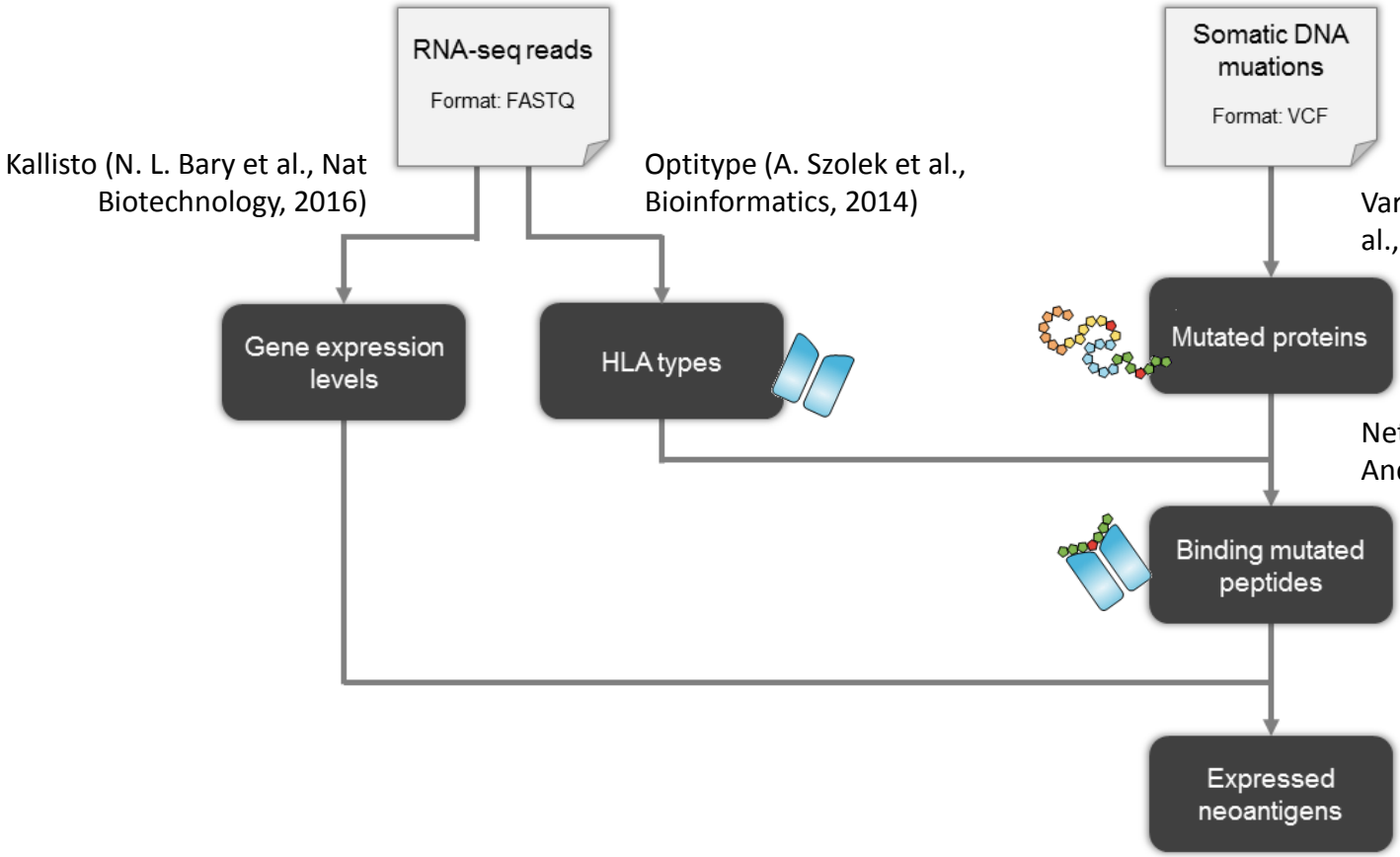
Variant annotation

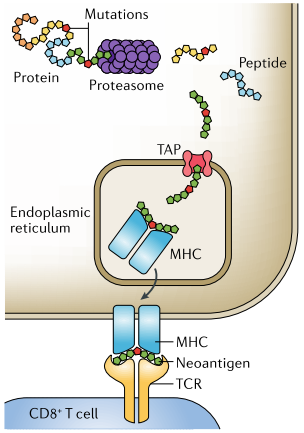
Francesca Finotello, Federica Eduati, and Pedro L. Fernandes

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

INSTITUTO GULBENKIAN DE CIÊNCIA

A pipeline for the prediction of class-I neoantigen





E Tappeiner, F Finotello et al., Bioinformatics, 2017

Variant annotation and interpretation

DNA mutations/variants can be identified from WGS and WES data from tumor and healthy tissue samples

Data analysis:

- Read mapping
- **Identification** of small and large alterations with detection tools
- **Annotation** and **interpretation** of variants individually (e.g. for functional impact) and collectively (e.g. to identify relevant gene pathways and networks)

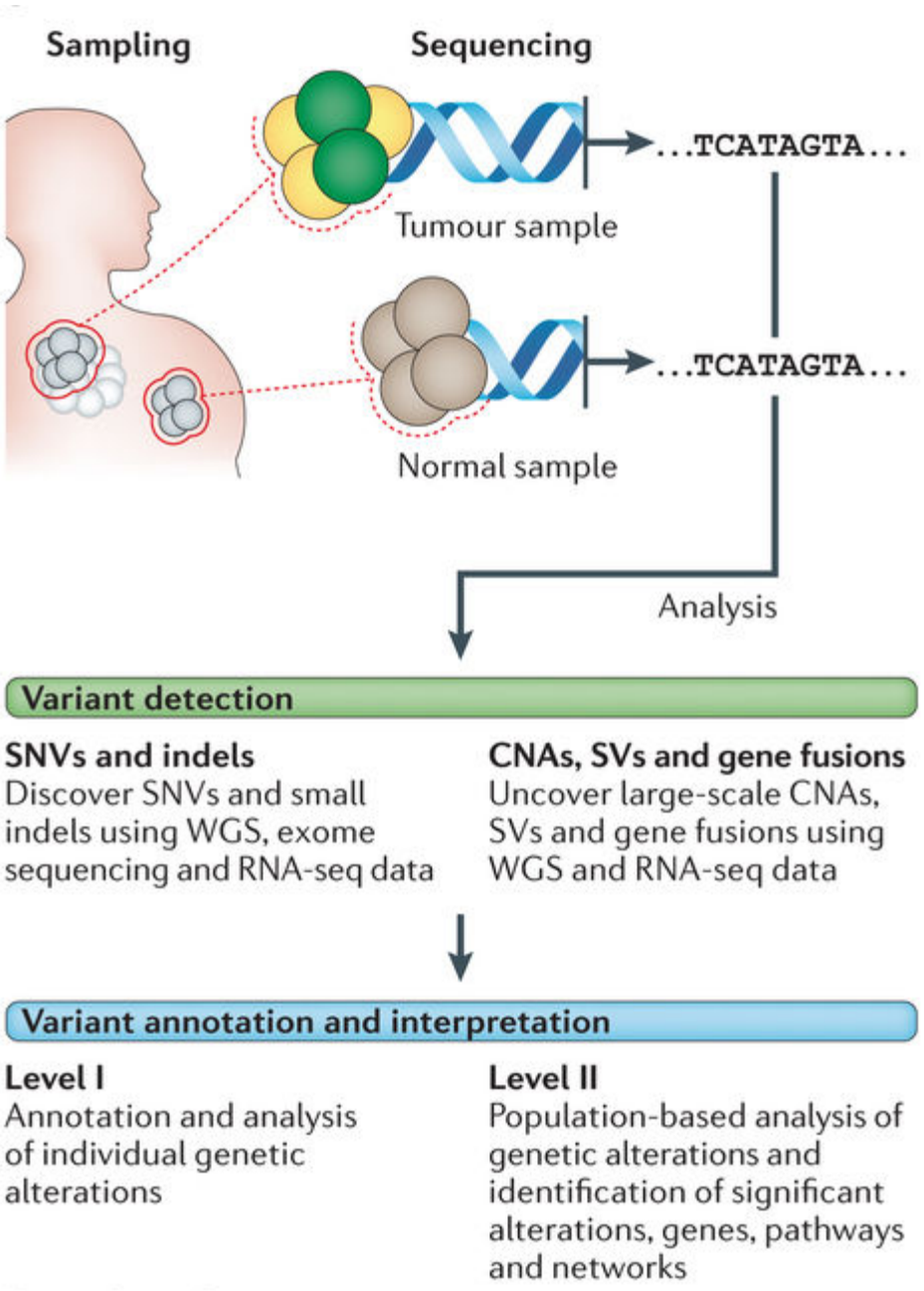


Figure adapted from L Ding et al., Nature reviews Genetics, 2014

Computational tools for variant annotation and interpretation

Level I annotation and interpretation		
ABSOLUTE	Purity, ploidy and clonality prediction	Optimization of logarithmic scores
ANNOVAR	Functional prediction	Annotation-based prediction
ASCAT	Purity, ploidy and clonality prediction	Goodness-of-fit ranking of candidate solutions
TUSON Explorer	Gene classification	Oncogene or tumour suppressor discovery using mutational signatures
CHASM	Functional prediction	Random forest classifier
MutationAssessor	Functional prediction	Conservation-based prediction (entropy score)
PolyPhen2	Functional prediction	Probability model based on structure and alignment
SciClone	Tumour clonality prediction	Bayesian mixture model
SIFT	Functional prediction	Conservation-based prediction
SNPeff	Functional prediction	Annotation and coding effect prediction
THetA	Purity, ploidy and clonality prediction	Maximum likelihood of mixture composition
VEP	Functional prediction	Annotation-based prediction

L Ding et al., Nature reviews Genetics, 2014

The Variant Effect Predictor (VEP)

VEP relies on up-to-date databases to determine the effect of DNA variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, protein sequences and regulatory regions

Different human genome assemblies are supported: GRCh37 and GRCh378

Main annotations reported:

- genes and transcripts affected by the variants
- location of the variants
- consequence of the variants on the protein sequence

VEP can use plugin modules written in Perl to add functionality to the software.
For instance, the “ProteinSeqs„ plugin outputs the complete sequence of the affected protein

VEP is available as standalone script and web interface:
http://www.ensembl.org/info/docs/tools/vep/index.html

W McLaren, et al, Genome biology, 2016

VCF format

Tab-delimited file of DNA mutations. Format:

- Multiple meta-information lines starting with “##”
- One header starting with “#”
- Multiple lines reporting the info about the mutations, one mutation per line

The file must have 8 mandatory columns:

- #CHROM: chromosome
- POS: position of the mutation on the chromosome
- ID: mutation identifiers (e.g. rs numbers from dbSNP)
- REF: reference base
- ALT: alternative/mutated base
- QUAL: Phred quality score for the ALT base
- FILTER: “PASS” if the position has passed the quality filters, list of failed filters otherwise
- INFO: additional information

Note: Unavailable information can be reported as “.”

https://samtools.github.io/hts-specs/VCFv4.2.pdf

VCF format

```
##fileformat=VCFv4.2
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G
```

https://samtools.github.io/hts-specs/VCFv4.2.pdf

Run VEP with TIminer

```
from TIminer import TIminerAPI

TIminerAPI.executeVep(…)
```

From TIminer documentation

http://icbi.i-med.ac.at/software/timiner/doc/index.html

`TIminer.TIminerAPI.executeVep`(*inputFile*, *subject*, *outputFile=None*, *mutatedSeqOutputFile=None*, *cacheDir=None*, *genomeVersion=38*)

This function takes as an input a file of DNA somatic mutations for each subject and predicts their impact on proteins using the Variant Effect Predictor (VEP). For each subject, it produces a file of annotated mutations and a FASTA file with the sequences of the corresponding mutated proteins. It considers only non-synonymous, Single-Nucleotide Polymorphisms (SNP).

Parameters:

- **inputFile** (*str*) –
  Path to the input file of somatic DNA mutations. The supported format is the VCF format.
- **subject** (*str*) – The subject ID that will appear in the output file.
- **outputFile** (*str*) – Path to the `output file` of annotated DNA mutations (optional, default = *mutated-proteins/subjectid_mutprotein_info.txt*).
- **mutatedSeqOutputFile** (*str*) – Path to the `FASTA output file` with the sequences of the mutated proteins (optional, default = *mutated-proteins/subjectid_mutprotein_seq.txt*).
- **cacheDir** (*str*) – Path to the genomic annotation to be used by VEP (optional, default = *path to installed database dir*).
- **genomeVersion** (*int*) – Version of the human genome used as reference annotation, being 37 for GRCh37, 38 for GRCh38 (optional, default = *38*).

Note: VEP@TIminer reports only one consequence per variant and only if its affects a coding region

VEP@TIminer output: annotated mutations

Tab-delimited text file with the annotated somatic DNA mutations. Format:

- Multiple meta-information lines starting with “##”
- One header starting with “#”
- Multiple lines reporting, for each mutation, the following annotations:

Uploaded_variation: subject ID (e.g. Patient_1)

Location: genomic location of the mutation, chromosome and position (e.g. 1:12884990)

Allele: reference allele (e.g. A)

Gene: Ensemble gene ID (e.g. ENSG00000204513)

Feature: feature ID (e.g. ENST00000535591)

Feature_type: feature type (e.g. Transcript)

Consequence: mutation type (e.g. missense_variant)

cDNA_position: position of the mutation in the cDNA (e.g. 1317)

CDS_position: position of the mutation in the CDS (e.g. 1121)

Protein_position: position of the mutation in the protein (e.g. 374)

Amino_acids: amino acid change (e.g. P/L)

Codons: codon change (e.g. cCt/cTt)

Existing_variation: information about existing variation (e.g. -)

Extra: additional annotations, including gene symbol and protein ID (e.g. IMPACT=MODERATE;STRAND=-1;**SYMBOL=PRAMEF11**;SYMBOL_SOURCE=HGNC;HGNC_ID=14086;**ENSP=ENSP00000439551**…)

Mutation type

missense_variant: a sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved

synonymous_variant: a sequence variant where there is no resulting change to the encoded amino acid

intron_variant: a transcript variant occurring within an intron

frameshift_variant: a sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three

inframe_deletion: an inframe non synonymous variant that deletes bases from the coding sequence

Note: TIminer pipeline considers only missense mutations for the next step of peptide-MCH binding prediction.

More mutation types annotated by VEP at:

http://www.ensembl.org/info/genome/variation/predicted_data.html#consequences

VEP@TIminer output: mutated proteins

Annotated mutation (outputFile)

#Uploaded_variation: Patient_1
Location: 19:43269699
Allele: T
Gene: ENSG00000124467
Feature: ENST00000306511
Feature_type: Transcript
Consequence: missense_variant
cDNA_position: 133
CDS_position: 35
Protein_position: **12**
Amino_acids: **R/H**
Codons: cGc/cAc
Existing_variation: -
Extra: … ENSP=**ENSP00000305005**…

Mutated protein (mutatedSeqOutputFile)

>ENSP00000305005.4:p.Arg12His
MGLLSAPPCTQ**H**ITWKGLLLTASLLNFWNPPTTAQVTIEAQPTKVSEGKDVLLLVHNLPQNLTGYIWYKGQIRDLYHYIT
SYVVDGQIIIYGPAYSGRETIYSNASLLIQNVTQEDAGSYTLHIIMGGDENRGVTGHFTFTLYLETPKPSISSSKLNPRE
AMEAVSLTCDPETPDASYLWWMNGQSLPMSHRLQLSETNRTLFLLGVTKYTAGPYECEIRNPVSASRSDPFTLNLLPKLP
KPYITINNLKPRENKDVLNFTCEPKSENYTYIWWLNGQSLPVSPRVKRPIENRILILPSVTRNETGPYQCEIRDQYGGIR
SYPVTLNVLYGPDLPRIYPSFTYYRSGEVLYLSCSADSNPPAQYSWTINGKFQLSGQKLFIPQITTKHSGLYACSVRNSA
TGKESSKSMTVKVSGKRIPVSLAIGI