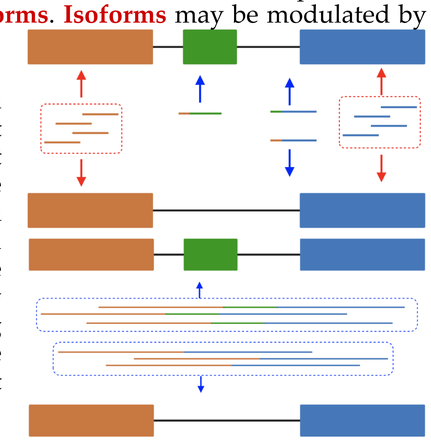
Count Splitting Controls For Type 1 Error in Differential Testing After Tree Merging of Gene Isoforms

Justin Landis[1], Michael Love[2, 3]

[1] Biological & Biomedical Sciences Program
[2] Department of Genetics, UNC Chapel Hill
[3] Department of Biostatistics, UNC Chapel Hill

Background

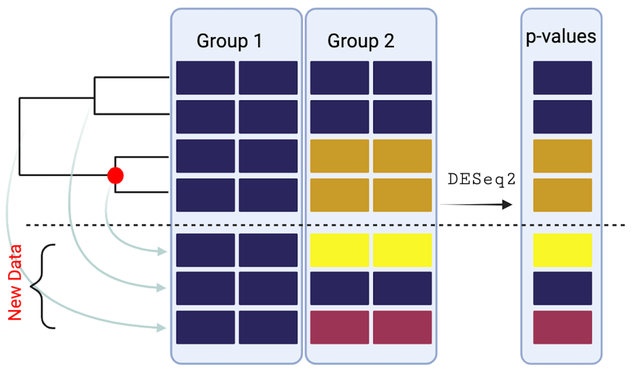
One **Gene** may be translated into *mRNA* and then spliced into multiple transcripts named **isoforms**. **Isoforms** may be modulated by splicing factors within the cell.[1]



RNAseq takes a snapshot of a cell's gene expression profile at the time of sequencing. Short read sequencing, 100-250 base pairs, required estimating which **isoform** is present since the read may be contained within one *exon*. **Long read sequencing** may span multiple *exons*, providing more confidence on the transcript **isoform** detected at the expense of read depth.

Goal: Develop an isoform-grouping method to facilitate isoform-level **differential expression (DE)** analysis using long-read sequencing data while controlling for False Positives for differential expression.

DE Testing of Inner Nodes



A gene with N isoforms implies $N-1$ inner nodes for its associated tree. These inner nodes are the sum of the leaves of a sample. Once we have our extended data, we perform `DESeq2`[2] and evaluate the resulting **pvalues** with our tree Climbing algorithm.

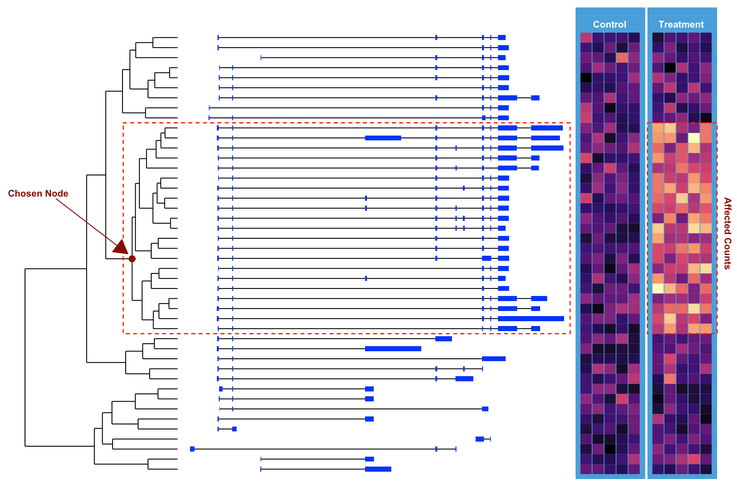
Cluster Tree Generation Method

To Generate Hierarchical clusters, we generated a **similarity metric** based on the similarities between transcripts as opposed to using data dependent counts. Let G represent a set of isoforms of size g. given any indexes $i,\ j \leq g$, define G_i and G_j as isoforms i and j from G such that they represent sets of exons of size N and M respectively. For any two i and j, we can define the similarity as:

$$S_{ij}(G_i, G_j) = \frac{2\sum_n^N \sum_m^M J(G_{i_n}, G_{j_m})}{N+M} \qquad J(G_{i_n}, G_{j_m}) = \frac{G_{i_n} \cap G_{j_m}}{G_{i_n} \cup G_{j_m}}$$

Simulation Methods

Choose an inner node within the tree and shift the mean of all leaves for a particular group by some delta. We can evaluate our Tree climbing algorithm based on how well it accurately chooses the known perturbed nodes.

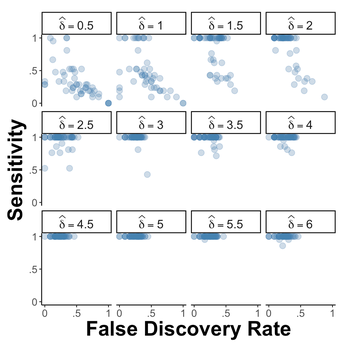


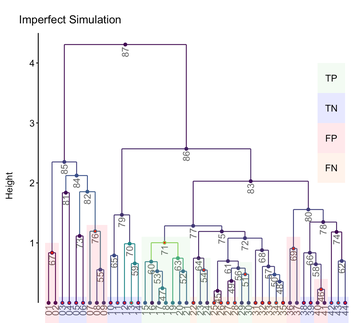
Let $\mu_{ij} = \mu_0 + \delta_{ij}$, $\mu_0 = 10$ and $\delta_{ij} = 0$ for i, j in the control group, and $\delta_{ij} = \hat{\delta}$ for i, j in the affected group. Each entry, X_{ij}, is sampled as follows.

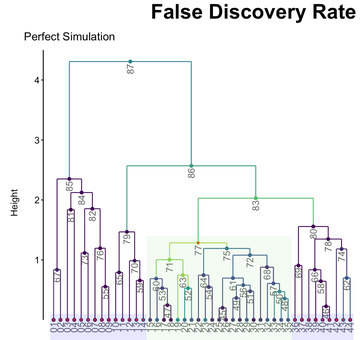
$$X_{ij} \sim \text{Nbinom}(\mu_{ij},\ \alpha = 100)$$

Simulation Sensitivity

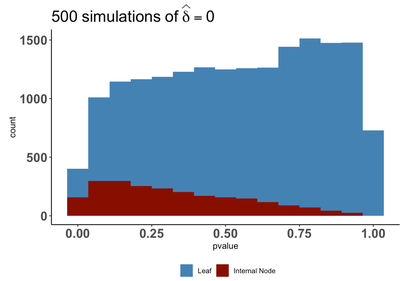
We conducted 50 simulations per $\hat{\delta} \in \{0.5, 1, 1.5, \ldots, 6\}$ and evaluated how often our tree climbing algorithm correctly merged the known perturbed data. Merged nodes with $\delta_{ij} = 0$ are **False Positives** and unmerged nodes of with $\delta_{ij} = \hat{\delta}$ are considered **False Negatives** in the tree climbing context.







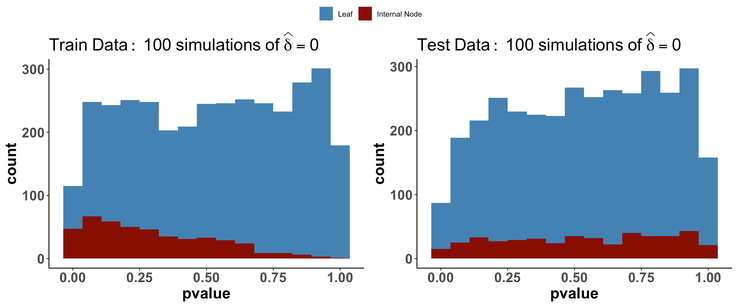
Count Splitting[3]



While at higher $\hat{\delta}$ shifts in our simulations are able to identify the correct inner node, but we also need to consider the case of the **null hypothesis**. To assess the distribution of pvalues under the assumption that there is no difference in the μ_{ij}, we ran simulations with $\hat{\delta} = 0$.

There is an **enrichment of low pvalues** among the **inner nodes** in the **null hypothesis simulations**. To correct this we can apply **count splitting**. The count splitting method is formulated for experiments that use the same data for feature selection as they use for analysis.[3]

$$X_{ij} \sim \text{Nbinom}(\mu_{ij}, \alpha = 100)$$
$$X_{ij}^{\text{train}} \sim \text{Bin}(X_{ij}, \theta = 0.5)$$
$$X_{ij}^{\text{test}} = X_{ij} - X_{ij}^{\text{train}}$$



Conclusions

1. Simulations of a single gene of 44 isoforms can reliably detect the correct node with $\log_2$ fold change of 0.5 between groups.
2. Utilization of count splitting controls for type 1 error under the null hypothesis.

Future Directions

1. Improve speed of data merging step.
2. Apply tree climbing and count splitting methods on real data sets.

References

1. Kim HS, Grimes SM, Hooker AC, Lau BT, Ji HP. Single-cell characterization of CRISPR-modified transcript isoforms with nanopore sequencing. *Genome Biology*. 2021;22(1):331. doi:10.1186/s13059-021-02544-1
2. Love M, Anders S, Huber W. *DESeq2: Differential Gene Expression Analysis Based on the Negative Binomial Distribution*.; 2023. doi:10.18129/B9.bioc.DESeq2
3. Neufeld A, Gao LL, Popp J, Battle A, Witten D. Inference after latent variable estimation for single-cell RNA sequencing data. Published online 2022. https://arxiv.org/abs/2207.00554
4. Thorne B. *Posterdown: Generate PDF Conference Posters Using r Markdown*.; 2019. https://github.com/brentthorne/posterdown
5. Created with biorender. https://Biorender.com