DELongSeq: An Efficient Isoform Expression Detection Method for long-read RNA Sequencing Data

Anagha Gouru (C '25), Yu Hu, Kai Wang
Department of Pathology and Laboratory Medicine, Perelman School of Medicine
Funded by: Penn Undergraduate Research Mentoring Program

Background

- RNA Sequencing tells us **which genes and isoforms are active** and **how much each splicing isoform is transcribed.**
- Done through isolating mRNA, filtering out ineffective reads, sequencing, and finally using differential expression (DE) analysis to detect genes. This can be taken one step further with functional enrichment, which shows the results' biological significance.
- Next-generation sequencing can either be short-read or long-read sequencing
  - Short-read: fragments the DNA, sequences, and ligates back together
  - Long-read: analyzes the full-length transcript without having to assemble again

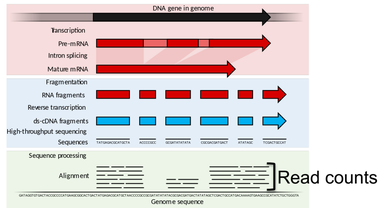

Figure 1: Common RNA-Seq Workflow (Transcriptomic Technologies, 2017)

Introduction

- Current short-read and long-read sequencing methods use **read counts** (see Background) to estimate expression levels.
- Commonly used long-read detection methods include:
  - **TALON**, which isolates novel genes by comparing them to existing transcript models.
  - **FLAIR**, which clusters alignments into groups and collapses into isoforms.
  - **LIQA**, which corrects sequencing bias in expression estimation.
- However, long-read methods still have uncertainty in expression estimation due to variability in precision across samples. Further, short-read methods cannot account for covariates or confounders, such as environment or age, that may affect gene expression.
- **DELongSeq** both accounts for uncertainty in isoform expression estimation and variation in precision of expression estimation across biological replicates, allowing for covariance.
- For optimal results, DELongSeq uses biostatistical techniques, employing the information matrix of the Expectation-Maximization (EM) Algorithm.

Materials/Methods

- DELongSeq quantifies the **uncertainty of isoform expression estimates**, performing maximum likelihood estimation in the presence of latent variables (not directly observed but inferred), as well as a random-effects regression model to account for variable uncertainty.
- We ran DELongSeq on both **simulated** and **real** RNA-Seq data, including esophageal cells (PRJNA15570), gastric cells (GSE157750), and neuroblastoma cells (GSE74886). See Results for more info.

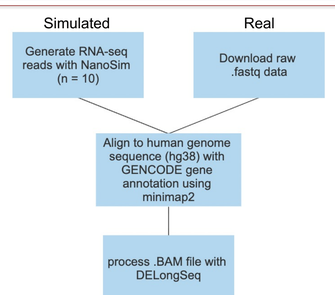

Figure 2: Overview of DELongSeq Workflow

Results – Simulated Data

To evaluate performance, we compared DELongSeq's' performance to two existing detection algorithms: TALON and FLAIR.

- **Uncertainty Quantification:** we evaluated the estimation accuracy of DELongSeq, TALON, and FLAIR with a set of measurements.
  - Measured the similarity between estimated isoform abundance and the ground truth using Spearman's correlations (Cor = 0.65)
  - Measured the 95% CI length of DELongSeq estimates and compared to the ground truth, estimating the similarity with Spearman's correlations (Cor = 0.81)
  - For major isoforms (highest fraction among all isoforms for a gene) with less than 10 reads mapped to the gene (read count = 10), the accuracy of TALON and FLAIR was **20% lower**, whereas DELongSeq's was less than 5%.

Isoform-Based Tests: we evaluated DELongSeq's performance of isoform detection, setting a threshold to find isoform level difference between cases and controls. We required this expression difference to 0.5 to declare it as a DE event. To make a fair comparison across all 3 methods, type I error rate and power were calculated with a threshold of 0.5 and 1.0.

False Positive	t_0 = 0.5		t_0 = 1.0	
	N = 10	N = 5	N = 10	N = 5
DELongSeq	0.015	0.025	0.021	0.035
FLAIR	0.026	0.028	0.024	0.038
TALON	0.037	0.051	0.048	0.086

Table 1: For isoform-based, DELongSeq had the lowest false positive rate compared to the other two methods.

Power	t_0 = 0.5		t_0 = 1.0	
	N = 10	N = 5	N = 10	N = 5
DELongSeq	0.639	0.580	0.643	0.586
FLAIR	0.537	0.481	0.591	0.537
TALON	0.432	0.394	0.428	0.381

Table 2: For isoform-based, the power was highest for DELongSeq and increased for all three methods with a higher threshold.

- **Gene-Based Tests:** we evaluated the performance of gene-based analysis using the same method as isoform-based tests, with thresholds of 0.5 and 1.0.

False Positive	t_0 = 0.5		t_0 = 1.0	
	N = 10	N = 5	N = 10	N = 5
DELongSeq	0.011	0.024	0.006	0.028
FLAIR	0.013	0.027	0.007	0.037
TALON	0.020	0.039	0.038	0.0

Table 3: For gene-based, DELongSeq had the lowest false positive rate compared to the other two methods.

Power	t_0 = 0.5		t_0 = 1.0	
	N = 10	N = 5	N = 10	N = 5
DELongSeq	0.695	0.631	0.756	0.685
FLAIR	0.585	0.523	0.695	0.628
TALON	0.470	0.428	0.505	0.445

Table 4: For gene-based, the power was highest for DELongSeq and increased for all three methods with a higher threshold.

- **1 Case vs. 1 Control Comparisons:** DELongSeq can provide estimation of variation even with one sample. We compared this to a simple linear regression model and divided the isoforms into 3 estimation variance quantiles (0-33%, 33-66%, 66-100%).
  - DELongSeq performed consistently better than the simple linear regression model, with power staying relatively stable as sample size decreased. For a decrease from n = 10 to n = 5, the powers were 0.61 and 0.55 for DELongSeq compared to 0.63 and 0.51 with the linear regression model.

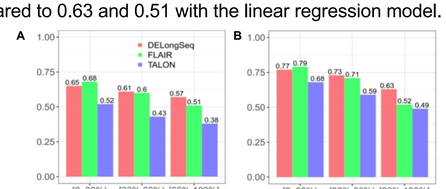

Figure 3: All methods appeared to decrease in power for isoforms with higher uncertainty on an (A)isoform and (B)gene level. However, DELongSeq and FLAIR's powers maintained relatively stable as uncertainty increased compared to TALON.

Results – Real Data

We also evaluated DELongSeq's isoform detection performance on three datasets.

- **Esophageal Squamous Epithelial Cells (ESCC):** we applied DELongSeq, TALON, and FLAIR on normal and cancer cells to detect differential isoform usage. (Fig. 4)

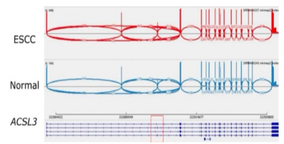

Figure 4: ESCC results indicated that the average read coverage at these genes was 21.6 for DELongSeq and 8.2 for the other methods.

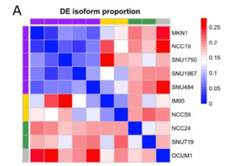

Figure 5: A heatmap with GC DE analysis results showed DELongSeq clearly differentiated the subtypes (see Figure 5) and detected subtype-specific isoform expression.

- **Gastric Cancer (GC):** we performed 1 vs. 1 sample comparisons on 10 different cell lines representing 4 cell subtypes (chromosomal unstable, Epstein-Barr virus positive, genome stable, microsatellite unstable). (Fig. 5)
- **Human SH-SY5Y Neuroblastoma Cells (5Y):** From 1 vs. 1 comparison analysis, most detected genes and isoforms had p-values < 0.05 and Spearman correlation of 0.68, suggesting DELongSeq's precision with a smaller sample size. DELongSeq also generated a smaller 95% CI.

Discussion

- Detection of genes and isoforms with mis-regulated expression is a **critical step** in transcriptomics studies.
- Existing methods ignore estimate variance in DE analysis, which can result in poor detection and higher false positive rate. The simulation study showed that DELongSeq produces a **higher power and lower false positive rate**, and this threshold can be further maximized in future studies.
- DELongSeq allows for adjustment of covariates and accounts for variation across samples. DELongSeq results showed that as sample size decreased, the **power stayed relatively stable** for both DE and non-DE samples.
- DELongSeq allows for **1 case vs. 1 control comparisons**, and simulations showed it was more powerful than existing methods

Acknowledgements/Further Information

Thank you to Prof. Kai Wang, Dr. Yu Hu, and the members of the Wang Lab for an insightful summer experience, and for challenging me every step of the way. This study is supported by the Penn Undergraduate Research Mentoring (PURM) program.

References

1. Yu Hu, Anagha Gouru, Kai Wang. DELongSeq for efficient detection of differential isoform expression from long-read RNA-Seq data. 2022. [SUBMITTED]
2. Yu Hu, Li Fang, Xuelian Chen, Jiang F. Zhong, Mingyao Li, Kai Wang. LIQA: Long-read Isoform Quantification and Analysis. 2020. bioRxiv doi: https://doi.org/10.1101/2020.09.09.289793
3. A technology-agnostic long-read analysis pipeline for transcriptome discovery and quantification. Dana Wyman, Gabriela Balderrama-Gutierrez, Fairlie Reese, et. al, bioRxiv 672931; doi: https://doi.org/10.1101/672931
4. Tang, A.D., Soulette, C.M., van Baren, M.J. et al. Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. Nat Commun 11, 1438 (2020). https://doi.org/10.1038/s41467-020-15171-6
5. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017 May 18;13(5):e1005457. doi: 10.1371/journal.pcbi.1005457. PMID: 28545146; PMCID: PMC5436640.