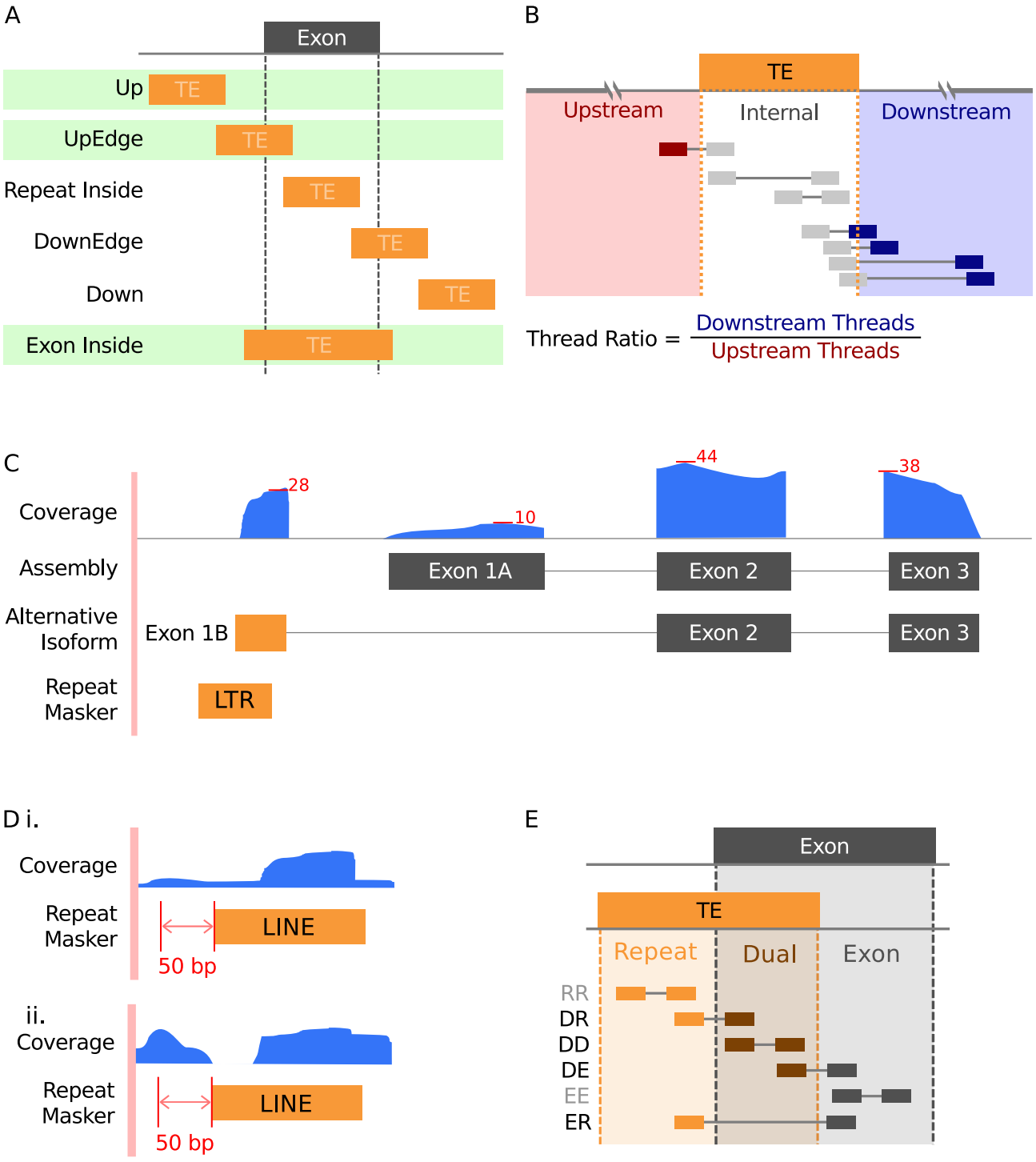
Supplementary Figure 1

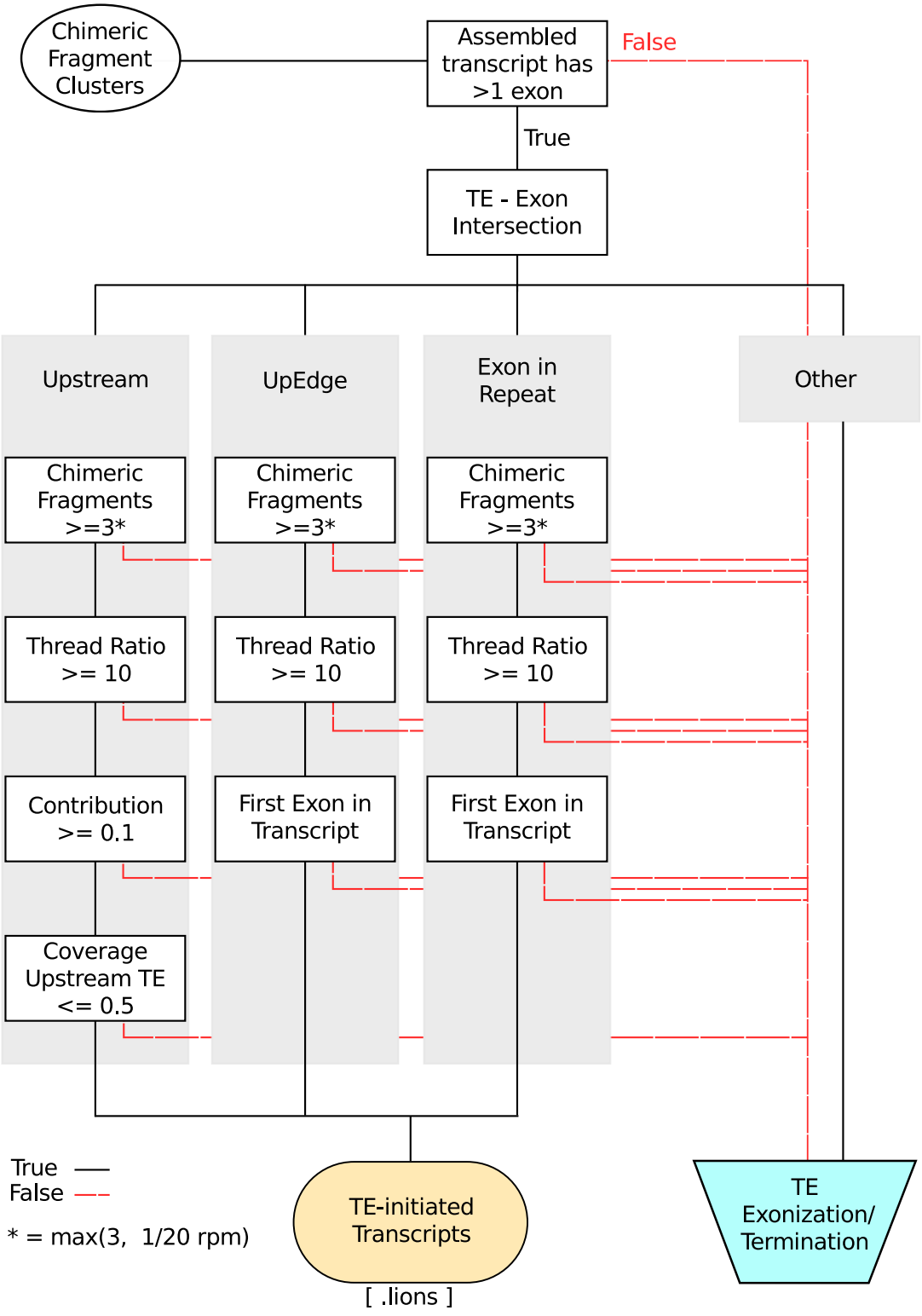


Supplementary Figure 1:

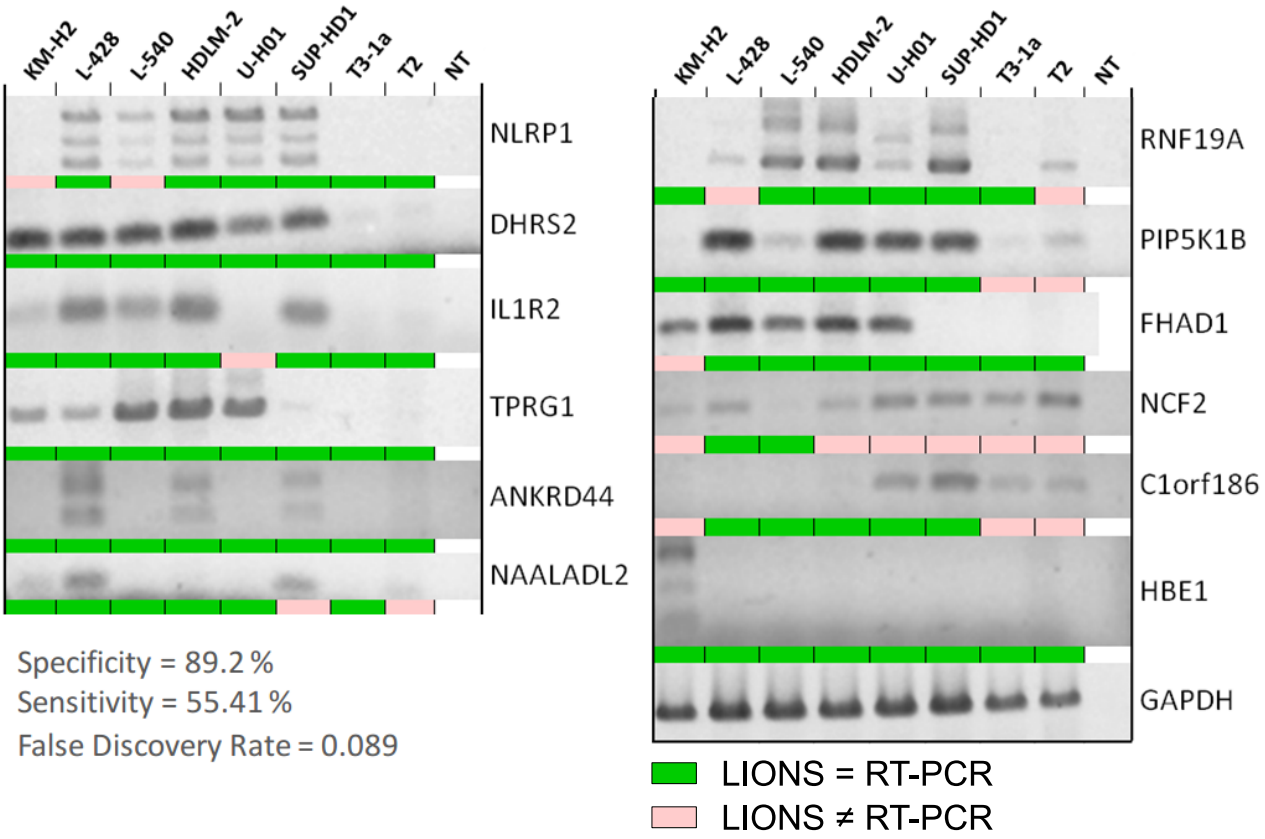
Calculated values for LIONS classification

To distinguish transposable element (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, contained within the exon 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 locus. For example, part “i” shows a locus with low levels of transcriptional readthrough but a potential initiation site present within the TE. In contrast, part “ii” indicates a locus in which there is an apparent gain of coverage within the LINE element which could be due to poor mapping quality at the 5' end of this LINE element. E) Chimeric fragment subclassification 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 sorting algorithms (Supplementary Figure 2).

Supplementary Figure 2



Supplementary Figure 3



Supplementary Figure 2:

Chimeric Fragment Clusters Sorting Algorithm for TE-initiated Transcripts

For each Transposable Element (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. Notably, the number of supporting chimeric fragments can be set to a function of the number of mapped reads (i.e. 1 fragment / 20 million mapped reads) to fairly compare libraries of varying 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 exon's 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.

Supplementary Figure 3:

Validation Set of Chimeric Transcripts in Hodgkin Lymphoma Cell Lines

LIONS was run on the set of RNA-seq libraries from HL cell lines and primary B cells listed in Supplementary Table 1. From the output of *LIONS*, the top-most HL recurrent (present in multiple HL libraries) and specific (absent from all B-cell control libraries). Chimeric transcripts (TE-initiated transcripts which splice into a protein coding gene in the sense orientation) were validated by reverse transcriptase PCR (RT-PCR) (primers in Supplementary Table 2). T3-1a and T2 are RNAs from primary B cell tonsil samples. NT indicates no template control. The other lanes are RNAs from various HL cell lines. A green bar indicates the *LIONS* classification of present/absent is concordant with the RT-PCR, a pink bar indicates discordance between the methods. RT-PCR is expectedly much more sensitive for low-abundance transcripts (note the fainter bands).