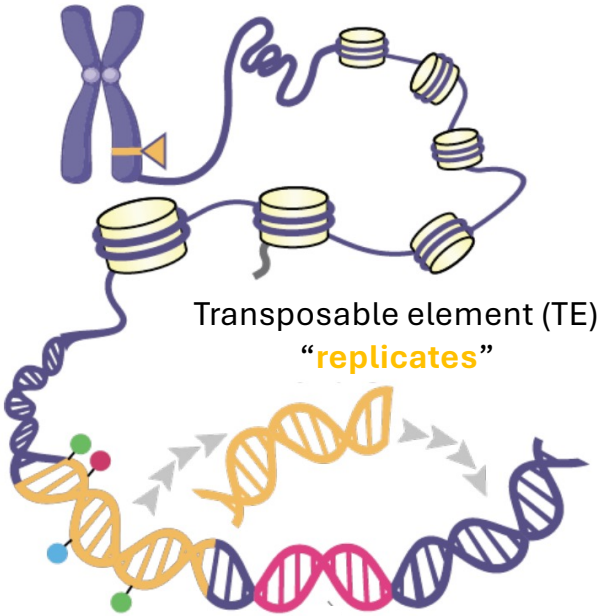
Non-coding RNAs and Transposable Elements



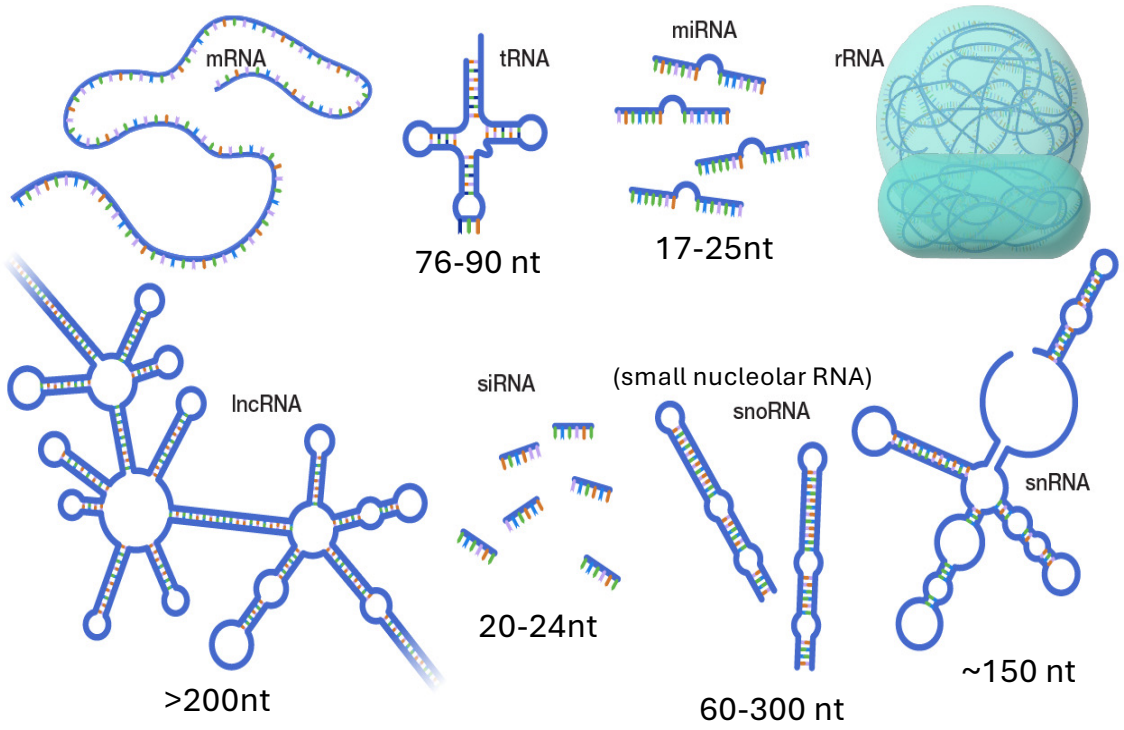
Simon Chu
The Wistar Institute
02/24/2026

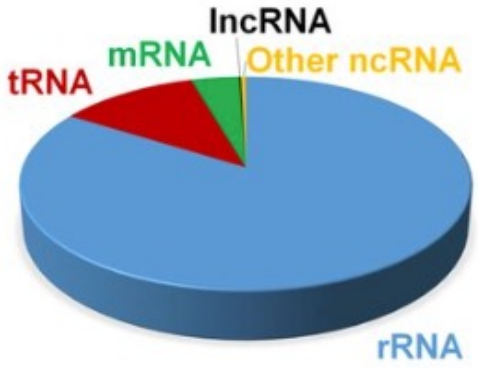
Outline

- **Non-coding RNAs**
  - Types of non coding RNAs
  - Summary of their functions

- **Transposable elements (TEs)**
  - TE annotation
  - TE insertion identification
  - TE RNA expression quantification
  - TE derived neoantigen for mRNA cancer vaccine

Different types of RNA





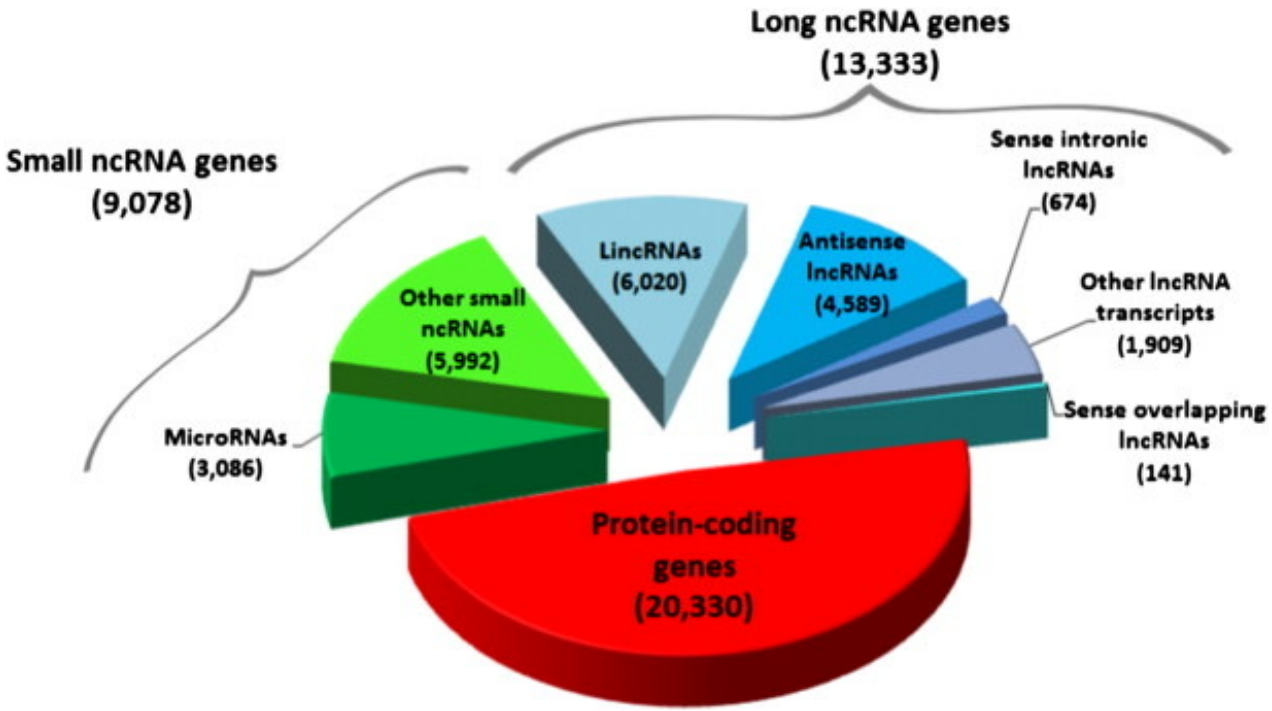
Mammalian cell

THE WISTAR INSTITUTE

Proportion of non-coding genes

In human genome

- ~2% of coding genes
- ~6% of non-coding genes
  - Due to large copies of ribosomal RNA genes

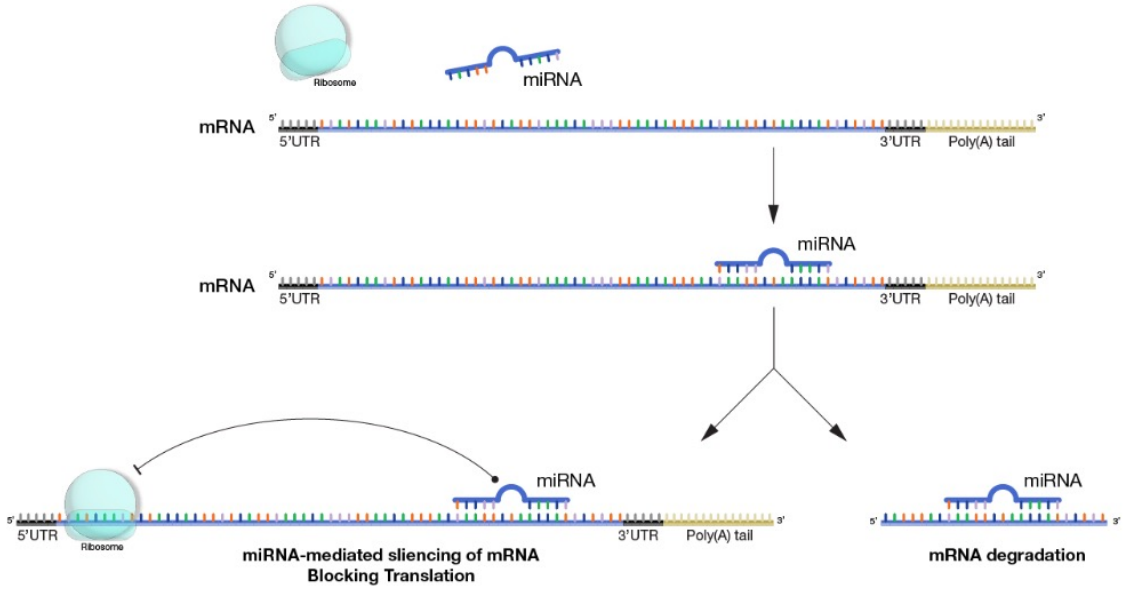


THE WISTAR INSTITUTE

Regulation roles of miRNA and siRNA

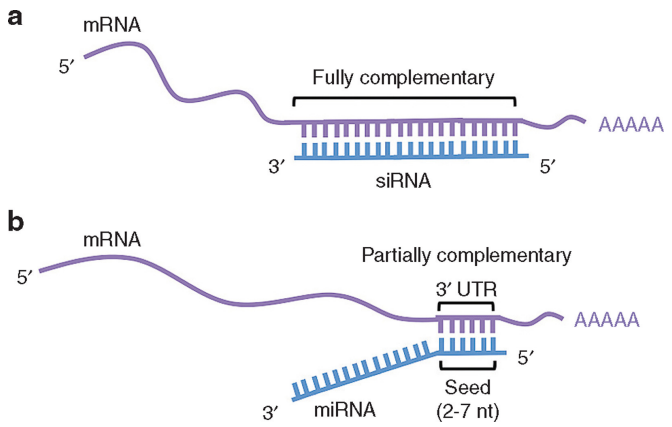
Micro ribonucleic acid (microRNA, miRNA)

- Small, single-stranded, non-coding RNA
- 17-25 nt



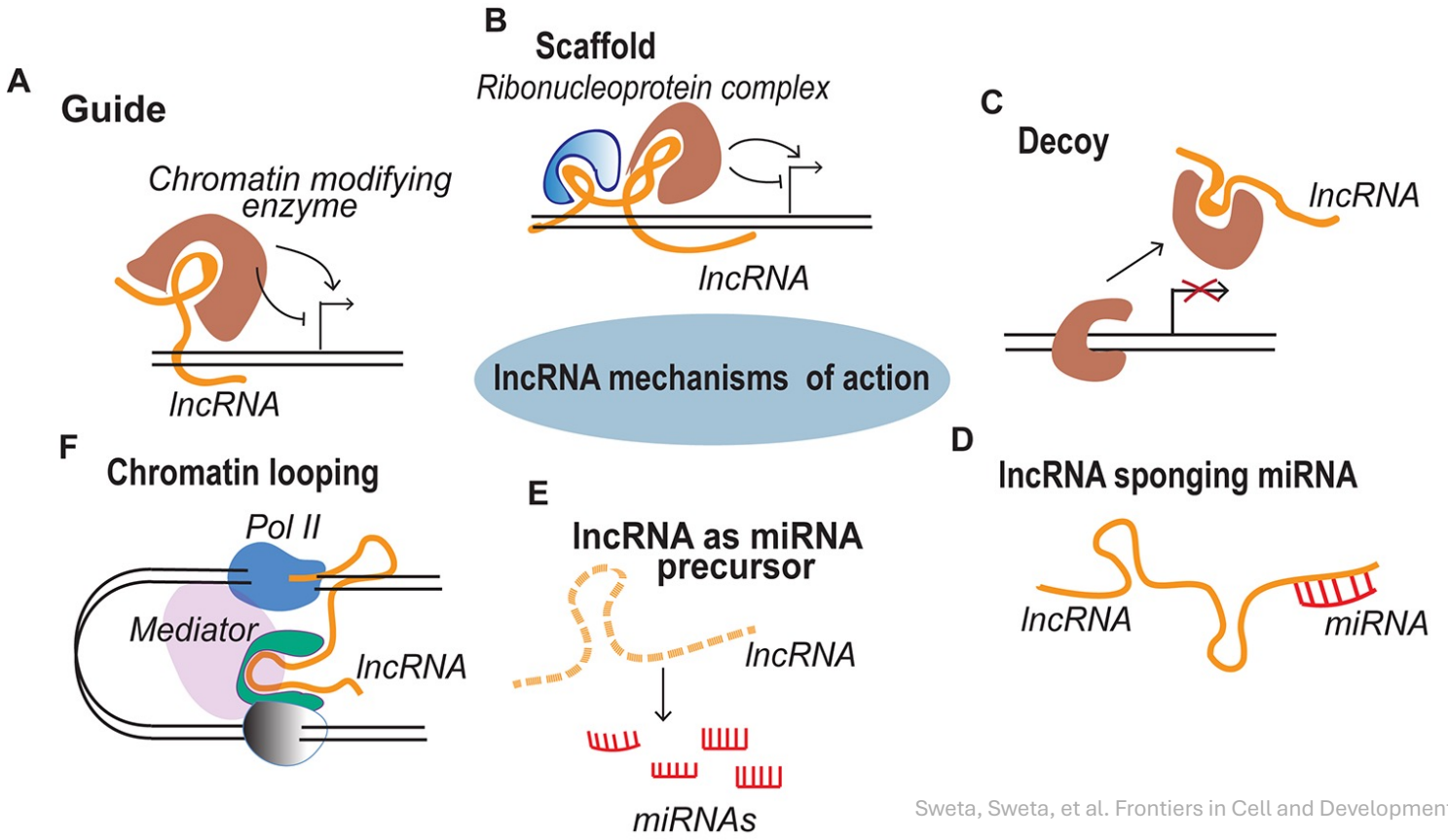
Small interfering RNA (siRNA)

- Small, double-stranded, non-coding RNA
- 20-24 nt



THE WISTAR INSTITUTE

lncRNA mechanisms of action

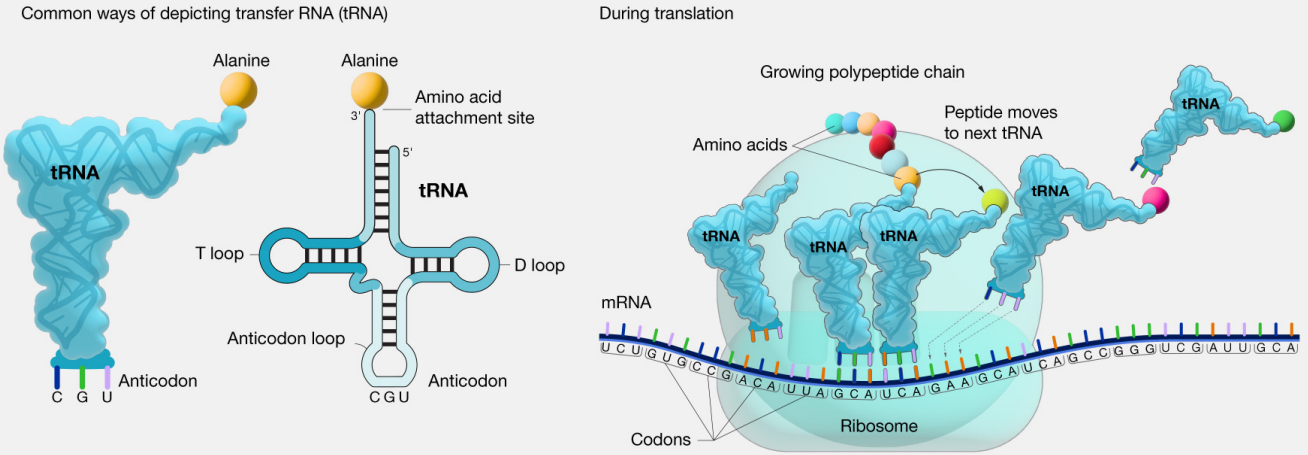


Sweta, Sweta, et al. Frontiers in Cell and Developmental Biology 2019

THE WISTAR INSTITUTE

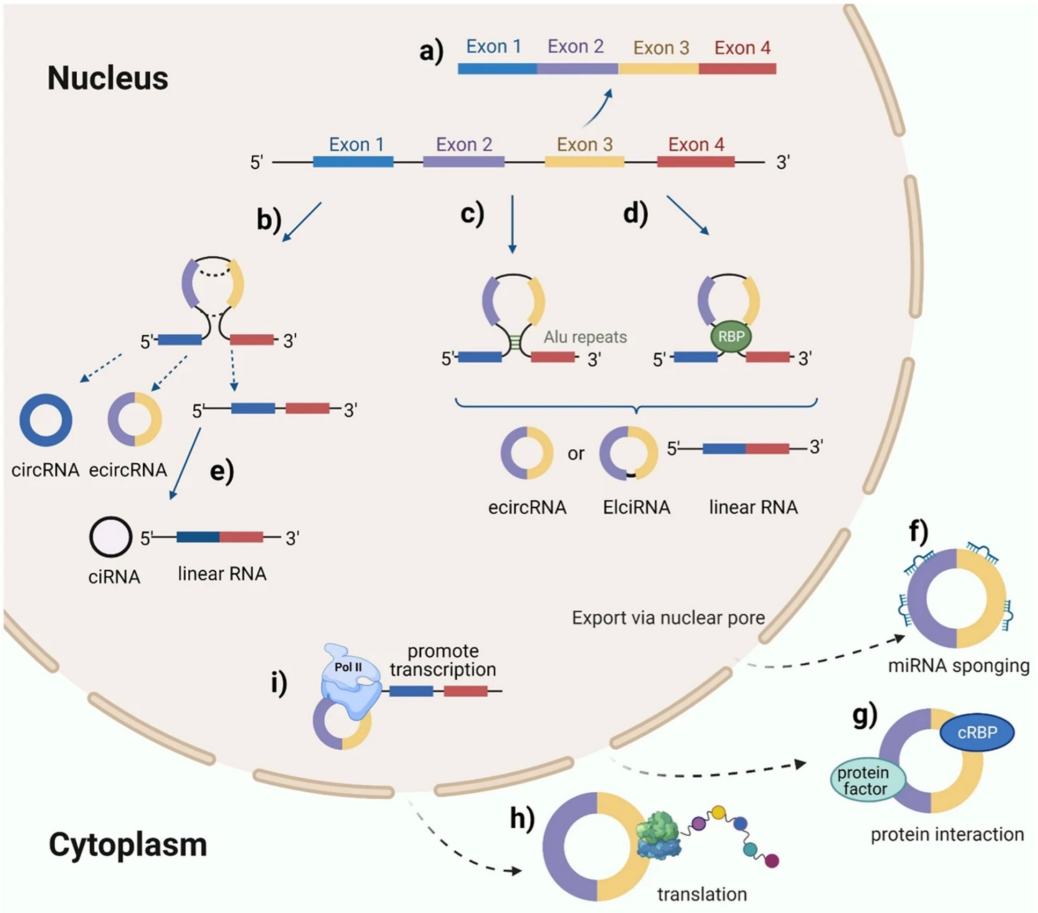
tRNA is essential in protein synthesis

- Transfer ribonucleic acid, typically 76-90 nt
- Provide physical link between mRNA and amino acid



THE WISTAR INSTITUTE

Circular RNAs: Formation and functional impact



Misir, S., Wu, N. and Yang, B.B., 2022. Cell Death & Differentiation

THE WISTAR INSTITUTE

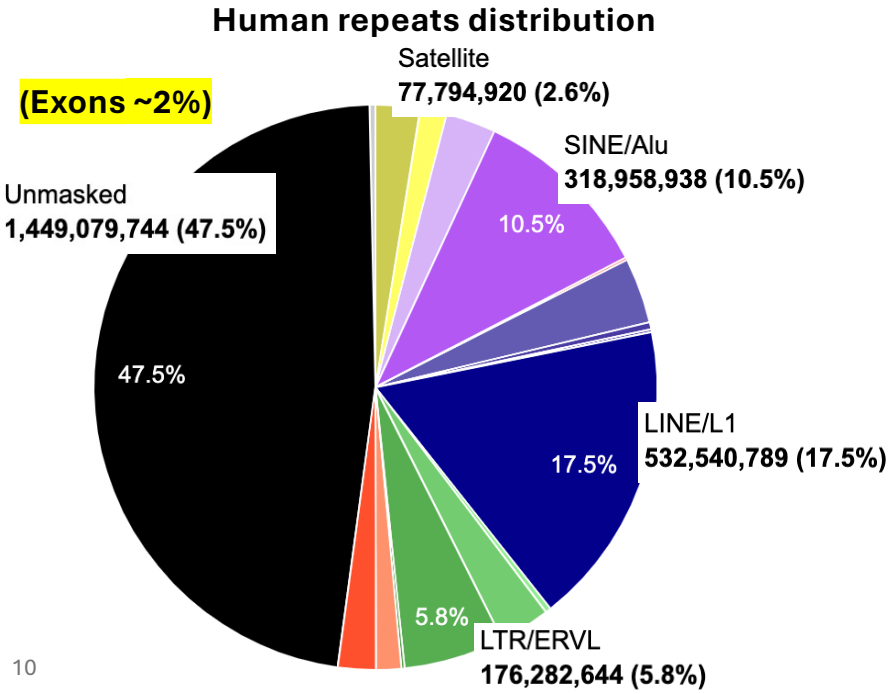
Outline

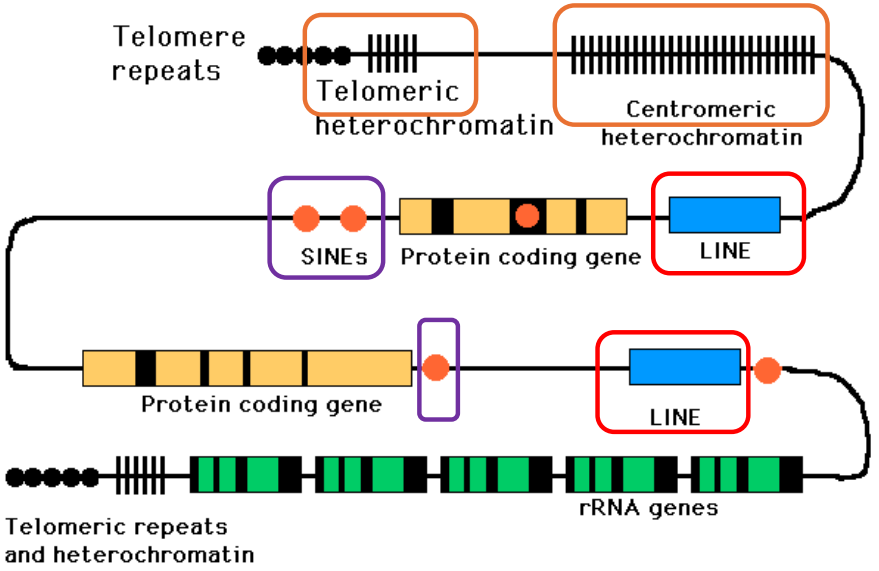
- **Non-coding RNAs**
  - Types of non coding RNAs
  - Summary of their functions

- **Transposable elements (TEs)**
  - TE annotation
  - TE insertion identification
  - TE RNA expression quantification
  - TE derived neoantigen for mRNA cancer vaccine

Genomic repeats take half of the human genome

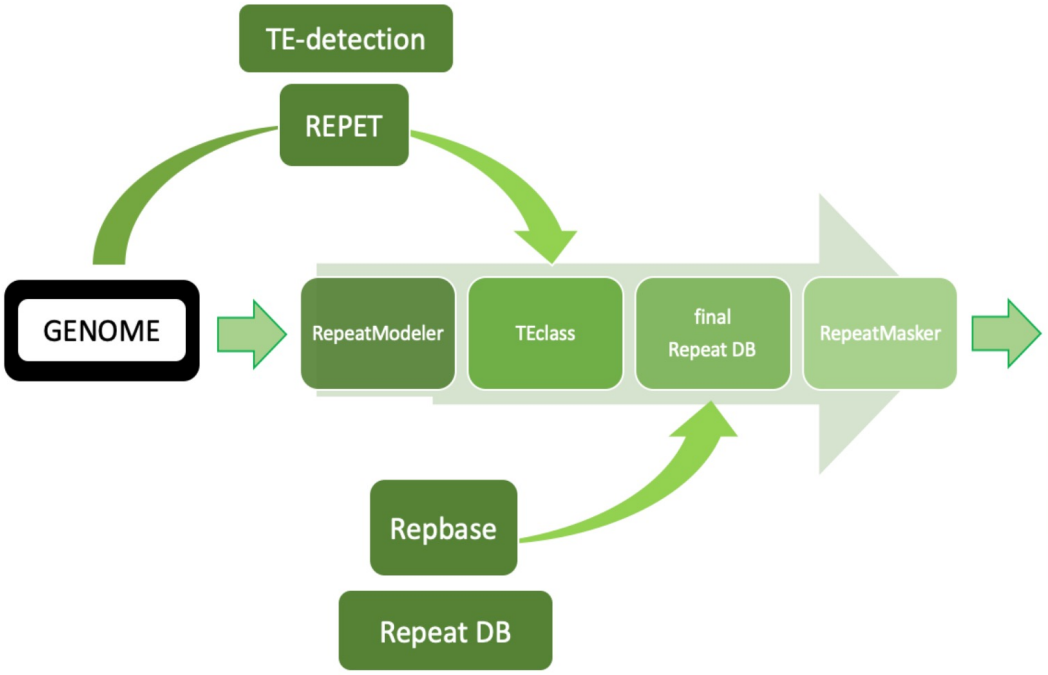
- Transposable elements (TEs):
  - Long interspersed elements (LINE), short interspersed elements (SINE), LTR, DNA transposons
- Tandem repeats and large satellite repeats





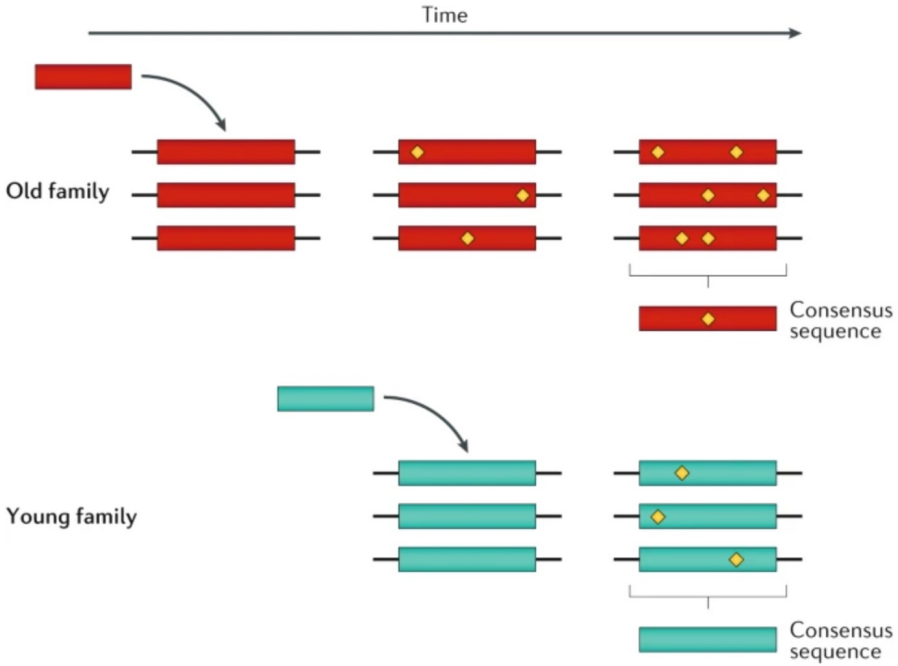
THE WISTAR INSTITUTE

Genomic repeats annotation with RepeatMasker



Repeat Annotation Results	
Total interspersed repeats	**15.89 %**
DNA	8.69 %
LINE	1.42 %
LTR	4.38 %
SINE	0.03 %
Unknown	1.37 %
Low complexity	**0.53 %**
Simple repeats	**3.98 %**
Total	**20.25 %**

Genomic repeats annotation with RepeatMasker

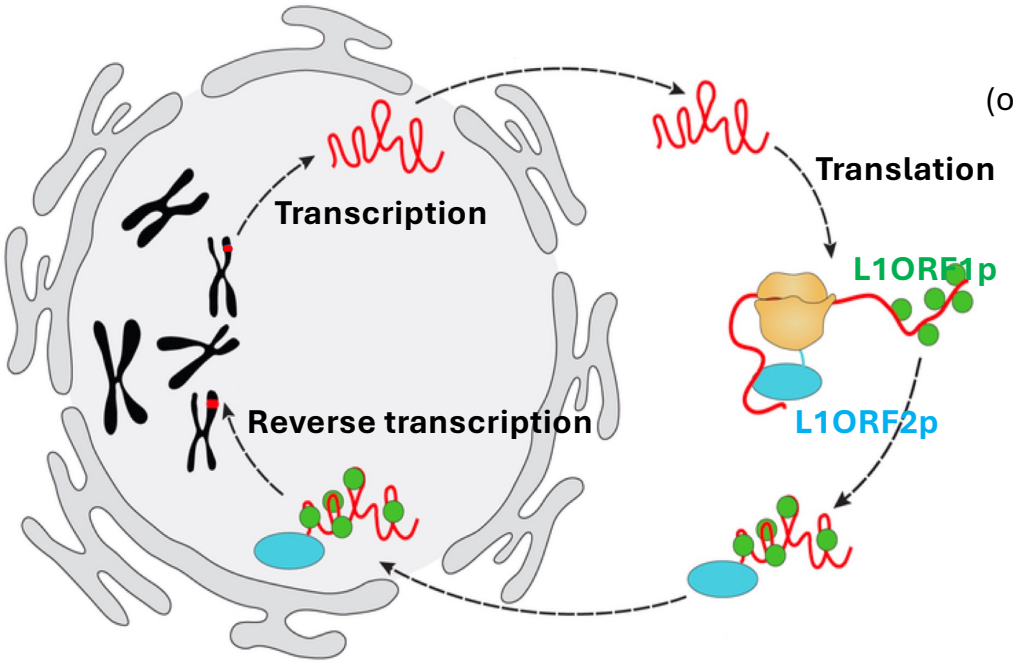



```
 1306 15.6  6.2  0.0 HSU08988  6563  6781  (22462) C  MER7A    DNA/MER2_type    (0)   336   103
12204 10.0  2.4  1.8 HSU08988  6782  7714  (21529) C  TIGGER1  DNA/MER2_type    (0)  2418  1493
  279  3.0  0.0  0.0 HSU08988  7719  7751  (21492) +  (TTTTA)n Simple_repeat      1    33    (0)
 1765 13.4  6.5  1.8 HSU08988  7752  8022  (21221) C  AluSx    SINE/Alu        (23)   289     1
12204 10.0  2.4  1.8 HSU08988  8023  8694  (20549) C  TIGGER1  DNA/MER2_type  (925)  1493   827
 1984 11.1  0.3  0.7 HSU08988  8695  9000  (20243) C  AluSg    SINE/Alu         (5)   305     1
12204 10.0  2.4  1.8 HSU08988  9001  9695  (19548) C  TIGGER1  DNA/MER2_type (1591)   827     2
  711 21.2  1.4  0.0 HSU08988  9696  9816  (19427) C  MER7A    DNA/MER2_type  (224)   122     2
```

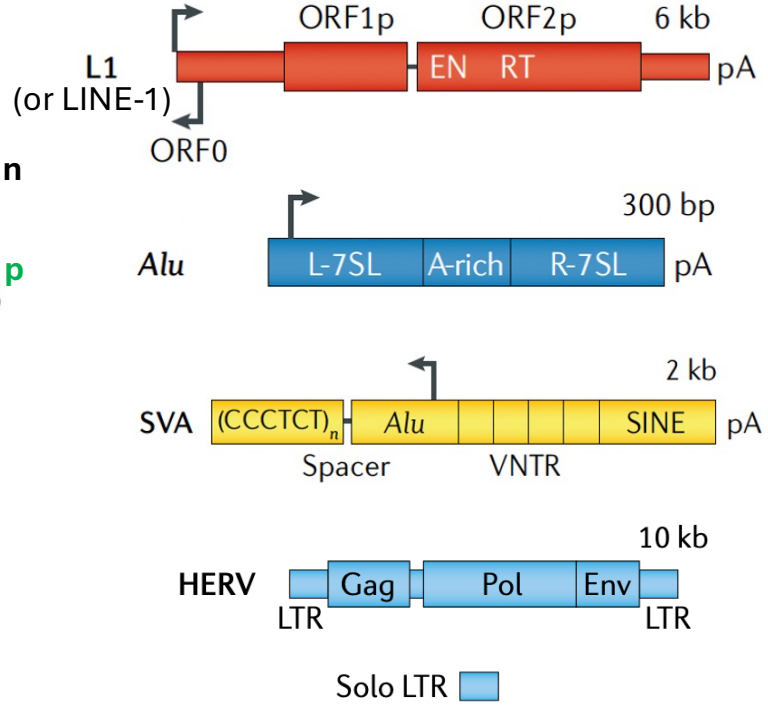
```
1306    = Smith-Waterman score of the match, usually complexity adjusted
      The SW scores are not always directly comparable. Sometimes
      the complexity adjustment has been turned off, and a variety of
      scoring-matrices are used.
15.6    = % substitutions in matching region compared to the consensus
6.2     = % of bases opposite a gap in the query sequence (deleted bp)
0.0     = % of bases opposite a gap in the repeat consensus (inserted bp)
HSU08988 = name of query sequence
6563    = starting position of match in query sequence
7714    = ending position of match in query sequence
(22462) = no. of bases in query sequence past the ending position of match
C       = match is with the Complement of the consensus sequence in the database
MER7A   = name of the matching interspersed repeat
DNA/MER2_type = the class of the repeat, in this case a DNA transposon
          fossil of the MER2 group (see below for list and references)
(0)     = no. of bases in (complement of) the repeat consensus sequence
          prior to beginning of the match (so 0 means that the match extended
          all the way to the end of the repeat consensus sequence)
2418    = starting position of match in database sequence (using top-strand numbering)
1465    = ending position of match in database sequence
```


THE WISTAR INSTITUTE

Retrotransposons "mobilize" from one location to another



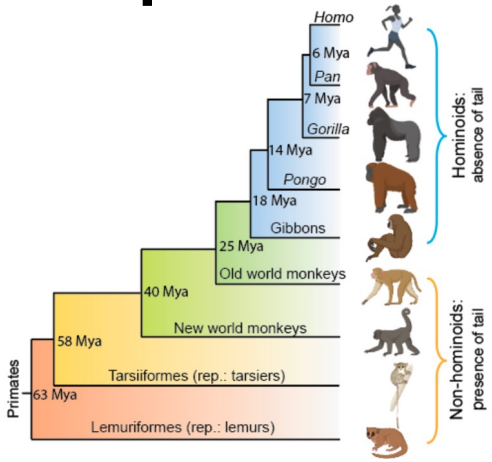
Transposable elements are ancient viruses

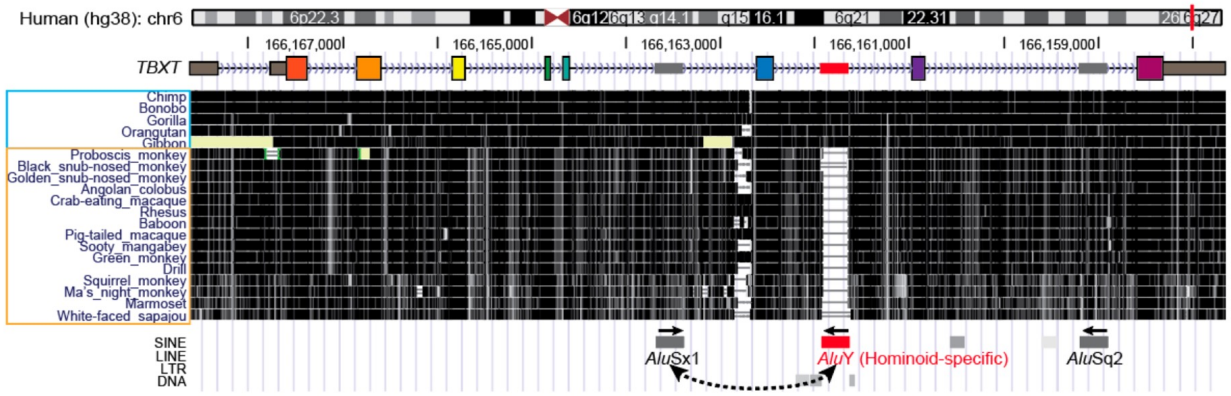


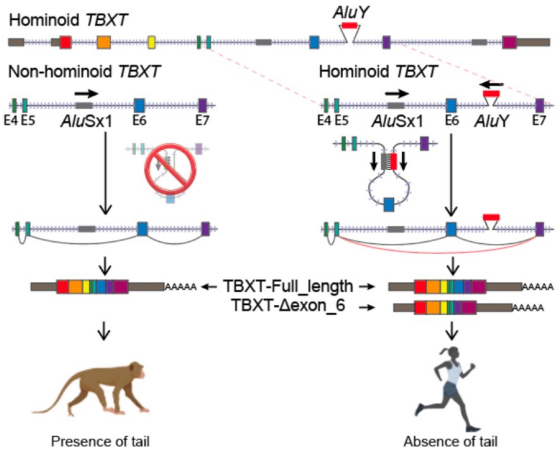
Payer and Burns, 2019, Nature Reviews Genetics

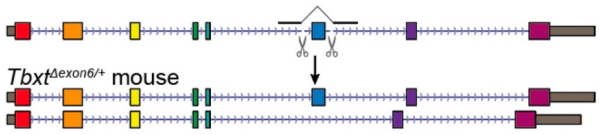
THE WISTAR INSTITUTE

Roles transposable elements are playing (Example: where is our tail?)





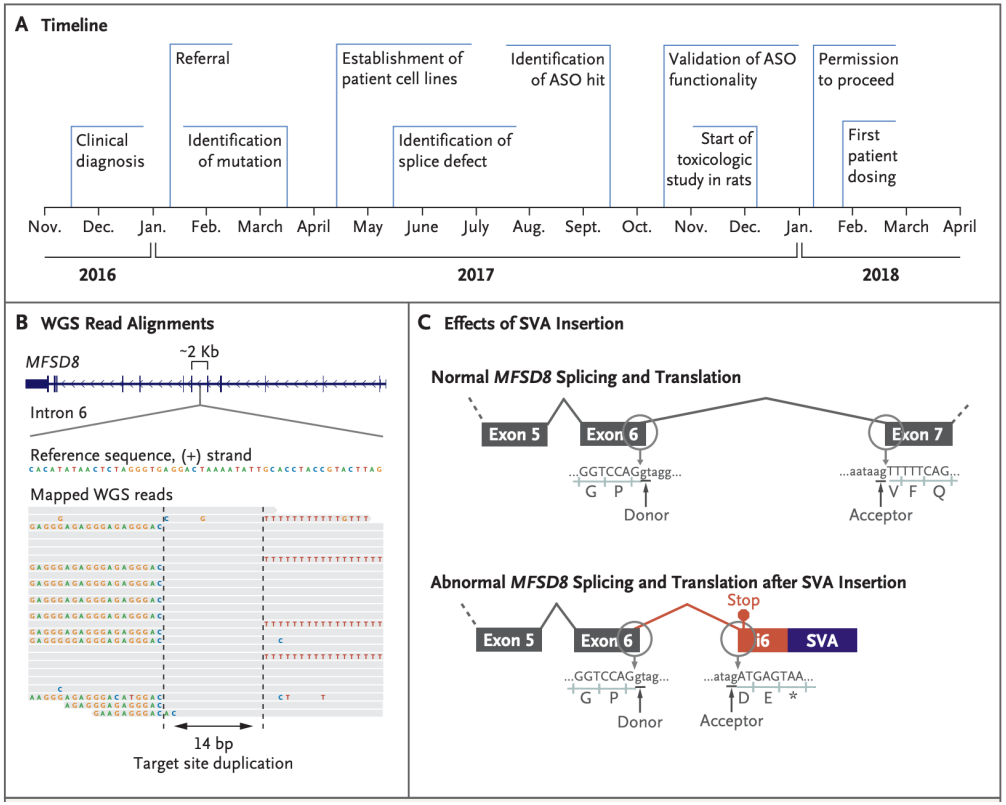


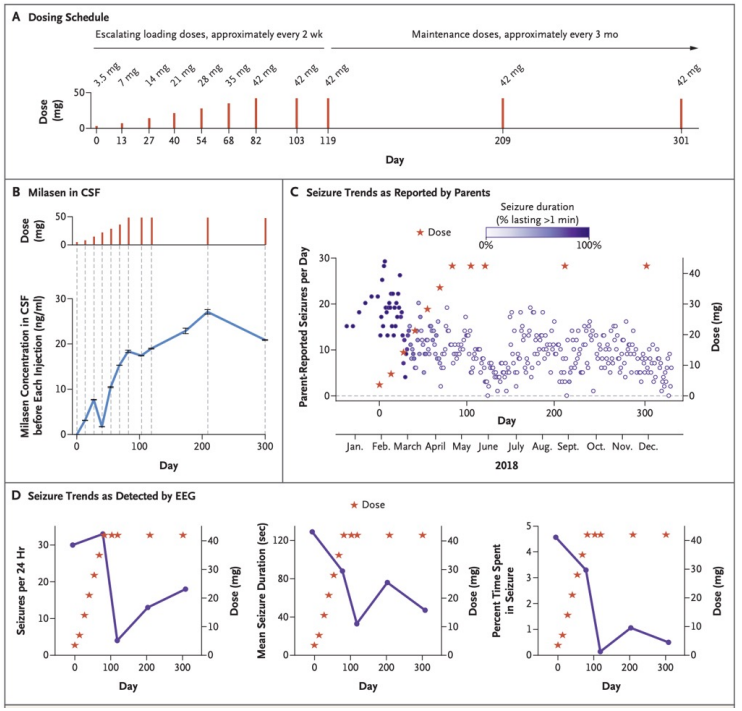


Xia, Bo et. al. *Nature* 2024

THE WISTAR INSTITUTE

An SVA insertion causes Batten disease (rare disease)

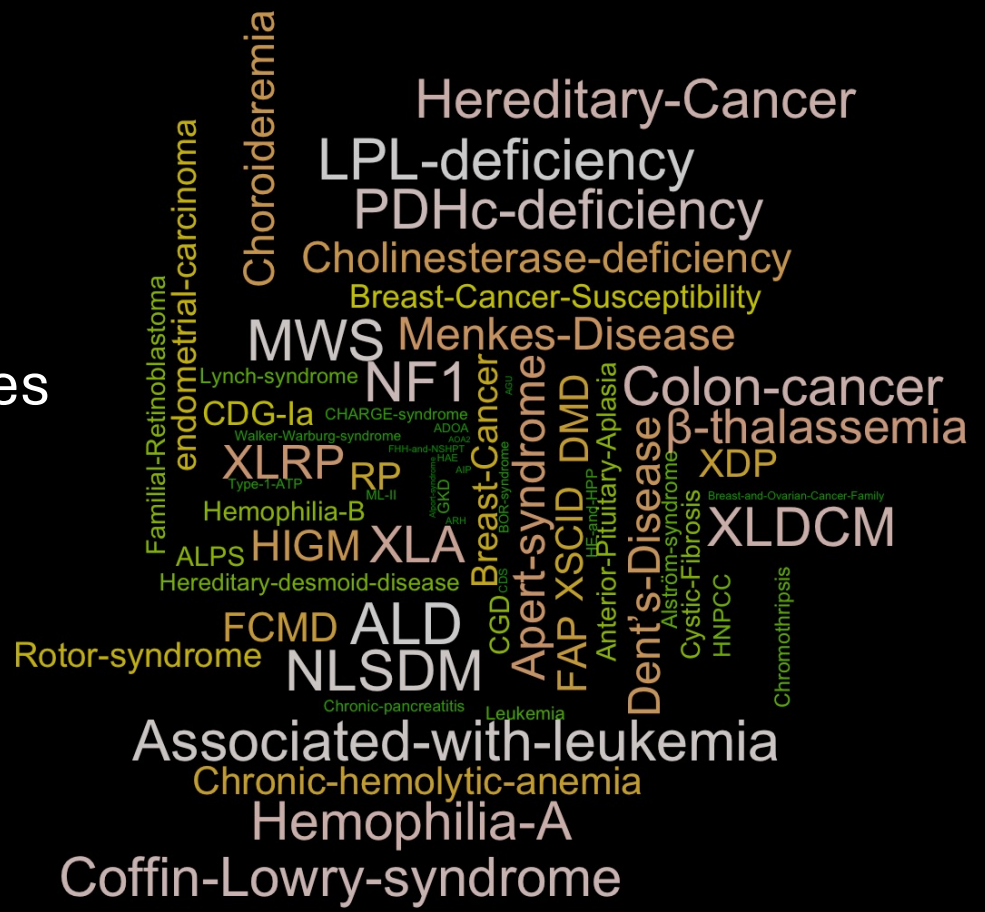




Kim, Jinkuk, et al. *New England Journal of Medicine* (2019).

THE WISTAR INSTITUTE

More than **65** human diseases directly caused by retrotransposon insertions



Dustin C. and Kazazian. *Mobile DNA*, 2016

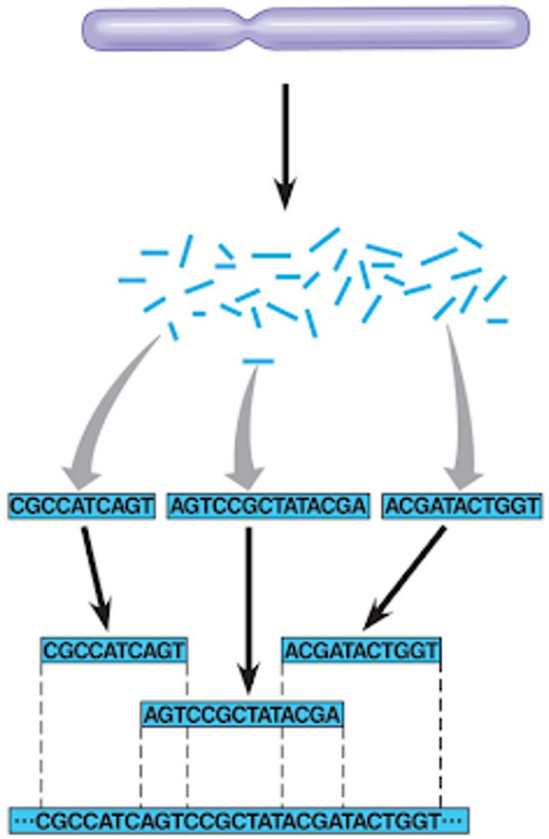
TE insertion identification from sequencing data

Cut the DNA into fragments

Clone the fragments

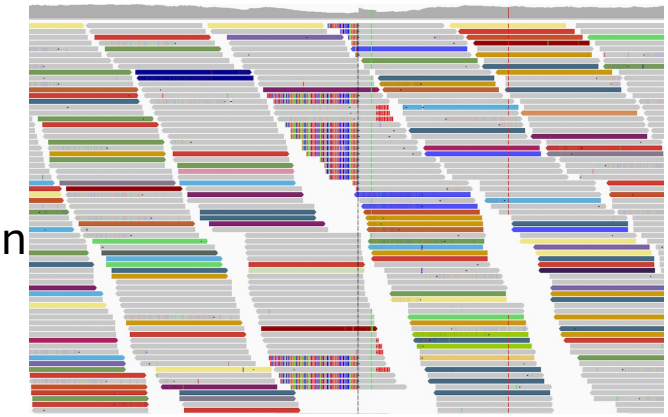
Sequence each fragment

Order/align the sequences

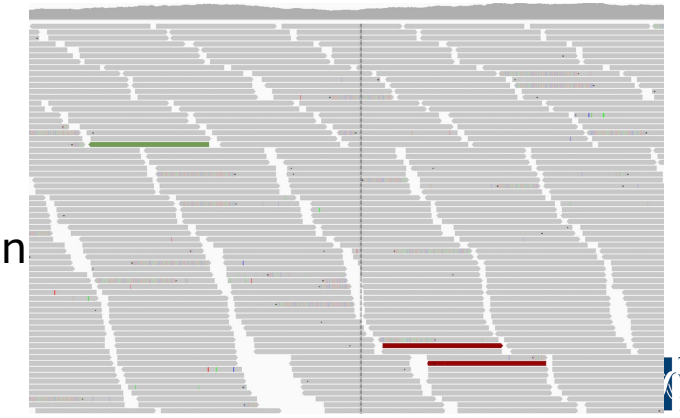


Different patterns

With insertion

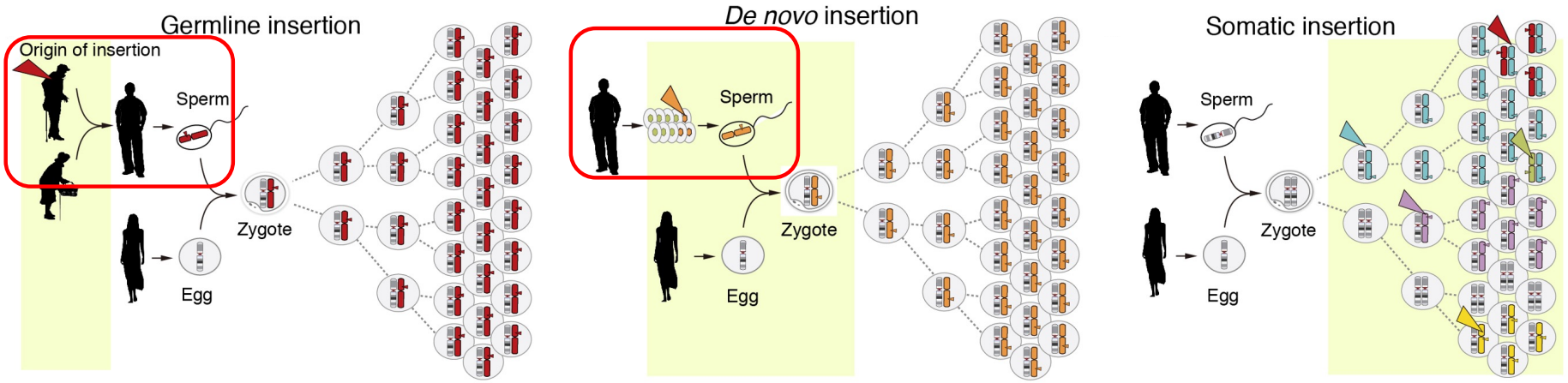


No insertion



THE WISTAR INSTITUTE

Challenges and motivation (different types of TE insertions)



- ~2,000 TE insertions each person
- Specificity is important in building a database

- *De novo* insertion is rare
- Sensitivity is more important in detection

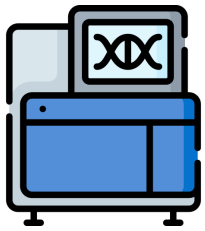
- Low allele-frequency if TE insertions happened late

Chu*, Zhao*, et al. *Current Protocols in Human Genetics* 2020

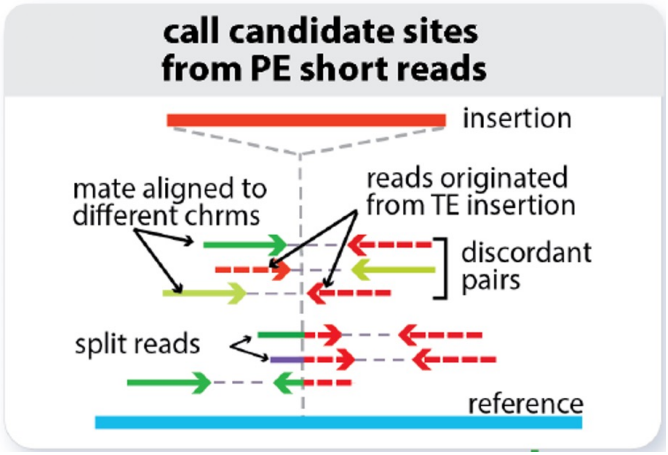
THE WISTAR INSTITUTE

xTea: TE insertion identification from short read sequencing data

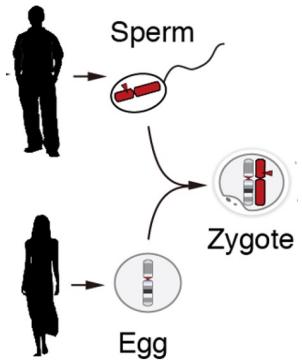
TE insertion **location** identification

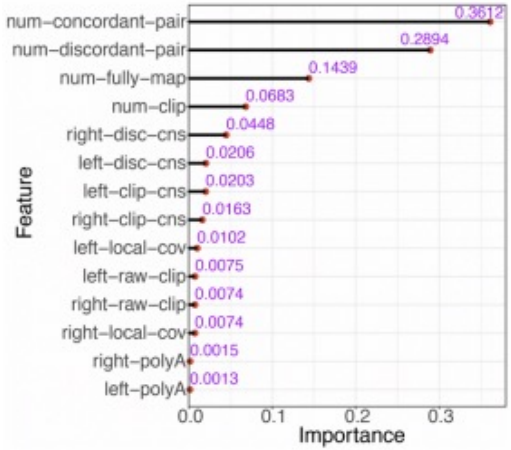


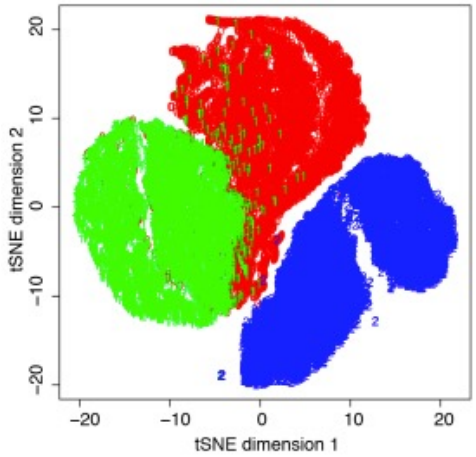
Personal sequencing data



Germline TE insertion **genotype** identification



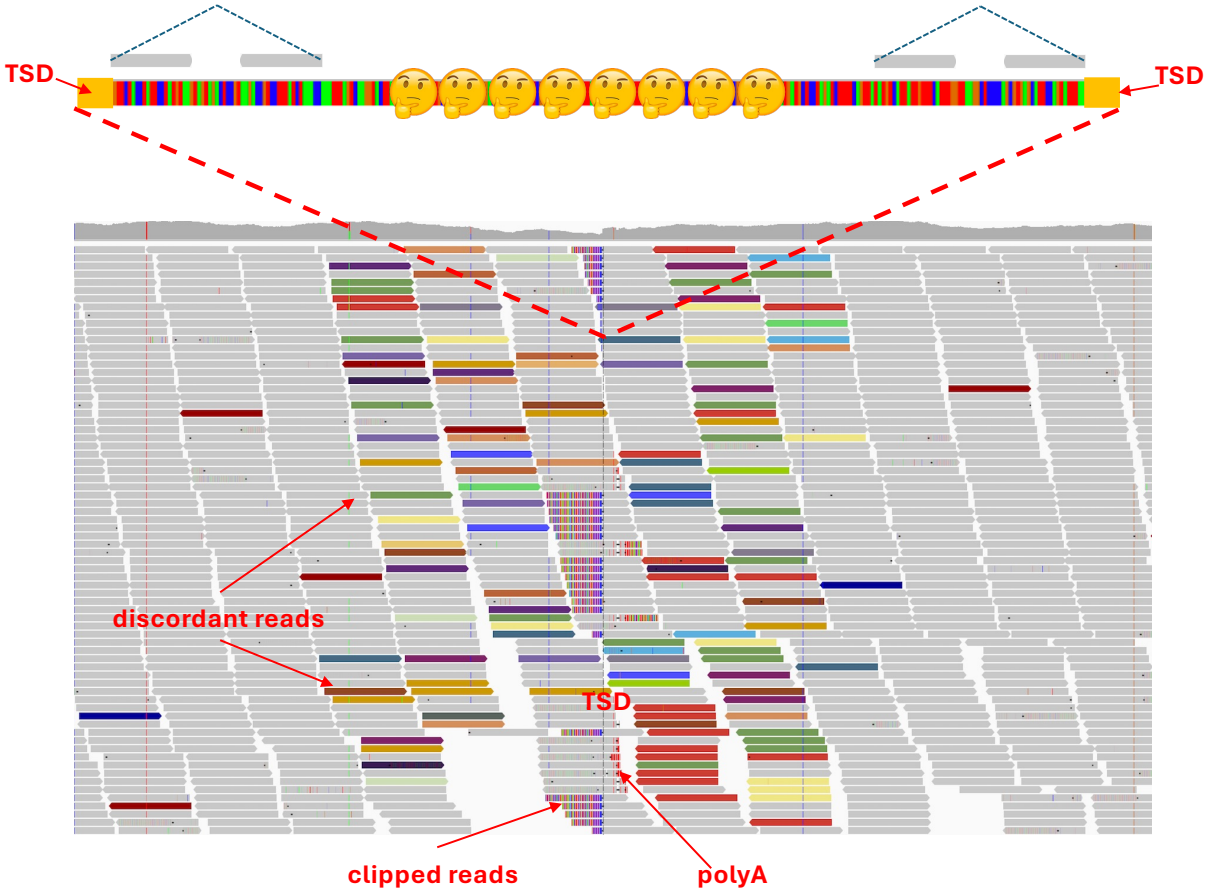




Chu et. al. *Nature Communications* 2021

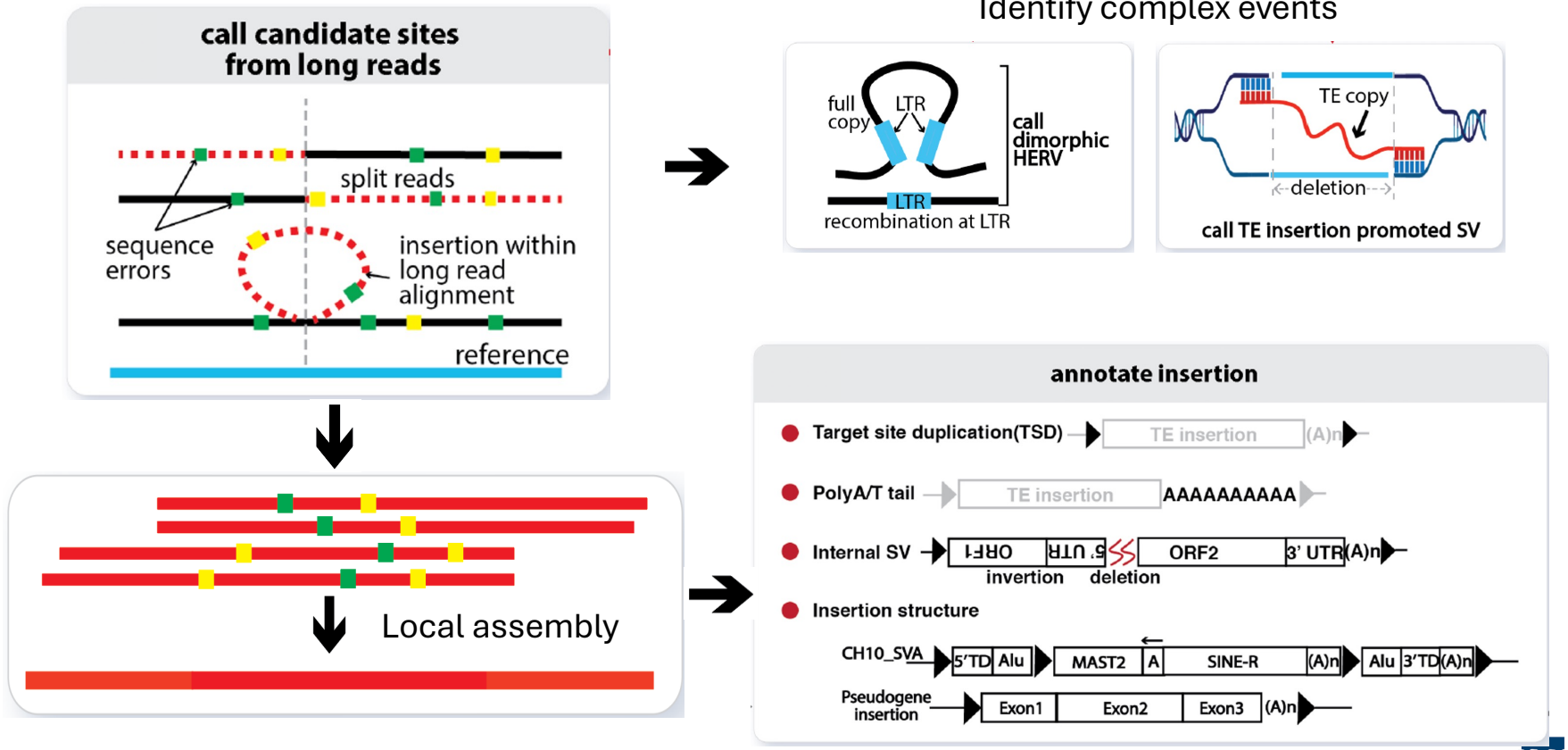
THE WISTAR INSTITUTE

We cannot construct the full insertion from short reads



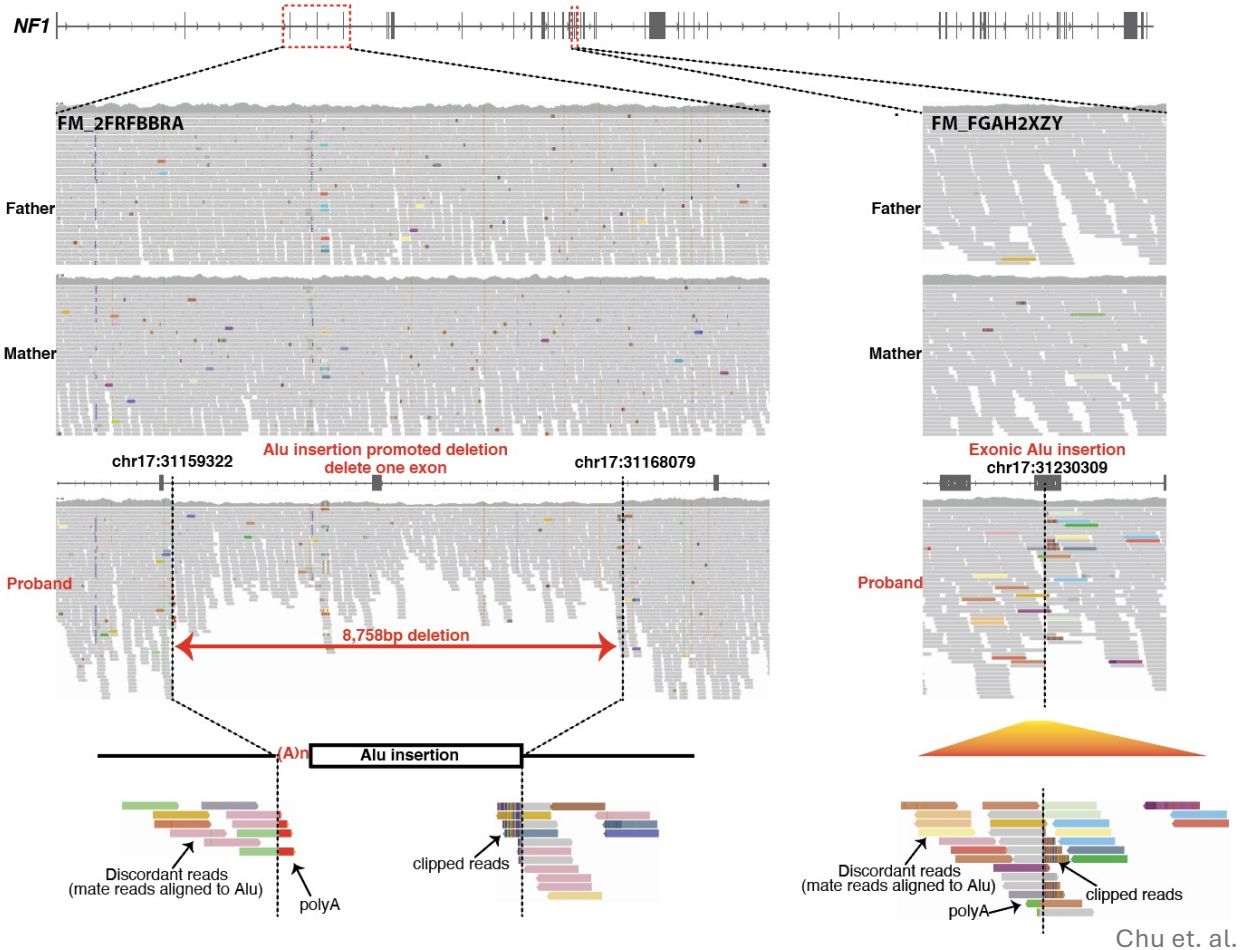
THE WISTAR INSTITUTE

xTea: TE insertion identification from long read sequencing data



 Chu et. al. *Nature Communications* 2021

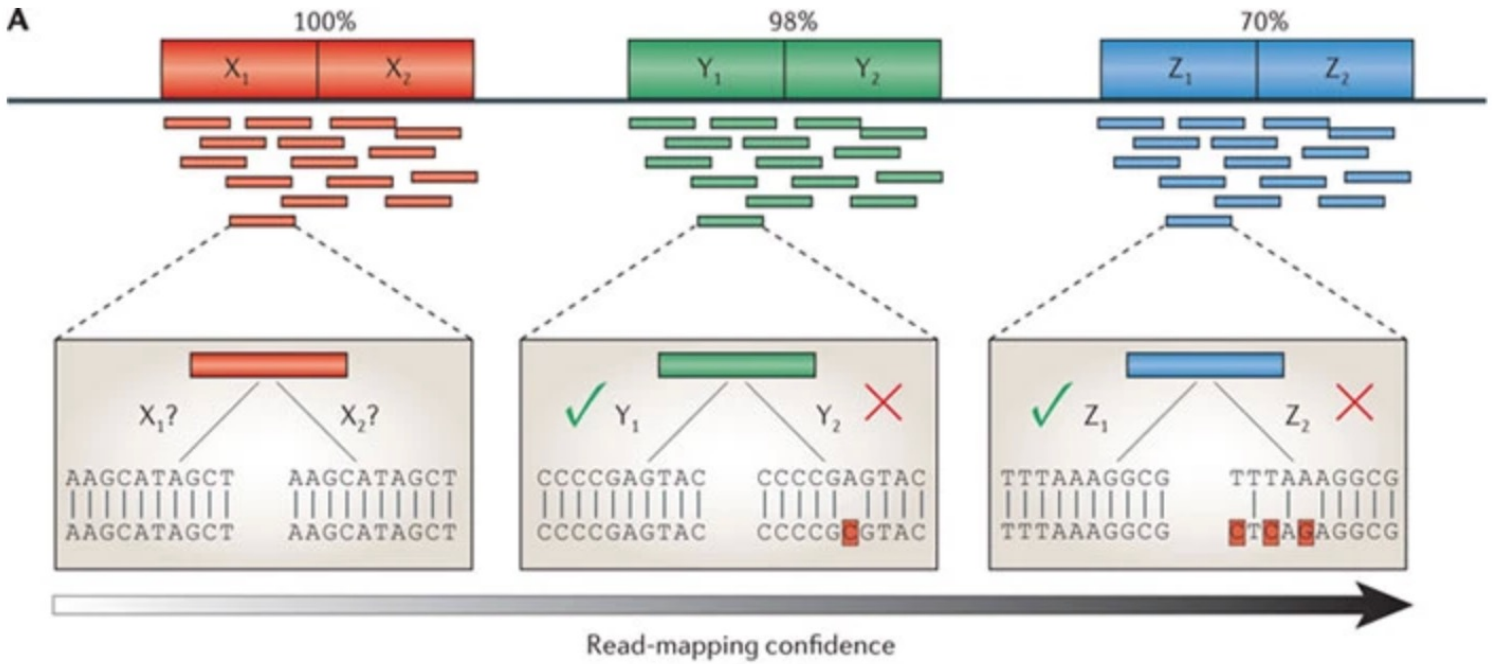
Example TE insertions cause brain tumors



Chu et. al. medrxiv, 2024

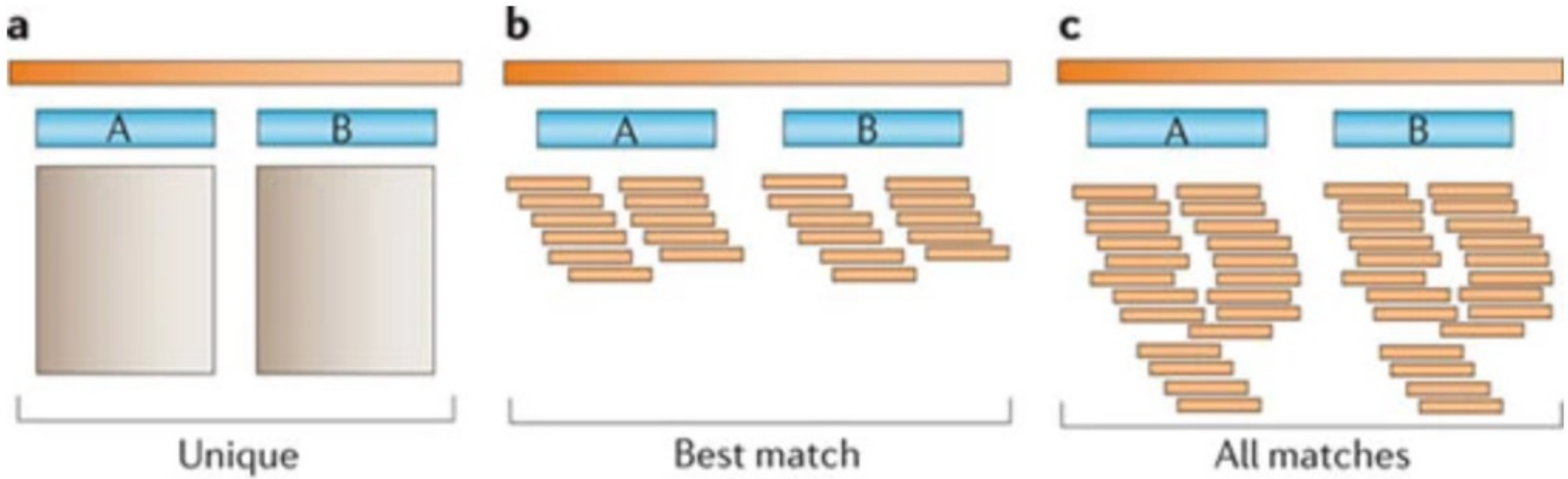
THE WISTAR INSTITUTE

TE RNA expression quantification is challenging



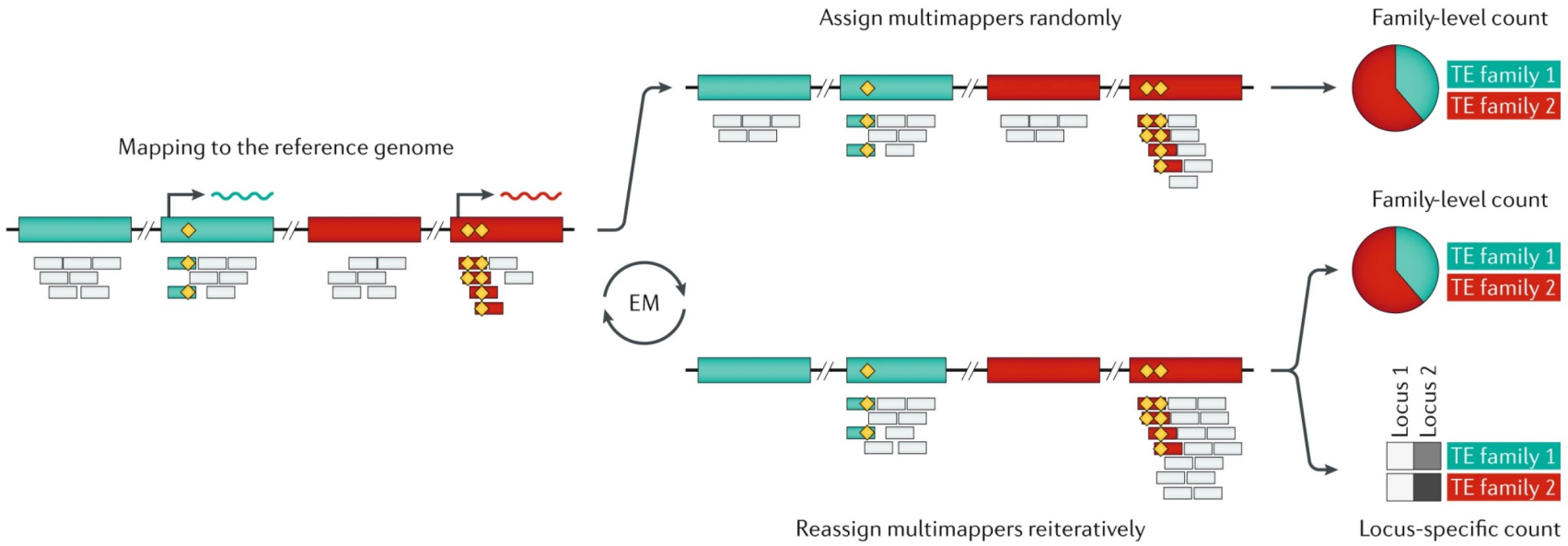
THE WISTAR INSTITUTE

Different strategies in dealing with multi-mapped reads



THE WISTAR INSTITUTE

TE RNA expression quantification approaches

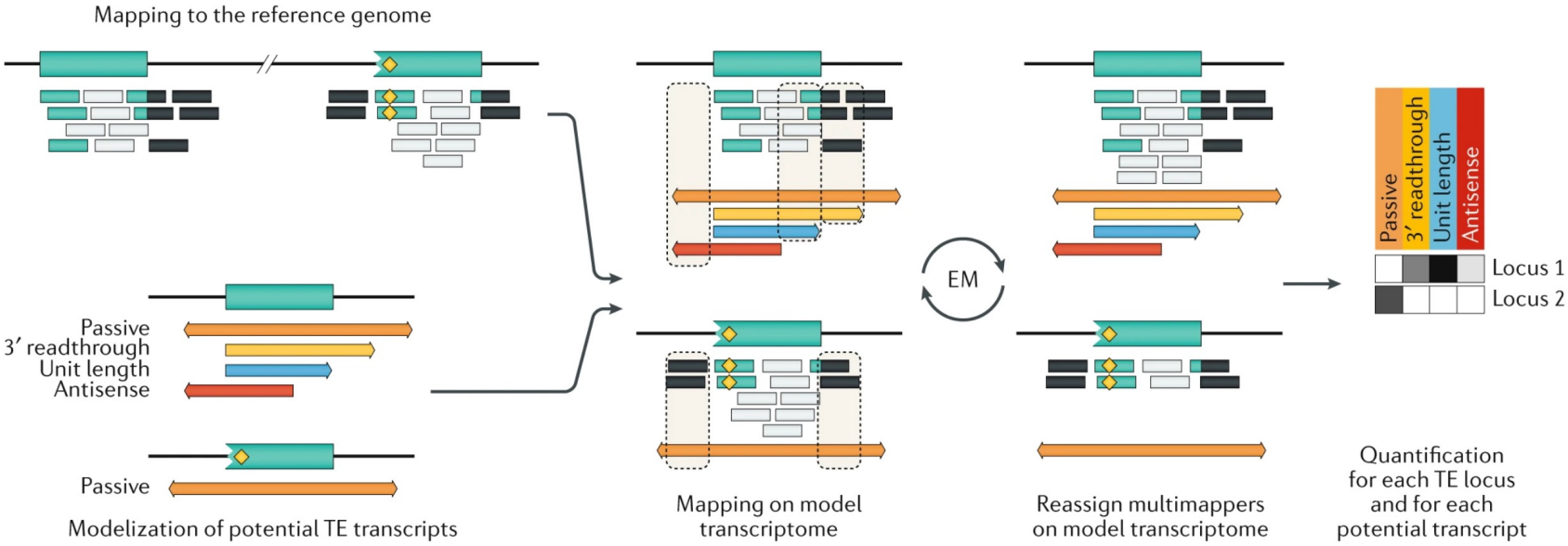


e.g. TEtranscript, SQuIRE

Lanciano and Gael, Nature Reviews Genetics, 2020

THE WISTAR INSTITUTE

TE RNA expression quantification approaches

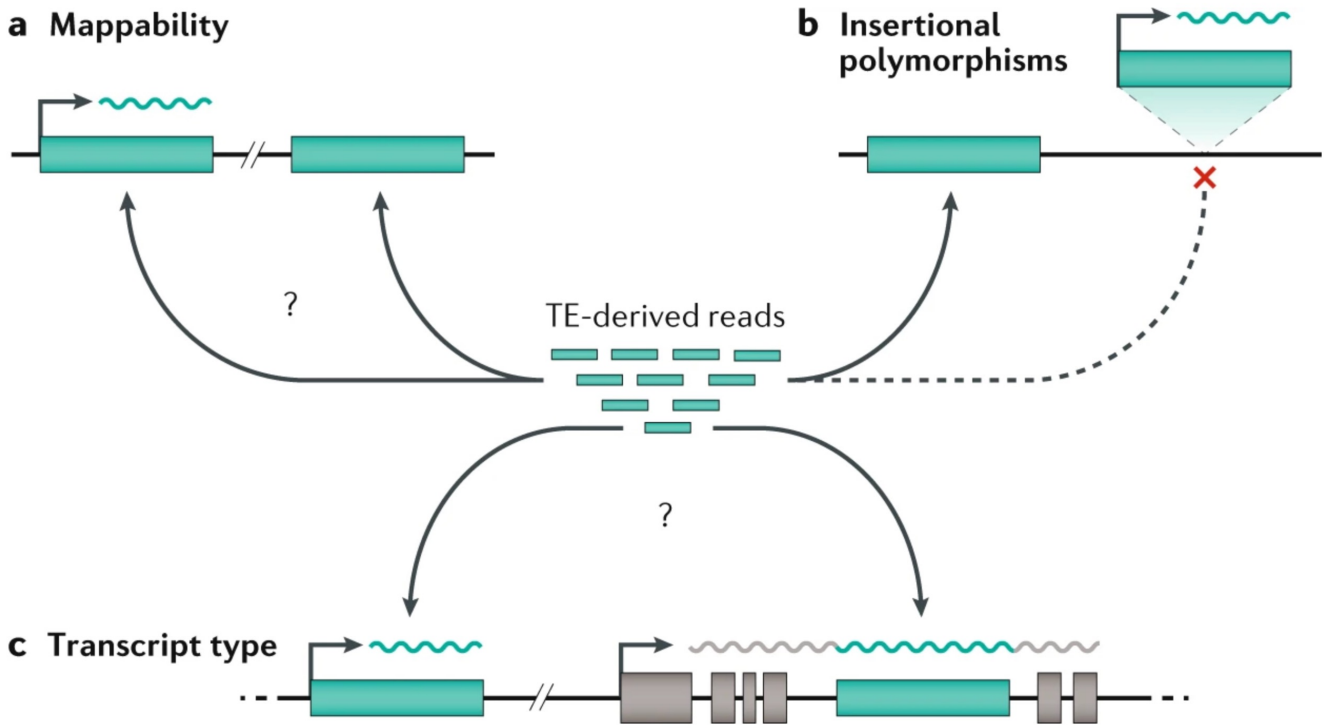


e.g. L1EM

Lanciano and Gael, Nature Reviews Genetics, 2020

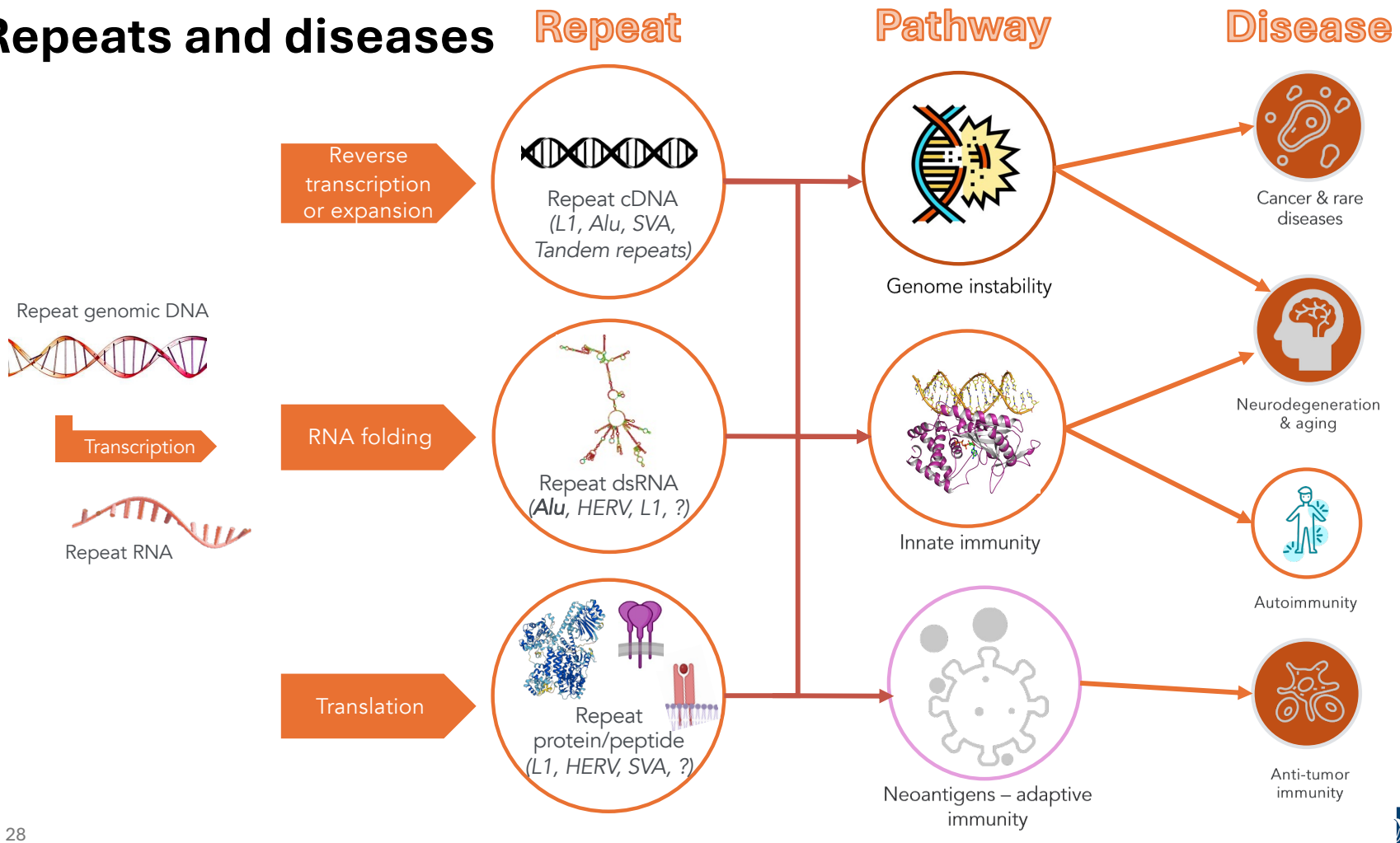
THE WISTAR INSTITUTE

TE RNA expression quantification remains challenging



Lanciano and Gael, Nature Reviews Genetics, 2020

Repeats and diseases



THE WISTAR INSTITUTE

Dark genome matters

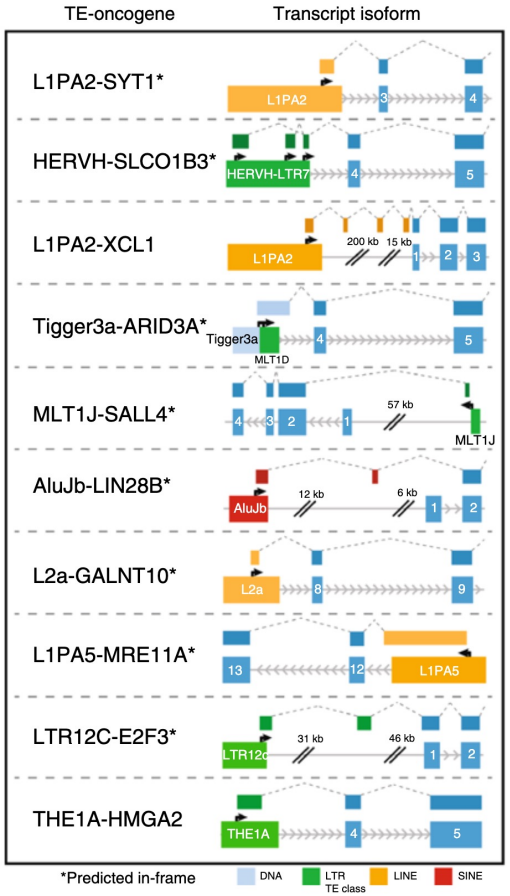
Transcribed smORFs

ORF class	RNA type	Median size (codons)	Translation[15]	Conservation	Coding features	Function
Intergenic ORFs	None	22	None	None[6,8]	Non-canonical AA	None
uORFs	5′ UTRs of mRNAs	22	Low	None[8,30]	• Nonrandom AA • No domains	• Non-coding • Translation regulation
lncORFs	lncRNAs	24	Low	None[8,10]	• Nonrandom AA • No domains	Non-coding or coding
Short CDSs	Short mRNAs	79	High	Class	• Positively charged AA • Transmembrane α-helices	• Coding • Regulators of canonical proteins
Short isoforms	Spliced mRNAs	79	High	Kingdom	• Canonical AA • Protein domain loss	• Coding • Small interfering peptides
Canonical ORFs	mRNAs	491	High	Kingdom	• Canonical AA • Multiple protein domains	• Coding[42] • Structural, enzymatic, regulatory

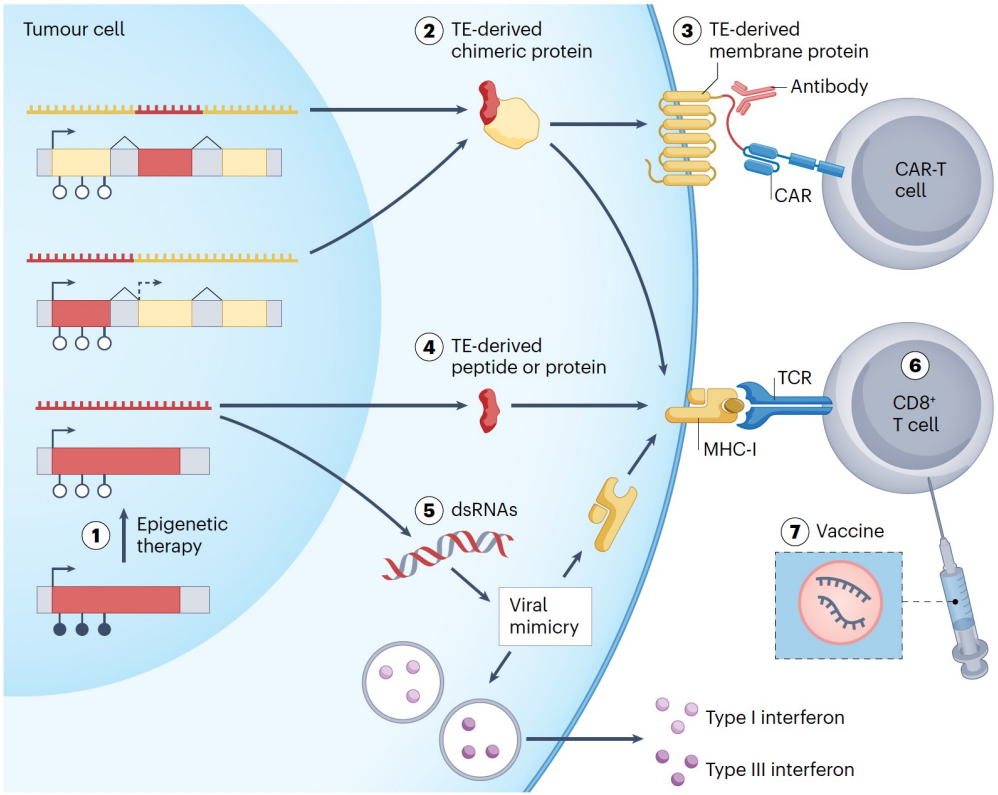
Untranslated region · ORFs · DNA · Other coding sequences · RNA splicing · Ribosome profiling signal

Couso, J.P. and Patraquim, P., 2017. Nature reviews Molecular cell biology

TE exonization as an enriched reservoir of novel targets



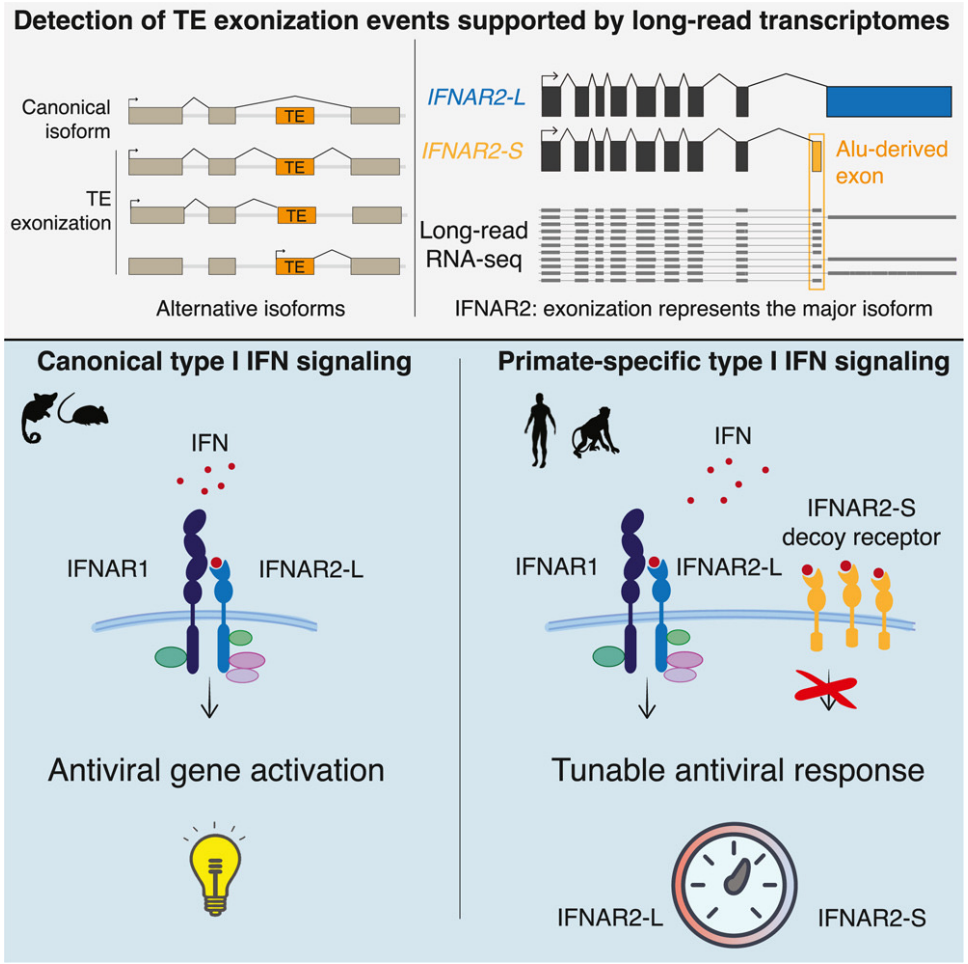
Jang, et. al., *Nature Genetics*, 2019



Liang, Yonghao, et al. *Nature Reviews Cancer*, 2024

THE WISTAR INSTITUTE

Functional TE exon-trapping



Pasquesi, Giulia Irene Maria, et al. *Cell,* 2024

31

TrapHunter: TE derived exon-trapping identification

Different types of exon-trapping events:

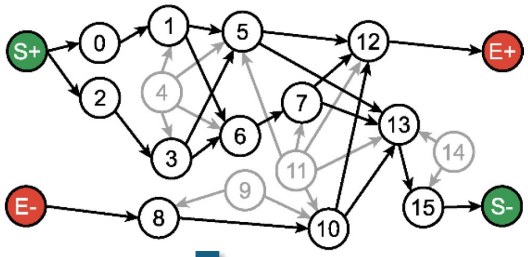
Initiated from repeats:

Spliced to repeats:

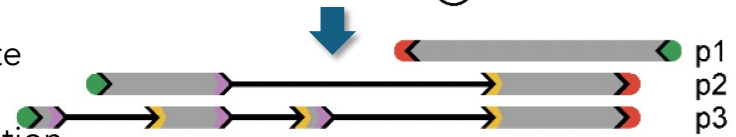
Terminated at repeats:

TrapHunter

Short-reads assembly

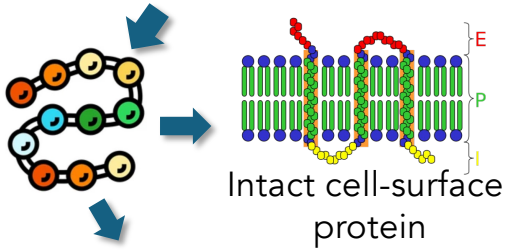


Candidate fusion identification



Second-stage contig assembly

Translate to aa sequence

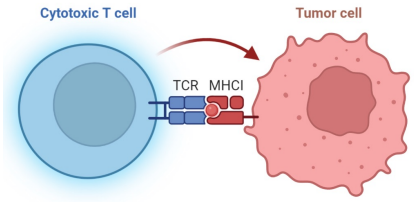


Chu et. al. in prep.

THE WISTAR INSTITUTE

Three major criteria in neoantigen discovery

Tumor specific

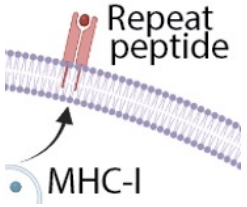


- Identified in tumor RNA but not normal
  - Differentially expressed

Prevalent in patients

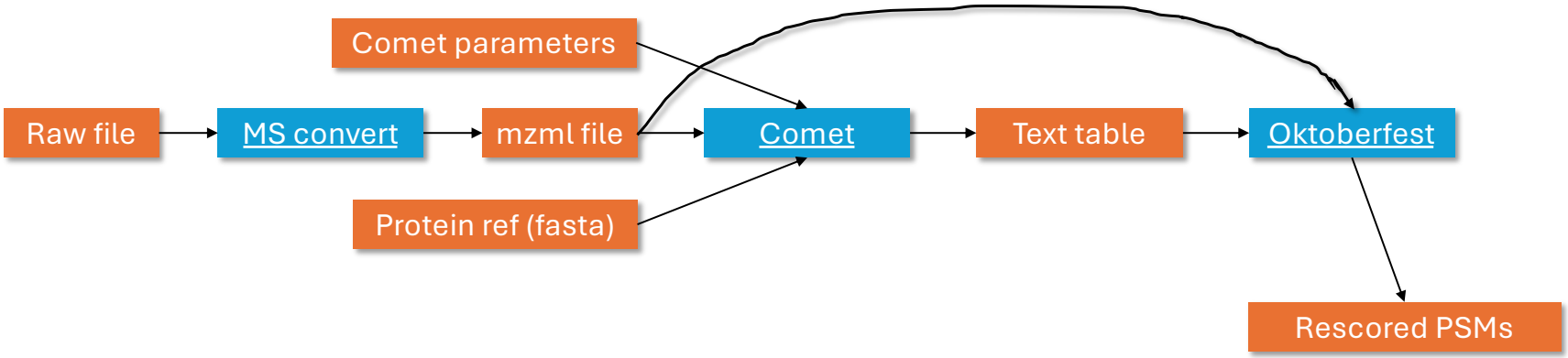
- Prevalent in tumors
  - Prevalent in TCGA tumors

On cell surface

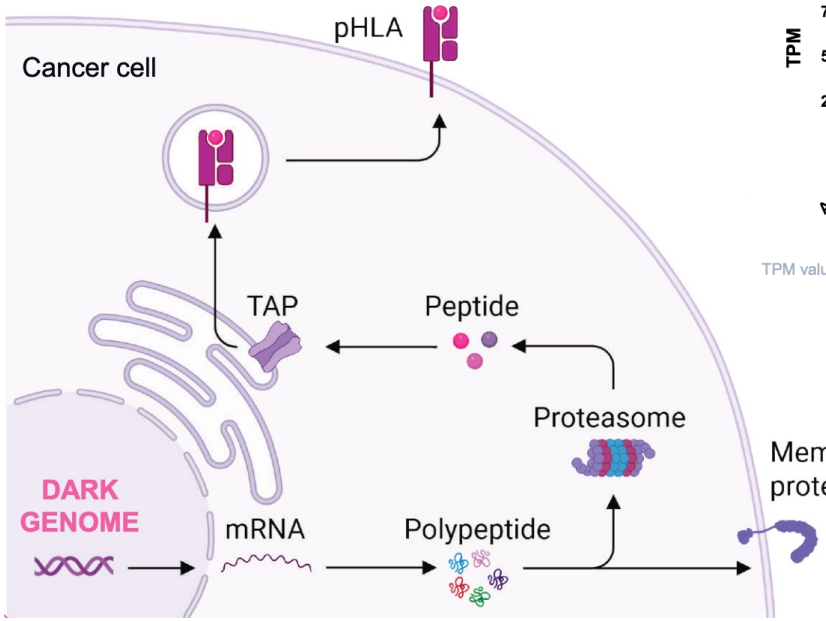


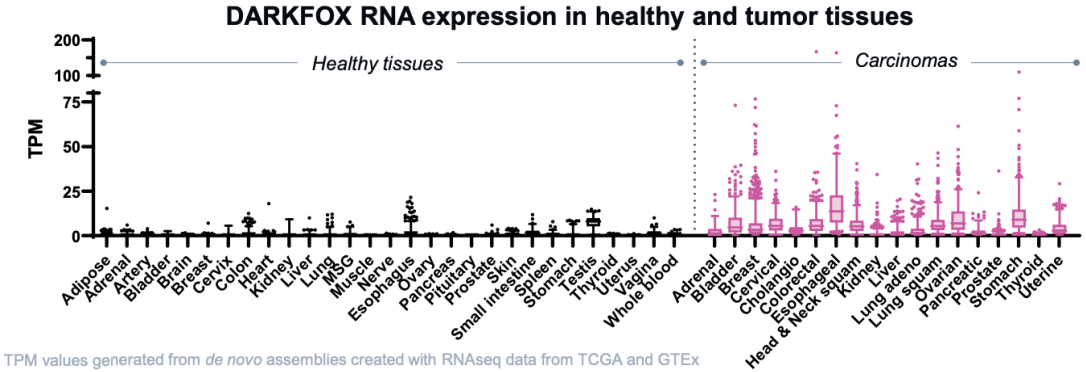
- Translated and on cell surface
  - Immunopeptidomics data

Neoantigen discovery pipeline with immunopeptidomics data

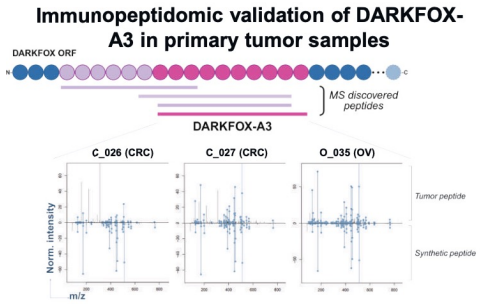


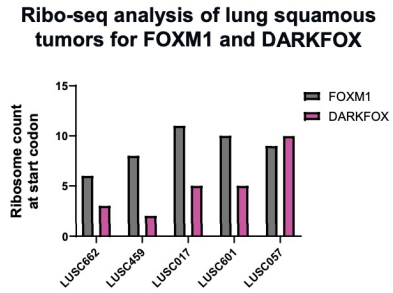
Implication example: Dark genome derived neoantigens for mRNA cancer vaccines





TPM values generated from *de novo* assemblies created with RNAseq data from TCGA and GTEx





Figures from Enara Biosciences

THE WISTAR INSTITUTE

Summary

- **Non-coding RNAs**
  - Different types of non coding RNAs
  - Summary of their functions

- **Transposable elements (TEs)**
  - TEs play important regulation roles
  - Sequence annotation with RepeatMasker
  - TE insertion identification with xTea
  - Current approaches in TE RNA expression quantification
  - TE derived neoantigen for mRNA cancer vaccine