Supplementary Material

Gene Scissors: a comprehensive approach to detecting and correcting spurious transcriptome inference due to RNAseq reads misalignment.

Zhaojun Zhang[1], Shunping Huang[1], Jack Wang[1], Xiang Zhang[2], Fernando Pardo Manuel de Villena[3], Leonard McMillan[1], and Wei Wang[4]
[1]Department of Computer Science, University of North Carolina at Chapel Hill.
[2]Department of Electrical Engineering and Computer Science, Case Western Reserve University.
[3]Department of Genetics, University of North Carolina at Chapel Hill.
[4]Department of Computer Science, University of California, Los Angeles.

Simulation study on expressed pseudogenes

To evaluate the performance of GeneScissors on expressed pseudogenes, we added some expressed pseudogenes in our simulation study. In addition to the 13000 non-pseudogenes used in our previous simulation study, we also randomly selected 179 pseudogenes (5% of the annotated pseudogenes in Ensembl database) as expressed genes for every simulated sample. We compared the TopHat pipeline and the GeneScissors (TopHat) pipeline. In addition to the three metrics used before, we also measured the recall of the expressed pseudogenes, which is the percentage of the 179 expressed pseudogenes that are correctly identified by GeneScissors. The results in Table S1 are averaged over 10 samples.

	TopHat Pipeline	GeneScissors (TopHat)
GenePrecision	41.8%	**48.5%**
GeneRecall	93.0%	93.0%
GeneF-measurement	57.9%	**63.6%**
PseudoRecall	**90.0%**	89.5%

Table S1: Comparison of TopHat pipeline and GeneScissors (TopHat) pipeline

The results in Table S1 are consistent with that in Table 3 in the paper. Overall, GeneScissors has 6.6% improvement on GeneF-measurement. Moreover, 89.5% of the 179 expressed pseudogenes are reported as expressed genes by GeneScissors pipeline, while 90% are reported by TopHat pipeline. Both rates are slightly less than the overall recall rate 93%, which suggests that the expressed pseudogenes are harder to detect, because most of them are from repetitive regions of the genome.

Feature selection study

We examined the importance of each of the five feature categories by excluding one at a

time and measuring its impact to the precision, recall, F1, and AUC measurements. We used RandomForests as the classification method. All scores were generated from a 5-fold cross-validation.

In Table S2, we show the scores of the complete model, the alternative models by excluding one feature category, and the baseline (which is the percentage of fragment attractors that are expressed genes in the simulation).

	Complete Model	Remove NE features	Remove NR features	Remove MF features	Remove MR features	Remove CM features	Base Line
Precision	0.896	0.872	0.873	0.811	0.893	0.890	0.57
Recall	0.877	0.861	0.858	0.808	0.872	0.875	0.57
F1	0.886	0.866	0.865	0.809	0.882	0.883	0.57
AUC	0.910	0.890	0.883	0.813	0.903	0.906	NA

Table S2: Summary of the feature selection study

The complete model that uses all features always scores the best among all alternative ones, suggesting that all features are necessary. However, these features are not independent. Removing any feature categories always leads to a drop in the scores. Since the RandomForests classification model does not require independent features, its result is not impaired by such dependencies. In our future study, we plan to further investigate the dependencies and their roles in predicting expressed genes.