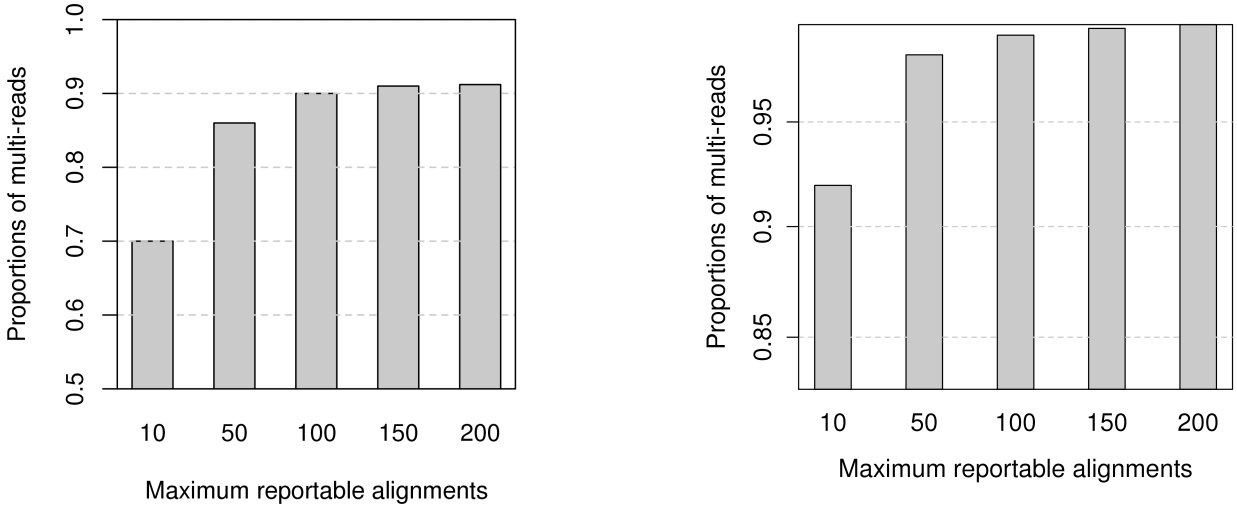
Supplementary materials

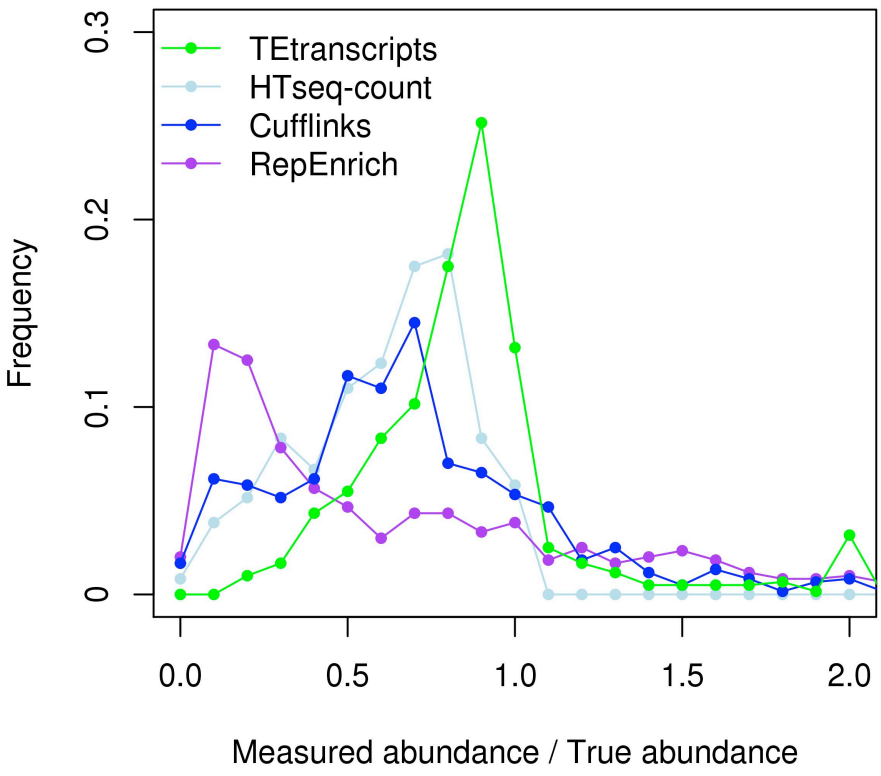
Supplementary Table 1: Flux Simulator parameter settings
Simulated RNA-seq samples were generated with Flux Simulator using the following parameters.

Parameter	Value
# Expression	
NB_MOLECULES	5000000
TSS_MEAN	50
LOAD_NONCODING	NO
POLYA_SCALE	NaN
POLYA_SHAPE	NaN
# Fragmentation	
FRAG_SUBSTRATE RNA	RNA
FRAG_METHOD	UR
FRAG_UR_ETA	350
# Reverse Transcription	
RTRANSCRIPTION	YES
RT_MOTIF	default
RT_LOSSLESS	YES
RT_MIN	500
RT_MAX	5500
# Amplification	
PCR_DISTRIBUTION	default
GC_MEAN	NaN
PCR_PROBABILITY	0.05
# Size Filtering	
FILTERING	NO
# Sequencing	
READ_NUMBER	2000000
READ_LENGTH	76
PAIRED_END	NO
FASTA	YES
UNIQUE_IDS	YES

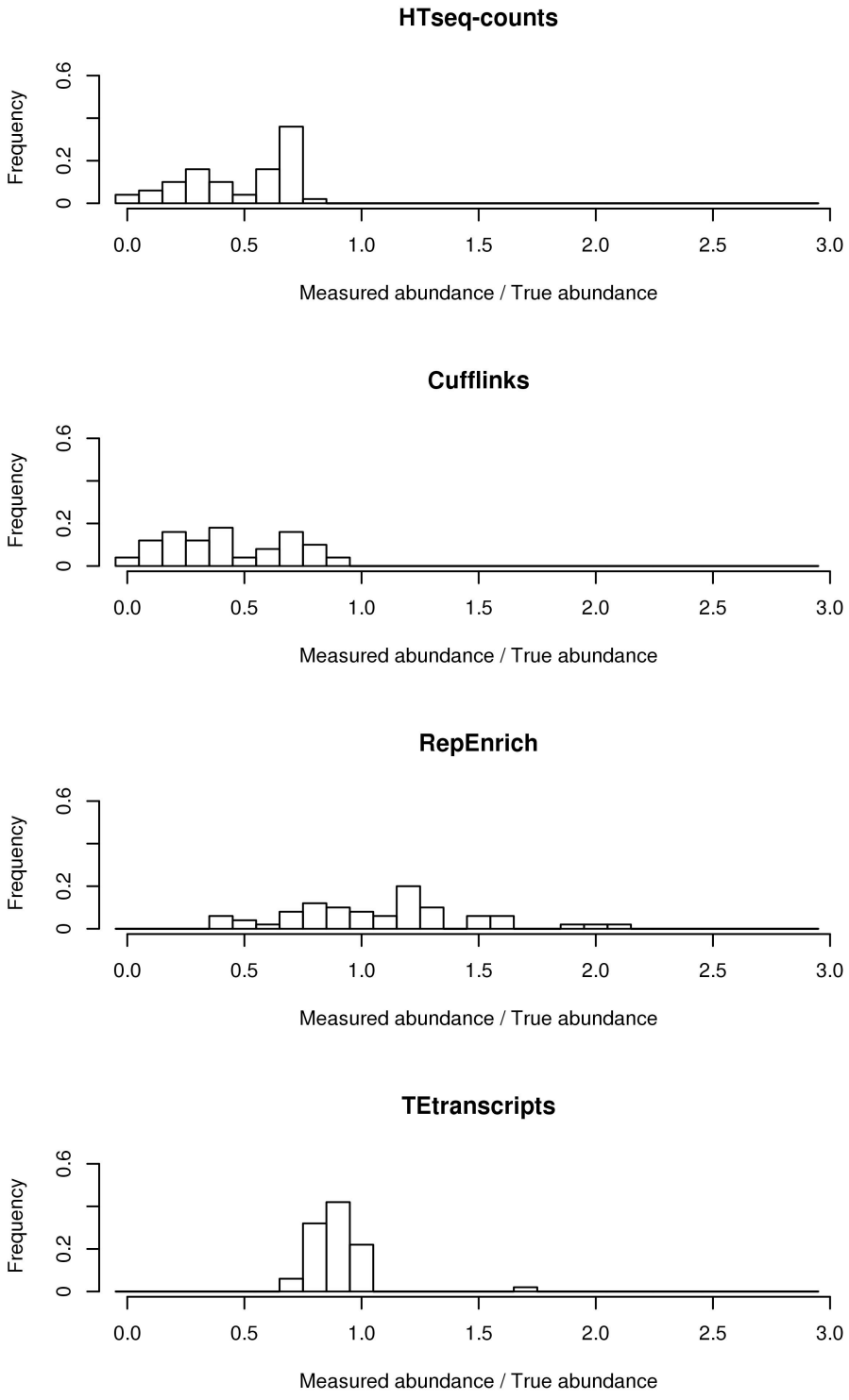


Supplementary Figure 1: Saturation analysis of multi-mapped reads. Left plot shows a published mouse RNA-seq dataset (GSE27394) used in the study of transposable elements mis-regulation in TDP-43-mediated neurodenerative disorders (Polymenidou et al. 2011); right graph shows one of the testis RNAseq data used in Figure 5 of the main text (GSE30352) (Brawand et al. 2011). Using different cut-offs of maximum alignments per read, from 10 to 200, we were able to show that the proportion of multi-reads increase dramatically from 10 to 100, with 90% of all multi-reads reported. However, the increase from 100 to 200 shows diminishing returns, with only about 1% additional reads recovered. Based on this analysis, we recommend using a cut-off of 100 maximum alignments for this dataset.

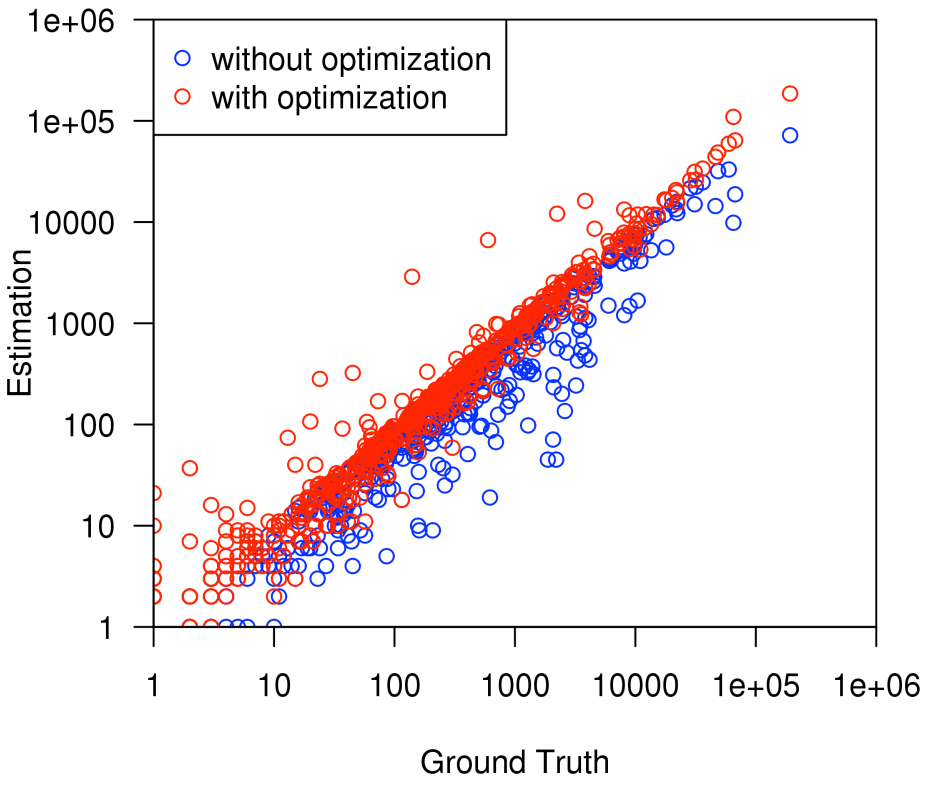
Many aligners support multi-reads alignments, and provide controls for the maximum number of multiple alignments per reads to output. For example, STAR has two parameters that play the most important role in the report of multi-mappers, --winAnchorMultimapNmax and –outFilterMultimapNmax. The author of STAR recommends setting winAnchorMultimapNmax =2 * outFilterMultimapNmax, but no less than 50. For example, in the above case, user can set –outFilterMultimapNmax 100 and –winAnchorMultimapNmax 200. However, increasing winAnchorMultimapNmax allows STAR to use shorter seed as anchors, which increases sensitivity for problematic alignments (with many/mismatches indels). Therefore, user may need to test different combinations of these two parameters and to find the one reports the most multi-reads and at mean time keeps the high alignment qualities.



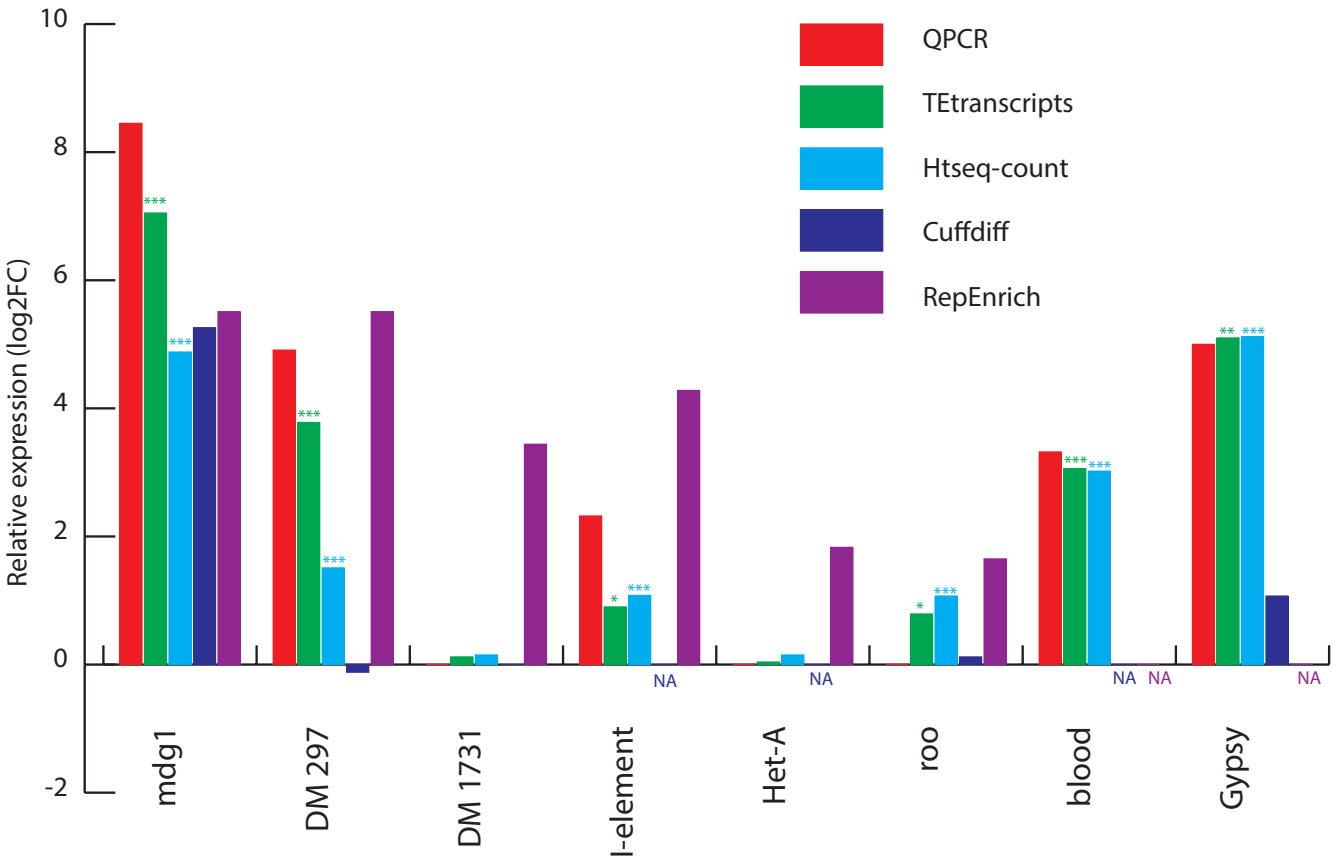
Supplementary Figure 2: TE abundance recovery rate using simulated data. Except HTSeq-count, all the other approaches over-estimate the abundances of some TEs. However, overall TEtranscripts outperforms all the other methods.



Supplementary Figure 3: TE abundance recovery rate on active TEs using simulated data. TEtranscripts performs much better than the other approaches in estimating abundances of active TEs.



Supplementary Figure 4: Comparison between with and without Expectation Maximization optimization. *TEtranscripts* was run on a simulated data (generated as described in the main text, based on mm9 genome with total number of reads 20M and 10% of TE reads.), with or without Expectation Maximization. Estimated read counts of TEs were plotted against the ground truth. Although we set the iteration parameter as *-i* 20 (i.e., up to 20 iterations), the algorithm actually converged in 6 iterations. The graph shows that EM optimization correlates better to the real counts than no optimization, and with minimal additional cost in running time.



Supplementary Figure 5: Comparison of Cuffdiff to other methods on published Drosophila RNA-seq data. We used Cuffdiff version 2.2.1 on the published Drosophila dataset (Ohtani et al. 2013), and compared it to the qPCR data and other approaches (Htseq-count + DEseq, RepEnrich + DEseq, and TEtranscripts) as described in the main text. Log2 fold changes of expression between wildtype and *piwi* knockdown are shown. "NA" denotes circumstances where expression could not be estimated. "*" represents the level of significance, * : FDR<0.05, **: FDR<0.01, ***: FDR<1e-5. We used the same Gene/TE annotation files for Cuffdiff and other approaches. Cuffdiff took 550Gb of memory and 4 days & 5 hours to call differential expressions, and failed to quantify several TEs.

References:

Polymenidou, M. et al. (2011). Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43. Nat Neurosci. 14, 459-468.