Proteogenomics Workflow for Neoantigen Discovery

XIE Lu, Shanghai Center for Bioinformation Technology

August 29-31, 2018. The 16th KJC Bioinformatics Symposium, Hayama, Japan

目录

CONTENTS

01

Proteogenomics

Definition

‘Proteogenomics’ refers to the correlation of the proteomic data with the genomic and transcriptomic data with the goal of enhancing the understanding of the genome

Why proteogenomics?

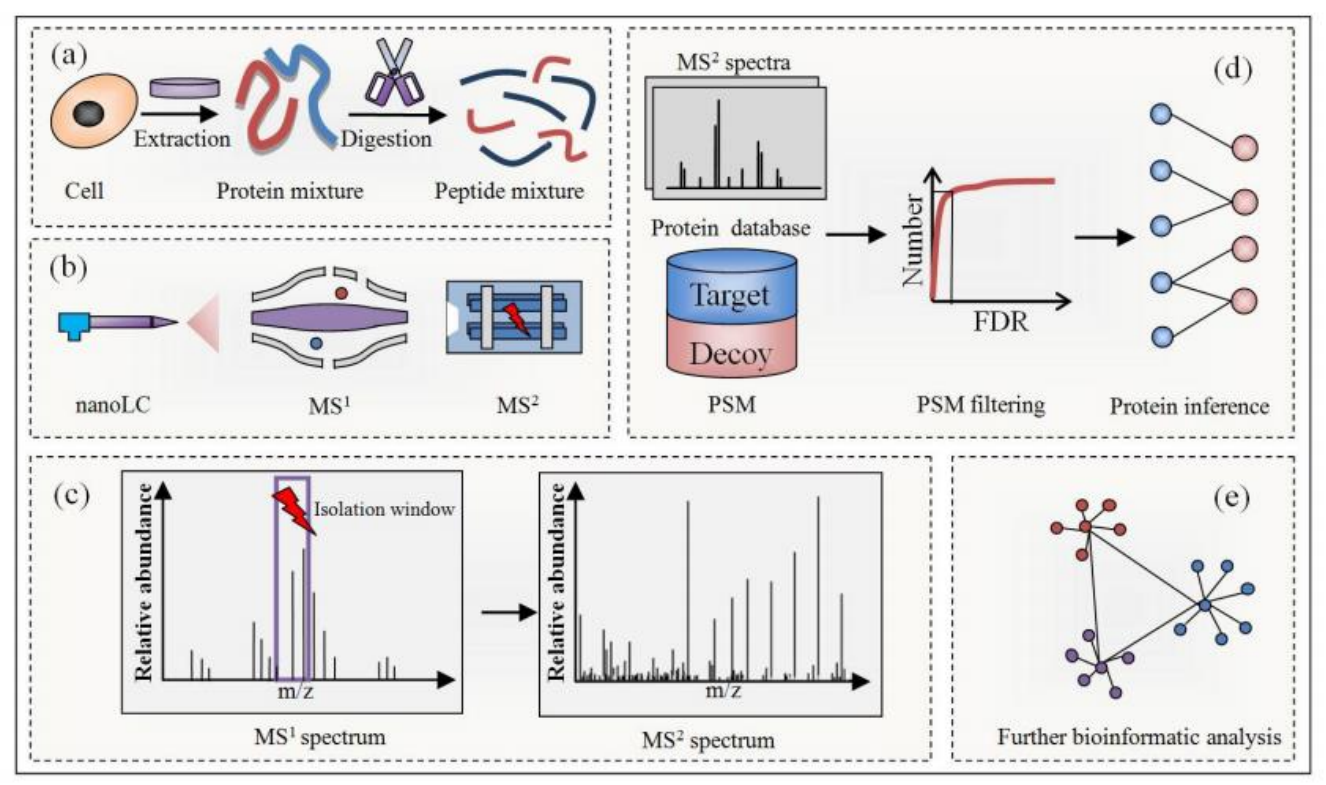
- Although gene prediction programs available, accurate gene identification decreases drastically from the nucleotide level to exons to whole gene structures
- Gene finding in prokaryotes is easier owing to their compact genomes with simple gene structure
- Gene finding in eukaryotes is difficult because of introns and complex regulatory regions

Santosh Renuse, Raghothama Chaerkady and Akhilesh Pandey. Proteogenomics. ***Proteomics***. 2011, 11, 620–630

Essentials for proteogenomics

- Availability of genome sequence data
- High-resolution and high accuracy mass spectrometry data
- Genome database search and annotation tools

MS/MS Proteomics for proteogenomics



Proteogenomics: Improving Genomes Annotation by Proteomics.

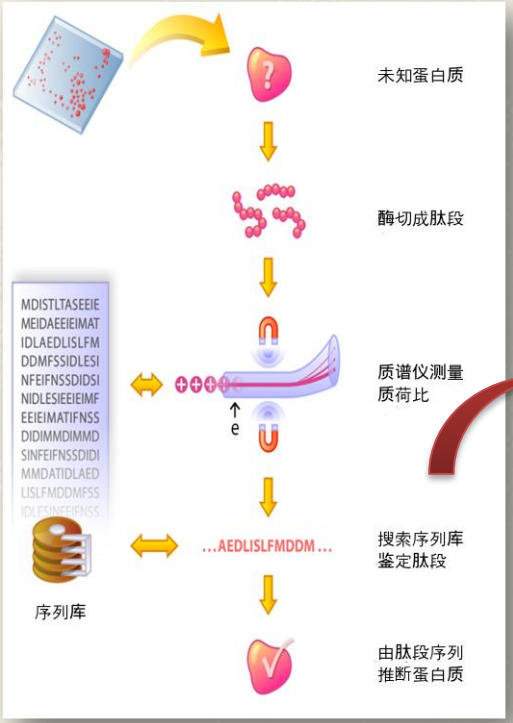
Kun Zhang, …Lu Xie, … Simin He. Prog. Biochem. Biophys, 2012

Genome reannotation by proteogennomics

Mass spectrometry-derived peptide data can be used to annotate genomes for the confirmation and/or correction of existing gene annotations

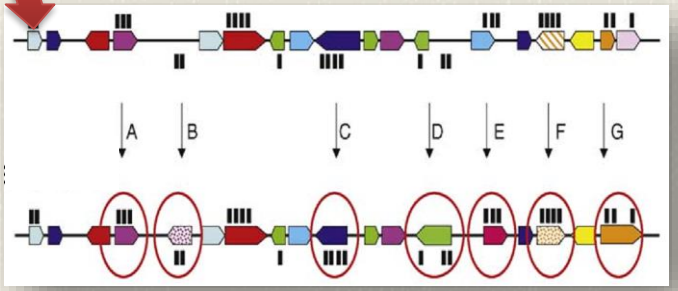
- Confirmation of existing gene models
- Correction of existing gene models
- Identification of novel genes

Genome annotation and revision

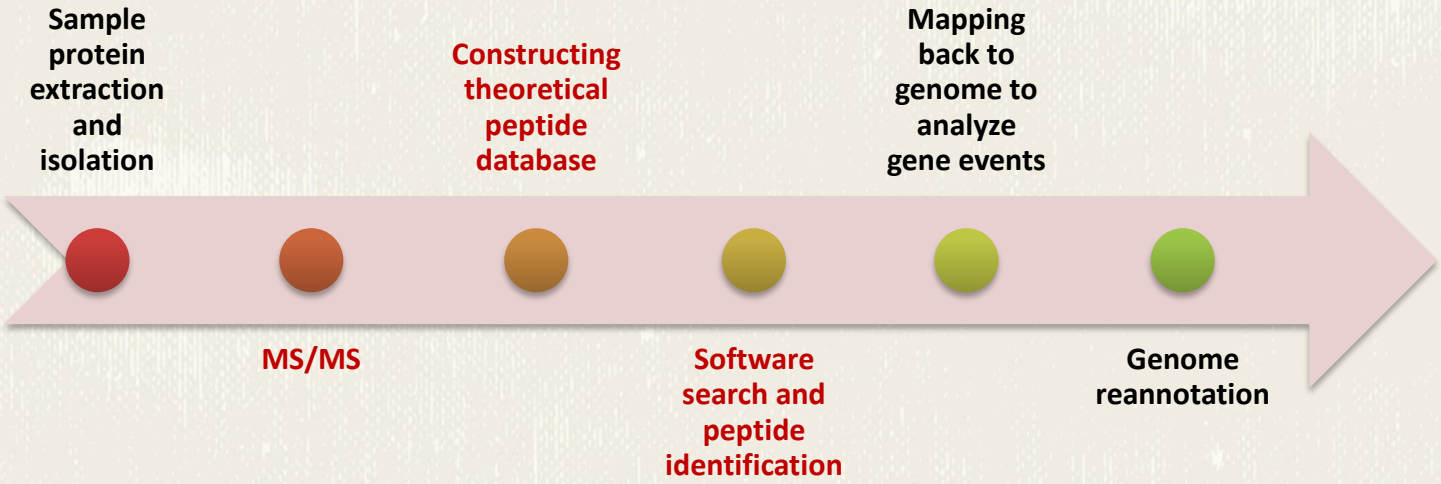


Gene event analysis

A To confirm pseudogenes
B To discover new genes
C/D Relocate gene translation initiation codon
E Correct reading frame
F Peptide from complement strand of gene
G Translation across stop codon



TECHNICAL PIPELINE



Difficulty：

Theoretical peptide database limits what kind of gene events can be identified. If only known proteins are included, novel gene events can not be discovered. To include all potential gene events is not possible.

Investigations: multiple level genome reannotation

Prokaryote genome:

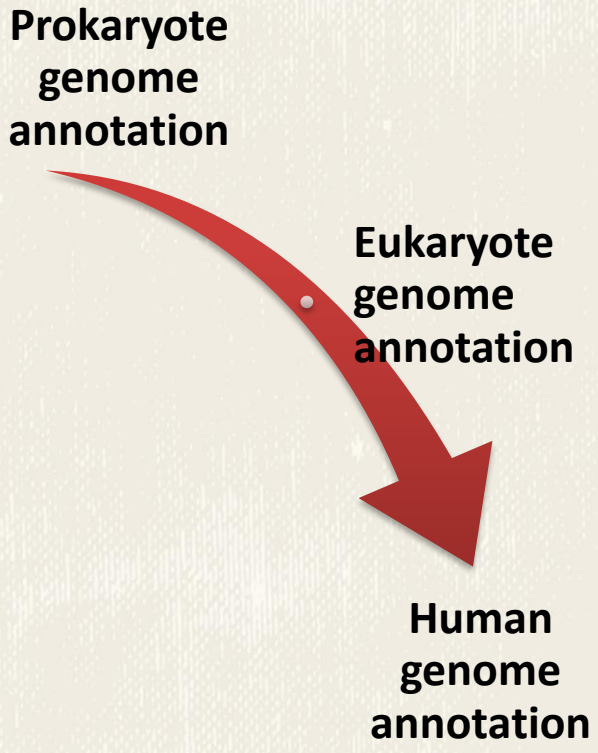
open reading frame（ORF）

Eukaryote genome:

ORF (UTRs, exomes, introns, inter gene regions); alternative splicings

Human genome:

confirmation of new genes, non-coding RNAs, alternative splicing; point mutation, fusion genes, small indels, chromosome structure variations, virus integration

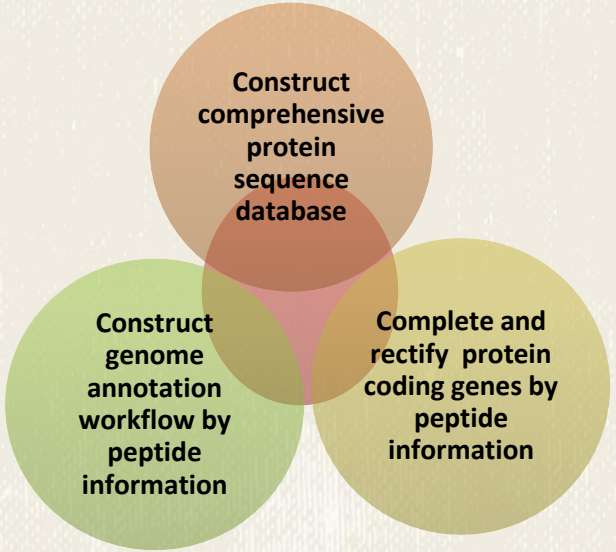


Historical works

1. Prokaryote genome reannotation *(unpublished)*
2. Supplement and revision of eukaryotic genome protein coding genes (***Genomics, 2011***)
3. Supplement and revision of human genome protein coding genes
   - Identifcation of fusion genes in cancer ***(BMC Genomics, 2013)***
   - Confirmation of de novo predicted new genes ***(Proteomics, 2014)***
   - Prediction and confirmation of predicted new transcripts on whole-genome scale ***(Scientific Reports, 2015)***
   - Identification of somatic mutated(altered) proteins ***(J Proteome Res, 2015)***

Proteogenomics for human disease research

1. **Studying disease mechanisms: fusion genes; novel splicing isoforms…**
2. **Finding genome variations: somatic mutated/altered proteins(SAPs)**
3. **Identifying potential biomarkers**



Prospect

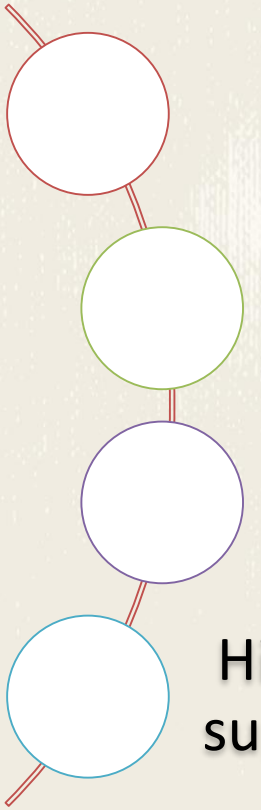
It may be suggested that every genome sequencing project should include proteogenomic analysis to provide a more accurate catalog of protein-coding genes

02

Genome reannotation

Supplement and revision

Prokaryotic genome reannotation



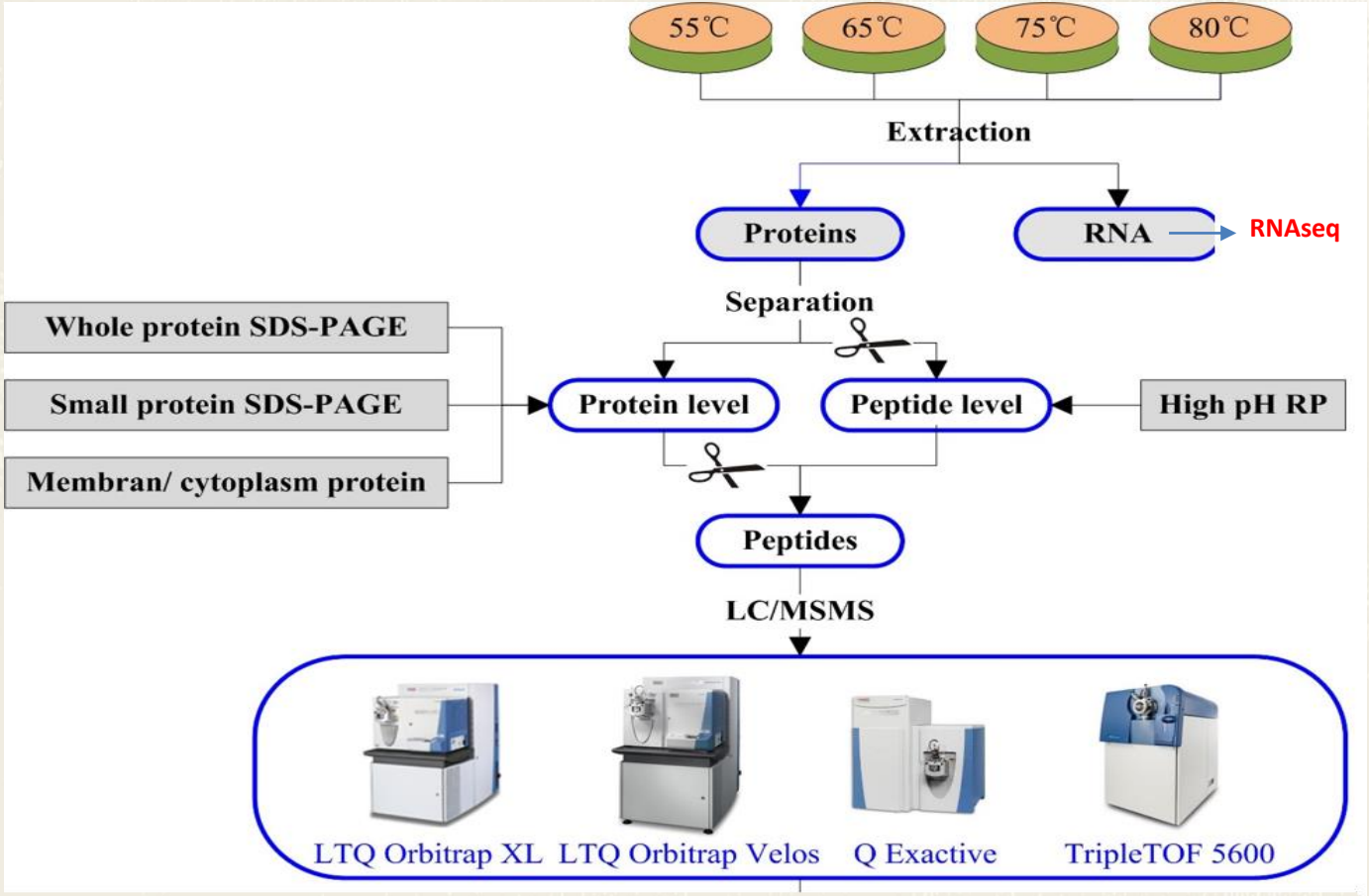
Thermophilic bacterium: Thermoanaerobacter tengcongensis (TTE)

The first complete genome profile sequenced in house (Genome Research, 2001)

2588 protein coding genes annotated (recorded by NCBI)

High resolution mass spectrometry identified peptides, for supplement and correction of this original annotation

High resolution accuracy and coverage peptide identification

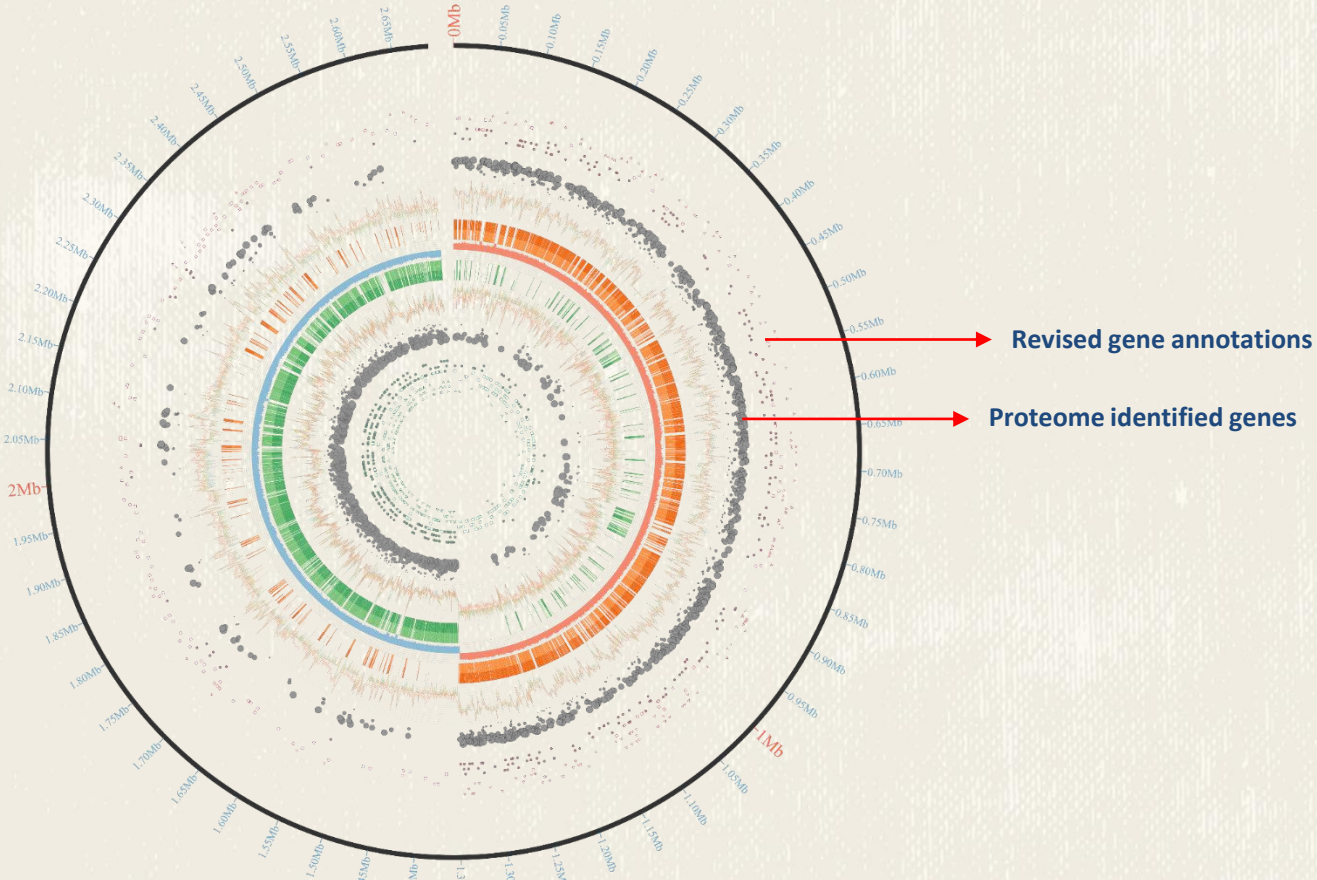


Genome reannotation by Augustus aided by peptides

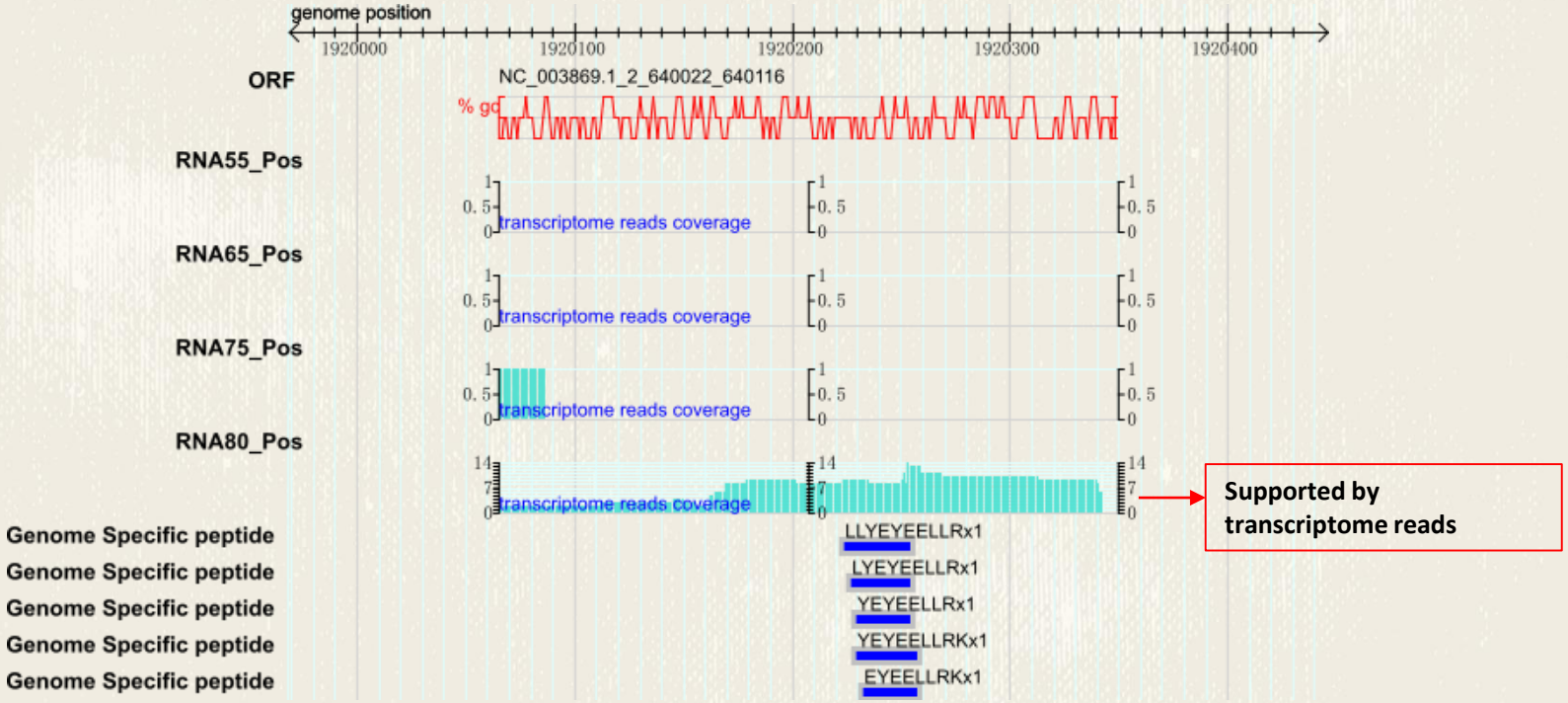
- Train Augustus software with MS identified peptides added to theoretical database
- Reannotated TTE genome: 2625 protein coding genes
- 312 novle genes, 483 corrected translation initiation codons, 368 corrected ORFs

Gene set	integrated peptides	integrated peptides (%)
NCBI	109494	96.9
AUGUSTUS (with NCBI & Peptides)	111179	98.4
CURATED	111234	98.4

The corrected TTE genome map profile



Some corrected genes expressed only at high temperature



Some peptides and RNA reads are expressed only at 80° TTE, suggesting they may represent the thermophilic features of TTE

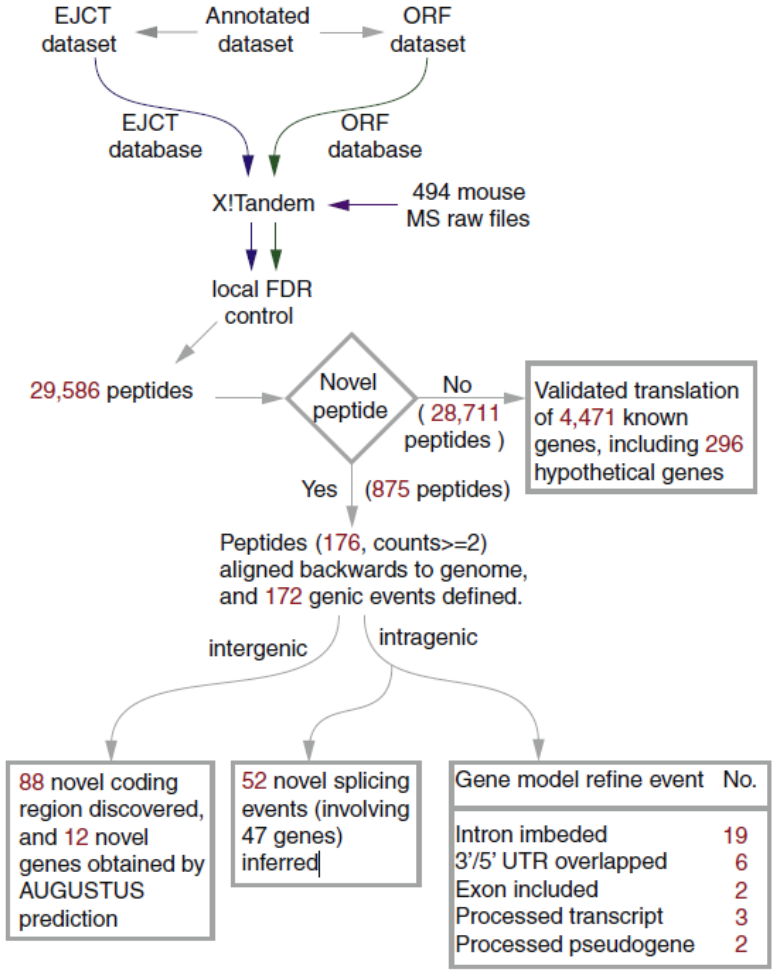
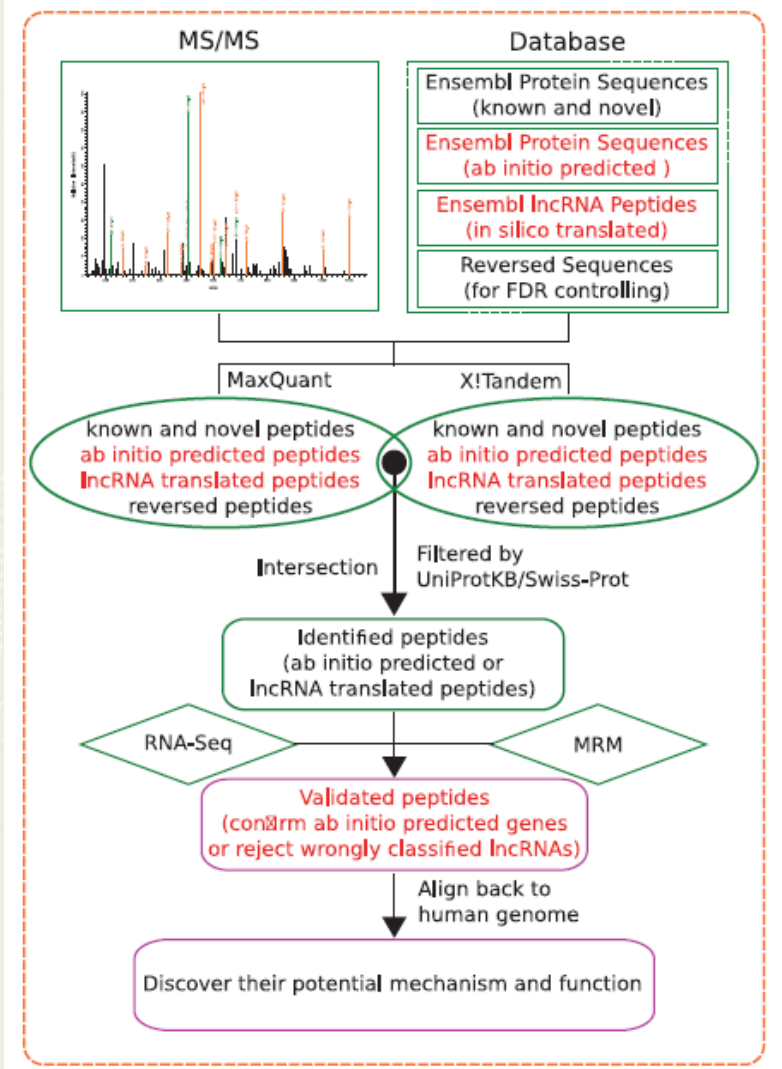


Fig. 1. The analysis pipeline for discovery of novel protein-coding features in mouse genome by proteogenomics strategy.

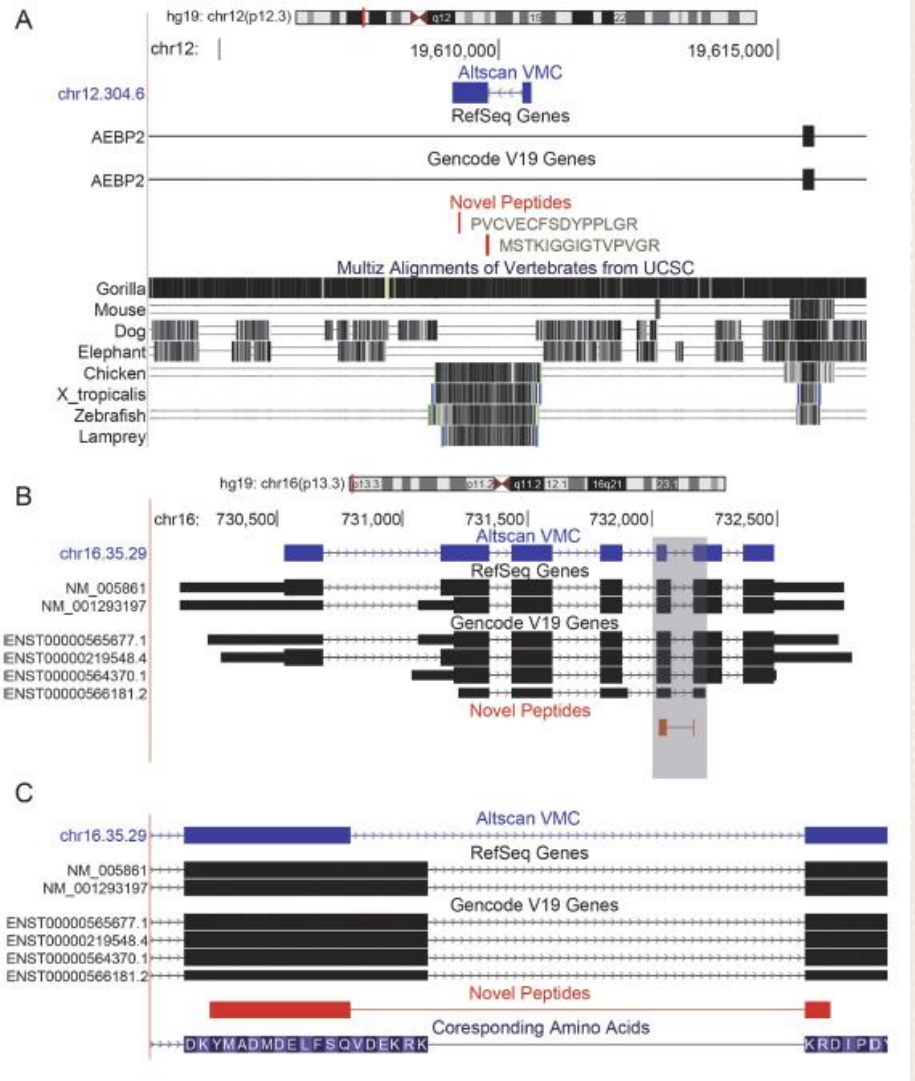
Proteogenomics (genome and peptides): mouse genome, exon junction and ORF

Xing X. B. , et al. Genomics 2011, 98: 343–351



Proteogenomics (genome and peptides): human genome, ab initio predicted genes, lncRNAs

Sun H., et al.
Proteomics 2014,
14, 2760–2768



Proteogenomics (genome and RNA-Seq and peptides): predicted transcripts, splicing isoforms

Hu Z.Q., …Qin G.R., et al. Scientific Reports 2015, 5:10940

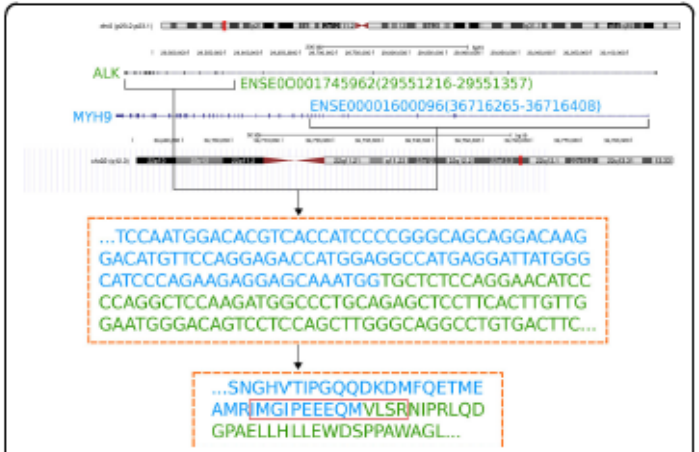


Figure 2 One example of gene fusion event between MYH9 and ALK. ENSE00001745962 of ALK is joined with ENSE00001600096 of MYH9. Both the DNA and peptide sequences are shown, with the identified peptide displayed in a red rectangle. The fusion point is separated by different colors.

important functions, such as cytokinesis, cell motility and maintenance of cell shape. Defects in this gene have been associated with non-syndromic sensorineural deafness autosomal dominant type 17 [22], Epstein syndrome and so on [23]. *ALK* normally locates on the complement strand of chr2, and encodes a receptor tyrosine kinase. Many translocations have been found with this *ALK* gene, including *EML4:ALK* which is responsi-

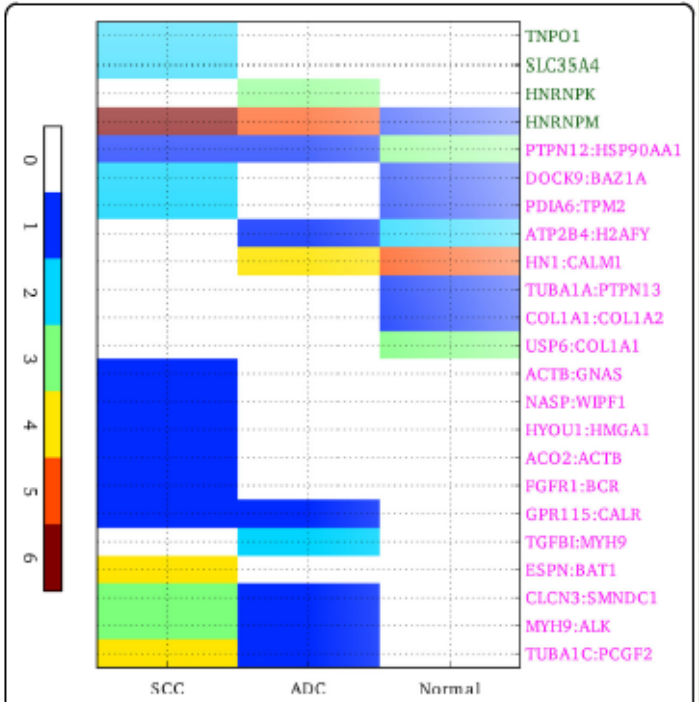


Figure 4 Distribution of the identified fusion or splicing events among subtypes of NSCLC: SCC (squamous cell carcinoma), ADC (adenocarcinoma), and Normal lung samples. The two genes in the fusion events are separated by colon and displayed in magenta and the genes related to alternative splicing are displayed in green. The value in the color bar indicates the number of spectrums of the identified peptide.

Proteogenomics (genome and peptides): gene fusion, splicing, human lung cancer

Sun H., et al. BMC Genomics 2013, 14(Suppl 8):S5

03

Identification of SAPs

Proteogenomics: for tumor genome mutations (SAPs)

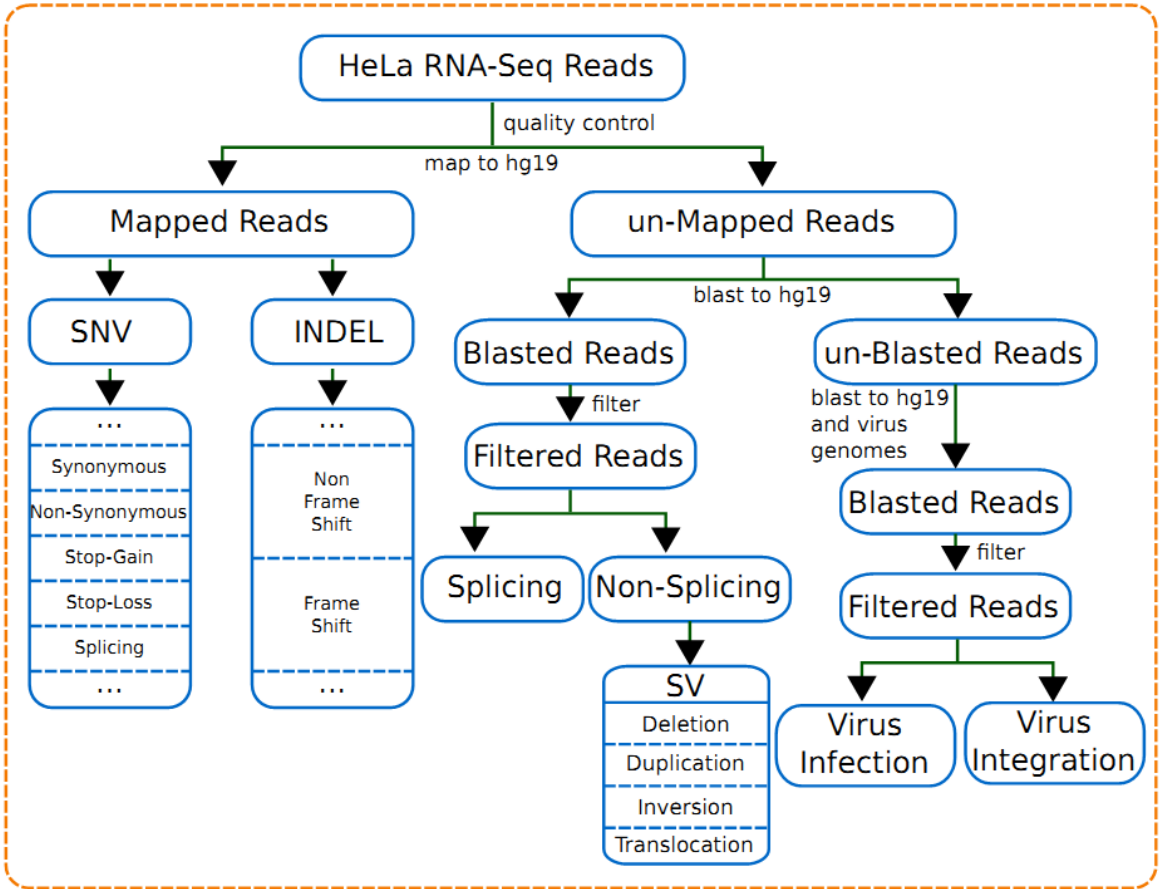
Hela cell line: human cervical carcinoma

Gene mutations and indels

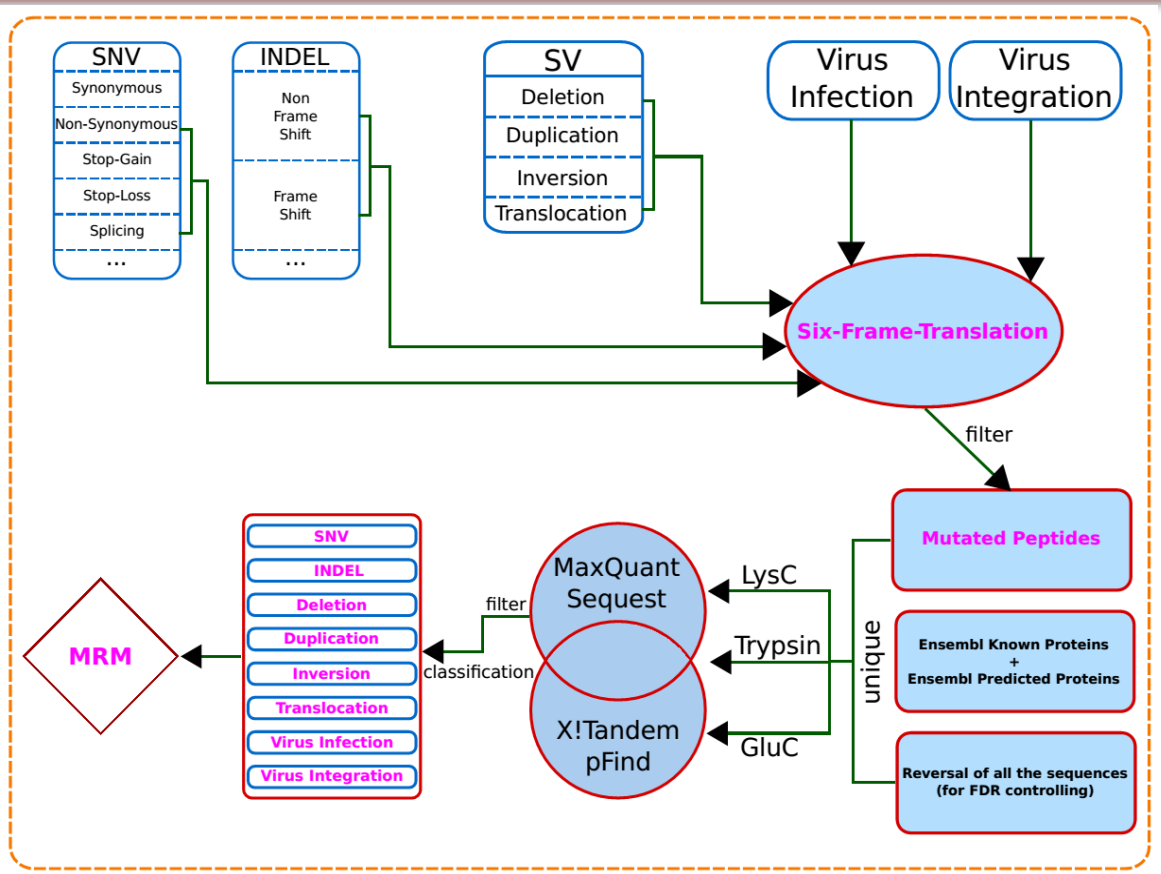
HPV infection and virus-human genome integration

Protein level evidence

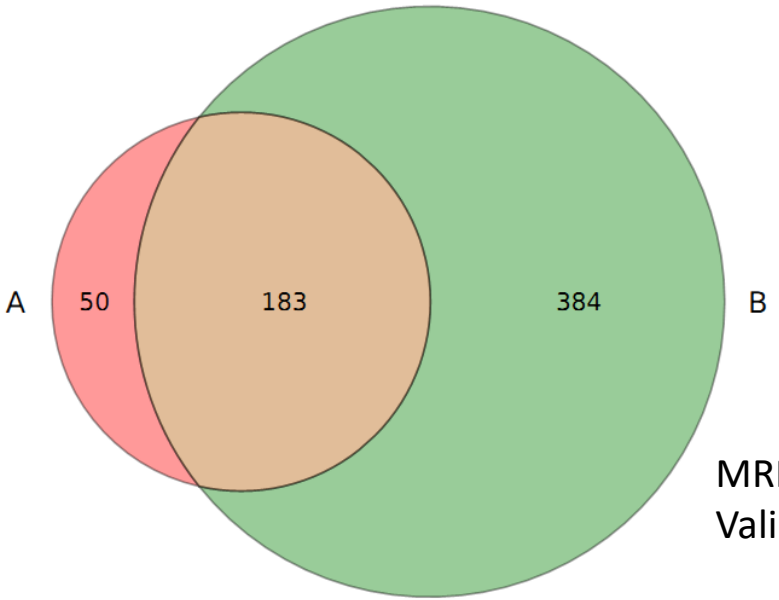
RNA-Seq: transcriptome expressed genome variations



MS/MS peptide confirmation of genome variation and virus integration



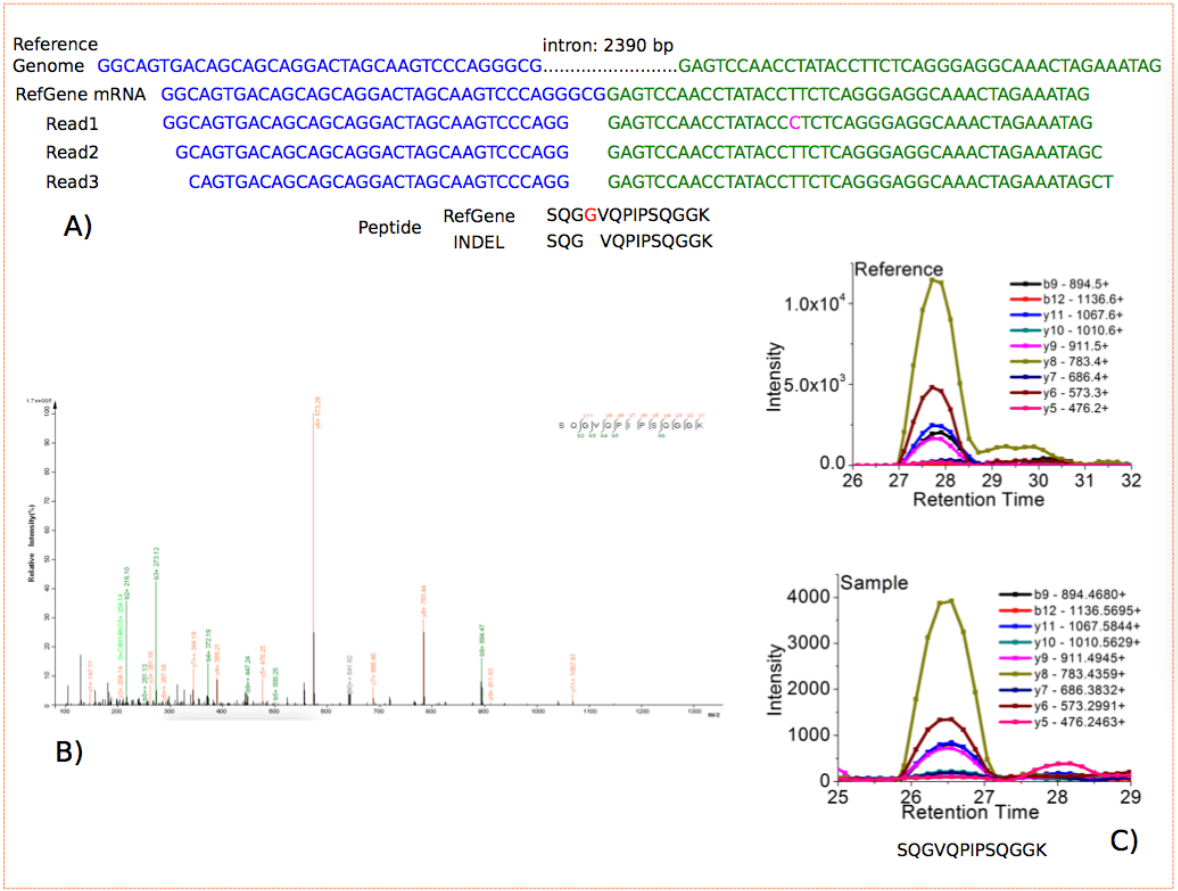
Identified mutated peptides by high accuracy MS/MS

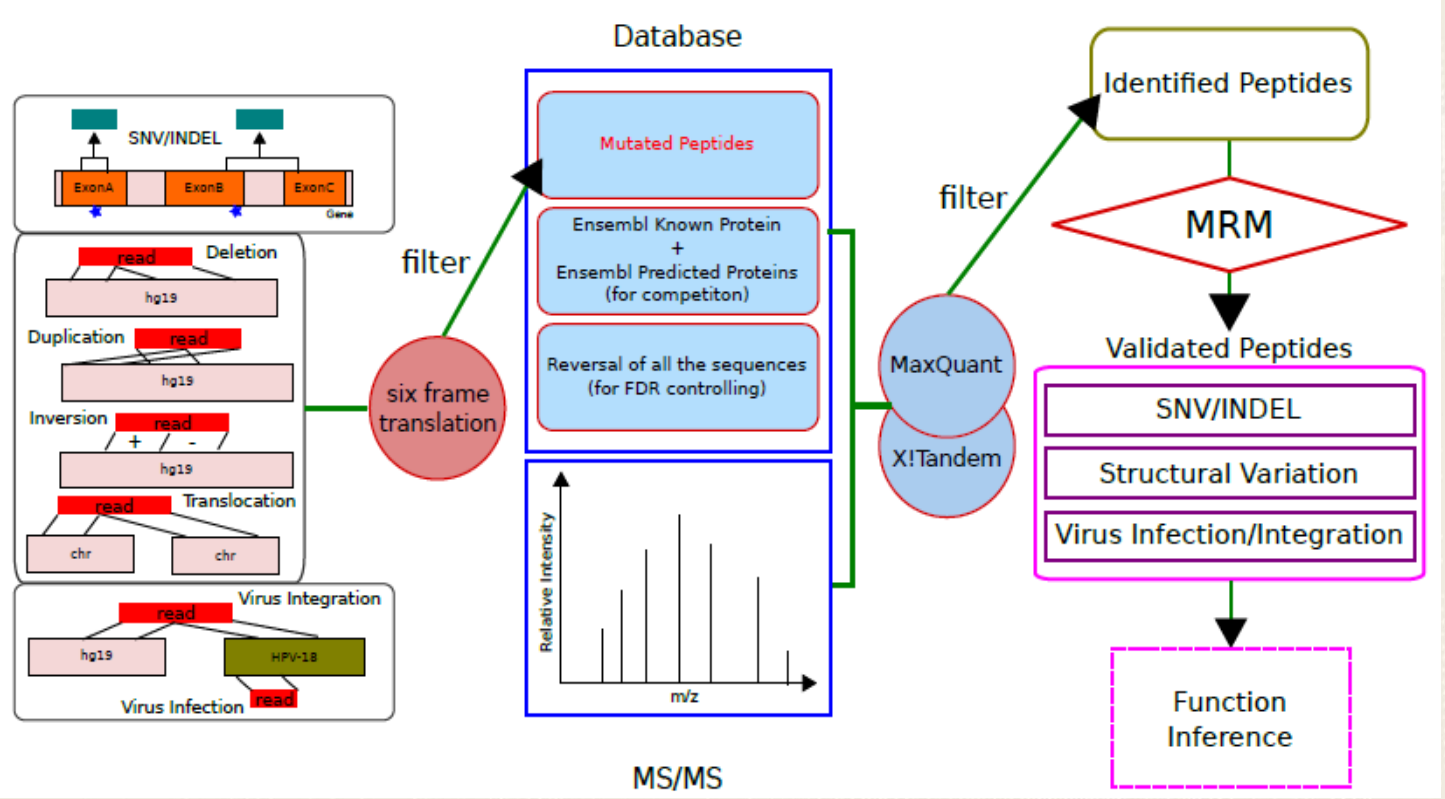


A: Maxquant & Sequest
B: X!Tandem & pFind

MRM: 183 peptides
Validated: 84

Fam120a deletion: RNA-SEQ, MS/MS and MRM evidences





Proteogenomics (RNA-Seq and peptides): mutation, virus integration, human cervical cancer

Sun H., et al. J. Proteome Res., **2015**, *14* (4), pp 1678–1686

04

Neoantigen discovery

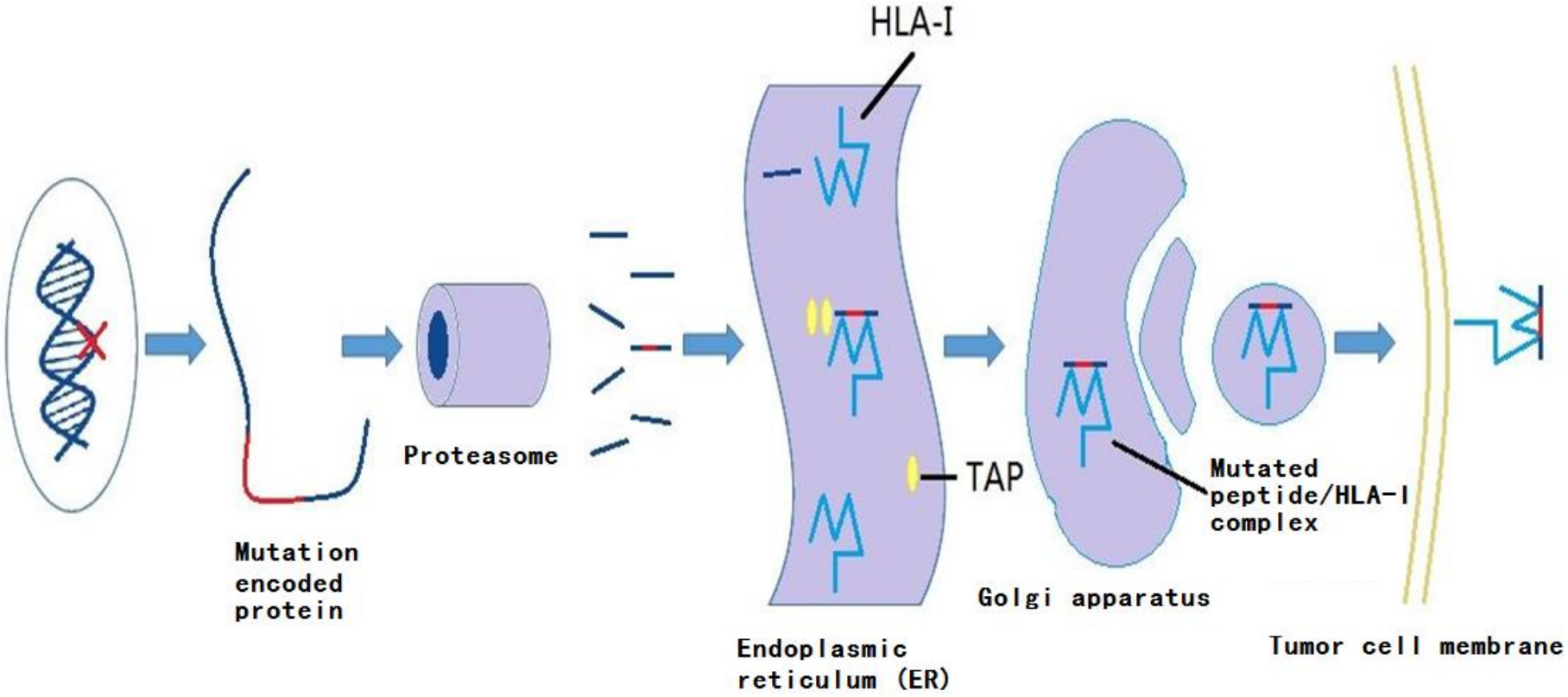
Neoantigen is SAP

- **Tumor neoantigens are altered proteins caused by tumor cell somatic mutations**
- **Neoantigens are predicted from tumor exome/transcriptome sequencing data**
- **Proteogenomics may provide peptide evidence for neoantigen prediction by mass spectrometry, may enhance tumor neoantigen prediction and verification**
- **Tumor neoantigen can induce specific anti-tumor cellular immunity; can enhance immunotherapy; are important targets for personalized immunotherapy**

SAP to become Neoantigen

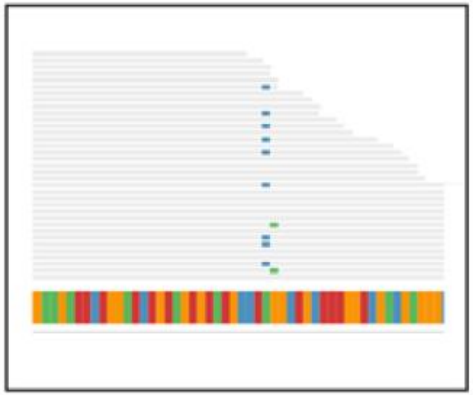
- **Able to be presented as antigen (binding to HLA-I)**
- **Able to be recognized by T cells (be compatible to TCR structure)**
- **Be immunogenic (similar to ectogenic microbial Ag to cause T cell reactivity)**
- **Do not cause destruction to other self proteins (dissimilar to endogenic proteins)**

Neoantigen presentation: HLA-I binding and transporting



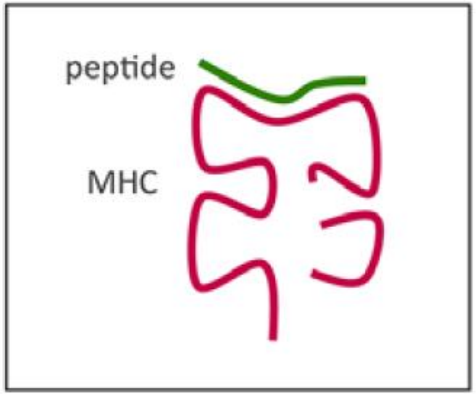
Neoantigen essentials: variant calling, HLA-binding, TCR recognization

What variant sequences are unique to the tumor?



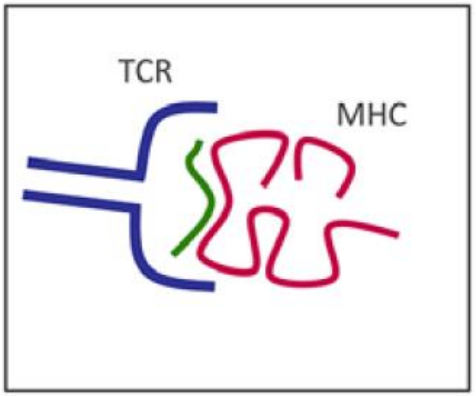
- Biopsy quality
- Sequencing strategy
- Mutation calling algorithms

Which peptides are most likely to be presented on MHC?



- Expression / turnover
- Processing
- MHC binding tumor-specific factors

Which peptides are most likely to drive a T cell response?



- Class I vs. class II
- Mutant vs. wildtype
- Clonal vs. subclonal
- Oncogene vs. passenger

Workflow of a neoantigen prediction pipeline

Exome

Tumor DNA
Normal DNA

(1) HLA typing

HLA-A1
HLA-A2
HLA-B1
HLA-B2
HLA-C1
HLA-C2

NetMHC/netMHCpan
Mut IC_{50} < 500 nmol/L

(2) Variant call

Somatic mutations

RNAseq
Tumor RNA

(3) Expression

Translate to amino acid, and generate 8 – 11-mer peptides

Possible neoepitopes

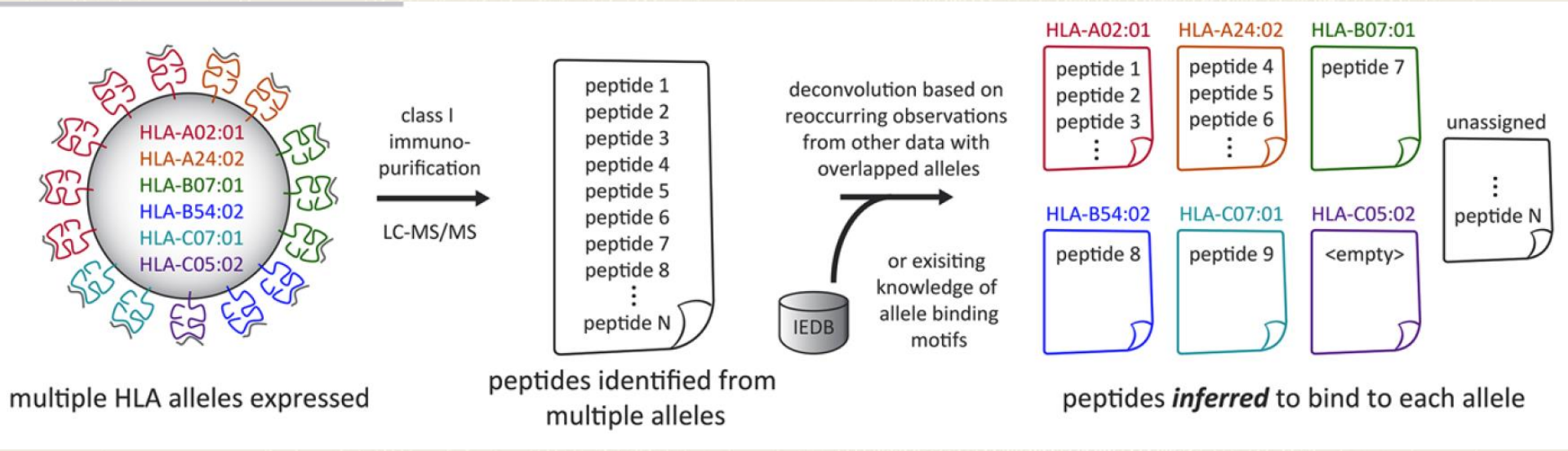
Amino acid substitution

(4) Neoantigen prediction

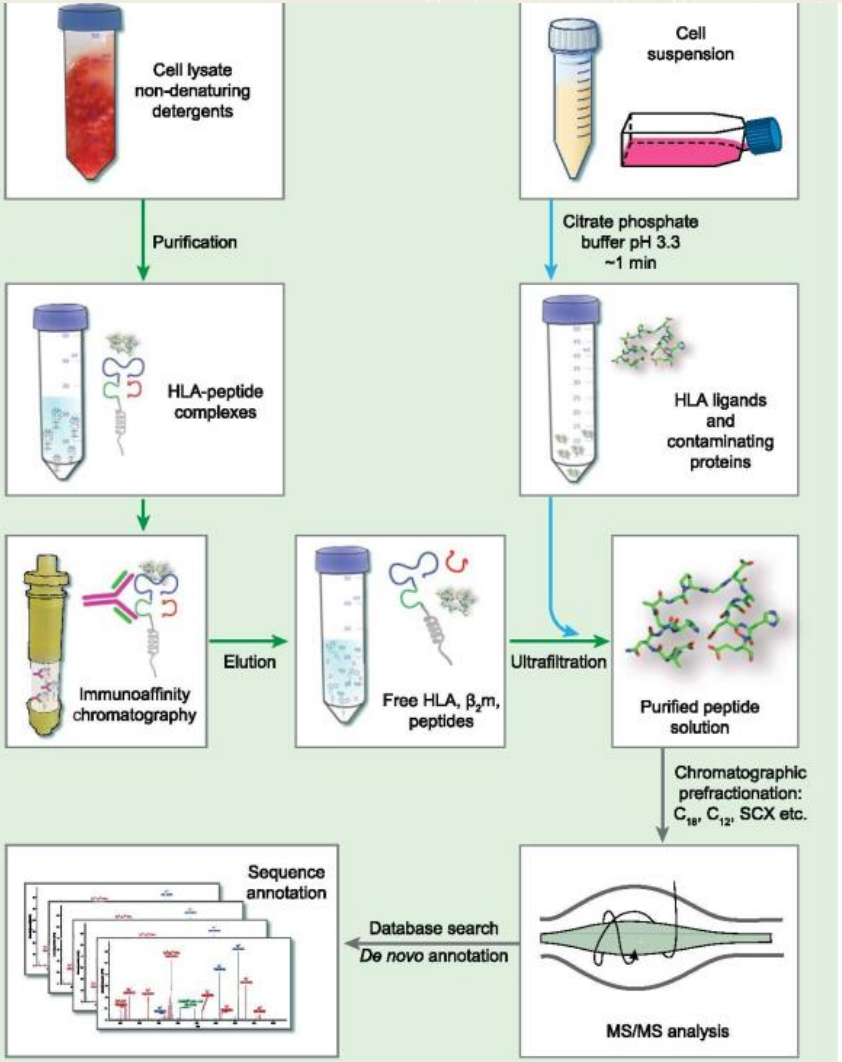
Mutation		Mut peptide (nM)		WT peptide (nM)	
PKD2L1	V548I	AYFVTYVFFI	31	AYFVTYVFFV	201
GPLD1	S387L	PYARLGWAMTL	56	PYARLGWAMTS	3844
RP11	S553F	KFLAAAHNF	56	KFLAAAHNS	19 392
TP53	C242Y	SYMGGMNRRPI	82	SCMGGMNRRPI	14,551
LAD1	R424W	SWGLPCTELF	84	SRGLPCTELF	12 175
LRP1B	L1903F	AFMPISGTSF	108	ALMPISGTSF	1689

Cancer Science. 2018;109:542–549

HLA-I binding peptides profiling

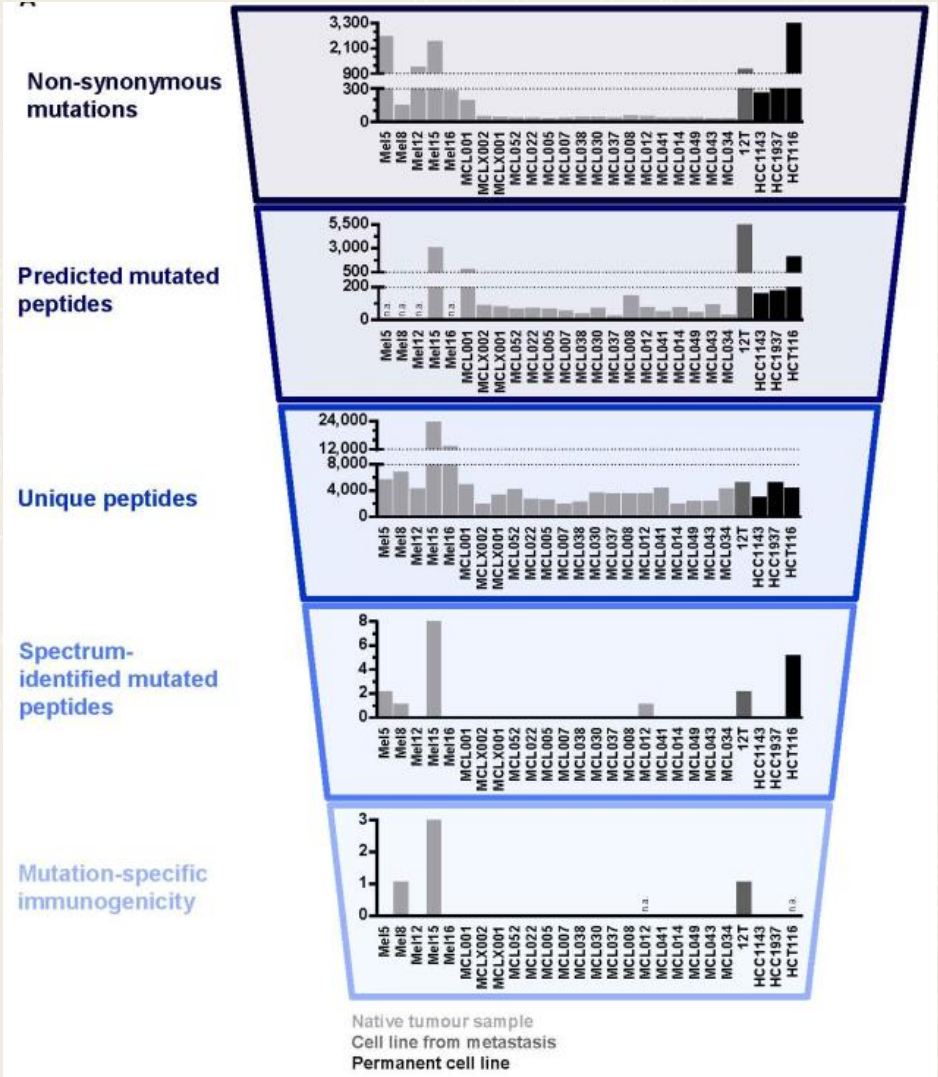


Proteomics. 2018, 18, 1700259



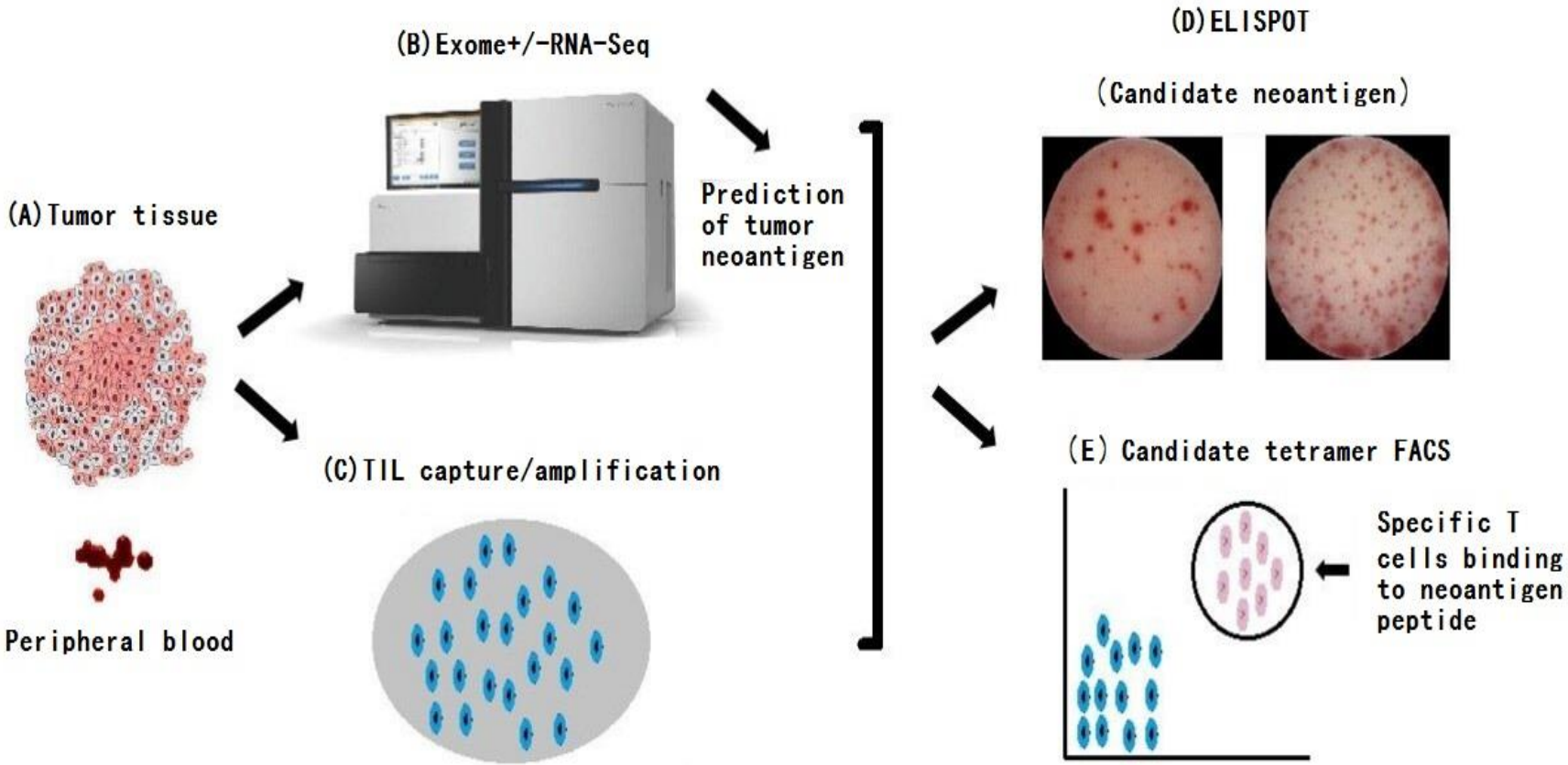
Proteomics identification of HLA-binding peptides enriched by immunoprecipitation(IP, left) or mild acid elution(MAE, right)

Immunology. 2018, doi: 10.1111/imm.12936

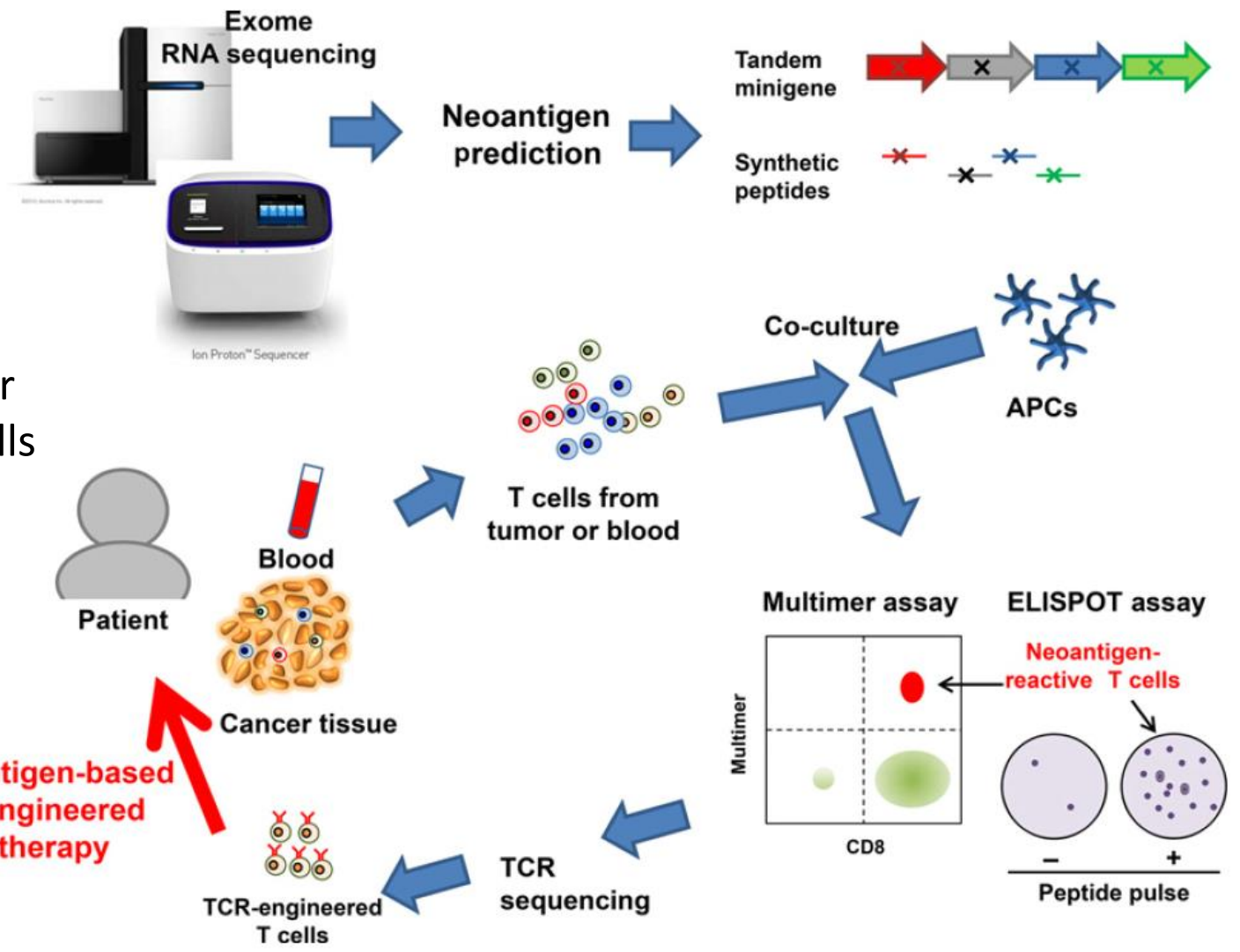


Five levels of class I neoantigen discovery

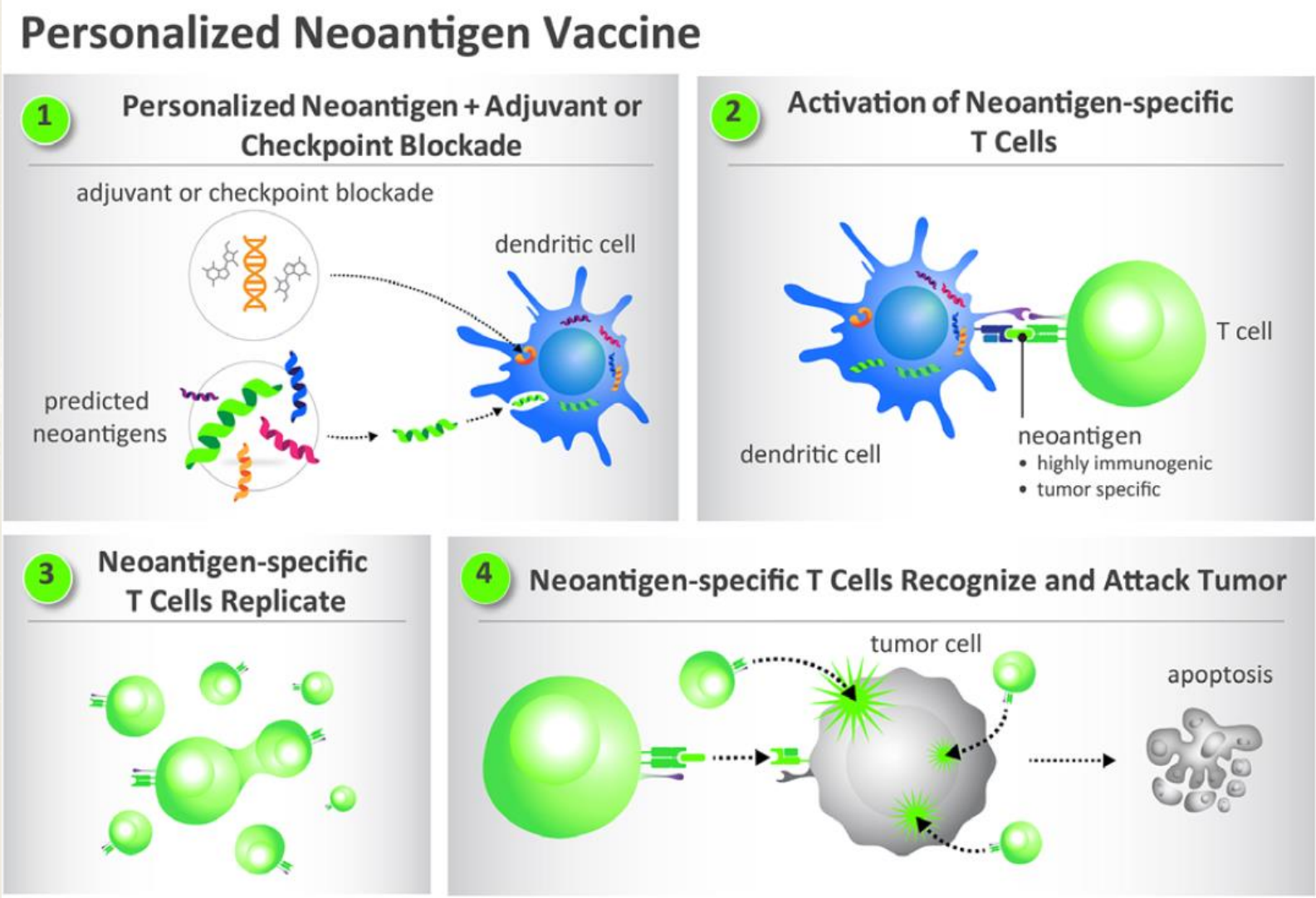
Neoantigen immunogenicity verification



Neoantigen based immunotherapy I: develop T-cell receptor (TCR)-engineered T cells



Neoantigen based immunotherapy II: development of personalized neoantigen vaccines




Thank you
for your
attention