bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

LIONS: Analysis Suite for Detecting and Quantifying Transposable Element Initiated Transcription from RNA-seq

Artem Babaian[1,2], Jake Lever[2,3], Liane Gagnier[1,2], and Dixie L. Mager[1,2]

[1]Terry Fox Laboratory, BC Cancer Agency, Vancouver, BC, Canada.

[2]Dept. of Medical Genetics, University of British Columbia, Vancouver, BC, Canada.

[3]Canada's Michael Smith Genome Sciences Centre, Vancouver, BC, Canada.

Dixie L. Mager

Terry Fox Laboratory

BC Cancer Agency

675 West 10th Avenue

Vancouver, BC, V5Z1L3, Canada

Email: dmager@bccrc.ca

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Transposable Elements (TEs), which comprise almost half of the human genome, can contribute to the evolution of novel transcriptional circuitry. Specific and biologically meaningful interpretation of TE-initiated transcripts has been marred by computational and methodological deficiencies. We developed the software suite *LIONS* (www.github.com/ababaian/LIONS) to analyze paired-end RNA-seq data for detecting and quantifying significant TE-contributions to a transcriptome. The *LIONS* suite serves as a platform for TE RNA-seq analysis and can be applied to a broad set of data sets for the study of development, stress treatments, aging and cancer.

List of Abbreviations

CAGE: Cap Analysis Gene Expression

ERV: Endogenous Retrovirus

LTR: Long Terminal Repeat

TE: Transposable Elements

TSS: Transcription Start Site

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

A major fraction of the human genome is composed of transposable elements (TEs), which along with sequences encoding transposition machinery, contain promoters, enhancers and other cis-regulatory sequences (1). Initially these regulatory sequences are necessary for TE retrotransposition, although over time they can be co-opted or exapted into host gene regulation (1–4). Even as mutations hamstring a TE's capability to mobilize, the intrinsic promoter in the elements may remain functional, such as the long terminal repeats (LTRs) flanking endogenous retroviruses (ERVs) or the 5' sense and anti-sense promoters of LINE-1 elements (5,6). Thus TEs can be viewed as a dispersed reservoir of regulatory sequences from which transcriptional innovation may arise.

The percentage of transcripts initiated within repetitive DNA as measured by Cap Analysis Gene Expression (CAGE) is substantial, ranging from ~3-15% depending on the tissue (7). These TE-initiated transcripts are enriched for long non-coding RNAs (lncRNA) (8,9). In human embryonic stem cells (hESCs), ERV transcription in particular is a marker of pluripotency in mice (10). There is also growing evidence that ERV-initiated transcripts are functionally involved in the evolution of the human stem cell transcriptome (11–14).

TEs in the vicinity of protein coding genes may gain function over evolutionary time as alternative tissue-specific promoters, like the THE1D LTR element that drives placental-specific transcription of human *IL2RB* (15). More interestingly, over the course of cancer evolution, normally dormant TE promoters can be exploited to express a protooncogene. Such "onco-exaptations" have been identified for the expression of *CSF1R* (16) and *IRF5* (17) in Hodgkin Lymphoma, *FABP7* (18) in Diffuse Large B-cell Lymphoma and *ALK* in melanoma (19) amongst others (20). While a number of cases of onco-exaptations have been documented, the mechanisms underlying these oncogenic events remains largely unexplored.

It has been proposed that TE invasions can function as evolutionary accelerants, promoting

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

adaptation and correlating with the radiation of species (21,22) and therefore there is a significant interest in understanding the extent and evolutionary mechanisms by which TEs contribute to a cell's transcriptome. Previous transcriptome-wide studies designed to detect TE-derived promoters have analyzed annotated mRNAs (23), ESTs (24), assembled transcripts (8,9,25), short Cap Analysis Gene Expression CAGE tags (7), Paired-end ditag sequences (26), paired-end 'chimeric fragment' RNA-seq screening (18,27,28), targeted TE events such as ERV9-driven (29) or L1-driven transcripts (30) and loci-gene correlation studies (31). While these methods have proved useful, they have significant limitations.

5' CAGE is the clearest measure of transcription start sites (TSSs) 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. The TE-exonization problem also creates high false-positive rates in chimeric fragment-based and hybridization-based methods that have gone unaddressed (27–29,32). Moreover, none of the aforementioned studies have attempted to quantify the strength or contribution of the putative 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 quantitative, genome-wide assessment of TEs initiating biologically significant transcription.

To quantitatively measure and compare the contribution of TE promoters to normal and cancer transcriptomes we developed a tool that incorporates features of previous methods but significantly builds upon them. We were motivated to use paired-end RNA-seq data alone, a broadly available data-type, to rapidly measure TE-initiations and transcriptome contributions. With a defined set of TE-initiated transcripts in each library, commonalities and differences between sets of data (biological replicates) can be determined. Together these analyses have been packaged to give rise to the *LIONS* suite (Figure 1).

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

To quantify the contribution of TE promoters to the transcriptome from RNA-seq data alone, we were motivated to develop the *LIONS* analysis suite. Briefly, RNA-seq data along with a reference genome, gene and repeat annotation are inputs for the classification and annotation of TE-initiated transcripts (Figure 1A). For each RNA-seq library, a standard (.lion) file of TE-initiated transcripts is the output, and then can be grouped into biological categories such as cancer versus normal controls, for comparison (Figure 1B). A detailed outline of the analysis is provided in materials and methods.

TEs intersect exons in three main categories; as initiations at the 5' end of a transcript; as exonizations either with or without being involved at a splice junction; and at the 3' end as a termination site for transcripts (Figure 2A). The core *LIONS* classification segregates the initiations from non-initiation events. This is biologically pertinent in the analysis of TE transcription since non-initiation events outnumber initiation events by three orders of magnitude (Figure 2B). Thus analyses based on chimeric read clusters alone, or TE-transcription levels alone do not necessarily reflect autonomous transcriptional activity of TEs but rather simply correlation or propensity to be transcribed as part of other transcripts. This is non-trivial as TEs have long been known to be enriched at 5' and 3' untranslated transcribed regions (UTRs) and within long-noncoding (lnc)RNAs (8,9).

To test the operating characteristics of *LIONS*, RNA-seq reads based on the ENCODE (33) K562 and H1 embryonic stem cell line transcriptomes were simulated at varying depths as a benchmark. Simulated TE-exon pair clustering of reads plateaus at ~52% sensitivity regardless of further increase in sequencing depth (Figure 3A). This plateau emphasizes the systemic difficulty of accurately determining either 5' or 3' ends of transcripts from RNA-seq data alone, but the undetected TE start sites correlate with lower overall expression (Figure 3B). TE promoter analysis is confounded by the basic biological properties of TE TSSs, they are weaker and more biologically irreproducible (have higher cell-cell variation) than their non-TE TSS counterparts in CAGE analyses (Supplementary

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1). From the fraction of TE TSSs which are measurable by chimeric fragments, the default *LIONS* parameters have a sensitivity of 36.35% and specificity of 98.63% (Figure 3C). The relative proportion of each class of TE TSS called largely matches the proportions of TE TSSs of the input transcriptomes, which rules out a systematic bias towards any one class of TE (Figure 3D). Altogether, while the set of TEs read-out by *LIONS* is not highly sensitive especially for lower expressed transcripts, it is highly specific and accurately reflects the underlying promoter activity of TEs.

To evaluate the accuracy of *LIONS*-classified TE-initiations, a set of Hodgkin lymphoma-specific and recurrent (relative to B-cell controls) chimeric transcripts were assayed by RT-PCR. *In silico* predictions were largely in agreement with RNA assayed by RT-PCR at 55.4% and 89.2% sensitivity and specificity respectively (Supplementary Figure 2).

Altogether *LIONS* is able to detect a specific set of TE-initiated transcripts from RNA-seq data alone. The detected set is enriched for higher expressed transcripts which, in a biological context such as cancer, are expected to be more relevant than the low expression / high variation TE-initiated transcripts.

Discussion

The preceding principles of local RNA sequencing analysis to distinguish TE-derived transcription initiation from exonization or termination can also be seen as a specific-case of the *ab initio* RNA-seq assembly problem. Local calculations used in *LIONS*, namely read threading and upstream coverage could be generalized to the entire transcriptome. Further refinement of these methods such as inclusion of aligned-strand bias measures (34), position-aware Hidden Markov Model or machine-learning trained sorting algorithms to detect the molecular signature of TSSs could be used to improve the accuracy of transcript assembly.

LIONS suite is limited in similar ways as other assembly methods are, namely in regions of high transcriptional complexity, especially if non-stranded data is used and there is bi-directional

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

transcription. The coverage around all transcript ends in RNA-seq is reduced relative to interior sequences (34) and confounded by lower overall expression and higher variability of expression of TE TSSs in general (7).

The focus of the *LIONS* suite on transcriptional initiation is the low-hanging fruit for TE-gene interactions. Additional analysis of chimeric read clusters may quickly yield TE sets which are incorporated into transcripts, such as TE-derived splice acceptors and donors in the newly classified characterized exitrons (35). Anecdotally, one of the largest difficulties in developing *LIONS* was distinguishing the true initiation events from exitron-like events that occur within a TE. This distinction is also one of the greatest limitations of previous studies looking at TE-derived transcriptomes (27,29,32), which did not make this distinction.

The source code for all *LIONS* components is available at www.github.com/ababaian/LIONS and all analyses are based on a standard .lions output file. Standardization was performed to encourage users to share down-stream analysis scripts such that graphs and statistics of TEs could be reproducible and applied to different data sets quickly.

Materials & Methods

Initialization, Alignment and Assembly

For an accurate measurement of TE initiated transcripts starting from whole transcriptome sequencing data the *East Lion* was developed (Figure 1A). The central principle in detecting transcription start sites within TEs is that a local analysis is performed to search for patterns of sequencing reads consistent with transcriptional initiation.

The primary *LIONS* input is a set of paired-end RNA sequencing data either in fastq or bam format. The datasets can be biologically or technically grouped for later comparisons or individual libraries can be run. Additionally a reference genome (hg19), a RepeatMasker (36) analysis of that genome (hg19 – 2009-04-24), and a set of reference protein-coding genes (UCSC Genes, 2013-06-14)

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

is required.

A workspace for the project is initialized on the system and an alignment is run with the splice-aware aligner tophat2 (v.2.0.13) (37) such that secondary alignments for multi-mapping reads are retained and flagged; *tophat2 --report-secondary-alignments.* On systems that support *qsub* parallelization and multiple CPU cores, each library is aligned in parallel with multiple threading allowing for rapid analysis of large datasets.

Following alignment, *ab initio* transcriptome assembly is performed on each library using repeat-optimized parameters of Cufflinks (v.2.2.1) (38); *cufflinks --min-frags-per-transfrag 10 --max-multiread-fraction 0.99 --trim-3-avgcov-thresh 5 –trim-3-dropoff-frac=0.1 --overlap-radius 50.* The use of an assembly substantially reduced false-positive TE-initiation calls relative to using a reference gene set since only transcript isoforms which exist in the data are considered, although it is possible to forego this step and use a reference gene set. The generated alignment and assembly is then processed to generate a bigwig coverage file for visualization and basic statistics for each exon and TE are calculated such as read-coverage and RPKM.

Chimeric Fragment Cluster Analysis

To search the sequencing data for potential TE-exon interactions, each TE-exon pair for which a chimeric fragment cluster exists are considered. Briefly, a chimeric fragment cluster is a set of reads in which one read maps to a TE and its pair maps to an exon from the assembly (Figure 2). These TE-exon pairs form the basis for classification into one of three cases; TE-initiation, -exonization or -termination of the transcript (Figure 2).

Classification requires the calculation of a series of values that are then fed into a classification algorithm. First, the relative position of the TE and exon boundaries with respect to the direction of transcription is compared. Only intersection cases in which the TE could initiate transcription are considered (Supplementary Figure 3A). A thread ratio is then calculated, the ratio of read pairs in which

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

one read maps outside of a TE in either the downstream or upstream direction. A high thread ratio distinguishes TE-initiation events from TE-exonizations, that is to say, if a TE initiates transcription then there should exist a strong bias towards the number of read-pairs downstream of the element (Supplementary Figure 3B).

For the detection of TE-initiated transcripts of most biological significance further restrictions are imposed. Single exon contigs are excluded from the analysis to reduce the false positive rate (retained introns, low abundance lincRNA). To quickly discard rare TE-initiated isoforms when an alternative, highly expressed isoform exists, TE contribution was estimated as the peak coverage within the TE divided by the peak coverage of it's interacting exon (Supplementary Figure 3C). Together these values form the basis on which TE-initiation, -exonization or -termination can be distinguished.

Classification of TE-exon interactions is performed by the sorting algorithm that can be customized (Supplementary Figure 4). The default set of parameters termed, '*oncoexaptation'* were manually defined by extensive manual inspection of the training ENCODE sequencing data and comparison with supporting ChIP-seq and CAGE data (Supplementary Figure 5). The default parameters are trained 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.

TE-initiated transcripts can be further sub-classified by their intersection to a set of protein-coding genes into; chimeric transcripts, TE-initiated transcripts which transcribe in the sense-orientation into a neighbouring protein-coding gene; anti-sense TE-transcripts, non-coding TE-initiated transcripts which run anti-sense to a protein-coding gene; or long intergenic non-coding (linc) TE-transcripts which don't overlap a known protein-coding gene. Of particular interest to cancer biology are chimeric transcripts that result in the overexpression of oncogenes, such as previously identified in Hodgkin Lymphoma for *IRF5* and *CSF1R* (16,17).

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Alternative filtering settings exist and are continually added based on the experimental demand such as; '*screen'*, a sensitive but error-prone (exonizations called as initiations) method or; '*drivers',* detection of TE-initiated transcripts which are exclusively transcribed from TEs. Each of these settings are customizable and should be tailored towards individual project requirements.

These analyses and filters are applied independent for each RNA-seq library and a standard .lion file is created. Sets of .lion files (that is sets of RNA-seq library analyses) are then grouped into a merged .lions file for set-based comparisons.

Operating Characteristics

To test the performance of the *LIONS* classification, a simulation of RNA-seq data was generated to benchmark the operating characteristics of the classifier. Starting with aligned RNA-seq from H1 embryonic stem cells and K562 chronic myeloid leukemia cell line, simulated transcriptomes were generated. First the top 20,000 expressed transcripts in *gencode (v19)* in K562 and the top 20,000 expressed contigs from the H1esc assembly were selected as the basis for the simulated transcriptome. *FluxSimulator (39)* was then used to generate paired-end fastq at 5, 30, 100 and 200 millions reads in H1esc and 25, 100 and 200 million reads in K562 using a reference *hg19* sequence. The simulated data are then processed by *LIONS* and compared to the expected input transcriptomes are defined as a 'ground truth'.

Recurrent and Group-specific TE-promoters

Grouping and comparing sets of TE-initiated transcripts is of central importance to understanding the biology of their activity. TE-initiated transcripts are more variable then non-TE transcripts across biological replicates (Supplementary Figure 1) and therefore the signals from individual transcriptomes are noisy. The reasoning then is that grouping TE-initiated transcripts across biological replicates and asking which transcripts are recurrent will enrich for TE-initiated transcripts of consequence. In a similar line of reasoning, comparing one biological group against another can

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

identify TE-initiated transcripts, or even classes of TEs that are more transcriptionally active in one grouping of transcriptomes against another.

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

A clear case in which recurrent and biological-group specific TE-initiated transcripts is significant is in cancer biology. The onco-exaptation hypothesis (20) predicts that the highly variable TE-initiated transcripts can be selected for during cancer evolution and therefore transcripts recurrent and cancer-specific are enriched for oncogenes or transcripts involved in the biology of the cancer.

RNA-seq Data Sets

ENCODE training RNA-seq fastq files was downloaded from the UCSC ENCODE ftp site. Hodgkin Lymphoma cell line and primary B-cell transcriptomes (17,40–43) bam files were converted to fastq for re-analysis by *LIONS*. Accession and library details are in Supplementary Table 1.

Hodgkin lymphoma cell line culture and RT-PCR.

Hodgkin Lymphoma cell culture, RNA isolation and cDNA synthesis was performed as previously described (17). Primers for RT-PCR are listed in Supplementary Table 2.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

1. Rebollo R, Romanish MT, Mager DL. Transposable elements: an abundant and natural source of regulatory sequences for host genes. Annu Rev Genet. 2012;46:21–42.

2. Rebollo R, Farivar S, Mager DL. C-GATE - catalogue of genes affected by transposable elements. Mob DNA. 2012 May 23;3(1):9.

3. Feschotte C. Transposable elements and the evolution of regulatory networks. Nat Rev Genet. 2008 May;9(5):397–405.

4. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 2017 Feb;18(2):71–86.

5. Jurka J, Kapitonov VV, Kohany O, Jurka MV. Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet. 2007;8:241–59.

6. Richardson SR, Doucet AJ, Kopera HC, Moldovan JB, Garcia-Perez JL, Moran JV. The Influence of LINE-1 and SINE Retrotransposons on Mammalian Genomes. Microbiol Spectr. 2015 Apr;3(2):MDNA3-0061-2014.

7. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009 May;41(5):563–71.

8. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13(11):R107.

9. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 2013 Apr 25;9(4):e1003470.

10. Macfarlan TS, Gifford WD, Driscoll S, Lettieri K, Rowe HM, Bonanomi D, et al. Embryonic stem cell potency fluctuates with endogenous retrovirus activity. Nature. 2012 Jul 5;487(7405):57–63.

11. Gerdes P, Richardson SR, Mager DL, Faulkner GJ. Transposable elements in the mammalian embryo: pioneers surviving through stealth and service. Genome Biol. 2016 May 9;17:100.

12. Lu X, Sachs F, Ramsay L, Jacques P-É, Göke J, Bourque G, et al. The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. Nat Struct Mol Biol. 2014 Apr;21(4):423–5.

13. Wang J, Xie G, Singh M, Ghanbarian AT, Raskó T, Szvetnik A, et al. Primate-specific endogenous retrovirus-driven transcription defines naive-like stem cells. Nature. 2014 Dec 18;516(7531):405–9.

14. Ramsay L, Marchetto MC, Caron M, Chen S-H, Busche S, Kwan T, et al. Conserved expression of transposon-derived non-coding transcripts in primate stem cells. BMC Genomics. 2017 Feb 28;18(1):214.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

15. Cohen CJ, Rebollo R, Babovic S, Dai EL, Robinson WP, Mager DL. Placenta-specific Expression of the Interleukin-2 (IL-2) Receptor β Subunit from an Endogenous Retroviral Promoter. J Biol Chem. 2011 Oct 14;286(41):35543–52.

16. Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med. 2010 May;16(5):571–579, 1p following 579.

17. Babaian A, Romanish MT, Gagnier L, Kuo LY, Karimi MM, Steidl C, et al. Onco-exaptation of an endogenous retroviral LTR drives IRF5 expression in Hodgkin lymphoma. Oncogene. 2016 May 12;35(19):2542–6.

18. Lock FE, Rebollo R, Miceli-Royer K, Gagnier L, Kuah S, Babaian A, et al. Distinct isoform of FABP7 revealed by screening for retroelement-activated genes in diffuse large B-cell lymphoma. Proc Natl Acad Sci U S A. 2014 Aug 26;111(34):E3534-3543.

19. Wiesner T, Lee W, Obenauf AC, Ran L, Murali R, Zhang QF, et al. Alternative transcription initiation leads to expression of a novel ALK isoform in cancer. Nature. 2015 Oct 15;526(7573):453–7.

20. Babaian A, Mager DL. Endogenous retroviral promoter exaptation in human cancer. Mob DNA. 2016;7:24.

21. Jurka J, Bao W, Kojima KK. Families of transposable elements, population structure and the origin of species. Biol Direct. 2011;6:44.

22. Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014 Sep 11;513(7517):195–201.

23. van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet TIG. 2003 Oct;19(10):530–6.

24. Nigumann P, Redik K, Mätlik K, Speek M. Many human genes are transcribed from the antisense promoter of L1 retrotransposon. Genomics. 2002 May;79(5):628–34.

25. Huda A, Bushel PR. Widespread Exonization of Transposable Elements in Human Coding Sequences is Associated with Epigenetic Regulation of Transcription. Transcr Open Access. 2013 Jun 19;1(1).

26. Conley AB, Piriyapongsa J, Jordan IK. Retroviral promoters in the human genome. Bioinforma Oxf Engl. 2008 Jul 15;24(14):1563–7.

27. Karimi MM, Goyal P, Maksakova IA, Bilenky M, Leung D, Tang JX, et al. DNA methylation and SETDB1/H3K9me3 regulate predominantly distinct sets of genes, retroelements, and chimeric transcripts in mESCs. Cell Stem Cell. 2011 Jun 3;8(6):676–87.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

28. Wang T, Santos JH, Feng J, Fargo DC, Shen L, Riadi G, et al. A Novel Analytical Strategy to Identify Fusion Transcripts between Repetitive Elements and Protein Coding-Exons Using RNA-Seq. PLOS ONE. 2016 Jul 14;11(7):e0159028.

29. Sokol M, Jessen KM, Pedersen FS. Human endogenous retroviruses sustain complex and cooperative regulation of gene-containing loci and unannotated megabase-sized regions. Retrovirology. 2015;12:32.

30. Cruickshanks HA, Tufarelli C. Isolation of cancer-specific chimeric transcripts induced by hypomethylation of the LINE-1 antisense promoter. Genomics. 2009 Dec;94(6):397–406.

31. Pavlicev M, Hiratsuka K, Swaggart KA, Dunn C, Muglia L. Detecting endogenous retrovirus-driven tissue-specific gene transcription. Genome Biol Evol. 2015 Apr;7(4):1082–97.

32. Pérot P, Mugnier N, Montgiraud C, Gimenez J, Jaillard M, Bonnaud B, et al. Microarray-based sketches of the HERV transcriptome landscape. PloS One. 2012;7(6):e40194.

33. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57–74.

34. Hillier LW, Reinke V, Green P, Hirst M, Marra MA, Waterston RH. Massively parallel sequencing of the polyadenylated transcriptome of C. elegans. Genome Res. 2009 Apr;19(4):657–66.

35. Staiger D, Simpson GG. Enter exitrons. Genome Biol. 2015 Jul 7;16(1):136.

36. RepeatMasker Home Page [Internet]. [cited 2015 Jul 22]. Available from: http://www.repeatmasker.org/

37. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36.

38. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc. 2012 Mar;7(3):562–78.

39. Griebel T, Zacher B, Ribeca P, Raineri E, Lacroix V, Guigó R, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012 Nov;40(20):10073–83.

40. Morin RD, Mendez-Lago M, Mungall AJ, Goya R, Mungall KL, Corbett R, et al. Frequent mutation of histone modifying genes in non-Hodgkin lymphoma. Nature. 2011 Jul 27;476(7360):298–303.

41. Steidl C, Shah SP, Woolcock BW, Rui L, Kawahara M, Farinha P, et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature. 2011 Mar 17;471(7338):377–81.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

42. Liu Y, Abdul Razak FR, Terpstra M, Chan FC, Saber A, Nijland M, et al. The mutational landscape of Hodgkin lymphoma cell lines determined by whole-exome sequencing. Leukemia. 2014 Nov;28(11):2248–51.

43. Twa DDW, Chan FC, Ben-Neriah S, Woolcock BW, Mottok A, Tan KL, et al. Genomic rearrangements involving programmed death ligands are recurrent in primary mediastinal large B-cell lymphoma. Blood. 2014 Mar 27;123(13):2062–5.

44. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511–5.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1: Schematic of the *LIONS* workflow.

The workflow for *LIONS* is divided into two main components. **A)** 'East Lion' analyzes individual transcriptomes starting with i) a .bam file(s) of paired-end reads, a reference genome, a RepeatMasker annotation and a reference set of protein coding genes. The reads are aligned to the genome with the spliced read mapper Tophat2 (37) and an *ab initio* transcriptome is assembled with Cufflinks (44). ii) These data are then analyzed per chimeric fragment cluster for transposable element (TE)-initiated transcripts (Figure 2). Briefly, fragment clusters consistent with transcriptional initiation (Orange) are enriched and those with passive exonization (Blue) or termination (not shown) are depleted. iii) The set of TE-initiated contigs are then intersected to reference set of protein coding genes and classified with respect to their intersection. Each transcriptome is analyzed independently and a standard .lions output file is generated. **B)** 'West Lion' performs set analysis on the .lions files. Transcriptomes are biologically grouped and analyzed individually and as part of a biological group (i.e. cancer vs. normal samples).

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2: Chimeric Fragment Clustering in *LIONS*.

A) The analysis space of *LIONS* consists of all Repeat-Exon combinations for which there exists a chimeric fragment; i.e. paired-end sequencing reads in which one read intersects a repeat and the read pair intersects an exon from the assembly. Chimeric fragments can result from an RNA molecule in three cases; i) TE-initiated transcripts (Repeat A:Exon 1 and Repeat B:Exon 2); ii) TE exonization in a transcript, either as a repeat contained within an exon or the repeat provides an exon splice site (Repeat C:Exon 1,2 and 3) or iii) TE terminated transcripts (Repeat E:Exon 3). Each chimeric fragment cluster is then classified by *LIONS* as either initiating or non-initiating with respect to a transcript. B) The number of chimeric fragments in K562, H1 or GM12878 transcripts that are classified as initiations compared to non-initiating clusters.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3: *LIONS* operating characteristics on simulated data.

Simulated RNA-seq data based on a reference H1 ESC (green) and K562 (blue) transcriptomes were used as a benchmark to test the sensitivity and specificity of the *LIONS* suite. A) In RNA-seq libraries simulated to varying depth, chimeric fragment clusters are limited in their capacity to detect TE-derived transcript start sites (TSSs), plateauing at ~52% sensitivity. B) The TE-TSSs which are detectable by chimeric fragment clusters (+) are more highly expressed (Welch T-test, p = 4.59e-8) than those for which chimeric fragment clusters are not detected (-). C) From the chimeric fragment cluster detectable TE-TSSs, default parameter *LIONS* has a 36.36% sensitivity and 98.63% specificity yielding a specific set of TE-TSSs. D) The relative proportion of *LIONS* called TE-initiated transcripts from each TE-class for each simulated data-sets at varying simulation depths, relative to their respective input transcriptome TE-class proportions (teal line).

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 1: Reproducibility of transposable element transcription start sites by CAGE.

5' Cap analysis gene expression (CAGE) transcription start site clusters were downloaded from the UCSC genome browser for GM12878 polyadenylated whole cell RNA (UCSC accession: wgEncodeEH001680). The center of each transcription start site (TSS) cluster was intersected against RepeatMasker to distinguish non-TE TSS (blue) and TE-TSS (orange). A) To test if TE-derived TSSs are more or less variable between biological replicates the irreproducible discovery rate (IDR) between the groups was compared. TE-derived TSSs are more variable between biological replicates (Welch's t-test, $p < 2.2e-16$) then non-TE TSSs. Reproducible clusters are those that pass an IDR cut-off of <0.05 (right of red line). B) Amongst the reproducible CAGE clusters, TE-derived TSSs have a lower (Welch's t-test, $p < 2.2e-16$) expression level by log fragments per kilobase per million mapped reads (FPKM). **C)** The TE-TSS clusters can further be striated by TE-class. Violin plot of the kernel density of the log(FPKM) is shown for each class overlaid with a bar graph of the count per TE-TSS.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 2: Reverse-transcription PCR validation of candidate TE-initiated transcripts.

From the Hodgkin Lymphoma (HL) RNA-seq datasets, TE-initiated transcripts with predicted intact coding sequences that occurred in at least 2/12 HL libraries and were absent from all nine primary B-cell libraries were selected as Hodgkin-specific and recurrent. Candidate genes were selected with potential involvement in cancer pathogenesis by a literature review. The TE-initiated isoforms were validated by reverse-transcription (RT-)PCR and compared to the *in silico* prediction from *LIONS.* The normal B-cell lines T2 and T3-1a were used as controls to test for HL specificity. The concordance between RT-PCR and *LIONS* was used as a measure of the sensitivity, 55.41% and specificity, 89.2% of the software.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 3: Calculated values for *LIONS* classification.

To distinguish TE-initiated transcripts from TE exonizations or TE-terminated transcripts several local values are calculated for each chimeric fragment cluster. A) The position of the TE (orange) relative to the exon (dark gray). Cases in which the TE is upstream, on the upstream edge, or contains the exon are considered for TE-initiation (highlighted green). B) The thread ratio for a TE considers direction bias in sequencing read pairs going upstream or downstream relative to the interacting exon. Upstream threads (red) are read pairs in which one read maps to within the TE and the pair maps upstream of the TE. Downstream threads (blue) are the converse to upstream threads while read pairs with both reads internal to the TE are not counted (gray). The thread ratio is the number of downstream threads divided by the number of upstream threads, or set to the cut-off threshold when no upstream threads are present for inclusion. C) The contribution score is an approximation of the TE promoter contribution to the expression of downstream exons for alternative or unassembled TE promoter. The maximum coverage within the TE, 28 reads, is divided by the maximum coverage within the interacting exon (exon 2), 44 reads, to yield an approximate contribution for the TE-exon interaction, 0.636. D) The read coverage for the 50 bp immediately upstream of the TE is divided by the coverage of the TE itself to measure the background level of transcription at this loci. i. A Locus with low levels of transcriptional read through but a potential initiation site present within the TE. ii. In contrast, a locus in which there is an apparent gain of coverage within the LINE element but could be due to poor mapping quality at the 5' end of this LINE element. E) Chimeric fragment sub-classification of whether a read intersects only a repeat (R), only an exon (E) or both (D). Chimeric fragments can thus be classified as DR, DD,DE or ER fragments. The ratio between the classifications can be used as a stringency cutoff for improving *LIONS* classification specificity. Taken together these values form the basis for *LIONS* classification of TE-initiated transcripts and are fed into the the sorting algorithms (Supplementary Figure 3).

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 4: Chimeric fragment clusters sorting algorithm for TE-initiated Transcripts.

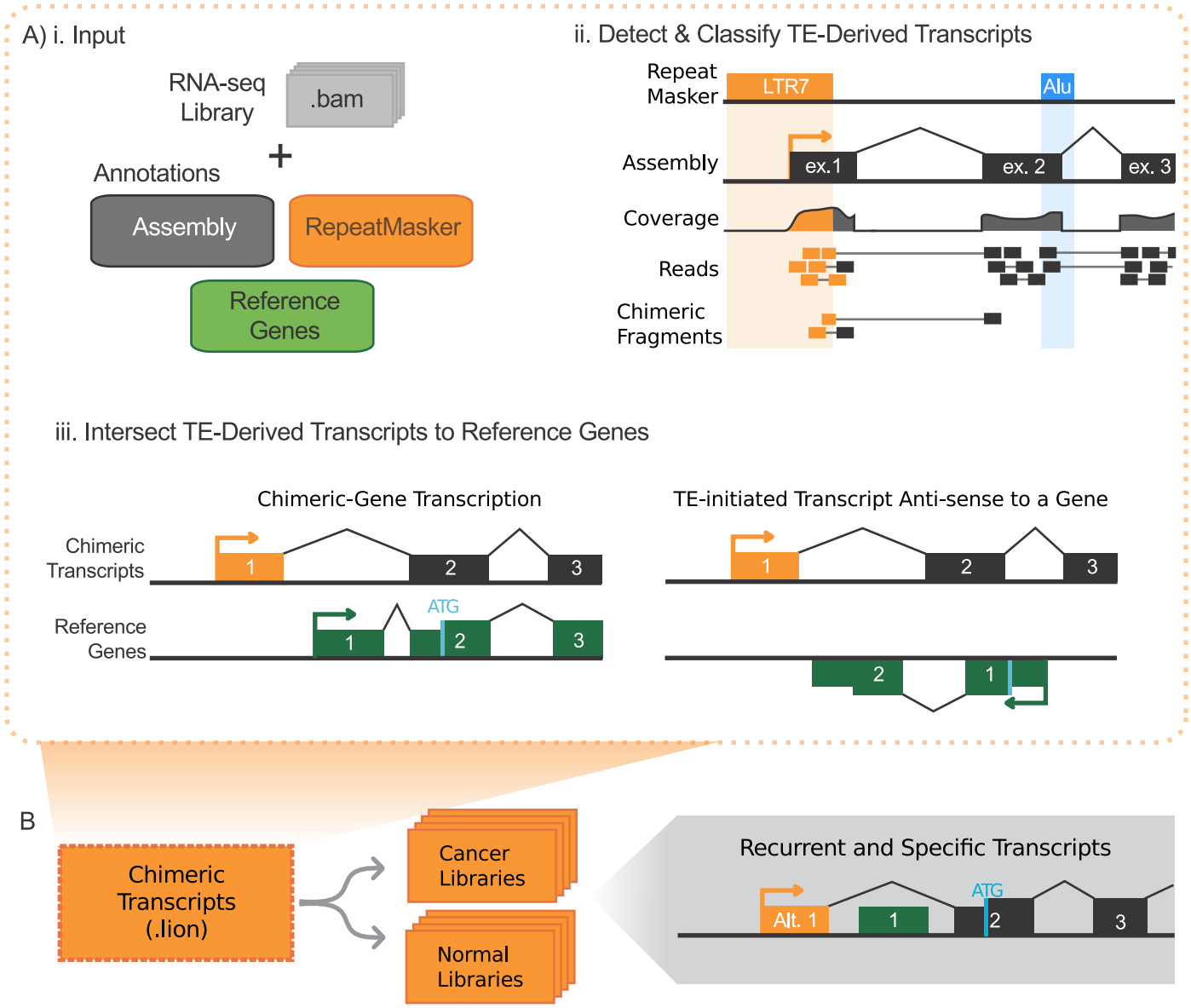
For each TE-Exon pair for which there exists chimeric fragment support, *LIONS* sorts TE-initiated transcripts from TE exonizations or TE-terminated transcripts. Default parameters are shown but can be manually changed by the user. The number of supporting chimeric fragments can be set as a function of the number of mapped reads (i.e. the greater of 3 fragments or 1 fragment / 20 million mapped reads) to fairly compare libraries of varying sequencing depth. Chimeric fragment clusters are filtered for those in which the assembled transcript has greater than one exon, then striated by the TE-exon intersection and then sub-sorted with different cut-offs; number of chimeric fragments, thread ratios for the TE, contribution of the TE-promoter to overall transcript levels, read coverage upstream of the repeat and the exons number within the assembled transcript. Separate cut-offs are necessary for different intersection cases to deplete the variety of non-initiation cases that arise. Final output of the sorting algorithm is the standard .lions file.

bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 5: UCSC genome browser view of a *LIONS* identified chimeric transcript in K562. An upstream MLT1K LTR element initiates transcription and splices into exon 2 of the *FHAD1* gene in which the coding sequence begins. The *Cufflinks* assembly contigs as well as the aligned reads and *tophat2* detected splice junctions are shown, this case would be classified as 'EInside' as the entire first exon is contained within the MLT1K LTR element. CAGE hidden Markov model clusters (UCSC accessions: whole cell wgEncodeEH001150; cytosol wgEncodeEH000332; nucleus wgEncodeEH000333), DNase-seq (wgEncodeEH000530) and ChIP-seq (H3K4me1 wgEncodeEH000046; H3K4me3 wgEncodeEH000048; H3K27me3 wgEncodeEH000044) coverage support that this is a promoter as well as being classified as a 'weak promoter' by the respective Broad ChromHMM model (wgEncodeEH000790).

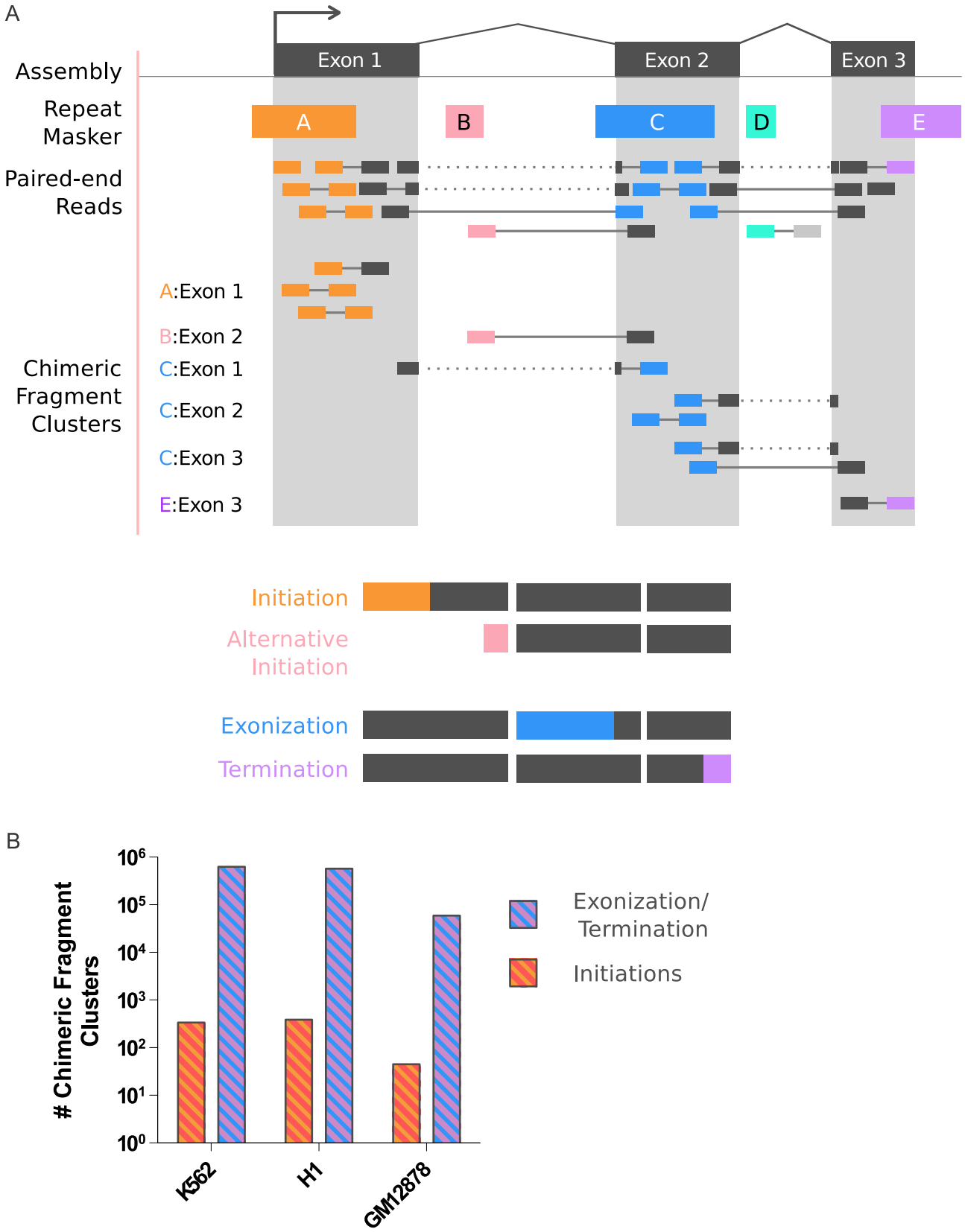
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 1



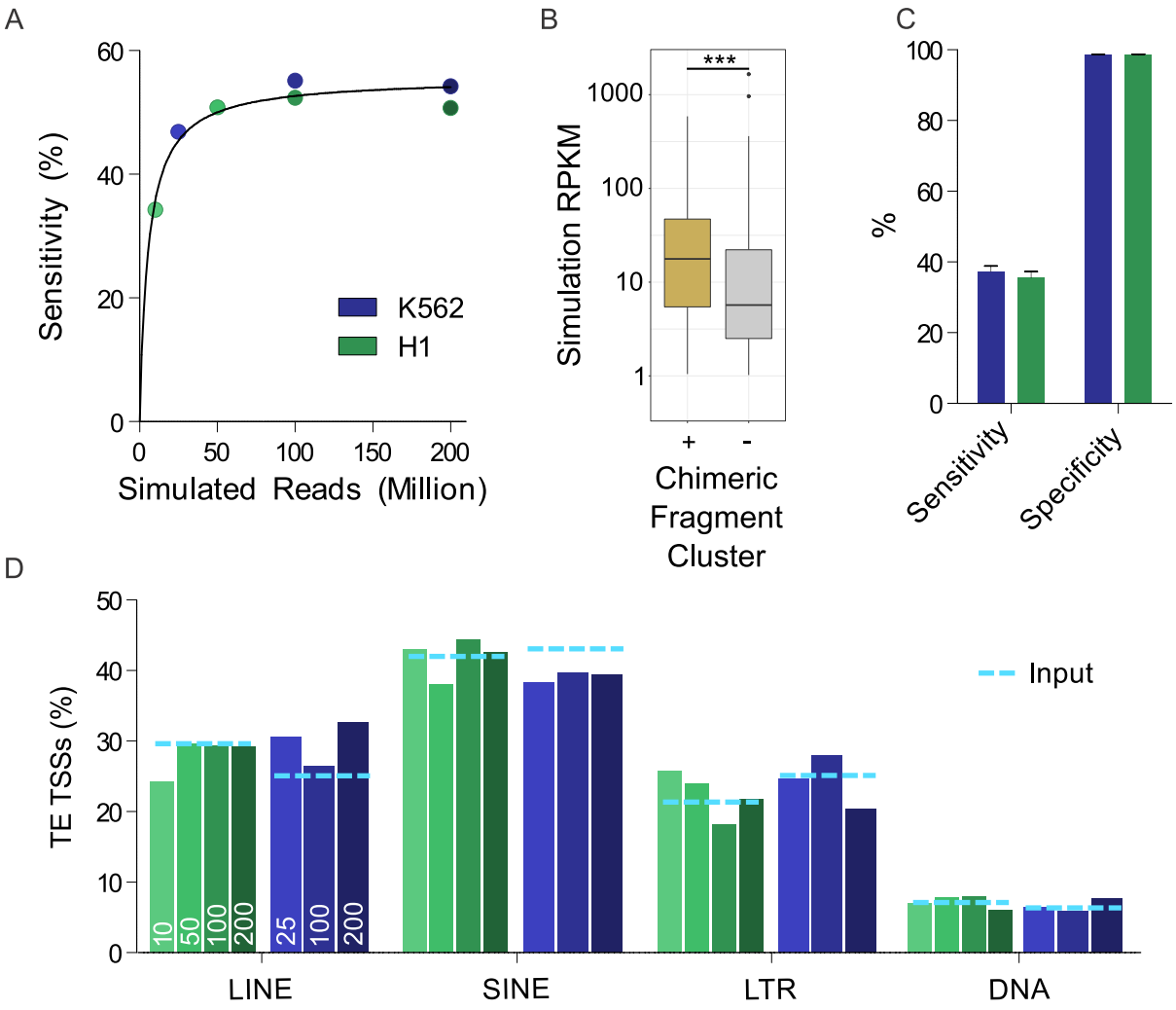
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 2



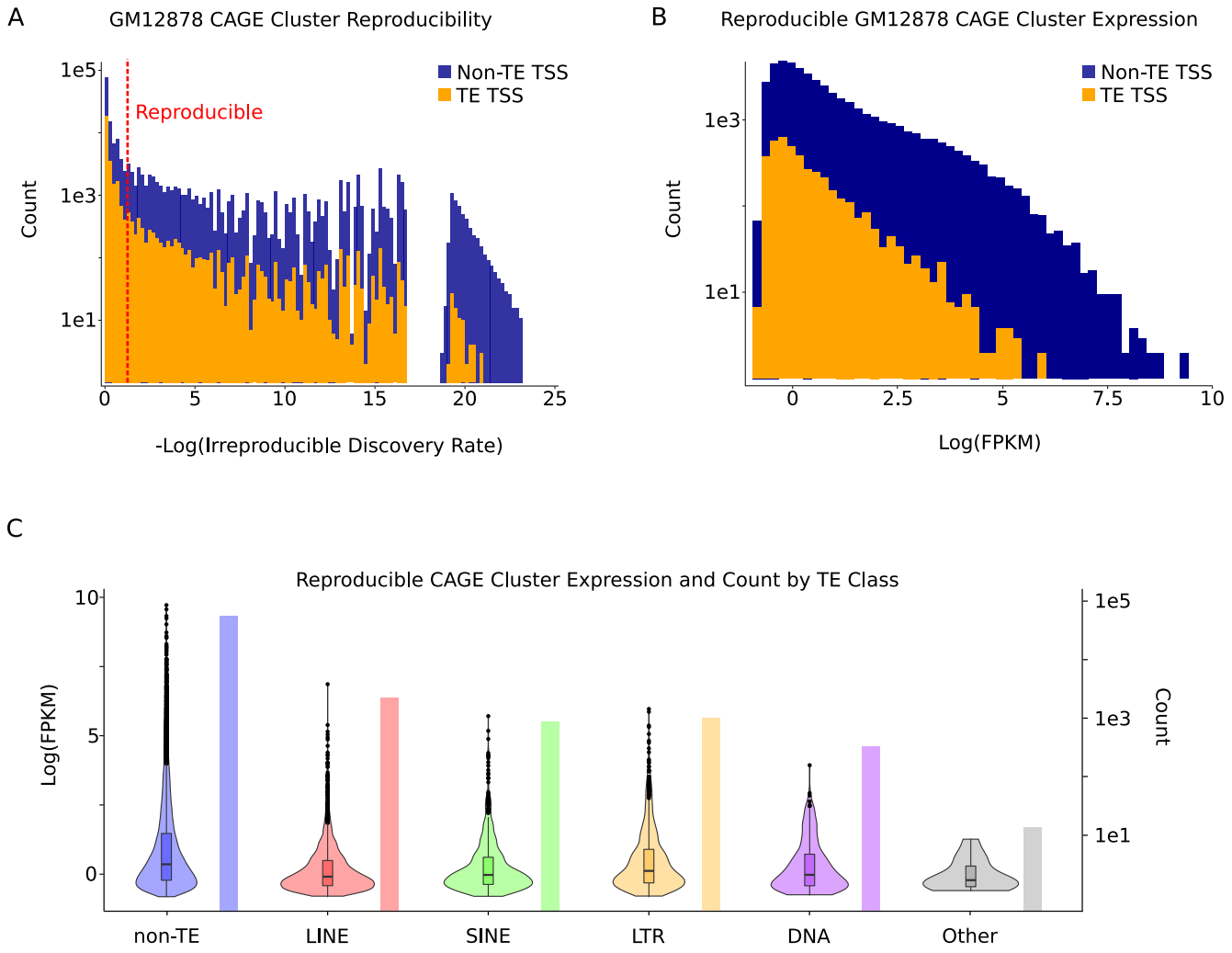
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Figure 3



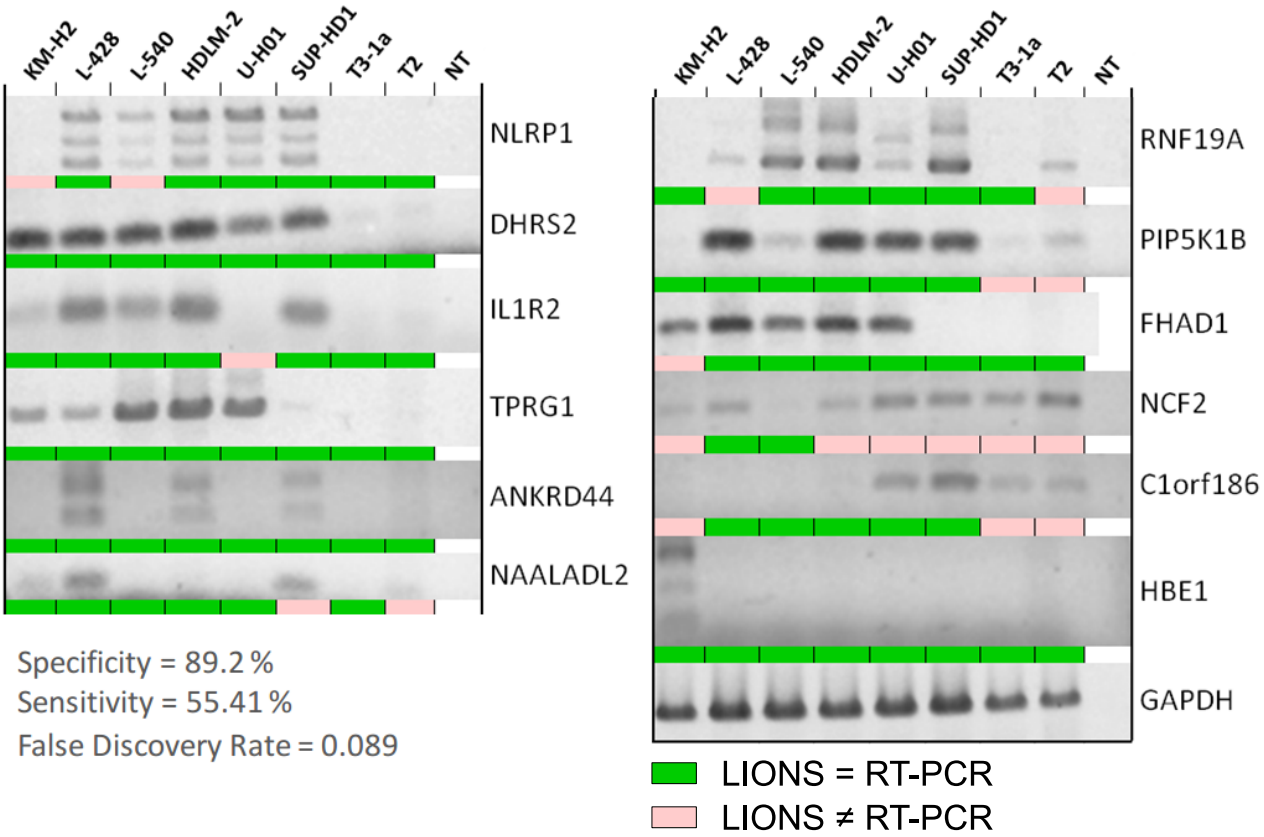
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 1



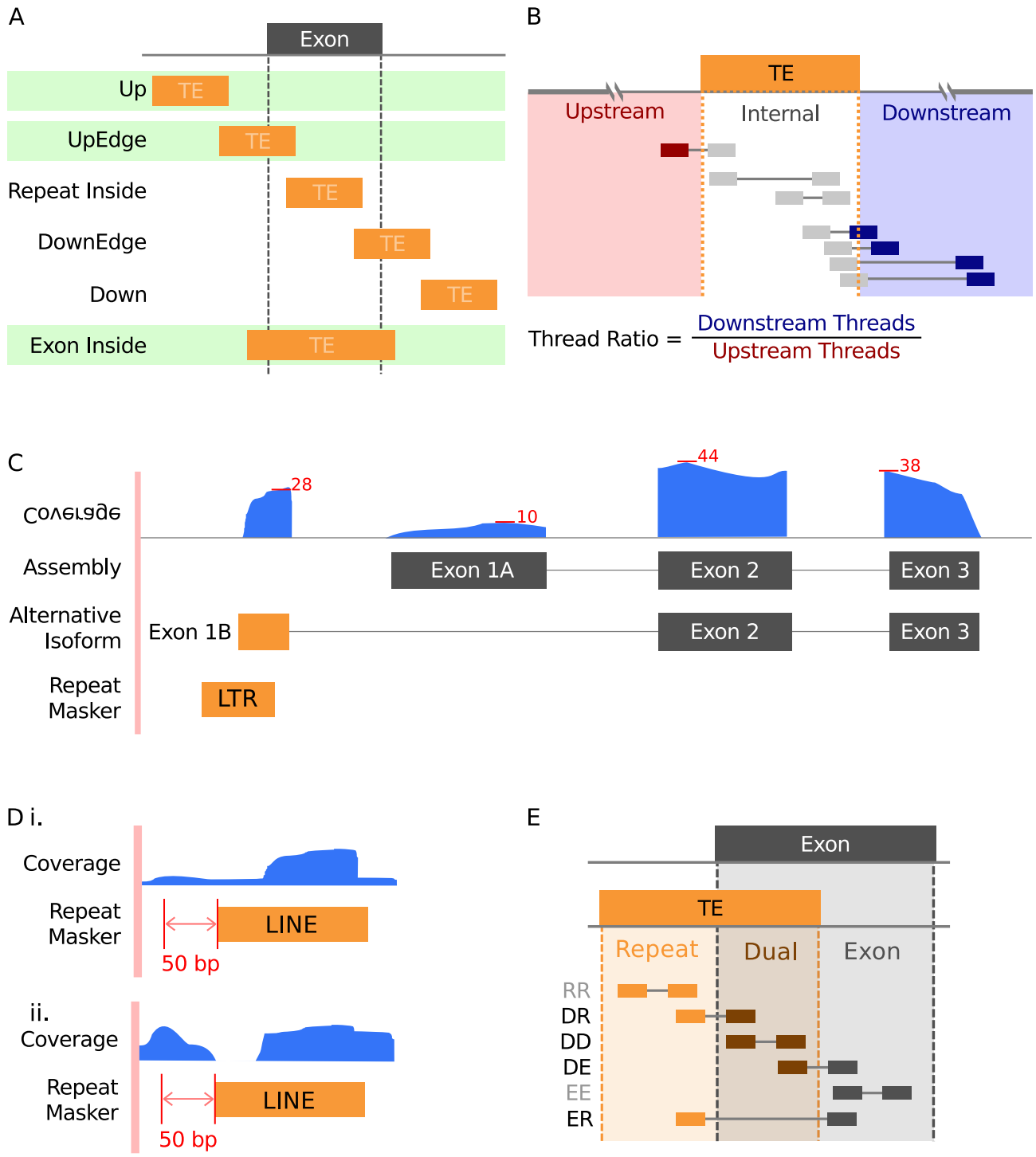
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 2



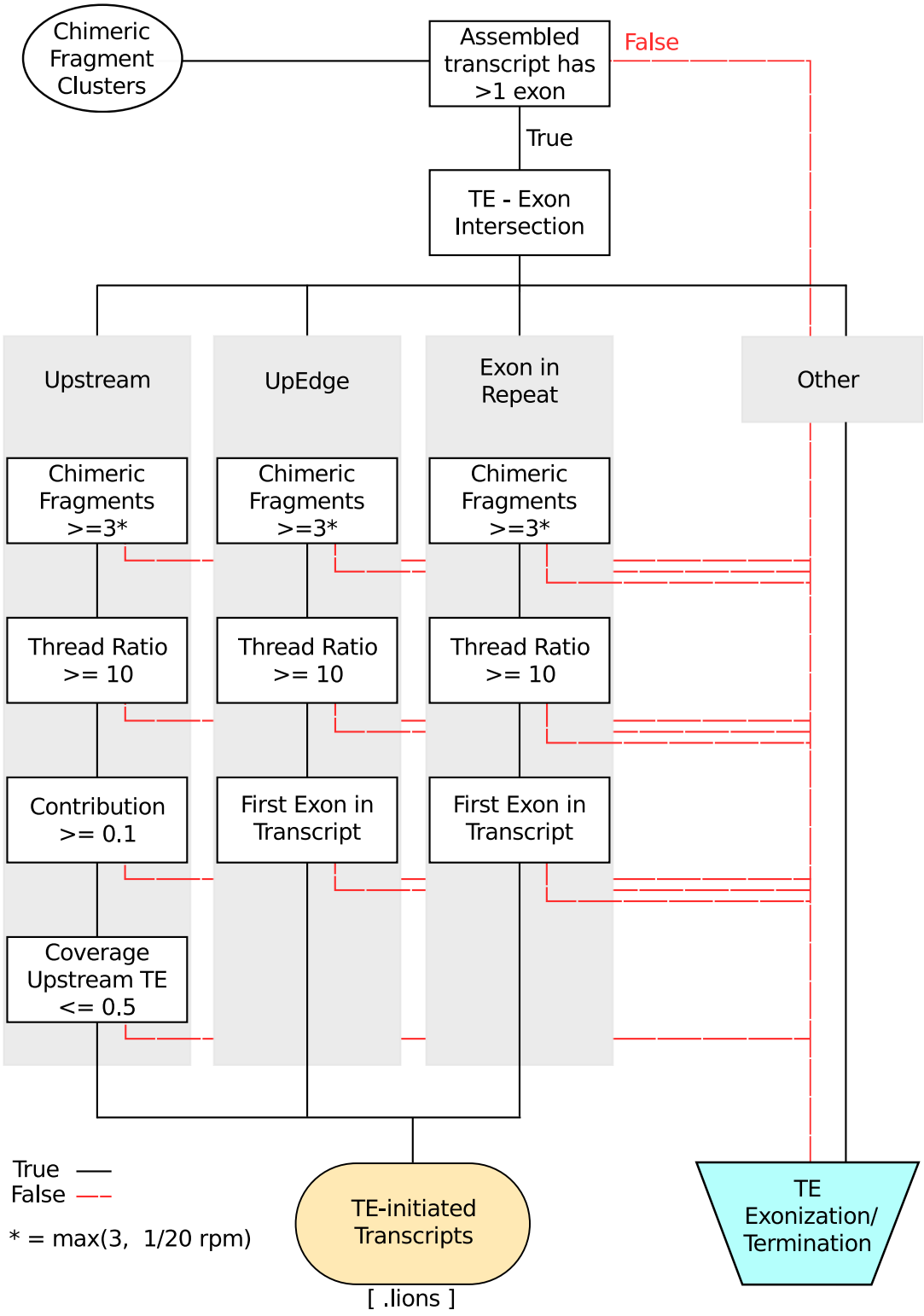
bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

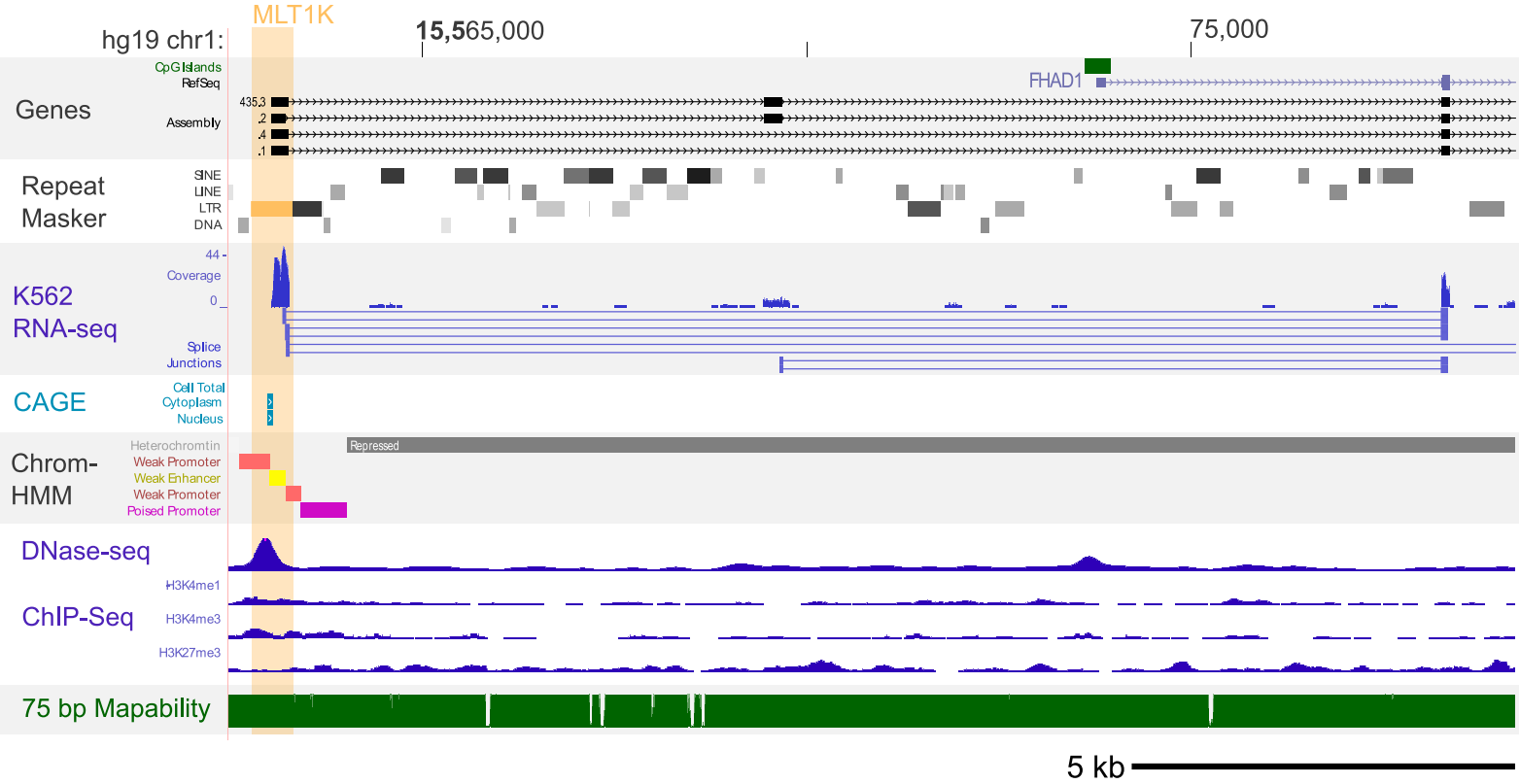
Supplementary Figure 3



bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 4



bioRxiv preprint doi: https://doi.org/10.1101/149864; this version posted June 14, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supplementary Figure 5