5-2024

Development of an Isoform Atlas in Pediatric Patients with Rare Diseases using Iso-seq

Boryana Koseva

Development of a Gene Isoform Atlas Across Perinatal and Pediatric Tissues

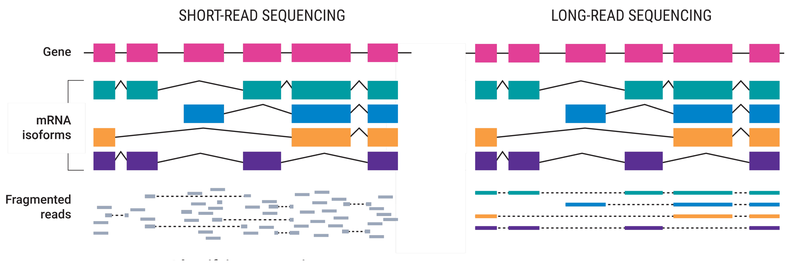
Boryana S. Koseva on behalf of Grundberg Lab and Pastinen Lab

Dept. of Pediatrics, Genomic Medicine Center, Children's Mercy Kansas City

Background

Short-read RNA sequencing has become the standard method for rapidly sequencing whole genomes, annotating transcriptomes and quantifying gene expression. However, short-read RNA sequencing can be bioinformatically challenging because the full transcript is inferred from short fragments, either by overlapping the sequenced fragments (*de novo*) or by aligning to a reference genome or transcriptome making it less than ideal to use in identifying and characterizing the biological diversity of transcripts (isoforms).

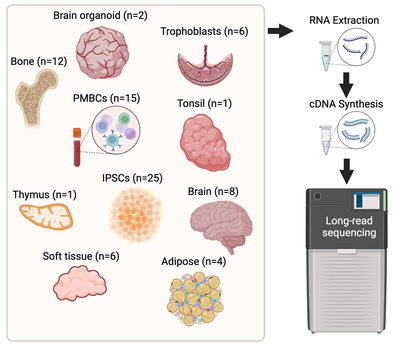
Full-length (FL) RNA sequencing has been developed in which a single molecule of up to 10 kilobase can be sequenced with high confidence, removing the need to infer the transcript from short fragments. This new approach gives us the ability to discover isoforms resulting from alternative splicing or gene fusion events, as well as detect allele-specific expression and single nucleotide variants.

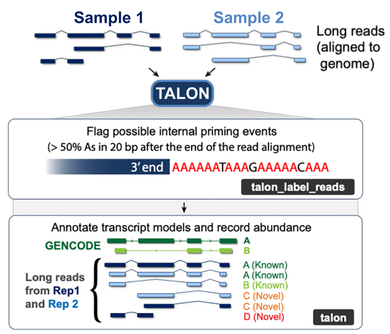


Project Goal

We aim to to build an atlas of gene isoforms found in a population (N=81) of pediatric patients with rare diseases by leveraging Iso-seq, the FL RNA sequencing method developed by Pacific Biosciences (PacBio), across 10 different tissue types.

Methods and Analysis

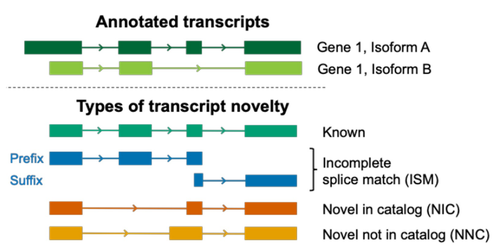


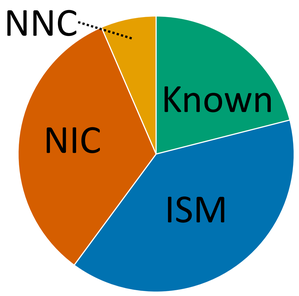


RNA was isolated from up to 10 tissue types for FL RNA (cDNA) sequencing. The sequenced reads for each sample were analyzed using the Isoseq V3 pipeline. Sequencing adaptors and concatemer reads were removed, and the resulting isoforms were clustered into a set of non-redundant high-quality isoforms. These isoforms were aligned to the Human reference genome (GRCh38). The mapping to the reference genome and the GENCODE annotation were used as inputs to the Talon software which categorized the isoforms based on how well they match the reference genome annotation and created a comprehensive isoform catalog. Novel transcripts were only retained if it was observed in at least 5 samples regardless of tissue origin.

Observed Isoform Diversity

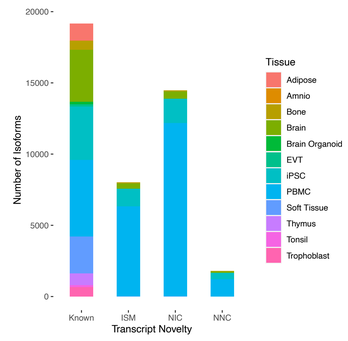
Our preliminary catalog contains a total of 240,584 unique transcripts. Of those, only 21% of isoforms are a perfect match to the reference genome annotation (i.e., known). The remaining 79% were not accurately represented in the reference annotation.





Tissue-specific Transcripts

To start addressing tissue-specific expression questions, we retrieved transcripts seen in a single tissue type. Known isoform models are included in our atlas when there is at leas one sample with the observed transcript. To exclude sequencing artifacts from the atlas we require that a novel isoform model is observed in at least 5 samples to be included.



Conclusions

Our preliminary observations suggest that there is more biological diversity in human transcriptomes than what has been detected using short-read sequencing. While there is a variety of human-related atlases that are publicly available, our study is the first to undertake the effort to catalog the full complement of isoforms in pediatric patients. Furthermore, the diversity of tissue types in our study also allow us to examine tissue-specific transcript novelty.

Future Directions

To improve the accuracy of the classification and the filtering step, we plan to incorporate orthogonal data such as CAGE and poly-A annotations. Our overall goal is to expand the atlas to include all GA4K and perinatal samples that have been sequenced using LR Sequencing (n = 265), and to highlight tissue-specific isoforms.

Acknowledgments

We would like to thank all families for participating in the Genomic Answer for Kids Study (GA4K). This work was made possible by the generous gifts to Children's Mercy Research Institute and GA4K at Children's Mercy Kansas City. Illustrations were created in BioRender.com or provided by PacBio and Talon authors.