IO17 | Large Scale Bioinformatics for Immuno-Oncology

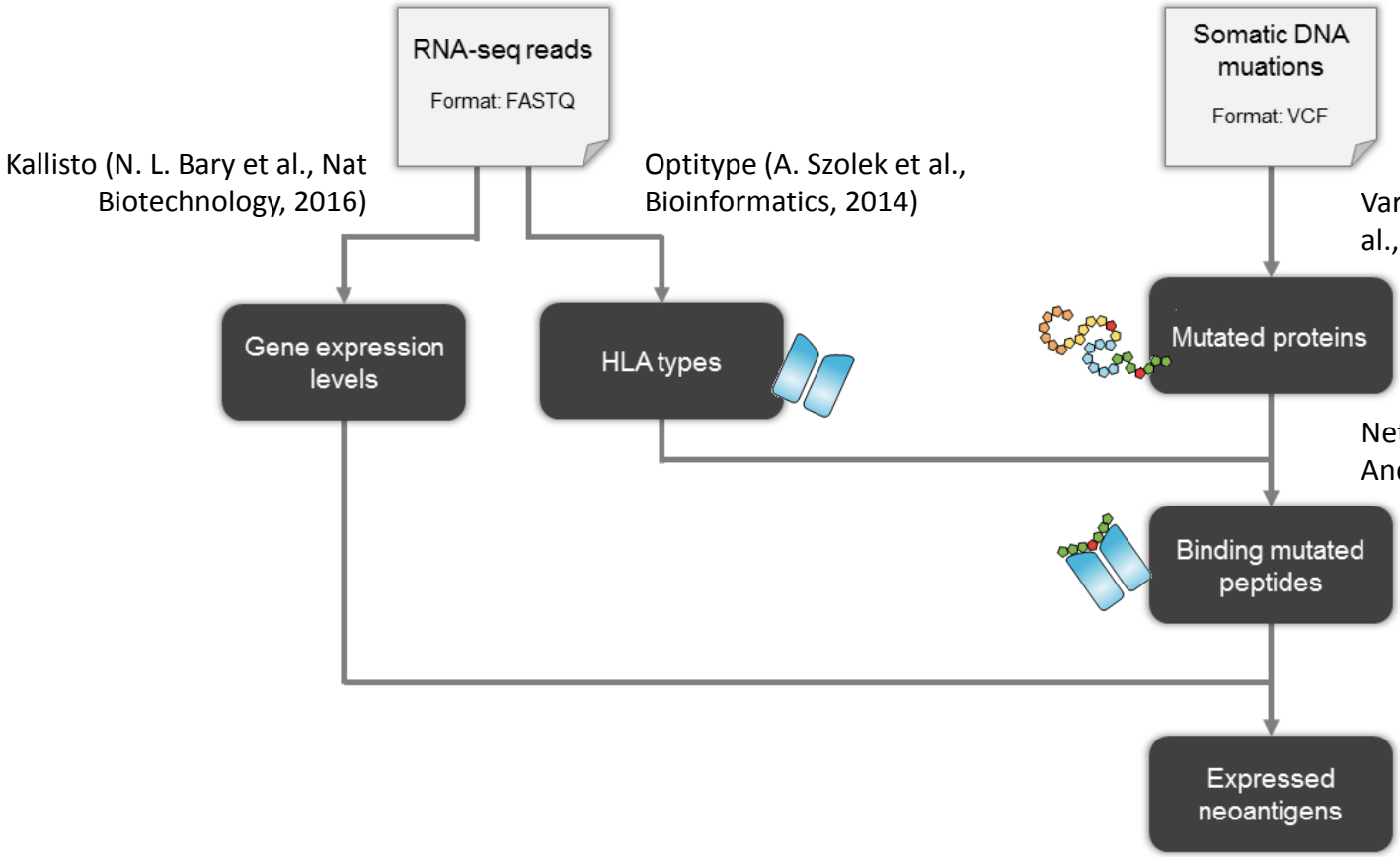
Peptide-MHC binding prediction

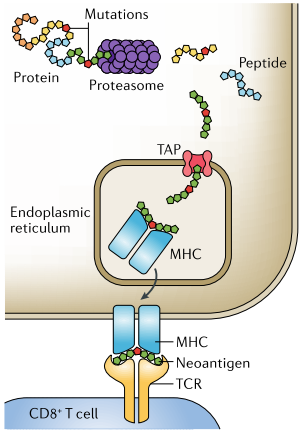
Francesca Finotello, Federica Eduati, and Pedro L. Fernandes

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

INSTITUTO GULBENKIAN DE CIÊNCIA

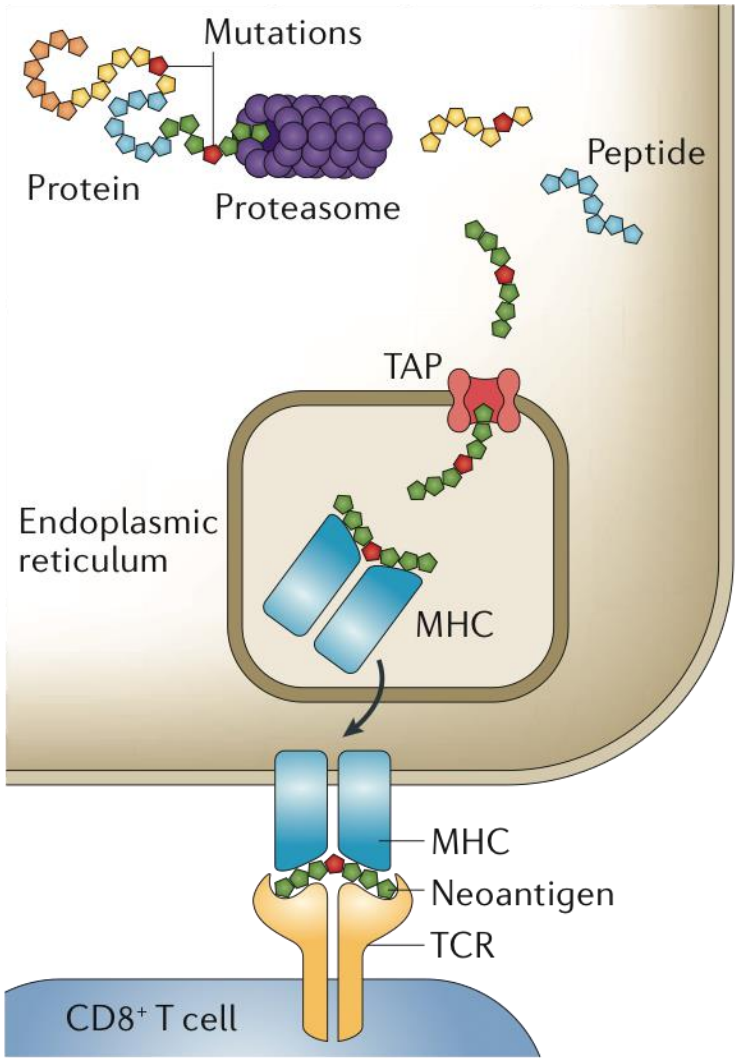
A pipeline for the prediction of class-I neoantigen





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

Neoantigen-MHC binding



Not all peptides bind to all MHC molecules (i.e. they can have high or low **binding affinity**)!

H Hackl*, P Charoentong*, F Finotello* et al. Nature Reviews Genetics, 2016

Prediction of peptide-MHC binding

Computational methods to predict peptide-MHC binding affinity are:

- Structure-based (3D structures) → threading/docking approaches
- Sequence-based (primary sequence) → more data available for training/validation

Modern sequence-based tools for peptide-MHC binding prediction use **machine learning** algorithms (e.g. neural networks) trained on data-sets of peptide-MHC/HLA with known **primary sequence** and experimentally measured **binding affinity**

Binding affinity is often measured as **half maximal inhibitory concentration** (**IC_{50}**)

IC_{50}: the concentration of the query peptide which inhibits 50% of a reference peptide binding the MHC (i.e. if a low concentration is required to outcompete the reference peptide, the query peptide has a strong binding affinity for the MHC)

Computational tools for peptide-MHC binding prediction

CONSENSUS	Consensus approach for prediction of pMHC-I and pMHC-II binding affinity hosted on the IEDB website	http://tools.immuneepitope.org/mhcii
netMHC	Machine-learning-based prediction of pMHC binding affinity to human and non-human MHC-I molecules	http://www.cbs.dtu.dk/services/NetMHC
netMHCcons	Consensus-based prediction of pMHC-I binding affinity integrating the predictions of NetMHC, NetMHCpan and PickPocket	http://www.cbs.dtu.dk/services/NetMHCcons
netMHCpan	Pan-specific version of netMHC	http://www.cbs.dtu.dk/services/NetMHCpan
netMHCstab	Machine-learning-based prediction of the stability of binding of small peptides to HLA-A and HLA-B molecules	http://www.cbs.dtu.dk/services/NetMHCstab-1.0
netMHCII	Machine-learning-based prediction of binding affinity to human and mouse class-II MHC molecules	http://www.cbs.dtu.dk/services/NetMHCII
netMHCIIpan	Pan-specific version of netMHCII	http://www.cbs.dtu.dk/services/NetMHCIIpan
PickPocket	Pan-specific predictor of pMHC-I binding affinity based on PSSM of peptides and MHC pockets	http://www.cbs.dtu.dk/services/PickPocket

H Hackl*, P Charoentong*, F Finotello* et al. Nature Reviews Genetics, 2016

NetMHCpan

Experimental data on peptide-MHC binding is lacking for the large majority of HLA alleles

Development of **pan-specific** methods, like NetMHCpan, which predict peptide binding to any class-I HLA molecules of known protein sequence

NetMHCpan represents HLA as pseudo-sequences (HLA residues in contact with bound peptides) and trains a neural network to output the affinity of a given HLA-peptide pair

NetMHCpan scored amongst the best performers, even compared to allele-specific methods

NetMHCpan 3.0 is used by TIminer for peptide-HLA binding prediction and is available as web-based and standalone tool: http://www.cbs.dtu.dk/services/NetMHCpan/

M Nielsen et al, PloS one, 2007
M Nielsen and M Andreatta, *Genome medicine*, 2016

Run NetMHCpan with TIminer

```
from TIminer import TIminerAPI

TIminerAPI.executeNetmhcpan(…)
```

From TIminer documentation

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

`TIminer.TIminerAPI.executeNetmhcpan`(*inputFile*, *mutatedProteinsInputFile*, *hlaInputFile*, *outputFile*, *threadCount=1*, *minPeptideLength=8*, *maxPeptideLength=11*, *affinityThresh=500*, *rankThresh=None*)

This function considers, a file of annotated mutations, a corresponding FASTA file with the sequence of the mutated proteins, and a list of class-I HLA types. It extracts short mutated peptides (default: 8-11 amino acids) and predicts their binding affinity to HLA types using NetMHcpan. Only mutated peptides with a binding affinity < 500 nM are reported in the output file.

Parameters:

- **inputFile** (*str*) –
  Path to the `input file` of annotated DNA somatic mutations generated by the Variant Effect Predictor (see the output files section for details).
- **mutatedProteinsInputFile** (*str*) –
  Path to the `FASTA file` with the sequences of the mutated proteins generated by the Variant Effect Predictor (see the output files section for details).
- **hlaInputFile** (*str*) –
  Path to the `file` containing the HLA types, that can be generated with Optitype or provided by the user (for details see the output files section).
- **outputFile** (*str*) – The path to the column based `output file` (see the output files section for details).
- **threadCount** (*int*) – Number of threads to be used (optional, default = *2*).
- **minPeptideLength** (*int*) – Minimal peptide length to test (optional, default = *8*).
- **maxPeptideLength** (*int*) – Maximal peptide length to test (optional, default = *11*).
- **affinityThresh** (*int*) – binding affinity threshold (optional, default = *500*. If set the rank threshold needs to be None).
- **rankThresh** (*int*) – ranking based binding threshold (optional, default = *None*. If set the binding affinity threshold needs to be None).

IC_{50} →

NetMHCpan@TIminer output

Tab-delimited text file with the information about all mutated peptides passing the binding-affinity threshold. Format:

#SubjectID: subject ID
Pos: genomic position of the mutation
GeneID: Ensemble gene ID
TranscriptID: Ensemble transcript ID
GeneSymbol: HGNC gene symbol
Protein: Ensemble protein ID
ProteinPos: position of the mutated amino acid in the protein sequence
Mut: amino acid change
VariantType: type of variant
HLA: HLA type
MutPeptide: sequence of the mutated peptide
MutAFF: IC50 affinity of the mutated peptide in nM
MutRank: rank of the mutated peptide affinity compared to a set of 400.000 random peptides
RefPeptide: sequence of the wild-type (i.e. non-mutated) peptide
RefAFF: IC50 affinity of the unmutated peptide in nM
RefRank: rank of the wild-type peptide affinity compared to a set of 400.000 random peptides