THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

Comparison of RNA Reconstruction Assemblers for Long-Read RNA Sequences

SAADYA RAO
FALL 2024

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Computer Science
with honors in Computer Science

Reviewed and approved* by the following:

Mingfu Shao
Professor of Computer Science
Thesis Supervisor

Ting He
Professor of Computer Science
Thesis Honors Adviser

* Electronic approvals are on file.

ABSTRACT

Ribonucleic acid (RNA) plays a crucial role in gene expression and protein synthesis, facilitating the regulation of biological processes through splicing. Accurately deciphering RNA sequences is essential for understanding gene activity and its implications in genetic diseases. This study focuses on evaluating the performance of six RNA transcript assemblers—StringTie, Scallop, PsiCLASS, StringTie2, Scallop2, and IsoQuant—on long-read RNA sequencing datasets generated from Oxford Nanopore Technology and Pacific Biosequences platforms. The datasets, comprising both annotated and unannotated samples, were processed using default parameters to ensure comparability. The assemblies were then assessed against GENCODE v36 annotations, utilizing GffCompare to calculate metrics sensitivity, or recall, and precision across various levels such as base, exon, intron, intron chain, transcript, and locus.

The results demonstrate that StringTie performed well, especially at the exon and intron levels, but showed some limitations in its sensitivity. StringTie2 improves upon its predecessor's precision, effectively reducing the number of false positives in transcript identification. In contrast, Scallop showed notably low sensitivity across all datasets, indicating significant challenges in capturing true transcripts. However, Scallop2 achieves a balance between sensitivity and precision by maintaining relatively high levels of both metrics, making it a more reliable option. IsoQuant consistently exhibits high precision across all levels, particularly when compared to Scallop2 and StringTie2, highlighting its utility for researchers prioritizing accuracy in transcript identification. Although the overall sensitivity of the assemblers varies, the analysis reveals that the performance metrics improve significantly with the use of annotated datasets. Additionally, the study underscores the importance of utilizing comprehensive reference genomes to enhance assembly quality.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to begin by expressing my appreciation to Dr. Mingfu Shao, my thesis supervisor and professor of computer science in the School of Electrical Engineering and Computer Science at the Pennsylvania State University. I am grateful to have worked under his mentorship and in collaboration with the other members of Shao Group. Dr. Shao's continued support and valuable insights have been instrumental to the development of this thesis and my growth as a researcher. I would also like to take a moment to thank my honors advisor Ting He for her assistance throughout my Schreyer Honors College experience. Finally, I want to express my gratitude to my family and close friends Rida Hamid, Sharon Liu, and Kenny Hermus for their guidance throughout my college career and the thesis-writing process.

Chapter 1

Introduction

Ribonucleic acid, or RNA, is a type of molecule containing four base pairs: adenine, cytosine, guanine, and uracil. These base pairs bond to make a single strand of genetic material in many living cells. RNA helps with regulating gene expression through splicing and protein synthesis. Splicing is the process of creating RNA only using certain parts of the gene, or section of DNA (Alberts et al., 2002). Splicing allows for the same gene to produce different RNAs, and hence different proteins. Understanding the expressed RNAs can provide great insight into the conditions that allow genes to turn on. Based on gene activity, researchers can determine how genetic diseases manifest and basic biological processes work. The process of understanding gene expression starts with deciphering RNA sequences that correspond with the gene of interest.

The complexity of RNA creates challenges for researchers trying to determine RNA sequences with accuracy. With its intricate structure, piecing together RNA sequences accurately can manifest itself akin to solving a molecular jigsaw puzzle. Existing wet-lab RNA sequencing technology only allows for the reading and sequencing of small snippets of RNA at a time. This means that researchers are left with fragments of RNA reads that might be overlapping, duplicated, or disconnected. The fragments are classified as short or long-reads depending upon how many base pairs are found in them. It is important to piece together these strands to obtain the ground truth RNA sequence.

These raw reads, or fragments of RNA, are often assembled and aligned to a region of a reference genome using alignment technologies. A reference genome is an annotated version of an organism's genes, clearly listing out the DNA bases in each chromosome. The reads are often stored in binary alignment maps (BAM) or sequence alignment maps (SAM). Then tools such as StringTie, Scallop, PsiCLASS, StringTie2, Scallop2, and IsoQuant take these BAM/SAM files and perform several actions

according to their algorithms to identify and splice together the final RNA transcript from the aligned long reads. These results are also matched to their appropriate gene. The transcripts are stored in gene transfer format (GTF) files.

Sequences of expressed RNA molecules can also be obtained from *de novo* RNA transcript assembly. Transcript assemblers such as Trinity, rnaSPAdes, and Trans-ABySS that specialize in *de novo* assembly do not require the raw reads to be aligned to a reference genome unlike regular assembly. This process is useful when there is no reference genome for the organism RNA being examined. However, *de novo* assembly produces more inaccurate results than regular assembly on average with almost any transcriptome assembler being used. This most likely occurs because genes can have copies in other areas of the genome and alternatively splice (different RNA sequences are created from the same gene). This means that when the reads are pieced together to make the final transcripts, it is difficult to tell how they fit together without a map, or reference genome, for guidance (Pertea et al., 2015).

Chapter 2

Literature Review

In the past decade, a variety of RNA-seq technologies emerged, each offering distinct methodologies to capture the complexity of the transcriptome. This section reviews these technologies and highlights any specific innovations designed to improve precision and sensitivity. It also details their experimental frameworks and results. By providing this context, the review lays the foundation for understanding how this thesis builds upon past advancements.

Tools and Their Methodologies

When developing transcriptome assembly software tools, researchers strive for precision, sensitivity, and computational efficiency. Sensitivity is a measure of how well the assembler can predict a transcript that matches the true RNA sequence. Precision is a measure of how many transcripts the assembler identifies that resemble the true RNA sequence. Tools with the highest and most balanced precision and sensitivity levels are said to be high in accuracy. Computational efficiency refers to the run time and memory usage of the tools.

Although some tools use similar algorithmic concepts, each of them processes the simulated and/or biological RNA-seq data in different ways. The input and output data for the algorithms as well as the structure of the procedures for each of the tools are uniquely designed to improve upon the goals listed earlier in a way that does not sacrifice one area for the sake of improving another.

StringTie

StringTie, introduced by Dr. Mihaela Pertea and Dr. Steven L. Salzberg in 2015, emerged as one of the earliest transcriptome sequencing tools. StringTie offers capabilities for using a reference genome

and performing *de novo* assembly (Pertea et al., 2015). It uses a generalized maximum flow network where all the nodes in the alternative splice graph (ASG) are linked to all the nodes in the network. To obtain all the ASG's for testing, fragments of kidney cell RNA were assembled and then mapped to the GRCh37/hg19 human reference genome. Then, for each gene loci, ASG's were built. Using the ASG's StringTie found the paths with the highest read per base coverage and determined the coverage level of those transcripts. Using the generalized maximum flow network algorithm, in which the number of fragments tied to a given transcript could maximize and update based on expression levels, determination of coverage level became possible (Pertea et al., 2015).

To understand StringTie's approach, understanding of the concept behind the generalized maximum flow network algorithm must come. A network is a set number of nodes connected by directed edges with a source (a node with no incoming edges) and sink node (a node with no outgoing edges) (Pertea et al., 2015). An edge connects two nodes in the graph if the corresponding alignment mappings start at one of the nodes and ends at the other. A generalized flow network, like a regular network, additionally accounts for potentially lost or gained flow in the network, mimicking a real-world scenario. Depending on the required outcome, there are many strategies to find the flow of the network. StringTie uses the value of the maximum flow (max $(||f|| = \sum_{(s,v)\in E} f((s,v)))$ for a couple of reasons: it is a novel approach to transcript assembly, it can be solved in polynomial time, and it provides the transcripts with the highest expression levels (Pertea et al., 2015).

Scallop

While StringTie introduces novelty through its use of flow networks, Scallop distinguishes itself by advancing the splice graph methodology. Scallop identifies multi-exon transcripts (portions of RNA that code for proteins) as well as transcripts expressed at low levels through the required use of a reference genome (Shao & Kingsford, 2017). In doing so, Scallop aims to strike a better balance between

sensitivity and precision. The algorithm for Scallop begins with organizing the aligned RNA-seq reads by gene loci and creating a splice graph with weighted edges and phasing paths for each locus. For each of these graphs, a set of paths (labeled as P) are identified and marked with a value *f(p)* that corresponds to them. Any path in the set of P represents an expressed transcript and *f(p)* is the frequency with which that transcript is expressed. P also covers all phasing paths. Phasing paths help to highlight which exons connect to form the given transcripts. False-positive phasing paths are also removed from the splice graphs. Once the splice graphs have been filtered, s-t paths are created from the graphs through analyzing the vertices on an individual level, also known as vertex decomposition (Shao & Kingsford, 2017).

Each vertex is decomposed differently depending on the influence of the phasing paths on it; the vertex is trivial if its in-degree or out-degree is one, otherwise, it is nontrivial (Shao & Kingsford, 2017). If the vertex is nontrivial, it can either be classified as 'splittable' or 'unsplittable'. These designations are created to classify and decompose the vertices using different linear programming instances. Although there are various linear programming instances with different objective functions, the end goal for these linear programming instances is to reduce the difference between the read coverage observed in the sequencing data and the expected coverage based on transcripts identified (Shao & Kingsford, 2017).

Upon decomposing all vertices, the s-t paths in P are rebuilt and the corresponding weights *f(p)* are reassigned. The final transcripts are obtained from running the procedures mentioned earlier on all gene loci. The collection of transcripts is filtered such that the remaining ones are highly expressed and long to ensure that all false-positive transcripts are removed (Shao & Kingsford, 2017).

PsiCLASS

Arguably the best and most versatile transcript assemblers were StringTie and Scallop before 2019. This did not stop other researchers from seeking out ways to improve upon the existing technologies, especially since there was a stronger need to improve sensitivity. Li Song, Sarven

Sabunciyan, Guangyu Yang, and Liliana Florea collaborated on a new RNA transcript assembler, PsiCLASS, that would do exactly this. Like Scallop, PsiCLASS requires the input RNA-seq reads to be aligned to a reference genome (Song et al., 2019). Using these reads, a graph of all sub-exons (portions of exons that may or may not be used for coding proteins) and splice variants is built for a sample of RNA-seq data from a given gene. A sub-exon graph is one where the vertices are subexons and the edges denote if the subexons are in the same region or connected by an intron (a portion of RNA that does not code for proteins) in the read data. It is important to note that subexons are chosen precisely using the Bayes formula $P_R(r_i) = \frac{\pi\Gamma_{\Theta_0,k_0}(r_i)}{\pi\Gamma_{\Theta_0,k_0}(r_i)+(1-\pi)\Gamma_{\Theta_1,k_1}(r_i)}$ in order to reduce the chances of the including intronic noise (Song et al., 2019). Since there are potentially many samples for a given gene, multiple sub-exon graphs can be created and combined to make a global graph strung together where any graphs share one or more introns. Using the noise information and intron coverage level $r_i = \min\left(\frac{c_i}{c_{i-1}}, \frac{c_i}{c_{i+1}}\right)$, a score is calculated for each sub-exon, indicating if it should or should not be included in the global graph (Song et al., 2019). PsiCLASS uses the global sub-exon graph to select the transcripts for output.

The selection process for transcripts starts with PsiCLASS identifying maximal paths in the global graph between source and sink nodes. Once all maximal paths are obtained, only a portion of transcripts are selected to end with the smallest number of transcripts given the constraints of the SET_COVER problem and dynamic programming algorithms used in transcript selection. Finally, all the selected transcripts are given a score based on how frequently they appear in the sample. A voting procedure takes place to obtain the final transcripts across the global sub-exon network (Song et al., 2019).

StringTie2

In 2021, there was a growing need for making transcript assemblers fast in execution time while maintaining low error rates given the rapid increase in RNA long-reads analysis projects. StringTie2 was

built on top of StringTie to improve the original speed and accuracy in transcript assembly. While StringTie2 still implements a generalized maximum flow network algorithm like StringTie, StringTie2 utilizes new data structures and storage mechanisms that allow the program to run faster and use less memory (Kovaka et al., 2019). For example, instead of storing all individual reads, StringTie2 stores the number of reads aligned to the same place in the genome, significantly reducing the amount of input data stored in memory. For storing the splice graphs, StringTie uses bit-vector representation while StringTie2 uses a sparse bit-vector data structure. This is especially important because many of the bits would be set to 0 in StringTie's implementation, taking up unnecessary space and reducing computational efficiency unlike StringTie2. Two new strategies are also employed in StringTie2 to improve high error rates: readjusting splice errors due to third generation sequencing instruments and pruning false positive edges (Kovaka et al., 2019).

Not only does StringTie2 perform fast and accurate transcript assembly for long-reads, but also it works for combining short reads into longer sequences, also known as super reads (Kovaka et al., 2019). Usually, the super reads sequencing technologies are prone to errors. StringTie2 tries to reduce these errors by modifying the error correction procedures such as the k-mer look up table from QuorUM (Kovaka et al., 2019).

Scallop2

Despite the numerous advances in transcriptome assemblers for precision, speed, memory usage, and sensitivity, transcript assembly remains a challenge. Although StringTie, Scallop, PsiCLASS, and StringTie2 have procedures to ensure that the final software assembled transcripts are identical to the true RNA sequence, accuracy rates are still relatively low. Scallop2 addresses this through its algorithmic procedures, which improve upon those found in Scallop. The algorithm for Scallop2 begins like Scallop where the aligned RNA-seq reads are organized by gene loci and a splice graph is created with P paths

and *f(p)* for each path (Zhang et al., 2021). The key improvements in Scallop2 lie in the next three main algorithmic steps. The first step involves the bridging algorithm where the phasing paths with many ends are combined, or bridged, together to make them single-ended. This is helpful in the process of finding paths that are closer to the ground truth sequence instead of paths created due to alternative splicing and alignment errors. Such erroneous paths will be eliminated because of the weighting system of the bridging algorithm, enhancing the accuracy of the system. The second step in the Scallop2 algorithm is to find false start and end vertices. Any reads that do not perform the bridging algorithm properly from the precious step will have missing edges between given nodes. This means that transcripts might form incorrectly due to false starting and ending vertices. Scallop2 then uses a special procedure to make sure that the false vertices will be kept in the splice graph, but the transcript fragments that are created from these vertices will be stored separately from the proper transcripts (Zhang et al., 2021).

IsoQuant

Recently, a new transcript assembler was produced specifically for long reads RNA analysis to improve upon precision and sensitivity in a novel way. The difference in algorithmic procedures of IsoQuant from the other transcript assemblers even allows for new transcripts to be identified that were not previously marked in a database. This gives researchers new opportunities to find out more about existing genes. IsoQuant follows a four-step algorithm to produce the transcripts after aligning the reads with a reference genome. The first step involves matching long-reads to isoform vectors (transcripts that are produced from the same gene through alternate splicing). This is done by comparing the exons of the known isoform to the read. Then, in the second step, the number of reads that are assigned to each isoform is noted as this will provide a direct relation to how frequent the transcripts are expressed. Before reconstructing and outputting the final transcript, IsoQuant properly aligns all splice junctions and manages any sites that do not have associated annotations. To get the final transcripts, isoquant creates

intron graphs much like StringTie creates splice graphs and then finds paths in the graph that match at least one aligned read. Using this method, new transcripts are found with high precision (Prjibelski et al., 2023).

Overall Comparison

Each of the examined tools addresses precision, sensitivity, and computational efficiency in unique ways. Despite their efforts to enhance these aspects through special algorithms, inherent trade-offs inevitably arise, resulting in strengths and limitations in performance across different tasks in transcript assembly.

In terms of precision and sensitivity, StringTie showed a 53% increase in transcripts assembled compared to one of the competing early transcript assemblers, Cufflinks, on RNA reads from human blood and a 20% increase on simulated RNA data (Pertea et al., 2015). However, Scallop later emerged with notable improvements, creating 34.5% more multi-exon transcripts and 67.5% lowly expressed transcripts with higher accuracy compared to StringTie (Shao & Kingsford, 2017). Interestingly, a comparative analysis revealed that while StringTie exhibited higher sensitivity, it displayed lower precision than Scallop in predicting single-exon transcripts (Shao & Kingsford, 2017). Building upon these findings, PsiCLASS introduced further refinements, particularly in precision, surpassing both StringTie and Scallop with a 71.4% precision rate on simulated data, marking a 15% and 28.3% improvement over StringTie and Scallop, respectively (Song et al., 2019). Despite StringTie's superior accuracy in single-exon transcript prediction compared to PsiCLASS, the latter demonstrated approximately 16.8% to 24.5% higher sensitivity and 21% to 31.8% greater precision for other transcripts (Song et al., 2019). Then StringTie2 emerged and significantly improved on sensitivity and precision compared to its predecessor StringTie and Scallop. Notably, on the same datasets that Scallop used, StringTie2 displayed a 3.9% increase in sensitivity and 47.3% increase in precision (Zhang et al., 2021).

On the *Zea mays* dataset, StringTie2 produced results with 24% increased sensitivity compared to Scallop, but Scallop had about 1% higher precision than StringTie2 (Kovaka et al., 2019). Shortly after, Scallop2 was introduced and showed consistently higher scores of precision and sensitivity in various datasets over StringTie2. For example, on the Smart-seq3 datasets Scallop2 produced 0.7% more matching transcripts than StringTie2 and increased precision by 107.1% and 12.5% for StringTie2 and Scallop (Zhang et al., 2021). IsoQuant was developed a few years later and although it did not compare in terms of precision and sensitivity to StringTie2 and Scallop2, it performed significantly better than StringTie and Scallop (Pardo-Palacios et al., 2023). IsoQuant was extremely strong in detecting and eliminating false positives (Prjibelski et al., 2023). In fact, IsoQuant showed a 2.5-fold decrease in the number of false positive transcripts compared to StringTie. Also, three other tools verified 70.1% of the transcripts that IsoQuant produced. Another strength of IsoQuant is that it produced 71% more new transcripts that were able to be confirmed through analyses whereas StringTie only produced 37% (Prjibelski et al., 2023).

When examining computational efficiency, StringTie took under 30 minutes to run on simulated data where the other two competing assemblers Cufflinks and Traph took 81 mins and 48 hours (Pertea et al., 2015). StringTie also used around 1.6 to 12 gigabytes (GB) of memory to run whereas Cufflinks and some of the other assemblers used 6.4 to 26.6 GB (Pertea et al., 2015). However, on a large set of 667 RNA-seq samples, PsiCLASS finished the tasks in 9 hours with a maximum of 14 GB used, while StringTie took 34 hours with 524 MB and Scallop took 75 hours with 5 GB (Song et al., 2019). StringTie2 significantly improved on StringTie's computational efficiency by running 1.8 times faster and using 17 times less memory compared to Scallop (Zhang et al., 2021). Later, Scallop2 improved Scallop's computational efficiency but fell a bit behind StringTie2 (Zhang et al., 2021). On 10 Illumina RNA-seq samples, Scallop2 took 30 minutes and 6 GB of memory, while the fastest test of StringTie2 took 8 minutes and 148 MB.

While some tools excel in precision and sensitivity, such as PsiCLASS and Scallop2, others prioritize computational efficiency, like StringTie2. In addition, tools like IsoQuant emphasize the generation of new, accurate transcripts. However, shortcomings in certain areas, such as precision, sensitivity, and computational efficiency are evident in several tools. Achieving a balance between these three areas remains a challenge for many tools. These limitations highlight the ongoing need for improvement in transcript assembly methodologies to meet the demands of diverse RNA-seq datasets and analytical requirements.

Chapter 3

Methods

Prior to conducting experiments, appropriate datasets of long-read RNA-seq data were gathered. The five datasets encompass data generated from Oxford Nanopore Technology and Pacific Biosequences platforms, as detailed in Table 1. The datasets are divided into two groups to assess each assembler's performance under different conditions. Group 1 consists of three human datasets from Nanopore: NA12878-cDNA, NA12878-DirectRNA, and NA12878-IVT_RNA. Group 2 involves additional human datasets from PacBio to further evaluate: ENCFF450VAU and ENCFF694DIE. All these datasets were leveraged in the form of sorted BAM files.

Table 1: Summary of Datasets

Group Number	Name	Protocol	Species (tissue)
Group 1	NA12878-cDNA	ONT SQK-PCS108	Human (blood)
Group 1	NA12878-DirectRNA	ONT SQK-RNA001	Human (blood)
Group 1	NA12878-IVT_RNA	ONT SQK-RNA001	Human (blood)
Group 2	ENCFF450VAU	PacBio long-read RNA-seq	Human (blood)
Group 2	ENCFF694DIE	PacBio long-read RNA-seq	Human (blood)

After dataset preparation and acquisition, each long-read RNA-seq dataset was processed through the six transcript assemblers: StringTie version 2.2.1, Scallop version 0.10.5, PsiCLASS version 1.0.3, StringTie2 version 1.3.6, Scallop2 version 1.1.2, and IsoQuant version 3.3.0. To ensure fair comparison the assemblers were run using default parameters. IsoQuant, in particular, required the use of a reference genome FASTA file; therefore, the GRCh38 Primary Assembly Genome file was chosen due to its reliability for being a comprehensive human reference genome, which provides a framework for accurate transcript assembly and annotation.

Assembler Overviews

StringTie is a transcriptome assembly tool that employs a flow network-based approach to reconstruct and quantify full-length transcripts. It builds an alternative splice graph from RNA-seq reads mapped to the GRCh37/hg19 human reference genome and uses a generalized maximum flow algorithm to identify the most expressed transcripts. This approach allows StringTie to simultaneously assemble transcripts and estimate their expression levels by finding paths in the graph that maximize read coverage, representing the most abundant isoforms in the sample.

Scallop is a reference-based transcript assembler that goes beyond traditional splice graph methodology to improve transcript reconstruction accuracy. It processes RNA-seq reads by building splice graphs with weighted edges and utilizing phasing paths to determine transcript connectivity. The tool employs a sophisticated vertex decomposition strategy, analyzing each junction point as either "splittable" or "unsplittable," and uses linear programming to optimize transcript identification. This helps Scallop effectively identify both multi-exon transcripts and low-abundance transcripts while maintaining a balance between sensitivity and precision in its final output.

PsiCLASS focuses on sensitivity through a multi-sample analysis by constructing sub-exon graphs and creating a global graph structure using a Bayesian approach. Then, using dynamic programming and SET_COVER, transcripts are identified and voted on to determine final transcript structures. This approach allows PsiCLASS to detect transcripts that may be weakly expressed in individual samples but consistently present across multiple samples.

StringTie2 maintains StringTie’s network flow-based approach while significantly improving accuracy and computational efficiency for long-read RNA-seq data. It uses optimized data structures such as sparse bit-vector representation and consolidated read storage. StringTie2 also introduces new error correction strategies for third-generation sequencing data and incorporates super-read technology for short reads, making it more robust in handling sequencing errors while maintaining fast execution times.

Scallop2's innovation in transcript assembly lies in its three-step approach. The first step involves the bridging algorithm, which combines multi-ended phasing paths into single-ended paths. The following steps implement a novel strategy to handle false start/end vertices. By implementing these improvements to splice graph analysis, Scallop2 more effectively distinguishes between true transcripts and transcripts created by alternative splicing or alignment errors.

IsoQuant utilizes a four-step approach to transcript identification and quantification. The tool matches long reads to known isoform vectors, quantifies expression levels based on read assignments, aligns splice junctions, and constructs intron graphs to identify both known and novel transcripts. This methodology allows IsoQuant to not only accurately reconstruct and quantify known transcripts but also discover previously unannotated isoforms with high precision.

Finalizing Outputs

The resulting transcriptome assemblies were generated in GTF format and compared against GENCODE v36 basic annotation for accuracy evaluation using GffCompare version 0.11.2. The full workflow is illustrated in Figure 1. GENCODE v36 basic annotation provided a reliable basis for comparison since it is abundant in RNA genes and transcripts. Observing the general statistics, there are 19962 protein coding genes, 85269 protein coding transcripts, 17958 long non-coding RNA genes, and 48734 long non-coding RNA loci transcripts (*GENCODE - Human Release 36 Statistics*, n.d.). Being one of the last major releases of the Human GENCODE Gene Set since the latest release, it is comprehensive and includes well-validated gene structures. GffCompare determined the quality of the assemblies by calculating various metrics such as sensitivity, precision, and F1-score at base, exon, intron, and transcript levels. Additionally, GffCompare provided statistics on structural matches at multiple levels, quantifying matching intron chains, transcripts, and loci, as well as the proportion of missed and novel features including exons, introns, and complete loci.

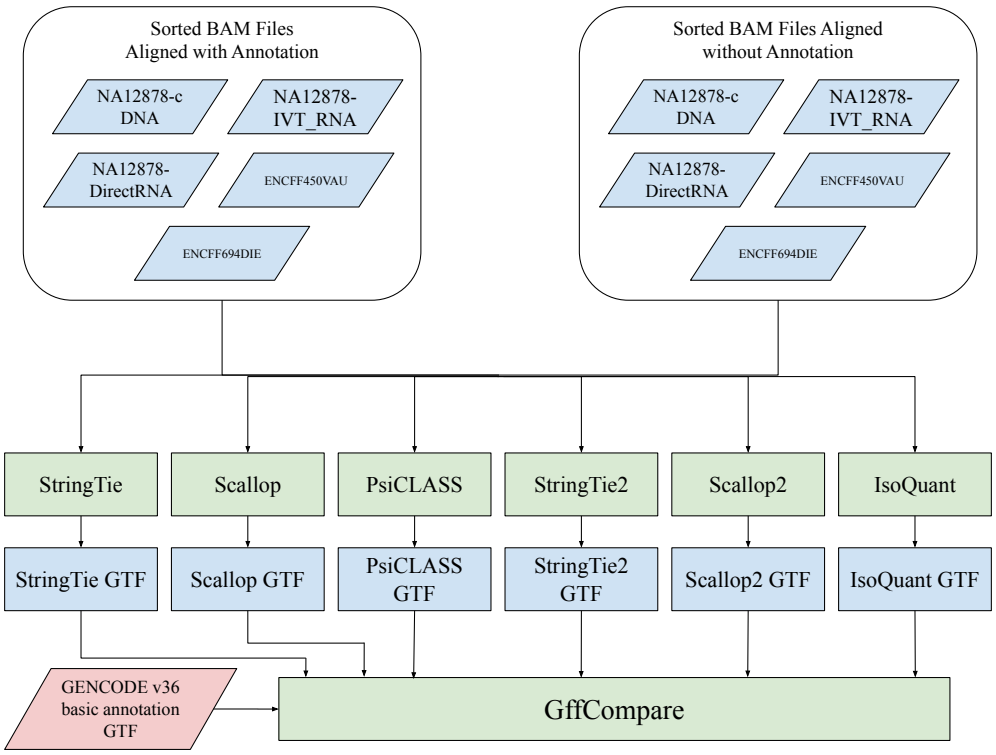

Figure 1: Pipeline for RNA Sequencing Data Processing and Analysis

Metrics

To assess the quality of the transcriptome assemblies generated from the data, two key evaluation metrics were employed, including sensitivity and precision. Sensitivity, or recall, was calculated as the proportion of true positive transcripts identified by the assembly compared to the total number of known transcripts in the GENCODE v36 reference annotation. This metric provides insight into the assembly's ability to capture true biological signals, which is crucial for understanding transcript expression accurately. Precision will be calculated as the ratio of true positive transcripts to the total number of transcripts reported by the assembly. High precision indicates that the assembled transcripts are more likely to be correct, minimizing the number of false positives.

The metrics were analyzed at multiple levels to provide a comprehensive evaluation of the assemblies. The base level analysis assesses the accuracy of individual nucleotide bases within the

assembled transcripts compared to the reference annotation, demonstrating the assembly's precision at the most fundamental level. The exon level metric measures the correct identification of exons (the coding segments of genes) allowing for an evaluation of how well the assembly captures the regions that contribute to protein coding. The intron level evaluates the accuracy of intron recognition, which are non-coding sequences that separate exons. The intron chain level assesses the assembly's ability to correctly identify chains of introns between exons, providing insights into the structural integrity of the transcript. Additionally, the transcript level measures the accuracy of complete transcript assemblies, focusing on the identification of full-length transcripts as they relate to the reference annotation. Finally, the locus level evaluates the assembly's performance at the gene locus level, ensuring that entire gene regions are correctly represented and accounted for in the transcriptome assembly.

Chapter 4

Results and Discussion

This section lays out the results and evaluation of the precision and sensitivity on different levels of transcript assembly across StringTie, Scallop, PsiCLASS, StringTie2, Scallop2, and IsoQuant across six different datasets aligned with and without annotation.

Table 2 summarizes the average sensitivity and precision of each tool across all the datasets as a percentage. These general results are the aggregate of the individual data points that will be examined.

Table 2: Summary Table

Tool	Average Sensitivity	Average Precision
StringTie	19.02	60.58
Scallop	0.27	39.01
StringTie2	13.68	67.35
Scallop2	16.46	63.52
IsoQuant	11.70	79.10

StringTie demonstrates moderate sensitivity, indicating that it successfully identifies a portion of true transcripts. Its precision, while higher than Scallop, suggests that there is room for improvement in minimizing false positives. Scallop's sensitivity and precision are very low for these datasets, indicating low effectivity in identifying true transcripts and a high rate of false positives when it makes predictions. These numbers may suggest that Scallop is better suited for short-read data, which may explain its challenges with the longer reads in datasets provided. Meaningful results for PsiCLASS could not be obtained within the available timeframe, and hence the results for PsiCLASS are not included in Table 2. StringTie2 demonstrates a sensitivity that is lower than StringTie, but its precision is better. This tradeoff shows that while it may identify fewer true transcripts, those it does identify are more likely to be accurate. Scallop2 performs better than its predecessor, Scallop, with improved sensitivity and precision. When compared to StringTie2, Scallop2 has a slightly higher sensitivity by about 3% and a lower precision by about 4%. On the other hand, Scallop2 shows a lower sensitivity and higher precision than

StringTie, illustrating a stronger balance between the tradeoff of recall and precision between both StringTie assemblers. IsoQuant stands out with the highest precision among the tools evaluated, suggesting that when it identifies a transcript, it is likely to be correct. However, its sensitivity is significantly low, indicating that it may miss many real transcripts. Taking an overall look at the data leads to the observation that no single tool excels in both sensitivity and precision. The difference in performance highlights the necessity of careful assembler selection based on whether high recall or precision is more important.

Detailed Results by Tool

StringTie

On average, sensitivity varies significantly across the three Nanopore datasets with averages of 15.47%, 28.03%, and 9.57% for NA12878-cDNA, NA12878-DirectRNA, NA12878-IVT_RNA respectively as shown in Table 3. However, they are slightly higher in the datasets aligned with annotations compared to those aligned without annotation upon examining the respective average sensitivities of 15.47%, 26.72%, and 9.20% in Table 5. The average precision levels are quite high for the Nanopore datasets even across annotation types most notably for the NA12878-IVT_RNA set at 75.25% in Table 3. Precision and sensitivity remain relatively constant across the datasets and annotation types for PacBio with an average sensitivity of 21.90% for ENCFF450VAU and 21.05% for ENCFF694DIE and an average precision of 55.63% for ENCFF450VAU and 52.52% for ENCFF694DIE in Table 4. Examining the exon level, both the Nanopore and PacBio datasets generally yield higher recall and precision percentages when aligned with annotation. The sensitivity of PacBio is generally higher than that of Nanopore but the opposite is true for precision. This is evident in the comparison of the lowest average precision of 58.98% from NA12878-cDNA in the Nanopore results in Table 3 versus the highest

average precision of 55.63% from ENCFF450VAU in the PacBio results in Table 4. At the intron level, sensitivity and precision are relatively high for all the datasets when compared to the different levels. In fact, there is almost perfect precision for NA12878-IVT_RNA at 98.1% in Table 3. There are slight improvements in these metrics for the annotated datasets. The base level metrics are a bit lower than the exon and intron level results for all datasets but still modest. Across the board, there is a dip in sensitivity and precision for intron chain levels, indicating challenges in identifying longer intronic structures. Although, StringTie performed better on this level with the PacBio datasets than those of Nanopore. Recall and precision percentages are even lower across the board at the transcript level, showing that StringTie struggles with full-length transcript identification and many identified transcripts may not be true positives. At the locus level, the only significant point to note is that recall and precision are higher with annotation. Overall, the Nanopore datasets show higher sensitivity in identifying transcripts at the exon and intron levels when annotated, while the PacBio datasets perform better at the locus level.

Table 3: StringTie Using Human Nanopore Data Aligned with Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
StringTie	Base Level	18.9	74.9	35.5	75.3	9.6	92.5
	Exon Level	22.4	79.9	37.6	83.3	14.5	92.1
	Intron Level	24.6	90.8	43.8	88.5	15.6	98.1
	Intron Chain Level	8.4	36.0	17.1	42.4	5.4	56.4
	Transcript Level	7.3	32.8	14.4	41.5	4.6	55.6
	Locus Level	11.2	39.5	19.8	60.4	7.7	56.8
	Average	15.47	58.98	28.03	65.23	9.57	75.25

Table 4: StringTie Using Human PacBio Data Aligned with Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
StringTie	Base Level	25.7	64.9	24.8	60.3
	Exon Level	31.4	74.9	30.0	72.2
	Intron Level	36.9	81.6	35.2	79.7
	Intron Chain Level	12.9	34.3	12.5	31.8
	Transcript Level	10.4	33.9	10.1	31.1
	Locus Level	14.1	44.2	13.7	40.0
	Average	21.90	55.63	21.05	52.52

Table 5: StringTie Using Human Nanopore Data Aligned without Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
StringTie	Base Level	18.9	74.9	34.6	76.5	9.5	92.4
	Exon Level	22.4	79.9	36.5	82.5	14.2	90.4
	Intron Level	24.6	90.8	42.4	87.9	15.1	97.0
	Intron Chain Level	8.4	36.0	15.5	39.3	4.9	49.2
	Transcript Level	7.3	32.8	13.1	38.8	4.3	49.1
	Locus Level	11.2	39.5	18.2	57.6	7.2	51.2
	Average	15.47	58.98	26.72	63.77	9.20	71.55

Table 6: StringTie Using Human PacBio Data Aligned without Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
StringTie	Base Level	25.7	64.9	24.8	60.3
	Exon Level	31.4	74.4	29.9	71.7
	Intron Level	36.8	81.1	35.1	79.3
	Intron Chain Level	12.7	33.3	12.3	31.0
	Transcript Level	10.2	32.8	10.0	30.3
	Locus Level	14.0	43.7	13.6	39.6
	Average	21.80	55.03	20.95	52.03

Scallop

Looking at all the levels of assembly, Scallop has notably low sensitivity levels for the Nanopore and PacBio datasets. That being said, Scallop performed better on the PacBio datasets as it was able to register more meaningful information, resulting in average sensitivities of 0.57% and 0.78% (Table 8). For the Nanopore datasets, the precision is notably low at the intron chain, transcript, and locus levels. It performed poorly on NA12878-IVT_RNA in particular both on the annotated and unannotated sets. However, the precision is high at the base level for NA12878-cDNA and NA12878-DirectRNA especially in the unannotated datasets at 88.3% and 92.1% (Table 7) respectively. The precision is also at 100.0% at the intron level (Table 7) for the annotated and unannotated NA12878-DirectRNA datasets. As for the precision in both the annotated and unannotated NA12878-cDNA datasets, -nan appears because there

was no appropriate intron information for Scallop to compute those metrics. The inconsistency in precision in different levels across the Nanopore datasets highlights the need for further testing on different datasets. While some predictions may be correct, the tool fails to identify many true transcripts in the Nanopore datasets. Upon examining the precision for ENCFF450VAU and ENCFF694DIE, it is generally higher when the dataset it annotated. For example, the average precision for ENCFF450VAU annotated was 68.33% (Table 8) and 68.05% (Table 10) for unannotated. The precision at the intron level—84.0% and 78.5% (Table 8)—is quite high and comparable to those in StringTie at that level, suggesting that when Scallop identifies intronic sequences, it is more likely to be correct. In general, the consistently low sensitivity and variable precision indicate that Scallop may not be the best choice when working with longer reads typical of Nanopore sequencing.

Table 7: Scallop Using Human Nanopore Data Aligned with Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Scallop	Base Level	0.0	83.4	0.0	87.6	0.0	0.0
	Exon Level	0.0	0.0	0.0	19.0	0.0	0.0
	Intron Level	0.0	-nan	0.0	100.0	0.0	0.0
	Intron Chain Level	0.0	-nan	0.0	0.0	0.0	0.0
	Transcript Level	0.0	0.0	0.0	5.6	0.0	0.0
	Locus Level	0.0	0.0	0.0	5.9	0.0	0.0
	Average	0	20.85	0	36.35	0	0

Table 8: Scallop Using Human PacBio Data Aligned with Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
Scallop	Base Level	0.6	69.8	1.0	63.1
	Exon Level	0.5	68.6	0.7	61.8
	Intron Level	0.4	84.0	0.6	78.5
	Intron Chain Level	0.7	62.4	0.8	51.2
	Transcript Level	0.5	57.4	0.7	45.6
	Locus Level	0.7	67.8	0.9	55.1
	Average	0.57	68.33	0.78	59.22

Table 9: Scallop Using Human Nanopore Data Aligned without Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Scallop	Base Level	0.0	88.3	0.0	92.1	0.0	0.0
	Exon Level	0.0	11.1	0.0	31.2	0.0	0.0
	Intron Level	0.0	-nan	0.0	100.0	0.0	0.0
	Intron Chain Level	0.0	-nan	0.0	0.0	0.0	0.0
	Transcript Level	0.0	0.0	0.0	15.4	0.0	0.0
	Locus Level	0.0	0.0	0.0	16.7	0.0	0.0
	Average	0	24.85	0	42.57	0	0

Table 10: Scallop Using Human PacBio Data Aligned without Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
Scallop	Base Level	0.6	70.0	1.0	63.0
	Exon Level	0.5	68.6	0.7	61.7
	Intron Level	0.4	83.3	0.6	78.4
	Intron Chain Level	0.7	61.8	0.8	51.1
	Transcript Level	0.5	56.8	0.7	45.4
	Locus Level	0.8	67.8	0.9	54.8
	Average	0.58	68.05	0.78	59.07

PsiCLASS

Due to the complexity of processing and aligning the Nanopore and PacBio data, the analysis for PsiCLASS could not be completed in a timely manner. This additional necessary processing resulted in an inability to obtain conclusive results for its performance metrics. Consequently, while this tool shows promise in handling RNA sequencing data, further exploration and analysis are necessary to fully assess its capabilities within this context.

Based on previous research, PsiCLASS has demonstrated significant advancements in its ability to handle RNA sequencing data, particularly in precision. It achieved a high precision rate on simulated

datasets, surpassing both StringTie and Scallop. This improvement highlights PsiCLASS's effectiveness in accurately identifying transcripts, making it a valuable tool for RNA analysis.

StringTie2

First, examining the average sensitivity, the percentages vary across the Nanopore annotated datasets with averages of 8.73%, 23.63%, and 6.22% (Table 11). This pattern in sensitivity is visible at the base, exon, and intron levels. The precision levels are higher on average at the base, exon, and intron levels compared to StringTie and Scallop. The difference between the annotated and unannotated datasets is not very significant, but like in the results of previous tools, sensitivity and precision increase with annotations. However, when comparing the results from the Nanopore data to PacBio, in general the sensitivity and precision are both higher for PacBio, indicating that StringTie2 is more reliable for PacBio data. This is shown in Table 12 with average recall of 16.02% and 15.52% for ENCFF450VAU and ENCFF694DIE and average precision of 72.12% and 67.68%. StringTie2 is quite strong in detecting accurate intron sequences with annotation at 94.9% and 94.1% (Table 12) for ENCFF450VAU and ENCFF694DIE. It struggles with sensitivity for all datasets at the intron chain, transcript, and locus levels. This highlights difficulties in capturing longer intronic structures, identifying full-length transcripts, and marking locus structures effectively. The tradeoff for high precision is evident especially in the PacBio data when comparing the locus and transcript level precision to those of StringTie. For example, in StringTie the sensitivity for the annotated ENCFF450VAU dataset at the locus level is 14.1% and precision is 44.2% (Table 4) where sensitivity is 11.9% and precision is 61.0% for StringTie2 (Table 12). The analysis of StringTie2 highlights the tool's strengths in identifying exon and intron features, particularly in PacBio datasets. However, the consistently low sensitivity across transcript levels suggests that StringTie2 lacks strong effectivity in identifying many true transcripts.

Table 11: StringTie2 Using Human Nanopore Data Aligned with Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
StringTie2	Base Level	11.5	60.1	31.7	72.6	5.9	88.7
	Exon Level	11.2	68.1	31.6	82.9	8.7	89.1
	Intron Level	11.9	85.4	37.0	91.3	9.1	96.6
	Intron Chain Level	5.1	42.1	13.2	50.0	3.9	75.4
	Transcript Level	4.7	29.9	11.2	45.0	3.5	66.1
	Locus Level	8.0	33.4	17.1	57.4	6.2	69.9
	Average	8.73	53.17	23.63	66.53	6.22	80.97

Table 12: StringTie2 Using Human PacBio Data Aligned with Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
StringTie2	Base Level	19.3	75.4	19.0	69.3
	Exon Level	22.2	86.9	21.4	84.4
	Intron Level	25.6	94.9	24.7	94.1
	Intron Chain Level	9.4	62.0	9.0	58.6
	Transcript Level	7.7	52.5	7.4	46.4
	Locus Level	11.9	61.0	11.6	53.3
	Average	16.02	72.12	15.52	67.68

Table 13: StringTie2 Using Human Nanopore Data Aligned without Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
StringTie2	Base Level	11.5	60.1	29.6	73.3	5.1	87.2
	Exon Level	11.2	68.1	29.0	80.9	7.2	86.5
	Intron Level	11.9	85.4	33.7	89.1	7.4	94.5
	Intron Chain Level	5.1	42.1	11.6	44.5	3.3	68.2
	Transcript Level	4.7	29.9	9.9	40.7	3.1	60.2
	Locus Level	8.0	33.4	15.5	54.7	5.5	66.0
	Average	8.73	53.17	21.55	63.87	5.27	77.10

Table 14: StringTie2 Using Human PacBio Data Aligned without Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
StringTie2	Base Level	19.2	75.3	18.8	69.1
	Exon Level	21.9	86.7	21.1	84.1
	Intron Level	25.3	94.8	24.4	93.8
	Intron Chain Level	9.2	61.4	8.9	57.8
	Transcript Level	7.6	51.9	7.3	45.8
	Locus Level	11.8	60.3	11.4	52.6
	Average	15.83	71.73	15.32	67.20

Scallop2

On the Nanopore datasets, Scallop2 performs similarly to StringTie2 in terms of precision but has increased sensitivity especially at the intron and exon levels. This indicates that Scallop2 might be more reliable on Nanopore data if a balance between sensitivity and precision is required in the assembly task. Precision at the base level is relatively high for the annotated NA12878-cDNA and NA12878-IVT_RNA datasets (80.1% and 96.3% in Table 15) and lower for the PacBio datasets ENCFF450VAU and ENCFF694DIE (73.9% and 70.3% in Table 16), indicating that when Scallop2 makes predictions, it is generally reliable, especially for Nanopore data. Just as it does for the Nanopore datasets, Scallop2 improves in sensitivity at the exon and intron levels for the PacBio data. At the intronic chain level, both recall and precision are lacking compared to all the other categories for all of the datasets. For example, sensitivity at the intron chain level is lower, with 7.7% for the annotated NA12878-cDNA dataset (Table 15) and 13.4% for the annotated ENCFF450VAU dataset (Table 16), indicating difficulties in capturing longer intronic structures. Using the same examples, precision is moderately lower at this level, at 38.8% for NA12878-cDNA (Table 15) and 29.3% for ENCFF450VAU (Table 16), showing some inaccuracies in predictions. Scallop2 also struggles on the transcript level with full-length transcript identification. However, it performs better on the Nanopore and PacBio datasets at this level compared to StringTie2. Despite difficulties projected at the intron chain and transcript level, recall and precision show improvement at the locus level suggesting better identification of locus structures compared to other levels. Overall, the performance metrics for PacBio datasets are consistently higher across most levels, suggesting that Scallop2 may be more suited to handle this type of sequencing data. The annotated datasets show improved sensitivity and precision metrics across all levels, indicating that prior knowledge enhances the tool's ability to identify transcript features effectively.

Table 15: Scallop2 Using Human Nanopore Data Aligned with Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Scallop2	Base Level	11.3	80.1	32.1	65.0	5.6	96.3
	Exon Level	17.3	80.8	34.3	72.0	10.0	92.4
	Intron Level	19.5	90.0	42.0	83.2	11.0	97.2
	Intron Chain Level	7.7	38.8	18.6	24.7	4.5	69.4
	Transcript Level	6.0	38.7	14.6	24.6	3.5	69.3
	Locus Level	8.5	60.8	16.9	71.9	5.5	81.9
	Average	11.72	64.87	26.42	56.90	6.68	84.42

Table 16: Scallop2 Using Human PacBio Data Aligned with Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
Scallop2	Base Level	20.6	73.9	19.7	70.3
	Exon Level	26.4	77.5	25.4	75.3
	Intron Level	32.2	83.6	31.1	81.8
	Intron Chain Level	13.4	29.3	13.1	26.8
	Transcript Level	10.5	29.3	10.2	26.8
	Locus Level	12.4	77.3	11.9	73.5
	Average	19.25	61.82	18.57	59.08

Table 17: Scallop2 Using Human Nanopore Data Aligned without Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Scallop2	Base Level	13.5	76.5	31.0	67.3	5.5	96.0
	Exon Level	19.4	72.9	33.4	68.8	9.9	88.8
	Intron Level	22.0	82.9	40.5	79.4	10.8	93.8
	Intron Chain Level	8.0	27.2	16.4	20.8	4.2	55.3
	Transcript Level	6.3	27.2	12.8	20.7	3.3	55.2
	Locus Level	9.0	53.5	15.7	67.1	5.1	76.4
	Average	13.03	56.70	24.97	54.02	6.47	77.58

Table 18: Scallop2 Using Human PacBio Data Aligned without Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
Scallop2	Base Level	20.6	74.0	19.6	70.4
	Exon Level	26.3	76.8	25.3	74.7
	Intron Level	32.1	83.3	30.9	81.6
	Intron Chain Level	13.1	28.4	12.8	26.1
	Transcript Level	10.2	28.4	10.0	26.1
	Locus Level	12.3	76.5	11.8	72.8
	Average	19.10	61.23	18.40	58.62

IsoQuant

For IsoQuant, the average resulting sensitivities on the Nanopore data in Table 19 of 6.90%, 19.67%, and 4.20% are quite low compared to other tool average sensitivities. That being said, IsoQuant shows higher sensitivity in PacBio data such as 14.55% and 13.80% in Table 20. The sensitivity at the base level for the Nanopore datasets is extremely low, suggesting that IsoQuant lacks effectivity in identifying many true transcripts on the individual nucleotide basis. Precision at the base level is relatively high for the Nanopore data and lower for the PacBio data, indicating that when predictions are made, they are generally reliable, especially for Nanopore data. However, at the exon level, IsoQuant performs better at detecting exon features in PacBio data, striking a better balance between recall and precision. Sensitivity improves at the intron level, for both data groups. For example, the annotated NA12878-cDNA data reaches 10.1% (Table 19) and 23.2% (Table 20) for the annotated ENCFF450VAU data at the intron level. This indicates a better capacity for identifying intronic sequences, especially in the PacBio data. Precision is also high, suggesting that introns are predicted with high accuracy. Like Scallop2 and StringTie2, sensitivity and precision at the intron chain and transcript levels are lower. This reduction seems to be a trend amongst the latest assemblers, showing that one of the most seemingly difficult tasks is to identify accurate full-length transcripts and long intronic sequences. However, the precision in these categories especially for the Nanopore datasets are much higher than that of Scallop2 and StringTie2 with IsoQuant's average precision on these datasets ranging from low 80% to mid 90% in Table 19. Like Scallop2, IsoQuant shows improvement in the locus level with precision much greater than Scallop2 on all of the datasets. IsoQuant's high precision sets it apart as a valuable tool for researchers prioritizing precision in their analyses.

Table 19: IsoQuant Using Human Nanopore Data Aligned with Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
Isoquant	Base Level	5.6	89.2	24.0	85.1	3.1	97.9
	Exon Level	9.8	89.7	26.8	89.3	6.0	95.5
	Intron Level	10.1	96.3	30.6	94.4	6.1	99.2
	Intron Chain Level	5.2	67.2	12.6	60.1	3.2	91.6
	Transcript Level	4.1	67.2	9.8	60.1	2.5	91.6
	Locus Level	6.6	81.2	14.2	83.9	4.3	95.1
	Average	6.90	81.80	19.67	78.82	4.20	95.15

Table 20: IsoQuant Using Human PacBio Data Aligned with Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
IsoQuant	Base Level	15.3	84.9	14.4	82.6
	Exon Level	20.4	84.8	19.3	83.1
	Intron Level	23.2	90.8	21.9	89.4
	Intron Chain Level	9.9	49.6	9.5	46.8
	Transcript Level	7.8	49.5	7.5	46.7
	Locus Level	10.7	77.9	10.2	75.2
	Average	14.55	72.92	13.80	70.63

Table 21: IsoQuant Using Human Nanopore Data Aligned without Annotation

		NA12878-cDNA		NA12878-DirectRNA		NA12878-IVT_RNA	
		Sensitivity	Precision	Sensitivity	Precision	Sensitivity	Precision
IsoQuant	Base Level	6.2	87.6	22.7	85.9	3.0	97.9
	Exon Level	10.4	87.6	25.4	89.2	5.7	94.8
	Intron Level	10.8	94.8	28.8	94.1	5.7	98.5
	Intron Chain Level	5.5	61.2	11.3	58.6	2.9	87.7
	Transcript Level	4.3	61.2	8.9	58.6	2.3	87.7
	Locus Level	6.8	75.8	13.0	80.1	4.0	91.2
	Average	7.33	78.03	18.35	77.75	3.93	92.97

Table 22: IsoQuant Using Human PacBio Data Aligned without Annotation

		ENCFF450VAU		ENCFF694DIE	
		Sensitivity	Precision	Sensitivity	Precision
IsoQuant	Base Level	15.3	84.8	14.3	82.6
	Exon Level	20.4	84.6	19.3	82.9
	Intron Level	23.2	90.7	21.9	89.3
	Intron Chain Level	9.8	49.1	9.4	46.3
	Transcript Level	7.7	49.1	7.4	46.3
	Locus Level	10.6	77.3	10.1	74.4
	Average	14.50	72.60	13.73	70.30

Comparison of Results

StringTie generally showed higher precision, particularly for exon and intron levels across the Nanopore and PacBio datasets. Its precision, although better than Scallop's, indicates that there is still potential for reducing the occurrence of false positives. StringTie struggled with full-length transcript identification, showing notably low precision and sensitivity at the transcript level. StringTie's performance was also challenged at the locus level, with sensitivity showing improvements when aligned with annotated datasets. That being said, previous research has highlighted that StringTie outperforms earlier assemblers like Cufflinks in terms of transcript assembly, showing a significant increase in the number of transcripts identified (Pertea et al., 2015).

Scallop performed poorly across all levels compared to the other tools, particularly in sensitivity, suggesting that it may not be well-suited for long-read datasets, such as those from Nanopore. It showed high precision in base-level transcript detection, but the tool struggled with more complex transcript features, especially at the intron chain and transcript levels. Previous studies have demonstrated that Scallop outperforms StringTie in some respects, such as producing 34.5% more multi-exon transcripts and 67.5% more lowly expressed transcripts with higher accuracy (Shao & Kingsford, 2017). However, this study was unable to capture similar results, and it may suggest that Scallop's lower sensitivity in capturing true transcripts limits its applicability to longer-read data, particularly from Nanopore. It also calls for further investigation of Scallop's abilities on more data.

StringTie2 improved significantly over its predecessor, particularly in precision, which was overall higher than StringTie by 6.77% according to Table 2. The algorithmic improvements made StringTie2 particularly strong in detecting exon and intron features, particularly in PacBio datasets. Sensitivity, however, remained low across transcript and locus levels, indicating difficulty in capturing

full-length transcripts and larger structural features. While StringTie2's performance was generally better than Scallop and StringTie, especially with PacBio data, it still had room for improvement in identifying more complex transcript structures. This is supported by previous research, which has shown that StringTie2 outperformed Scallop, with a 3.9% increase in sensitivity and a 47.3% increase in precision on the same datasets used by Scallop (Zhang et al., 2021).

Scallop2 performed better than Scallop, showing a balance between sensitivity and precision. It demonstrated improved performance over StringTie2, particularly in PacBio datasets, where Scallop2 achieved higher sensitivity across exon and intron levels. Compared to Scallop and StringTie2, Scallop2's overall performance was more reliable when a balance between precision and recall is required. This aligns with findings from earlier research, which showed that Scallop2 improved by 0.7% in matching transcripts and increased precision by 12.5% over Scallop and 107.1% over StringTie2 (Zhang et al., 2021).

IsoQuant stood out for its exceptionally high precision, particularly in detecting true positives across the datasets. While IsoQuant showed one of the lowest sensitivities out of all tools at 11.70% shown in Table 2, it demonstrated decent overall performance at the exon and intron levels, particularly with PacBio data. IsoQuant's key strength of eliminating false positives is verified by previous research as it produced significantly fewer false positive transcripts compared to StringTie (Prjibelski et al., 2023). In this study, IsoQuant's precision, especially for Nanopore datasets, was much higher than that of Scallop2 and StringTie2, further emphasizing its strength in transcript identification. Moreover, IsoQuant produced 71% more new transcripts that were confirmed through additional analysis, showcasing its utility for researchers prioritizing precision in RNA sequencing (Prjibelski et al., 2023).

Limitations

There are several limitations that may affect the results and interpretations of the RNA transcript assembly performance. Potential biases may arise from the inherent characteristics of the datasets used. The analysis relied on a limited number of RNA sequencing datasets from Nanopore and PacBio, which may not fully represent the diversity of RNA molecules. The size and variability of the datasets may also influence the robustness of the findings, as larger and more diverse datasets could provide a more comprehensive assessment of each tool's capabilities.

Furthermore, specific tools exhibited limitations that impacted their performance assessments. For example, Scallop demonstrated notably low sensitivity when applied to long Nanopore data, suggesting it may not be ideally suited for handling such sequencing technology. This limitation highlights the need for careful consideration when selecting tools for different RNA sequencing platforms.

Additionally, the extra processing and alignment of existing data required for PsiCLASS hindered the ability to conduct a thorough analysis of the assembler. This resulted in incomplete evaluations and a lack of conclusive results for PsiCLASS. Future research would benefit from allocating time for pre-processing tasks to fully explore the capabilities of PsiCLASS and to mitigate biases.

Future Directions

Future research should incorporate a wider variety of long read RNA-seq datasets, including those from different biological contexts and additional sequencing platforms. This would provide a more comprehensive evaluation of the performance of the transcript assemblers, allowing for a better understanding of their strengths and weaknesses across diverse RNA molecules.

Given the observed limitations of tools like Scallop when applied to long Nanopore data, further studies could investigate its performance on other long-read datasets. This could also involve optimizing

existing algorithms such as the splice graph methodology, phasing path creation, or vertex decomposition to handle the unique characteristics of long-read RNA sequencing technologies. Additionally, exploring Scallop's capabilities in the context of short-read data may provide more helpful insights into its performance and utility in different sequencing environments. Additionally, the time constraints faced during this study limited the ability to fully analyze PsiCLASS. Future investigations should prioritize the extra processing and alignment of Nanopore and PacBio data to conduct a thorough assessment of PsiCLASS's capabilities. This would provide valuable insights for comparing PsiCLASS against other tools.

Future research should also consider testing the assembled GTF files against alternative reference annotations, such as Ensembl or RefSeq, to further evaluate the robustness of transcript identification. Utilizing different reference annotations can provide a broader context for assessing the quality of the assemblies. Comparing results across these annotations could highlight discrepancies and improve understanding of the complexities present in transcriptomic data. Additionally, such an approach may show the strengths and limitations of assembler in relation to the chosen reference, potentially helping researchers select the most appropriate methodologies for their specific applications.

Developing more refined evaluation metrics that capture the nuances of transcriptomic data could enhance the assessment of transcript assemblers. Instead of relying solely on metrics like sensitivity and precision, future research could incorporate tools such as SQANTI3, which provides a detailed, quality assessment of transcript assemblies. Additionally, a more in-depth evaluation of memory usage and runtime for each tool could provide valuable information about the computational efficiency and scalability of each assembler. Doing so may aid in selecting the most suitable tool for large-scale transcriptomic studies. By leveraging such comparison tools, researchers can gain deeper insights into the performance of different assemblers, helping to identify which tools are most effective for accurately representing biological occurrences.

Chapter 5

Conclusion

This study has provided an evaluation of six RNA transcript assemblers—StringTie, Scallop, PsiCLASS, StringTie2, Scallop2, and IsoQuant—utilizing long-read RNA sequencing datasets from both Oxford Nanopore Technology and Pacific Biosequences platforms. The analysis revealed distinct performance characteristics among the assemblers, in terms of sensitivity and precision, which are critical for accurate transcript identification.

StringTie demonstrated strong performance at the exon and intron levels, although it faced limitations in sensitivity that could hinder its effectiveness in certain applications. StringTie2 improved precision over its predecessor, effectively minimizing false positives. In contrast, Scallop's low sensitivity highlights its challenges, particularly when applied to long-read data, suggesting that it may not be the ideal choice for all sequencing technologies. However, Scallop2 appeared as a balanced option, maintaining relatively high levels of both sensitivity and precision, making it a suitable alternative for diverse research needs. IsoQuant stood out for its consistently high precision across all analyzed levels, outperforming both Scallop2 and StringTie2 on this level.

The findings highlight the significance of using annotated datasets and a comprehensive reference genome, which can enhance the performance of RNA transcript assemblers. The insights gained from this research can inform future efforts to refine transcript assembly methods and improve the accuracy of gene expression studies. Overall, this work contributes to the ongoing advancement of RNA transcriptomic analysis to understand gene regulation and its implications in health and disease.

BIBLIOGRAPHY

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). From DNA to RNA. In *Molecular Biology of the Cell. 4th edition*. Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK26887/

GENCODE - Human Release 36 Statistics. (n.d.). Retrieved October 31, 2024, from https://www.gencodegenes.org/human/stats_36.html

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

Liu, Z., Quinones-Valdez, G., Fu, T., Huang, E., Choudhury, M., Reese, F., Mortazavi, A., & Xiao, X. (2023). L-GIREMI uncovers RNA editing sites in long-read RNA-seq. *Genome Biology*, *24*(1), 171. https://doi.org/10.1186/s13059-023-03012-w

Pardo-Palacios, F. J., Wang, D., Reese, F., Diekhans, M., Carbonell-Sala, S., Williams, B., Loveland, J. E., De María, M., Adams, M. S., Balderrama-Gutierrez, G., Behera, A. K., Gonzalez, J. M., Hunt, T., Lagarde, J., Liang, C. E., Li, H., Jerryd Meade, M., Moraga Amador, D. A., Prjibelski, A. D., … Brooks, A. N. (2023). Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *bioRxiv*, 2023.07.25.550582. https://doi.org/10.1101/2023.07.25.550582

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-C., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295.

Prjibelski, A. D., Mikheenko, A., Joglekar, A., Smetanin, A., Jarroux, J., Lapidus, A. L., & Tilgner, H. U. (2023). Accurate isoform discovery with IsoQuant using long reads. *Nature Biotechnology*, *41*(7), 915–918.

Shao, M., & Kingsford, C. (2017). Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, *35*(12), 1167–1169.

Song, L., Sabunciyan, S., Yang, G., & Florea, L. (2019). A multi-sample approach increases the accuracy of transcript assembly. *Nature Communications*, *10*(1), 5000. https://doi.org/10.1038/s41467-019-12990-0

Workman, R. E., Tang, A. D., Tang, P. S., Jain, M., Tyson, J. R., Razaghi, R., Zuzarte, P. C., Gilpatrick, T., Payne, A., Quick, J., Sadowski, N., Holmes, N., de Jesus, J. G., Jones, K. L., Soulette, C. M., Snutch, T. P., Loman, N., Paten, B., Loose, M., … Timp, W. (2019). Nanopore native RNA sequencing of a human poly(A) transcriptome. *Nature Methods*, *16*(12), 1297–1305. https://doi.org/10.1038/s41592-019-0617-2

Zhang, Q., Shi, Q., & Shao, M. (2021). Scallop2 enables accurate assembly of multiple-end RNA-seq data. *BioRxiv*, 2021–09

Academic Vita

Saadya Rao

EDUCATION

Pennsylvania State University, Schreyer Honors College, College of Engineering August 2021 – Dec 2024
Bachelor of Science in Computer Science | *University Park, PA*

Relevant Coursework: Data Structures & Algorithms, Systems and Socket Programming, Deep Learning for Computer Vision, Object Oriented Programming, Computer Organization, Numerical Analysis

Technology: Google Cloud, Amazon Web Services, Linux, Git, Jupyter Notebook, OpenCV
Programming Languages: C/C++, Kotlin, Java, Python, Flutter, SQL, R

EXPERIENCE

Capital One | *Android Development* June – August 2024
Software Engineer Intern Plano, TX

- Led the development of a highly customizable and injectable WebView component for Capital One's AutoNavigator mobile application, enabling seamless integration across various features, including the Stand Alone Trade-in module
- Researched and utilized certificate pinning concepts to ensure secure loading of authorized URLs in the WebView
- Implemented advanced session transfer and route management for seamless native and external navigation
- Expanded functionality for thousands of Android users by deploying a new feature through the injectable component

MIAX Exchange Group | *Trading Systems Development* June – July 2023
Software Engineer Intern Princeton, NJ

- Built a full stack matching engine and interface application using C++ STL, Boost Libraries, and UNIX Socket API
- Developed complex order storing and matching algorithms to handle client trades
- Added functionality to support best bid and offer calculations to display after every transaction

Pennsylvania State University | *Nittany AI Advance* January 2023 – February 2024
Software and AI Engineer Intern University Park, PA

- Designed an application to detect falls and overcrowding by applying machine learning models to Septa CCTV footage
- Assembled a live alert system for Septa authorities as an initiative to protect train users and prevent crime
- Trained YOLOv8's object detection model on custom platform data provided by Septa

PROJECTS

Reference-Based RNA Transcript Assembly January – November 2024
Undergraduate Research Assistant

- Under supervision of Prof. Mingfu Shao, evaluated the performance of six RNA transcript assemblers on long-read RNA sequencing datasets from Oxford Nanopore and Pacific Biosequences platforms, focusing on sensitivity and precision

- Analyzed the results to write a comprehensive thesis, contributing to the understanding of transcript assembler performance across various RNA sequencing datasets

Oris October 2022 – October 2023

Co-Founder and Lead AI Engineer

- Engineered a 94% accurate convolutional neural network to detect Gingivitis in oral images taken by phone
- Designed a device to take oral phone images to give the model, built using CAD and approved by dentists
- Won 2nd place out of 60 teams in the Nittany AI Challenge for delivering an iOS system as a service app-store ready product

Project Eleos January – September 2022

Software Engineer

- Constructed a 92% accurate NLP app, to aid therapy services, using GCP, Flutter, and asynchronous programming
- Established design goals to reflect an understanding of customer base via communicating with stakeholders and experts

LEADERSHIP

Penn State Natya January 2023 – May 2024

Fundraising Director

- Coordinated events to raising funds for Natya, Penn State's competitive Indian classical dance team
- Created monetary goals to ensure that competition dues are met for each dancer