

Gene expression

LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq

Artem Babaian^{1,2,*}, I. Richard Thompson³, Jake Lever^{2,4},
Liane Gagnier¹, Mohammad M. Karimi^{5,*} and Dixie L. Mager^{1,2,*} 

¹Terry Fox Laboratory, BC Cancer, Vancouver, BC V5Z1L3, Canada, ²University of British Columbia, Vancouver, BC V6T1Z1, Canada, ³Qatar Biomedical Research Institute, Hamad Bin Khalifa University, Doha, PO 34110, Qatar, ⁴Michael Smith Genome Sciences Centre, BC Cancer, Vancouver, BC V5Z 4S6, Canada and ⁵MRC London Institute of Medical Sciences, Imperial College, London, W12 0NN, UK

*To whom correspondence should be addressed.

Associate Editor: Bonnie Berger

Received on August 23, 2018; revised on January 14, 2019; editorial decision on February 17, 2019; accepted on February 19, 2019

Abstract

Summary: Transposable elements (TEs) influence the evolution of novel transcriptional networks yet the specific and meaningful interpretation of how TE-derived transcriptional initiation contributes to the transcriptome has been marred by computational and methodological deficiencies. We developed *LIONS* for the analysis of RNA-seq data to specifically detect and quantify TE-initiated transcripts.

Availability and implementation: Source code, container, test data and instruction manual are freely available at www.github.com/ababaian/LIONS.

Contact: ababaian@bccrc.ca or mahdi.karimi@lms.mrc.ac.uk or dmager@bccrc.ca

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

1 Introduction

A major fraction of the human genome is composed of transposable elements (TEs), which can contain promoters, enhancers and other *cis*-regulatory sequences. TEs can be viewed as a dispersed reservoir of regulatory sequences from which transcriptional innovation arises (Chuong *et al.*, 2017; Rebollo *et al.*, 2012).

TEs near genes may gain regulatory function over evolutionary time as alternative promoters. Interestingly, during cancer evolution, normally dormant TE-promoters can be co-opted to express a proto-oncogene. Such ‘onco-exaptations’ have been identified for the expression of *CSF1R* and *IRF5* in Hodgkin lymphoma (HL), *FABP7* in diffuse large B-cell lymphoma, *IL-33* in colorectal cancer and *ALK* in melanoma, among others (reviewed in Babaian and Mager, 2016).

Previous transcriptome-wide studies designed to detect TE-derived promoters have analyzed annotated mRNAs (van de Lagemat *et al.*, 2003), expressed sequence tags (Nigumann *et al.*, 2002), assembled transcripts (Huda and Bushel, 2013; Kapusta

et al., 2013; Kelley and Rinn, 2012), short cap analysis gene expression (CAGE) tags (Faulkner *et al.*, 2009), paired-end ditag sequences (Conley *et al.*, 2008), ‘chimeric-fragment’ RNA-seq screening (Karimi *et al.*, 2011; Wang *et al.*, 2016), targeted TE-initiation events such as ERV9- (Sokol *et al.*, 2015) or L1-driven transcription (Cruikshanks and Tufarelli, 2009).

While these methods have proved insightful, they each have limitations. For instance, 5' CAGE is the clearest measure of transcription start sites but provides insufficient information on the resultant transcript structure. RNA-seq assembly methods may not identify the true 5' end of transcripts or suffer from a high false positive rate due to TE-exonization events. Moreover, none of the aforementioned studies quantify the relative contribution of the TE-initiated isoforms to overall transcript expression when alternative promoters exist. Therefore, effective TE-initiating transcript screens have required extensive human-inspection and have failed to provide a comprehensive and quantitative assessment of TEs initiating transcription.

To measure and compare the contribution of TE-promoters between transcriptomes (of the same species), we developed *LIONS* which uses paired-end RNA-seq data to detect and measure TE-initiated transcripts. *LIONS* generates a list of TE-initiated transcripts for each sequencing library. This list can be merged into a biological group, and groups can then be compared for TE-usage, such as between a set of treatment and control, or cancer and normal libraries.

2 Materials and methods

Aligned or unaligned RNA-seq data and reference annotations are input for the classification of TE-initiated transcripts (Fig. 1A). For each RNA-seq library (Supplementary Table S1), a standardized output file (.lion) of TE-initiated transcripts are generated (Supplementary Table S2), which can then be biologically grouped for comparison (i.e. cancer versus normal). *LIONS* consists of two main modules; East Lion for data initialization and classification of TE-initiation events; and West Lion for the comparison between biological groups (Fig. 1B). Detailed installation and running instructions are available in the *LIONS* user manual online (https://github.com/ababai/LIONS/blob/master/user_guide/LIONS_UG.pdf).

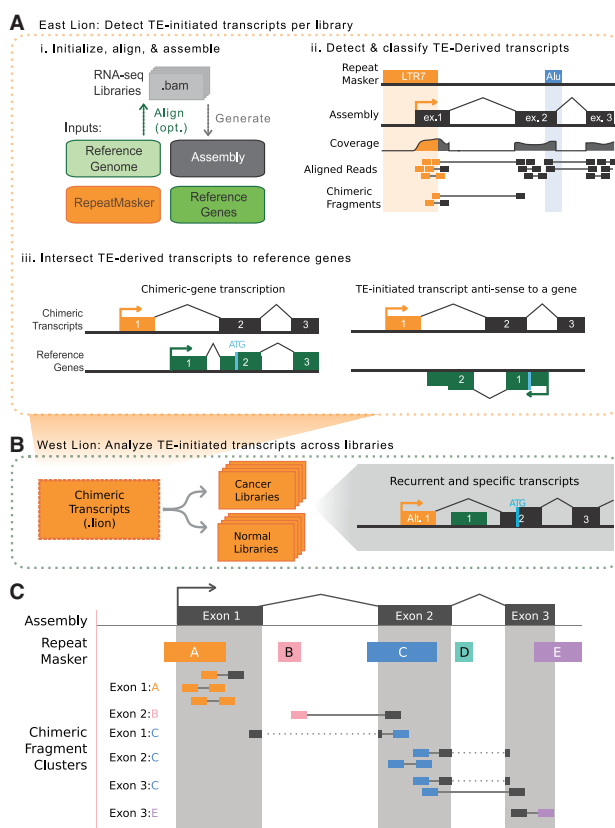


Fig. 1. The *LIONS* workflow in two modules; (A) ‘East Lion’ scripts (i) initialize data, re-align RNA-seq (optional) and assemble contigs; (ii) classify TE-initiated transcripts and (iii) cross-reference to a protein coding gene set. (B) ‘West Lion’ groups and analyses biological sets of TE-initiations detected by East Lion. (C) Chimeric fragment clusters consistent with transcriptional initiation (orange) are enriched and those with passive exonization (blue) or termination (purple) are depleted. Instances in which a TE initiates a minor isoform (pink) can be included or excluded based on the ‘TE-contribution’ parameter (see Supplementary Text for additional information) (Color version of this figure is available at *Bioinformatics* online)

3 Results

3.1 Initialization, alignment and assembly

For an accurate measurement of TE-initiated transcripts from RNA-seq, the East Lion module was developed (Fig. 1A). Input is a set of paired-end RNA sequencing data in fastq or bam format, a reference genome, a RepeatMasker table (<http://www.repeatmasker.org/>) and a reference coding gene set. The datasets can be biologically or technically grouped or analyzed individually. Optional alignment/assembly is performed with *tophat2/cufflinks* suite (with parameters optimized for TE-initiation detection) (Kim et al., 2013) and secondary alignments for multi-mapping reads are retained and flagged. On systems supporting *qsub* parallelization, each library is aligned in parallel with multiple threading allowing for rapid analysis of large datasets. An optional *Cufflinks ab initio* assembly is constructed to reduce false-positive TE-initiation calls (relative to using a reference gene set only). The alignment and assembly is then processed to generate basic statistics for each exon and TE, such as read-coverage and RPKM expression for the repeat and genic exons which are used for transcript classification.

3.2 Detection and classification of TE-initiated transcripts

To search the sequencing data for potential TE-exon interactions, each TE-exon pair for which a chimeric fragment cluster exists is considered. Briefly, a chimeric fragment cluster is a set of paired-end reads in which one read maps to a TE and its pair maps to an exon from the assembly. These TE-exon pairs form the basis for classification into one of three cases; TE-initiation, TE-exonization or TE-termination of the transcript (Fig. 1C).

Classification requires a series of values which are then processed by the classification algorithm (Supplementary Figs S1 and S2). First, the relative position of the TE and exon boundaries with respect to the direction of transcription is compared and only intersection cases which support TE-initiation are retained. A ‘thread ratio’ is calculated as the ratio of the number of read-pairs in which one read maps outside and downstream of a TE to the number read-pairs in which one read maps outside and upstream of a TE. A high thread ratio distinguishes TE-initiations from TE-exonizations, i.e. to say, if a TE initiates transcription, then there should exist a strong bias towards the number of read-pairs downstream of the element. For the detection of TE-initiated transcripts of high biological significance, further restrictions are imposed: single exon contigs (i.e. retained introns and incompletely assembled lincRNA) are excluded. To discard rare TE-initiated isoforms when a highly expressed isoform exists, TE-contribution is estimated by peak-coverage within the TE divided by the peak-coverage of its interacting exon. Together these values form the basis on which TE-initiation, TE-exonization or TE-termination can be resolved.

The parameters for the classification algorithm of TE-exon interactions can be user-customized (Supplementary Fig. S2). The default set of parameters termed ‘oncoexaptation’ were manually defined by extensive human-inspection of the training ENCODE sequencing data and cross-referenced with ChIP-seq and CAGE data (Supplementary Table S2). The default parameters are selected to specifically detect high-abundance isoforms of TE-initiated transcripts with a significant contribution to overall gene expression. These are conservative but offer the most biologically relevant results with respect to cancer biology. In addition, ‘simuOptimal’ parameter option is available which was globally optimized based on simulated RNA-seq data (Supplementary Fig. S4 and Supplementary Text).

TE-initiated transcripts can be further sub-classified by their intersection to a set of protein-coding genes into (i) chimeric transcripts: TE-initiated transcripts which transcribe in the sense-orientation into a neighbouring protein-coding gene; (ii) anti-sense TE-transcripts: non-coding TE-initiated transcripts which run anti-sense to a protein-coding gene and (iii) long intergenic non-coding TE-transcripts which do not overlap a known protein-coding gene. Of particular interest to cancer biology are chimeric transcripts that result in the overexpression of potential oncogenes. Alternative filtering settings exist and are continually added based on the experimental demand. Two such settings are ‘screen’, which is sensitive but error-prone (e.g. exonizations called as initiations), and ‘drivers’, which detects TE-initiated transcripts that are exclusively transcribed from TEs. Each of these settings is customizable and can be tailored towards individual project requirements.

These analyses are performed independently for each RNA-seq library and a standardized output .lion file is created. Sets of .lion files (i.e. sets of RNA-seq library analyses) are then grouped into a merged .lions file for set-based comparisons.

3.3 Analysis of Recurrent and Group-specific TE-initiated transcripts

To detect recurrent TE-initiated transcripts between libraries with different assemblies, the West Lion was developed. Given the set of TEs that initiate transcription, the ‘recurrence’ parameter is the number of libraries within a biological group that a given TE-initiating transcription is required to be present. In contrast, the ‘specificity’ parameter is the number of comparison (control) libraries the initiating TE is present in. Together, TEs which have greater than the recurrent cut-off and less than the specificity parameter cut-off are considered recurrent and specific TE-initiated transcripts for a group (Fig. 1B).

Grouping and comparing sets of TE-initiated transcripts are of central importance to understanding the biology of their activity. TE-initiated transcripts are more variable than non-TE-transcripts across biological replicates (Faulkner *et al.*, 2009) and therefore the signals from individual transcriptomes are noisy. It is reasonable that grouping TE-initiated transcripts across biological replicates and asking which transcripts are recurrent will enrich for TE-initiated transcripts of consequence (see review: Babaian and Mager, 2016). In a similar line of reasoning, comparing one biological group against another can identify TE-initiated transcripts, or even classes of TEs, that are more transcriptionally active in one group of transcriptomes against another (Fig. 1B).

4 Evaluation and conclusion

To evaluate the accuracy of LIONS-classified TE-initiations, a set of HL specific and recurrent (relative to B-cell controls) chimeric transcripts were assayed by RT-PCR (Supplementary Fig. S3). HL and B-cell control cell culture, RNA isolation and cDNA synthesis were performed as previously described (Babaian *et al.*, 2016). *In silico* predictions by LIONS were largely in agreement with RNA assayed by RT-PCR at 70.2 and 89.2% sensitivity and specificity, respectively (Supplementary Fig. S3). It is expected that the sensitivity of RT-PCR which reaches single-molecule sensitivity is substantially greater than RNA-seq. Altogether, LIONS is able to detect a highly specific set of TE-initiated transcripts from RNA-seq data. The

detected set is enriched for higher expressed transcripts which, in a biological context such as cancer, are expected to be more relevant to oncogenesis and warrant further biological investigation.

Acknowledgements

We thank Matt Lorincz and Rita Rebollo for helpful suggestions during the course of this work. We also thank Frances Lock for help with RNA preparation.

Funding

This work was supported by a grant from the Natural Sciences and Engineering Council of Canada (to D.L.M.). A.B. was supported by a Natural Sciences and Engineering Council of Canada Alexander Graham Bell Graduate Scholarship and a Roman Babicki Fellowship in Medical Research from the University of British Columbia.

Conflict of Interest: none declared.

References

- Babaian,A. and Mager,D.L. (2016) Endogenous retroviral promoter exaptation in human cancer. *Mob. DNA*, **7**, 24.
- Babaian,A. *et al.* (2016) Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. *Oncogene*, **35**, 2542–2546.
- Chuong,E.B. *et al.* (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.
- Conley,A.B. *et al.* (2008) Retroviral promoters in the human genome. *Bioinformatics*, **24**, 1563–1567.
- Cruikshanks,H.A. and Tufarelli,C. (2009) Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. *Genomics*, **94**, 397–406.
- Faulkner,G.J. *et al.* (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nat. Genet.*, **41**, 563–571.
- Huda,A. and Bushel,P.R. (2013) Widespread Exonization of Transposable Elements in Human Coding Sequences is Associated with Epigenetic Regulation of Transcription. *Transcr. Open Access*, **1**, 1000101.
- Kapusta,A. *et al.* (2013) Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet.*, **9**, e1003470.
- Karimi,M.M. *et al.* (2011) DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. *Cell Stem Cell*, **8**, 676–687.
- Kelley,D. and Rinn,J. (2012) Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol.*, **13**, R107.
- Kim,D. *et al.* (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Gen. Biol.*, **14**, R36.
- Nigumann,P. *et al.* (2002) Many human genes are transcribed from the antisense promoter of L1 retrotransposon. *Genomics*, **79**, 628–634.
- Rebollo,R. *et al.* (2012) Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.*, **46**, 21–42.
- Sokol,M. *et al.* (2015) Human endogenous retroviruses sustain complex and cooperative regulation of gene-containing loci and unannotated megabase-sized regions. *Retrovirology*, **12**, 32.
- van de Lagemaat,L.N. *et al.* (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.*, **19**, 530–536.
- Wang,T. *et al.* (2016) A Novel Analytical Strategy to Identify Fusion Transcripts between Repetitive Elements and Protein Coding-Exons Using RNA-Seq. *PLoS One*, **11**, e0159028.