Transposable elements quantification using RNA sequencing

Dingjie Wang, PhD

Department of Biomedical Informatics
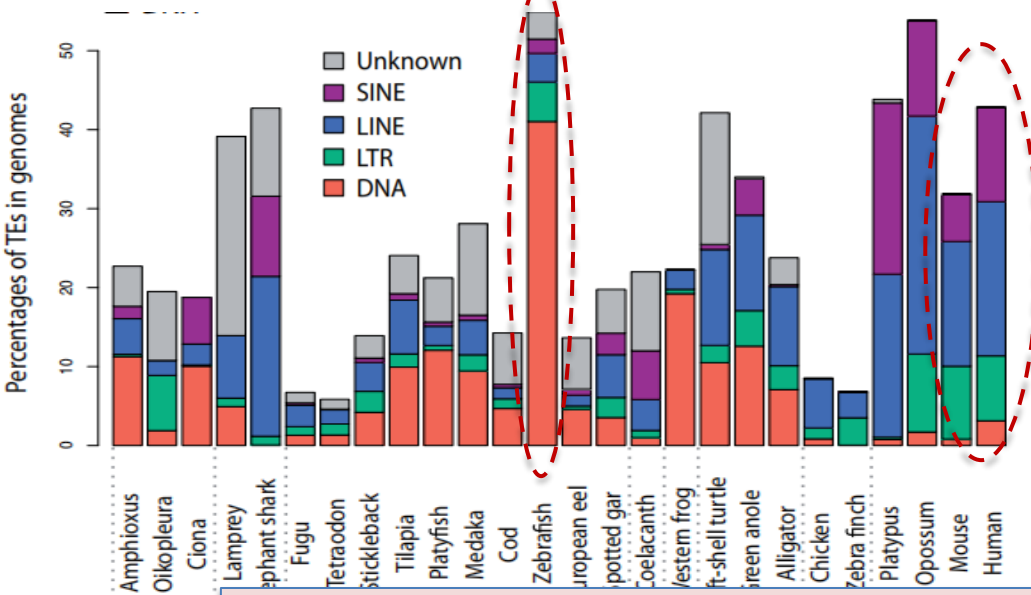**Ohio State University**

Feb. 19, 2021

THE OHIO STATE UNIVERSITY
WEXNER MEDICAL CENTER

Background

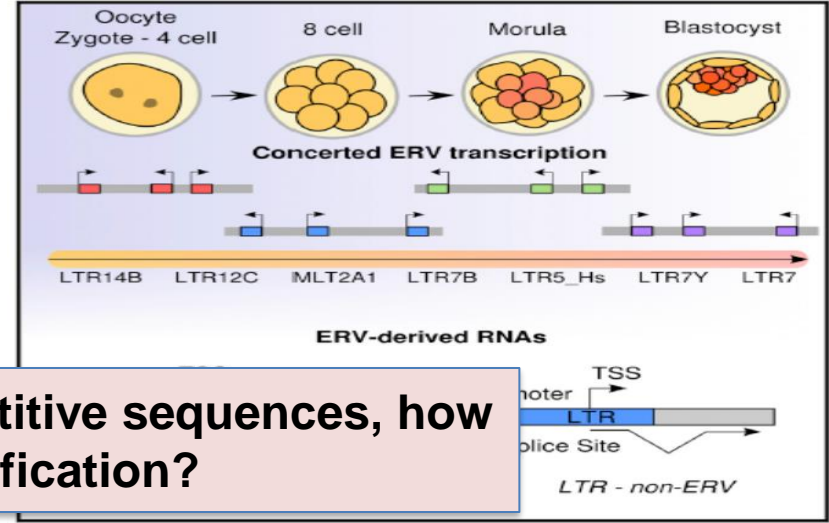**Percentages of TEs in the different vertebrate and nonvertebrate genomes**

Ref. Chalopin D

Question: For these TEs with high repetitive sequences, how to achieve accurate quantification?
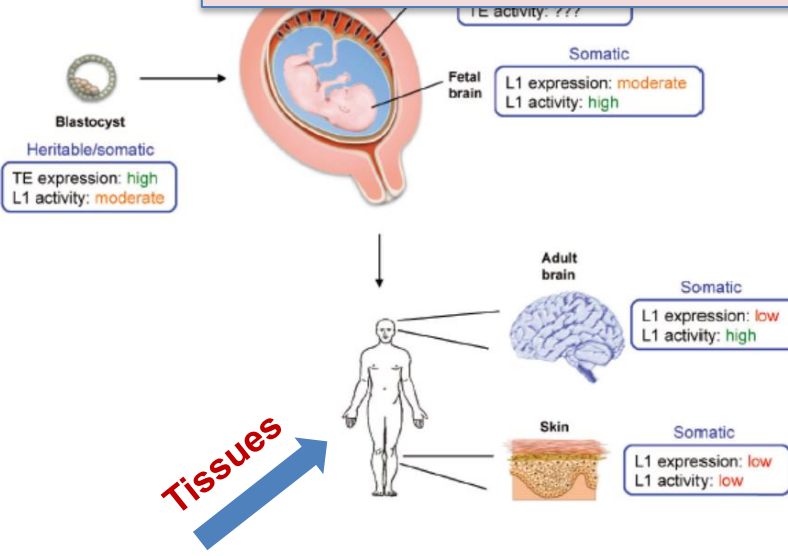
LINE-1 expression and activity in humans: from tissues to cells.

Tissues

Cell Lines

Garcia-Perez J L et al. Development, 2016, 143(22): 4101-4114.

Recent research advances

> *Review paper:*

Measuring and interpreting transposable element expression

Sophie Lanciano & Gael Cristofari ✉

> *Quantification at subfamily-level*
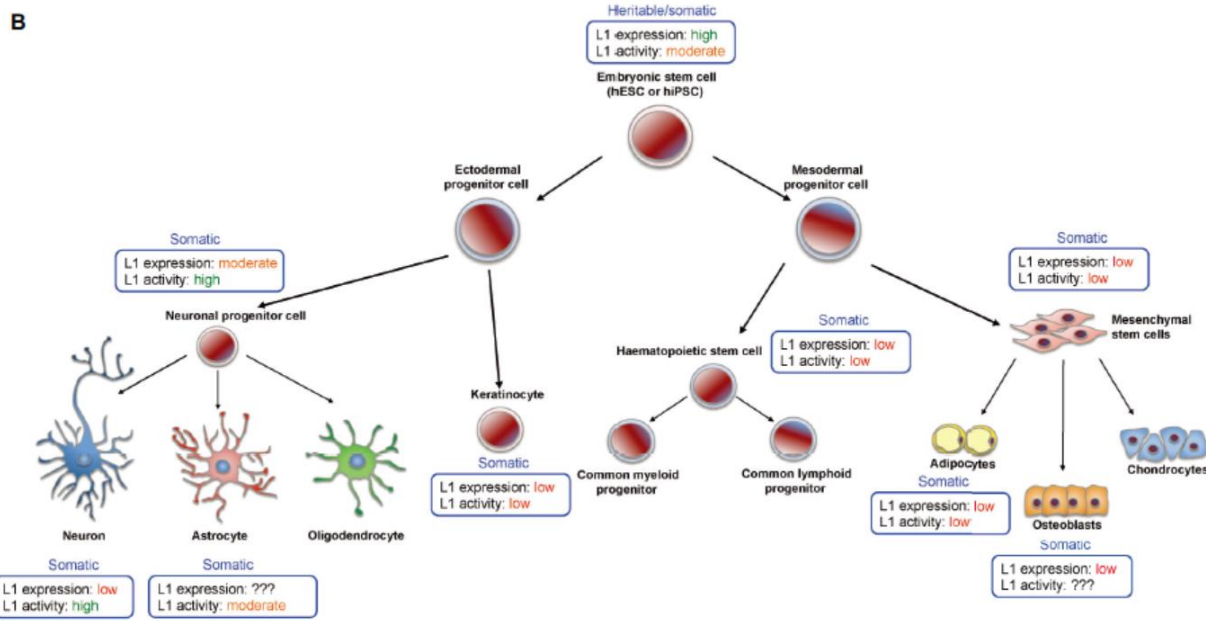
TEtranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets (FREE)

One of the most used tools, tested on a wide variety of organisms (citation >150)

Ying Jin, Oliver H. Tam, Eric Paniagua, Molly Hammell ✉   Author Notes

Transposable element expression in tumors is associated with immune infiltration and increased antigenicity

Yu Kong, Christopher M. Rose, Ashley A. Cass, Alexander G. Williams, Martine Darwish, Steve Lianoglou, Peter M. Haverty, Ann-Jay Tong, Craig Blanchette, Matthew L. Albert, Ira Mellman, Richard Bourgon, John Greally, Suchit Jhunjhunwala & Haiyin Chen-Harris ✉

Transposable elements drive widespread expression of oncogenes in human cancers

Hyo Sik Jang, Nakul M. Shah, Alan Y. Du, Zea Z. Dailey, Erica C. Pehrsson, Paula M. Godoy, David Zhang, Daofeng Li, Xiaoyun Xing, Sungsu Kim, David O'Donnell, Jeffrey I. Gordon & Ting Wang ✉

TETOOLS facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes 🔓

Emmanuelle Lerat ✉, Marie Fablet ✉, Laurent Modolo ✉, Hélène Lopez-Maestre, Cristina Vieira ✉

ERVmap analysis reveals genome-wide transcription of human endogenous retroviruses

Maria Tokuyama, Yong Kong, Eric Song, Teshika Jayewickreme, Insoo Kang, and (ID) Akiko Iwasaki

> *Quantification at locus-level*

SQuIRE reveals locus-specific regulation of interspersed repeat expression 🔓

Wan R Yang, Daniel Ardeljan, Clarissa N Pacyna, Lindsay M Payer ✉, Kathleen H Burns ✉   Author Notes

● **TElocal:** A package for quantifying transposable elements at a locus level of RNA-seq datasets.
https://github.com/mhammell-laboratory/TElocal

> *Quantification in single-cell RNA-seq dataset*

GENOME RESEARCH

Transcript assembly improves expression quantification of transposable elements in single-cell RNA-seq data

Wanqing Shao[1,2] and Ting Wang[1,2,3]

Recent research advances

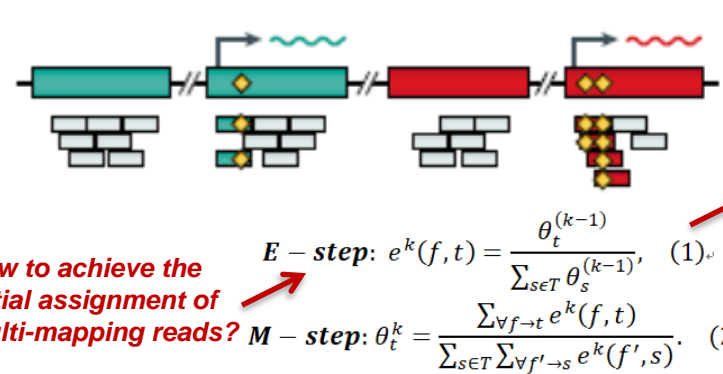Table 1 | Computational tools and approaches to analyse TE unit expression from RNA-seq data

Tools or approaches	Mapping or pseudo-mapping to	Fate of multimappers	Type of quantification	Distinguishes unit-length transcripts from other TE-derived transcripts	Includes polymorphic TE expression	Notes	Ref.
TEtools	TE pseudogenome	Randomly assigned	F	–	–	Applicable to unassembled genomes	136
SalmonTE	Consensus transcriptome	EM algorithm	F	–	–	Fast pseudomapping	152
REdiscoverTE	Model transcriptome	EM algorithm	F	+	–	Uses SalmonTE algorithm	153
TEtranscripts	Reference genome	EM algorithm	F	–	–	One of the most used tools, tested on a wide variety of organisms	148
RepEnrich	Reference genome	Remapped on TE pseudogenome	F	–	–	–	144
TeXP	Reference genome	Randomly assigned	F	+/–	–	Subtracts signal from pervasive transcription but not from other forms of chimeric transcripts	100
ERVmap	Reference genome	Discarded	L	–	–	Uses a curated full-length human ERV database	157
Random assignment of multimappers	Reference genome	Randomly assigned	L	–	–	Locus-specific transcription not reliable on youngest TEs	134
TEcandidates	Reference genome	Remapped on partially masked reference genome	L	–	–	–	146
SQuIRE	Reference genome	EM algorithm	L	–	+/–	Polymorphic insertion can be added as extra chromosome if internal sequence known	145
Manual curation	Reference genome	Discarded	L	+	–	Difficult to generalize	99
Telescope	Reference genome	EM algorithm	L	+	–	–	149
L1EM	Reference genome and model transcriptome	EM algorithm	L	+	–	Proof of principle on human L1 elements, could be generalized	65
Multi-omics 1	Reference genome	NA	L	+	+	Combines targeted DNA sequencing, RNA-seq and ChIP–seq	154
Multi-omics 2	Reference genome	NA	L	+	+	Combines whole-genome sequencing and RNA-seq	33

➢ **How to treat multi-mapping reads:**
(1) **Discarded**
(2) **Randomly assigned**
(3) **EM algorithm**

Lanciano S, Cristofari G. Nature Reviews Genetics, 2020, 21(12): 721-736.

a Mapping on a reference genome

Assign multimappers randomly

Family-level count

Mapping to the reference genome

TE family 1
TE family 2

Family-level count

TE family 1
TE family 2

EM

How to achieve the initial assignment of multi-mapping reads?

$$E-step:\ e^k(f,t)=\frac{\theta_t^{(k-1)}}{\sum_{s\in T}\theta_s^{(k-1)}},\quad (1)$$

$$M-step:\ \theta_t^k=\frac{\sum_{\forall f\to t}e^k(f,t)}{\sum_{s\in T}\sum_{\forall f'\to s}e^k(f',s)}.\quad (2)$$

Lanciano S, Cristofari G. Nature Reviews Genetics, 2020, 21(12): 721-736.

Locus 1
Locus 2

TE family 1
TE family 2
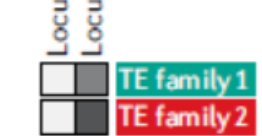
Reassign multimappers reiteratively

Locus-specific count

TEtranscripts: equal assignment

Inputs
- Alignment files (SAM/BAM)
- Gene annotations (GTF)
- TE annotations (GTF)

Indexing
- Gene index
- TE index

Quantification
- TE or Gene Reads? — TE / Gene
- TE mode — multi / uniq
- Equal Weighting
- TE expression profile (multi-reads) — Expectation Maximization
- Resolve Ambiguity
- Resolve Ambiguity

Output
- TE count table
- Gene count table
- Differential Analysis

Jin Y, Tam O H, Paniagua E, et al. Bioinformatics, 2015, 31(22): 3593-3599.

SQuIRE: the initial assignment is based on unique reads proportion

A RNA-seq
- TE A
- TE A
- TE B
- TE C

B Read labeling after alignment to TE loci the genome
- * unique sequence
- unique reads
- multi-mapping reads

$\frac{4\ unique\ reads}{2\ unique\ positions}=2$

$\frac{1\ unique\ reads}{1\ unique\ positions}=1$

Normalized by the number of unique read positions

4 unique reads TE A
1 unique read TE B
0 unique reads TE C

C Multi-mapping read fraction assignment

$\frac{2}{3}\times\frac{2}{2+1}=\frac{4}{9}$ → 4/9 TE A

$\frac{2}{3}\times\frac{1}{2+1}=\frac{2}{9}$ → 2/9 TE B

1/3 TE C

D Total read count estimation

5.78 total read counts TE A
$\frac{4}{9}\times4+4=5.78$

1.89 total read counts TE B
$\frac{2}{9}\times4+1=1.89$

1.33 total read counts TE C
$\frac{1}{3}\times4=1.33$

E Expectation-Maximization

E-step

5.78 TE A
1.89 TE B
1.33 TE C

6.0 total read counts TE A
2.0 total read counts TE B
1.0 total read counts TE C

M-step

Yang W R, et al. Nucleic acids research, 2019, 47(5): e27-e27.

The formula shown in the figure:

$$ARR = \frac{\text{Estemated Abundance}}{\text{Real Abundance}}$$
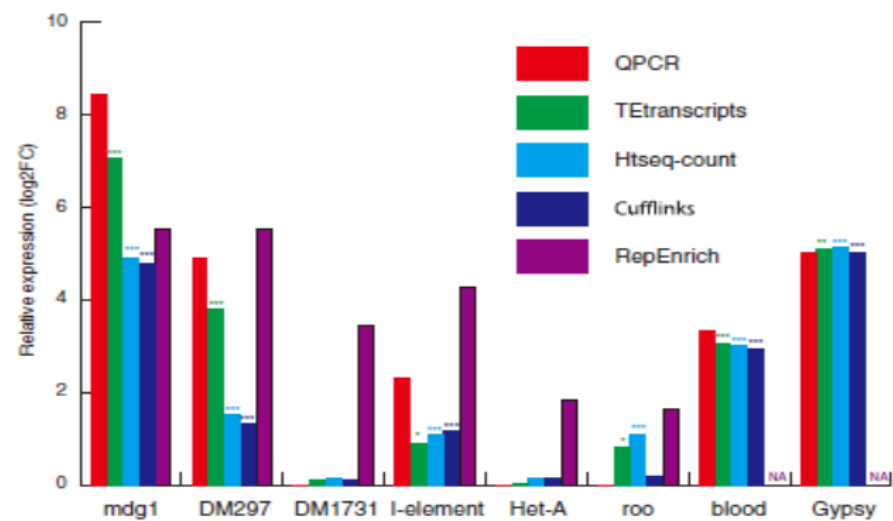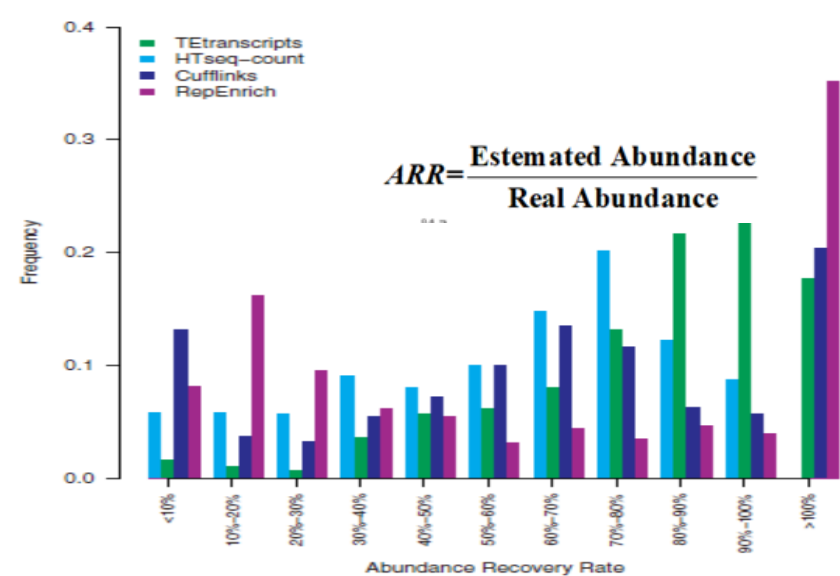
Jin Y, Tam O H, Paniagua E, et al. Bioinformatics, 2015, 31(22): 3593-3599.
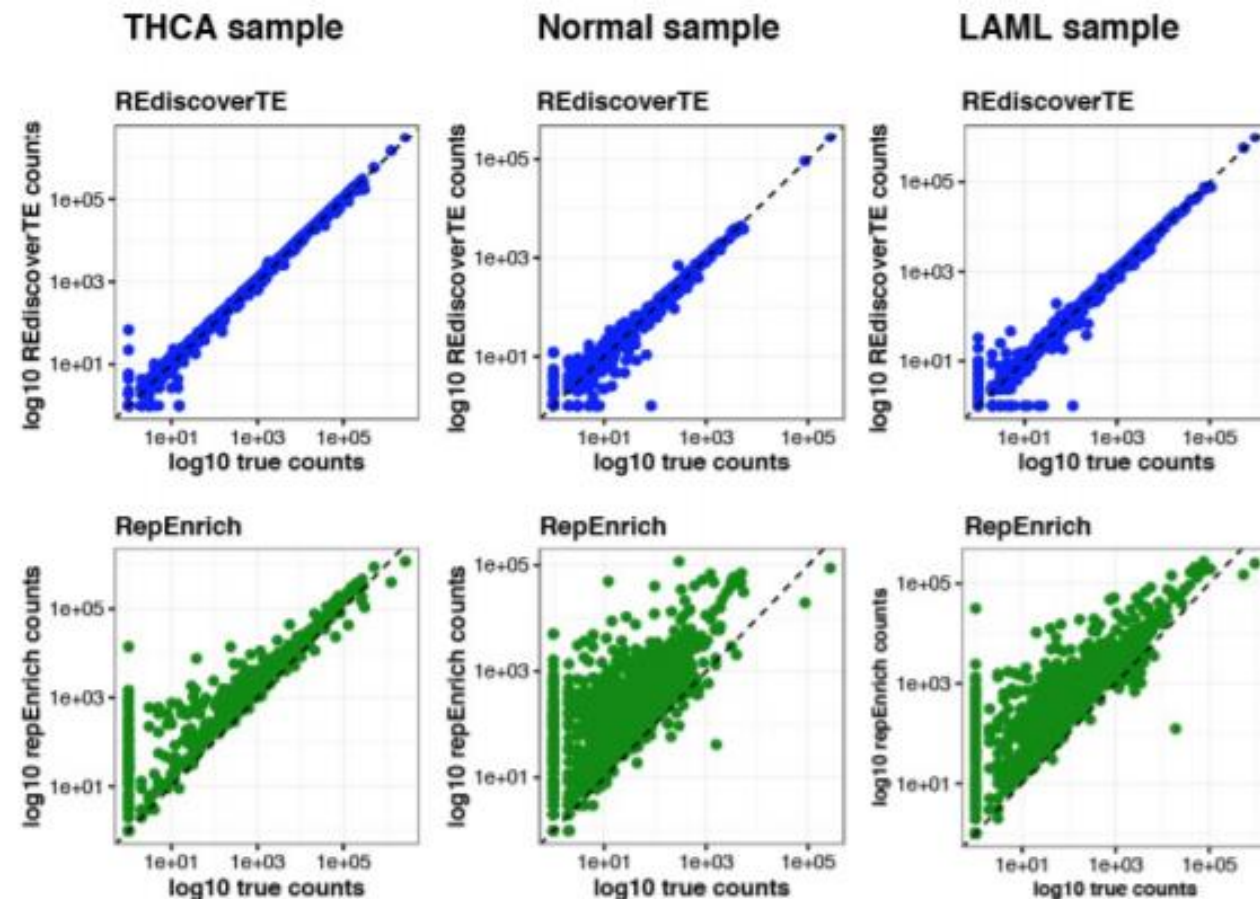
● **Limitation of current methods:**

(1) All methods are based on short read RNA-seq dataset;

(2) Most of methods only focus on the quantification at subfamily-level;

(3) There is no good methods to assign multi-mapping reads to each TE loci.

(4) Lack of method based LR or LR+SR for TE quantification.

Kong Y, et al. Nature communications, 2019, 10(1): 1-14.