An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data

Hyun-Hwan Jeong[1,2], Hari Krishna Yalamanchili[1,2], Caiwei Guo[2,3],
Joshua M. Shulman[1,2,3,4], Zhandong Liu[2,5,†]

[1]*Department of Molecular and Human Genetics, Baylor College of Medicine,*
[2]*Jan and Dan Duncan Neurological Research Institute, Texas Childrens Hospital,*
[3]*Department of Neuroscience, Baylor College of Medicine,*
[4]*Department of Neurology, Baylor College of Medicine,*
[5]*Department of Pediatrics, Baylor College of Medicine,*
Houston, Texas 77030, USA
[†]*E-mail: zhandonl@bcm.edu*

Transposable elements (TEs) are DNA sequences which are capable of moving from one location to another and represent a large proportion (45%) of the human genome. TEs have functional roles in a variety of biological phenomena such as cancer, neurodegenerative disease, and aging. Rapid development in RNA-sequencing technology has enabled us, for the first time, to study the activity of TE at the systems level.

However, efficient TE analysis tools are not yet developed. In this work, we developed `SalmonTE`, a fast and reliable pipeline for the quantification of TEs from RNA-seq data. We benchmarked our tool against `TEtranscripts`, a widely used TE quantification method, and three other quantification methods using several RNA-seq datasets from Drosophila melanogaster and human cell-line. We achieved 20 times faster execution speed without compromising the accuracy. This pipeline will enable the biomedical research community to quantify and analyze TEs from large amounts of data and lead to novel TE centric discoveries.

Keywords: Transposable Element; Quasi-Mapping; RNA-seq; Next Generation Sequencing; Large Scale Genome Analysis

1. Introduction

Transposable elements (TEs) are DNA elements which can be mobilized or inserted into the genome and represent a significant proportion of most eukaryotic genomes.[1] Most of the TEs in the genome are not functional and had been considered as 'junk DNA,' except for a few that retain intact functions such as transcription and mobilization.[2] Furthermore, the mobilization of TEs can disrupt normal gene structure in the genome, sometimes leading to disease such as cancer[3,4] neurodegenerative diseases,[1] and aging.[5]

Recent development of high-throughput Next Generation Sequencing (NGS) technologies, like RNA-seq, enables genome-wide study for TEs.[6–9] Toward this end, several algorithms and pipelines were proposed to analyze reads files from TE studies.[10–16] However, most of the

tools share some common limitations: 1) discordant read mapping due to increased chance of multiple mapping in repetitive elements from TEs in the same clade, 2) limited scalability for large-scale analysis, and 3) small coverage for the entire TEs defined in the human genome, i.e., a tool used in [16] only considered LINE 1 (Long Interspersed Nuclear Element 1) elements.[17]

Among the existing tools, `TEtranscripts` has performed well on various datasets.[14] Nonetheless, The scalability of `TEtranscripts` is a critical limiting factor for large systems biology studies because it cannot handle `FASTQ` files directly and needs `SAM` (Sequence Alignment Map)/`BAM` (Binary Sequence Alignment Map) files generated from raw `FASTQ` files. Since there are many tuning parameters on handling repetitive sequence among different RNA-seq mapping algorithms, this step will be highly variable depending on the mapping parameters and sometimes even generate artifactual results if a unique mapping parameter is superimposed by a previous analyst who handled the mapping.

Although `TEtranscripts` is the fastest tool for TE quantification,[14] the interval tree algorithm,[18] which is used to find the interval of genes or TEs on the reference genome, performed poorly in terms of running time in practice, making `TEtranscripts` suboptimal for large-scale TE analysis.

In recent studies, many large-scale analysis of public meta RNA-seq datasets offered new insight and findings that cannot be discovered in each dataset alone.[19] However, a meta-study on TE without using a large number of high-performance computing cluster is not yet feasible given the time complexity of current algorithms. Toward this end, we developed a new pipeline called `SalmonTE`. It deploys a low time-complexity quantification method, `Salmon`,[20] and contains various statistical models for TEs quantification. Moreover, `SalmonTE` provides a rich set of built-in functions for data pre-processing from raw `FASTQ` files. In the results section, we demonstrate the running speed of `SalmonTE` outperforms all other methods including `TEtranscripts` and delivers a reliable quantification result as well.

2. Methods

The proposed pipeline consists of three parts: library preparation, quantification, and statistical analysis (Figure 1). To increase the usability and to enable parallel processing for multiple RNA-seq reads files, we adopted the `Snakemake` workflow system and wrote a script based on the execution rule of `Snakemake` for the TE quantification.[21] In contrast to TEtranscripts, SalmonTE starts with raw RNA-seq files, and does not need any additional pre-processing for a given sequence file. Moreover, TEtranscripts requires a modified GTF files based on RepeatMasker database.[22] SalmonTE only needs the FASTA file of cDNA (complementary DNA) sequences of each TE. The entire source code and executable scripts are available at `https://github.com/hyunhwaj/SalmonTE`.

2.1. *Transposable Element Library Preparation*

To build the index library for the quasi-mapping, `SalmonTE` takes the FASTA file of cDNA sequences from TE databases such as Repbase (version 22.06).[23] In the current version, the index files for *Homo sapiens* and *Drosophila melanogaster* are available. We reasoned that it is hard to estimate TEs which replicate without an RNA intermediate from RNA-seq sample.

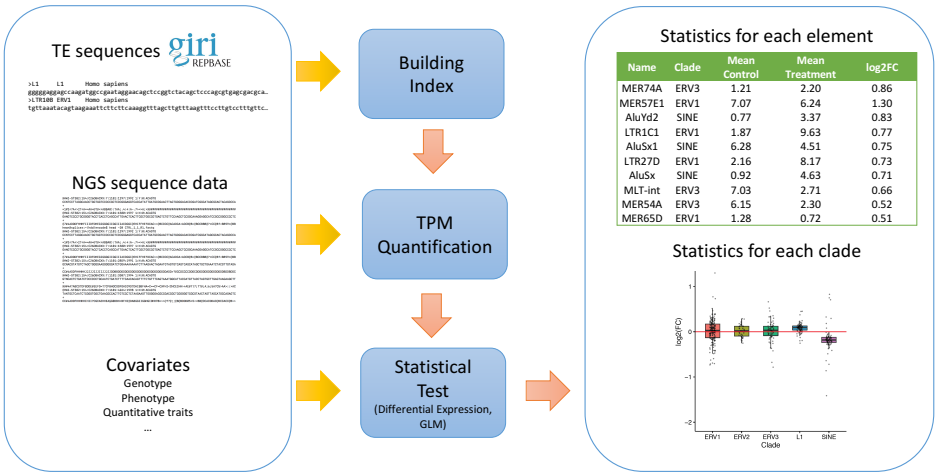

Fig. 1. An illustration of the `SalmonTE` pipeline. Left Panel: Input from Repbase to build the mapping index, raw `FASTQ` file, and covariates for statistical testing. Middle Panel: The workflow of `SalmonTE` consists of three parts: building the index based on Repbase or user-input cDNA sequences of TEs, quantification based on `FASTQ` file, and statical test through the generalized linear model or differential expression analysis. Right Panel: Example output including the statistical report and box plot on estimated log_2 fold-change.

Therefore, we excluded the following elements: simple repeats and multi-copy genes, and DNA transposable. After collecting the cDNA sequences, we manually curated clades of each TE based on the repeat class annotation from Repbase.
As a result, the generated TE library index database contains 687 TEs for *Homo sapiens* and 163 TEs for *Drosophila melanogaster*.

2.2. *Salmon quantification algorithm*

We adopted the `Salmon` [20] algorithm to estimate the relative TE abundance from a given RNA-seq sample. `Salmon` enables a fast and accurate quantification of TE expression from RNA-seq reads with a light-weight mapping, online initial expression estimation phase, and offline inference for the estimation refinement.[20,24–26] Salmon quantifies the relative abundance of each TE given a set of TE sequences T and a set of sequenced fragments (reads) F. Suppose that we have M TEs and the set of underlying true TE counts are given as $T = \{(t_1, \ldots, t_M), (c_1, \ldots, c_M)\}$, where t_i is the nucleotide sequence of i-th TE in the set and c_i is the true count of the corresponding TE. If T contains a complete count, we can calculate the nucleotide fraction η_i of each t_i from (1),

$$\eta_i = \frac{c_i \cdot \widetilde{l_i}}{\sum_{j=1}^{M} c_j \cdot \widetilde{l_j}} \tag{1}$$

where $\widetilde{l_i}$ is the effective transcript length of t_i.[27]

We can also calculate the Transcripts Per Million (TPM) using (2),

$$TPM_i = \frac{\frac{\eta_i}{\tilde{l}_i}}{\sum_{j=1}^{M} \frac{\eta_j}{\tilde{l}_j}} \times 10^6 \tag{2}$$

where TPM_i is used as a relative abundance of each transposable element in a given sample.

It is difficult to directly estimate the η and TPM given T and F, so `Salmon` performs the following processes. First, `Salmon` runs a quasi-mapping procedure which is initially proposed in [24]. A quasi-mapping specifies the target of each given read and also determines the position and the orientation of the read concerning the target by computing the Maximum Mappable Prefix (MMP) [28] and Next Informative Position (NIP) [24] of the read. This mapping procedure uses a generalized suffix array[29] and enables a fast and accurate mapping as compared to other mapping tools, such as `Bowtie 2`, `STAR`, and `Kalisto`.[24] The mapping also provides a possible mapping locations for each read.

The maximum-likelihood objective model for a set of reads F is defined as follows:

$$Pr\{F|\eta, Z, T\} = \prod_{j=1}^{N} \sum_{i=1}^{M} Pr\{t_i|\eta\} \cdot Pr\{f_i|t_i, z_{ij} = 1\} \tag{3}$$

where $z_{ij} = 1$ if j-th read in F is derived from i-th TE. Since $Pr\{f_i|t_i, z_{ij} = 1\}$ is unknown, `Salmon` uses the following auxiliary terms to define conditional model to estimate the probability:

$$Pr\{f_j|t_i\} = Pr\{l|t_i\} \cdot Pr\{p|t_i, l\} \cdot Pr\{o|t_i\} \tag{4}$$

where $Pr\{l|t_i\}$ is the probability of drawing a read of the inferred length l given t_i, $Pr\{p|t_i, l\}$ is the probability of the read starting at position p on t_i, $Pr\{o|t_i\}$ is the probability of obtaining a read alignment with the given orientation o to t_i, and this model accounts for sample-specific parameters and biases.

With these probabilistic models, `Salmon` performs online inference to estimate read counts α and nucleotide fraction η using a variant of stochastic collapsed variational Bayesian inference (See Supplementary Algorithm in [20]).[26] In addition to the inference algorithm, `Salmon` constructs equivalence classes for a given F. We assign any pair of reads mapped to same set of target TEs in the same equivalence class. This construction shrink the representation of the sequencing experiment and greatly reduce the running time of offline phase.[20]

Next, `Salmon` starts the offline phase. Given the set of equivalence classes of F, an EM algorithm was used to refine the previous estimation for each equivalence class with following objective function L:

$$L\{\alpha|F, Z, T\} = \prod_{j=1}^{N} \sum_{i=1}^{M} \hat{\eta}_i Pr\{f_j|t_i\} \tag{5}$$

, where $\hat{\eta}_i = \frac{\alpha_i}{\sum_j \alpha_j}$. Once the offline phase is done, `Salmon` outputs the estimation of each TE abundance for F.

2.3. *Statistical tests*

We provide a statistical analysis function to identify differentially expressed TEs from the counts table as the last step of the pipeline. Differential analysis using DESeq2 can handle binary covariates such as binary genotype: phenotype and gender.[30] To handle quantitative covariates such as age, we apply the General Linear Model (GLM).[31] The statistical analysis will produce two statistics to represent associations between the TEs and the covariates: the first one is the test statistics for each TE, and the second one is the summary of the statistics for each clade. The output files are provided with various file formats, such as tab-separated values file (TSV), XML spreadsheet file format (XLS, XLSX), R object file (Rdata), and Portable Document Format (PDF) file.

3. Results

3.1. *Datasets*

Two datasets were used for our comparison to other methods. The first dataset is the RNA-seq data from Gene Expression Omnibus (accession no. GSE47006) which includes wild-type and *Piwi* (P-element Induced WImpy testis) knockdown flies. This dataset was used as a benchmark dataset in the `TEtranscripts` paper as well.[6] We compared the performance in terms of running time and quantification accuracy between our proposed pipeline and other tools, including `TEtranscripts`, `HTSeq-count`, `Cuffdiff` and `RepEnrich`.[14,32–34]

In the second dataset, we seek to identify new TEs that are associated with Amyotrophic Lateral Sclerosis (ALS). We applied our pipeline to a K562 cell-line RNA-seq dataset from ENCODE (Encyclopedia of DNA Elements, `http://encodeproject.org`) Consortium (accession ID: ENCBS555BYH).[35] The dataset consists of two biological replicates of shRNA (short hairpin RNA) knockdown (KD) targeting *TARDBP* (TAR DNA Binding Protein, as known as TDP-43) gene and two biological replicates of controls (a shRNA inserted but targets no genes). It has been reported that loss of *TDP-43* function causes ALS.[7,36] To measure scalability with the dataset. we also ran `TEtranscripts` to compare running time of both methods. We also performed an integrative analysis for highly differentially expressed TEs for further understanding of any new mechanism of ALS.

3.2. *Computational experiment setup*

Generating BAM files from `FASTQ` files are mandatory to `TEtranscripts`, `HTSeq-count`, `Cuffdiff`, and `RepEnrich`, we applied `STAR` [37] to generate the files with the following parameters: `--outFilterMultimapNmax 100` and `--winAnchorMultimapNmax 100`. Sixteen threads were used for both `SalmonTE` and `STAR` . We also used the same parameter setup for each quantification tool similar to the `TEtranscripts` paper.

All of the computational experiments were done in a workstation with `Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz` (10 cores and maximum 40 threads) and `128GBytes` RAM.

3.3. *SalmonTE guarantees a reliable TE expression estimation*

For the quantification accuracy comparison, we first took estimated abundance of 8 TEs from each quantification tool. To validate the results, Reverse Transcription-quantitative Polymerase Chain Reaction (RT-qPCR) was done on these 8 TEs [6]. We observed `SalmonTE` outperformed all other tools ($r^2 = 0.98$, Figure 2 and Table 1). We also found that `SalmonTE` identified a weak down-regulation of DM1731_I and HETA which was missed by `TEtranscripts`.

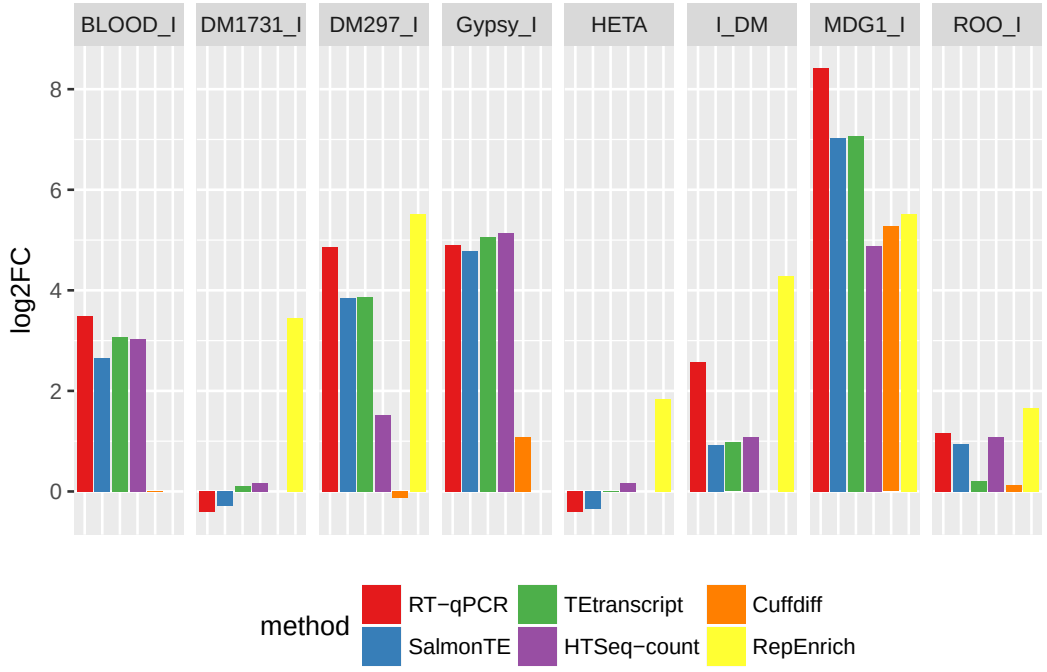

Fig. 2. Comparison of Drosophila TE expression estimation. Four computational methods were compared to `SalmonTE`. RT-qPCR was used to validate the expression levels of the 8 TEs in the Drosophila samples.

Table 1. Pearson Correlation between RT-qPCR and computational TE quantification methods.

Method	`SalmonTE`	`TEtranscripts`	HTSeq-count	Cuffdiff	RepEnrich
r^2	0.98	0.97	0.85	NA	NA

Next, we compared the estimated log_2FC of `SalmonTE` to `TEtranscripts`on each transposable element for a deeper investigation. Our data shows that the estimated TE abundance of both methods are highly correlated ($r^2 = 0.98$), and we also observed there is a high concordance in the direction of fold-changes between `SalmonTE` and `TEtranscripts` (Figure 3). We also measured the correlations of normalized read counts between `SalmonTE` and `TEtranscripts`, and we observed that the calculated read counts from those methods are highly correlated in each sample as well ($r^2 = 0.92$ for wild-type (WT) sample and $r^2 = 0.91$ for

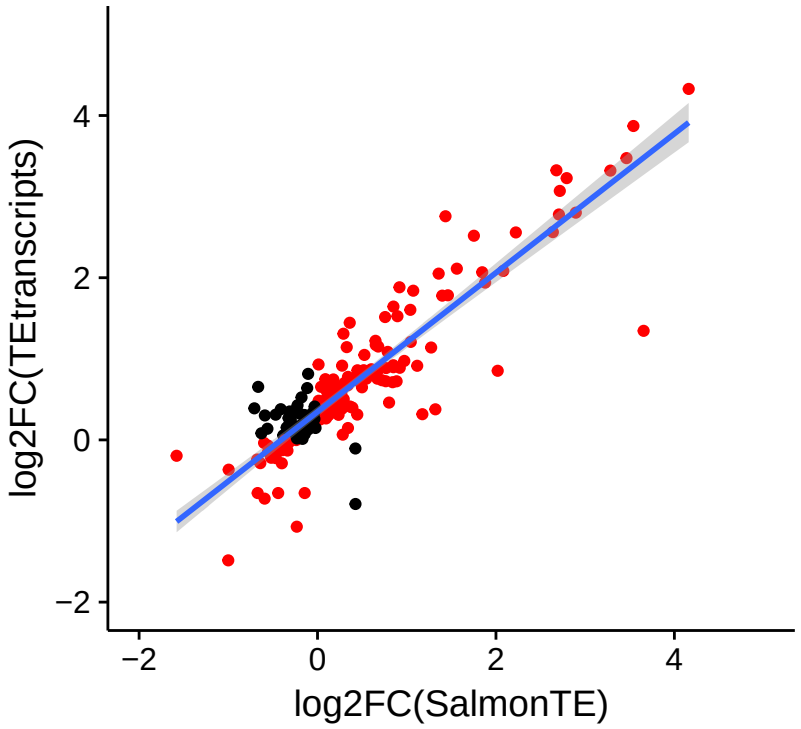

Fig. 3. Correlation of log_2FC ($\frac{Piwi}{WT}$) for each transposable element between `SalmonTE` and `TEtranscripts`. Red points represent TEs with the same fold change direction between `SalmonTE` and `TEtranscripts`.

Piwi KD sample). From this observation, we conclude that both tools generate a similar estimation result. It is not a surprising result because `TEtranscripts` deploys `RSEM` algorithm,[38] and previous studies have demonstrated that transcripts count estimations from `RSEM` and `Salmon` are very correlated.[39,40]

3.4. *SalmonTE shows a better scalability in the speed benchmark dataset*

We measured the speed of `SalmonTE` and`TEtranscripts` on two different datasets (Table 2). Compared to `TEtranscripts`, `SalmonTE` showed a 19x to 27x fold increase in speed. In this analysis, we demonstrate that `SalmonTE` outperformed `TEtranscripts` in processing speed. Our pipeline finishes in less than 5 minutes, while `TEtranscripts` needs about 2 hours to process a single sample. Moreover, our benchmark shows that estimated cost of our pipeline in the cloud computing environment is for the thousands of samples 22 times cheaper than `TEtranscripts` in the computing environment (Table 3).

3.5. *Discover differentially expressed TEs in ALS cell line*

Next, we applied `SalmonTE` pipeline to the *TDP-43* knockdown dataset. We identified 23 transposable elements that are differential expressed between TARDBP knockdown and control cell lines (Table 4) with the threshold of $|log_2FC| \geq 0.5$. No statistical test were performed because the number of replicates in the dataset are small.

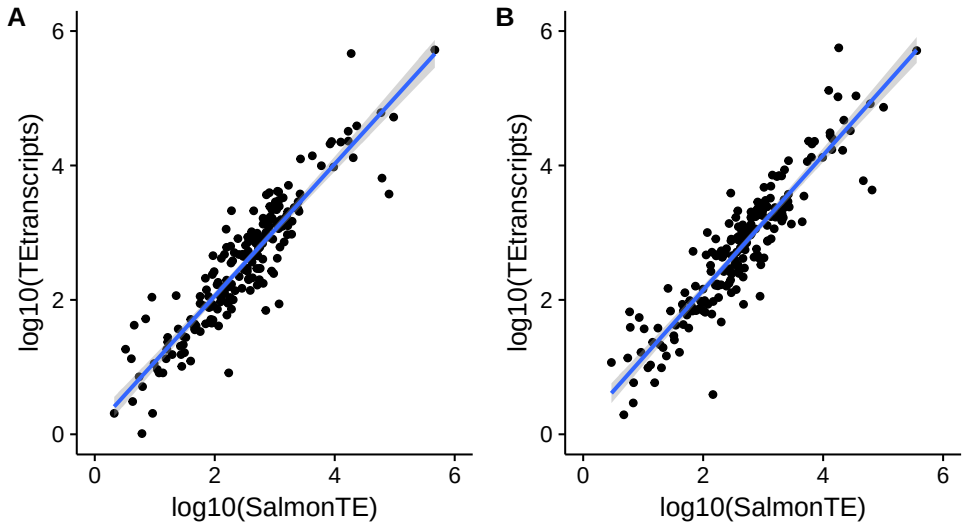

Fig. 4. Sample correlation of count for each transposable element between `SalmonTE` and `TEtranscripts`. **A**. WT sample, **B**. Piwi KD sample.

Table 2. Running speed comparison between `SalmonTE` and `TEtranscripts`.

Dataset	Piwi KD [6]	K562 *TDP-43*
Total number of samples	2	4
RNA-seq file type	Single end	Paired ends
Total number of reads	90,411,467	309,701,182
`SalmonTE` runtime (hh:mm:ss)	0:05:33	0:17:13
`TEtranscripts` runtime (hh:mm:ss)	1:45:26	7:49:40
Speedup	19.00x	27.28x

Table 3. Price estimation of `SalmonTE` and `TEtranscripts` in cloud computing environment (Amazon Elastic Compute Cloud (EC2), and Amazon Elastic Block Store (EBS)). We assume that the size of a `FASTQ` file for a sample is 20GB for the calculations.

Methods	`SalmonTE`	`TEtranscripts`
Estimated using 1000 samples	90 hours	2,000 hours
The price of Amazon EC2 (m4.10xlarge, US Oregon region) [41]	$180	$ 4,000
The price of Amazon EBS (gp2 40TB, US Oregon region) [42]	$500	$ 11,111
Total price	$680	$ 15,111

We can see that most of the differentially expressed features are Endogenous Retrovirus (15 of 23) in *TDP-43* cell-line sample, and we hypothesize that some of the differentially Endogenous Retrovirus TEs are associated with ALS. *TDP-43* is an established and well-studied DNA and RNA binding protein, and could potentially regulate transposable elements at multiple levels.[43] To facilitate a mechanistic understanding of the underlying regulatory mechanism of *TDP-43* and to substantiate the identified differentially expressed transposable, we performed an integrative analysis by combining RNA-seq and *TDP-43* binding data. We obtained DNA binding (ChIP-Seq [44] data) and RNA binding (CLIP-Seq [45] data) datasets

of *TDP-43* in the same K562 cell line from the ENCODE consortium. For illustration, we choose MER74A and AluJo elements that are highly up and down regulated respectively and are also found in Dfam database.[46] We quantified the number of overlapping *TDP-43* ChIP/CLIP peaks with MER74A and AluJo annotations from Dfam. We observed that AluJo element which is down regulated in *TDP-43* knockdown samples is enriched for *TDP-43* ChIP and CLIP peaks as shown in Figure 5, which might indicate that *TDP-43* positively regulate AluJo elements. On the other hand, we did not find any enrichment of *TDP-43* binding for MER74A elements. This preferential binding of *TDP-43* substantiates the differentially expressed transposable elements by our pipeline.

Table 4. 23 Differentially expressed transposable elements in the ENCODE TARDBP data

Name	Clade	log2FC
MER74A	ERV3	1.68
MER57E1	ERV1	1.30
AluYd2	SINE	0.83
LTR1C1	ERV1	0.77
AluSx1	SINE	0.75
LTR27D	ERV1	0.73
AluSx	SINE	0.71
MLT-int	ERV3	0.66
MER54A	ERV3	0.52
MER65D	ERV1	0.51
LTR28	ERV1	-0.59
LTR1F	ERV1	-0.63
FLAM	SINE	-0.64
MER21	ERV3	-0.68
MER101	ERV1	-0.69
LTR26B	ERV1	-0.70
MER83C	ERV1	-0.71
AluJo	SINE	-0.72
LTR06	ERV1	-0.73
MLT2D	ERV3	-0.78
AluYf5	SINE	-0.86
AluYd3	SINE	-1.41
THER2	SINE	-2.03

To identify if there is any general differential expression trend on subfamilies of TEs, we grouped all the TEs based on their clade information. We excluded all of the CR1 (Chicken Repeat 1) since the number of such elements in the clade is small. We found that SINE (Short Interspersed Nuclear Elements) are mostly down expressed, and elements in L1 (Long interspersed nuclear element 1) are generally over expressed in *TDP-43* knockdown samples. This result provides a working hypothesis that knocking-down of *TDP-43* repress the expression of SINE elements and induce the expression of L1 elements.

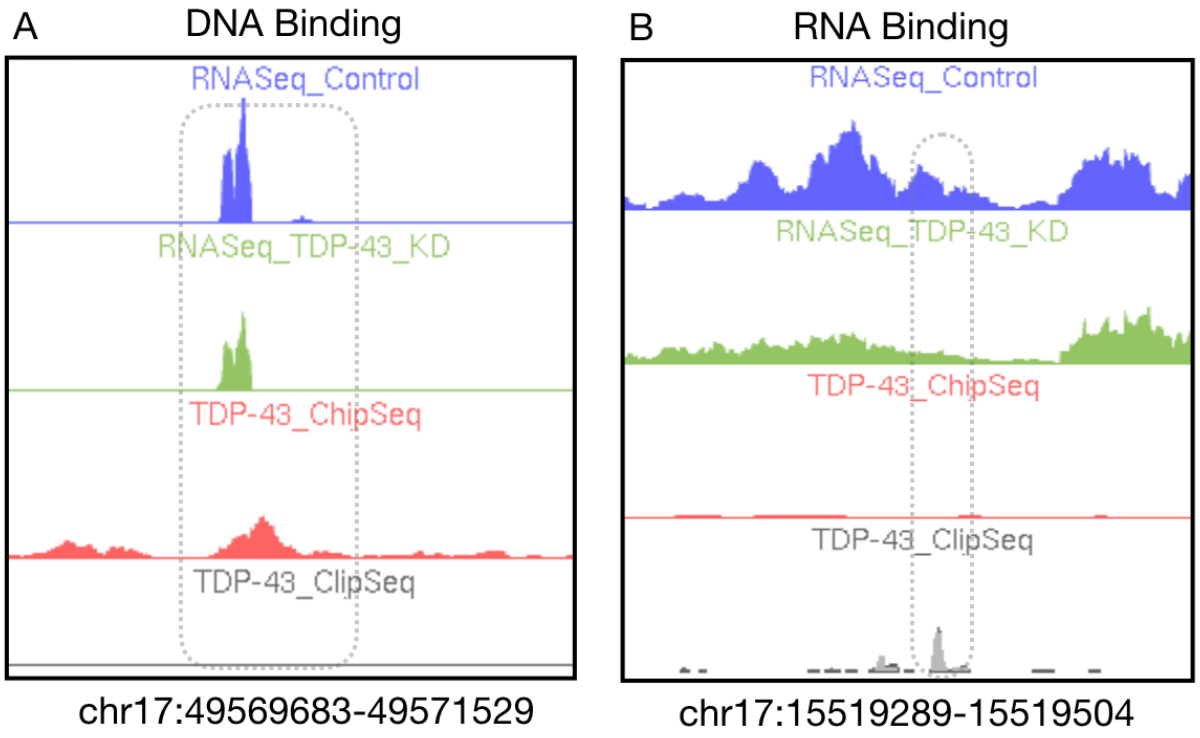

Fig. 5. **A**. Showing down-regulation of AluJo with *TDP-43* ChIP-seq peak, **B**. Showing down-regulation of AluJo with *TDP-43* CLIP-seq peak.

4. Conclusion

In this work, we developed `SalmonTE`, a fast and reliable pipeline for quantification of TEs from NGS data. Our results of `SalmonTE` on the various datasets have shown a large speed-up in computing time relative to `TEtranscripts`, while preserving an accurate quantification on TEs. Therefore, we expect this pipeline will enable the biomedical research community to rapidly quantify and analyze TEs from large amounts of data generated over the past years that are otherwise lost due to genome-masking. Our tool could help the research community to discovery many TE associated with diseases.

There are still several remaining features that could be implemented in the future to improve the usability of `SalmonTE`. For example, prediction of genomic locations, which contain the differentially expressed TEs, is useful in many TE studies. Several methods were developed toward this end,[15,47] but these tools share the scalability issue and require massive computing power for a large-scale TE study. Moreover, alignment free algorithms are intrinsically limited to addressing this question. Therefore, we foresee a novel algorithm which extends and improves the current alignment-free methods.

Acknowledgments

This work has been supported by National Institute of General Medical Sciences R01-GM120033, National Science Foundation - Division of Mathematical Sciences DMS-1263932,

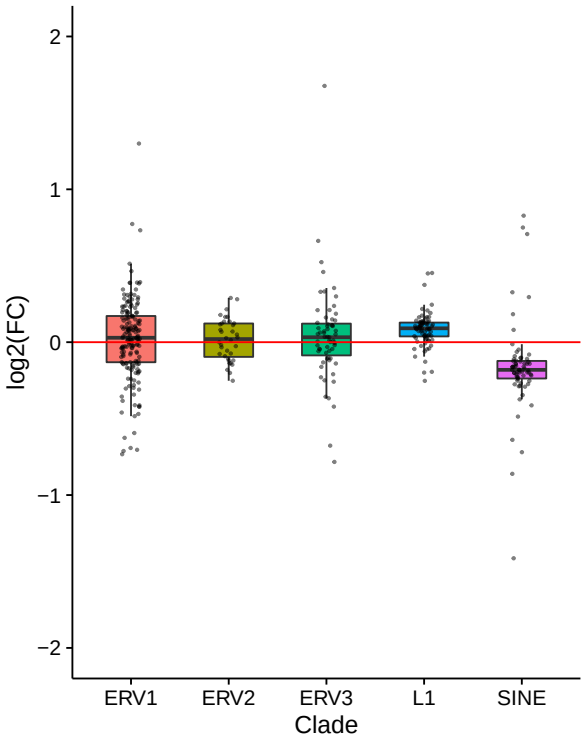

Fig. 6. A boxplot of log_2FC for each clade in the ENCODE *TDP-43* data

Cancer Prevention Research Institute of Texas RP170387, Houston Endowment (Z.L.), and the Alzheimer's Association (J.M.S.). We thank Kala Pham and Rami Al-Ouran for comments that greatly improved this manuscript.

References

1. J. A. Erwin, M. C. Marchetto and F. H. Gage, *Nature Reviews Neuroscience* **15**, 497 (2014).
2. C. Biémont and C. Vieira, *Nature* **443**, 521 (2006).
3. V. P. Belancio, D. J. Hedges and P. Deininger, *Genome research* **18**, 343 (2008).
4. R. L. Jirtle and M. K. Skinner, *Nature reviews. Genetics* **8**, p. 253 (2007).
5. J. G. Wood and S. L. Helfand, *Frontiers in genetics* **4** (2013).
6. H. Ohtani, Y. W. Iwasaki, A. Shibuya, H. Siomi, M. C. Siomi and K. Saito, *Genes & development* **27**, 1656 (2013).
7. S. P. Mihevc, M. Baralle, E. Buratti and B. Rogelj, *Scientific reports* **6**, p. 33996 (2016).
8. W. Li, Y. Jin, L. Prazak, M. Hammell and J. Dubnau, *PloS one* **7**, p. e44099 (2012).
9. L. Krug, N. Chatterjee, R. Borges-Monroy, S. Hearn, W.-W. Liao, K. Morrill, L. Prazak, N. Rozhkov, D. Theodorou, M. Hammell *et al.*, *PLoS genetics* **13**, p. e1006635 (2017).
10. E. Lee, R. Iskow, L. Yang, O. Gokcumen, P. Haseley, L. J. Luquette, J. G. Lohr, C. C. Harris, L. Ding, R. K. Wilson *et al.*, *Science* **337**, 967 (2012).
11. A. Platzer, V. Nizhynska and Q. Long, *Biology* **1**, 395 (2012).
12. E. Helman, M. S. Lawrence, C. Stewart, C. Sougnez, G. Getz and M. Meyerson, *Genome research* **24**, 1053 (2014).
13. E. Hénaff, L. Zapata, J. M. Casacuberta and S. Ossowski, *BMC genomics* **16**, p. 768 (2015).
14. Y. Jin, O. H. Tam, E. Paniagua and M. Hammell, *Bioinformatics* **31**, 3593 (2015).

15. J. R. de Ruiter, S. M. Kas, E. Schut, D. J. Adams, M. J. Koudijs, L. F. Wessels and J. Jonkers, *Nucleic Acids Research* (2017).
16. Z. Tang, J. P. Steranka, S. Ma, M. Grivainis, N. Rodić, C. R. L. Huang, I.-M. Shih, T.-L. Wang, J. D. Boeke, D. Fenyö *et al.*, *Proceedings of the National Academy of Sciences* **114**, E733 (2017).
17. A. D. Ewing, *Mobile DNA* **6**, p. 24 (2015).
18. H. Samet, *The design and analysis of spatial data structures* (Addison-Wesley Reading, MA, 1990).
19. A. Nellore, A. E. Jaffe, J.-P. Fortin, J. Alquicira-Hernández, L. Collado-Torres, S. Wang, R. A. Phillips III, N. Karbhari, K. D. Hansen, B. Langmead *et al.*, *Genome biology* **17**, p. 266 (2016).
20. R. Patro, G. Duggal, M. I. Love, R. A. Irizarry and C. Kingsford, *Nature Methods* **14**, 417 (2017).
21. J. Köster and S. Rahmann, *Bioinformatics* **28**, 2520 (2012).
22. Amazon, Amazon ebs pricing (2017), `http://www.repeatmasker.org/`.
23. G. I. R. Institute, Repbase (2017), `http://www.girinst.org/repbase/update/browse.php`.
24. A. Srivastava, H. Sarkar, N. Gupta and R. Patro, *Bioinformatics* **32**, i192 (2016).
25. C. M. Bishop, *Pattern recognition and machine learning* (springer, 2006).
26. J. Foulds, L. Boyles, C. DuBois, P. Smyth and M. Welling, Stochastic collapsed variational bayesian inference for latent dirichlet allocation, in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013.
27. B. Li, V. Ruotti, R. M. Stewart, J. A. Thomson and C. N. Dewey, *Bioinformatics* **26**, 493 (2009).
28. H. Li, *Bioinformatics* **28**, 1838 (2012).
29. U. Manber and G. Myers, *siam Journal on Computing* **22**, 935 (1993).
30. M. I. Love, W. Huber and S. Anders, *Genome biology* **15**, p. 550 (2014).
31. R. Johnston, *Multivariate statistical analysis in geography; a primer on the general linear model*, tech. rep. (1980).
32. S. Anders, P. T. Pyl and W. Huber, *Bioinformatics* **31**, 166 (2015).
33. C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn and L. Pachter, *Nature biotechnology* **31**, 46 (2013).
34. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy and N. Neretti, *BMC genomics* **15**, p. 583 (2014).
35. E. P. Consortium *et al.*, *Nature* **489**, p. 57 (2012).
36. C. Yang, H. Wang, T. Qiao, B. Yang, L. Aliaga, L. Qiu, W. Tan, J. Salameh, D. M. McKenna-Yasek, T. Smith *et al.*, *Proceedings of the National Academy of Sciences* **111**, E1121 (2014).
37. A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras, *Bioinformatics* **29**, 15 (2013).
38. B. Li and C. N. Dewey, *BMC bioinformatics* **12**, p. 323 (2011).
39. H. Jin, Y.-W. Wan and Z. Liu, *BMC bioinformatics* **18**, p. 117 (2017).
40. C. Zhang, B. Zhang, L.-L. Lin and S. Zhao, *BMC genomics* **18**, p. 583 (2017).
41. Amazon, Amazon ec2 instance pricing (2017), `https://aws.amazon.com/ec2/pricing/on-demand/`.
42. Amazon, Amazon ebs pricing (2017), `https://aws.amazon.com/ebs/pricing/`.
43. Q. Tan, H. K. Yalamanchili, J. Park, A. De Maio, H.-C. Lu, Y.-W. Wan, J. J. White, V. V. Bondar, L. S. Sayegh, X. Liu *et al.*, *Human molecular genetics* **25**, 5083 (2016).
44. D. S. Johnson, A. Mortazavi, R. M. Myers and B. Wold, *Science* **316**, 1497 (2007).
45. R. B. Darnell, *Wiley Interdisciplinary Reviews: RNA* **1**, 266 (2010).
46. R. Hubley, R. D. Finn, J. Clements, S. R. Eddy, T. A. Jones, W. Bao, A. F. Smit and T. J. Wheeler, *Nucleic acids research* **44**, D81 (2015).
47. S. W. Criscione, Y. Zhang, W. Thompson, J. M. Sedivy and N. Neretti, *BMC genomics* **15**, p. 583 (2014).