MHC-Peptide Binding Prediction

Peter Ziribagwa Sabakaki

@CancerGenomicsWorkingGroup

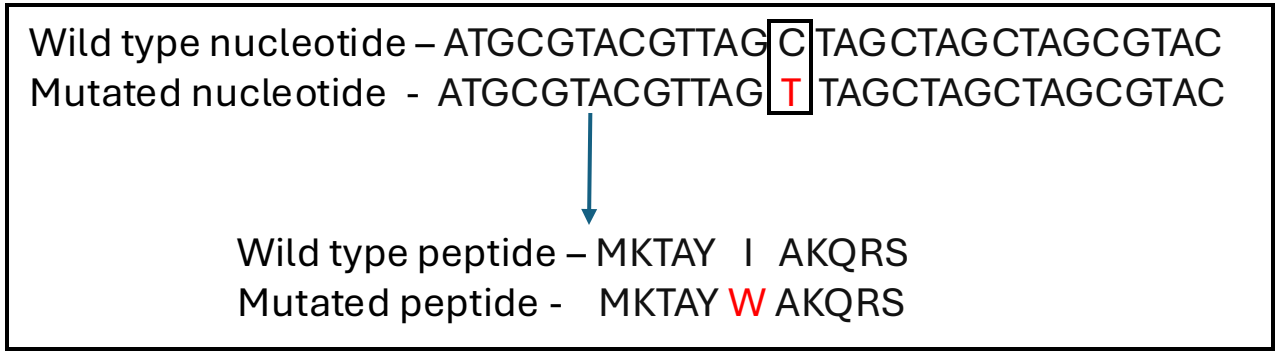
What are we doing?

- Identify **mutated proteins** that can bind to any of the **genotyped HLA alleles** with a high binding affinity

VCF

	#CHROM	POS	ID	REF	ALT	QUAL		FILTER	INFO	F
61	chr1	116549	.	C	T	.		SuspiciousHomAlt		
62	chr1	120458	.	T	C	.		SuspiciousHomAlt		
63	chr1	125271	.	C	T	.		SuspiciousHomAlt		
64	chr1	126113	.	C	A	.		PASS	MTD=isaac	
65	chr1	128798	.	C	T	.		SuspiciousHomAlt		
66	chr1	129963	.	T	A	.		SuspiciousHomAlt		
67	chr1	139967	.	T	C	.		SuspiciousHomAlt		
68	chr1	172595	.	G	A	.		SuspiciousHomAlt		
69	chr1	173173	.	A	G	.		SuspiciousHomAlt		
70	chr1	229673	.	A	C	.		SuspiciousHomAlt		
71	chr1	241369	.	C	T	.		SuspiciousHomAlt		
72	chr1	321466	.	G	T	.		SuspiciousHomAlt		
73	chr1	356443	.	A	G	.		SuspiciousHomAlt		
74	chr1	356537	.	G	A	.		SuspiciousHomAlt		
75	chr1	523471	.	T	C	.		SuspiciousHomAlt		

Extract Mutated peptide



HLA-genotypes

A*33:01	A*01:02	B*49:01	B*14:02	C*07:01	C*08:02

What is Binding affinity?

- BA – is the strength of interaction between HLA and peptide
- BA is measured as 50% inhibitory concetration (IC50) in nanomolar units (nM)
- IC50 – is the concentration of a peptide that is required to inihibit 50% of the binding of that peptide to an MHC molecule
- pMHC is considered strong if IC50 > 150nM
- pMHC is considered week if IC50 is between 150 – 500nM
- pMHC is non-binding if IC50 is < 500nM

Tools that can predict pMHC binding affinity

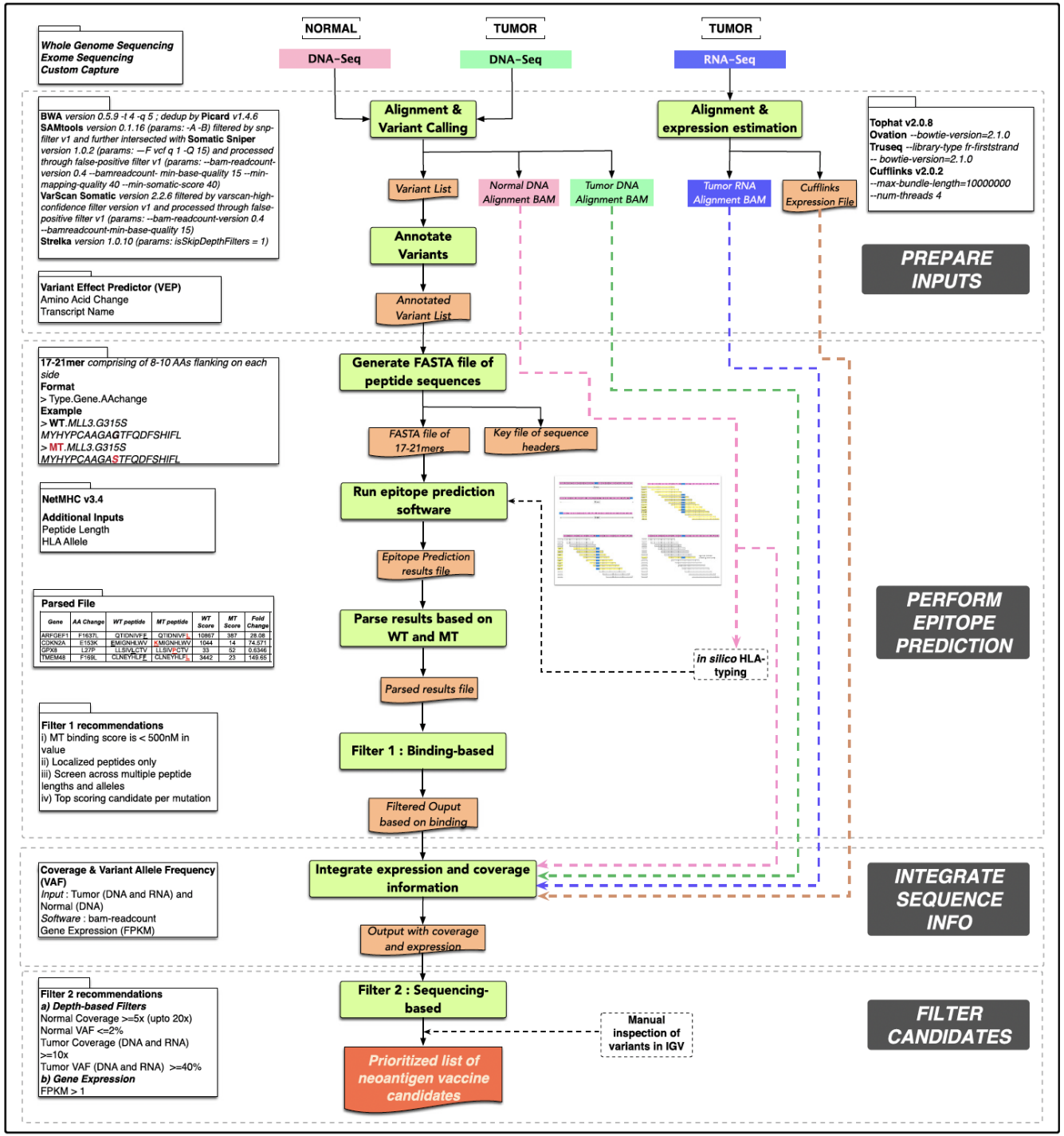
- pVAC-seq
- TSNAD
- CloudNeo
- MuPeXI
- INTEGRATE-Neo
- NeoanitigenR
- HLAthena

NB: All these tools use **NetMHCpan** machine learning tool

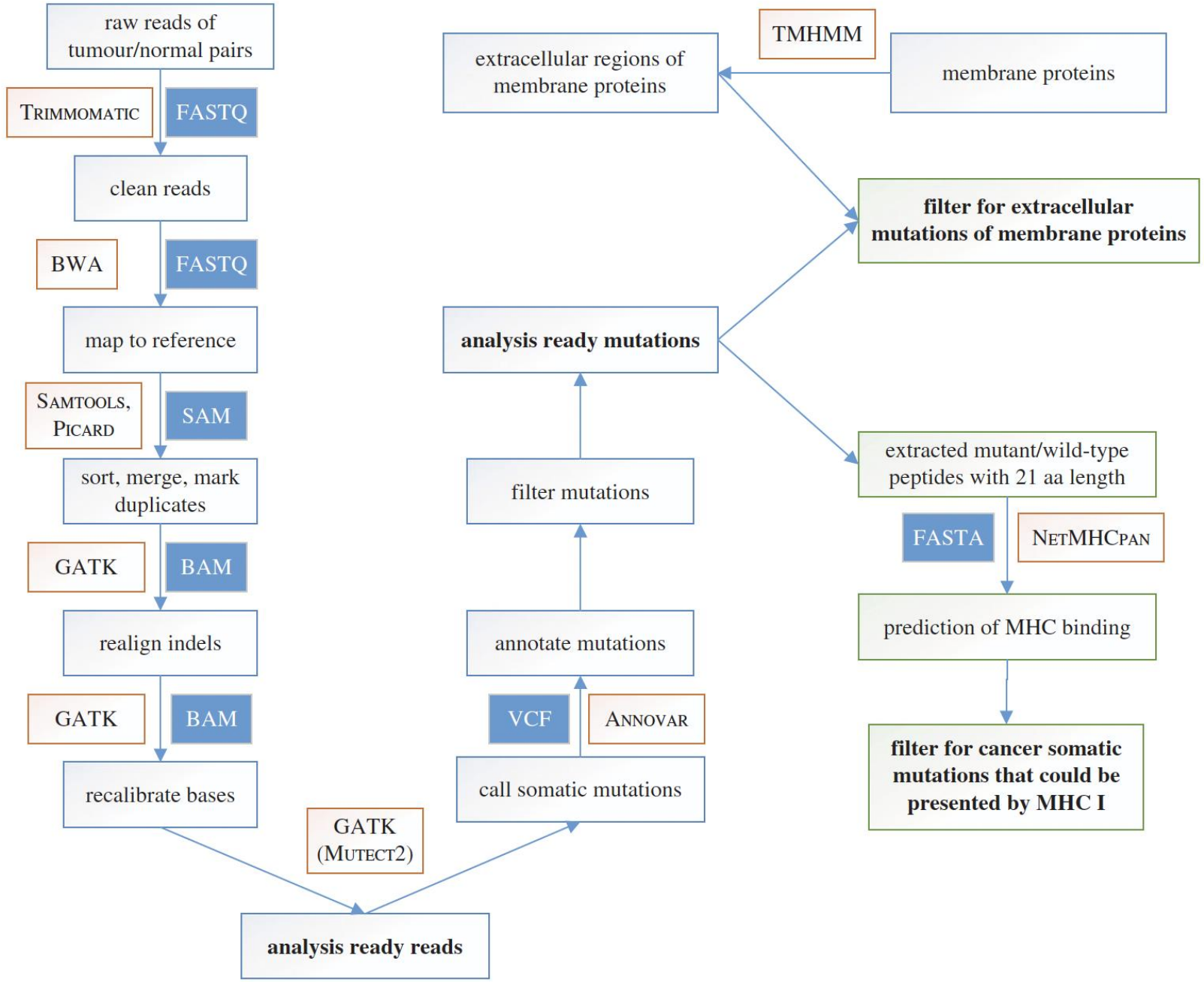
How NetMHCpan Works!

- **Data:** NetMHCpan was trained with **a large database of known peptide-MHC binding interactions**, with experimentally determined binding affinities, expressed as **IC50** values.
- **Feature Extraction**: Utilises relevant features from both **the peptide sequences** and the **MHC molecule** sequences. These features include amino acid properties and sequence motifs that are important for binding.
- **Model Training**: Using these features, NetMHCpan trained an artificial neural networks to predict the binding affinity of **any given peptide** to any **MHC** molecule.
- **Prediction**: the trained model can now predict the binding affinity (IC50 value) for new peptide-MHC pairs.
- **Validation**: The predictions are validated against experimental data to ensure reliability. This involves cross-validation and prospective validation with new data.

PVAC-seq


Whole Genome Sequencing
Exome Sequencing
Custom Capture
NORMAL
DNA-Seq
TUMOR
DNA-Seq
TUMOR
RNA-Seq
BWA version 0.5.9 -t 4 -q 5 ; dedup by Picard v1.4.6
SAMtools version 0.1.16 (params: -A -B) filtered by snp-filter v1 and further intersected with Somatic Sniper version 1.0.2 (params: — F vcf q 1 -Q 15) and processed through false-positive filter v1 (params: --bam-readcount-version 0.4 --bamreadcount- min-base-quality 15 --min-mapping-quality 40 --min-somatic-score 40)
VarScan Somatic version 2.2.6 filtered by varscan-high-confidence filter version v1 and processed through false-positive filter v1 (params: --bam-readcount-version 0.4 --bamreadcount-min-base-quality 15)
Strelka version 1.0.10 (params: isSkipDepthFilters = 1)
Alignment & Variant Calling
Alignment & expression estimation
Tophat v2.0.8
Ovation --bowtie-version=2.1.0
Truseq --library-type fr-firststrand -- bowtie-version=2.1.0
Cufflinks v2.0.2
--max-bundle-length=10000000
--num-threads 4
Variant List
Normal DNA Alignment BAM
Tumor DNA Alignment BAM
Tumor RNA Alignment BAM
Cufflinks Expression File
Variant Effect Predictor (VEP)
Amino Acid Change
Transcript Name
Annotate Variants
Annotated Variant List
PREPARE INPUTS
17-21mer comprising of 8-10 AAs flanking on each side
Format
> Type.Gene.AAchange
Example
> WT.MLL3.G315S
MYHYPCAAGAGTFQDFSHIFL
> MT.MLL3.G315S
MYHYPCAAGASTFQDFSHIFL
Generate FASTA file of peptide sequences
FASTA file of 17-21mers
Key file of sequence headers
NetMHC v3.4
Additional Inputs
Peptide Length
HLA Allele
Run epitope prediction software
Epitope Prediction results file
Parsed File
Gene
AA Change
WT peptide
MT peptide
WT Score
MT Score
Fold Change
ARFGEF1
F1637L
QTIDNIVFF
QTIDNIVFL
10867
387
28.08
CDKN2A
E153K
EMIGNHLWV
KMIGNHLWV
1044
14
74.571
GPX8
L27P
LLSIVLCTV
LLSIVPCTV
33
52
0.6346
TMEM48
F168L
CLNEYHLFF
CLNEYHLFL
3442
23
149.65
Parse results based on WT and MT
in silico HLA-typing
PERFORM EPITOPE PREDICTION
Parsed results file
Filter 1 recommendations
i) MT binding score is < 500nM in value
ii) Localized peptides only
iii) Screen across multiple peptide lengths and alleles
iv) Top scoring candidate per mutation
Filter 1 : Binding-based
Filtered Ouput based on binding
Coverage & Variant Allele Frequency (VAF)
Input : Tumor (DNA and RNA) and Normal (DNA)
Software : bam-readcount
Gene Expression (FPKM)
Integrate expression and coverage information
Output with coverage and expression
INTEGRATE SEQUENCE INFO
Filter 2 recommendations
a) Depth-based Filters
Normal Coverage >=5x (upto 20x)
Normal VAF <=2%
Tumor Coverage (DNA and RNA) >=10x
Tumor VAF (DNA and RNA) >=40%
b) Gene Expression
FPKM > 1
Filter 2 : Sequencing-based
Manual inspection of variants in IGV
Prioritized list of neoantigen vaccine candidates
FILTER CANDIDATES

TSNAD


raw reads of tumour/normal pairs
TRIMMOMATIC
FASTQ
clean reads
BWA
FASTQ
map to reference
SAMTOOLS, PICARD
SAM
sort, merge, mark duplicates
GATK
BAM
realign indels
GATK
BAM
recalibrate bases
GATK (MUTECT2)
analysis ready reads
call somatic mutations
VCF
ANNOVAR
annotate mutations
filter mutations
analysis ready mutations
extracellular regions of membrane proteins
TMHMM
membrane proteins
filter for extracellular mutations of membrane proteins
extracted mutant/wild-type peptides with 21 aa length
FASTA
NETMHCPAN
prediction of MHC binding
filter for cancer somatic mutations that could be presented by MHC I

What next!

- All these tools require a fasta file of mutated peptides.
- Lets' Prepare mutated fasta files for all samples (we can use MuPeXI)
- Where are the VCF files?

END

THANKS