Beyond Split-Reads: Leveraging Pseudoalignment for Enhanced Splicing Event Detection in Low-Coverage Data with Leafcutter

Xingpei Zhang
xz3149

Advisor: Dr. David A. Knowles
Assistant Professor of Computer Science

Thesis
submitted in partial
fulfillment of the requirements
for
M.S. in Computer Science

December 2023

Table of Contents

Abstract

The precise definition and detection of alternative splicing (AS) have long posed challenges for researchers. Previous studies have introduced various methods to address this issue, including rMATS, DEXSeq, and Leafcutter (Anders et al., 2012; Shen et al., 2014; Li et al., 2018). Leafcutter, in particular, uses intron excision ratios to identify differential splicing events. While these methods perform adequately on high-coverage bulk RNA-seq data, their performance deteriorates on low-coverage data, particularly single-cell data, due to limited supporting reads for differential splicing events. To tackle this issue, we leverage Salmon, a pseudoalignment tool, to enhance the detection capabilities of Leafcutter in low-coverage bulk and single-cell pseudobulk data (Patro et al., 2017; Li et al., 2018). Here, we present LeafcutterITI (for "Isoform to Intron"), a modified version of Leafcutter, to quantify introns from the abundance of corresponding isoforms, rather than relying on split reads alone. This method permits more reads to support the intron excision events than what a split-read approach can offer. We validate the performance of this modification through simulations and on real data. Lastly, we analyze single-cell pseudobulk data from the Tabula Muris project (Tabula Muris Consortium et al., 2018) for seven non-myeloid cell types in the mouse brain and identify the intron excision events that are differential used across these cell types.

Introduction

The process of mRNA alternative splicing (AS) plays pivotal roles in various cellular functions, including differentiation, protein diversity, and regulation of gene expression (Han et al., 2013; Nilsen & Graveley, 2010; Salz, 2011). To better understand this process, a precise definition and quantification of AS events across conditions and cell-types is critical. However, this is a challenge with short-read RNA-seq data.

To address this, researchers have devised various methods, such as rMATS, DEXSeq, and Leafcutter (Anders et al., 2012; Li et al., 2018; Shen et al., 2014). While the majority of these methods rely only on reads aligned to local exons or junctions, they perform adequately with high-coverage bulk RNA-seq data. However, their performance diminishes due to a lack of sufficient supporting reads at low coverage, especially in single-cell data. Additionally, as the number of supporting reads decreases, the results become more susceptible to fluctuations from technical noise, making their accuracy questionable. For instance, rMATs and MISO demonstrated limited discovery capability when tested on single-cell data from human HCT166 cells (Huang & Sanguinetti, 2017).

While there are methods like BRIE2 and SingleSplice, specifically designed for single-cell data, they come with certain drawbacks (Huang & Sanguinetti, 2021; Welch et al., 2016). For instance, BRIE2 requires a predefined set of splicing events, and SingleSplice requires spike-in transcripts to quantify the levels of technical noise (Huang & Sanguinetti, 2021; Wen et al., 2020). In addition, they inherit from bulk methodology the limitation of only using a subset of informative reads. Thus, detecting differential splicing events in low-coverage data remains a challenge.

It would be valuable to augment the AS quantification capabilities of existing tools that excel with high-coverage bulk RNA-seq, adapting them to low-coverage data. Among available methods, Leafcutter stands out as a prime candidate for this enhancement, given its flexible representation of complex AS events and the benefit of not requiring annotations for splicing events (Li et al., 2018). In terms of enhancing detection in low-coverage data, SUPPA introduced a compelling concept: instead of relying on exon or split reads, it utilizes transcripts per million (TPM) values from pseudoalignment methods for the isoforms that support each splicing outcome in an event (Alamancos et al., 2015). This strategy enables the utilization of information from more distal parts of the isoform pertaining to a splicing event. As it is based on isoforms, this approach can harness more reads to support a specific splicing event than methods relying only on local exons and split reads. This increased read support becomes particularly crucial when coverage is limited. However, it is important to note that the effectiveness of this method hinges on the accuracy of isoform abundance estimation and therefore the comprehensiveness of the transcriptome annotation. The recent advent of pseudoalignment tools (e.g., Salmon and Kallisto) and transcriptome assemblers (e.g., StringTie2) renders this approach more feasible than before (Bray et al., 2016; Kovaka et al., 2019; Patro et al., 2017).

In this context, we propose a refined version of Leafcutter that employs pseudoalignment-based TPM values and expected count for isoforms to quantify introns, which we refer to as LeafcutterITI (Isoform to Intron). Then, we demonstrate the enhanced splicing event detection capacity of LeafcutterITI and accuracy compared with Leafcutter through simulation and pseudobulk data generated by Smart-seq2. Lastly, we utilize LeafcutterITI to analyze non-myeloid cell types in the mouse brain and identify cell type-specific intron excision events using data from the Tabula Muris project (Tabula Muris Consortium et al., 2018).

Materials & Methods

Simulation and Validation

Simulated RNA-sequencing data in FASTA format was generated using Polyester, based on the human transcriptome Release 44 from GENCODE (Frankish et al., 2021; Frazee et al., 2015). Only genes with more than two annotated isoforms were selected for this simulation. Four different coverage levels — 5X, 20X, 50X, and 100X — were tested. Pair-end reads, with a read length of 100 base pairs (bp) and an expected fragment length of 250 base pairs, were generated following the default settings of Polyester. The expected number of reads for isoform *j* in gene *i* was calculated using the following formula and then input into Polyester:

$$Exp\left(Read_{ij}\right) = P_{ij} * coverage * \frac{isoform\ len_{ij}}{read\ len}$$

$$\sum_{j} P_{ij} = 1$$

where P_{ij} is the proportion of gene abundance that is allocated to isoform j of gene i.

Adapted from Patro et al. (2017), for gene *i*, the proportion of abundance allocated to isoforms will be generated based on the following rules: either (i) split according to a flat Dirichlet distribution (α = (1,...,1)) (i.e., uniform) or (ii) attributed to a single isoform. Alignment for FASTA files was done with STAR to generate bam files for Leafcutter (Dobin et al., 2013).

Two distinct sets of simulations were generated based on these rules that will be denoted as simulation 1 and simulation 2 and serve different validation proposes.

In Simulation 1, for all genes, the proportion of abundance allocated to isoforms will be generated based on a flat Dirichlet Distribution (α = (1,...,1)). Four replicates for each coverage

level (5X, 20X, 50X, 100X) will be produced. Both Leafcutter and LeafcutterITI will be applied to simulated data to generate intron clusters. The minimum read cutoff for Leafcutter clusters is set at 20, 20, 50, and 50 for each respective coverage level. For LeafcutterITI, the TPM cutoff for cluster inclusion is 5, regardless of coverage level. Δ PSI for introns within these clusters will be calculated as the difference between the true PSI value and the average PSI value of four replicates calculated by the pipeline, providing a measure of the overall accuracy of PSI estimation by Leafcutter and LeafcutterITI. The influence of incomplete transcriptome reference on LeafcutterITI is also tested. An incomplete transcriptome reference was generated by the following rule: when the abundance of an isoform is less than 10% of that for the gene, there is a 30% chance that it will be missing in the reference transcriptome. LeafcutterITI run with TPM value in this round of simulation.

Simulation 2 involves a control group and a test group, each comprising five samples per coverage level. For the control group, the proportion of gene abundance allocated to isoforms for all genes is determined by a flat Dirichlet Distribution ($\alpha = (1,...,1)$). In the test group, for a given gene i, there is an 80% probability that the proportion of abundance allocated to isoforms will match the control and a 20% chance that abundance will be allocated to a single isoform.

Any intron cluster mapped to a gene exhibiting differential abundance allocation between control and test groups is considered a differential usage cluster. Conversely, if the allocation is identical, the null hypothesis holds. Both Leafcutter and LeafcutterITI were applied to simulated data to generate intron clusters. LeafcutterITI used normalized counts instead of TPM (details below). The minimum read cutoff for Leafcutter and LeafcutterITI clusters is set at 20, 20, 50, and 50 for each respective coverage level (5X, 20X, 50X, 100X). For Leafcutter_ds, the differential intron excision events detection method for Leafcutter, the minimum read count

cutoff for each group across these coverage levels is 5, 20, 50, and 50, respectively, with a minimum of three samples per intron for all tests (Li et al., 2018).

Performance was evaluated using receiver operating characteristic (ROC) and Presicion-recall (PR) curves, considering only testable clusters. The model score was calculated as 1 – p_BH_corrected, where p_BH_corrected represents p-values adjusted to control the False Discovery Rate using the Benjamini-Hochberg (BH) correction (Benjamini & Hochberg, 1995). Permutations were conducted at the intron level, randomly shuffling labels (control vs test) for each intron. The distribution of p-values was visually represented using a quantile-quantile (QQ) plot on a -log10 scale.

Implementation of the pseudoalignment-based version of Leafcutter (LeafcutterITI)

LeafcutterITI was implemented as a Python-based pipeline. For LeafcutterITI, the introns were quantified as follows: **1**) isoforms to intron excision events map will be generated based on GENCODE reference (Frankish et al., 2021), **2**) TPM for isoforms were generated by Salmon (Patro et al., 2017) **3**) each intron was quantified by summing the TPM or normalized count for isoforms that include this intron, **4**) intron clustering and filtering **5**) downstream differential splicing event usage detection between conditions.

Intron Clustering

LeafcutterITI can use a slightly different intron clustering procedure than Leafcutter. For Leafcutter, an intron excision event in a cluster satisfies: **1)** overlap with at least one other intron

excision event in this cluster, **2)** have a shared intron splice site with at least one other intron excision event in this cluster (Li et al., 2018). In LeafcutterITI, an intron excision event in a cluster satisfies: **1)** overlap with at least one other intron excision event in this cluster, **2)** have a shared intron splice site with at least one other intron excision event in this cluster, or **3)** connected to an exon that another intron excision event in this cluster is also connected to. Where not otherwise specified, LeafcutterITI uses this alternative intron clustering procedure. LeafcutterITI can also use the same intron clustering procedure as Leafcutter.

Statistical Test

Two different statistical tests were used to detect differential splicing event usage between conditions. The first test method was the direct use of Leafcutter_ds, which assumes the counts are Dirichlet-multinomial distributed (Li et al., 2018). The second method is based on the assumption that the percent spliced in (PSI) values follow a Dirichlet distribution. This method treats the PSI values of an intron cluster for a sample as a J-element probability vector, where J equals the number of introns in that cluster. Then, testing whether these PSI values from two conditions originate from different Dirichlet distributions using a log-likelihood ratio test. Additionally, we applied an optional “bio significance” filter to exclude significant clusters whose PSI values come from two distinct Dirichlet distributions but have a close expected value (i.e. small effect size). Where not otherwise specified, a 5% FDR threshold was used.

Normalization method

As the TPM does not reflect the actual number of reads that support an isoform, a normalization method that accounts for the TPM ratio and actual read count is beneficial. The normalized count for gene i, isoform j was computed as

$$Normalized\ Count_{ij} = (\sum_{j} Count_{ij}) \frac{TPM_{ij}}{\sum_{j} TPM_{ij}}$$

These normalized counts will be the default input for Leafcutter (alternative input will be the TPM directly). The intron count will be computed as the sum of normalized count or TPM of isoforms that contain this intron by LeacutterITI. Where not otherwise specified, normalized count was used.

Tabula Muris data analysis

Three distinct datasets were generated from Tabula Muris pseudobulk data to serve different purposes (Tabula Muris Consortium et al., 2018).

1. Five brain non-myeloid oligodendrocyte and five heart fibroblast pseudobulk samples were generated by randomly extracting 100 cell barcodes without replacement from the brain non-myeloid oligodendrocyte and heart fibroblast pseudobulk data from Tabula Muris. This dataset aimed to assay the performance of LeafcutterITI in high-coverage pseudobulk datasets similar to bulk data.

2. Five brain non-myeloid oligodendrocyte and five heart fibroblast pseudobulk samples were generated by randomly extracting ten cell barcodes without replacement from the brain non-myeloid oligodendrocyte and heart fibroblast pseudobulk data from Tabula

Muris. This dataset aimed to assay the performance of LeafcutterITI in low-coverage pseudobulk datasets.

3. Seven non-myeloid cell types in the mouse brain were selected from Tabula Muris pseudobulk data. As oligodendrocytes and endothelial have a much greater number of cell barcodes than other cell types, five oligodendrocytes and five endothelial samples were generated by randomly extracting 100 cell barcodes without replacement. Five samples each of astrocytes, neurons, oligodendrocyte precursors, pericytes, and Bergmann glial were generated by randomly dividing the barcodes in the Tabula Muris data for the respective cell types. This dataset aimed to discover differential intron excision events between cell types in the mouse brain.

Data Analysis and Visualization

Data analyses were performed with Python 3.10.11 using Jupyter Notebook. Graphs were made using Matplotlib 3.7.1 and Seaborn 0.12.2 (Hunter, 2007; Waskom, 2021).

Results

Overview of LeafcutterITI

To address the problem of limited split reads supporting the existence of certain intron excision events in low-coverage data, we utilize pseudoalignment to a reference transcriptome to infer and count intron excision events based on the abundance of isoforms consistent with these events. We refer to this as the 'isoform to intron' approach, abbreviated as ITI.

LeafcutterITI is a modified version of Leafcutter that employs the ITI approach to quantify intron excision events while retaining Leafcutter's intron clustering algorithm (Li et al., 2018). Since TPM does not directly correspond to the actual number of reads mapped to an isoform, LeafcutterITI uses normalized count-like values to enable count-based modeling. These are generated by multiplying the expected gene count by the TPM ratio of isoforms for a gene, reflecting both the ratio of the isoform and the total reads supporting the gene (Methods).

The LeafcutterITI procedure involves the following steps: 1) Generating an isoform to intron excision events map, providing information about the intron excision events each isoform supports. This step is only required once for each transcriptome reference. 2) Quantifying isoforms using Salmon (Patro et al., 2017). 3) Obtaining normalized count-like values for each isoform. 4) Computing the abundance of intron excision events by mapping isoform abundance to intron excision abundance. 5) Filtering and clustering introns. 6) Detecting differential splicing event usage between conditions (e.g., cell types) (Figure 1). Unless specified otherwise, all references to LeafcutterITI in this thesis follow these six steps. Two differential splicing event detection methods are available for LeafcutterITI: the Leafcutter_ds from Leafcutter, which uses a Dirichlet-multinomial generalized linear model, and a method based on the assumption that percent spliced in (PSI) values for each intron cluster follow a Dirichlet distribution, referred to as differential Dirichlet tests. The differential Dirichlet tests also include an optional “bio significance” filter to ensure the PSI values for the intron clusters in different conditions are different by at least 10% for at least one intron. The default test method is Leafcutter_ds in this thesis.

With the similar definition and clustering method for intron clusters, LeafcutterITI is compatible with all downstream analysis methods of Leafcutter, including LeafViz visualization (Li et al., 2018).

Model performance assessment with simulations

We performed two distinct rounds of simulation, as described in the methods section, to compare the overall performance of LeafcutterITI with Leafcutter. This comparison aimed to demonstrate an enhancement in detection capability in low-coverage data without sacrificing accuracy. In the first simulation, we compared the difference (ΔPSI) between the actual PSI and the estimated PSI from Leafcutter and LeafcutterITI for intron clusters. We also tested the influence of an incomplete transcriptome reference on LeafcutterITI's performance. Our observations indicated that LeafcutterITI can detect more intron clusters than Leafcutter, especially at low coverage levels, and is less sensitive to changes in coverage level (Figure 2). As the coverage level increases, both LeafcutterITI and Leafcutter provide more precise estimations of intron PSI values (Figure 3 & 4). Additionally, LeafcutterITI showed more accurately estimated intron PSI values than Leafcutter across all coverage levels. With an incomplete reference, LeafcutterITI exhibited a slight reduction in detectable intron clusters and small influence on the overall PSI estimation accuracy (Figures 2 - 4).

In the second simulation, we generated five samples for each of the control and test groups, with some genes showing differential splicing event usage (methods). We ran both Leafcutter_ds and different Dirichlet tests on the clusters. The model performances were analyzed using receiver operating characteristic (ROC) curves and precision-recall (PR) curves. The area under the curve (AUC) was used as a numeric assessment of model performance. LeafcutterITI was capable of detecting and testing more intron clusters than Leafcutter at low

coverage levels and found more significant clusters across all coverage levels (Figure 5). LeafcutterITI and Leafcutter detected approximately 26,000 intron clusters and tested most of them at 20X, 50X, and 100X coverage levels. While Leafcutter detected only 14,246 clusters, LeafcutterITI found a similar number of clusters to those in other coverage levels at 5X coverage. The deviation of p-values for samples with real labels and permuted labels showed little inflation in the statistics using Leafcutter_ds for 5X and 100X coverage levels in both LeafcutterITI and Leafcutter (Figure 6). Similar results were obtained for 20X and 50X coverage levels (results not shown).

We compared the ROC and PR curves for LeafcutterITI using Leacutter_ds and the differential Dirichlet test with Leafcutter. LeafcutterITI performed similarly to Leafcutter across all coverage levels, except it was slightly less accurate at 5X coverage (Figures 7 and 8). In terms of detection methods used, Leafcutter_ds performed better than the differential Dirichlet test without the bio-significance filter and similarly to the differential Dirichlet test with the bio-significant filter at 5X coverage. Interestingly, Leafcutter_ds performed similarly to the differential Dirichlet test without the bio-significant filter and better than the different Dirichlet test with the bio-significant filter at 100X coverage. The different Dirichlet tests lost some accuracy with the bio-significant filter at 100X coverage. With the bio-significance filter, LeafcutterITI used the differential Dirichlet test found fewer significant clusters than using Leafcutter_ds (Figure 5 and 9). Moreover, the detection power of the differential Dirichlet test decreased as coverage increased, likely because it is unaware of overall differences in coverage. Thus, we decided to use Leafcutter_ds for all other analyses in this thesis.

Model performance assessment with real data

To assess model performance with real sequencing data, we downloaded pseudobulk data generated using Smart-seq 2 for brain non-myeloid oligodendrocytes and heart fibroblasts from the Tabula Muris project (Tabula Muris Consortium et al., 2018). To simulate low and high-coverage pseudobulk data, we created two groups of samples: one group contained 10 cell barcodes per sample, and the other contained 100 cell barcodes per sample. We divided the pseudobulk data into five pseudobulk samples for each cell type and for each group (methods section). Subsequently, we compared the differential intron cluster usage between brain non-myeloid oligodendrocytes and heart fibroblasts.

LeafcutterITI demonstrated superior performance in detecting more clusters and identifying more significant clusters in samples with 10 cells compared to Leafcutter, and they performed similarly in samples with 100 cells (Figure 10). At the 10-cell level, LeafcutterITI identified 4,748 testable clusters and 750 significant clusters, while Leafcutter identified only 2,055 testable clusters and 269 significant clusters. Although the ratio of testable clusters to significant clusters was similar, LeafcutterITI detected more than three times as many significant clusters as Leafcutter. At the 100-cell level, LeafcutterITI identified slightly fewer clusters than Leafcutter but detected more testable and significant clusters. We also assessed the deviation of p-values for samples with real labels and permuted labels. We observed a slight inflation of p-value in results with permuted labels for both LeafcutterITI and Leafcutter (Figure 11). Still, they are distinguishable from the true label and mostly corrected after FDR control (e.g. at the 10-cell level, LeafcutterITI yields 750 significant clusters with true labels and only 23 significant clusters with permuted labels).

Unique Intron Excision Pattern Identification in Mouse Brain

To evaluate the suitability of LeafcutterITI for detecting differential splicing and identifying unique intron excision patterns between cell types, we analyzed seven mouse brain cell types: oligodendrocytes, endothelial cells, astrocytes, neurons, oligodendrocyte precursors, pericytes, and Bergmann glial cells. We performed pairwise comparisons of differential intron cluster usage between these cell types using pseudobulk samples generated as described in the methods. The samples were obtained from the Tabula Muris project and processed as described in the methods section (Tabula Muris Consortium et al., 2018).

Applying a 10% False Discovery Rate, we discovered that out of 20,022 clusters, 8,411 clusters are differentially used in at least one pair of cell types. These clusters contain at least one intron excision event with a PSI value difference of greater than 0.2 and an absolute effect size greater than 1.5 in the cell type pairs where the cluster is significant. Moreover, we identified 17,057 intron excision events out of 58,781 total intron excision events that exhibited a PSI value difference of at least 0.2 and an absolute effect size greater than 1.5 in at least one pair of cell types. We successfully identified cell type-specific intron excision patterns for these cell types (Figure 12). Additionally, we observed that pseudobulk samples for the same cell type exhibited more similar intron excision patterns compared to samples from different cell types.

Discussion

In conclusion, we confirm the accuracy and usefulness of LeafcutterITI. Our studies demonstrate that LeafcutterITI can detect more intron clusters in low-coverage data compared to Leafcutter, while maintaining a similar level of accuracy in high-coverage data, as shown through both simulation and sequencing data analysis. Although LeafcutterITI exhibits slightly

lower accuracy than Leafcutter in low-coverage data, its superior PSI estimation and a greater number of testable clusters enhance our ability to identify differential splicing events in such data (Figures 2 - 8). Additionally, we demonstrate that an incomplete transcriptome does not significantly deteriorate LeafcutterITI's performance, likely due to the redundancy in transcriptome annotation (Figures 2 - 4).

We have also identified specific intron excision patterns for different cell types and confirmed that samples from the same cell type exhibit more similar intron excision patterns compared to those from different cell types (Figure 12). These findings underscore the potential of clustering single cells based on intron excision patterns. Moreover, opens up the possibility of constructing a database that documents the maker intron excision events for cell types. This approach may be able for researchers to obtain a finer resolution for cell types.

Although LeafcutterITI shows promising performance on pseudobulk data generated by Smart-seq2, its effectiveness with other full transcript coverage single-cell RNA sequencing techniques, such as SPLiT-seq, should be explored (Rosenberg et al., 2018). We note that LeafcutterITI’s accuracy greatly depends on the completeness of the transcriptome and the precision of isoform quantification. While pseudoalignment methods provide relatively accurate isoform quantification, the incompleteness of the transcriptome, particularly in disease samples and non-canonical organisms, could limit the accuracy and detection capacity of LeafcutterITI. We plan to use transcriptome augmentation methods, such as Stringtie2, in conjunction with long-read RNA sequencing data to generate a more complete transcriptome and enable LeafcutterITI to detect novel intron excision events.

There is further potential for the utilization of LeafcutterITI. For instance, Leafcutter cannot capture alternative first and last exon usage events due to the absence of intron excision

events to indicate their presence. However, by manually creating virtual introns upstream and downstream of the isoform, we can capture these events with LeafcutterITI. Knowing which intron excision events are associated with which isoforms, and with transcriptome annotations providing information about isoform types (e.g., protein-coding and lincRNA), we can trace back the possible RNA type that the intron clusters contribute to, offering a novel approach to considering differential splicing events.

Ultimately, the goal for LeafcutterITI is to be directly applied to single-cell data, rather than single-cell pseudobulk data. We will continue testing its application in single-cell data and likely develop new statistical analysis methods tailored for LeafcutterITI that account for the sparsity of single-cell data.

Acknowledgments

First of all, I thank Dr. Knowles for welcoming me to his lab. Without his selfless support and encouragement, it would have been impossible for me to finish this thesis. His insightful advice and feedback have significantly shaped my research approach and methodology. His mentoring and expertise in the field have profoundly impacted my career path and led me to find my real interest in computational biology.

I would like to extend my deepest gratitude to my thesis committee members for their time and suggestions. I would like to thank my lab colleagues. I am thankful for their excellent

lab meeting presentations, as they deepen my understanding of the field and inspire my research. I also enjoy the lunchtime before or after the lab meeting.

I would like to thank Dr. Fiszbein, who mentored my undergraduate honors thesis. In her lab, I practiced my research skills and learned to become an independent researcher.

I would like to thank the great courses taught by Dr. Knowles, Dr. Pe'er, Dr. Sims, Dr. Shen, and Dr. Zhang on subjects related to my thesis that provided me with the required knowledge to complete the work in this thesis. I would also like to thank those who have helped further my research journey.

Furthermore, I would like to thank all my friends for their friendship and support. Lastly, I am more than grateful to my family for their support that allowed me to grow up as an independent person and for their unconditional love.

Figures

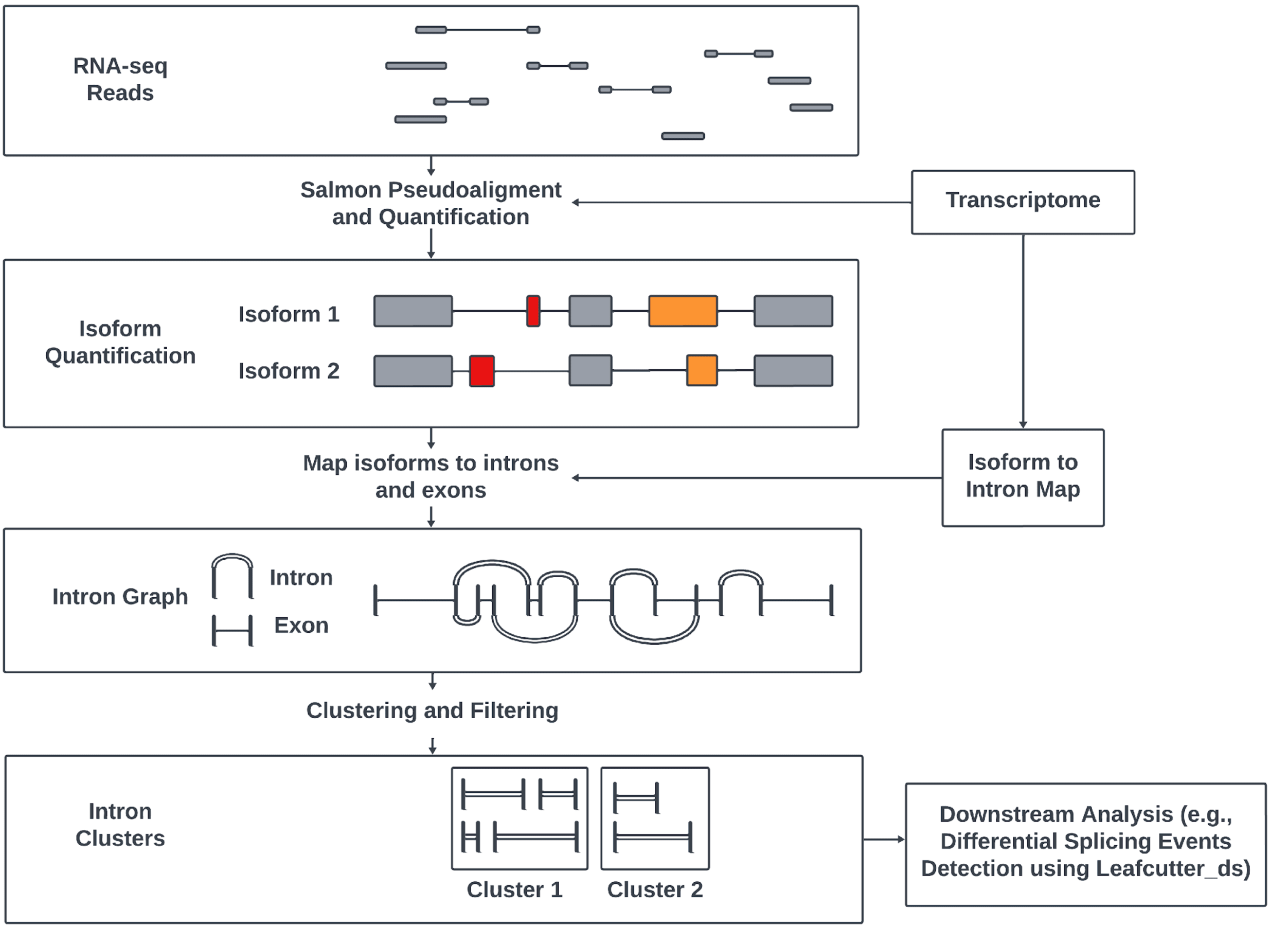

Figure 1. Overview of LeafcutterITI. LeafcutterITI utilizes transcriptome to compute isoform to intron map. Then, the count of intron excision events is computed by mapping isoform counts to intron excision events. Based on the intron excision events, LeafcutterITI generates the intron clusters. In this example, LeafcutterITI identifies two intron clusters.

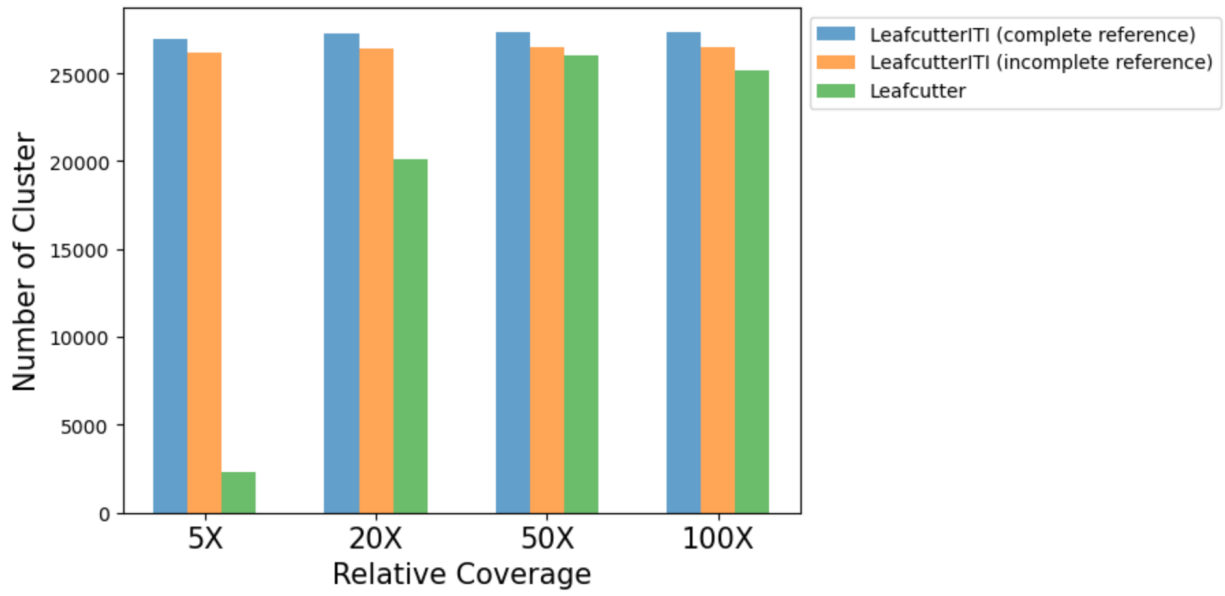

Figure 2. Number of clusters identified by LeafcutterITI and Leafcutter at different coverage levels in simulation 1.

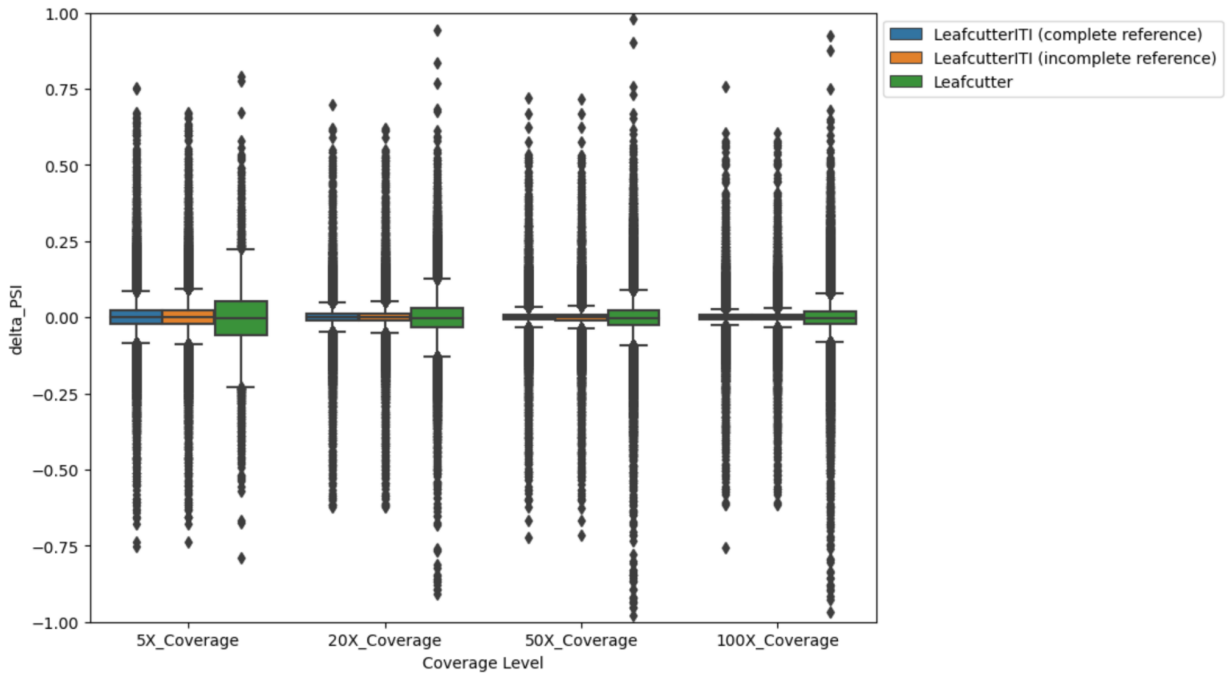

Figure 3. ΔPSI between true PSI value and the average PSI value of four replicates for intron clusters.

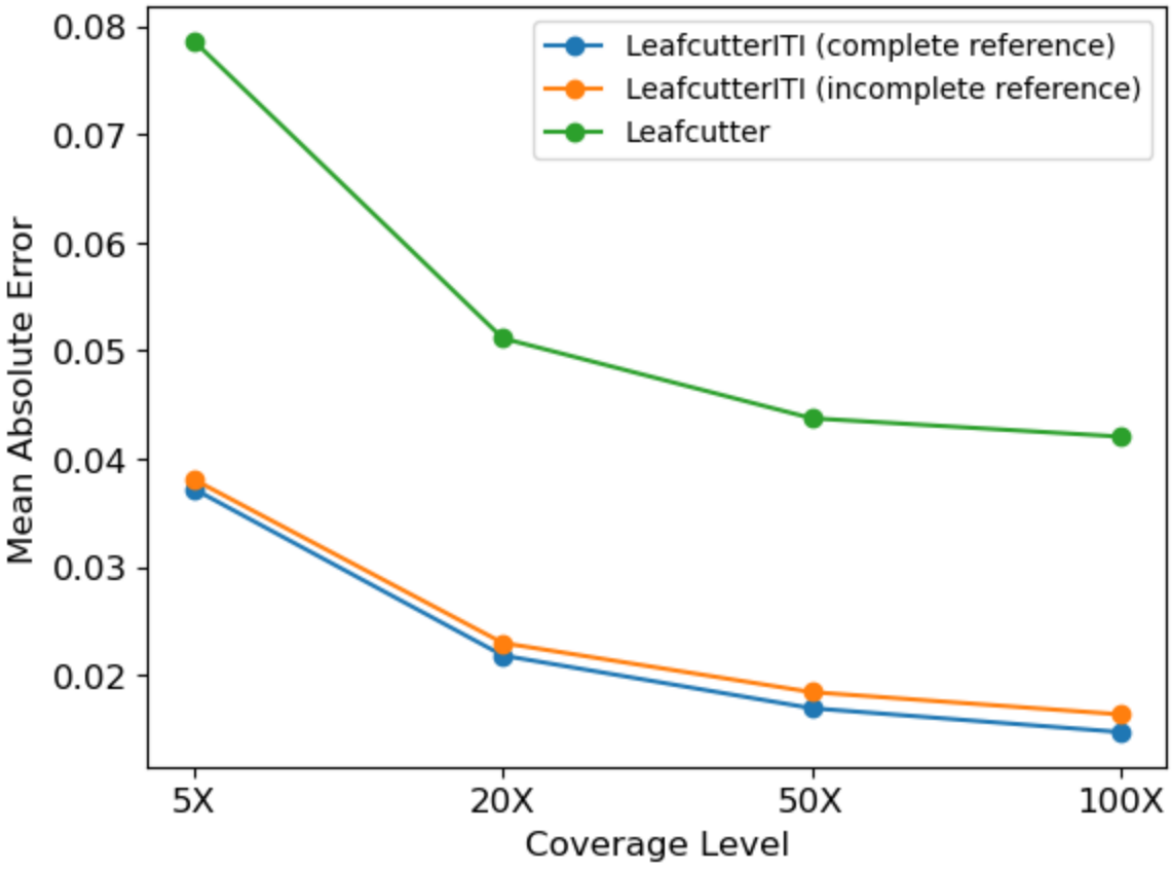

Figure 4. Mean Absolute Errors between different of true PSI values and the average PSI value of four replicates for intron clusters.

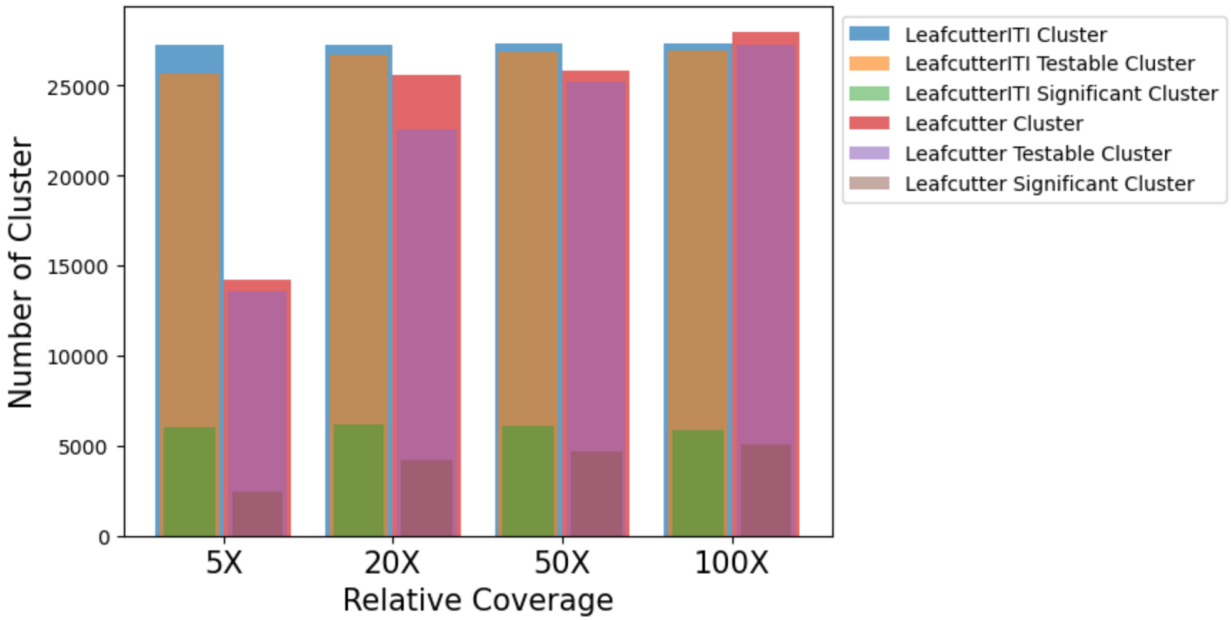

Figure 5. Number of clusters, testable clusters, and significant clusters by LeafcutterITI and Leafcutter in different coverage levels in simulation 2

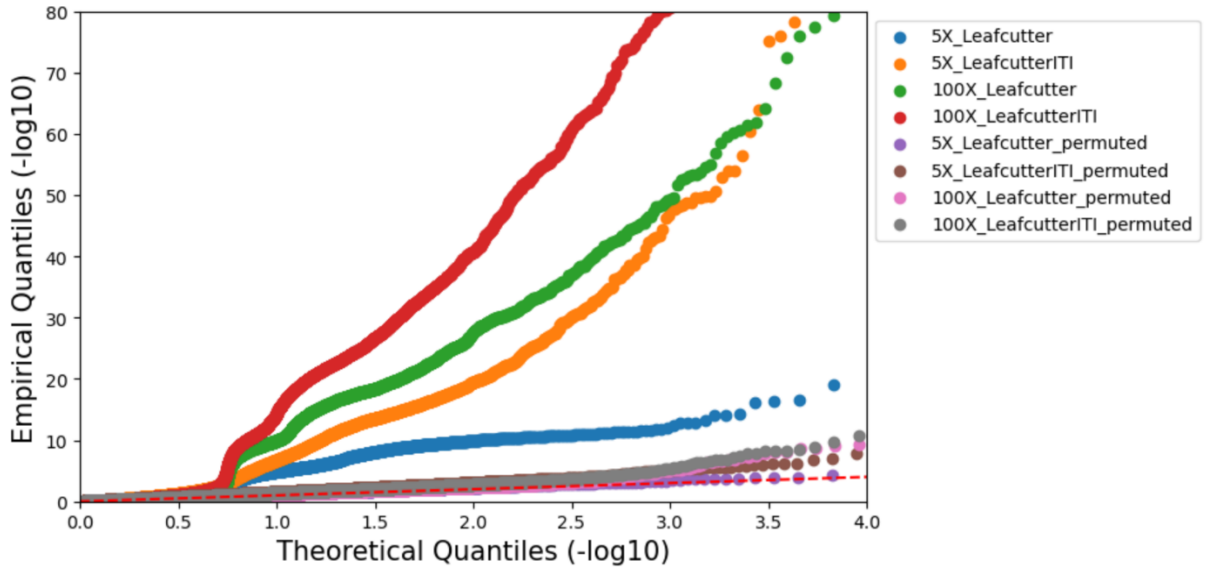

Figure 6. Empirical Quantiles and Theoretical Quantiles for p-values generated by Leacutter_ds for simulation 2. The red dotted line represents line X=Y.

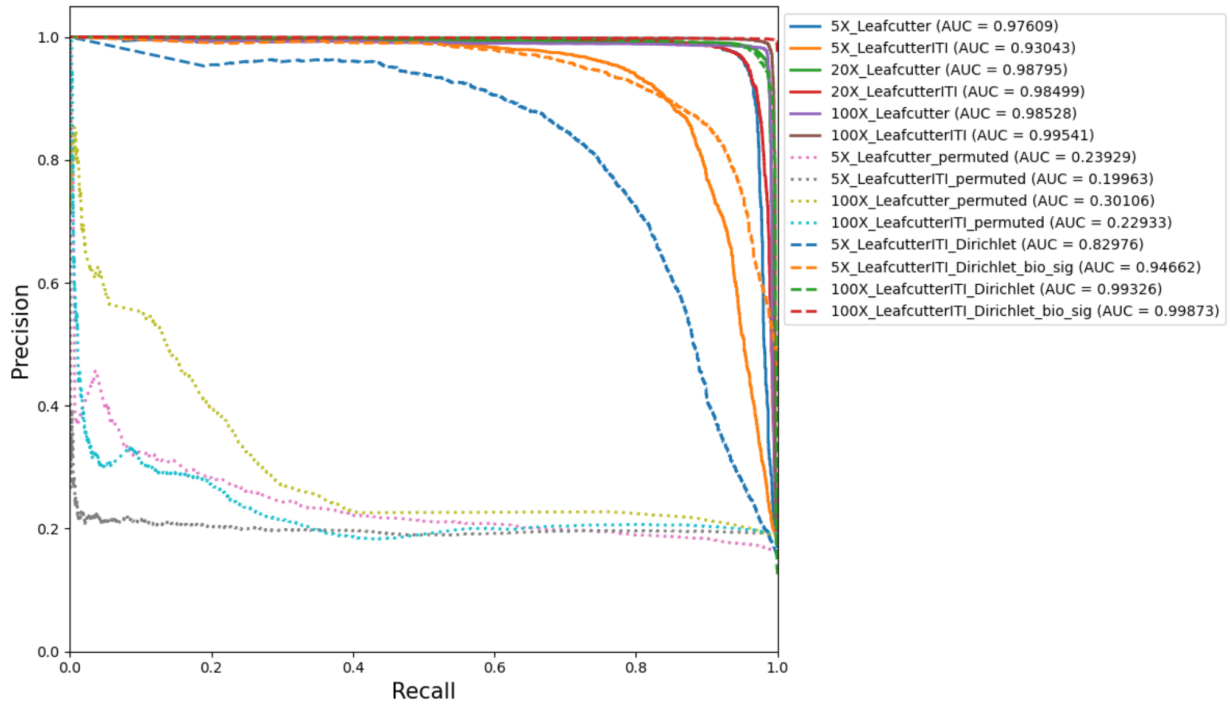

Figure 7. Precision and recall for LeafcutterITI and Leafcutter using leafcutter_ds and different Dirichlet tests for different coverage levels. Dirichlet in legend indicates the usage of the different Dirichlet test. 50X curves are not included due to their similarity to 20X and 100X curves.

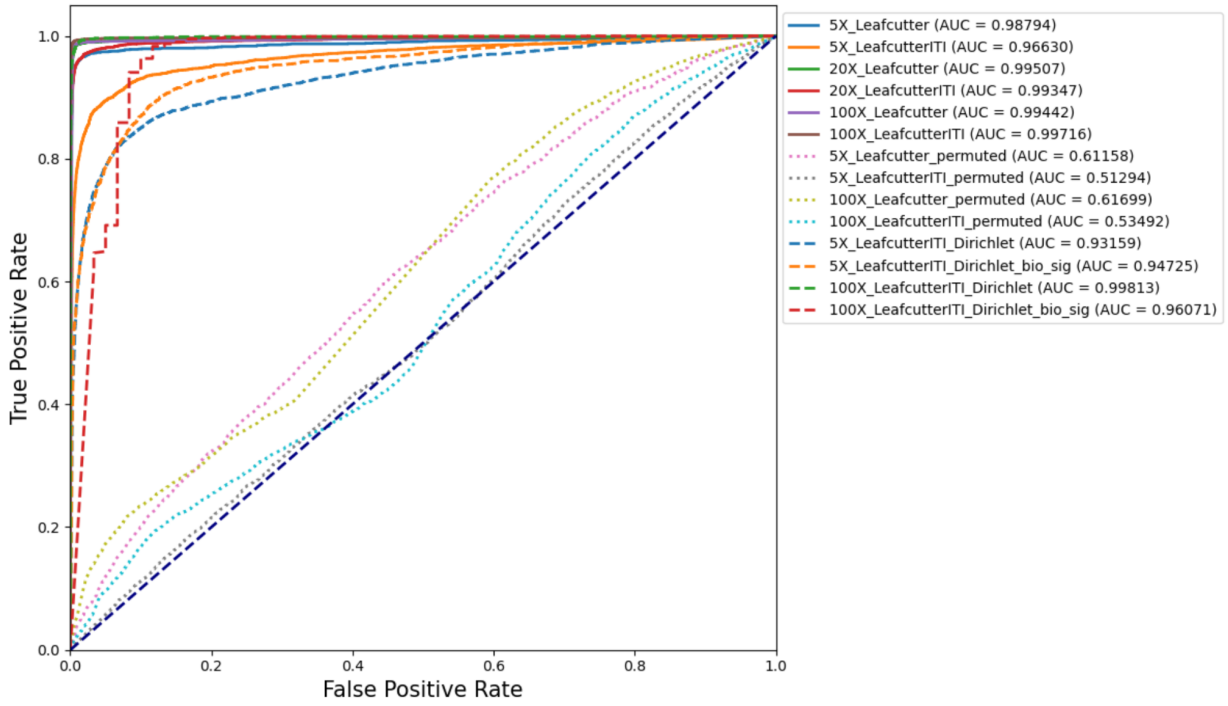

Figure 8. True positive rate and false positive rate for LeafcutterITI and Leafcutter using leafcutter_ds and different Dirichlet tests for different coverage levels. Dirichlet in legend indicates the usage of different Dirichlet tests. 50X curves are not included due to their similarity to 20X and 100X curves.

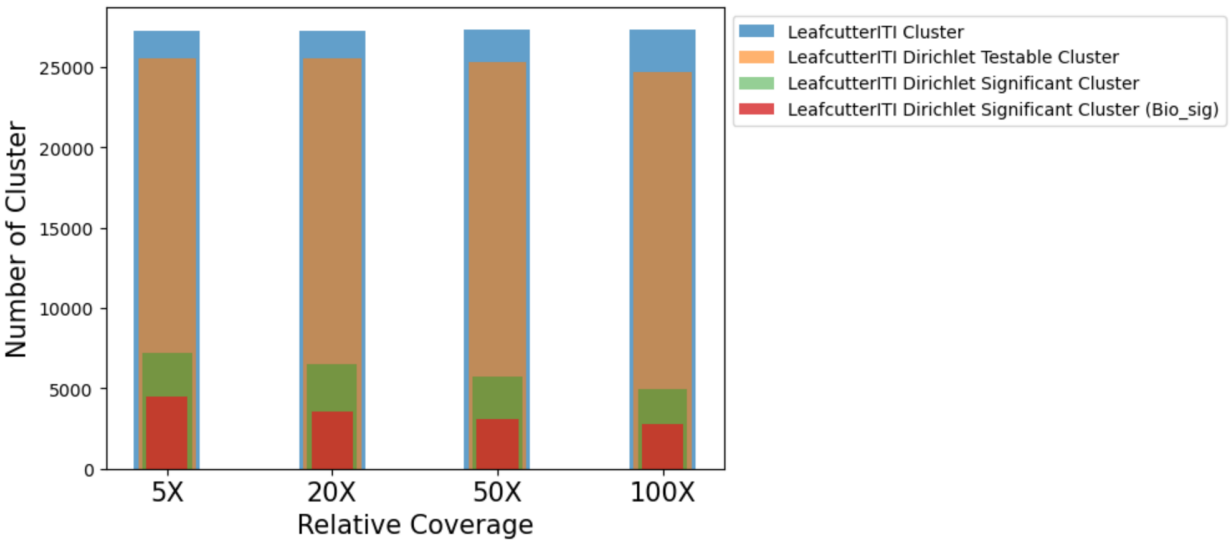

Figure 9. Number of clusters, testable clusters, and significant clusters by LeafcutterITI using different Dirichlet test in different coverage levels in simulation 2

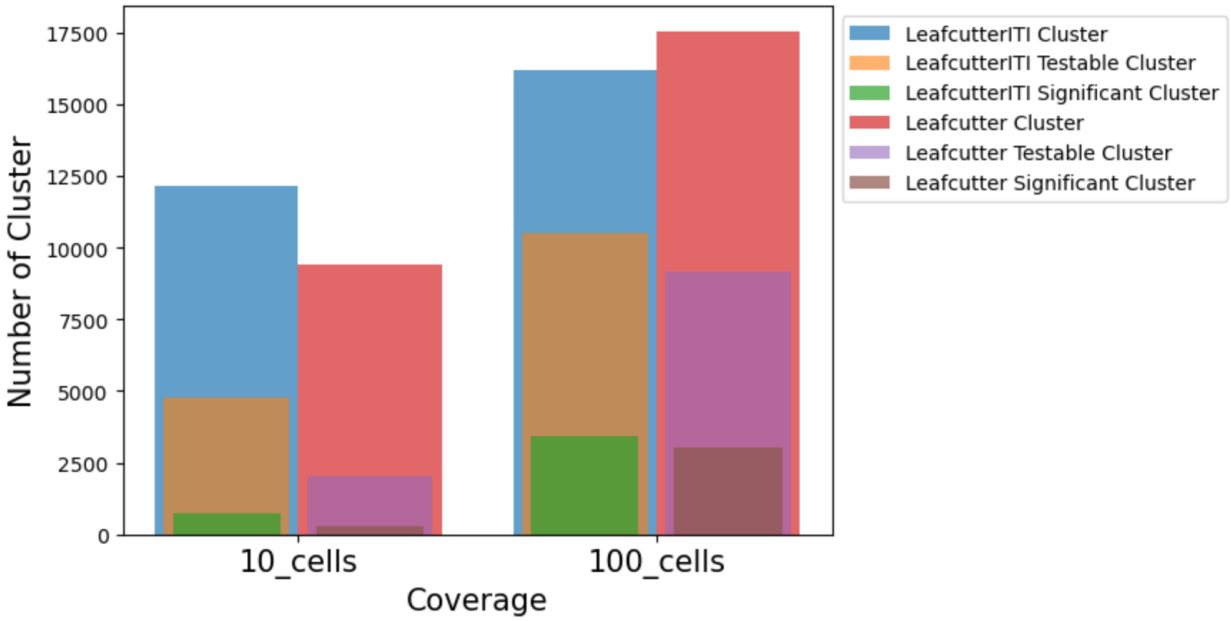

Figure 10. Number of clusters, testable clusters, and significant clusters by LeafcutterITI and Leafcutter in different coverage levels in brain non-myeloid oligodendrocyte cell versus heart fibroblast cell.

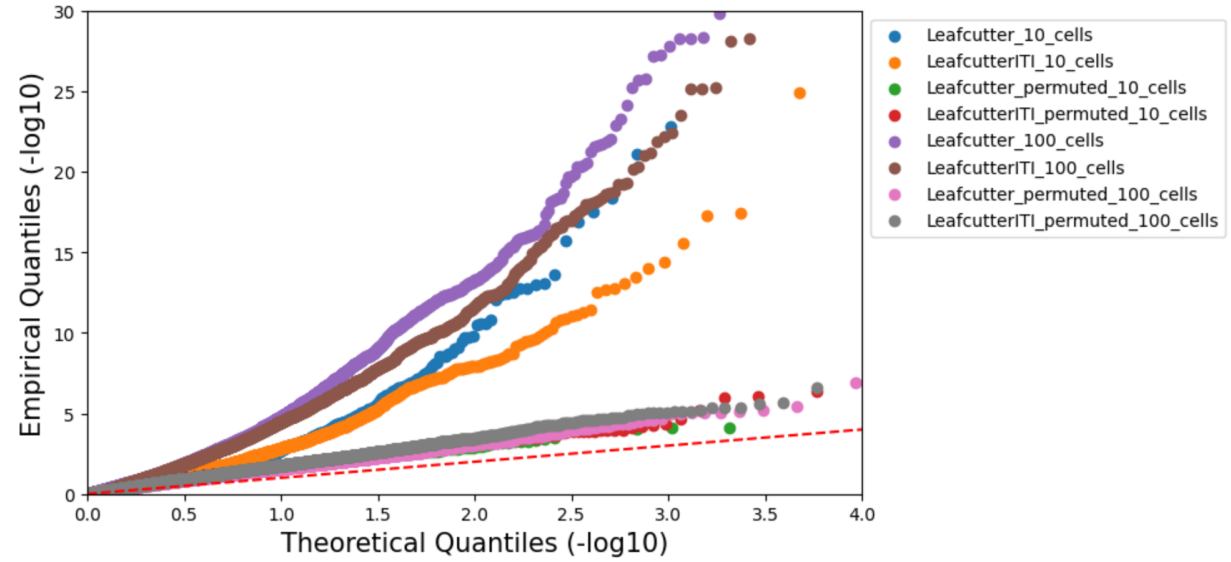

Figure 11. Empirical Quantiles and Theoretical Quantiles for p-values generated by Leacutter_ds for brain non-myeloid oligodendrocytes versus heart fibroblasts. The red dotted line represents line X=Y.

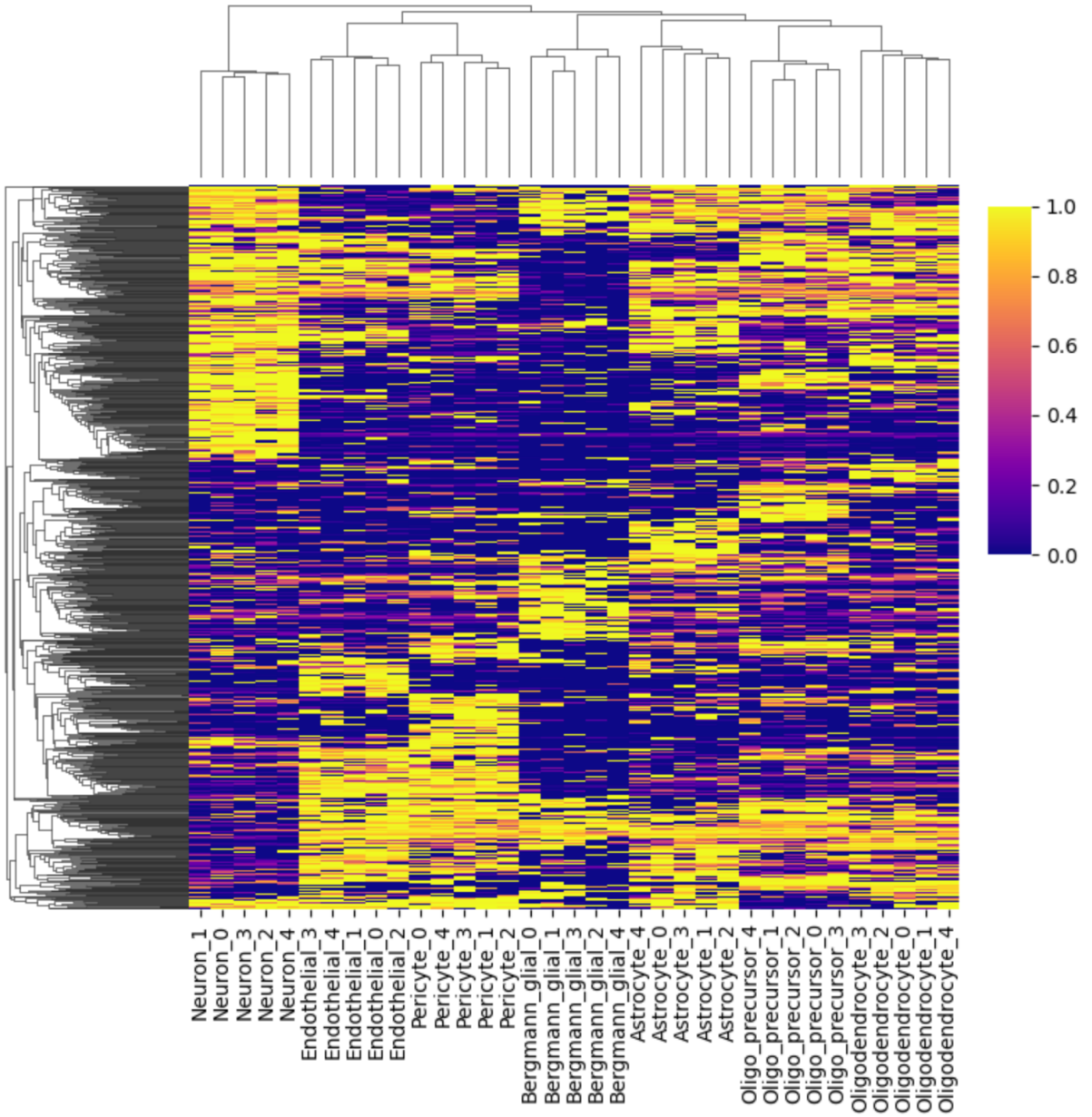

Figure 12. LeafcutterITI identified cell types of specific intron splicing events from mouse brain pseudobulk samples. Shown is a heatmap of the intron excision ratios of the top 1000 introns that were found to be differentially spliced in at least one cell type pair. Cell types examined included oligodendrocytes, endothelial, astrocytes, neurons, oligodendrocyte precursors, pericytes, and Bergmann glial cells. The color represents the PSI value for the intron.

References

Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N., & Eyras, E. (2015). Leveraging transcript quantification for fast computation of alternative splicing profiles. *RNA* , *21*(9), 1521–1531. https://doi.org/10.1261/rna.051557.115

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research*, *22*(10), 2008–2017. https://doi.org/10.1101/gr.133744.111

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, *34*(5), 525–527. https://doi.org/10.1038/nbt.3519

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* , *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Frankish, A., Diekhans, M., Jungreis, I., Lagarde, J., Loveland, J. E., Mudge, J. M., Sisu, C., Wright, J. C., Armstrong, J., Barnes, I., Berry, A., Bignell, A., Boix, C., Carbonell Sala, S., Cunningham, F., Di Domenico, T., Donaldson, S., Fiddes, I. T., García Girón, C., … Flicek, P. (2021). GENCODE 2021. *Nucleic Acids Research*, *49*(D1), D916–D923. https://doi.org/10.1093/nar/gkaa1087

Frazee, A. C., Jaffe, A. E., Langmead, B., & Leek, J. T. (2015). Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics* , *31*(17), 2778–2784. https://doi.org/10.1093/bioinformatics/btv272

Han, H., Irimia, M., Ross, P. J., Sung, H.-K., Alipanahi, B., David, L., Golipour, A., Gabut, M., Michael, I. P., Nachman, E. N., Wang, E., Trcka, D., Thompson, T., O'Hanlon, D., Slobodeniuc, V., Barbosa-Morais, N. L., Burge, C. B., Moffat, J., Frey, B. J., … Blencowe, B. J. (2013). MBNL proteins repress ES-cell-specific alternative splicing and reprogramming. *Nature*, *498*(7453), 241–245. https://doi.org/10.1038/nature12270

Huang, Y., & Sanguinetti, G. (2017). BRIE: transcriptome-wide splicing quantification in single cells. *Genome Biology*, *18*(1), 123. https://doi.org/10.1186/s13059-017-1248-5

Huang, Y., & Sanguinetti, G. (2021). BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome Biology*, *22*(1), 251. https://doi.org/10.1186/s13059-021-02461-5

Hunter. (2007). *Matplotlib: A 2D Graphics Environment*. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55

Kovaka, S., Zimin, A. V., Pertea, G. M., Razaghi, R., Salzberg, S. L., & Pertea, M. (2019). Transcriptome assembly from long-read RNA-seq alignments with StringTie2. *Genome Biology*, *20*(1), 278. https://doi.org/10.1186/s13059-019-1910-1

Li, Y. I., Knowles, D. A., Humphrey, J., Barbeira, A. N., Dickinson, S. P., Im, H. K., & Pritchard, J. K. (2018). Annotation-free quantification of RNA splicing using LeafCutter. *Nature Genetics*, *50*(1), 151–158. https://doi.org/10.1038/s41588-017-0004-9

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*(7280), 457–463. https://doi.org/10.1038/nature08909

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Rosenberg, A. B., Roco, C. M., Muscat, R. A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L. T., Peeler, D. J., Mukherjee, S., Chen, W., Pun, S. H., Sellers, D. L., Tasic, B., & Seelig, G. (2018). Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, *360*(6385), 176–182. https://doi.org/10.1126/science.aam8999

Salz, H. K. (2011). Sex determination in insects: a binary decision based on alternative splicing. *Current Opinion in Genetics & Development*, *21*(4), 395–400. https://doi.org/10.1016/j.gde.2011.03.001

Shen, S., Park, J. W., Lu, Z.-X., Lin, L., Henry, M. D., Wu, Y. N., Zhou, Q., & Xing, Y. (2014). rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(51), E5593-601. https://doi.org/10.1073/pnas.1419161111

Tabula Muris Consortium, Overall coordination, Logistical coordination, Organ collection and processing, Library preparation and sequencing, Computational data analysis, Cell type annotation, Writing group, Supplemental text writing group, & Principal investigators. (2018). Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, *562*(7727), 367–372. https://doi.org/10.1038/s41586-018-0590-4

Waskom, M. (2021). seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. https://doi.org/10.21105/joss.03021

Welch, J. D., Hu, Y., & Prins, J. F. (2016). Robust detection of alternative splicing in a population of single cells. *Nucleic Acids Research*, *44*(8), e73. https://doi.org/10.1093/nar/gkv1525

Wen, W. X., Mead, A. J., & Thongjuea, S. (2020). Technological advances and computational approaches for alternative splicing analysis in single cells. *Computational and Structural Biotechnology Journal*, *18*, 332–343. https://doi.org/10.1016/j.csbj.2020.01.009