UNIVERSITY OF CALIFORNIA
SANTA CRUZ

# COMPARATIVE ANALYSIS OF LONG-READ TRANSCRIPTOME ASSEMBLY PIPELINES

A thesis is submitted in partial satisfaction
of the requirements for the degree of

MASTER OF SCIENCE

in

BIOMOLECULAR ENGINEERING & BIOINFORMATICS

by

Danilo Dubocanin

September 2020

The Thesis of Danilo Dubocanin
is approved:

_____
Professor Chris Vollmers, Chair

_____
Professor Russ Corbett-Detig

_____
Professor Ali Shariati

_____
Quentin Williams
Vice Provost and Dean of Graduate Studies

**TABLE OF CONTENTS**

## LIST OF FIGURES

## LIST OF TABLES

**Abstract**

**COMPARATIVE ANALYSIS OF LONG-READ TRANSCRIPTOME ASSEMBLY PIPELINES**

**Danilo Dubocanin**

Long-read sequencing can overcome some of the barriers in transcriptome assembly that plague short-read based technologies. Due to their short length, short-reads fail to span entire transcripts, and this leads to difficulties in discerning proper splice junctions. Conversely, long-read sequencing can span entire transcripts end-to-end, and thus can circumvent issues in inferring splice junctions. Multiple long-read transcriptome assembly pipelines have been developed in recent years but there is no comprehensive analysis comparing the various pipelines. Some of these pipelines implement novel approaches to generating transcriptomes using long-reads, while other pipelines adapted methods originally developed for short-read based transcriptome assembly. We show that there are significant differences in transcriptomes assembled on the same data, using different assembly pipelines. Our analysis further shows that high-level summary statistics can be misleading about transcriptome quality, as well as the importance of using internalized controls to validate transcriptome assemblies.

Dedication

To my father and mother, the most selfless and kind people I know.
You taught me to always be curious and the importance of hard work and sacrifice.

Acknowledgments

First, I'd like to acknowledge Professor Chris Vollmers for being an amazing mentor. He was always patient, and his excitement for science was palpable, and in turn it helped fuel my passion for science as well.

I'd also like to acknowledge Roger Volden for helping me on countless occasions. He is someone who I consider a mentor and someone I look up to as a bioinformatician.

## Introduction

RNA molecules are an integral part of the Central Dogma of Biology. The cumulative RNA of an individual is referred to as one's transcriptome. Understanding the transcriptomic landscape of an individual can lead to important insights into disease and answer fundamental biological questions. An important element of a thorough and complete transcriptome assembly is accurately identifying isoforms. The majority of protein-coding eukaryotic genes have multiple isoforms. Isoforms are generated through mechanisms of alternative splicing, which may rearrange introns and exons, or may generate isoform-specific introns and exons. Through alternative splicing, organisms are capable of expanding their protein library from a set number of genes. Alternative splicing explains why humans have ~3.5x more distinct proteins than we have distinct genes (Wang 2015). Many higher eukaryotes exhibit alternative splicing events. More than 95% of human genes have been proven to undergo alternative splicing (Wang 2015). Alternative splicing is a dynamic process that can change the transcriptomic landscape across time, tissue type, and within individuals. Aberrations in alternative splicing have been implicated in a variety of diseases, including Cancer and Alzheimer's Disease (Love 2015, Raj and Jager 2018, Li 2019). There is debate on whether a majority of alternative splicing is a product of noise, or whether it is a functionally relevant evolutionary driver of proteome diversity (Pickrell 2010, Rotival 2019). Regardless, alternative splicing plays an important role in shaping an individual's unique transcriptome, and understanding the isoforms expressed in an individual is a crucial element of understanding gene expression.

**Why use long reads for transcriptome assembly?**

RNA-sequencing has opened the door to understanding transcriptomic landscapes. RNA-sequencing consists of RNA extraction, mRNA enrichment or ribosomal RNA depletion, cDNA synthesis (this step may soon become obsolete given that ONT has recently demonstrated nascent RNA-seq), and preparation of an adaptor-ligated sequencing library (Kukurba and Montgomery 2015, Hrdlickova 2017). Up until recently, the state-of-the-art for sequencing was using Sequencing By Synthesis (SBS) short-read sequencing technologies, such as Illumina paired-end sequencing. Short-read technologies are highly accurate but suffer from a fundamental pitfall due to the length of the reads, typically <300 bases (Mantere 2019). Short-reads fail to resolve ambiguous sequences in our highly repetitive genome. They similarly also fail to solve ambiguous regions in transcriptomes, such as exons that are shared between multiple isoforms, and repetitive regions within mRNA molecules (Byrne 2017). Nevertheless, a sleuth of short-read transcriptome assemblers have been developed with varying degrees of success. All short-read assemblers can be categorized in two broad groups, *de novo* assemblers (genome agnostic) and reference guided assembler (Holzer and Marz 2019, Ungaro 2017). Transcriptome assembly using short-reads is essentially a graph reduction problem because a graph, for example, can represent the multiple relationships an exon can share with multiple transcripts (Lu 2013).

Ultimately, many of the issues with short-read transcriptome assemblers can be resolved using long-read technologies (Byrne 2017, Mantere 2019). Initially, the

issue with long-read sequencing was a high error rate, but as new methods, such as the use of Circular Consensus Sequencing (CCS) arose, an unprecedented opportunity for the assembly of high-accuracy transcriptomes arose with them. CCS methods to improve long-reads have been developed to be used on PacBio long-reads (SMRT-seq) as well as on Oxford Nanopore long-reads (R2C2) (Eid 2009, Ardui 2018, Volden 2018). Another issue with long-read sequencing for transcriptome assembly was the low throughput of first generation long-read sequencing pipelines. With recent advances in sequencing chemistry and workflows, the throughput for long-read sequencing has reached levels where it can be effectively harnessed for transcriptome assembly. The advent of long-read sequencing technologies capable of being used for transcriptomics has generated the need to develop software capable of efficiently generating isoforms from these reads. To date, multiple long-read transcriptome assembly pipelines have been developed yet there is no comprehensive analysis comparing them. Given the fact that long-reads may span the entire transcript, the problem of assembling transcriptomes can be framed and tackled differently than the traditionally used graph reduction problem. The transcriptome assembly pipelines we analyzed are Stringtie2, TALON, scallop-LR, FLAIR , Mandalorion, and IsoSeq3. These different transcriptome assembly pipelines approach the same problem in significantly different ways.

**What are the current long read transcriptome assemblers?**

Stringtie2 is designed to function with both short-read and long-read sequencing technologies. Stringtie2 is the second version of Stringtie, initially developed to build transcriptomes using elements of both reference-guided and *de novo* assemblies. With some changes, stringtie2 attempts to extend these same principles to be used with long reads in order to take advantage of reads that fail to span entire transcripts. Namely, stringtie2 aims to generate a splice graph of transcripts and assembles transcripts most strongly supported by the splice graph by constructing a flow network. Stringtie2 was developed before long-read CCS reads were readily available. Thus, Stringtie2 attempts to account for the high error rate long-reads by correcting for splice sites not supported by any low-error alignment reads. To do this correction they search for a supported splice site within 10 bp and shift the alignment accordingly and remove splice sites with low read support in the graph down to a particular threshold (Pertea 2015, Kovaka 2019).

In a similar fashion to stringtie2, Scallop-LR is also a short-read transcriptome assembler modified to take advantage of long-read sequencing technologies. Scallop-LR is specifically designed for PacBio long reads. Scallop-LR also utilizes a splice graph and decomposes the splice graph to find the most likely paths to represent a transcript. In order to retain information encoded in an aligned long-read (i.e. number of exons), scallop-LR represents long reads as long phasing paths and preserves these paths in assembly when generating transcripts (Shao and Kingsford 2017, Tung 2019).

FLAIR (Full Length Isoform Analysis of RNA) is a long-read specific transcriptome assembler that can optionally use short reads to increase splice site accuracy by using the short-reads to correct error prone long reads. FLAIR collapses transcripts into isoform groups by splice sites, and then further collapses those isoform groups by looking at windows of TSSs (transcription start sites) and collapsing all isoforms within that TSS window to the one that is most frequently represented. This same logic is applied to finding TESs (transcription end sites). FLAIR then groups these 'first-pass' isoforms by TSSs and TESs and aligns reads to these first pass isoforms, if a threshold number of raw-reads aligns to a particular isoform it is validated and considered a valid, observed transcript (Tang 2020)

IsoSeq3 is PacBio's proprietary transcriptome assembly pipeline built to work on PacBio reads specifically. IsoSeq3 relies on a hierarchical clustering algorithm to group reads amongst themselves based on their similarity. According to PacBio, similar sequences are defined as sequences with less than a 100 bp 5' end discrepancy, less than a 30 bp 3' end discrepancy and contains gaps less than 10 bp long. There is no limit on the number of gaps between two transcripts. Following clustering, IsoSeq3 performs a partial order alignment with clusters to generate an isoform consensus sequence. IsoSeq3 is the only non-reference guided transcriptome assembler we analyzed (Gordon and Tseng 2015).

Mandalorion identifies positions based on the raw reads where TSSs/TESs are highly likely. Small bins are then placed around these positions to include the highest number of read alignment ends possible. Mandalorion then identifies likely splice

sites in the raw reads, and groups these reads based on TSS, TES, and splice sites. A partial order alignment is performed on the grouped reads to generate a consensus isoform. In this manner, Mandalorion does not use partial reads not spanning the entire transcript to generate isoforms. An isoform is only called if it has raw read support (Byrne 2017).

TALON is a technology agnostic pipeline that relies on raw read support for calling isoforms. TALON is the only pipeline analyzed that utilizes a database that can be updated in order to track and utilize biological replicates. TALON error corrects reads, filters them so they don't contain sequencing artifacts, and then stores them in a database. If a near-identical read occurs in multiple replicates, then it is identified as an isoform. TALON relies on raw reads, but unlike Mandalorian, it does not group reads and does not construct a consensus sequence representative of an isoform (Wyman and Balderrama-Gutierrez 2020).

With the exception of IsoSeq3, all the transcriptome assembly pipelines we analyzed are reference-based. All the pipelines utilize different underlying algorithms to identify isoforms and we hypothesize that the resulting transcriptomes will be significantly different With increasingly accurate reads, and increasingly high throughput, purely looking at the per base or per transcript accuracy of the isoforms generated by a pipeline does not give a holistic picture of the actual performance of the pipeline. Various downstream analytical pipelines have been developed to provide insight into the nuanced details of individual transcriptomes. We use one such pipeline, SQANTI, for our initial downstream analysis of each transcriptome

assembled (Tardaguila and de la Fuente 2018). In this analysis, we will use high quality PacBio SMRT-seq data to compare the performance of these pipelines to one another.

## Materials and Methods

### PacBio Data

All sequencing data is publicly available at https://downloads.pacbcloud.com/public/dataset/Alzheimer2019_IsoSeq/ (Alzheimer's Disease Brain) and https://downloads-ap.pacbcloud.com/public/dataset/UHR_IsoSeq/ (Universal Human Reference). The Alzheimer's Disease Brain data was generated from 300 ng of total RNA (BioChain lot #507294) and prepared with Iso-Seq® Express Template Preparation for Sequel® and Sequel® II Systems. RNA library was subsequently sequenced on Sequel II System with Sequel II Binding Kit 1.0 and Sequel II Sequencing Kit 1.0 (4 rxn) for a duration of 24hr movie + 2hr pre-extension. The Universal Human Reference data consisted of Universal Human Reference RNA (Agilent) + SIRV Isoform Mix E0 (Lexogen). The sequencing library was prepared using Iso-Seq Template Preparation for Sequel Systems (PN 101-070-200) and subsequently sequenced using Sequel System II with "Early Access" binding kit (101-490-800) and chemistry (101-490-900). To clean the data for pipeline usage we aligned reads using pbmm2 and PacBio lima de-multiplexer. (https://github.com/PacificBiosciences/pbmm2,https://github.com/PacificBiosciences/barcoding)

**IsoSeq3 Data**

All IsoSeq3 data is available at https://downloads.pacbcloud.com/public/dataset/Alzheimer2019_IsoSeq/ and https://downloads-ap.pacbcloud.com/public/dataset/UHR_IsoSeq. No changes were made to the existing IsoSeq3 datasets.The Alzheimer's Brain dataset was analyzed with SMRTlink 8.0 "IsoSeq" protocol, followed by mapping to hg38 reference genome and collapsed into a non-redundant transcript set. The UHR dataset was specifically analyzed with SMRTlink 7.0 "IsoSeq With Mapping protocol" with hg38+SIRV combined reference genome and collapsed into a non-redundant transcript set.

**Scallop-LR**

Scallop-LR v0.9.2 and other versions of scallop are publicly available at https://github.com/Kingsford-Group/scallop/releases. Scallop-LR command that was used: scallop-lr -i <bam file> -o <gtf annotation file> -c <ccs-header-file> and all default settings. CCS header file was generated using: grep '>' input.fasta >ccs_headers.

**Stringtie2**

Stringtie2 is available at https://github.com/skovaka/stringtie2. Stringtie2 was run with default settings and the -L option for long reads and we provided an annotation file to guide the assembly process (-G option).

**Mandalorion**

Mandalorion v3.5 was used for Mandalorion transcriptome assembly. Run with specific command options -O 0,40,0,40 -t 110 -l 150 -n 2 -w 1 -r 0.01 -i 1-A 0. Mandalorion v3 is available at https://github.com/rvolden/Mandalorion-Episode-III.

**FLAIR**

Flair is available at https://github.com/BrooksLabUCSC/flair. Flair align, correct, and collapse were run with default settings.

**SQANTI**

SQANTI v1.2 was used for downstream analysis of transcriptomes. sqanti_qc.py command was run with default settings. This version of SQANTI is available at https://bitbucket.org/ConesaLab/sqanti/src/master/.

## Results

### Analysis of an Alzheimer's Disease Brain Sample

The Alzheimer's Disease Brain dataset contained 3485026 circular consensus sequencing reads. The median read accuracy was 99.83% and the average read accuracy was 99.39% percent. Each CCS read had a median of 2 insertions/deletions and 2 mismatches. The average number of mismatches and insertions deletions per read is significantly higher, with an average of 3.6 mismatches per read and 11.6 inserted/deleted bases. The median length was 2651 bases while the average length was 2917. The standard deviation for read length was +- 1203 bases. The longest read was 16203 bases long while the minimum read length was 50 bases. Summary statistics for the isoforms identified by pipelines utilizing consensus calling (IsoSeq3 and Mandalorion) are summarized in Table 1.

| Pipeline | Median Acc (%) | Mean Acc (%) | Median in/dels per sequence | Mean in/dels per sequence | Median mismatch per sequence | Mean mismatch per sequence |
|---|---|---|---|---|---|---|
| IsoSeq3 | 100 | 99.98893 | 0 | 0.3 | 0 | .04 |
| Mandalorion | 100 | 99.98238 | 0 | 0.425 | 0 | .041 |

**Table 1:** Accuracy, insertions/deletions, and mismatch summary for pipelines utilizing consensus calling.

All pipelines produced extremely accurate (>99.98%) transcripts based on alignment to the human reference genome (hg38) using minimap2. Given that long-PacBio reads utilize CCS technology and that the initial reads were highly accurate (99.39%), this is to be expected. All pipelines significantly reduced the number of

mismatches and insertions/deletions detected. Based on this initial analysis, basic metrics such as accuracy and mismatches clearly do not fully recapitulate the intricacies and differences in performance between these pipelines. Thus, we further analyzed and filtered our results using the SQANTI package. A summary of these results are shown in Figure 1 A-C.

There were clear discrepancies in quantity, type and length of isoforms recognized by each transcriptome assembly pipeline. PacBio's IsoSeq3 pipeline recognized the most isoforms, with 262424 isoforms identified. Mandalorion recognized the least isoforms, 26398. IsoSeq3 identified approximately 10x more isoforms than Mandalorion. SQANTI characterizes isoforms into a few distinct categories, most notably full-splice matches (FSM), incomplete splice matches (ISM), novel in catalog (NIC), and novel not in catalog (NNC) (Tardaguila and de la Fuente 2018). FSM are defined as isoforms which contain known splice sites in a previously observed order (present in the reference annotation). ISM refers to isoforms which contain consecutive observed splice sites in a known order but are also missing some splice sites. NIC refers to isoforms containing cataloged splice sites in a novel splice junction, and NNC isoforms refer to isoforms containing novel splice sites. Isoforms found in novel genes can be classified into a variety of groups, but for our analysis we grouped them under 'others'. There were significant discrepancies in the number of types of isoforms identified by each transcriptome assembly pipeline.

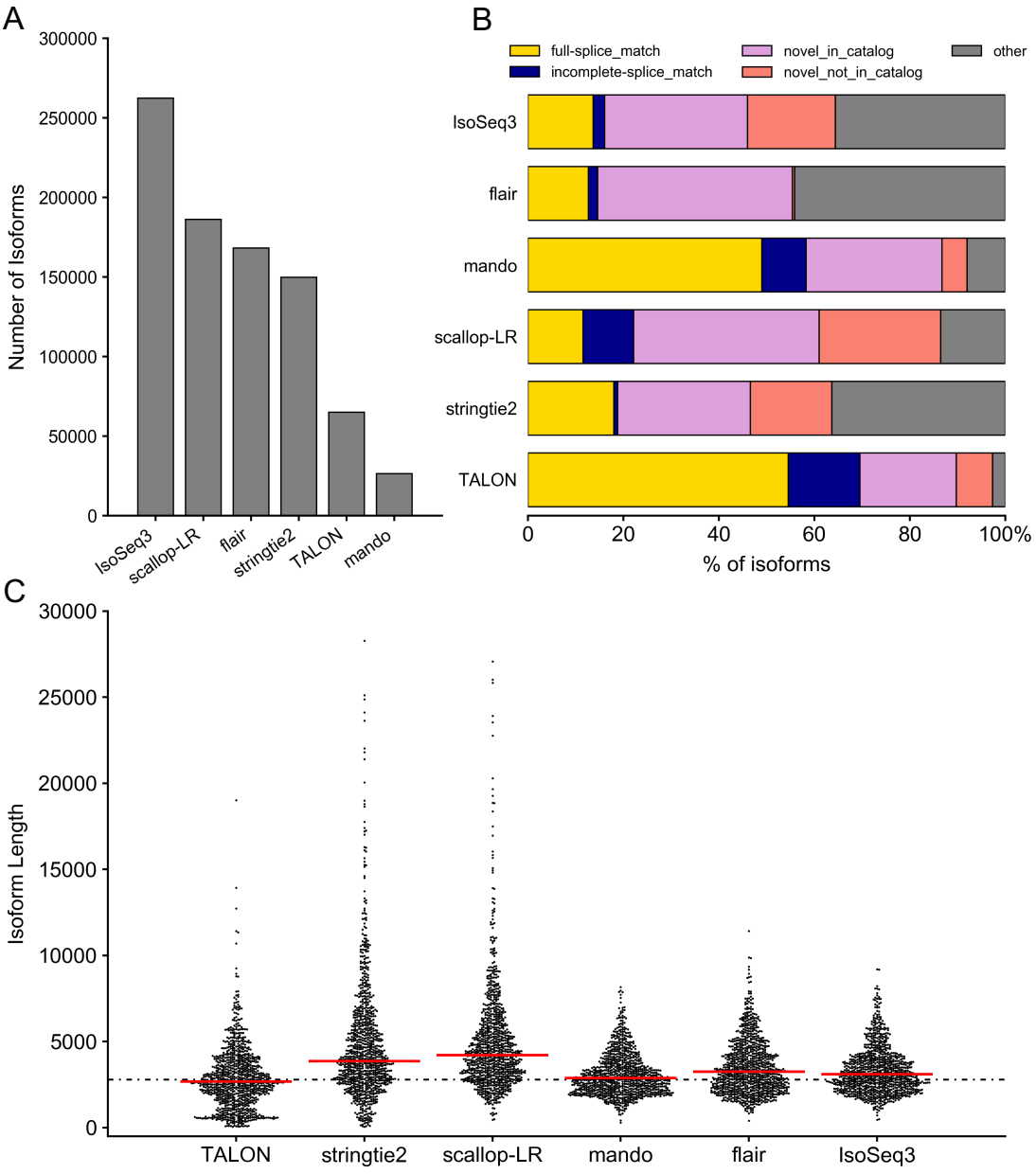


**Figure 1**: (A) Number of isoforms generated by each pipeline. (B) Percentage of isoforms generated by each pipeline belonging to a particular structural category as classified by SQANTI. (C) Swarmplot representing lengths of isoforms identified by each pipeline. Red line indicates pipeline specific median isoform length, dashed line indicates median transcript length in human (Piovesan 2016).

Both Mandalorion and TALON generated isoforms that most commonly fell into the FSM category, whereas the other pipelines generated isoforms that were distributed amongst all categories. The length distribution of isoforms identified by each pipeline also differed drastically. Theoretically, pipelines that rely on raw read support cannot generate isoforms longer than the longest read supplied into the pipeline. This is shown in Figure 1, where the only pipelines that generated isoforms longer than the longest read (16203 bp) were Stringtie2 and Scallop-LR. Both rely on a splicing graph to generate isoforms and thus it is possible to generate isoforms longer than the longest read. Mandolorion, IsoSeq3, TALON, and FLAIR all had a calculated median isoform length that is relatively consistent with prior reports on the median length of isoforms in the human genome (2787 bp) (Piovesan 2016). Stringtie2 and Scallop-LR both generated isoforms with a more dispersed length distribution, particularly generating isoforms that are longer than current estimates for the median transcript length in humans.

**Discrepancies in Isoforms Generated by different pipelines from the same Brain**

In order to elucidate why the transcriptome assembly pipelines generate such a significant difference in quantity and types of isoforms, we looked at isoform diversity across the transcriptome, to see if whether a few regions in the genome cause the majority of disparate isoforms or whether the different pipelines generate different numbers of isoforms across the genome. To visualize this, we looked at the isoform diversity for genes in a particular window of the genome by dividing the

number of distinct isoforms mapped to a region by the number of genes in that particular region.

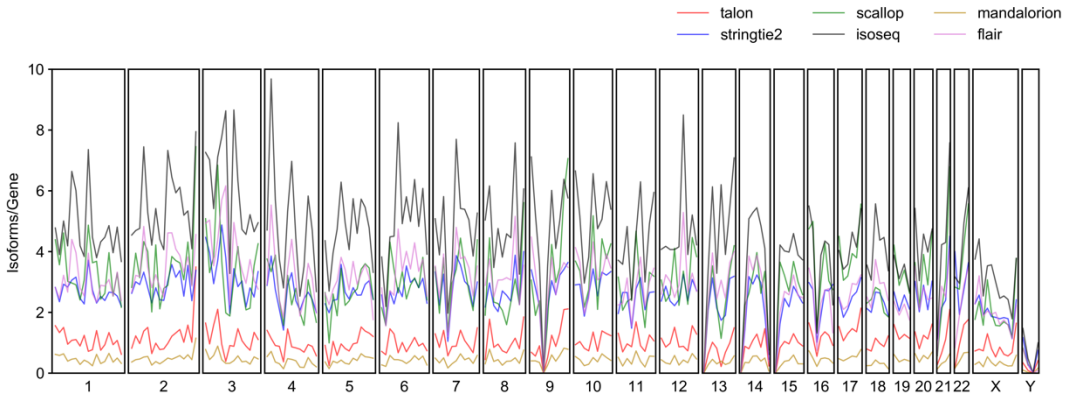


**Figure 2**: Density plot showing the isoforms/gene in a 15 million base window. Each box represents a chromosome and each line represents a pipeline.

We found that discrepancies in the number of isoforms assembled to be consistent across the transcriptome (Figure 2). IsoSeq3 consistently yielded the highest number of isoforms/gene, followed by scallop-lr, FLAIR, stringtie2, TALON, and Mandalorion. To more clearly understand the differences in the transcriptome outputs of these pipelines we analyzed individual gene loci and compared the generated isoforms to both the reference annotation and the read support present from sequencing data. Particularly, we analyzed isoforms that may be of clinical relevance in the Alzheimer's Disease brain (Figure 3). HSBP1 has been implicated in neurodegenerative disorders and has been shown to play a role in deterring the formation of tau protein aggregates in the aging brain (Baughman 2017). Stringtie2 and scallop-LR fail to capture the strongest transcript signature in our raw read data. Alternatively, Mandalorion and TALON fail to pick up on a weaker transcript

signature representing an elongated exon at the c-terminus as does FLAIR, albeit FLAIR incorrectly captures the elongated c-terminal exon as a mono-exonic isoform. IsoSeq3 identifies both isoforms present in the read data but also generates spurious isoforms that contain no read support.

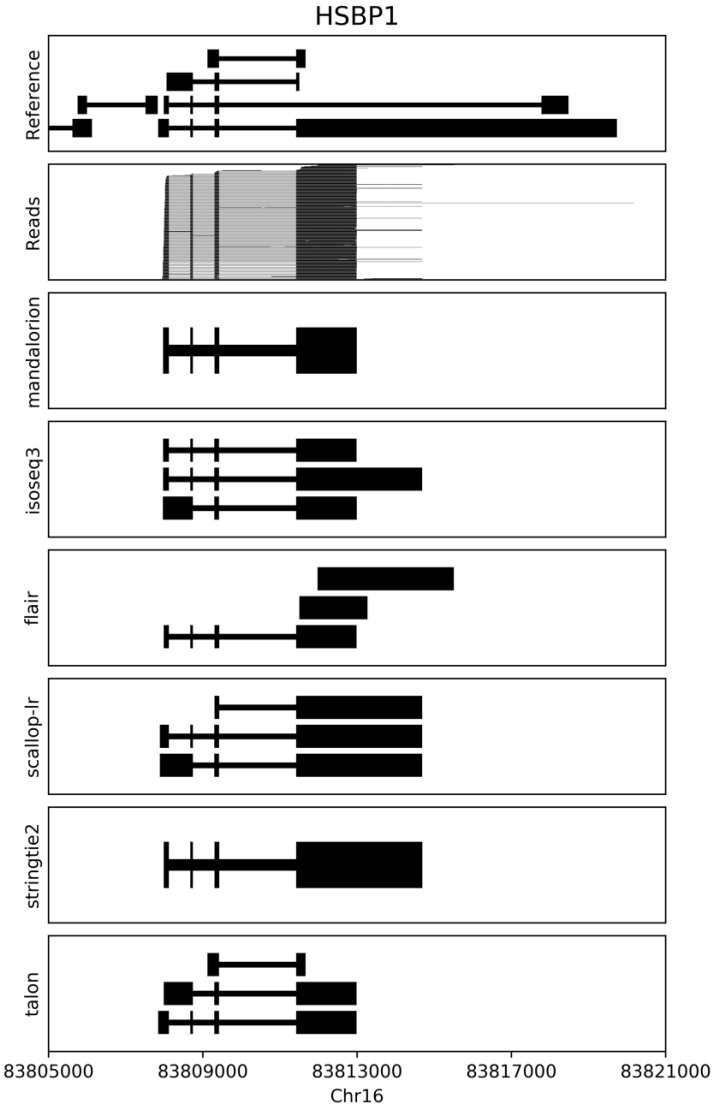


**Figure 3**: Genome Brower view of HSBP1 gene loci. Top panel indicates reference annotation isoforms, 2nd from the top panel represents raw reads sequenced, and subsequent panels indicate isoforms identified by different pipelines.

Our analysis on the Alzheimer's Disease Brain dataset showed that in order to discern a conclusion on pipeline performance we would need to analyze a dataset with some form of internalized control.

**Analysis of a Universal Human Reference dataset with an internalized control**

Another publicly available PacBio dataset contained a Lexogen SIRV spike-in which we can leverage as an internalized control in our sample. SIRVs are synthetically made RNA molecules that have sequences unique to the organism at hand, and they represent computationally difficult isoforms to assemble due to the lack of characteristic transcript features such as differential base composition in the splice site vicinity (Paul 2016). Given that we know the exact structure of these RNA molecules that we know are present in our data, we can gauge transcriptome assembly performance on how well each pipeline captures these RNA molecules from the data. We extended our analysis of the Alzheimer's Disease Brain dataset to this Universal Human Reference dataset containing a Lexogen SIRV spike-in (UHR).

This dataset contained 5,739,168 total reads. The median and average accuracy for each read was 99.8675% and 99.6458%, respectively. The median number of mismatches per read was 1 and the average number of mismatches per read of 1.906. The median number of insertions/deletions per read was 1 with the average number of insertions/deletions per read being 5.3397. The median length of the reads was 1804 bases and an average length of 1965 bases per transcript with a standard deviation of +- 1187 bases. The longest observed read was 19375 bases

long. Summary statistics for the transcriptomes assembled by pipelines using consensus calling are summarized in Table 2.

| Pipeline | Median Acc (%) | Mean Acc (%) | Median in/dels per sequence | Mean in/dels per sequence | Median mismatch per sequence | Mean mismatch per sequence |
|---|---|---|---|---|---|---|
| IsoSeq3 | 100 | 99.968 | 0 | .721 | 0 | .0017 |
| Mandalorion | 100 | 99.984 | 0 | .29 | 0 | .043 |

**Table 2:** Accuracy, insertions/deletions, and mismatch summary for pipelines utilizing consensus calling.

High-level statistics such as accuracy, mismatch, and insertion/deletion statistics continued to draw little insight into pipeline performance due to the fact that every pipeline generates highly accurate isoforms at the base level. The results of our post-SQANTI filtering analysis can be seen in Figure 4 A-C. While the UHR dataset contains significantly more reads, there is a notable decrease in the number of isoforms detected by IsoSeq3, Stringtie2, scallop-LR and FLAIR. TALON and Mandalorion both detected more isoforms. The decrease in isoform detection in this dataset for particular pipelines can be explained by the nature of the data. Due to the fact that this dataset contains DNA from multiple individuals across multiple tissue types, ubiquitously expressed isoforms will have much stronger read support than individual or tissue specific isoforms. Thus, there may be more raw reads in the data, but those raw reads more strongly support a smaller number of isoforms. TALON and Mandalorion rely on grouping reads to call isoforms, thus if there are more reads, representing less isoforms, the pipelines will be able discern more of these isoforms with increasing accuracy.

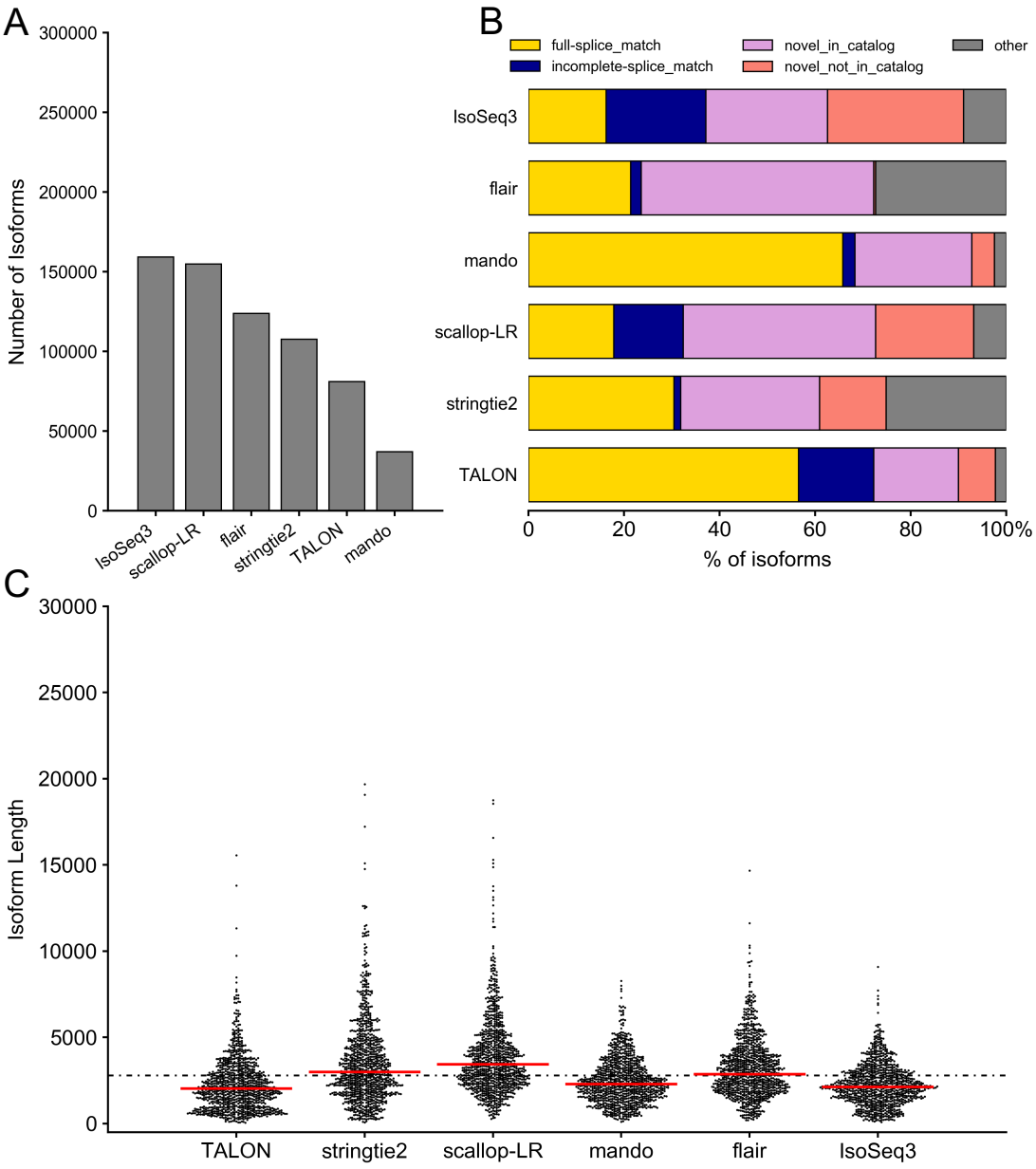


**Figure 4**: (A) Number of isoforms generated by each pipeline. (B) Percentage of isoforms generated by each pipeline belonging to a particular structural category as classified by SQANTI. (C) Swarmplot representing lengths of isoforms identified by each pipeline. Red line indicates pipeline specific median isoform length, dashed line indicates median transcript length in human (Piovesan 2016).

The structural categorization of isoforms generated from each pipeline is consistent with our prior findings with Mandalorion and TALON generating primarily full splice matches, while other pipelines generate a significant number of incomplete splice matches and novel in catalog splice matches. Numerous prior studies have shown that housekeeping genes and genes expressed broadly across multiple tissue types are under selective pressure for compactness (Eisenberg and Levanon 2003). We observe this trend in our data with all pipelines capturing shorter transcripts in the UHR dataset than they did in the Alzheimer's Brain dataset. Transcriptome-wide, there is less variance between each pipeline's output in terms of the density of isoforms per gene (Figure 5).

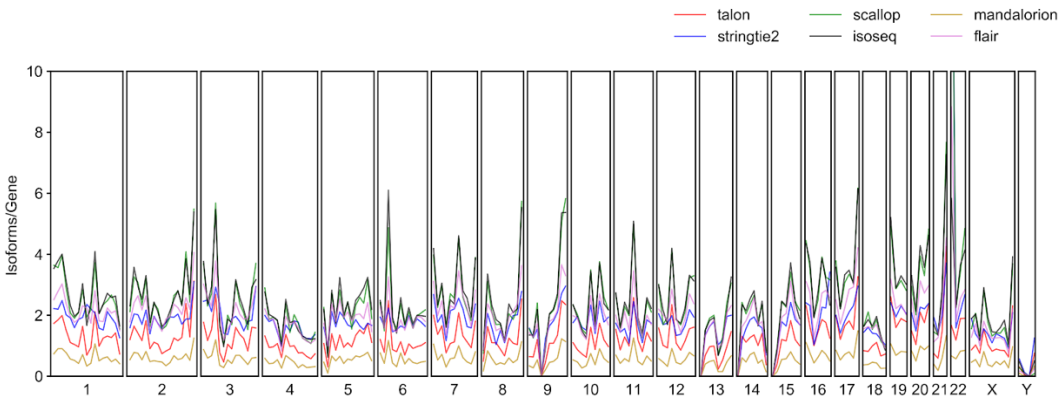


**Figure 5**: Density plot showing the isoforms/gene in a 15 million base window. Each box represents a chromosome and each line represents a pipeline.

**Mandalorion and Stringtie2 most clearly discern the internalized control**

65 artificial SIRV isoforms were inserted into this RNA pool. We analyzed the isoforms identified by each pipeline aligning to the artificial SIRV chromosomes. Mandalorion and Stringtie2 both correctly captured 54 out of the 65 artificial

isoforms present, and both captured 57 total isoforms. TALON correctly captured 53 isoforms but captured an additional 205 erroneous isoforms. Figure 6 summarizes the performance of the pipelines in capturing SIRV isoforms.

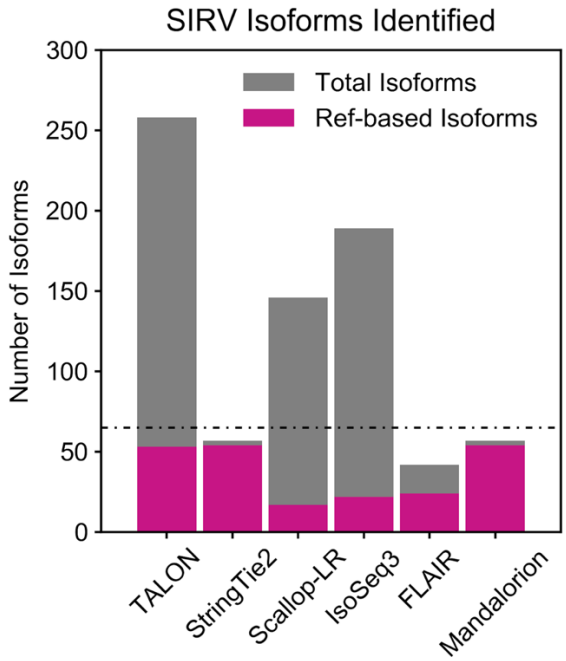


**Figure 6:** Bar plot representing the number of SIRV transcripts identified by each pipeline. Gray bar represents the total number of isoforms identified aligning to artificial SIRV chromosomes, magenta bar represents the number of isoforms that match an internal control SIRV isoform. Dashed line represents actual number of SIRV isoforms inserted into the RNA pool.

A genome browser view of the isoforms detected in the SIRV2 artificial chromosome provide a telling visual depiction of the general behavior of each pipeline. Mandalorion and Stringtie2 find a majority of isoforms, while other pipelines do not find a majority, or find many dubious isoforms as well.

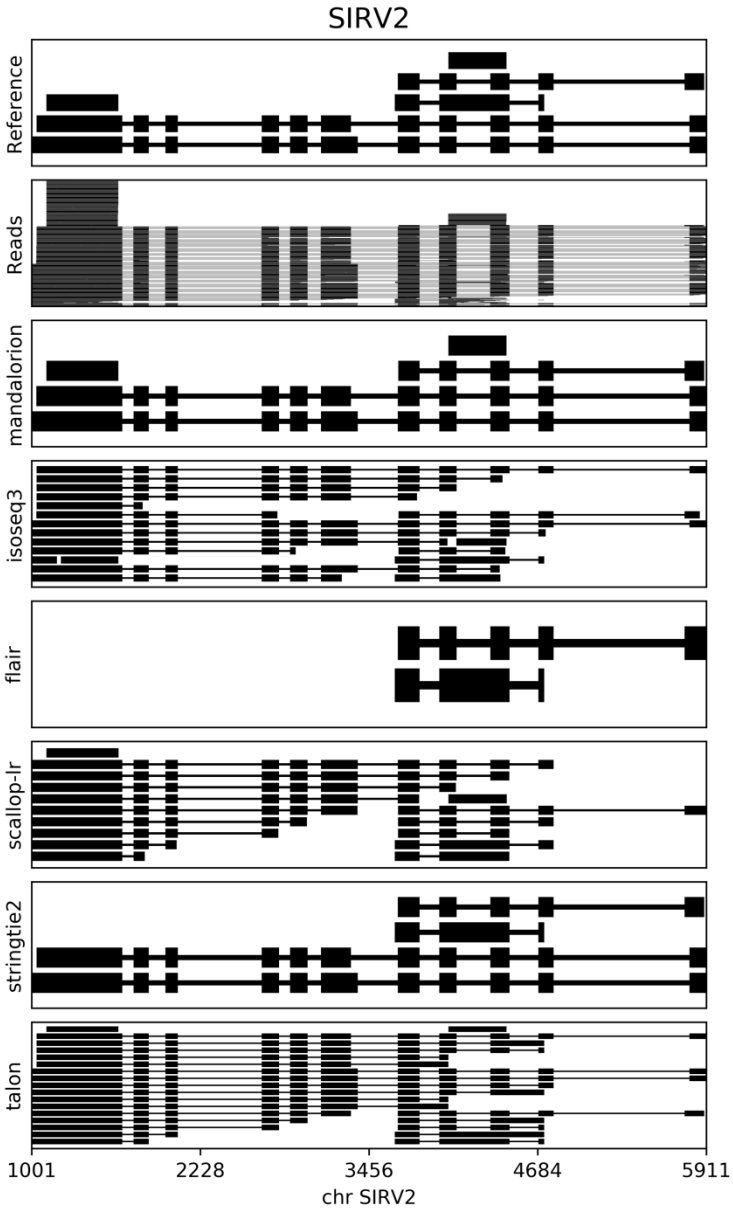


**Figure 7**: Genome Brower view of a SIRV2 artificial gene loci. Top panel indicates internal control reference isoforms that were inserted in our RNA pool, 2nd from the top panel represents raw reads, and subsequent panels indicate isoforms identified by different pipelines.

## Conclusion

Our analysis indicates that high level heuristics for determining transcriptome quality are not insightful to discern the ability of transcriptome assembly pipelines to accurately capture isoforms. TALON and Mandalorion were the only two pipelines which identified Full Splice Matches as the largest structural category of isoforms. While not all isoforms are, or should be, Full Splice Matches, in a species with a relatively well documented transcriptome, such as humans, we must expect that any given individual transcriptome should have a significant degree of similarity to the reference transcriptome. This is indirectly measured by the frequency of Full Splice Matches in each generated transcriptome.

Our analysis also showed the importance of having an internalized control in order to measure transcriptome quality between different pipelines. Stringtie2 and Mandalorion best recapitulated the control isoforms present in one of our samples.

This analysis did not leverage the full capabilities of TALON and FLAIR because we aimed to measure the performance of each pipeline given the same input. Because TALON uses a running database that can be updated, and it relies on biological replicates to call isoforms, TALON may be a more viable option when biological replicates are readily available. Conversely, FLAIR can leverage short, highly accurate Illumina reads to correct splice junctions and can increase performance by utilizing this hybrid transcriptome assembly method.

Ultimately, a variety of factors need to be considered when deciding on a transcriptome assembly pipeline to use. Mandalorion generated the most reliable and

experimentally validated transcriptome, but it also generated the most conservative transcriptome assembly. Mandalorion matches IsoSeq3 accuracy but also generates much better isoform models. Mandalorion relies on raw read support to generate isoforms, and it may not be able to capture longer isoforms due to sequencing limitations. On the other hand, if trying to capture long (>15kb) isoforms, Stringtie2 may be a more viable option. We analyzed two splice-graph based transcriptome assembly pipelines, Stringtie2 and Scallop-LR, and Stringtie2 significantly outperformed Scallop-LR. Finally, if access to biological replicates or short reads are available, it may be worthwhile to consider utilizing TALON or FLAIR, because they can leverage these technologies for more precise transcriptomes.

As long read sequencing becomes more accurate and more widely adopted, we hope that this analysis can assist biologists in picking a transcriptome assembly tool that suits their needs the best.

**Bibliography**

Ardui, Simon, Adam Ameur, Joris R Vermeesch, and Matthew S Hestand. 2018. "Single Molecule Real-Time (SMRT) Sequencing Comes of Age: Applications and Utilities for Medical Diagnostics." *Nucleic Acids Research* 46 (5): 2159–68. https://doi.org/10.1093/nar/gky066.

Eid, John, Adrian Fehr, Jeremy Gray, Khai Luong, John Lyle, Geoff Otto, Paul Peluso, et al. 2009. "Real-Time DNA Sequencing from Single Polymerase Molecules." *Science* 323 (5910): 133–38. https://doi.org/10.1126/science.1162986.

Gordon, Sean P., Elizabeth Tseng, Asaf Salamov, Jiwei Zhang, Xiandong Meng, Zhiying Zhao, Dongwan Kang, et al. 2015. "Widespread Polycistronic Transcripts in Fungi Revealed by Single-Molecule MRNA Sequencing." Edited by Deyou Zheng. *PLOS ONE* 10 (7): e0132628. https://doi.org/10.1371/journal.pone.0132628.

Henson, Joseph, German Tischler, and Zemin Ning. 2012. "Next-Generation Sequencing and Large Genome Assemblies." *Pharmacogenomics* 13 (8): 901–15. https://doi.org/10.2217/pgs.12.72.

Hölzer, Martin, and Manja Marz. 2019. "De Novo Transcriptome Assembly: A Comprehensive Cross-Species Comparison of Short-Read RNA-Seq Assemblers." *GigaScience* 8 (5). https://doi.org/10.1093/gigascience/giz039.

Hrdlickova, Radmila, Masoud Toloue, and Bin Tian. 2017. "RNA-Seq Methods for Transcriptome Analysis." *WIREs RNA* 8 (1): e1364. https://doi.org/10.1002/wrna.1364.

Kovaka, Sam, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. 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.

Kukurba, Kimberly R., and Stephen B. Montgomery. 2015. "RNA Sequencing and Analysis." *Cold Spring Harbor Protocols* 2015 (11): 951–69. https://doi.org/10.1101/pdb.top084970.

Li, Shengli, Zhixiang Hu, Yingjun Zhao, Shenglin Huang, and Xianghuo He. 2019. "Transcriptome-Wide Analysis Reveals the Landscape of Aberrant Alternative Splicing Events in Liver Cancer." *Hepatology* 69 (1): 359–75. https://doi.org/10.1002/hep.30158.

Love, Julia E., Eric J. Hayden, and Troy T. Rohn. 2015. "Alternative Splicing in Alzheimer's Disease." *Journal of Parkinson's Disease and Alzheimer's Disease* 2 (2). https://doi.org/10.13188/2376-922X.1000010.

Lu, BingXin, ZhenBing Zeng, and TieLiu Shi. 2013. "Comparative Study of de Novo Assembly and Genome-Guided Assembly Strategies for Transcriptome Reconstruction Based on RNA-Seq." *Science China Life Sciences* 56 (2): 143–55. https://doi.org/10.1007/s11427-013-4442-z.

Mantere, Tuomo, Simone Kersten, and Alexander Hoischen. 2019. "Long-Read Sequencing Emerging in Medical Genetics." *Frontiers in Genetics* 10. https://doi.org/10.3389/fgene.2019.00426.

Paul, Lukas, Petra Kubala, Gudrun Horner, Michael Ante, Igor Holländer, Seitz Alexander, and Torsten Reda. 2016. "SIRVs: Spike-In RNA Variants as External Isoform Controls in RNA-Sequencing." *BioRxiv*, October, 080747. https://doi.org/10.1101/080747.

Pertea, Mihaela, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. 2015. "StringTie Enables Improved Reconstruction of a Transcriptome from RNA-Seq Reads." *Nature Biotechnology* 33 (3): 290–95. https://doi.org/10.1038/nbt.3122.

Pickrell, Joseph K., Athma A. Pai, Yoav Gilad, and Jonathan K. Pritchard. 2010. "Noisy Splicing Drives MRNA Isoform Diversity in Human Cells." *PLOS Genetics* 6 (12): e1001236. https://doi.org/10.1371/journal.pgen.1001236.

Piovesan, Allison, Maria Caracausi, Francesca Antonaros, Maria Chiara Pelleri, and Lorenza Vitale. 2016. "GeneBase 1.1: A Tool to Summarize Data from NCBI Gene Datasets and Its Application to an Update of Human Gene Statistics." *Database: The Journal of Biological Databases and Curation* 2016 (December). https://doi.org/10.1093/database/baw153.

Raj, Towfique, Yang I. Li, Garrett Wong, Jack Humphrey, Minghui Wang, Satesh Ramdhani, Ying-Chih Wang, et al. 2018. "Integrative Transcriptome Analyses of the Aging Brain Implicate Altered Splicing in Alzheimer's Disease Susceptibility." *Nature Genetics* 50 (11): 1584–92. https://doi.org/10.1038/s41588-018-0238-1.

Rotival, Maxime, Hélène Quach, and Lluis Quintana-Murci. 2019. "Defining the Genetic and Evolutionary Architecture of Alternative Splicing in Response to Infection." *Nature Communications* 10 (1): 1671. https://doi.org/10.1038/s41467-019-09689-7.

Shao, Mingfu, and Carl Kingsford. 2017. "Accurate Assembly of Transcripts through Phase-Preserving Graph Decomposition." *Nature Biotechnology* 35 (12): 1167–69. https://doi.org/10.1038/nbt.4020.

Tang, Alison D., Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. 2020. "Full-Length Transcript Characterization of SF3B1 Mutation in Chronic Lymphocytic Leukemia Reveals Downregulation of Retained Introns." *Nature Communications* 11 (1): 1438. https://doi.org/10.1038/s41467-020-15171-6.

Tardaguila, Manuel, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector del Risco, Marc Ferrell, et al. 2018. "Corrigendum: SQANTI: Extensive Characterization of Long-Read Transcript Sequences for Quality Control in Full-Length Transcriptome Identification and Quantification." *Genome Research* 28 (7): 1096–1096. https://doi.org/10.1101/gr.239137.118.

Tung, Laura H., Mingfu Shao, and Carl Kingsford. 2019. "Quantifying the Benefit Offered by Transcript Assembly with Scallop-LR on Single-Molecule Long Reads." *Genome Biology* 20 (1): 287. https://doi.org/10.1186/s13059-019-1883-0.

Ungaro, Arnaud, Nicolas Pech, Jean-François Martin, R. J. Scott McCairns, Jean-Philippe Mévy, Rémi Chappaz, and André Gilles. 2017. "Challenges and Advances for

Transcriptome Assembly in Non-Model Species." *PLOS ONE* 12 (9): e0185020. https://doi.org/10.1371/journal.pone.0185020.

Volden, Roger, Theron Palmer, Ashley Byrne, Charles Cole, Robert J. Schmitz, Richard E. Green, and Christopher Vollmers. 2018. "Improving Nanopore Read Accuracy with the R2C2 Method Enables the Sequencing of Highly Multiplexed Full-Length Single-Cell CDNA." *Proceedings of the National Academy of Sciences* 115 (39): 9726–31. https://doi.org/10.1073/pnas.1806447115.

WANG, YAN, JING LIU, BO HUANG, YAN-MEI XU, JING LI, LIN-FENG HUANG, JIN LIN, et al. 2015. "Mechanism of Alternative Splicing and Its Regulation." *Biomedical Reports* 3 (2): 152–58. https://doi.org/10.3892/br.2014.407.

Wyman, Dana, Gabriela Balderrama-Gutierrez, Fairlie Reese, Shan Jiang, Sorena Rahmanian, Stefania Forner, Dina Matheos, et al. 2019. "A Technology-Agnostic Long-Read Analysis Pipeline for Transcriptome Discovery and Quantification." Preprint. Genomics. https://doi.org/10.1101/672931.