




Sequence analysis

ExplorATE: a new pipeline to explore active transposable elements from RNA-seq data

Martin M. Femenias ^{1,*}, Juan C. Santos², Jack W. Sites Jr^{3,†}, Luciano J. Avila¹ and Mariana Morando¹

¹Consejo Nacional de Investigaciones Científicas y Técnicas, Instituto Patagónico para el Estudio de los Ecosistemas Continentales (IPEEC-CONICET), Puerto Madryn, CT U9120ACD, Argentina, ²Department of Biological Sciences, St. John's University, Queens, NY 11439, USA and ³Department of Biology and M.L. Bean Life Science Museum, Brigham Young University (BYU), Provo, UT 84602, USA

*To whom correspondence should be addressed.

†Present address: Department of Biology, Austin Peay State University, Clarksville, Tennessee 37044, USA.

Associate Editor: Christina Kendzierski

Received on July 15, 2021; revised on May 3, 2022; editorial decision on May 17, 2022; accepted on May 19, 2022

Abstract

Motivation: Transposable elements (TEs) are ubiquitous in genomes and many remain active. TEs comprise an important fraction of the transcriptomes with potential effects on the host genome, either by generating deleterious mutations or promoting evolutionary novelties. However, their functional study is limited by the difficulty in their identification and quantification, particularly in non-model organisms.

Results: We developed a new pipeline [explore active transposable elements (ExplorATE)] implemented in R and bash that allows the quantification of active TEs in both model and non-model organisms. ExplorATE creates TE-specific indexes and uses the Selective Alignment (SA) to filter out co-transcribed transposons within genes based on alignment scores. Moreover, our software incorporates a Wicker-like criteria to refine a set of target TEs and avoid spurious mapping. Based on simulated and real data, we show that the SA strategy adopted by ExplorATE achieved better estimates of non-co-transcribed elements than other available alignment-based or mapping-based software. ExplorATE results showed high congruence with alignment-based tools with and without a reference genome, yet ExplorATE required less execution time. Likewise, ExplorATE expands and complements most previous TE analyses by incorporating the co-transcription and multi-mapping effects during quantification, and provides a seamless integration with other downstream tools within the R environment.

Availability and implementation: Source code is available at <https://github.com/FemeniasM/ExplorATEproject> and https://github.com/FemeniasM/ExplorATE_shell_script. Data available on request.

Contact: mmfemenias@gmail.com

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

Transposable elements (TEs) are ‘jumping’ DNA sequences that change their position within the host genome. TEs are ubiquitous in eukaryotes (Chalopin *et al.*, 2015; Kapusta *et al.*, 2017; Pasquesi *et al.*, 2018) and their presence drive genome evolution in many lineages (Bourque *et al.*, 2008; Jurka *et al.*, 2011; Lynch *et al.*, 2015; Sotero-Caio *et al.*, 2017; Zeng *et al.*, 2018). Although TEs can become immobile ‘molecular fossils’ after accumulating mutations that impede their mobilization (Lanciano and Cristofari, 2020; Sotero-Caio *et al.*, 2017), many remain active and may impact the gene expression in their host genomes (Lanciano and Cristofari, 2020; Teissandier *et al.*, 2019). Recent studies show that TEs are

involved in important processes that alter gene-expression dynamics, which include changes in regulatory networks (Bourque *et al.*, 2008; Chuong *et al.*, 2017; Feschotte, 2008; Kokošar and Kordiš, 2013; Lynch *et al.*, 2015), the activation of cellular signaling pathways (Aravin *et al.*, 2001; De Cecco *et al.*, 2019; Pasquesi *et al.*, 2020) and the alteration of chromatin accessibility (Jachowicz *et al.*, 2017). These TEs that bypass genome control mechanisms (i.e. end up being expressed) constitute a significant fraction of host transcriptomes and can be recovered in RNA-seq experiments, providing valuable information about the impact of transposons on the host genome expression patterns (Faulkner and Carninci, 2009; Garcia-Perez *et al.*, 2016; Nishihara, 2020; Pasquesi *et al.*, 2020; Sundaram

and Wysocka, 2020). However, the expression data associated with TEs are usually discarded without further analysis or exploration, and their impact on the host gene-expression dynamics is ignored (Jin et al., 2015; O'Neill et al., 2020; Slotkin, 2018).

High-throughput sequencing has allowed further studies on TEs, including their functional impact and their role in gene regulation (Teissandier et al., 2019). However, the characterization of repetitive sequences from raw reads remains a challenge. The main drawbacks are the multi-mapper reads (Teissandier et al., 2019; Treangen and Salzberg, 2012), and the identification of source sequences, either autonomous TE uni-length transcripts, or embedded repeats in gene-derived transcripts [e.g. those derived from co-expression and pervasive transcription, as reviewed by Lanciano and Cristofari (2020)]. Given the repetitive nature and co-existence of related TE families in the same reference, short reads can ambiguously map to different genomic regions or different transcripts that share some sequence similarity. Therefore, these multi-mapper reads cannot be assigned to a particular sequence of origin, hindering the subsequent alignment procedure.

To avoid multi-mapping, some tools use reads that map once (uni-mappers) underestimating expression levels in young TE families (Chung et al., 2019; Lanciano and Cristofari, 2020). Alternatively, the multi-mapped reads can be reassigned using the expectation maximization (EM) algorithm (Li and Dewey, 2011). This algorithm can handle the ambiguity of multi-mapper reads, and it is used in most TE analysis software (Lanciano and Cristofari, 2020). In addition, the expression of gene transcripts carrying TEs can result in the overestimation of some TE families. This co-expression of TEs occurs when they are within the intronic regions of a gene or when the transcription transcends the boundaries of the gene. Thus, non-removal of the intronic region or weak polyadenylation sites can lead to transcription of TEs from gene promoters. The identification and exclusion of co-transcribed TEs prior to abundance estimates might not be trivial, as a substantial fraction of such TEs could come from protein-coding transcripts carrying repeats within introns or UTR regions (Chung et al., 2019; Faulkner and Carninci, 2009).

To date, many bioinformatics tools have been developed for TE identification and quantification from RNA-seq data [comparisons of explore active transposable elements (ExplorATE) with other published tools and approaches are summarized in Supplementary Table S1]. Most of these algorithms are based on a model transcriptome or a reference genome. For example, Tetranscripts (Jin et al., 2015) is one of the most widely used tools; its algorithm focus on sequence similarities in different TE hierarchical levels, and properly handle multiple assignments of reads using the EM algorithm. Chung et al. (2019) modified the original Tetranscripts pipeline to reduce TE read counts based on the reads coverage of the surrounding introns, and thus partially corrects the overestimation caused by reads derived from pre-mRNA or introns retained in mature mRNA.

Other Salmon-based software (Patro et al., 2017) for TE analyses include SalmonTE (Jeong et al., 2018) and REdiscoverTE (Kong et al., 2019). For instance, SalmonTE estimates abundance uncertainty due to random sampling and the ambiguity introduced by multi-mappers reads. Moreover, SalmonTE uses a Rebase-based TE model transcriptome for fast quantification with pre-built indexes, but this approach may introduce TE age-related biases and/or lead to false mappings in samples with divergent or missing TEs in the model transcriptome (Kong et al., 2019). In addition, fragments derived from an un-annotated genomic locus can be falsely mapped to an annotated transcript (Srivastava et al., 2020).

Another promising Salmon-based software is REdiscoverTE (Kong et al., 2019) which has reduced mapping ambiguity by incorporating an extensive model transcriptome. This reference includes transcripts and introns from GENECODE, and RepeatMasker elements derived from human reference sequences. However, a general limitation of SalmonTE and REdiscoverTE is the use of quasi-mapping approach (a default in older versions of Salmon), which may generate spurious mappings and reduce the accuracy compared

to alignment-based approaches (Sarkar et al., 2018; Srivastava et al., 2020; Vuong et al., 2018).

As mentioned above, current tools for TE analysis are based on traditional aligners (like Bowtie, STAR or HISAT) or Salmon's lightweight mapping. Although lightweight mapping achieves faster and accurate quantification in most cases, this approach loses sensitivity and specificity particularly when sequence similarity is high. The Selective Alignment (SA) (Sarkar et al., 2018; Srivastava et al., 2020) is described as a hybrid approach that improves the sensitivity and specificity of previous lightweight mapping by reducing false exact matches between reads and target sequences. This procedure (default in Salmon versions > 1.0.0) selects alignments by scores comparison using an extension alignment scoring algorithm (Li, 2018; Suzuki and Kasahara, 2018; the SA procedure is summarized in Supplementary Note S1). Overall, the SA shows marked improved performance over the lightweight mapping approaches without compromising execution speed (Srivastava et al., 2020). A comparison of traditional alignment, lightweight mapping and SA strategies is shown in Supplementary Table S2.

Here, we propose a novel method to alleviate above-described shortcomings by identifying and quantifying active TEs in both model and non-model organisms. Our method adapts the SA score comparisons to simultaneously handle TE co-transcription (retained in introns regions or UTRs) and multi-mapping problems. Our software ExplorATE, which can be run as bash code or as an R package, uses RepeatMasker output files and gene models to identify potentially active TEs (target TEs) and decoy sequences (co-transcribed TEs or other non-TEs-derived repetitive sequence) from genomes or *de novo* transcriptomes. During its performance, ExplorATE uses the identified target TEs and decoy sequences to create TE-specific indexes and execute the SA algorithm (Srivastava et al., 2020). Thus, exact matches between TEs and sequence reads are filtered by scores and read sequence information (i.e. expected distance between paired-end reads and orientation), differentiating between target TEs and decoy sequences before Salmon estimations. Finally, ExplorATE incorporates functions to import estimates, and creates ready-to-use datasets for further analyses (e.g. differential expression) in R.

2 System and methods

2.1 Pipeline overview

ExplorATE algorithm is summarized in three steps: (i) identification of target TEs and decoy TE sequences; (ii) quantification of each identified TE using the Salmon algorithm with the SA option and (iii) importation of TE counts into the R environment. The ExplorATE algorithm defines target TEs as those sequences that were annotated as transposons but are not overlapped with gene-coding transcripts. All TEs overlapping with gene-coding transcripts and other sequences that share some homology with transposons are defined as 'decoy sequences'. The ExplorATE algorithm can perform two alternative analyses based on the availability of a reference genome (Fig. 1).

If no reference genome is provided ('non-model organisms' mode, 'nmo' in bash script), ExplorATE takes the repeats of active elements identified by RepeatMasker from a *de novo* transcriptome. The transcriptome-derived RepeatMasker file is processed to resolve overlapping repeats within each transcript. ExplorATE allows to choose from three criteria for overlapping resolution: (i) based on the highest score, (ii) longest length and (iii) lower divergence relative to the reference library used in RepeatMasker. To establish the identity of TEs, ExplorATE also allows to set a criterion similar to Wicker's rule (Wicker et al., 2007). Under this criterion, the algorithm assigns target transcripts based on the percentage of identity for a class/family of TEs, the percentage for each TE class/family as the ratio between TE class/family length with respect to the transcript length, and a minimum of transcript length. For example, the default '80-80-80' Wicker-like rule is a selection criterion where the transcripts will be considered targets if they have a TE class/family annotation with >80% identity (from RepeatMasker file),

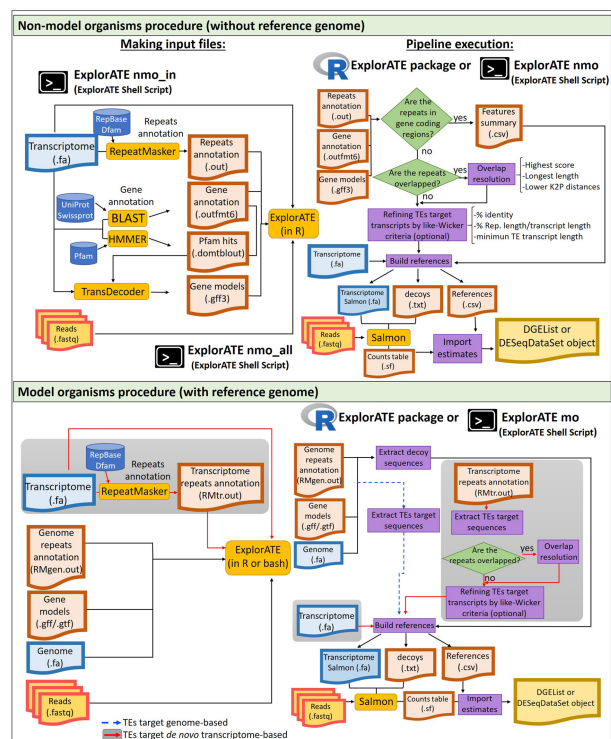


Fig. 1. ExplorATE workflow summary for with or without reference genome procedures: (top panel) non-model organisms without reference genome requires external input files that include a *de novo* transcriptome and gene annotations to identify target TEs for Salmon quantification. (Bottom panel) Model organisms with reference genome procedure allows users to define the target TEs from genome-based intergenic regions (dotted arrows), or use a *de novo* transcriptome with RepeatMasker annotations (shaded boxes). The ExplorATE pipeline can be applied by R package functions or by a shell script (see Section 2.1)

represents >80% of the transcript length, and target transcripts must have at least 80 bp in length. The user can change the default values to make this rule more or less stringent.

Finally, ExplorATE uses a gene annotation file (e.g. derived from BLAST) and a 'gene models' (e.g. as those derived from TransDecoder output files), to identify TEs within genes to be used as decoy sequences.

If ExplorATE is run under a reference genome modality ('model organisms' mode, 'mo' in bash script), the user has two alternatives to run the TEs analysis: (i) using a set of TEs from intergenic regions of the genome as target or (ii) using a *de novo* transcriptome (and its RepeatMasker annotations) to define target TEs. In both cases, a genome-derived RepeatMasker file will be required to extract decoy sequences from genic regions. Unlike other Salmon-based programs (e.g. SalmonTE or RediscoverTE), our method takes advantage of Salmon's SA strategy (Sarkar *et al.*, 2018; Srivastava *et al.*, 2020).

As described by Srivastava *et al.* (2020), SA is based on the collection of maximum exact matches between reads and transcripts (uni-MEM), and the subsequent application of the chaining algorithm from Minimap2 (Li, 2018). During the chaining procedure, a score for each alignment is obtained, which allows filters to be applied to discard spurious alignments. For example, alignments with scores less than a threshold of 0.65 are removed, and scores between target and decoy sequences are compared. Thus, filters applied by the software allow retains only those alignments whose score is better in target sequences than decoy sequences and exceed the threshold of 0.65. All valid alignments are used during the quantification phase. A summary of the SA procedure is provided in Supplementary Note S1.

The implementation of the SA filters above-described allows ExplorATE exclude spurious alignments (e.g. those that do not reach a threshold score of at least 0.65), and alignments derived

from co-transcribed TEs before applying the EM in the Salmon quantification. These filters are needed because they reduce mapping ambiguity, making the application of the EM algorithm more efficient during counts estimation. The resulting counts can be imported into the R environment with ExplorATE's functions generating ready-to-use normalized count files that can be streamlined toward differential expression analyses in edgeR or DESeq2 (Love *et al.*, 2014; Robinson *et al.*, 2010).

2.2 ExplorATE implementation

We used both simulated and real data to evaluate the performance of our program and for comparisons with other similar pipelines. For instance, we compared our ExplorATE results under 'mo' modality with those Tetranscripts and SalmonTE, while we compared the 'nmo' modality with the results of TETOOLS (see Supplementary Methods).

2.2.1 Simulated data analyses

ExplorATE is designed to exclude spurious mappings of co-transcribed TEs with genes and correcting for counts of non-co-transcribed TEs. To evaluate our software accuracy, we simulated data from a human ovary sample (SRA: ERR579132) and the ExplorATE's performance was estimated as the abundance of the non-co-transcribed TEs. We filtered such abundances in a procedure that allowed us to make comparisons with other software (see Supplementary Methods); and evaluated the agreements between the simulated and estimated abundances using Spearman's correlations, and the mean absolute relative difference (MARD).

2.2.2 Real data analyses

We compared ExplorATE with SalmonTE and Tetranscripts programs under the 'mo' modality, and with TETOOLS for 'nmo' alternative. The source data for 'mo' analyses were two paired-end samples (SRA: SRR4421820 and SRR4421821) of K562 human cell-line treated with a shRNA knockdown (KD) against TDP-43. We also included the *Drosophila melanogaster* ovarian cell single-end dataset (accession number GSE47006) from Ohtani *et al.* (2013). For this dataset, we included a control sample and a piwi KD; the latter is a sample from DmGTSF1 that was altered to cause TEs deregulation. We compared the estimates from all these pipelines through Spearman's correlations (see Supplementary Methods).

For 'nmo' analyses, we benchmarked ExplorATE with the 'TEcounts.py' module of TETOOLS (Lerat *et al.*, 2017) using the RNA-seq reads of the lizard *Liolaemus parthenos* from the ovary, brain and liver tissues as test dataset (SRA: SRR14320877-79). Details of the RNA-seq procedures are provided in the Supplementary Information.

2.2.3 Computation time benchmarking

To compare execution times, we ran ExplorATE with 'mo' and 'nmo' modalities; and compared these with Tetranscripts and TETOOLS performances using the above-described data. All programs were run on a local computer [CPU: Intel (R) Core (TM) i9-7900X CPU @ 3.30 GHz, 64 GB RAM] allocating 12 threads maximum. To evaluate time dependency of the ExplorATE performance as a function of the library size and the number of threads, we subsampled libraries to 4M, 10M and 20M total reads. Then, we executed independent runs assigning 4, 8, 12, 16 and 20 threads in the 'nmo' procedure.

3 Results

3.1 Simulated data

Based on simulated data, ExplorATE was highly accurate in the abundance estimation for each TE subfamily and at the transcripts level [$r > 0.80$, P -values < 0.001 and MARD < 0.24 ; Fig. 2, Supplementary Fig. S1, confidence intervals (CI) in Supplementary Table S3]. For the count comparisons, ExplorATE achieved a slightly better fit than

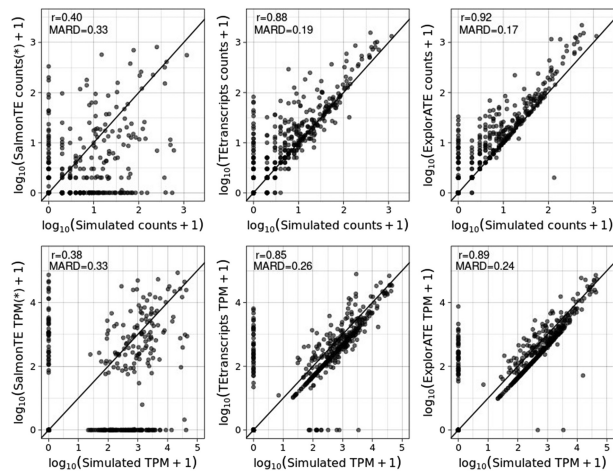


Fig. 2. Benchmarking counts (top) and TPM (bottom) estimations of non-co-transcribed TEs at the subfamily level. From left to right, comparisons of RSEM versus SalmonTE, Tetranscripts and ExplorATE estimates. Performance precision was measured using Spearman's correlation coefficient (r , P -values < 0.001) and MARD. Identity line is shown as reference on each comparison. The asterisk indicates that for SalmonTE was not possible to isolate the counts for non-co-transcribed TEs and only the overall estimate for each subfamily from the model transcriptome is shown (details in [Supplementary Methods](#))

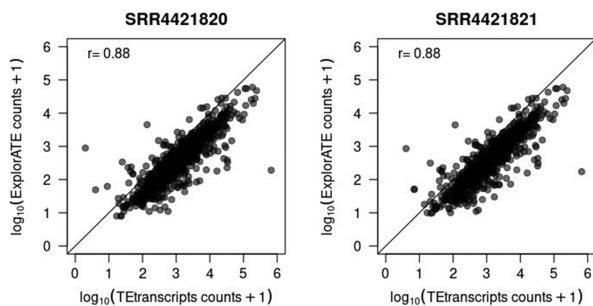


Fig. 3. Counts comparisons between Tetranscripts and ExplorATE for the K562 human cell-line samples. Performance precision was measured using the Spearman's correlation coefficient (r , P -values < 0.001). The identity line is shown as reference

Tetranscripts on this set of non-co-transcribed elements ($r = 0.916$, CI: 0.900–0.929 for ExplorATE versus RSEM; $r = 0.879$, CI: 0.858–0.897 for Tetranscripts versus RSEM; P -values < 0.0001 ; [Fig. 2](#), [Supplementary Table S3](#)), yet the TPM estimates from both tools were highly congruent ($r = 0.887$, CI: 0.866–0.904 for ExplorATE versus RSEM; $r = 0.845$, CI: 0.819–0.869 for Tetranscripts versus RSEM; P -values < 0.0001 ; [Fig. 2](#), [Supplementary Table S3](#)).

3.2 Real data

Our analyses showed that ExplorATE and Tetranscripts estimates were strongly correlated when we used real datasets under the 'mo' approach ($r > 0.750$, P -value < 0.001 ; [Fig. 3](#), [Supplementary Figs S2 and S3](#)). In contrast, the ExplorATE and SalmonTE estimates were only strongly correlated with *D.melanogaster* dataset ([Supplementary Fig. S4](#)). For this one, all three software (i.e. ExplorATE, Tetranscripts and SalmonTE) are strongly correlated with each other when pairs of complete observations were used ($r > 0.750$, P -values < 0.001 ; [Supplementary Table S3](#)). However, when missing values are coded as zeros, ExplorATE and Tetranscripts were strongly correlated ($r = 0.863$, CI: 0.811–0.901 for the control sample and $r = 0.847$, CI: 0.790–0.889 for the piwi-KD sample; P -values < 0.001), while SalmonTE correlates with ExplorATE and Tetranscripts were moderate to strong ($r < 0.671$ with upper limits < 0.752 , P -values < 0.001 ; [Supplementary Table S3](#)).

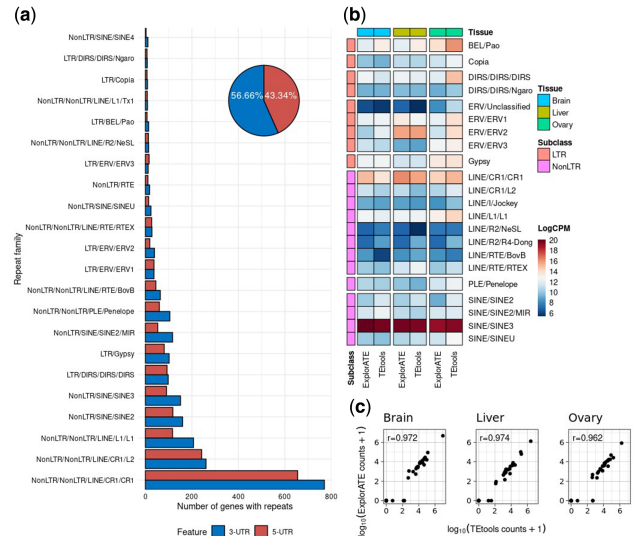


Fig. 4. Benchmarking of ExplorATE and TETOOLS at the superfamily level using the *L.parthenos* tissues. (a) Counts of repeats families co-transcribed with genes in the UTR regions identified by ExplorATE. Pie-chart shows the percentage of co-transcribed repeats for each UTR region and the bar-plot shows the repeat counts in each gene region given by each retrotransposon family. (b) Abundance heatmap of active retrotransposons for each tissue. (c) Correlation panels of TE counts between ExplorATE and TETOOLS. Log CPM: logarithm of counts per million mapped reads. r , Spearman's correlation coefficient (P -value < 0.001)

ExplorATE achieved accurate estimates for missing subfamilies in SalmonTE ([Supplementary Fig. S3](#)). As expected, SalmonTE's underlying lightweight mapping errors were reduced with the SA heuristic implemented in ExplorATE. This outcome was evidenced in the human K562 cell-line dataset, where the overall mapping ambiguity was greater than in the *D.melanogaster* dataset. For the K562 dataset, SalmonTE failed to capture the diversity of TEs in the samples ([Supplementary Figs S3 and S4](#)), whereas ExplorATE achieved more precise estimates and consistent with those of the Tetranscripts ([Fig. 3](#) and [Supplementary Fig. S3](#)).

Likewise, our analyses under the 'mo' approach showed that the ExplorATE results are highly congruent with the alignment-based TETOOLS approach for the *L.parthenos* dataset from all tissues ($r \geq 0.962$ with P -value < 0.001 , [Fig. 4](#); lower limit of CI ≥ 0.914 for all tissues, [Supplementary Table S3](#)). ExplorATE, unlike TETOOLS, identified co-transcribed repetitive elements in UTR regions in at least 3346 decoy transcripts ([Fig. 4a](#)). Therefore, ExplorATE achieved more conservative estimates for these TEs subfamilies that overlapped with coding genes, particularly for ovarian tissue ([Fig. 4b](#)).

3.3 Execution times

We found that ExplorATE was 3.5–32 times faster than alignment-based tools for all datasets ([Supplementary Table S4](#)). Likewise, the execution times of ExplorATE are considerably reduced in parallel processing regardless of the library size and time improvement was 1–4 min less per core when using 4–12 threads ([Supplementary Fig. S5 and Table S5](#)). However, this may vary depending on the size of the annotation file processed.

4 Discussion

We introduce a new R package ExplorATE that allows the identification and quantification of TEs from RNA-seq data with or without a reference genome. This novel pipeline uses the SA to simultaneously address two key issues while analyzing TEs from RNA-seq data, including multi-mapping and co-transcription of TEs. ExplorATE uses gene models to filter fragments with TEs in genic and intergenic regions, and then uses these TE sequences as

decoys or targets. During the SA procedure, chaining scores are compared, and spurious alignments derived from introns or UTR regions are removed prior to the EM application in the Salmon quantification estimation.

After excluding all co-transcribed elements, we found that ExplorATE results strongly correlated those of RSEM at the sub-family level (Fig. 2). Moreover, we found that ExplorATE seldom underestimates non-co-transcribed TE subfamilies, which supports the ExplorATE's ability to exclude spurious alignments derived from genic regions prior to quantification. Chung *et al.* (2019) found contrasting co-expression profiles of intronic and intergenic TEs that are distant from genes, and suggested that a by depth reads correction of co-transcribed elements prior to EM execution would be a more accurate quantification of TEs. We followed this observation and demonstrated that exclusion of spurious alignments before EM improves the mapping performance. In addition, instead of reducing counts proportionally to depth reads, ExplorATE selects the best alignments between reads and TEs for a set of non-co-transcribed target TEs. Therefore, the EM algorithm is applied only to the set of best alignments, achieving an overall better fit and lower overestimation of non-co-transcribed TEs than published tools that lack any correction (Fig. 2).

Although it was not possible to filter non-co-transcribed counts in SalmonTE analyses (Supplementary Methods), our SalmonTE's estimates using simulated data varied significantly from the estimates of non-co-transcribed elements (Fig. 2). Furthermore, analyses of real data derived from model organisms showed that ExplorATE results are strongly correlated with those of Tetranscripts, particularly for the paired-end dataset (Fig. 3 and Supplementary Figs S3 and S4). When paired-end libraries are available, the SA procedure uses sequence information (orientation and distances between paired reads) to filter alignments and improve overall mapping performance. As Kong *et al.* (2019) have demonstrated, the implementation of model transcriptome in SalmonTE is a limiting factor in the TE estimation for samples with ample diversity of these elements. These authors further noted that the homology between TEs and canonical human genes can lead to significant bias in the estimation, and they propose the use of an extensive model transcriptome based on RepeatMasker sequences and GENCODE sequences. This extensive model transcriptome is an accurate solution when reference sequences are available. We consider the combination of the extensive model transcriptome with the SA approach to be very convenient if using human or mouse datasets. However, this approach cannot be generalized to phylogenetically distant organisms (e.g. non-model organisms in which only *de novo* transcriptomes were studied). Alternatively, we propose to identify non-co-transcribed TEs and decoy sequences from the data, either using a genome or a *de novo* transcriptome. Our results for human and *D.melanogaster* data show that ExplorATE obtained accurate counts for TEs families that were not identified by SalmonTE, overcoming the reference bias by the model transcriptome of this program (Supplementary Fig. S3).

Like other quantification tools, the quality of the estimates by ExplorATE will depend on the TEs annotation inputs. RepeatMasker provides annotation files for several model organisms (<https://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html>) and emerging TE annotation tools could be easily coupled to the ExplorATE pipeline (e.g. Berthelier *et al.*, 2018; Flynn *et al.*, 2020; Ou *et al.*, 2019). Most users might make *de novo* annotations of TEs (from raw reads or genome-based; Goerner-Potvin and Bourque, 2018), or use them to complement RepeatMasker annotations. Our 'mo' procedure allows to design TE targets either from intergenic regions of the genome, or from a 'masked' *de novo* transcriptome. The *de novo* transcriptome and TE identification in RepeatMasker may be the only alternative for a first TEs overview in non-model organisms, and it is an essential step in our 'nmo' procedure. For this reason, ExplorATE incorporates specific heuristics for the identification of target TEs based on a Wicker-like rule that could reduce spurious mapping. Furthermore, with RNA-seq data from non-model organisms, we recommend using libraries of TEs from closely related species, combined with *de novo* TEs libraries.

In sum, the ExplorATE package presented here provides specific functions to identify target and decoy sequences, and takes advantage of SA to correct spurious mappings derived from co-transcribed elements and other genomic regions. ExplorATE achieved accurate estimates both with and without a reference genome, with a low computational cost, and it showed that the alignment scores comparisons are useful to handle the TEs co-expression. The TE-derived alignments selection procedure of ExplorATE can be used to advance research on TEs quantification, e.g. combining specific heuristics for target TEs identification with fast and more accurate sequence aligners.

Acknowledgements

We thank Dr Todd A. Castoe and Dr Giulia Pasquesi for their help with lizard TEs libraries.

Funding

This work was supported by the Fondo para la Investigación Científica y Tecnológica [PICT2011-0784 to L.J.A.; PICT2015-1252 to M.M.]; National Science Foundation [DEB-2016372 to J.C.S.] and St. John's University start-up funds [to J.C.S.].

Conflict of Interest: none declared.

References

- Aravin,A.A. *et al.* (2001) Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.*, **11**, 1017–1027.
- Berthelier,J. *et al.* (2018) A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*. *BMC Genomics*, **19**, 14.
- Bourque,G. *et al.* (2008) Evolution of the mammalian transcription factor binding repertoire via transposable elements. *Genome Res.*, **18**, 1752–1762.
- Chalopin,D. *et al.* (2015) Comparative analysis of transposable elements highlights mobilome diversity and evolution in vertebrates. *Genome Biol. Evol.*, **7**, 567–580.
- Chung,N. *et al.* (2019) Transcriptome analyses of tumor-adjacent somatic tissues reveal genes co-expressed with transposable elements. *Mob. DNA*, **10**, 39.
- Chuonq,E.B. *et al.* (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.
- De Cecco,M. *et al.* (2019) L1 drives IFN in senescent cells and promotes age-associated inflammation. *Nature*, **566**, 73–78.
- Faulkner,G.J. and Carninci,P. (2009) Altruistic functions for selfish DNA. *Cell Cycle*, **8**, 2895–2900.
- Feschotte,C. (2008) Transposable elements and the evolution of regulatory networks. *Nat. Rev. Genet.*, **9**, 397–405.
- Flynn,J.M. *et al.* (2020) RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. USA*, **117**, 9451–9457.
- Garcia-Perez,J.L. *et al.* (2016) The impact of transposable elements on mammalian development. *Development*, **143**, 4101–4114.
- Goerner-Potvin,P. and Bourque,G. (2018) Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.
- Jachowicz,J.W. *et al.* (2017) LINE-1 activation after fertilization regulates global chromatin accessibility in the early mouse embryo. *Nat. Genet.*, **49**, 1502–1510.
- Jeong,H.H. *et al.* (2018) An ultra-fast and scalable quantification pipeline for transposable elements from next generation sequencing data. In: *Pacific Symposium on Biocomputing, Kona Coast, Hawaii*, World Scientific Publishing Co. Pte Ltd, pp. 168–179.
- Jin,Y. *et al.* (2015) Tetranscripts: a package for including transposable elements in differential expression analysis of RNA-seq datasets. *Bioinformatics*, **31**, 3593–3599.
- Jurka,J. *et al.* (2011) Families of transposable elements, population structure and the origin of species. *Biol. Direct.*, **6**, 44.
- Kapusta,A. *et al.* (2017) Dynamics of genome size evolution in birds and mammals. *Proc. Natl. Acad. Sci. USA*, **114**, E1460–E1469.
- Kokošar,J. and Kordis,D. (2013) Genesis and regulatory wiring of retroelement-derived domesticated genes: a phylogenomic perspective. *Mol. Biol. Evol.*, **30**, 1015–1031.

- Kong, Y. et al. (2019) Transposable element expression in tumors is associated with immune infiltration and increased antigenicity. *Nat. Commun.*, **10**, 5228.
- Lanciano, S. and Cristofari, G. (2020) Measuring and interpreting transposable element expression. *Nat. Rev. Genet.*, **21**, 721–736.
- Lerat, E. et al. (2017) TEtools facilitates big data expression analysis of transposable elements and reveals an antagonism between their activity and that of piRNA genes. *Nucleic Acids Res.*, **45**, e17.
- Li, H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
- Li, B. and Dewey, C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.
- Love, M.I. et al. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
- Lynch, V.J. et al. (2015) Ancient transposable elements transformed the uterine regulatory landscape and transcriptome during the evolution of mammalian pregnancy. *Cell Rep.*, **10**, 551–561.
- Nishihara, H. (2020) Transposable elements as genetic accelerators of evolution: contribution to genome size, gene regulatory network rewiring and morphological innovation. *Genes Genet. Syst.*, **94**, 269–281.
- Ohtani, H. et al. (2013) DmGTSF1 is necessary for piwi-piRISC-mediated transcriptional transposon silencing in the *drosophila* ovary. *Genes Dev.*, **27**, 1656–1661.
- O'Neill, K. et al. (2020) Mobile genomics: tools and techniques for tackling transposons. *Philos. Trans. R. Soc. B Biol. Sci.*, **375**, 20190345.
- Ou, S. et al. (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.*, **20**, 275.
- Pasquesi, G.I.M. et al. (2018) Squamate reptiles challenge paradigms of genomic repeat element evolution set by birds and mammals. *Nat. Commun.*, **9**, 2774.
- Pasquesi, G.I.M. et al. (2020) Vertebrate lineages exhibit diverse patterns of transposable element regulation and expression across tissues. *Genome Biol. Evol.*, **12**, 506–521.
- Patro, R. et al. (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.
- Robinson, M.D. et al. (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
- Sarkar, H. et al. (2018) Towards selective-alignment: bridging the accuracy gap between alignment-based and alignment-free transcript quantification. In: ACM-BCB 2018—Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Association for Computing Machinery, Inc., New York, NY, pp. 27–36.
- Slotkin, R.K. (2018) The case for not masking away repetitive DNA. *Mob. DNA*, **9**, 15.
- Sotero-Caio, C.G. et al. (2017) Evolution and diversity of transposable elements in vertebrate genomes. *Genome Biol. Evol.*, **9**, 161–177.
- Srivastava, A. et al. (2020) Alignment and mapping methodology influence transcript abundance estimation. *Genome Biol.*, **21**, 239.
- Sundaram, V. and Wysocka, J. (2020) Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. *Philos. Trans. R. Soc. B Biol. Sci.*, **375**, 20190347.
- Suzuki, H. and Kasahara, M. (2018) Introducing difference recurrence relations for faster semi-global alignment of long sequences. *BMC Bioinformatics*, **19**, 33–47.
- Teissandier, A. et al. (2019) Tools and best practices for retrotransposon analysis using high-throughput sequencing data. *Mob. DNA*, **10**, 52.
- Treangen, T.J. and Salzberg, S.L. (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, **13**, 36–46.
- Vuong, H. et al. (2018) A revisit of RSEM generative model and its EM algorithm for quantifying transcript abundances. *bioRxiv*, 503672.
- Wicker, T. et al. (2007) A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, **8**, 973–982.
- Zeng, L. et al. (2018) Transposable elements and gene expression during the evolution of amniotes. *Mob. DNA*, **9**, 17.