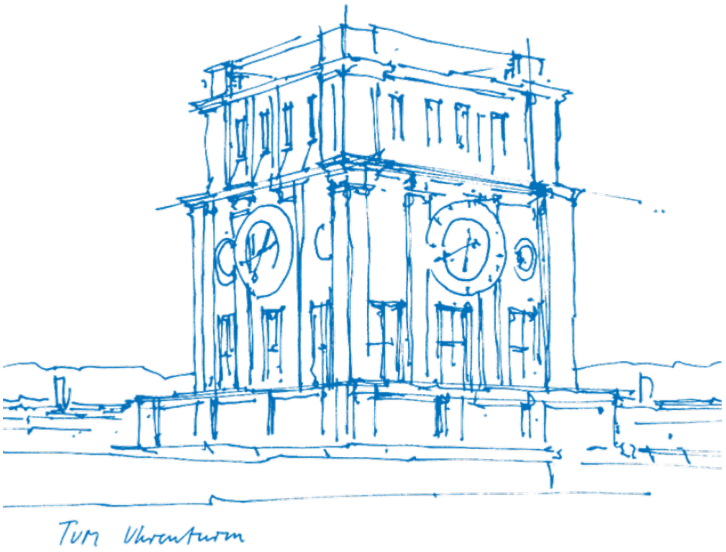
TUM School of Life Sciences
Technical University of Munich

# Towards uncovering Transcriptomic Dynamics through Computational Analysis

Chit Tong Lio



TUM School of Life Sciences
Technische Universität München

# Towards uncovering Transcriptomic Dynamics through Computational Analysis

Chit Tong Lio

Vollständiger Abdruck der von der TUM School of Life Sciences der Technischen Universität München zur Erlangung einer

**Doktorin der Naturwissenschaften (Dr. rer. nat.)**

genehmigten Dissertation.

**Vorsitz:**
Prof. Dr. Mathias Wilhelm

**Prüfende:**

1. Prof. Markus List, Ph.D.
2. Prof. Dr. Jan Baumbach

Die Dissertation wurde am 14.06.2024 bei der Technischen Universität München eingereicht und durch die TUM School of Life Sciences am 14.08.2024 angenommen.

*This thesis is dedicated to my family and loved ones, who supported this journey.*

# Acknowledgement

First, I would like to express my greatest appreciation to my family, who supported me on this journey. My family in Macau has been providing me with love and education so that I have enough resilience and patience to finish my dissertation today. And my other family in Germany, who provided me the greatest support from my life partner Timotheus Fischer and my best friends of all time and space, Amit Fenn and Vanda Marosi.

I sincerely appreciate Prof. Dr. Jan Baumbach for giving me the opportunity to make this PhD happen and for providing me with resources and an inspiring environment. I want extened my deepest thanks to Prof. Dr. Markus List, Dr. Olga Tsoy and Prof. Dr. Tim Kacprowski for their supervision and invaluable expertise in shaping this work. I also want to express my gratitude to Prof. Dr. Dmitrij Frischman for his invaluable suggestions and comments during my TAC meetings.

I want to thank all my friends and fellow colleagues who made this journey extra fun, especially those who spent time reviewing my thesis: Zakaria, Manuela, Klaudia, Lena-Maria, Johannes, Lena, and Lisi.

I have a special appreciation for my daughter (I call her Peanut for now), who came into my world at the end of this PhD journey.

# Abstract

The study of ribonucleic acid (RNA) contents in cells aims to elucidate the relationships among gene expression, cell states, and cell functions under specific conditions. Ultimately, the study field addresses the question of which genes drive the cell's condition. To answer this question, researchers have developed experimental and computational tools to facilitate the field of study. RNA sequencing allows the quantification of RNA expression in biological samples. Applying computational methods in systems biology allows us to study the interactions of each component inside a cell. The interactions of these components within a system usually happen dynamically, e.g., the expression of a transcription factor can affect the transcription of target genes over time. These dynamic components are often ignored in many experiments. Methods for time series analysis still need to be improved, especially in the study of alternative splicing.

Three key research areas are discussed in this dissertation: 1) the investigation of potential snoRNA and miRNA candidates associated with Alzheimer's disease using differential co-expression network analysis, 2) the development of Spycone—a tool facilitating splicing-aware time-series network analysis at the transcript level, and 3) benchmarking analysis of existing tools for differential transcript usage (DTU) detection in both two-condition experiments, time-series data, and single-cell data. Additionally, guidelines for DTU analysis of bulk RNA-seq data are provided.

In the first publication, I demonstrated the use of a data-driven network - differential gene co-expression network - in deciphering the involvement of small RNAs in the Alzheimer's mouse model. One of the main difficulties in identifying potential snoRNA and miRNA candidates is the lack of prior knowledge, namely interaction information and functional annotation. Traditionally, differential gene expression analysis can help us identify genes that change expression levels between conditions. However, by treating genes independently, this method ignores the co-dependence environment within a biological system. I constructed a data-driven network specific to the Alzheimer's disease Tg4-42 mouse model. The resulting network represents the differential relationship of small RNAs between the wild-type and Alzheimer's mouse models. Extracting important nodes using centrality measures allows the identification of potential small RNA biomarkers involved in developing Alzheimer's disease-like phenotype in the mouse model. These findings can help to establish new connections to the known mechanisms of Alzheimer's disease progression.

In the second publication, I introduced Spycone, a framework for systematic analysis in alternative splicing for time series data. In systems biology, gene-level analysis is typically given more attention; however, alternative splicing plays a crucial role in determining protein diversity and function. Additionally, alternative splicing is a dynamic process that changes during the development of an organism or in a disease like cancer. Time-series data are often used to study dynamic processes. However, only one tool, TSIS, is specific for isoform switch detection in time series data. One of the challenges is that no tool provides downstream analysis in alternative splicing for time series data. I developed Spycone, which includes a novel algorithm for detecting isoform switch events in time-series transcriptomics and a framework for systematically analyzing time-series data. The framework consists of four main analyses: clustering integrating multiple algorithms, functional enrichment incorporating NEASE specialized for splicing, network enrichment incorporating DOMINO, and splicing factor motifs enrichment. This work outperformed the competing tool TSIS with simulated data based on a Hidden Markov Model. I provided evidence of the biological relevance of the findings for analyzing a time series dataset from a SARS-Cov-2 infected cell line.

In an unpublished work, I conducted a benchmark analysis on published DTU detection tools. These tools aim to identify genes whose transcription distribution changes between conditions. These changes are assumed to be due to alterations in alternative splicing patterns. In this analysis, I evaluated the tools in three types of data: simulated and real-world bulk RNA-seq data, real-world time series data and simulated single-cell RNA-seq in multiple settings. I addressed the following questions: How do methods for pairwise comparison compare to time series methods, and how do DTU tools perform in single-cell datasets? I provided an updated perspective and guidelines for performing DTU analysis in these scenarios.

This dissertation strives to advance our understanding of RNA-related processes and their implications in various biological contexts. I showed that using data-driven network reconstruction can compensate for the lack of prior knowledge and extract possible snoRNAs involved in the etiology of Alzheimer's disease. Meanwhile, in Spycone, I developed a novel Python package to analyze alternative splicing in time series data and aid biological interpretation of the results. Finally, the benchmark analysis provides an updated view of the state-of-the-art DTU analysis for bulk RNA-seq data (static and dynamic) and a new perspective for applying DTU analysis in single-cell experiments.

# Kurzfassung

Die Untersuchung der Ribonukleinsäure (RNA)-Gehalte in Zellen zielt darauf ab, die Beziehungen zwischen Genexpression, Zellzuständen und Zellfunktionen unter bestimmten Bedingungen zu erhellen. Letztlich geht es in diesem Bereich um die Frage, welche Gene den Zustand der Zelle steuern. Um diese Frage zu beantworten, haben die Forscher experimentelle und rechnerische Instrumente entwickelt, die das Studienfeld erleichtern. Die RNA-Sequenzierung ermöglicht die Quantifizierung der RNA-Expression in biologischen Proben. Die Anwendung von Berechnungsmethoden in der Systembiologie ermöglicht es uns, die Wechselwirkungen zwischen den einzelnen Komponenten innerhalb einer Zelle zu untersuchen. Die Interaktionen dieser Komponenten innerhalb eines Systems laufen in der Regel dynamisch ab, z. B. kann die Expression eines Transkriptionsfaktors die Transkription von Zielgenen im Laufe der Zeit beeinflussen. Diese dynamischen Komponenten werden bei vielen Experimenten oft ignoriert. Die Methoden für die Zeitreihenanalyse müssen noch verbessert werden, insbesondere bei der Untersuchung des alternativen Spleißens.

In dieser Dissertation werden drei Hauptforschungsbereiche behandelt: 1) die Untersuchung potenzieller snoRNA- und miRNA-Kandidaten, die mit der Alzheimer-Krankheit in Verbindung gebracht werden, unter Verwendung einer differenziellen Koexpressionsnetzwerkanalyse, 2) die Entwicklung von Spycone - einem Tool, das eine spleißfähige Zeitserien-Netzwerkanalyse auf Transkriptebene ermöglicht, und 3) eine Benchmarking-Analyse bestehender Tools für die Erkennung der differenziellen Transkriptnutzung (DTU) in Experimenten mit zwei Bedingungen, Zeitseriendaten und Einzelzelldaten. Zusätzlich werden Leitlinien für die DTU-Analyse von Massen-RNA-seq-Daten bereitgestellt.

In der ersten Veröffentlichung habe ich die Verwendung eines datengesteuerten Netzwerks - eines differenziellen Genkoexpressionsnetzwerks - zur Entschlüsselung der Beteiligung kleiner RNAs am Alzheimer-Mausmodell demonstriert. Eine der Hauptschwierigkeiten bei der Identifizierung potenzieller snoRNA- und miRNA-Kandidaten ist das Fehlen von Vorwissen, d. h. von Interaktionsinformationen und funktioneller Annotation. Traditionell kann uns die differenzielle Genexpressionsanalyse bei der Identifizierung von Genen helfen, deren Expressionsniveau sich zwischen den Bedingungen ändert. Da bei dieser Methode die Gene jedoch unabhängig voneinander behandelt werden, wird die Co-Abhängigkeit innerhalb eines biologischen Systems ignoriert. Ich habe ein datengesteuertes Netzwerk speziell für das Tg4-42-Mausmodell der Alzheimer-Krankheit erstellt. Das resultierende Netzwerk stellt die unterschiedlichen Beziehungen zwischen kleinen RNAs im Wildtyp- und im Alzheimer-Mausmodell dar. Die Extraktion wichtiger Knoten mit Hilfe von Zentralitätsmaßen ermöglicht die Identifizierung potenzieller kleiner RNA-Biomarker, die an der Entwicklung eines der Alzheimer-Krankheit ähnlichen Phänotyps im Mausmodell beteiligt sind. Diese Erkenntnisse können dazu beitragen, neue Verbindungen zu den bekannten Mechanismen des Fortschreitens der Alzheimer-Krankheit herzustellen.

In der zweiten Veröffentlichung habe ich Spycone vorgestellt, einen Rahmen für die systematische Analyse des alternativen Spleißens bei Zeitreihendaten. In der Systembiologie wird der Analyse auf Gen-Ebene in der Regel mehr Aufmerksamkeit geschenkt; alternatives Spleißen spielt jedoch eine entscheidende Rolle bei der Bestimmung der Proteinvielfalt und -funktion. Außerdem ist alternatives Spleißen ein dynamischer Prozess, der sich während der Entwicklung eines Organismus oder bei einer Krankheit wie Krebs verändert. Zeitreihendaten werden häufig zur Untersuchung dynamischer Prozesse verwendet. Es gibt jedoch nur ein Tool, TSIS, das speziell für die Erkennung von Isoformwechseln in Zeitreihendaten geeignet ist. Eine der Herausforderungen besteht darin, dass kein Tool eine nachgeschaltete Analyse des alternativen Spleißens für Zeitreihendaten bietet. Ich habe Spycone entwickelt, das einen neuartigen Algorithmus zur Erkennung von Isoform-Switch-Ereignissen in Zeitserien-Transkriptomdaten und einen Rahmen für die systematische Analyse von Zeitseriendaten umfasst. Der Rahmen besteht aus vier Hauptanalysen: Clustering unter Einbeziehung mehrerer Algorithmen, funktionelle Anreicherung unter Einbeziehung von NEASE, das auf Spleißen spezialisiert ist, Netzwerkanreicherung unter Einbeziehung von DOMINO und Anreicherung von Spleißfaktormotiven. Diese Arbeit übertraf das konkurrierende Tool TSIS mit simulierten Daten, die auf dem Hidden Markov Model basieren. Ich habe die biologische Relevanz der Ergebnisse bei der Analyse eines Zeitreihendatensatzes von einer mit SARS-Cov-2 infizierten Zelllinie nachgewiesen.

In einer unveröffentlichten Arbeit habe ich eine Benchmark-Analyse der veröffentlichten DTU-Erkennungswerkzeuge durchgeführt. Diese Werkzeuge zielen darauf ab, Gene zu identifizieren, deren Transkriptionsverteilung sich zwischen den Bedingungen ändert. Es wird davon ausgegangen, dass diese Änderungen auf Veränderungen der alternativen Spleißmuster zurückzuführen sind. In dieser Analyse bewertete ich die Tools anhand von drei Datentypen: simulierte und reale Massen-RNA-seq-Daten, reale Zeitreihendaten und simulierte Einzelzell-RNA-seq-Daten in verschiedenen Einstellungen. Ich habe die folgenden Fragen untersucht: Wie schneiden die Methoden für den paarweisen Vergleich im Vergleich zu Zeitreihenmethoden ab, und wie schneiden die DTU-Tools bei Einzelzelldatensätzen ab? Ich stellte eine aktualisierte Perspektive und Leitlinien für die Durchführung von DTU-Analysen in diesen Szenarien vor.

Diese Dissertation zielt darauf ab, unser Verständnis von RNA-bezogenen Prozessen und deren Auswirkungen in verschiedenen biologischen Kontexten zu verbessern. Ich habe gezeigt, dass eine datengesteuerte Netzwerkrekonstruktion den Mangel an Vorwissen ausgleichen und mögliche snoRNAs extrahieren kann, die an der Ätiologie der Alzheimer-Krankheit beteiligt sind. In der Zwischenzeit habe ich mit Spycone ein neues Python-Paket entwickelt, um alternatives Spleißen in Zeitreihendaten zu analysieren und die biologische Interpretation der Ergebnisse zu unterstützen. Schließlich bietet die Benchmark-Analyse einen aktualisierten Überblick über den Stand der Technik der DTU-Analyse für Massen-RNA-seq-Daten (statisch und dynamisch) und eine neue Perspektive für die Anwendung der DTU-Analyse in Einzelzellexperimenten.

# Contents

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Motivation

The rapid development of RNA sequencing technology enables scientists to quantify RNA contents in cells, especially messenger ribonucleic acids (mRNAs). The presence of mRNAs indicates gene expression; some of them are post-transcriptionally modified and translated to proteins. Beyond the coding world, non-coding RNAs play pivotal roles in regulating cellular functions despite their lack of translation into proteins. Researchers use RNA sequencing to analyze RNA behavior under certain conditions (e.g., disease, development). Most biological processes and disease developments are dynamic. Where deterministic temporal changes lead to a change to phenotype [1]. Time series data can capture changes and fluctuations in response to stimuli, treatments, disease progression, and biological processes. Therefore, it provides valuable insights into underlying mechanisms and patterns of change. Circadian rhythm is an example of using time series data; 43% of protein-coding genes follow circadian rhythms in their expressions throughout the day [2]. Time series data also facilitate the inference of dynamic gene regulatory networks based on the co-expression of the genes and their targets. For example, key transcriptional regulators drive different stages in the human myeloid differentiation process [3]. Networks are often used in systems biology to help us understand the underlying biological mechanism in conditions.

In systems biology, we study a model organism as a whole. That is to investigate the components (e.g., transcriptome, proteome) and the interaction between these elements within a system. This methodology is based on the understanding that cellular components do not operate in isolation but as modules grouped by functionality. Systems biology often employs network analysis to facilitate the discovery of marker genes under a condition by incorporating established networks. These networks include known molecular interaction networks (e.g., protein-protein interaction (PPI) network, RNA-RNA interaction network). Despite pre-constructed networks, we can also use network construction techniques to create a condition-specific data-driven network (e.g., a co-expression network).

In Alzheimer's disease, many studies are investigating the involvement of miRNAs [4]. For example, downregulating miR-29 and miR-107 can increase the production of Amyloid-$\beta$ protein [5, 6]. As another type of small RNA that can regulate the mRNA abundance [7], little is known about the involvement of snoRNA in the progression of Alzheimer's disease. Moreover, the Tg4-42 mouse model is one of the few that develops neuron death in the hippocampus region. Hence, it can give us new insight into the mechanism of the etiology of Alzheimer's disease. Another problem is that we know less about snoRNA compared to miRNA and other coding genes. Prior knowledge is limited in terms of interaction databases and gene ontology annotation. This limitation makes it harder to study snoRNA. Here, I employed a data-driven approach to compensate for the lack of prior knowledge about small nucleolar RNA interaction and function in mice. In the first publication, I applied a differential co-expression network analysis to investigate small nucleolar RNA in the hippocampus of the Tg4-42 Alzheimer's mouse model. After building a differential co-expression network, nodes with high centrality measures indicate the potential functional role in the disease mechanism of the mouse model. Findings of the biomarkers indicate, indeed, some of them are associated with Alzheimer's disease mechanism. In addition, performing functional enrichment analysis of their interacting genes indicates the novel small nucleolar RNA biomarkers and the association of Alzheimer's disease.

Deep-sequenced RNA sequencing data allows a glance at alternative splicing that carries out pre-mRNA processing and produces different protein isoforms. There are a few limitations to studying alternative splicing with RNA sequencing. First, read depth is a critical factor in inferring splicing. Second, many splicing tools (more details in the Background section) are characterized into event-based and exon/isoform-based methods. In this dissertation, I focused on isoform-based methods using two types of data: static and dynamic data. Static data are the typical case-control experiments. Dynamic data is longitudinal data with multiple time points. However, most of the isoform-based tools are developed for static data. Only a few tools are designed for dynamic data. For example, TSIS was created to detect isoform switch (IS) events within a time series data [8]. However, it does not provide further analysis to characterize the detected IS events. More work is needed to improve IS detection in time series data and provide a tool for systematic analysis.

Understanding the function of alternative splicing and its upstream and downstream regulation is important, as dysregulation of splicing could be a primary cause of disease. For example, Spinal muscular atrophy is caused by the loss of SMN1 gene function due to alternative splicing [9]. Splicing regulation involves splicing factors, splicing enhancers, or splicing inhibitors (see background section). The outcome of splicing alterations can potentially rewire the connection within a PPI network. Such changes can disrupt or enhance a series of interactions within the PPI network, indicating a profound impact on gene function alternation due to splicing on molecular interactions. Using systems biology to study the effect of a splicing event on molecular networks will help researchers decipher the underlying molecular mechanism causing a condition.

The second publication introduces the Spycone framework, a splicing-aware time course network enricher that focuses on transcript-level data. This work contains two main parts: 1) develop a novel isoform switch detection algorithm. 2) incorporate clustering, network enrichment, and functional enrichment to analyze the isoform-switched genes. The novel isoform switch detection algorithm aimed to overcome the limitation of the other tool, the isoform switch detection tool, for time course data TSIS. Our analysis showed that TSIS detects switched isoforms with low-level expression. Lowly expressed isoforms could have less impact on the functional changes during isoform switch events. In addition, Spycone offers the detection of differential domain inclusion and exclusion. However, to use Spycone, researchers must do deep RNA sequencing from time series data. The cost of generating this type of data will be high. It is, therefore, essential to be able to perform this type of analysis in pairwise settings.

Numerous tools exist to perform differential transcript usage (DTU) analysis in pairwise data. Several benchmarking analyses have been performed to compare these tools. However, they are either applied in plant systems [10] or using an outdated aligner for the analysis [11]. In addition, more differential transcript tools are not benchmarked. Therefore, I performed a comprehensive benchmarking analysis for twelve tools (six of which were not benchmarked prior to this study) with simulated and real human datasets. I covered different experimental settings, including time series data and single-cell data. The third publication (see Unpublished results) provides a guideline and recommendation for performing DTU analysis based on different experimental setups.

## 1.2 Aim of dissertation

In the introduction section, I have discussed the current state of transcriptomics data analysis, specifically in the context of alternative splicing. It is true that while differential transcript usage and isoform switch analysis are powerful techniques for identifying differentially spliced genes, they are primarily designed for static comparisons, where the dynamic or time component is ignored. This is a significant limitation because alternative splicing events can vary dynamically depending on cellular contexts, developmental stages, or environmental cues. Furthermore, while systems biology approaches have been widely applied to gene-level studies, comparatively fewer studies focus on alternative splicing. Given that alternative splicing can significantly impact the functional diversity of gene products, it is critical to expand our understanding of this process at the transcript level. As discussed in section 2.6, alternative splicing can rewire the protein-protein interaction network. These limitations highlight the need for further methodological developments to fill the gap.

In this dissertation, the objectives are as follows: 1) investigate the involvement of snoRNA in the etiology of Alzheimer's disease mouse model Tg4-42; 2) develop a systematic analysis framework for alternative splicing analysis in time series data; 3) perform a comprehensive benchmark analysis for DTU analysis and provide an updated view of the current state of DTU analysis and guideline. Specifically, I will discuss two papers and one ongoing work. The first paper investigates potential snoRNA and miRNA candidates involved in the etiology of Alzheimer's disease using differential co-expression analysis. In the second paper, I aimed to study the impact of alternative splicing in time series transcriptomics data. However, I found no suitable tools available for this purpose. To address this gap, I developed Spycone, a novel tool that enables splicing-aware time series network analysis at the transcript level. Finally, to compensate for the shortcomings of the previous benchmark analysis, I performed a comprehensive benchmark with the existing DTU detection tools for two-condition experiments, time series DTU analysis, and single-cell experiments. I provided guidelines for DTU analysis of bulk RNA-seq data. Overall, this dissertation aims to overcome the limitations of existing techniques for transcriptomics data analysis and provide new insights into transcript-level analysis. By accomplishing these objectives, I hope to contribute to advancing network analysis in the field of transcriptomics, shedding light on crucial aspects of gene regulation and its implications in diseases like Alzheimer's.

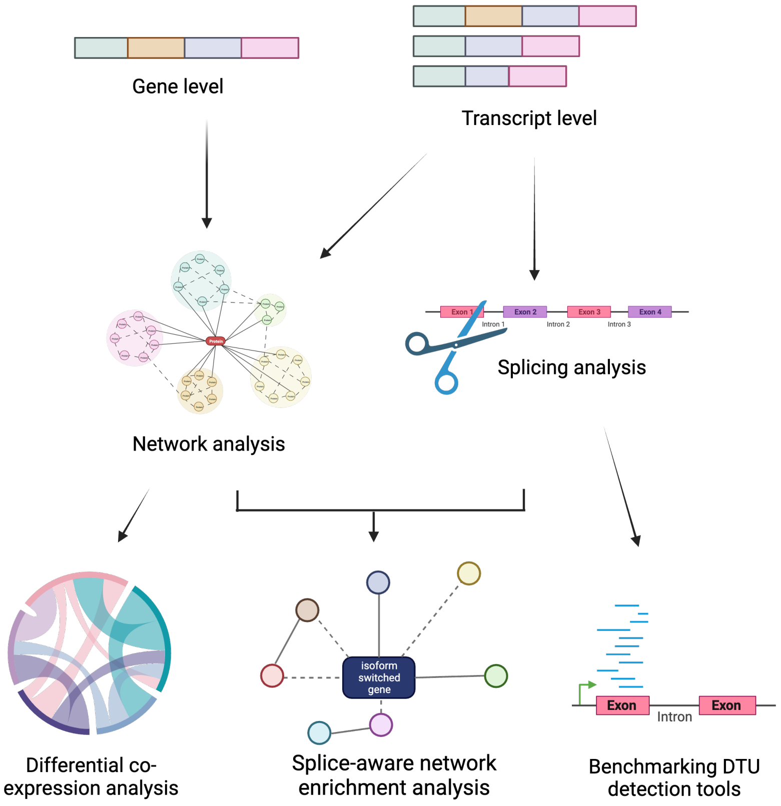


**Figure 1.1** Focus of this dissertation.
In this work, I first focused on studying snoRNA involvement in Alzheimer's mouse model Tg4-42 using a systems biology approach. Then, I investigate the possibility of applying network analysis in transcript-level resolution. In the second focus, I aimed to fill the gap of applying systems biology approaches in transcript-level resolution, where I developed Spycone. I performed a comprehensive benchmark analysis in existing DTU tools to further understand the current state-of-the-art differential transcript usage analysis.

# 2 Background

## 2.1 Ribonucleic acid (RNA) molecules: types and roles

Nucleic acid is an abundant macromolecule that is found in the nucleus of the cells, as well as in the mitochondria. In 1869, Friedrich Miescher first isolated DNA molecules from white blood cells [12]. Later, in 1955, James Watson and Francis Crick uncovered the structure of deoxyribonucleic acid (DNA) based on the X-ray image produced by Rosalind Franklin [13, 14, 15]. One of the interesting questions after the discovery of DNA was, how are proteins synthesized? Francis Crick first enunciated the Central Dogma theorem in 1958 [14]. The theorem consists of three main units: DNA, RNA, and protein, and it describes the information flow from one unit to another, including DNA to DNA, DNA to RNA, RNA to protein, and RNA to RNA (Figure. 2.1).

From the genetic information stored in DNA molecules, RNA molecules carry partial information, which is either translated into cellular units (i.e., proteins) or functions as enzymes (known as ribozymes). In 1968, Robert Holley was awarded the Nobel Prize in Medicine for discovering the structure of transfer RNA (tRNA), which links protein synthesis and messenger RNA (mRNA) [16]. The major structural difference between RNA and DNA is that RNA consists of sugar ribose rather than deoxyribose (which is ribose lacking one oxygen atom). RNA molecules consist of two purine-derived nucleobases (adenine and guanine) and two pyrimidine-derived nucleobases (uracil (instead of thymine in DNA) and cytosine). The presence of uracil contributes to the non-Watson-Crick base pairing structure of RNA molecules [17]. Nucleic acids have been long suspected to contribute to protein synthesis and cell growth [18]. While RNA is thought to be the origin of life, it is not the primary genetic material for most of the multicellular organisms [19]. A major transition in evolution is believed to be when DNA replaced RNA as the primary genetic information storage [20]. RNA molecules act as intermediate information carriers in heredity.

Heredity involves transferring genotype and phenotype to the offspring. As a unit of heredity, a gene is a DNA region containing promoter regions, untranslated regions, exonic and intronic regions in eukaryotes. A gene with introns was previously thought to be disrupted since the "mature gene" contains only exons and produces proteins. However, this definition no longer applies to our understanding of genes today. Many genes do not result in a functional protein; these are non-coding genes. One gene region can produce one or more functional products, such as proteins and non-coding RNAs. The production of these functional products is referred to as gene expression. Gene expression is a highly regulated process involving complex regulatory mechanisms, including chromosome architecture, chromatin modification, transcription, mRNA processing, degradation, translation, protein folding and modification.

The process of converting genetic information in RNA to protein is called translation. Various types of RNA are involved in translation, namely mRNA, tRNA and ribosomal RNA (rRNA) [21]. mRNA carries genetic information, while tRNA and rRNA act as ribozymes [22]. The loop region of the hairpin structures of

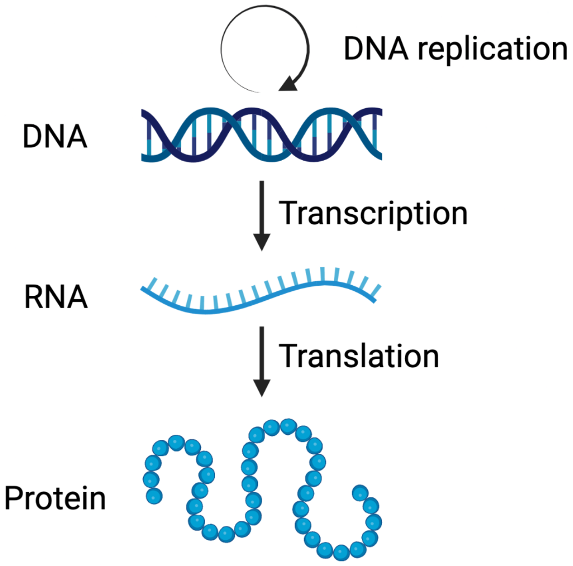


**Figure 2.1** Central dogma of molecular biology

The flow of genetic information from DNA, RNA to protein. Source: created with BioRender.com.

tRNA contains anticodons that recognize the codons on the mRNA (Figure 2.2). The corresponding tRNA loads desired amino acids to form peptides. Both tRNA and the generated peptide are held in place by the ribosome, which consists of proteins and rRNA. The resulting peptide undergoes folding to form a mature protein. Therefore, the genetic information carried by mRNA determines the protein product and function. The expression profile of mRNA plays a role in cellular protein diversity [23].

mRNA plays a role in cellular protein diversity via post-transcriptional modification. Post-transcriptional modification is a process where primary mRNA is modified by alternative splicing, degradation, or chemical modification. For instance, microRNAs (miRNAs) regulate mRNA expression through RNA silencing. miRNAs are predicted to be encoded in 2% of human genes, but they regulate up to 80% of human genes [24]. Most miRNAs are found in the intergenic regions or the antisense strand to functional genes [25, 26]. Other miRNAs are located in the intronic regions of a gene. They are a class of small non-coding RNAs with 19-25 nucleotides. MiRNA biogenesis involves the transcription of primary miRNA, followed by a canonical pathway to generate mature miRNA. First, miRNA is transcribed by RNA polymerase II into a primary miRNA and cleaved by Drosha at the hairpin structure of the miRNA. After exporting to the cytoplasm, Dicer, an RNase III endonuclease, removed the terminal loop and produced mature miRNA [27].

Another small non-coding RNA species primarily involved in post-transcriptional modification is small nucleolar RNA (snoRNA), a non-coding RNA that is 60-300 nucleotides long. Depending on the type of snoRNAs, they are essential for rRNA modifications. C/D box snoRNAs are highly conserved and are responsible for 2'-O-methylation of rRNA, while H/ACA box snoRNAs are responsible for pseudouridine modifications. These modifications are essential for the stability of RNAs. Researchers have been studying these roles in snoRNA for a long time. Recently, more studies have described the potential roles of

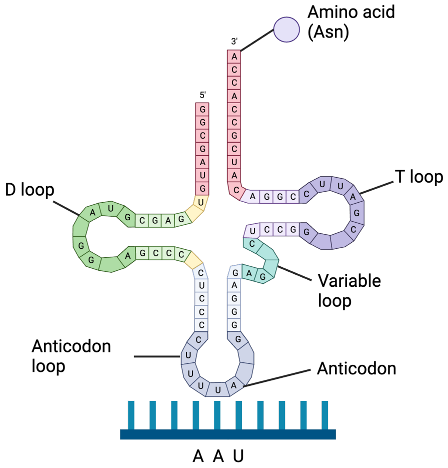


**Figure 2.2** Structure of a tRNA
The amino acid Asparagine (Asn) is used as an exmample here. The amino acid is binded to the 3'-end of the tRNA. Anticodon is located at the anticodon loop where the nucleic acid complementary to the codon is found. Source: created with BioRender.com

snoRNAs in rRNA acetylation. For instance, snoRNA snR4 and snR45 can interact with an RNA cytidine acetyltransferase and complementary to the acetylated site on 18s rRNA in yeast [28]. Knock-out and mutated models suggested the validity of the finding on snoRNA's rRNA acetylation role. Another recent study discovered a set of snoRNAs co-purified with a 3'-mRNA processing complex. These snoRNAs interact with a component of cleavage and polyadenylation specificity factor, FIP1. One of the snoRNA SNORD50 could block the polyadenylation of 3'-mRNA, affecting the mRNA turnover, intracellular localization, and translation efficiency [7].

## 2.2 Introduction of Alternative splicing

After a eukaryotic gene is transcribed into a pre-mRNA, it undergoes a series of processing steps, including 5' capping, 3' polyadenylation (polyA tail), and splicing - the removal of introns (non-coding regions) and merging of exons (coding regions). The discovery of alternative splicing revolutionized our understanding of gene expression and provided insights into the molecular basis of many biological processes. The importance of alternative splicing was recognized by awarding the Nobel Prize in Medicine to Philip Sharp and Richard Roberts in 1993 [29]. They described how the process of splicing allows for the removal of non-coding intronic regions from pre-mRNA transcripts, producing mature mRNA that can be translated into protein. When a gene has multiple exons, splicing might merge different combinations of exons - this process is called alternative splicing. Alternative splicing enables the production of multiple protein isoforms from a single gene. This mechanism increases the functional diversity of the proteome and contributes to the complexity of biological processes. For example, Tcf/Lef family isoforms are expressed in a cell-type-specific manner. The isoforms contribute to the expression of Wnt/b-catenin target genes [30]. AS events of FOXP1 regulate the transcription of target genes responsible for stem cell differentiation [31]. The regulation of alternative splicing is a highly coordinated process involving many factors, including splice site selection, splicing factors, splicing enhancers and silencers, and other RNA binding

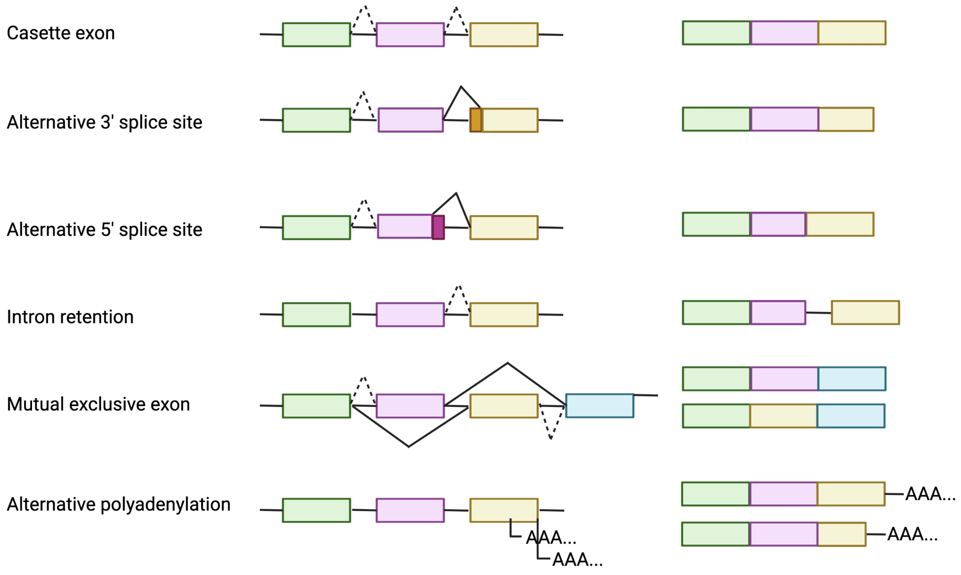

Casette exon
Alternative 3' splice site
Alternative 5' splice site
Intron retention
Mutual exclusive exon
Alternative polyadenylation
AAA...
AAA...
AAA...
AAA...

**Figure 2.3** The types of alternative splicing.
Cassette exon (also known as exon skipping, describe events where an exon is spliced out or retained in a transcript. Alternative 3'/5' splice site has a different 3'/5' splice site than the canonical transcript. Intron retention refers to transcripts with intron being retained. Source: created with BioRender.com

proteins (more details in the following section). Dysregulation of alternative splicing has been associated with various human diseases, including cancer [32], neurodegenerative disorders [33], and developmental abnormalities (e.g., myotonic dystrophy [34]).

### Mechanisms of alternative splicing

The splicing machinery, known as the spliceosome, catalyzes the splicing process and consists of five small nuclear ribonucleoproteins (snRNPs) - U1, U2, U4, U5, and U6 [35]. The main components of snRNPs are small nuclear RNA (snRNAs). Once transcribed, snRNAs are 5' capped and exported to the cytoplasm with the help of a pre-export protein complex. During export, snRNAs are directed to snRNA-rich Cajal bodies in the nucleus before being exported through the nuclear pore [36]. Upon entering the cytoplasm, the pre-export complex disassociates from the snRNAs [37]. The survival motor neuron (SMN) protein complex then recruits snRNAs and Sm proteins to form a ring structure around the snRNAs, which is thought to stabilize them and initiate RNA processing and trimming of the 3' end before they are re-imported to the nucleus [38]. The reason for this export and re-import process has yet to be fully understood. However, it is thought to be a quality assurance mechanism to ensure proper maturation of snRNPs before they are assembled into the spliceosome. Upon re-entering the nucleus, snRNPs are processed and remodeled in Cajal bodies before being transported to the nucleoplasmic subcompartment known as nuclear speckles, which are believed to be involved in pre-mRNA splicing [39].

The regions around the exon-intron boundary contain specific sequences: the 5' end of the intron contains a 5' splice site (donor site) with two nucleotides, GU. In comparison, the 3' end contains a 3' splice site (acceptor site) that is an AG terminal site and a branchpoint adenosine located upstream of the 3' splice site by 15-50 base-pairs [40]. The first step of splicing involves U1 snRNP recognizing and binding to the 5’ donor site, which is stabilized by SR protein and the cap-binding complex [41]. Simultaneously, U2 snRNP recognizes the base-pairing near the branchpoint adenosine. U1 and U2 snRNPs interact and bring the two exons close to each other, forming a pre-spliceosome complex [42]. After forming the pre-spliceosome complex, U4, U5, and U6 snRNPs assemble into a pre-catalytic complex [43].

Next, the pre-catalytic complex is activated, triggering the release of U1 and U4 snRNPs, leading to the formation of the catalytic complex [43]. The catalytic complex consists of a U2-U6 snRNA dimer that holds the 5’ donor site and the 3’ acceptor site proximal to each other [40]. It carries out the first transesterification step, where the phosphate at the donor site is cleaved by the 2’ hydroxyl group of the branchpoint adenosine, causing the detachment of the 5’ exon. At this point, the 5’ end of the intron ligates to the 2’ hydroxyl group of the branchpoint adenosine, forming an intermediate structure called a lariat. In the second step, the phosphate of the 3’ splice site is cleaved by the 3’ end of the detached exon, bringing the two proximal exons together. This step is completed by ligation, resulting in the release of the intron lariat structure and the formation of the post-spliceosome complex. Finally, U2, U5, and U6 snRNPs are released, and the process is ready to repeat for the following intron [40].

The different types of alternative splicing depend on how the mRNA is spliced (Figure. 2.3). Cassette exons are a set of exons that can be included or excluded during splicing. Exon skipping is when a cassette exon is excluded. An alternative 5’ or 3’ splice site is when the exons are included with a modified splice site. The mutually exclusive event happens when two exons are never spliced or skipped in the same mature mRNA. Intron retention is when an intron is kept in the mature mRNA.

### Regulation of alternative splicing

Alternative splicing, like transcription, is highly regulated. This regulation includes the selection of exons (or splice sites), which determine the final composition of the mature mRNA and the resulting protein product. Similar to transcription, splicing is also regulated by both cis- and trans-acting elements. Exon (or intron) splicing enhancers (ESE/ISE) and exon (or intron) splicing silencers (ESS/ISS) are examples of cis-elements. In contrast, proteins interacting with cis-elements are trans-acting elements, such as SR proteins. In addition, transcription and splicing are tightly linked together. For example, transcriptional elongation pauses or lower rates favor the exon skipping of alternative exons [44].

ESEs, typically found in constitutive exons, are associated with regular splicing, leading to exon inclusion. Conversely, the absence of ESE causes exon skipping. ESEs help amplify the splice site’s splicing signal, aiding the splice site’s recognition by splicing factors. Most ESEs contain binding motifs for SR family proteins, splicing factor proteins with an RNA-recognition motif (RRM) domain and an Arg-Ser (RS) rich domain at the C-terminal. The RS domain is associated with constitutive splicing by interacting with snRNPs of the spliceosome. ISEs are intron sequences that drive the usage of neighboring or distal splice sites. In the example of hnRNP A1 pre-mRNA, there is a highly conserved region between exon 7 and alternative exon 7B. The interaction of hnRNP A1 with the ISE region promotes the usage of the distal 5’ splice site, hence promoting the splicing of exon 7 [45].

ESSs function mechanistically similar to ESEs, containing binding motifs for heterogeneous nuclear ribonucleoproteins (hnRNPs). hnRNPs are splicing repressors that contain an RRM domain. Silencer-bound proteins often bind to ESS to inhibit the splicing of an exon, acting as an antagonist and blocking the binding

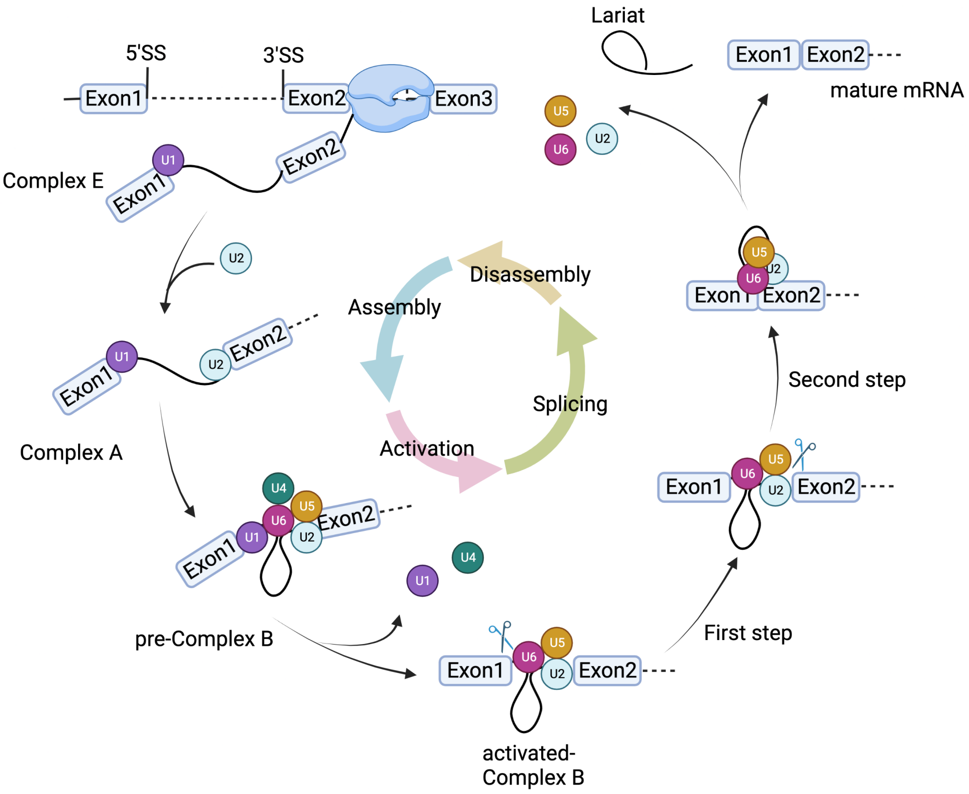


**Figure 2.4** The molecular model of assembly of spliceosome.
The four stages of splicing begin with the assembly and recruitment of U1 and U2 SNPs. The activation stage forms and activates complex B, which brings the two exons together. Splicing occurs by first cleaving the 5’-end of exon two and then the 3’-end of exon 1. This stage ends with the ligation of the two exons. Finally, the splicesome disassembles and releases the intron lariat and mature mRNA. Source: created with BioRender.com

of SR proteins to ESEs, thereby preventing splicing. hnRNPI (polypyrimidine-tract-binding protein (PTB)) is often bound to the 3' splice site, acting as a blockage for U2 snRNPs. ISSs are the intronic element that blocks splicing. In the example of FGF-R2, ISS, located upstream of exon 3B of the FGF-R2 gene, contains the binding site of PTB. The binding of PTB with this region represses the splicing of exon 3B [46].

The spliceosome and many regulatory proteins are expressed universally and at high concentrations (Figure. 2.4). However, splicing, like transcription, is highly specific. So, how is splicing a specific transcript in a gene regulated when the regulatory machinery is universal? It turns out that there are multiple ways to achieve this. In addition to selectively targeting the regulatory proteins to the designated transcript, the combinatorial effect of SR and hnRNPs elements also plays a role [47].

For instance, in regulating exon-3 of the tat gene in HIV, hnRNPA1 binds to the ESS of the exon [48]. It only inhibits splicing when multiple hnRNPA1 molecules bind and propagate to the 3' splice site in the upstream intron. This propagation ultimately blocks the interaction of the 3' splice site with U2 snRNPs of the spliceosome, thereby preventing splicing. However, if SF2/ASF binds to the ESE located between the 3' splice site and the hnRNPA1-bound ESS, this action will stop the propagation and allow the splicing of exon-3. Thus, not only does the selective effect of the regulatory protein play a role, but the concentration of the proteins also affects the outcome of splicing [35].

## 2.3 Functional impact of alternative splicing

Alternative splicing affects both the abundance and diversity of transcriptome and proteome [23]. The effect includes various aspects: protein function, transcriptome diversity, protein localization, and 5'- and 3' untranslated region (UTR) processing.

Splicing can alter protein function by producing different protein isoforms with varying domains. These domains are functional regions that typically define the protein's function, making them essential for protein diversity. In humans, 81,837 proteins are listed in the latest Uniprot database, whereas there are only around 20,000 protein-coding genes [49]. Splicing is crucial for producing certain small RNA species, such as snoRNA and circRNA. These small RNAs are often found in intronic or intragenic regions, while others are intergenic and independent of host genes. During transcription of host genes, splicing of the intron where the small RNA is located occurs. Host transcripts are susceptible to Nonsense-Mediated Decay (NMD) [50] if a premature translation stop codon is detected in the typical open-reading frame. Many discoveries have also shown that other types of RNAs, such as tRNA and rRNA, also undergo splicing [51, 52]. Splicing can significantly impact protein localization. For instance, NMDA receptors, glutamate-gated ion channels expressed on the postsynaptic membrane, play a crucial role in synaptic plasticity and neurological diseases. Before being transported to the membrane, NMDA receptor subunits are co-translationally assembled in the endoplasmic reticulum (ER). Alternative splicing of the C terminal of NMDA generates various transcripts, including one spliced with a C1 domain that contains an ER retention/retrieval motif, leading to receptor suppression and retention in the ER [53]. Splicing can also result in alternative 5' or 3' UTRs, upstream regions of a gene's start codon (AUG). These regions provide binding sites for translation-regulating factors. For example, $\beta$-catenin, an oncogene, has multiple splice variants containing alternative 3'UTRs. One variant's 3'UTR contains an AU-rich element that maintains mRNA stability and translation efficiency. [54]. The length of 3'UTR is also associated with immune cell differentiation, in which shortening of 3' UTR is more prominent in activated lymphocytes [55]. 3'UTR contains cellular signals that guide mRNA localization, mRNA stability, and translation efficiency.

## 2.4 Sequencing technologies

In the 1980s, splicing was investigated using Expressed Sequence Tag (EST) [56]. ESTs are short sequences that consist of a few hundred base pairs. They are derived from a cDNA library's 5' or 3' end. While ESTs were useful in the early stages of splicing research, they have limitations, including incomplete coverage of the gene region, limited transcriptome coverage and inability to distinguish among different transcripts. The development of high-throughput sequencing of DNA and RNA has revolutionized biomedical research. High-throughput sequencing technologies include next-generation sequencing (NGS), which can generate millions of short reads in parallel, and third-generation sequencing (TGS), which can produce longer reads. With these new technologies, we could identify transcripts and genomic variants associated with diseases using the latest sequencing technology.

Before the emergence of high-throughput sequencing, Sanger sequencing was widely used to identify DNA fragments [57]. It is beneficial for the validation of plasmid constructs or PCR products. Sanger sequencing is based on the dideoxy method to sequence a single-stranded DNA. The reagent mixture contains deoxyribonucleoside triphosphates (dNTPs) and fluorescence-labeled dideoxyribonucleoside triphosphates (ddNTPs) of all four types. DNA polymerases are added to synthesize the DNA template. The elongation will be terminated when a ddNTP is added to the strand. At this point, the mixture contains DNA templates of different lengths. With electrophoresis and fluorescence to visualize the DNA sequence, this method can only sequence short DNA strands (100-1000 base pairs). Shortly before the Human Genome Project started, shotgun sequencing was first introduced in 1988 [58]. Shotgun sequencing can sequence longer nucleotide fragments than Sanger sequencing and was also used for sequencing in the Human Genome Project. This method breaks up long DNA strands into random fragments, followed by chain termination. We obtain a library of overlapping reads by running several rounds of shotgun sequencing. These reads can be assembled by running computer software based on finding the overlapping ends of reads.

### High-throughput sequencing

Illumina, one of the key players in the current sequencing industry, uses sequencing by synthesis. In this method, library preparation requires tagmentation and indexing steps (Figure. 2.5). The DNA molecules are fragmented into reads of lengths ranging from 150-500 bp, and the tagmentation step tags the DNA fragments with adaptors. After ligation of the adaptors to the template, indexes and oligonucleotides required for the sequencing step are added to both ends of the fragment, following the adaptors. Before sequencing, the fragments are amplified using bridge amplification, which generates clones of fragments on glass flow cells. The flow cells are coated with oligonucleotides that are complementary to the oligonucleotides on the fragments, and the fragments then bind to the oligonucleotides. In the first round of synthesis by polymerase, the fluorescence signal emitted after annealing is recorded by the sequencing machine, and a base-calling algorithm identifies the correct base. This parallel sequencing can sequence millions of fragments simultaneously, producing an extensive database. Different generations of Illumina

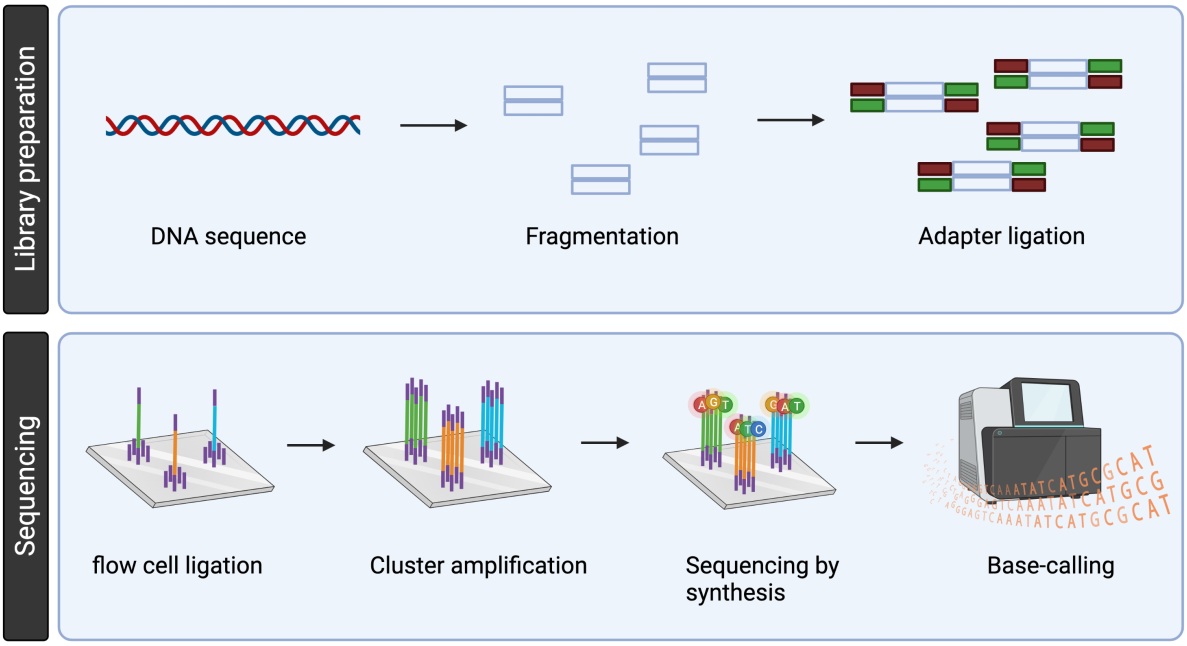


**Figure 2.5** Schematic overview of illumina sequencing.
Library preparation step includes fragmentation of the targeted DNA sequence and ligation of adapters to each of the fragments. Illumina carries out sequencing by synthesis (SBS), the fragments are ligated to the flow cell through the adapters, cluster amplification creates one million copies of the fragment through PCR. Sequencing by synthesis performs by DNA polymerases which glows every time a base is added to the strand. Source: created with BioRender.com

sequencing machines exist, such as MiniSeq, MiSeq, NextSeq, and NovaSeq. For example, the latest NovaSeq system can sequence up to 1.2 billion reads with a 300 bp read length and produce 360 GB of data. The limitation of sequencing technologies is that they are highly dependent on hardware. We can only generate as much data as we can afford for storage and analysis. Fortunately, computer storage and speed have drastically improved since the early 2000s. We can now generate terabytes of data and obtain more and longer reads, enabling us to study splicing and other biological processes in unprecedented detail.

However, short reads pose challenges for analyzing repetitive regions and structural variations due to their limited read lengths (<300 bp). Short read lengths cannot detect variants in repetitive regions, especially introns. This leads to inaccurate quantification of splicing events. Third-generation sequencing is focused on obtaining long reads of DNA molecules, even capable of producing a whole transcript with a single read [59]. One of the key players in this field is Pacific Biosciences (PacBio). Two of their long-read sequencing systems are the Revio and Sequel systems, which produce HiFi reads (Figure 2.6) (Revio system generates a higher number of reads). HiFi reads are produced based on a sequencing technology called Single Molecule Real-Time (SMRT) sequencing, which can sequence up to 30-50 kb long DNA fragments. The main feature of this technology is the SMRT flow cell, which contains zero-mode waveguides (ZMWs) - nanophotonic chambers capable of holding only one cDNA fragment each. HiFi reads are produced using circular consensus sequencing (CCS), where double-stranded cDNA fragments are ligated at both ends, forming a circular bell shape, which, as a whole, is called the SMRTbell library. The

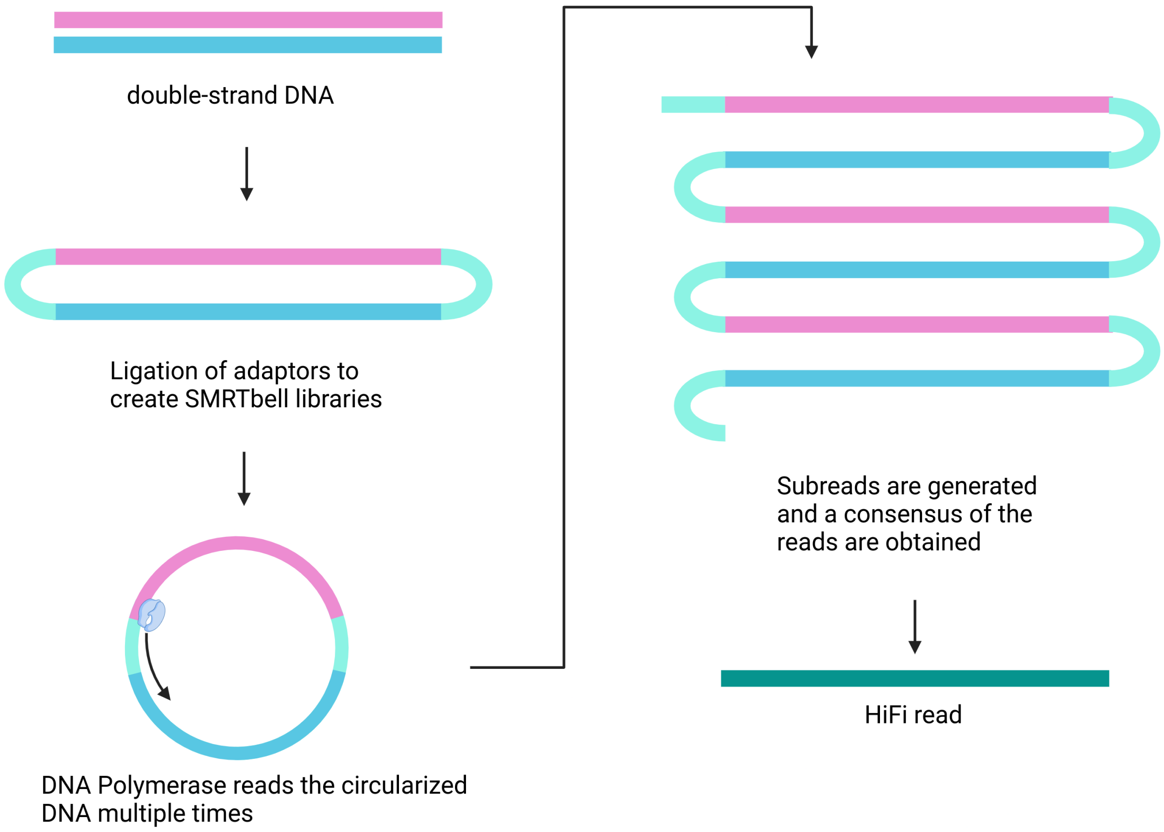

double-strand DNA
Ligation of adaptors to create SMRTbell libraries
DNA Polymerase reads the circularized DNA multiple times
Subreads are generated and a consensus of the reads are obtained
HiFi read

**Figure 2.6** Schematic overview of HiFi-reads technology from PacBio.
Circular consensus sequencing (CCS) creates a consensus HiFi read after sequencing is performed multiple passes on the circular DNA ligated with adaptors. Source: created with biorender.com

SMRTbell template is then sequenced repeatedly multiple times, producing up to 16x the length of the template. The read is then compiled, and sequencing errors are corrected based on the consensus of the read.

Nanopore-based sequencing developed by Oxford Nanopore Technologies (ONT) is another type of third-generation sequencing technology [60]. There are two main branches of nanopore sequencing: biological membrane systems and solid-state sensor systems. In the biological membrane system, transmembrane nanopore proteins inspired by natural systems are used, with alpha-hemolysin and *Mycobacterium smegmatis* porin A (MspA) being commonly used. DNA template strands are moved through the nanopore proteins with the help of motor proteins, DNA helicases, and DNA exonucleases. In contrast, the solid-state system uses metal and metal alloy substances for DNA passage. ONT offers systems such as MinION, GridION, and PromethION. The portable MinION, for example, has up to 512 nanopore channels, each equipped with a membrane and a nanopore and controlled and detected by an application-specific integrated circuit (ASIC). As a DNA strand passes through the nanopore, the ASIC circuit detects the disruption of the current applied to the nanopore, generating a signal (or squiggly line). A base-calling algorithm is used to identify the nucleotides of the sequence from the generated signal since nanopore se-

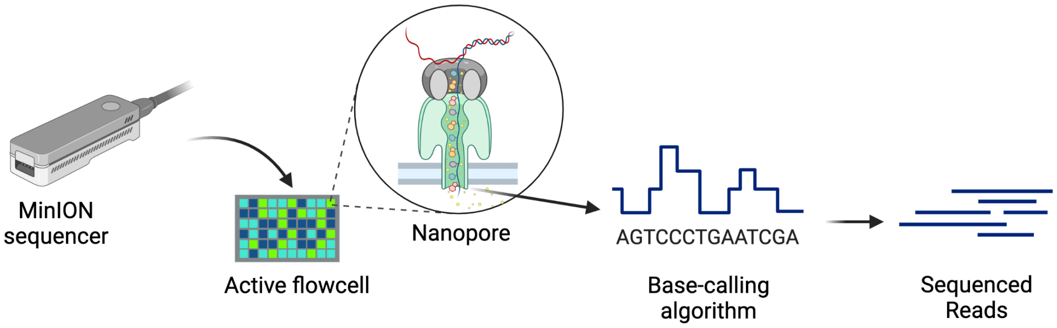


**Figure 2.7** Nanopore sequencing technology

The flowcell placed in the sequencer contains active nanopores can allows DNA fragments to pass through. While the DNA pass through the nanopore, the ASIC circuit applied to the nanopore is disrupted and hence producing an electrical signal that indicates the bases. Source: created with BioRender.com

quencing generates signals instead of actual nucleotides (Figure 2.7). Some algorithms can detect RNA sequences, while others are designed to identify RNA modifications [61].

### 2.4.1 RNA sequencing

#### mRNA sequencing

RNA sequencing has revolutionized the field of transcriptomics and has become the most widely used NGS method for gene expression analysis. RNA sequencing primarily aims to identify the functional genes involved in biological processes and diseases. Before NGS emerged, hybridization-based methods such as microarrays were commonly used for gene expression quantification. Microarrays could quantify thousands of genes using probes complementary to them. However, they can only detect annotated genes with known sequences and provide relative gene abundance. Alternative methods, serial analysis of gene expression (SAGE), and cap analysis of gene expression (CAGE) allow for the quantification of absolute gene expression without prior knowledge of gene sequences and the use of probes [62]. Both methods depend on capping of 5'-end and tagging of 3'-end, respectively, resulting in incomplete gene body coverage. This limitation can lead to inaccurate detection of RNA spliced variants. Advancements in NGS technologies include Illumina-based sequencing, allowing longer read length and deeper sequencing depth. Paired-end sequencing can obtain reads with information from both 5'- and 3'-end of the transcript, leading to more accurate splicing patterns detection.

Before sequencing, RNAs need to be reverse transcribed to cDNA using reverse transcriptase. rRNAs are depleted first in the sample preparation for cDNA library for sequencing. The depletion is usually achieved by using oligo-dT beads to enrich for polyadenylated mRNA or selectively degrade rRNA using exonuclease, which mRNA is protected by 5'-cap. Then, fragmentation is required to reduce the size of mRNA in order to obtain reads. In the Illumina protocol, adaptors specific for 5' and 3' are used to distinguish the direction of the strand. The ligation products are then reverse-transcribed and amplified. Significant bias can emerge during the PCR amplification step because not all fragments can be amplified with the same efficiency. For instance, GC-rich or AT-rich fragments have lower amplification efficiency than GC-neutral. As a result, GC-rich or AT-rich regions will be underestimated [63]. A PCR-free protocol was developed to mitigate this bias [64].

Bulk RNA-seq analysis ignores the heterogeneity derived from the functional difference between cell types, the stage of the cell cycle and cell age, etc. Single-cell approaches are used to compare transcriptome profiles across cell types in tissue and heterogeneity within a cell type. It can also be used to identify new cell types and observe stochastic gene expression within a cell population, though it is susceptible to the stage of cell cycle [65]. The rapid development of single-cell sequencing technologies allows us to analyze more cells with a higher sequencing depth. Each cell is represented by a barcode and sequenced in a pooled, multiplexed manner. Microfluidics technology has emerged to efficiently isolate single cells for further investigation. Popular techniques include droplet-based and plate-based methods. Droplet-based methods encapsulate cells in oil droplets, while plate-based methods use cell-sized microwells that trap cells for high capture efficiency. Chromium 10x, a commercialized single-cell protocol from 10x Genomics, enables sequencing up to 80,000 cells in a single run. Chromium 10x uses droplet-based cell sorting, leading to a higher dropout rate than the plate-based method. Moreover, Chromium 10x specializes in 3'-tagged transcripts sequencing, prone to 3' coverage bias [66]. On the other hand, smart-based protocols like Smart-seq2 [67], Smart-seq3 [68] and Smart-seq3xpress [69] are plate-based methods. They have a lower throughput (up to  6000 cells) but a lower dropout rate. In addition, smart-based protocols employ a template-switching step that captures the full-length transcript information, hence minimizing 3'-coverage bias [70]. In general, single-cell sequencing captures fewer genes due to each cell's limited abundance of transcripts. In some protocols (e.g., Chromium 10x, Smart-seq3), unique molecular identifiers (UMI) are added to improve the quality of gene identification and reduce amplification biases during library preparation [66].

Single-cell long reads protocols aim for better transcript identification, as well as novel transcript detection; examples are R2C2 in Oxford Nanopore sequencing [71] and Multiplexed Arrays Sequencing (MASseq) using the PacBio platform. cDNA libraries are prepared by concatenating multiple cDNAs (e.g., 15 cDNA molecules) to a single molecule, following barcode ligation. The ligated cDNA underwent CCS library construction and sequencing in the PacBio Sequel II platform (it is also possible with the PacBio Revio platform). [72]. Although these protocols are still limited due to lower sequencing depth and high sequencing error rate.

### small RNA-seq

Small RNA-seq allows for profiling small non-coding RNAs (ncRNAs) [73]. Small ncRNAs are less than 200 nucleotides and do not encode proteins, yet they form a regulatory network that interferes with many cellular functions. Some of the most well-known small ncRNAs are transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs), which play crucial roles in translation. Another important class of small ncRNAs are microRNAs (miRNAs), short, single-stranded RNAs usually around 22 nucleotides long. They are involved in the negative regulation of gene expression at the post-transcriptional level. Small nucleolar RNAs (snoRNAs) are less well-known but essential as they are involved in rRNA modification and splicing.

The crucial steps in obtaining small RNAs are total RNA isolation, RNA size selection, and cDNA library preparation and sequencing. One challenge of small RNA-seq is that small ncRNAs can have highly

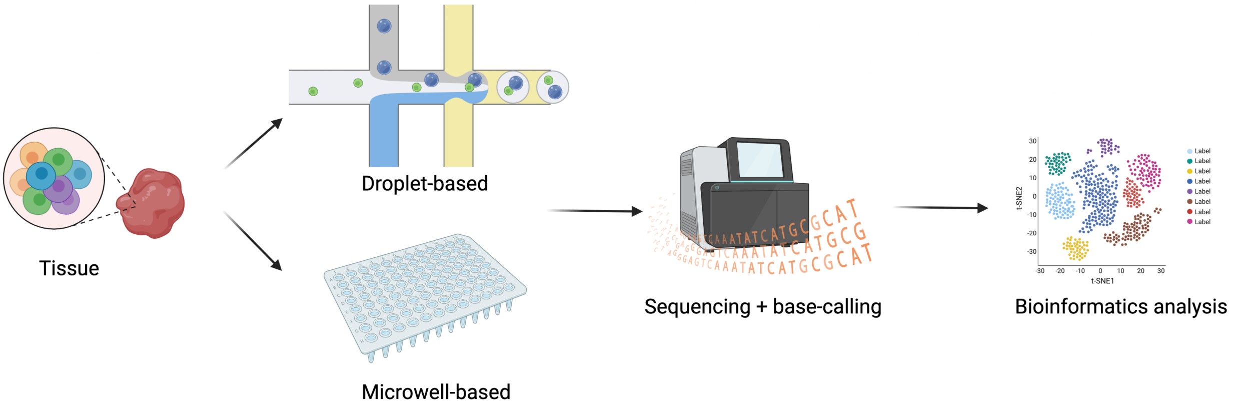


**Figure 2.8** Summary of single-cell sequencing workflow.
A tissue containing multiple cell types can be sampled with droplet-based or microwell-based protocols, followed by sequencing and bioinformatics analysis. Source: created with BioRender.com

similar sequences, making it difficult to accurately distinguish between small RNA species. Additionally, several biases can emerge from the sequencing process. For instance, ligation of adapters to the small RNA can lead to intramolecular ligation, forming circularized RNAs. Moreover, secondary structure affects the selection efficiency, meaning loss of highly structured small RNA during extraction [74]. As in mRNA sequencing, PCR in the amplification step will lead to PCR bias. This bias can be mitigated using unique molecular identifiers (UMIs). Despite these challenges, small RNA-seq is a powerful tool for studying the complex regulatory networks involving small ncRNAs and their roles in various biological processes.

## 2.5 Computational analysis of transcriptomic data

### 2.5.1 Experimental design

The experimental design is tailored to the research question and is constrained by available resources. Comparing different conditions (e.g., healthy and tumor) using a snapshot of the samples is commonly called static conditions. In contrast, time course conditions capture dynamic patterns by providing snapshots of multiple time points of the samples. During sample preparation, batch effects could arise from multiple sources, including experiment timing, handling person, sequencing machine and location, and chemical reagents. Multiplex sequencing can sequence multiple samples in a single run to minimize batch effects in sequencing experiments. However, this approach may only sometimes be successful due to errors in barcode recognition that can render some reads unusable. Therefore, it is essential to perform quality control checks to ensure the reliability of the sequencing data. Additionally, sample size is another critical consideration: insufficient sample size can result in high variability, which can mask the actual signal and produce false positive results. Thus, carefully considering experimental design and sample size is crucial for obtaining reliable and informative results.

### 2.5.2 Alignments and Quantification

RNA-seq experiments produce fastq files with read sequences. Raw fastq reads are typically mapped to a reference genome or transcriptome, and quality checks should be performed before and after mapping. FastQC [75] is a useful command line tool for short-read data, while longQC [76] is designed for long-read data. These bioinformatics tools provide basic statistics about the sample, such as read length distribution, total number of reads, and GC content distribution.

After mapping, several aspects should be evaluated. The percentage of mapped reads directly indicates the sample purity; high percentages of unmapped reads can indicate sample contamination. Additionally, the GC content and read length distribution are crucial for accuracy in the quantification step since they may reveal biases arising from PCR amplification. For transcript-level analysis, read coverage should be uniformly distributed along the transcript. There are multiple tools available for checking these qualities, such as RSeQC [77], Qualimap [78], and bamtools (stat module) [79].

Aligners can be categorized into two main groups: direct alignment and pseudo-alignment. One of the widely-used tools - STAR - is an example of a direct alignment tool [80]. STAR can map reads to splice junctions with high precision [81]. The main advantage of STAR is the ability to align reads with mismatches, insertions, deletions, and splice junction reads from non-contiguous genomic regions.

Salmon and Kallisto are popular pseudo-alignment tools. They allow for fast quantification of transcript abundance without actually producing read alignments. Both tools use de Bruijn graph-based algorithms to search seed loci for each read using k-mers [82]. Salmon first performs quasi-mapping to obtain a binary matrix indicating which fragments are mapped to transcripts [83]. It then estimates transcript abundance in two steps: (1) using variational Bayesian inference of read counts per transcript to estimate the approximate posterior probability, and (2) optimizing the estimates with an expectation-maximization (EM) algorithm [84]. In contrast, Kallisto uses a maximum-likelihood function followed by an EM algorithm to estimate the probability of which fragments are selected from transcripts after quasi-mapping [85]. Both methods offer bootstrapping (and Gibbs sampling for Salmon) to further confirm the confidence level of the estimation after convergence of the EM algorithm. For single-cell RNA libraries, alignment can be done similarly to bulk RNA-seq but with additional steps for barcode detection and demultiplexing. A UMI correction step is also necessary for protocols with UMIs.

### 2.5.3 Data normalization and batch effects

After the sequencing reads are aligned to the reference genome, the quantification step in the analysis pipeline is followed. The final output of gene or transcript quantification is a count matrix containing expression values for every gene or estimated count values for every transcript in each sample. Multiple factors can affect these counts. Firstly, the intrinsic amount of starting material, including cell number, cell volume, and intracellular localization of transcripts, cannot be fully represented by the reads [86]. Second, technical variations in the library preparation steps can introduce biases into the sample. Different batches of cell cultivation, reagents, sequencing batches, and even handling personnel can influence the number of transcripts. Thus, it is crucial to control the experimental protocol. For this reason, not all variations among samples are biological, and it is essential to apply the appropriate normalization method and batch effect correction when one dataset is susceptible. After obtaining the raw count matrix, we can observe the difference in library sizes by summing up the columns. Library size refers to the number of mapped reads in a sample. The comparison might only be meaningful if we compare sample counts by rescaling the library sizes. For example, 10 out of 100 equals 20 out of 200. Rescaling the counts by getting the proportion to the library size is the easiest normalization method. However, in RNA-seq data, we also need to consider the length of the transcript. The longer the transcript, the more reads can cover it. RPKM (reads per kilobase per million reads) is a measure that considers both the number of reads mapped to the gene and the length of the gene [87]:

$$\text{RPKM} = \frac{r_g \times 10^9}{fl_g \times R} \tag{2.1}$$

where $r_g$ is the number of reads mapped to gene $g$, $fl_g$ is the effective length of gene $g$ (usually calculated from the exonic region in the gene) and $R$ is the library size. Another commonly used normalization measure is TPM (transcripts per million). TPM is a slightly modified version of RPKM:

$$\text{TPM} = \frac{r_g \times rl \times 10^6}{fl_g \times T} \tag{2.2}$$

$$\text{T} = \sum_{g \in G} \frac{r_g \times rl}{fl_g} \tag{2.3}$$

where $rl$ is the average read length and T is the total number of reads transcripts.

TPM and RPKM are within-sample normalization methods, while between-sample normalization is necessary for differential expression analysis. One commonly used method is the trimmed mean of M-values (TMM), implemented in edgeR [88]. TPM and RPKM can perform poorly when the transcript distribution is skewed, meaning they are suitable for within-sample comparison [89]. When a gene is highly expressed in a biological sample, the sequencing depth is limited for the rest of the genes. Even if the counts are normalized by library size, there is still an intrinsic bias toward highly expressed genes. The TMM method corrects this bias by trimming the M-values to minimize the log fold change differences of the sample while preserving the ranking of genes. The TMM method calculates a normalization factor that takes the trimmed average of M-values of genes in sample k using reference sample l:

$$M_g = \text{log}_2 \frac{\frac{r_{gk}}{R_k}}{\frac{r_{gl}}{R_l}} \tag{2.4}$$

M-value is the gene-wise log2 fold change. Trimmed mean of M-values is the average of M-values after trimming 30% of the upper bound and the lower bound by default. The resulting normalization factor is the weighted, log2 transformed trimmed-mean of M-values:

$$\text{log}_2(\text{TMM}^l_k) = \frac{\sum_{g \in G} W^l_{gk} M^l_{gk}}{W^l g_k} \tag{2.5}$$

where weight is an inverse variance that due with the fact that highly expressed genes have lower variance than lowly expressed ones.

The normalized counts from TMM method gives counts-per-million (CPM):

$$\text{Normalized count} = \frac{r_{gk}}{\text{TMM}^l_k} * 10^6 \tag{2.6}$$

Another normalization method that corrects for sequencing depth and RNA composition is the relative log-expression (RLE) method, implemented in DESeq2 [90]. This method first calculates a sample-wise geometric mean for each gene in every sample K with R replicates:

$$Y_g = \sqrt{\prod_{k=1}^{K} \prod_{r=1}^{R} X_g kr} \tag{2.7}$$

Next, it calculates size factors for each sample: the ratio of counts and the geometric mean as the reference. The size factor for each sample is obtained as the median of all ratios.

$$M_{kr} = \text{Median}_g \left( \frac{X_{gkr}}{Y_g} \right) \tag{2.8}$$

Finally, the raw counts are normalized by dividing each column by the corresponding size factor.

$$\text{Normalized count} = \frac{X_{gkr}}{M_{kr}} \tag{2.9}$$

### 2.5.4 Differential gene expression analysis

Differential gene expression is performed on top of normalized counts. At the transcript level, this analysis can be performed on top of transcript counts, and we can gain insights into differential transcript expression. However, a major challenge of this analysis is the uncertainty of variance within samples. To tackle this, edgeR and DESeq2 employ a negative binomial (NB) distribution to model the count matrix [90]. The NB distribution generalizes Poisson distribution by introducing over-dispersion parameters. The over-dispersion parameter is estimated using an empirical Bayes method before fitting. A gene-wise dispersion is first estimated and then fitted with a smooth curve to generate an expected dispersion for each gene. The estimates are then shrunk toward the expected values, and the extent of shrinkage is automatically adjusted based on the data and sample size. They also shrink the log2 fold change towards zero to account for the high variance of lowly expressed genes. The resulting shrunken log2 fold change and standard errors are tested for differential expression using a Wald test.

### 2.5.5 Differential transcript usage analysis

Alternative splicing analysis can be categorized into two main approaches: the first approach is an exon-/isoform-based approach, and the second approach is event-based. The exon-/isoform-based approach quantifies reads based on the mapping to the annotated set of exons and transcripts. Event-based methods map sequencing reads into the annotated genome. By comparing the genomic features such as exonic and intronic regions, alternative splice sites, and alternative 5'/3' terminal, we can characterize the variations underlying sequence changes derived from alternative splicing. The identified events are further quantified using PSI (percentage spliced-in) values. Both approaches rely on genome annotation to some degree; the isoform-based approach allows better biological interpretation of the result since transcripts can directly correlate with protein products. In this dissertation, I mainly focused on the isoform-based approach.

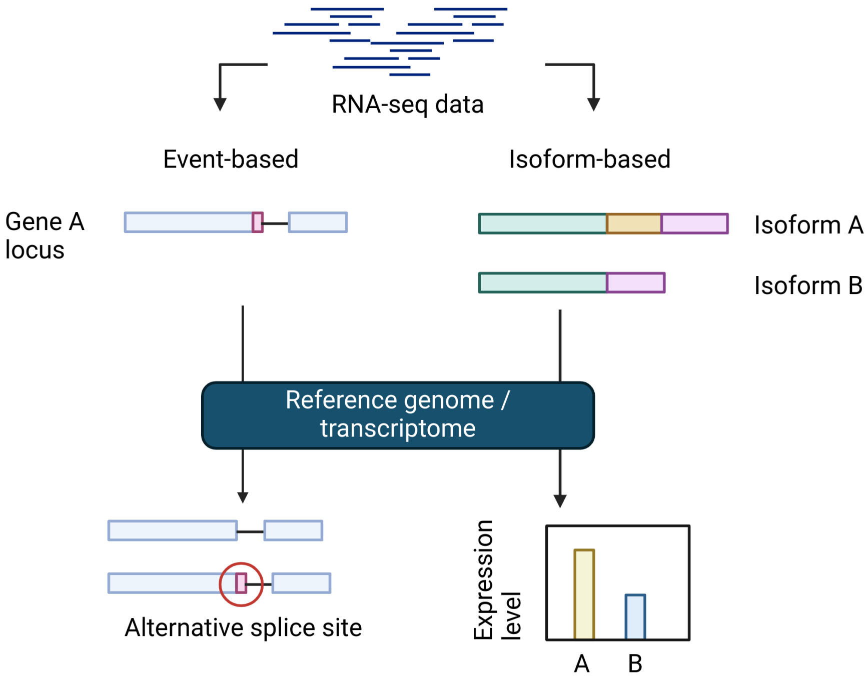


**Figure 2.9** Differences of event-based and isoform-based approach to alternative splicing analysis
In the event-based approach, the loci in Gene A with evidence in the RNA-seq data are compared to the annotated splice site in the reference genome and an alternative splice site is identified. For the isoform-based approach, the expression level of two isoforms, A and B, from the same gene are quantified by mapping RNA-seq reads to the reference transcriptome. Source: created with BioRender.com

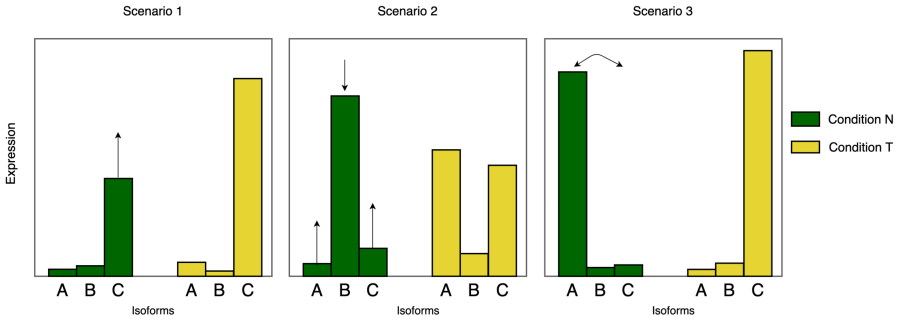


**Figure 2.10** Differences of differential transcript expression and differential transcript usage
Scenario 1 are the genes that has one differentially expressed transcript, which are usually detected as differentially expressed genes as well. Scenario 2 are the genes that changes the distribution of the transcript abundance between conditions. This change can be involving multiple transcripts. Scenario 3 indicates an isoform switch event. The arrows indicate the direction of changes of the transcript expression. Source: taken from [91] under CC-BY-NC-ND 4.0 International license

Differential transcript usage (DTU) analysis has the potential to identify changes in transcript expression within a gene that are specific to alternative splicing. DTU differs from differential transcript expression (DTE), illustrated in figure 3.4. DTU analysis focuses on detecting genes with differing distributions of transcript abundance between conditions. Here, transcript abundance refers to the ratio of transcript expression to the total expression of the gene. Genes with DTE are usually differentially expressed genes (DEG). In the left plot of the figure, only transcript C is differentially expressed between conditions. This kind of change is primarily due to the activation of transcription. In the right plot, however, transcript A is highly expressed in condition N while transcript C is highly expressed in condition T. Transcription of this gene is activated in both conditions; only the splicing pattern has changed to give rise to different transcripts. This redistribution behavior is called DTU. In addition, if two transcripts are involved in this abundance change, it is also known as isoform switch (IS).

Given that alternative splicing can significantly impact the proteome and other cellular processes, understanding these changes is crucial. However, traditional differential expression analysis alone cannot pinpoint which genes undergo alternative splicing. To address this, we can turn to DTU analysis, a quantitative method that effectively identifies changes in the proportion of transcripts within a gene.

Numerous tools are available to conduct DTU analysis. By leveraging these tools, researchers can gain a comprehensive understanding of the complex interplay between alternative splicing and gene expression, shedding light on the regulatory mechanisms that govern cellular processes.

In table 2.1, exon-centric tools focus on counting bins of reads overlapping with junctions. JunctionSeq uses generalized linear models (GLMs) to test for differential usage of splice junctions. Transcript-centric tools start with transcript counts. DEXSeq is an extension of DESeq2 intended to detect both differential exon usage [92] and DTU [93]. Like DESeq2, it uses a GLM on shrinkage dispersion to estimate the parameters for the Wald test. JunctionSeq [94] uses splice junction, and exon read counts from QoRT [95] for the DTU test, and DSGSeq calculates negative binomial statistics by comparing exon-counts between two conditions [96]. SeqGSEA combines the DSGseq and DESeq methods to generate a normalized metric of differential splicing (DS) and differential expression (DE) score using a rank-based strategy [97]. DRIMSeq uses a Dirichlet distribution to model the transcript abundance within a gene [98]. Dirichlet distribution takes a vector of density with k-dimension. It resulted in a k-vector of probabilities that can be added up to one, which makes it ideal to estimate the transcript ratio within the gene. DTUrtle extends it with extra post-hoc filtering steps prior to using StageR [99], allowing gene-wise false discovery correction [100]. Iso-KTSP tests switching transcript pairs by using a classifier [101]. The classifier determines if an event is 'tumor' or 'normal' by scoring every pair of transcripts in a gene. Each transcript pair is scored based on the frequency of one transcript having a higher expression than the other. The permutation test is applied to obtain empirical p-values. satuRn uses a quasi-binomial generalized linear model to model gene counts [102]. Cufflinks/cuffdiff performs de novo transcript alignment and differential expression analysis at both gene and transcript levels, using a Poisson model to model the variability of biological replicates regarding fragment counts in a transcript.

| Tool | Implementation | Year | Reference | Principle idea | excluded reason |
|---|---|---|---|---|---|
| DEXSeq | R | 2012 | [103] | Negative binomial generalized linear model to model transcript counts | |
| DRIMSeq | R | 2016 | [98] | Use dirichlet-multinomial model to model relative abundance of transcript | |
| seqGSEA | R | 2014 | [97] | negative binomial models in DSGSeq and DESeq | |
| DTUrtle | R | 2021 | [100] | dirichlet-multinomial model from DRIMSeq | |
| JunctionSeq | R | 2016 | [94] | Negative binomial generalized linear model to model junction counts | |
| NBsplice | R | 2020 | [104] | Negative binomial generalized linear model to model transcript counts, and test with a linear hypothesis | |
| satuRn | R | 2021 | [102] | Quasi-binomial generalized linear model to model transcript counts | |
| limmaDS | R | 2013 | [105] | Apply generalized least squares approach to detect transcript changes | |
| Cuffdiff2 | C++ | 2012 | [106] | Use a Poisson model to estimate changes in transcript counts | |
| iso-KTSP | Java | 2014 | [101] | Classify isoforms to condition specific based on the change of transcript abundance | |
| DSGSeq | R | 2013 | [96] | negative binomial statistics to detect transcript changes | |
| edgeR | R | 2010 | [88] | Negative binomial generalized linear model to model transcript counts | |
| IUTA | R | 2014 | [107] | Estimate transcript usage, followed by testing DTU under Aitchison geometry | Incompatible with STAR |
| IsoDOT | R | 2015 | [108] | Estimate transcript usage with a penalized regression method | Long run time |
| rSeqDiff | R | 2013 | [109] | Apply linear poisson model to estimate transcript counts | Not supporting replicates |

**Table 2.1** Table showing the DTU tools published after 2010

One DTU analysis is known as isoform switching, which focuses on a pair of transcripts from the same gene and explores their switch in abundance between two conditions. The goal is to detect genes that underwent splicing, which resulted in different transcripts between two conditions. To facilitate this analysis, IsoformSwitchAnalyseR incorporates DEXSeq and DRIMSeq to identify genes exhibiting isoform switching dynamics [110]. IsoformSwitchAnalyseR assesses each transcript pair within a gene and calculates a differential isoform fraction (dIF) value, providing valuable insights into the extent of transcript abundance changes between the two conditions. By integrating genomic references and annotations, this tool also discerns the specific consequences of isoform switching, such as alternative splice sites, alternative 5' or 3' UTRs, nonsense-mediated decay predictions, and polyA tails' characterization.

## 2.6 Network approaches and systems biology

Systems biology is an approach that aims to comprehend complex biological systems wherein all components interact functionally as modules. In this context, modules refer to components that share functional similarities. Much work is needed to understand the data before conducting advanced analyses like functional enrichment and network analyses of high-dimensional data like transcriptomics. Unsupervised methods enable us to examine the data and detect patterns based on the underlying variance of the samples. In this thesis, I introduce several prevalent methods used in systems biology: clustering, enrichment analysis, and co-expression analysis.

### 2.6.1 Clustering analysis

Clustering analysis is an unsupervised machine learning method aimed at finding common patterns among a group of similar objects. It is commonly applied to two major types of biological data: omics data (such as RNA-seq and ChIP-seq) and patient data. One common feature of both types of data is their high dimensionality. In RNA-seq data, clustering is often applied to extract features (e.g., genes) with similar expression patterns. However, only a few use cases have been applied to transcript-level data. Here, we will explore using clustering to study alternative splicing in RNA-seq data.

Clustering algorithms for biological data fall into five categories: partitioning k-means, hierarchical, density-based, model-based, and graph-based. K-means separate clusters into several clusters $k$, and a vector of mean or median represents each cluster. K-means aims to iteratively assign objects to the closest cluster to minimize the mean similarity between objects and clusters. Hierarchical clustering represents clusters as nodes of a tree. If two clusters have a common parent node, they are similar. The similarity will depend on the distance of the parent node to the clusters. Hierarchical clustering can form clusters from two methods: 1) Initially, each object is a cluster that merges into larger clusters based on similarity. 2) Initially, all objects are in one cluster, separated into smaller clusters. A linkage parameter determines the similarity. For example, a single linkage takes the minimum distance between two clusters, while a complete linkage is the maximum distance. Average distance takes the mean distance of the clusters' objects. Density-based clustering aims to find clusters of dense regions, enabling the discovery of clusters with different shapes in 2D space. A typical algorithm of density-based clustering is DBSCAN [111]. DBSCAN clusters a dataset spatially by categorizing each data point as either a core point (belonging to a cluster) or a border point (not part of any cluster). A core point is defined as a point that falls within a specified radius distance of another point and has a minimum number of neighboring points within that distance.

In RNA-seq data, gene clusters with similar expression patterns are assumed to be co-regulated by regulatory factors, such as transcription factors, or participate in a similar cellular function. For example, three phases are defined in the yeast metabolic cycle, and each phase is assigned a gene group using k-means clustering on RNA-seq data [112]. Clustering analysis can also stratify patients and discover new gene candidates in disease etiology. In [113], hierarchical clustering was used to identify cancer subgroups with distinct gene expression patterns for each group. K-means clustering and hierarchical clustering are standard algorithms applied to RNA-seq data. However, several aspects could affect the quality of the resulting clusters. For instance, k-means is susceptible to noise and outliers, leading to incorrect clustering. On the other hand, sparsity in single-cell data can lead to under-representation of the cluster, in which the clusters are very dense in lower dimensions [114]. In other words, clustering quality highly depends on the data and the algorithm, and there is no gold standard clustering algorithm for RNA-seq data.

Several metrics are used to evaluate clustering performance. Depending on whether ground truth is provided, these metrics can be categorized into internal and external metrics. Internal metrics are used without ground truth and indicate how similar the objects are within a cluster and how dissimilar they are between objects outside the cluster. Examples of internal metrics include the silhouette index and the Davies-Bouldin index. External metrics such as the adjusted Rand and Jaccard indexes compare the predicted clusters with the ground truth.

### 2.6.2 Time series analysis

Acquiring time course data is essential to understanding dynamic processes such as disease progression. By pinpointing clusters of genes that exhibit comparable temporal expression or AS/IS (alternative splicing/inclusion) patterns, we can analyze the mechanism of the disease's development. Research on mouse retinal development demonstrated that genes displaying analogous temporal exon usage patterns also shared similar biological functions and specificity for particular cell types [115]. Time series data present unique challenges for RNA-seq analysis with the extra time component. Consider the left plot in figure 2.11; when comparing two time series conditions, conventional pairwise comparison methods like DESeq2 (full model) will detect changes in expression levels between different time points. However, if we compare the time series, the patterns between the two conditions are similar. In the right plot, the time series patterns differ between the two conditions. To effectively analyze time series data, the time series analysis methods need to account for these time point differences and focus on comparing the conditions based on the changes observed over time. Several methods address the time component in differential gene expression analysis. DESeq2 introduced a likelihood ratio test for time series data, which uses a reduced model that removes the interaction between time and treatments and tests only for differences over time [90]. Next-maSigPro identifies temporally differential expressed genes between two conditions with a generalized linear model; this method enables the comparison of gene expression patterns between two time series conditions [116]. When studying time-dependent changes in transcript usage, Iso-maSigPro and TSIS are popular tools. Iso-maSigPro is an extension of Next-maSigPro that can identify differentially expressed transcripts and splicing events within genes and compares transcript usage patterns between two time series conditions [116]. While ANOVA can test differential expression in time series data, Next-maSigPro specifically accounts for transcriptomics data by modeling with negative binomial distribution instead of Gaussian distribution. TSIS, on the other hand, is designed for studies with only one time series condition and identifies pairs of transcripts that switch abundance over time by calculating switching probability, expression differences before and after the switch, correlation coefficient, and performing statistical tests [8]. However, TSIS resulted in a lot of low-expressed switching transcripts. I aimed to improve the isoform switch detection algorithm to tackle this issue.

### 2.6.3 Enrichment analysis

Functional enrichment analysis is a powerful tool for interpreting the results of biological data analysis and obtaining functional insights. It falls under a collective of methods with three main categories: over-representation analysis, gene set enrichment analysis, and topology-based analysis [117]. Over-representation analysis typically tests for a set of DEGs or differentially spliced genes (DSG). A Fisher

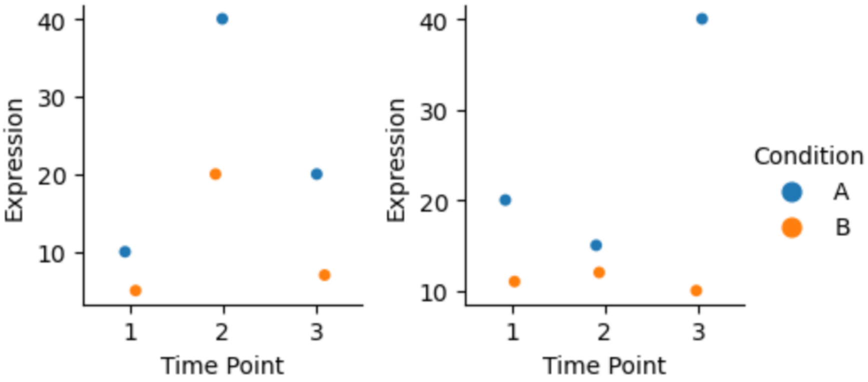


**Figure 2.11** Challenge of time series differential expression analysis
The plots show two time series expression patterns in conditions A and B, indicated with blue and orange colors. The left plot shows no significant pattern change between the two conditions. The right plot demonstrates the change of patterns over time.

exact test is often used to determine if the DEGs are enriched in a given gene set of a functional term, such as a biological pathway.
In gene set enrichment analysis, the DEGs are first ranked by log2 fold change or mean expression. The ranked gene list is then compared to a gene set, a set of genes associated with a pathway to obtain an enrichment score [118]. The enrichment score represents the over-representation of the gene set of a pathway or a GO term random walking along a ranked gene list (e.g., ranked based on p-values). Topology-based analysis incorporates interaction knowledge bases (e.g., protein-protein interaction) to extract the association of DEGs to the genes within a functional term.

However, the methods above focus on the gene level, where a set of genes is given as an input. In splicing, we could decipher the functions of genes that underwent alternative splicing without knowing the actual effect of the splicing events. The network-based enrichment method for AS Events (NEASE) uses different knowledge bases to determine the edges of a pathway affected by differentially spliced genes (DSG).[119].

In the cellular environment, proteins function cooperatively through interactions with each other, e.g., in a signaling cascade where receptors receive signals from the extracellular environment, which activates the effector. The activated effector then triggers a cascade of signaling kinases that activate the downstream transcription factor, initiating the transcription of target genes. However, RNA-seq data is often noisy and can identify more co-expressed genes than expected, leading to a problem in finding false-positive genes. To overcome this issue, interaction networks of prior knowledge (i.e., interactome) can be used to identify such genes. An interactome consists of biological entities, such as proteins or RNAs, interacting in biological systems. The edges between nodes in the network indicate the interaction between entities, which are either experimentally validated or predicted. Popular interactomes are STRING [120], BioGRID [121], I2D [122].

NEASE first retrieves domain information of DSGs from domain-domain interaction databases (DDI), such as 3did [123] and DOMINE [124], which contain the information about interactions between domains of different proteins [119]. If a splicing event of a DSG results in the disruption of a protein domain or protein binding motifs, the interactions of the DSG with other proteins will also be disrupted [125]. NEASE performs a hypergeometric test to determine if alternative splicing affects the interactions with a particular pathway more than by chance. With different knowledge bases, NEASE offers a comprehensive approach to splicing-aware functional enrichment analysis for RNA-seq data.

Network enrichment analysis involves projecting the gene set of interest onto the interactome in order to identify active or disease modules [126]. By applying statistical tests (e.g., fisher-exact test), significant active modules shed light on the connection between the module members and the disease context [127]. This approach enables researchers to gain valuable insights into the molecular mechanisms and interactions that are crucial to the disease under investigation.

One example of active module identification is the DOMINO algorithm, which aims to overcome bias related to the non-specificity of GO terms. This bias is revealed by comparing GO terms enrichment results from different network enrichment methods using randomized gene sets. DOMINO slices the interactome into highly connected modules using the Louvain modularity algorithm. Each module is sliced further to obtain subslices enriched in the gene set of interest using a fast prize-collecting Steiner tree (PCST) algorithm. PCST algorithm finds the connected subnetwork that minimizes the cost function, which is the sum of all prizes assigned to nodes and weights assigned to edges [128]. Another network enrichment example is DIAMOnD. The DIAMOnD algorithm starts with an input set of seed nodes, and all proteins that are connected to the seed nodes will be ranked by their connectivity p-values. This p-value represented the probability of getting the same or more connections of the seed protein than expected. The protein with the most significant p-value will then be added to the set of seed nodes. This process will repeat iteratively until the algorithm spanned through the whole network.

However, recent benchmark studies pointed out that the existing network enrichment algorithms suffer from biases due to the inherited properties of the PPI network [129]. The pitfalls will be discussed in detail in the discussion section.

### 2.6.4 Co-expression network analysis

Using a biological interactome can lead to unwanted false positives. These false positives are those connections in the biological interactome that are not interacting in the disease-context environment. For example, the interaction network in a disease condition may differ from that of a healthy condition. Furthermore, most biological networks are designed for general purposes. However, transcriptome profiles are tissue-specific. Co-expression analysis is employed to identify functional modules associated with the disease of interest based on omics data to address this issue. For transcriptomics, the analysis involves finding genes with correlated gene expression values, indicating that they are expressed simultaneously and likely to be co-regulated across samples. Co-expression analysis has been applied to various types of gene expression data, including transcriptomics [130], epigenomics [131], proteomics [132], and metabolomics [133], and has been used to study various biological systems, including complex diseases, developmental processes, and responses to environmental stress [134]. For instance, incorporating genomics data can extract expression quantitative trait loci (eQTL) signals that correlate to the expression profile [135]. This can further strengthen the findings of co-expressed transcripts and their contributions to

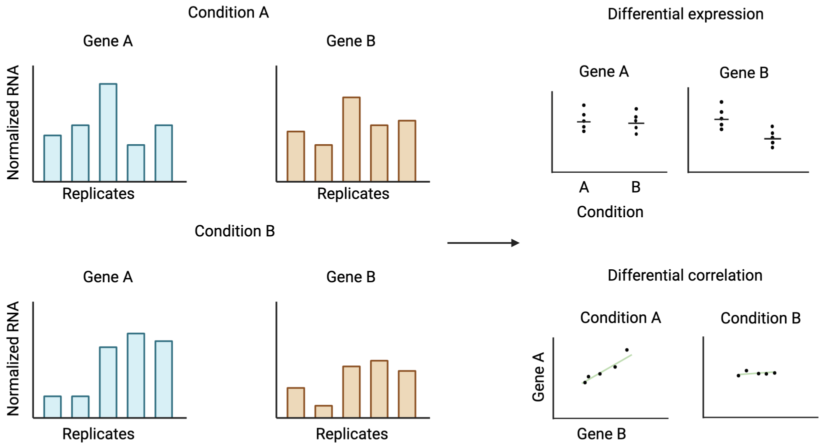


**Figure 2.12** Differential gene co-expression analysis
Differential gene co-expression analysis considers the relationship between a pair of genes. The bar plots show the count distribution of gene A and gene B replicates in conditions A and B. In differential expression analysis, gene A and gene B might have been detected down-regulated in condition B compared to condition A. In differential correlation, gene A and gene B are positively correlated in condition A, while the correlation is lost in condition B.

the disease risk.

Co-expression analysis is performed by comparing the expression levels of pairs of genes across samples and computing a similarity metric, such as Pearson correlation or Spearman rank correlation. Weighted correlation network analysis (WGCNA) utilizes a weighted adjacency matrix to represent gene relationships based on pairwise correlation coefficients. It applies a hierarchical clustering algorithm to identify clusters of highly co-expressed genes. The resulting gene networks can be visualized as dendrograms and used to identify modules of co-expressed genes associated with particular biological processes. Differential Gene Co-expression Analysis (DGCA) is used to identify groups of genes that are differentially co-expressed across different conditions, such as disease and normal states, and to identify gene interactions that are likely to be functionally related [136]. DGCA quantifies the expression levels of individual genes across different samples and uses statistical methods to identify pairs of genes with significantly different co-expression patterns (Figure. 2.12).

### Centrality measures

After constructing co-expression networks, a common subsequent analysis involves the identification of hub nodes. Hub nodes are defined as nodes that exert a significant influence on the overall topology of a network when they are removed. These nodes play an important role in maintaining the structural integrity and connectivity of the network. In a biological context, hub nodes often symbolize key regulators within a module or pathway. For instance, transcription factors typically oversee the expression of multiple genes, while specific kinases can activate numerous downstream effectors. These regulators play a pivotal role in the development of diseases. To identify hub nodes, four common types of centrality measures can be used to quantify the connections: 1) degree centrality, 2) closeness centrality, 3) betweenness centrality, and 4) eigenvector centrality.

Degree centrality refers to the count of connections a node possesses within a network. Closeness centrality involves computing the average shortest path length between a particular node and all other nodes in the network. The greater a node's centrality, the closer its connections are to other nodes. Betweenness centrality considers how often a node acts as a bridge between two other nodes within the network. Eigenvector centrality assigns a relative score to each node, wherein a node's centrality is proportional to the cumulative centrality of its neighboring nodes. In simpler terms, a node becomes significant if it is connected to essential neighbors.

# 3 General methods

## 3.1 Small RNA network analysis

This analysis aims to extract potential small RNA biomarkers in Alzheimer's disease mouse model Tg4-42 using differential co-expression network analysis; figure 3.1 illustrates the workflow. The small RNA expression dataset is obtained from a previous study [137] that performed small RNA sequencing in transgenic Alzheimer's mouse model Tg4-42 and wild type. The dataset contains miRNA and snoRNA expression values. I performed differential expression analysis and applied differential gene co-expression analysis (DGCA) to construct a differential co-expression network and centrality measures to extract potential biomarkers.

### Data normalization

In this analysis, I employed DESeq2 to obtain normalized counts [90]. Raw counts are first filtered for small RNAs with more than 10 counts across samples. The raw count table is given to DESeq2, which uses the RLE normalization method. The resulting normalized count table is then transformed using the variance-stabilization method [138].

### Differential co-expression analysis

The small RNAs' co-expression network is built using DGCA based on the Pearson correlation for the mouse model Tg4-42 and wild-type mouse. Each gene pair is calculated a correlation coefficient for gene expression from gene $x$ and gene $y$.

$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} \tag{3.1}$$

The correlation coefficient is transformed into a z-score using the Fisher z-transformation method [139]. First, the correlation coefficient is normalized and a natural logarithm function is applied:

$$z = \frac{1}{2}\text{log}_e(\frac{1+r}{1-r}) \tag{3.2}$$

The underlying distribution of the z-scores is assumed to be normally distributed. The variance can then be calculated using:

$$s = \frac{1}{n-3} \tag{3.3}$$

$n$ is the sample size used to calculate the Pearson correlation coefficients. The differential z-scores of a gene pair are then given by:

$$dz = \frac{(z_1 - z_2)}{\sqrt{|s_1^2 - s_2^2|}} \tag{3.4}$$

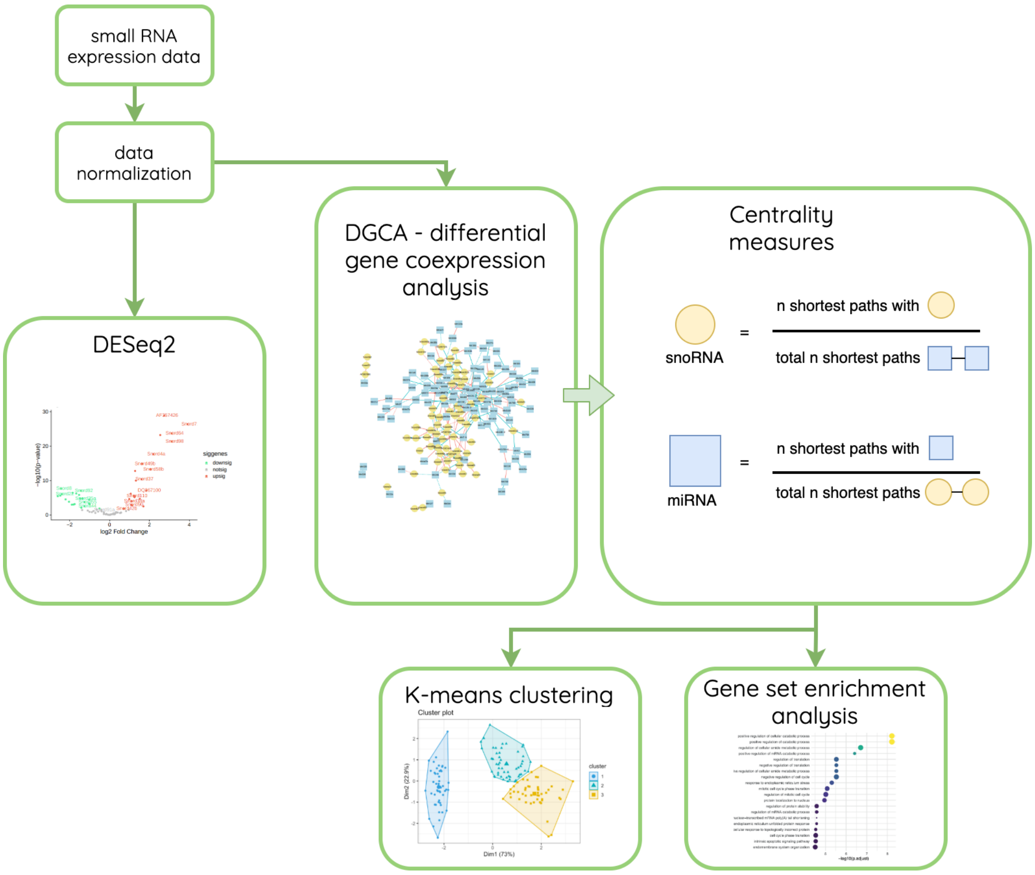


**Figure 3.1** Workflow of network analysis in small RNA-seq data.
small RNA expression data was obtained from small RNA sequencing. Differential gene expression analysis was performed using DESeq2. I then used DGCA to construct a differential gene co-expression network and centrality measures for all nodes for k-means clustering and gene set enrichment analysis.
Source: own work

The differential co-expression network is built using the differential z-scores of each small RNA pair [136]. In the network, nodes represent small RNAs, and edges indicate a change of correlation coefficients between the two conditions (e.g., positive correlation to negative correlation). An empirical p-value is calculated for each edge by shuffling the gene expression values, and multiple testing corrections are performed using the Benjamini-Hochberg method [140]. The interesting nodes are extracted from this network using centrality measures of the nodes mentioned in the following section.

#### Centrality measures

The concept of centrality generally quantifies the frequency with which a node appears on the shortest paths connecting all other node pairs in a network (see 2.6.4). In the publication, I employed a betweenness-derived centrality measure for each node, focusing either on miRNA or snoRNA (Figure 3.1). The count of shortest paths relative to the alternate type of small RNA determines this centrality. For example, the miRNA-centric centrality for each miRNA is calculated based on the number of shortest paths connecting each pair of snoRNAs, and the reverse applies for snoRNA-centric centrality. Small RNAs with high centrality are considered potential biomarkers.

Furthermore, to explore which pairs of small RNA will have similar correlation coefficient changing patterns, I conducted clustering analysis on the correlation coefficients of small RNA pairs using the k-means clustering method.

#### Gene-set enrichment analysis

I performed gene-set enrichment analysis on small RNAs with high centrality measures using GO terms [141] and KEGG [142] databases. Here, the interactors of these small RNAs are used since functional annotations of the small RNAs are limited. For miRNAs, miRDB obtains potential miRNA interactors with prediction scores higher than 0.7 [143]. While snoDB is used to obtain a list of curated snoRNA interacting genes [144].

## 3.2 Spycone framework

The second publication describes the framework for Spycone, a splicing-aware time-course network enricher. The framework includes an improved time course isoform switch detection algorithm, clustering analysis, active module identification, functional enrichment analysis, and splicing factor analysis. Figure 3.2 provides a detailed overview of Spycone's main components. The tool provides an exploratory analysis and functional characterization of alternative spliced genes. The following sections describe each analysis in detail.

### 3.2.1 Transcript-level analysis

There is no limitation on the type of omics data as input. The only requirement is a normalized dataset. First, transcripts are filtered out based on the desired expression level before IS detection (e.g., only transcripts with more than 10 TPM are analyzed). Isoform switch detection aims to extract pairs of transcripts

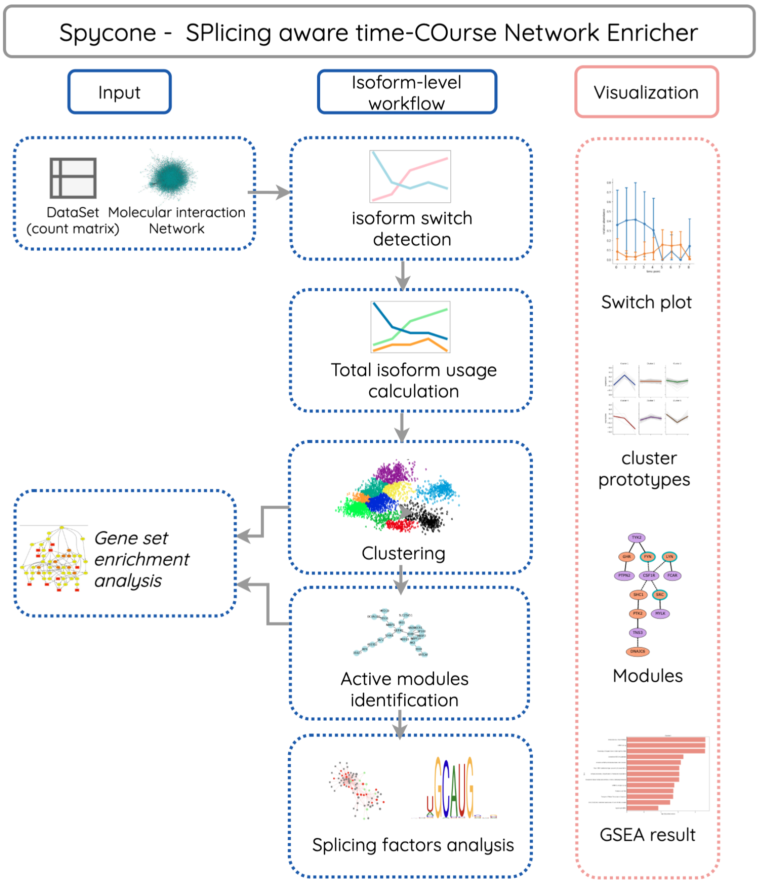

Spycone - SPlicing aware time-COurse Network Enricher
Input
Isoform-level workflow
Visualization
DataSet (count matrix)
Molecular interaction Network
isoform switch detection
Total isoform usage calculation
Clustering
Gene set enrichment analysis
Active modules identification
Splicing factors analysis
Switch plot
cluster prototypes
Modules
GSEA result

**Figure 3.2** Schematic overview of Spycone.

Spycone takes input as a count matrix of expression data and a pre-constructed molecular interaction network (e.g., PPI networks). The isoform-level workflow starts with isoform switch detection, which allows the visualization of detected isoform pairs. Then, the calculation of total isoform usage indicates the splicing patterns change, where clustering can group genes with similar splicing patterns. The clusters can then be visualized. DOMINO is integrated for active module identification. Clustering and DOMINO results can be further analyzed with gene set enrichment analysis. In addition, splicing factors analysis enriches potential splicing factors that directly bind to the switching isoforms. Source: taken from [145] under Creative Commons CC BY license that belong to the same gene and switch abundance between any time points. This abundance switch indicates that the underlying splicing pattern has changed by producing different gene versions.

### 3.2.2 Protein–protein interaction network and domain-domain interaction

Users can use a PPI network of their choice. By default, Spycone uses the protein-protein interaction (PPI) network from BioGRID (v.4.4.208) [121]. The domain-domain interaction network is obtained from 3did (v2019_01) [123]. The edges of the default PPI network are weighted according to the number of interactions between the domains of the protein (nodes of PPI). Weighting PPIs with domain-based information can result in a functionally more interpretable network in diseases and pathways [146].

#### Isoform switch detection algorithm

Isoform switch detection starts with the detection of switch points, where two time series of transcripts from the same gene are compared. For each pair of transcripts with a switch point, Spycone calculates six features: 1) switching probability, 2) significance of switch points, 3) difference of relative abundance, 4) event importance, 5) dissimilarity coefficient, and 6) domain inclusion or exclusion. Finally, p-values from the significance of switch points are corrected with multiple testing corrections.

- **Detection of switch points** Switch points are defined when two time series intersect. In Spycone, a switch point is defined where at least 60% of the replicates present the intersection. All switch points detected here are considered for the next steps.

- **Switching probability** Similar to TSIS, Spycone computes the switching probability for each IS event. The switching probability is determined by averaging the ratios of samples in which the relative abundance $I$ of isoform $i$ is higher than isoform $j$ before the switch $T1$, and vice versa, the ratio of samples in which the relative abundance $I$ of isoform $i$ is lower than isoform $j$ after the switch $T2$. For IS events where two isoforms switched between time intervals $T1$ and $T2$, the switching probability between isoform $i$ and isoform $j$ is calculated as follows:

$$P_x(\text{switch}) = [P(\sum_{t=x}^{T_y}(I_{i,t} > I_{j,t})) + P(\sum_{t=x}^{T_y}(I_{i,t} < I_{j,t}))]/2 \tag{3.5}$$

- **significance of switch points** A two-sided Mann-Whitney U test is applied to the replicates to test for the significance of the switch. The relative abundance of the transcripts before and after the switch point is tested.

- **Difference of relative abundance** Spycone measures the magnitude of changes during IS by calculating the average difference in relative abundance before and after a switch point. When replicates are present, Spycone computes the average change in relative abundance. A cutoff of 0.1 has been chosen to ensure that the changes in relative abundance account for at least 10% of the total gene expression. The difference in relative abundance $I$ between switching isoforms $i$ and $j$ at time point $t$ is defined as follows:

$$\text{Diff}_{i,j,s} = \Big[\sum_{r=1}^{R}(I^r_{i,s+1} - I^r_{i,s})/R + \sum_{r=1}^{R}(I^r_{j,s+1} - I^r_{j,s})/R\Big]/2 \tag{3.6}$$

- **Event importance** Event importance reflects the expression level of the transcripts involved in the switching events. I considered the event 'important' when the expression level of the transcripts is relatively high within the gene. Due to transcriptional noise, transcripts are often lowly expressed in the dataset. To address this, event importance is defined as:

$$\text{event importance} = \sum_{r=1}^{R}\left[\left(\frac{I^r_{aGt}}{max(I^r_{Gt})} + \frac{I^r_{aGt+1}}{max(I^r_{Gt+1})} + \frac{I^r_{bGt}}{max(I^r_{Gt})} + \frac{I^r_{bGt+1}}{max(I^r_{Gt+1})}\right)/4\right]/R \quad (3.7)$$

  where $I^r_{Gt}$ represents the relative abundance of isoform $a$ and isoform $b$ of gene $G$ at time point $t$, and $R$ is the total number of replicates. Each $I$ value is normalized to the highest relative abundance max($I^r_{Gt}$) observed at the corresponding time point. The metric calculates the average of the relative abundance of isoforms $i$ and $j$ before and after the switch event.

- **Dissimilarity coefficient** The underlying assumption is that when transcripts switch from one isoform to another, it results in a decrease in the expression of the latter isoform. The dissimilarity $d$ of transcripts is calculated from the Pearson coefficient:

$$r = \frac{cov(E_i, T_i)}{\sigma_E, \sigma_T} \quad (3.8)$$

$$d = \frac{1-r}{2} \quad (3.9)$$

- **Domain inclusion or exclusion** Pfam database v.35.0 [147] is used to map isoforms to their corresponding domains. Every pair of switching isoforms is compared in Spycone to detect any loss/gain domain in the IS events.

- **Multiple testing** I performed multiple testing corrections for IS detection using three available methods: Bonferroni [148], Holm–Bonferroni [149], and Benjamini–Hochberg [140] false discovery rate. By default, Spycone employed the Benjamini–Hochberg method.

#### Total isoform usage

The genes that undergo splicing changes are further analyzed. Each isoform-switched gene is transformed into a metric called total isoform usage. This metric sums up the magnitude of changes in each transcript within a gene.

$$\Delta\text{total.isoform.usage} = \sum_{A=0}^{n}\left|\left(\frac{I_{AGt1}}{\sum_{A=0}^{n}(I_{AGt1})} - \frac{I_{AGt0}}{\sum_{A=0}^{n}(I_{AGt0})}\right)\right| \quad (3.10)$$

Here, $I$ represents the expression of isoform $A$ of gene $G$ at time points $t1$ and $t0$, and $n$ is the total number of all isoforms for gene $G$.

### 3.2.3 Downstream analysis

#### Clustering

Clustering can be performed on direct transcript expression or total isoform usage—the former clusters transcripts of all genes by looking at the similarity of expression patterns. The latter clusters genes based on the similarity of splicing patterns since changes in total isoform usage imply changes in splicing. The clustering algorithms are imported from scikit-learn (v0.23.2) [150] and tslearn (v0.5.1.0) library [151].

### Functional enrichment analysis

For a transcript-level analysis, users can obtain clusters of genes with similar total isoform usage patterns or transcripts with similar expression patterns. Subnetworks are further extracted through the DOMINO algorithm [152]. Gene-set enrichment analysis (GProfiler) [153] and splicing-aware functional enrichment analysis (NEASE) [119] are implemented following isoform switch analysis.

### Splicing factor co-expression and motif enrichment analysis

This functionality aims to detect splicing factors that regulate the splicing of isoform-switched genes. I assume that the splicing factor that regulates the splicing of a gene cluster will have a similar expression pattern as the total isoform usage: when splicing factor expression increases, more genes undergo splicing. Spycone calculates the correlation between the expression of each pair of isoforms and the splicing factor. The splicing factors with correlation coefficient > 0.7 or < -0.7 are further investigated by calculating a PSSM score, indicating the potential of binding to the regulated RNA transcript. To calculate the PSSM score, the position weight matrix (PWM) of each splicing factor binding motifs are obtained from the mCross database ([154]). For a given position of a sequence and a nucleotide, the log-odd ratio of finding a specific nucleotide is indicated in the PWM. The PWM is matched with the sequence of interest (e.g., the 5'- and 3'- splice site flanking regions of spliced exons detected by Spycone). The PWM is matched to the sequencing at every position, which calculates the score by summing up the log-odd ratio of a specific nucleotide for each position. Hence, the higher the PSSM score of a given position, the higher the probability of the splicing factor binding to the motif on the exon region. This function is implemented using the motif module of the Biopython library [155].

## 3.3 Differential transcript usage detection benchmark

This benchmarking analysis aims to comprehensively evaluate the existing methods for detecting differential transcript usage (DTU) across static and dynamic data. Previous benchmark analyses have conducted comparisons of diverse workflows for splicing analysis, with a specific emphasis on DTU analysis. Liu et al. conducted a study where they compared eight differential splicing detection tools [10]. In this investigation, DEXSeq [103] and DSGSeq [96] achieved an area under the ROC curve (AUC) of approximately 0.8. Notably, Cufflinks [156]exhibited superior performance with a precision of 0.9 in de novo discovery. Another analysis, centered on human systems, was performed by Merino et al., where they evaluated differential expression and splicing tools across varying noise levels [11]. Their findings indicated that DEXSeq and LimmaDS [105] are the optimal tools for DTU detection. However, it is worth noting that the pipeline utilized TopHat [156] as the aligner, despite the established superiority of STAR [80] in splice-aware alignment [157, 158]. In the study by Fenn et al., the DICAST tool was introduced to benchmark eight event-based splicing detection tools [81]. Additionally, Jiang et al. conducted a large-scale comparison involving 21 event-based detection tools [159]. In this section of the thesis, I conducted an extensive comparison of DTU detection tools across various simulated scenarios and real-world datasets.

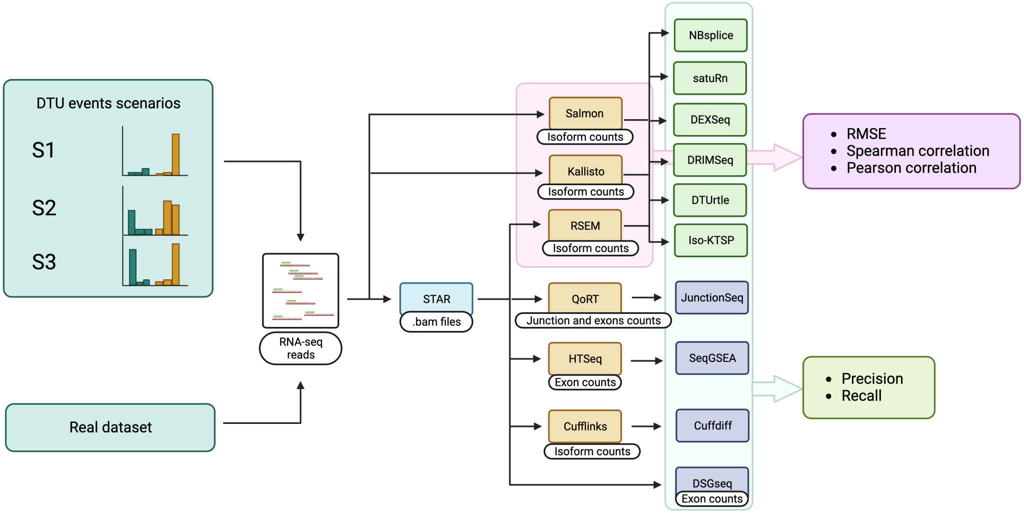


**Figure 3.3** Workflow for DTU detection.
Both simulated data and real transcriptomic data are utilized in this analysis. STAR is used as the aligner for this analysis. Different transcription quantification tools are used depending on the requirement for the DTU tools. For tools that don't work with a specific transcript count output, three popular choices are applied here: Salmon, kallisto and RSEM. The groundtruth and the estimated counts are compared using RMSE and correlations. The results of DTU tools are evaluated with precision, recall and F1 score.
Source: taken from [91] under CC-BY-NC-ND 4.0 International license

### 3.3.1 Evaluation of DTU detection

The performances of DTU detection methods are evaluated based on simulated datasets and a real-world transcriptomic dataset. I applied the DTU methods on the deep sequenced transcriptomics dataset (GSE222260) [160], sequenced in short-read paired-end sequencing. This dataset has also been used as an evaluation dataset in the previous benchmarking analysis [11]. The dataset contains twenty prostate cancer tumor samples and ten normal tissue samples and performed short-read RNA-seq. Figure 3.3 shows the overview of this analysis.

#### Simulation

The simulation process closely follows the methodology proposed by Merino et al., where I utilized RSEM (v1.3.3) to simulate both single-end and paired-end data. This was based on estimated abundances inferred from sequencing model parameters of real datasets, and a reference transcriptome [11]. I used the 'rsem-calculate-expression' function to estimate model parameters from actual datasets [161], collecting statistics such as the number of reads, alignment to multiple and unique loci, read and fragment length distribution, and quality score distribution. I estimated model parameters from GSE157490 [162] for single-end data, a cell line dataset with SARS-Cov2 infection sequenced at 100M reads. I used GSE162562, a patient dataset with SARS-Cov2 infection, for paired-end data sequenced at 100M reads [163]. Each dataset was simulated with 50 million reads, the minimum depth considered robust for DTU detection [164], and extended to 100 million reads.

I took baseline transcript expression levels from the SARS-CoV2 datasets and adjusted the transcript counts to create simulated data for three DTU scenarios (Figure 3.4). Recognizing that changes in transcript expression and DTU are intertwined with overall gene expression changes, I simulated both effects concurrently. I considered a random fold change between 2 and 5 for each gene across conditions. The transcript ratios were generated using a Dirichlet distribution, which calculates probabilities for k categories within a k-dimensional density distribution. This method is ideal for simulating transcript ratios as the sum of the vectors equals 1. Here, k represents the number of transcripts in a gene, with each assigned an expression value. The higher the probability for transcript i, the greater its expression value. To enhance statistical power in detecting DTU transcripts, I simulated DTU transcripts at higher expression levels. The expression value for each transcript i in condition j was simulated using the following formula:

$$\text{Transcript expression}(i,j) = \text{baseline gene expression} * \text{fold change} * \text{transcript ratio} \tag{3.11}$$

For baseline (e.g., a control), the fold change is 1, while for the condition of interest, I considered the random fold change. To create replicates with measurement noise, I computed the expression values using a negative binomial distribution, with dispersion for each transcript estimated via DESeq2:

$$\text{Transcript expression of replicate x}(i,j) = \text{negative binomial distribution}(\text{transcript expression}, 1/\text{dispersion}) \tag{3.12}$$

I distributed the genes across scenarios to achieve a mixed final dataset. While I could have considered datasets focusing on individual scenarios, this would not have provided a realistic dataset for tool evaluation. Next, I modified the transcript ratios according to the scenarios. In scenario S1, only a single transcript per gene changes expression. In scenario S2, more than two transcripts change relative abundance. Scenario S3 involves swapping the relative abundance of two transcripts, indicating an isoform switch event.

I simulated three background levels, 0, 0.1, and 0.5, representing increasing fractions of genes whose expression remains unchanged. The modified transcript results were then used for further simulation. The ’rsem-simulate-reads’ command facilitated this process, with the RSEM reference created using the human genome GRCh38 and the theta0 parameter set to a noise proportion of 0.1 in the background. I simulated four conditions, integrating different parameter combinations as outlined in Table 1. Simulations with 100M reads were conducted only with four replicates at a background level of 0.5.

### Simulation for single-cell data

I used two simulators for single-cell data simulation: RSEM and scDesign3 [165]. A demultiplexed Smart-seq2 dataset from human cells is used for parameter estimation [27]. Before the simulation, I followed the Seurat workflow for grouping cell types [166] by first embedding the cells to a k-nearest neighbor graph and applying the Louvain algorithm [167] to cluster the cells. For RSEM, we employed an identical simulation workflow for the bulk RNA-seq data, except that the single-cell-prior parameter is used. Our simulation method was adjusted to accommodate single-cell transcript counts. Specifically, we utilized the parameters designated for single-cell analysis in RSEM to replicate sparse matrices. To evaluate the methods using straightforward single-cell data, we employed two cell types with a collective population

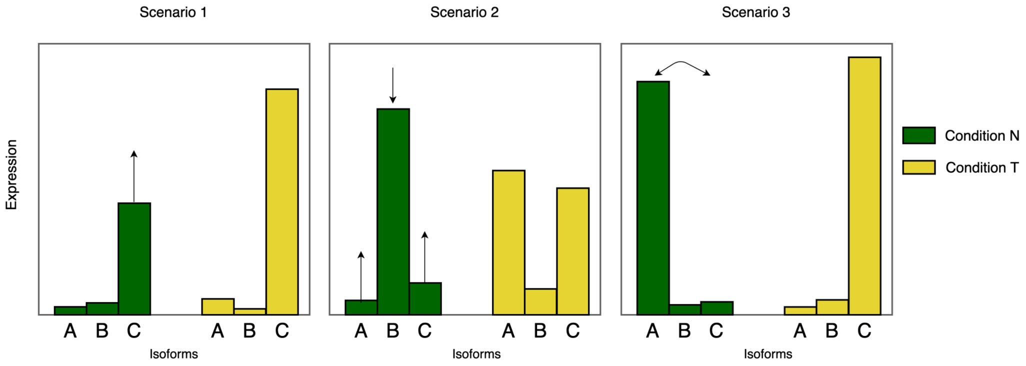


**Figure 3.4** Simulation of different DTU scenarios.
DTE refers to genes with a single differentially expressed transcript, often identified as differentially expressed genes too. DTU, on the other hand, involves genes that exhibit a shift in the distribution of transcript abundance across different conditions, potentially affecting multiple transcripts. A specific case of DTU is the isoform switch (IS), where the abundance is redistributed between two transcripts.

of 900 for simulation. We generated balanced datasets comprising 20, 50, 100, 200, 500, and 1000 cells, with an equal number of cells for each cell type. We incorporated differential transcript usage (DTU) events within the single-cell simulations with two background levels - 0 and 0.1. Here, the background level denotes the proportion of expressed genes that remain unaltered between the two cell types. In addition, scDesign3 is used for another simulation, which is dedicated to single-cell data simulation [165]. scDesign3 uses negative binomial distribution to fit each gene's marginal distribution and a Gaussian copula to model high-dimensionality in single-cell data. After estimating the parameters for the simulation, the mean count of the genes for each cell is modified to assign DTU events (as described in the Simulation section). I simulated balanced and unbalanced datasets with 0 and 0.1 backgrounds. Each balance dataset contains two groups of cell types, each with the same number of cells (ranging from 50 to 700). Each unbalanced dataset contains two to seven groups of cell types, and each cell type consists of a random number of cells.

### Differential transcript usage detection

I performed a comprehensive literature review using keywords like "differential transcript usage," "differential isoform usage," and "isoform switch" in PubMed and Google Scholar. This search identified 19 DTU detection tools published since 2010. However, I excluded Iso-DOT [108] due to its lengthy runtime without parallelization, rSeqDiff [109]for not accounting for replicates, and IUTA [107] due to its incompatibility with STAR output bam files and lack of sufficient documentation. The search revealed various tools for detecting differential transcript usage (DTU), including exon/junction-centric tools such as JunctionSeq [94], seqGSEA [97], DSGSeq [96], and transcript-centric tools like DEXSeq [103], DRIM-Seq [98], DTUrtle [100], iso-KTSP [101], satuRn [102], NBsplice [104], LimmaDS [105], edgeR [88], along with assembly-based tools such as Cufflinks/cuffdiff [106].

In DEXSeq, the perGeneQValue is utilized to obtain adjusted p-values. DRIMSeq employs the dmTest function, DTUrtle uses the posthoc_and_stager function, edgeR applies the diffSpliceDGE function, Limma uses the diffSplice function, and satuRn and JunctionSeq use the testDTU and runJunctionSeqAnalyses functions, respectively. NBSplice employs the NBTest function. For tools that provide adjusted p-values, genes with values below 0.05 are deemed significant. For tools that yield gene scores, I set the threshold for significant genes at the recommended level (iso-KTSP at 0.8 and DSGSeq at 2).

### Evaluation

To assess the performance of various tools on simulated data, I used precision, recall, and F1 score as key metrics. We compiled a list of genes marked as having DTU, either by having an adjusted p-value below 0.05 or exceeding a set threshold for tools like iso-KTSP and DSGseq. These genes are considered positive. True positives (TP) are identified as those positive genes also present in the simulated ground truth, with the remainder classified as false positives (FP). Genes that are simulated with DTU but not detected by the tools are considered false negatives (FN). We calculate each event scenario separately for stratification analysis with the corresponding ground truth. In different DTU event scenarios, P represents genes simulated under the respective scenario. The formulas for precision and recall are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.13}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.14}$$

$$\text{F1 score} = \frac{2TP}{2TP + FP + FN} \tag{3.15}$$

For the transcript counts derived from kallisto, RSEM and Salmon, several measures are calculated to evaluate the accuracy of isoform quantification.
The Root mean square error (RMSE) can be calculated using the following formula:

$$\text{RMSE} = \sqrt{\sum (E_i - T_i)^2/n} \tag{3.16}$$

Here, $E_i$ represents the estimated count of transcript $i$, $T_i$ represents the ground truth count of transcript $i$, and $n$ is the total number of transcripts.
The Spearman correlation can be calculated using the following formula:

$$\text{Spearman} = \frac{6\sum(d_i)^2}{n(n^2 - 1)} \tag{3.17}$$

Where $d_i$ represents the difference in rank between the ground truth counts and the estimated counts. The differences are summed up and scaled to obtain the final correlation coefficient based on the number of transcripts $n$. The Pearson correlation can be calculated using the following formula:

$$\text{Pearson} = \sum \frac{((E_i - M_E)(T_i - M_T))}{((n-1)S_E(S_T))} \tag{3.18}$$

Here, $E_i$ represents the estimated count of transcript $i$, $T_i$ represents the ground truth count of transcript $i$. Standard deviation is computed for both the estimated counts ($S_E$) and ground truth counts ($S_T$), while $M_E$ and $M_T$ are the corresponding means.

# 4 Publications

## 4.1 Small RNA Sequencing in the Tg4–42 Mouse Model Suggests the Involvement of snoRNAs in the Etiology of Alzheimer's Disease

### Citation

The article titled "Small RNA Sequencing in the Tg4–42 Mouse Model Suggests the Involvement of snoRNAs in the Etiology of Alzheimer's Disease" has been published online at Journal of Alzheimer's Disease on 14 June 2022.

### Full citation:

"Lio, Chit Tong, Tim Kacprowski, Maik Klaedtke, Lars R. Jensen, Yvonne Bouter, Thomas A. Bayer, and Andreas W. Kuss, Small RNA Sequencing in the Tg4–42 Mouse Model Suggests the Involvement of snoRNAs in the Etiology of Alzheimer's Disease, Journal of Alzheimer's Disease 87, vol. 87, no. 4, 14 June 2022, Pages 1671-1681, 10.3233/JAD-220110"

### Summary

The paper investigates the role of small nucleolar RNAs (snoRNAs) in Alzheimer's disease (AD) using the Tg4-42 mouse model. At the same time, dysregulation of miRNAs in AD is often studied as potential biomarkers, and snoRNA involvement in AD needs to be better understood. The Tg4-42 mouse model develops AD-typical neurological phenotypes, including synaptic hyperexcitability, glucose metabolism loss and gliosis, as one of the few models that develop neuron death in the hippocampus without plaque formation. This mouse model also expressed exclusively typical wild-type human A$\beta_{4-42}$ sequence; we believe this model will provide us valuable insights into the etiology of AD.

In this study, we aimed to investigate changes in the Tg4-42 mouse model that lead to AD-like phenotypes. Wild type and Tg4-42 mice of different ages (3 months and eight months old) are compared to elucidate the change before and after the onset of hippocampal neural loss, one of the hallmarks of AD. We employed differential co-expression analysis of small RNA sequencing data to construct a differential co-expression network. Unlike in differential gene expression analysis, where changes in genes are tested independently. Differential gene co-expression analysis allows us to extract pairs of small RNAs that change their co-expression behavior. This change in co-expression behavior indicates the underlying change in regulatory mechanism. The resulting differential co-expression network represents the relationship between each pair of small RNAs compared to wild-type and Tg4-42 mice.

We identified seven snoRNAs and five miRNAs that show high centrality. Centrality measures are often applied to networks to find the most important node. In addition, some of those miRNAs are associated with

AD. We found clusters of pairs of small RNAs with differential co-expression relationships between wild-type and Tg4-42 mice. The direction of changes in the co-expression relationship defines these clusters. We further investigate the potential functionality of these snoRNAs and miRNAs by performing gene set enrichment analysis on the gene interactors based on available databases. The results further highlighted several snoRNAs and microRNAs, previously not linked to AD, suggesting new avenues for research. The study adds to understanding AD's molecular basis and suggests that snoRNAs play a significant role in its etiology.

### Availability

The dataset used in this research is publicly available on the European Nucleotide Archive (https://www.ebi.ac.uk/ena) with the accession identification number of the project PRJEB39314.

### Contribution

I designed and wrote all the code for the analysis, from the exploratory data analysis which I tried different approaches to finalizing the analysis workflow. I generated all the figures in the manuscript. I contributed in writing the manuscript as a collaborated effort. Following submission and review, I addressed the reviewer's comments by doing additional analysis.

### Rights and permissions

Reprinted from Journal of Alzheimer's Disease, vol. 87, no. 4, Lio, Chit Tong, Tim Kacprowski, Maik Klaedtke, Lars R. Jensen, Yvonne Bouter, Thomas A. Bayer, and Andreas W. Kuss, Small RNA Sequencing in the Tg4–42 Mouse Model Suggests the Involvement of snoRNAs in the Etiology of Alzheimer's Disease, Pages 1671-1681, Copyright (2022), with permission from IOS Press

### Additional supplementary material

DOI 10.3233/JAD-220110
IOS Press

# Small RNA Sequencing in the Tg4-42 Mouse Model Suggests the Involvement of snoRNAs in the Etiology of Alzheimer's Disease

Chit Tong Lio[a,b], Tim Kacprowski[c,d], Maik Klaedtke[e], Lars R. Jensen[e], Yvonne Bouter[f], Thomas A. Bayer[f,*] and Andreas W. Kuss[e,*]

[a] *Chair of Experimental Bioinformatics, Technical University of Munich, Freising, Germany*
[b] *Chair of Computational Systems Biology, University of Hamburg, Hamburg, Germany*
[c] *Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Braunschweig, Germany*
[d] *Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig, Germany*
[e] *Department of Functional Genomics, Human Molecular Genetics Group, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany*
[f] *Department of Psychiatry and Psychotherapy, Division of Molecular Psychiatry, University Medical Center Goettingen (UMG), Georg-August-University, Goettingen, Germany*

Handling Associate Editor: Yuk Yee Leung

Accepted 5 April 2022
Pre-press 2 May 2022

**Abstract**.

**Background:** The Tg4-42 mouse model for sporadic Alzheimer's disease (AD) has unique features, as the neuronal expression of wild type N-truncated A$\beta_{4-42}$ induces an AD-typical neurological phenotype in the absence of plaques. It is one of the few models developing neuron death in the CA1 region of the hippocampus. As such, it could serve as a powerful tool for preclinical drug testing and identification of the underlying molecular pathways that drive the pathology of AD.

**Objective:** The aim of this study was to use a differential co-expression analysis approach for analyzing a small RNA sequencing dataset from a well-established murine model in order to identify potentially new players in the etiology of AD.

**Methods:** To investigate small nucleolar RNAs in the hippocampus of Tg4-42 mice, we used RNA-Seq data from this particular tissue and, instead of analyzing the data at single gene level, employed differential co-expression analysis, which takes the comparison to gene pair level and thus affords a new angle to the interpretation of these data.

**Results:** We identified two clusters of differentially correlated small RNAs, including Snord55, Snord57, Snord49a, Snord12, Snord38a, Snord99, Snord87, Mir1981, Mir106b, Mir30d, Mir598, and Mir99b. Interestingly, some of them have been reported to be functionally relevant in AD pathogenesis, as AD biomarkers, regulating tau phosphorylation, TGF-β receptor function or Aβ metabolism.

*Corresponding authors: Andreas W. Kuss, Human Molecular Genetics Group, Department of Functional Genomics, Interfaculty Institute for Genetics and Functional Genomics, University Medicine Greifswald, Greifswald, Germany. Tel.: +49 0 3834 4205814; E-mail: andreas.kuss@uni-greifswald.de and Thomas A. Bayer, Division of Molecular Psychiatry, Department of Psychiatry and Psychotherapy, University Medical Center Goettingen (UMG), Georg-August-University, Goettingen, Germany. Tel.: +49 (0)551 39 22912; E-mail: thomas.bayer@medizin.uni-goettingen.de.

ISSN 1387-2877/$35.00 © 2022 – IOS Press. All rights reserved.

**Conclusion:** The majority of snoRNAs for which our results suggest a potential role in the etiology of AD were so far not conspicuously implicated in the context of AD pathogenesis and could thus point towards interesting new avenues of research in this field.

Keywords: Alzheimer's disease, miRNA, NGS, RNA-Seq, small RNA, snoRNA, Tg4-42, transgenic mouse model

## INTRODUCTION

While an increasingly large number of studies are focusing on miRNAs and their putative involvement in the etiology of Alzheimer's disease (AD) and/or their potential as biomarkers for AD (reviewed in [1–5]), almost nothing is known about small nucleolar RNAs (snoRNAs) in this context. These are small RNA molecules of 60–300 nucleotides in length, predominantly found in the nucleolus, which play a role in the posttranscriptional modification of other RNAs, e.g., ribosomal RNAs. Most snoRNAs belong to either the group of box C/D snoRNAs or box H/ACA snoRNAs (reviewed in [6]). These groups are defined by conserved signature-sequence elements and characteristic secondary structures [7, 8]. C/D as well as H/ACA RNAs are important for protein translation, rRNA acetylation, mRNA abundance, splicing as well as translational efficiency, genome stability, and/or other cellular processes [9]. While H/ACA RNAs are involved in pseudo-uridinylation or the conversion of uridine to pseudo-uridine, C/D RNAs predominantly mediate 2′-O-ribose methylation [10].

In this study, we investigated small RNA NGS data of the hippocampi of Tg4-42 mice [11], specifically focusing on snoRNAs. The Tg4-42 mouse line is one of only few mouse models developing an AD-typical neurological phenotype starting with synaptic hyperexcitability [12–14] and loss of glucose metabolism [15] followed by behavioral deficits, severe neuron loss, gliosis, and metabolic changes of the glutamate/4-aminobutyrate-glutamine axis [14, 16–18]. It is well established that N-truncated $A\beta_{4-42}$ is highly abundant in the brain of AD patients [19, 20] and even represents a major species in plaques of affected individuals [21, 22]. Previously, we have compared the transcriptomes of the 5XFAD model, which shows early plaque formation, intraneuronal $A\beta$ aggregation, neuron loss, and behavioral deficits [23], and the Tg4-42 model with intraneuronal N-truncated $A\beta_{4-42}$ accumulation, neuron loss as well as behavioral deficits, without plaque formation [24]. Using deep sequencing, differentially expressed genes (DEGS) were identified. While many DEGs were identified in 5XFAD or in Tg4-42 mice, 36 DEGs were found in both mouse models indicating common disease pathways associated with behavioral deficits and neuron loss [24].

While 5XFAD and all other commonly used murine models are associated with familial forms of AD, Tg4-42 mice express wild type N-truncated $A\beta_{4-42}$ without any mutations under the control of the murine neuron-specific Thy1-promoter and are thusly ideal for modelling sporadic forms of AD. Therefore, in this study we opted for Tg4-42.

Using a differential co-expression analysis approach, our aim was to identify differences in the networks of small RNAs between Tg-42 and wild type (WT) hippocampi, which could point to hitherto unknown molecular players in the pathogenesis of AD. Along with several miRNAs, this led to the identification of nine snoRNAs, which seem to play a role in the molecular basis of the AD phenotype and were hitherto not known to be involved in the etiology of this disorder.

## MATERIALS AND METHODS

### *NGS-dataset of small RNAs in mouse hippocampus*

In order to elucidate the influence of small RNAs involved in neuron loss and associated memory decline in AD, we used 3-month-old and 8-month-old Tg4-42 mice, which correspond to time points before and after onset of hippocampal neuron loss and memory deficits in the Tg4-42 model [14]. We based our analyses on our previously reported NGS-datasets of small RNAs [11] for eight WT and eight Tg4-42 mouse hippocampi, respectively. Each of the two groups (WT and Tg4-42) contained samples from four young (three months of age) and four aged (eight months of age) animals.

### *Differential expression analysis*

During quality control of the mapped data, we excluded RNAs with less than 10 read-counts across all samples and collapsed technical replicates after heatmap inspection as previously reported [11]. In our previous work, we employed DESeq2 [25], that compares expression values of an individual gene at different conditions, where dysregulated miRNAs were determined by the level of log2 fold change and its significance after hypothesis testing. In this work, however, instead of single gene level analysis, we employed differential co-expression analysis with DGCA [26] that takes the comparison to the level of gene-gene pairs (see below). With that, we constructed a differential co-expression network where we assume that there are changes in the regulatory relationship between two connected small RNAs at different conditions and that they are likely to be regulated through the same mechanism or share similar functions. Dysregulated small RNAs are those with high centrality within the network that indicates its importance to the network integrity.

### *Differential co-expression and network analysis for miRNA and snoRNA*

In order to establish a correlation network, we used *DGCA*, an R-package that performs differential correlation analysis for two conditions [26]. We carried out all possible comparisons with *ddcorAll* functions using Pearson correlation. DGCA calculates correlation coefficients for each small RNA pair and transforms them into z-scores while *p*-values are calculated by permutation. Pairs with an adjusted *p*-value < 0.05 (using the Benjamini-Hochberg method) and a differential z-score of more than three have been selected for further analyses.

With this approach, we compared the eight WT with the eight Tg4-42 samples, using a variance-stabilized expression matrix, containing both snoRNAs and miRNAs. A network was established using the z-score differential coefficient. We determined centrality values for each type of small RNAs with respect to the other type. That is, betweenness centrality reflects how many times a snoRNA is part of the shortest paths between all pairs of miRNAs in the network, and *vice versa*. We then performed a clustering analysis of centrality values using the Heatmap function in *ComplexHeatmap* [27], setting the *row_km* parameter at 2, thus splitting the heatmap in two clusters by k-means clustering [28].

### *Functional enrichment analysis of small RNAs*

Small RNAs with a high centrality and their interaction partners were further analyzed using *clusterProfiler* R package [29]. *clusterProfiler* provides an interface for gene ontology [30] and pathway enrichment analysis such as KEGG [31]. It uses a hypergeometric model to test for enrichment of biological functions in a given gene list. Interaction information was extracted from the RNAInter database [32], which comprises RNA-associated interactions based on computational prediction and/ or experimental validation as well as from other databases such as miRDB [33]. We extracted the potential RNA interactors of the small RNAs with prediction score higher than 0.7. This score is calculated based on experimental confidence (number of publications supported), scientific community confidence (number of citations of the supporting publications) and types of tissues/cells (number of tissue and cell types the interaction being observed in). For snoRNAs, snoDB was used to obtain a curated list of snoRNA interacting genes [34].

## RESULTS

In order to find age-independent significant alterations between WT and Tg4-42 mice, we performed a differential correlation analysis comparing groups of eight WT and eight Tg4-42 individuals, which in each case comprised four young (three months of age) and four old (eight months of age) animals. We used a variance-stabilized expression matrix, containing both snoRNAs and miRNAs and generated a correlation network representation as shown in Fig. 1. The network consists of 68 gene pairs with significant correlations in at least one of the conditions (WT or Tg4-42) with a differential z-score, i.e. difference in correlation between the two conditions, higher than 3 or lower than –3.

To identify similar regulation patterns, we then performed k-means clustering with all pairs of small RNA in the network and thus identified two clusters of small RNA pairs, based on a correlation shift between WT and Tg4-42. Cluster 1 comprises pairs with a shift from positive to negative correlation, while the pairs forming cluster 2 show the opposite behavior (Fig. 2; see Supplementary Table 1 for specific values).

We also determined centrality values for the individual small RNAs included in our correlation network with respect to the other type of small RNAs (Supplementary Table 2 for miRNAs, Supplementary Table 3 for snoRNAs). By identifying all the shortest paths within one small RNA type and then counting how many times each node of another type falls

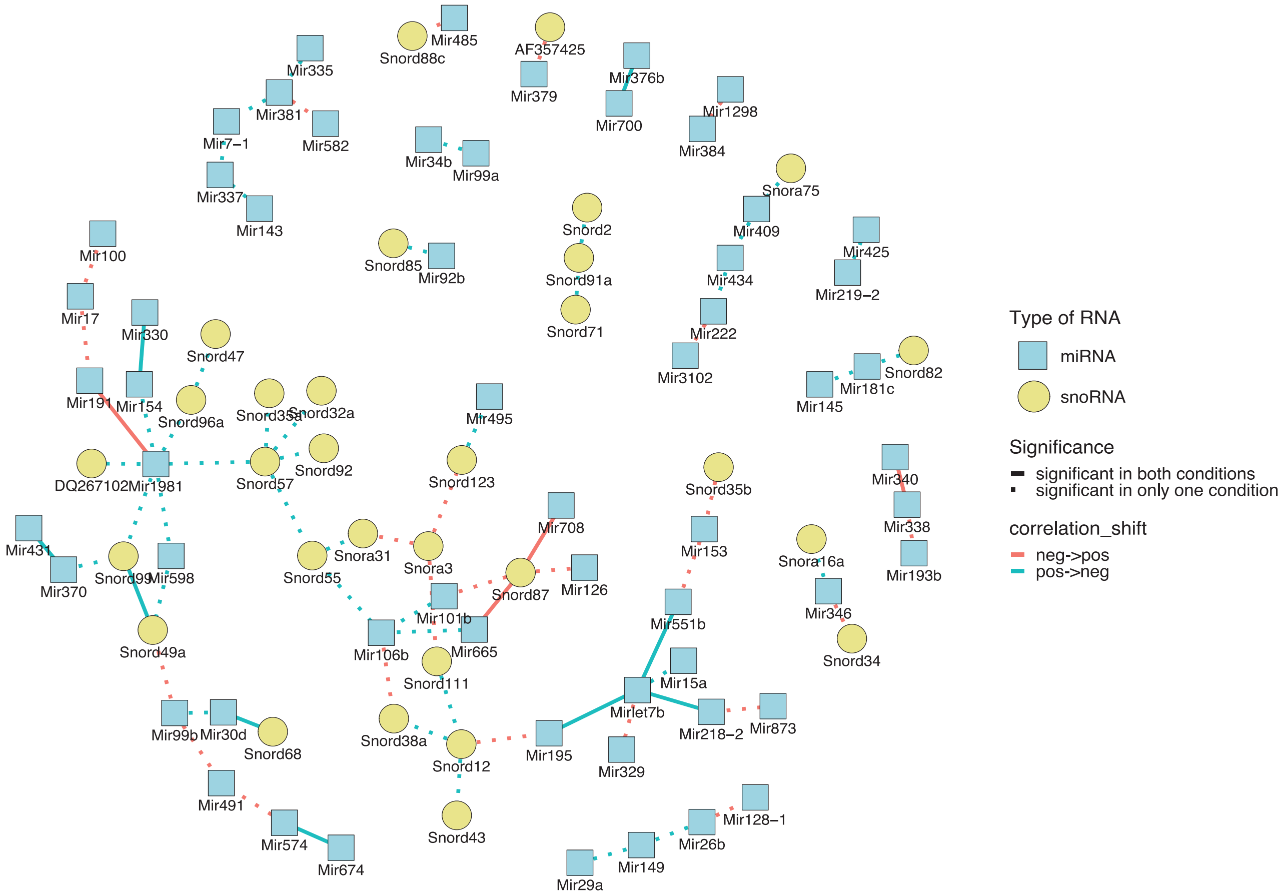


Fig. 1. Correlation network of significantly differentially correlated gene pairs when comparing WT and Tg4-42 mice. The type of RNA is represented by the shape and color of the nodes: miRNAs are depicted as blue squares and snoRNAs as yellow circles. The type of correlation shift is represented by the color of the connecting lines: red indicates the shift from a negative correlation in WT to a positive correlation in Tg4-42, the opposite is indicated by blue connecting lines. Solid lines connect small RNA pairs that were found significant in both, WT and Tg4-42, dotted lines indicate significance in only one, WT or Tg4-42.

on one path, this measure identifies small RNAs that can be seen to act as quasi bridges between small RNAs of the other type. As such, betweenness centrality values indicate the influence of a node on the interaction between other nodes in the network. In total we observed seven snoRNAs and five miRNAs with a centrality value higher than 0.1 (Table 1), and we could observe that 22 small RNA-pairs in Cluster 1 and nine pairs in Cluster 2 contain at least one partner with a high centrality (>0.1), yet only in nine pairs from Cluster 1 and two from Cluster 2 both partners had a centrality above 0.1 (Fig. 2). Interestingly, the majority of these pairs (21 of 31) contain at least one snoRNA and, on average, the seven snoRNAs had a higher centrality as compared to the five miRNAs. Since the relatively high centrality values connected with individual RNAs point towards a more prominent functional role of these molecules in the investigated context, we performed a literature search focusing on all the small RNAs with a centrality higher than 0.1 that we observed in this study. This yielded results only for three miRNAs and one snoRNA (Table 1), so that to the best of our knowledge, with the exception of Snord49a, all other snoRNAs we describe here are hitherto completely unknown in the context of AD.

Interestingly, one miRNA (Mir1981) correlates significantly with five of the other small RNAs, four times highly positive in the hippocampi of WT and strongly negative in the hippocampi of Tg4-42 animals (Fig. 2). What is more, three of the negative correlations of Mir1981 in Tg4-42 mice involve small RNAs (Mir598, Snord57, Snord99) that showed centrality values, which were also comparatively high (Fig. 2, Table 1).

Another interesting observation we made among the small RNA pairs with low centrality values of both partners concerns pairs containing Mir222 (centrality

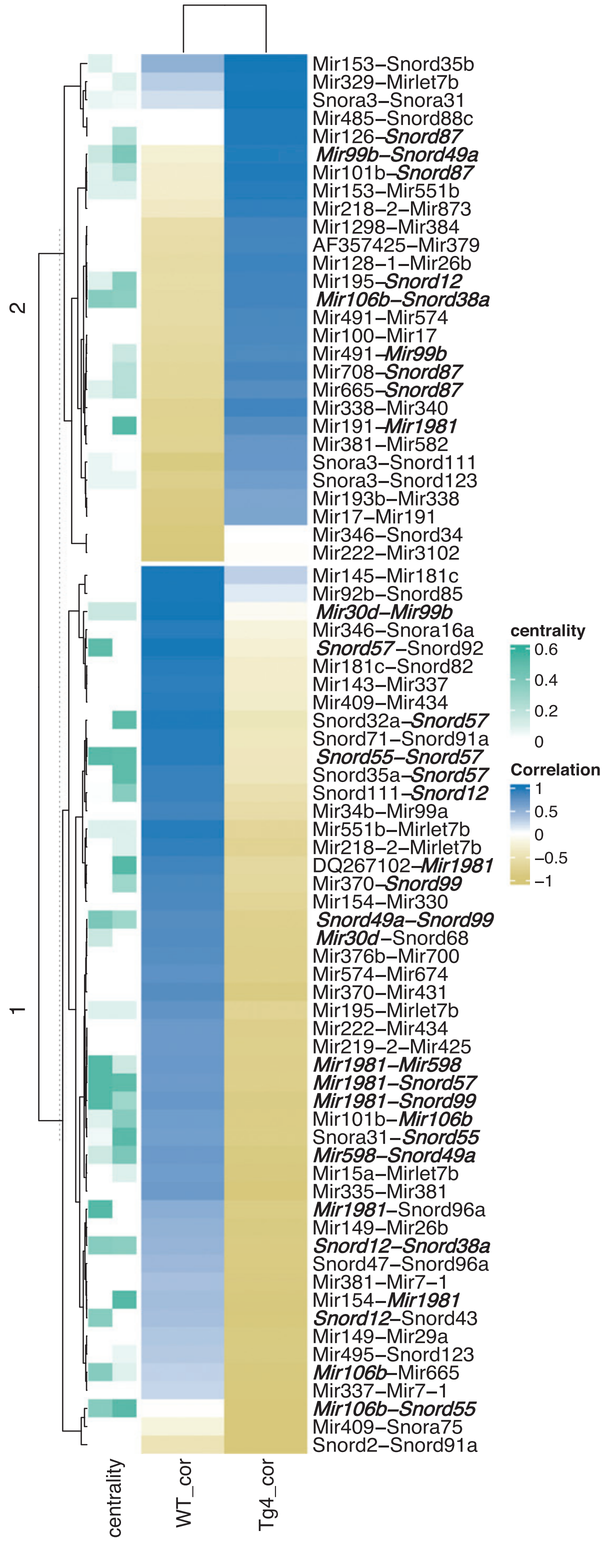


Fig. 2. Heatmap showing the 2 clusters of differentially correlated small RNAs. The yellow to blue color scale represents the degree of positive (blue) or negative (yellow) correlation between the pairs of small RNAs in WT ("WT_cor") and Tg4-42 ("Tg4_cor"). The green shaded annotation column on the left denotes the centralities of the small RNAs in the corresponding positions on the right. Small RNAs with centrality > 0.1 (see also Table 1) are shown in boldface italics.

Table 1
Small RNAs with centrality > 0.1

| Small RNA | Node centrality | AD-related References |
|---|---|---|
| | *snoRNAS* | |
| Snord55 | 0.5153273 | – |
| Snord57 | 0.5037283 | – |
| Snord49a | 0.3927092 | [53] |
| Snord12 | 0.3579122 | – |
| Snord38a | 0.3446562 | – |
| Snord99 | 0.2783761 | – |
| Snord87 | 0.2021541 | – |
| | *miRNAs* | |
| Mir1981 | 0.5189394 | – |
| Mir106b | 0.3598485 | [37, 38, 77] |
| Mir30d | 0.1439394 | [11] |
| Mir598 | 0.1363636 | [35] |
| Mir99b | 0.1439394 | [11, 36] |

0) and Mir 346 (centrality 0.003). These show a flip between positive and negative correlation values, possibly depending on the differently correlated small RNA in the respective combinations (Fig. 2): Mir222-Mir3102, and Mir346-Snord34 show a correlation of nigh on –0.96 in WT and a correlation value of -0.06 in Tg4-42 animals, while Mir222 is positively correlated (0.68) with Mir434 in WT, yet negatively (–0.83) in Tg4-42 samples and the same can be seen for the correlation between Mir346 and Snora16a (correlation 0.94 in WT and –0.22 in Tg4-42).

Finally, we observed Mir181c (centrality 0), which shows a positive correlation of 0.27 with Mir145 yet a negative correlation of –0.33 with Snord82, both in Tg4-42 hippocampi (Fig. 2).

To learn more about the possible functional implications of our findings, we performed functional enrichment analyses on the identified small RNAs with high centrality values. This led to an enrichment (-log10[padjust] = 2.24697) of the GO terms "sensory perception of sound" and "sensory perception of mechanical stimulus" for Mir106b, Mir30d, and Mir99b). Since the number of snoRNA annotations is still quite limited, we expanded this analysis to the predicted target genes of the small RNAs. After filtering for a prediction score higher than 0.7 (see Methods section), a total of 376 potential interactors for miRNAs and 21 for snoRNAs were left (Supplementary Tables 4 and 5). These were submitted to GO enrichment analysis and KEGG pathway enrichment analysis. For the analysed miRNA-interactors, numerous GO terms were found to be enriched (see Supplementary Table 6) and the top 20 are

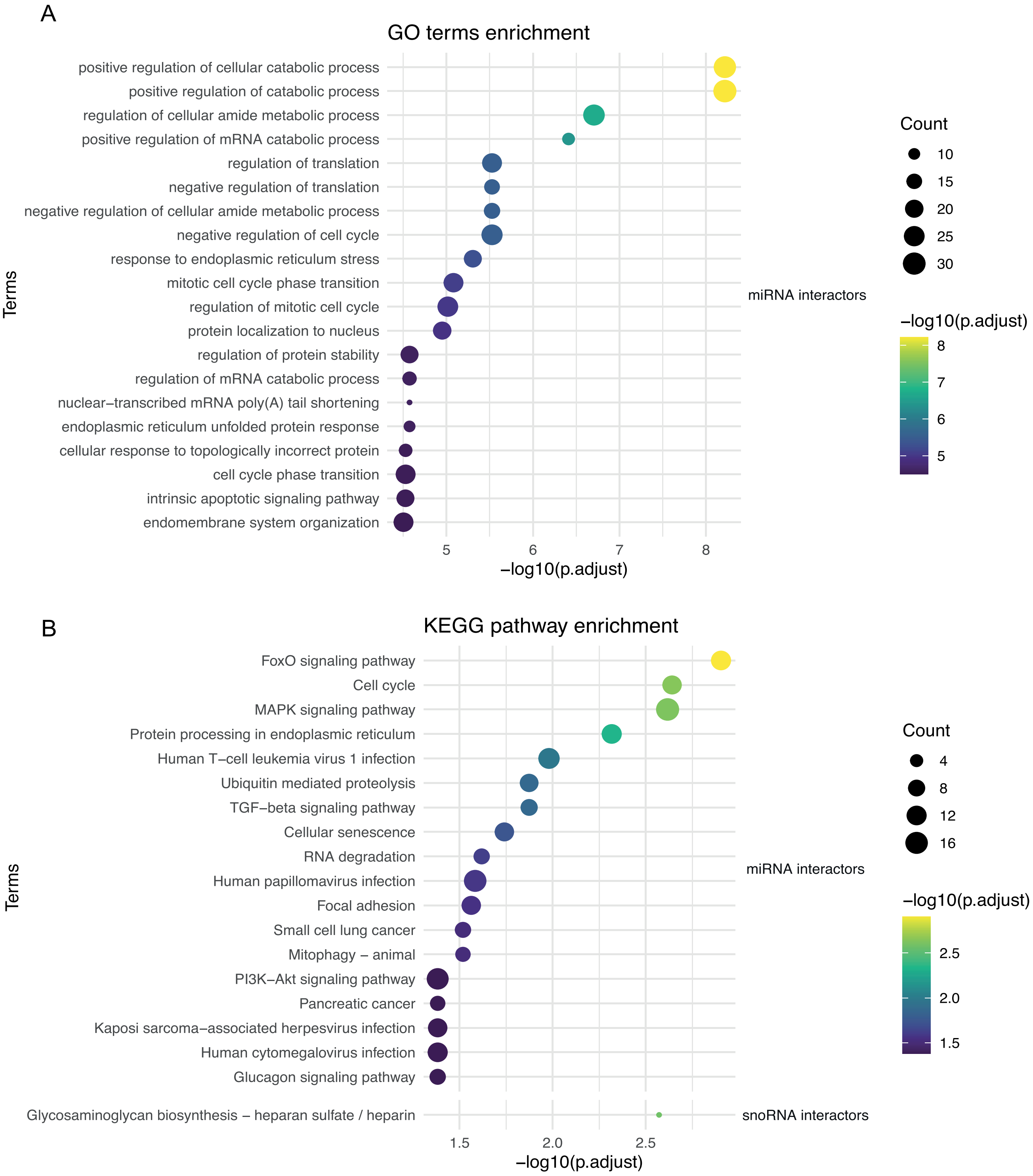


Fig. 3. A) GO enrichment results for interactors of miRNAs with centrality > 0.1. B) KEGG pathway enrichment results for interactors of miRNAs and snoRNAs with centrality > 0.1.

shown in Fig. 3A. For snoRNA-interactors, no significant enrichments were observed. In Fig. 3B, the results for KEGG pathway enrichment are presented. Among all terms, "FoxO signaling pathway" is the most prominently enriched pathway for miRNA-interactors while "Glycosaminoglycan biosynthesis -heparan sulfate / heparin" is the only term enriched for snoRNA-interactors.

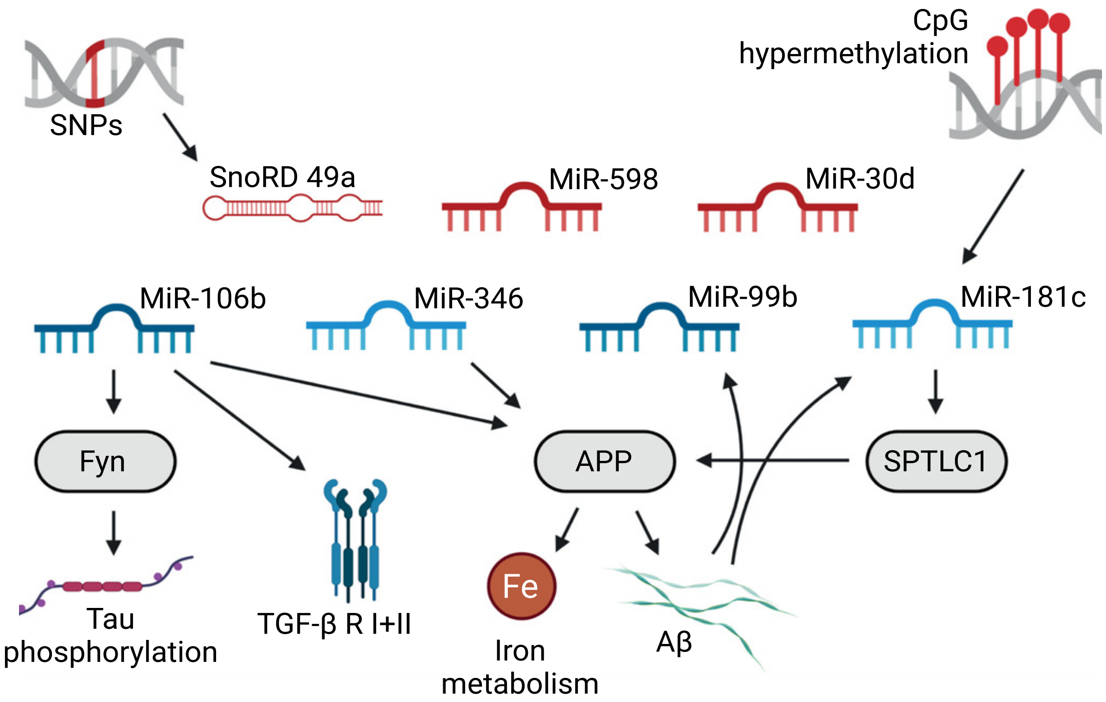


Fig. 4. Literature based snoRNA/mirRNA associations with AD-pathogenesis. Red RNAs are discussed as AD biomarkers, light blue RNAs have low centrality values. Created with BioRender.com.

## DISCUSSION

The Tg4-42 mouse model does not fully recapitulate the pathology of AD (reviewed in [19]). The transgenic mouse model Tg4-42 expresses exclusively normal wildtype human $A\beta_{4-42}$ sequence, predominantly in pyramidal neurons in the CA1 area of the hippocampus, associated with synaptic hyperexcitability, reactive micro- and astroglia, reduction in glucose metabolism is detected by $^{18}$F-PET/MRI, loss of degenerating CA1 pyramidal neurons, and loss of spatial reference memory [13–15]. However, no plaques and neurofibrillary tangles are observed.

Using the large numbers of differentially expressed small RNAs between WT and Tg4-42 hippocampi in a network analysis (Fig. 1), we observed pairs of small RNAs with a pronounced AD-specific shift in correlation. Such pairs comprised one or two partners with a high centrality in the network (Fig. 2, Table 1), suggesting a more prominent role in the genesis of the Tg4-42 AD-related phenotype.

We have previously shown that CA1 neurons in the hippocampus, not other cell types like interneurons, microglia or astroglia, express $A\beta_{4-42}$ [14]. Therefore, all observed changes in small RNA expression are a consequence of neuronal $A\beta_{4-42}$ activity.

A synopsis of what is already known about those small RNAs for which previous AD related findings are available (Table 1) is given in Fig. 4. Mir30b [11] and Mir598 [35] have previously been discussed as potential biomarkers for AD. Functional implications are known for Mir99b and Mir106b. For Mir99b, Ye et al. found that miR-99b-5p affects neuron survival by targeting mTOR and suggested a dynamic change of miR-99b-5p levels during Aβ-associated AD pathogenesis [36].

For Mir106b, Liu et al. report data that suggest that miR-106b has an inhibitory effect on $A\beta_{1-42}$-induced tau phosphorylation by targeting Fyn [37]. Moreover, it is surmised to be involved in the regulation of APP expression in neuronal cell lines [38] and was found to target TβR II and thus may possibly have an influence on TGF-β signaling [39], for which a key role in the etiology of AD has long been suggested [40–42].

Mir346 was shown to play a role in the upregulation of APP in the brain and to participate in maintaining APP-related regulation of Fe homeostasis [43], which is disrupted in AD (see [44]).

Mir181 seems to be downregulated by Aβ in hippocampal cultures [45] and has been found to show altered expression in AD patients or mice [2, 46] and its inhibition rescues memory deficits in a murine AD model [47].

Interestingly, Mir1981, which correlates significantly with five of the other small RNAs, four times highly positive in the hippocampi of WT and strongly negative in those of Tg4-42 animals (Fig. 2), has so far not been implicated in the pathogenesis of AD. It has, however, been found to belong to a group of non-canonical miRNAs that are highly expressed in the brain and for which a functional role in post-mitotic neurons has been suggested [48]. Mmu-Mir1981 was also found upregulated after alcohol treatment of TLR4 knockout mice in a study of alcohol-induced neuroinflammation [49] and the critical involvement of neuroinflammatory processes in the pathogenesis of AD has long been discussed [50–52].

The only snoRNA for which previous observations provide a link to AD is Snord49a, where genome-wide significant associations between risk variants and AD have been observed [53]. Some of the others have so far only surfaced in cancer related investigations. Snord87, for example, was observed to be upregulated in hepatocellular carcinoma [54] whereas Snord55, was found to be decreased in tumor-educated platelets from patients with non-small cell lung cancer [55] and Snord12 has been shown to be associated with the survival of patients with uveal melanoma and to be part of a four-snoRNA signature for survival prediction [56]. Still, the findings pertaining to Snord55 and Snord12 are in keeping with the general observation of an inverse association between cancer and AD (for review, see [57, 58]). Interestingly, piRNA-54265, a fragment of Snord57, is being discussed as a biomarker for colorectal cancer [59]. Snord57 was also found in our analysis and if one were to assume that full-length Snord57 has a protective effect, the observation that Snord57 fragments are elevated in cancer patients could also be in agreement with an inverse correlation between AD and cancer.

With respect to Snord38A or Snord99, our literature search did also not yield conclusive evidence as to an involvement in human brain disorders, yet also little else pointing to an involvement in other diseases, seems to be known.

Still, taken together, the majority (4 out of 5) of miRNAs we found, as well as one snoRNA have already been implicated in the molecular basis of AD (see Table 1, Fig. 4). This provides substantial support for our conclusion that the small RNAs with high centrality values in our analysis are indeed involved in cellular processes that play a role in the pathogenesis of AD. Moreover, it strongly underpins the likelihood of the involvement of the six snoRNAs and one miRNA that were previously not observed in connection with AD.

All observed changes in small RNA differences are a consequence of the A$\beta_{4-42}$ driven pathology in the Tg4-42 mouse model, representing downstream events. The intention of the current work was to identify small RNA changes in response to the toxic effect of A$\beta_{4-42}$ in mouse brain.

Our GO-analysis points to an involvement of Mir106b, Mir30d, and Mir99b in "sensory perception of sound" and "sensory perception of mechanical stimulus". This is interesting, as sensory impairments are part of the clinical spectrum of AD and are even being discussed as early disease markers (e.g., [60, 61]).

With respect to the putative interactors of the small RNAs we identified here, GO-analysis shows no enrichment for snoRNAs, which might be owed to the fact that still much less is known about snoRNAs as compared to miRNAs. For the miRNAs from our study the most prominent enriched GO terms by far were "positive regulation of cellular catabolic process" and "positive regulation of catabolic process". This is of note as various catabolic processes have been found to play a role in AD aetiology and progression, such as tau catabolism (e.g., [62]) or Aβ catabolism (e.g., [63, 64]).

Our KEGG pathway analysis revealed the "FoxO signaling pathway" as the most prominently enriched pathway for miRNA interactors. Forkhead box O (FoxO) transcription factors play a role in diverse biological processes, such as cell metabolism, cell proliferation, DNA repair, autophagy, the reaction to oxidative stress, etc. (e.g., [65, 66]). More importantly, however, there is a considerable body of evidence pointing to their crucial involvement in the aetiology of age-related diseases including AD (e.g., [67–71]).

Also, the "MAPK signaling pathway" features prominently among the results of this analysis. This is noteworthy, since MAPKs are being discussed as therapeutic targets for neurodegenerative disorders for quite some time (e.g., [72, 73]). Moreover, p38 MAPK has been found to be activated at early stages of AD [74] and it was recently shown that p38 MAPK-mediated loss of nuclear RNase III enzyme Drosha underlies Aβ-induced neuronal stress in AD [75].

We found only "Glycosaminoglycan biosynthesis -heparan sulfate/heparin" enriched for snoRNA interaction partners. Still, glycosaminoglycan is known to be involved in the pathogenesis of AD by contributing to the formation of amyloid fibrils [76] so that this finding supports a link between the snoRNAs we describe and AD.

Taken together, the results of the NGS-analysis of the hippocampi of Tg4-42 mice which we present here add further evidence for Aβ-driven down-stream molecular profiles. Of note, as the model does not generate amyloid plaques, we found evidence that soluble $A\beta_{4-42}$ expressed in the Tg4-42 mouse model triggers specific AD pathways known to be involved in tau-phosphorylation, TGF-β signaling and Aβ biology. This observation supports the current discussion in the AD field, that soluble amyloid peptides are instrumental in AD etiology and that tau pathology is triggered by Aβ toxicity. Moreover, we provide evidence for a functional role of snoRNAs in AD pathogenesis beyond the implication as biomarkers, which opens up new avenues of research in this context, for example the elucidation of how exactly the small RNAs resulting from our study are involved in the AD-typical neurological phenotype as presented by the Tg4-42 animals.

## ACKNOWLEDGMENTS

This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the e:Med research and funding concept (grant 01ZX1908A). We thank Robert Weissmann and Stefan Weiss for bioinformatics support in primary data analysis and Corina Jensen and Christian Sperling for excellent technical assistance.

Authors' disclosures available online (https://www.j-alz.com/manuscript-disclosures/22-0110r1).

## SUPPLEMENTARY MATERIAL

The supplementary material is available in the electronic version of this article: https://dx.doi.org/10.3233/JAD-220110.

## REFERENCES

[1] Wang M, Qin L, Tang B (2019) microRNAs in Alzheimer's disease. *Front Genet* **10**, 153.

[2] Silvestro S, Bramanti P, Mazzon E (2019) Role of miRNAs in Alzheimer's disease and possible fields of application. *Int J Mol Sci* **20**, 3979.

[3] Salta E, De Strooper B (2017) Noncoding RNAs in neurodegeneration. *Nat Rev Neurosci* **18**, 627-640.

[4] Salta E, De Strooper B (2012) Non-coding RNAs with essential roles in neurodegenerative disorders. *Lancet Neurol* **11**, 189-200.

[5] Angelucci F, Cechova K, Valis M, Kuca K, Zhang B, Hort J (2019) microRNAs in Alzheimer's disease: Diagnostic markers or therapeutic agents? *Front Pharmacol* **10**, 665.

[6] Kufel J, Grzechnik P (2019) Small nucleolar RNAs tell a different tale. *Trends Genet* **35**, 104-117.

[7] Samarsky DA, Fournier MJ, Singer RH, Bertrand E (1998) The snoRNA box C/D motif directs nucleolar targeting and also couples snoRNA synthesis and localization. *EMBO J* **17**, 3747-3757.

[8] Ganot P, Caizergues-Ferrer M, Kiss T (1997) The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev* **11**, 941-956.

[9] Bratkovič T, Božič J, Rogelj B (2019) Functional diversity of small nucleolar RNAs. *Nucleic Acids Res* **48**, 1627-1651.

[10] Watkins NJ, Bohnsack MT (2012) The box C/D and H/ACA snoRNPs: Key players in the modification, processing and the dynamic folding of ribosomal RNA. *Wiley Interdiscip Rev RNA* **3**, 397-414.

[11] Bouter Y, Kacprowski T, Rossler F, Jensen LR, Kuss AW, Bayer TA (2020) miRNA alterations elicit pathways involved in memory decline and synaptic function in the hippocampus of aged Tg4-42 mice. *Front Neurosci* **14**, 580524.

[12] Hinteregger B, Loeffler T, Flunkert S, Neddens J, Birner-Gruenberger R, Bayer TA, Madl T, Hutter-Paier B (2020) Transgene integration causes RARB downregulation in homozygous Tg4-42 mice. *Sci Rep* **10**, 6377.

[13] Dietrich K, Bouter Y, Muller M, Bayer TA (2018) Synaptic alterations in mouse models for Alzheimer disease-a special focus on N-truncated Abeta 4-42. *Molecules* **23**, 718.

[14] Bouter Y, Dietrich K, Wittnam JL, Rezaei-Ghaleh N, Pillot T, Papot-Couturier S, Lefebvre T, Sprenger F, Wirths O, Zweckstetter M, Bayer TA (2013) N-truncated amyloid beta (Abeta) 4-42 forms stable aggregates and induces acute and long-lasting behavioral deficits. *Acta Neuropathol* **126**, 189-205.

[15] Bouter C, Henniges P, Franke TN, Irwin C, Sahlmann CO, Sichler ME, Beindorff N, Bayer TA, Bouter Y (2018) (18)F-FDG-PET detects drastic changes in brain metabolism in the Tg4-42 model of Alzheimer's disease. *Front Aging Neurosci* **10**, 425.

[16] Wagner JM, Sichler ME, Schleicher EM, Franke TN, Irwin C, Low MJ, Beindorff N, Bouter C, Bayer TA, Bouter Y (2019) Analysis of motor function in the Tg4-42 mouse model of Alzheimer's disease. *Front Behav Neurosci* **13**, 107.

[17] Sichler ME, Low MJ, Schleicher EM, Bayer TA, Bouter Y (2019) Reduced acoustic startle response and prepulse inhibition in the Tg4-42 model of Alzheimer's disease. *J Alzheimers Dis Rep* **3**, 269-278.

[18] Hinteregger B, Loeffler T, Flunkert S, Neddens J, Bayer TA, Madl T, Hutter-Paier B (2021) Metabolic, phenotypic, and neuropathological characterization of the Tg4-42 mouse model for Alzheimer's disease. *J Alzheimers Dis* **80**, 1151-1168.

[19] Bayer TA (2021) N-truncated Abeta starting at position four-biochemical features, preclinical models, and potential as drug target in Alzheimer's disease. *Front Aging Neurosci* **13**, 710579.

[20] Bayer TA, Wirths O (2014) Focusing the amyloid cascade hypothesis on N-truncated Abeta peptides as drug targets against Alzheimer's disease. *Acta Neuropathol* **127**, 787-801.

[21] Masters CL, Simms G, Weinman NA, Multhaup G, McDonald BL, Beyreuther K (1985) Amyloid plaque core protein in Alzheimer disease and Down syndrome. *Proc Natl Acad Sci U S A* **82**, 4245-4249.

[22] Portelius E, Bogdanovic N, Gustavsson MK, Volkmann I, Brinkmalm G, Zetterberg H, Winblad B, Blennow K (2010) Mass spectrometric characterization of brain amyloid beta isoform signatures in familial and sporadic Alzheimer's disease. *Acta Neuropathol* **120**, 185-193.

[23] Oakley H, Cole SL, Logan S, Maus E, Shao P, Craft J, Guillozet-Bongaarts A, Ohno M, Disterhoft J, Van Eldik L, Berry R, Vassar R (2006) Intraneuronal beta-amyloid aggregates, neurodegeneration, and neuron loss in transgenic mice with five familial Alzheimer's disease mutations: Potential factors in amyloid plaque formation. *J Neurosci* **26**, 10129-10140.

[24] Bouter Y, Kacprowski T, Weissmann R, Dietrich K, Borgers H, Brauss A, Sperling C, Wirths O, Albrecht M, Jensen LR, Kuss AW, Bayer TA (2014) Deciphering the molecular profile of plaques, memory decline and neuron loss in two mouse models for Alzheimer's disease by deep sequencing. *Front Aging Neurosci* **6**, 75.

[25] Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550.

[26] McKenzie AT, Katsyv I, Song WM, Wang M, Zhang B (2016) DGCA: A comprehensive R package for differential gene correlation analysis. *BMC Syst Biol* **10**, 106.

[27] Gu Z, Eils R, Schlesner M (2016) Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics* **32**, 2847-2849.

[28] Hartigan JA, Wong MA (1979) Algorithm AS 136: A K-means clustering algorithm. *J R Stat Soc Ser C Appl Stat* **28**, 100-108.

[29] Wu T, Hu E, Xu S, Chen M, Guo P, Dai Z, Feng T, Zhou L, Tang W, Zhan L, Fu X, Liu S, Bo X, Yu G (2021) clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**, 100141.

[30] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene Ontology: Tool for the unification of biology. *Nat Genet* **25**, 25-29.

[31] Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2015) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-D462.

[32] Kang J, Tang Q, He J, Li L, Yang N, Yu S, Wang M, Zhang Y, Lin J, Cui T, Hu Y, Tan P, Cheng J, Zheng H, Wang D, Su X, Chen W, Huang Y (2021) RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Res* **50**, D326-D332.

[33] Chen Y, Wang X (2019) miRDB: An online database for prediction of functional microRNA targets. *Nucleic Acids Res* **48**, D127-D131.

[34] Bouchard-Bourelle P, Desjardins-Henri C, Mathurin-St-Pierre D, Deschamps-Francoeur G, Fafard-Couture É, Garant J-M, Elela SA, Scott MS (2019) snoDB: An interactive database of human snoRNA sequences, abundance and interactions. *Nucleic Acids Res* **48**, D220-D225.

[35] Riancho J, Vázquez-Higuera JL, Pozueta A, Lage C, Kazimierczak M, Bravo M, Calero M, Gonalezález A, Rodríguez E, Lleó A, Sánchez-Juan P (2017) MicroRNA profile in patients with Alzheimer's disease: Analysis of miR-9-5p and miR-598 in raw and exosome enriched cerebrospinal fluid samples. *J Alzheimers Dis* **57**, 483-491.

[36] Ye X, Luo H, Chen Y, Wu Q, Xiong Y, Zhu J, Diao Y, Wu Z, Miao J, Wan J (2015) MicroRNAs 99b-5p/100-5p regulated by endoplasmic reticulum stress are involved in Abeta-induced pathologies. *Front Aging Neurosci* **7**, 210-210.

[37] Liu W, Zhao J, Lu G (2016) miR-106b inhibits tau phosphorylation at Tyr18 by targeting Fyn in a model of Alzheimer's disease. *Biochem Bioph Res* **478**, 852-857.

[38] Hébert SS, Horré K, Nicolaï L, Bergmans B, Papadopoulou AS, Delacourte A, De Strooper B (2009) MicroRNA regulation of Alzheimer's Amyloid precursor protein expression. *Neurobiol Dis* **33**, 422-428.

[39] Wang H, Liu J, Zong Y, Xu Y, Deng W, Zhu H, Liu Y, Ma C, Huang L, Zhang L, Qin C (2010) miR-106b aberrantly expressed in a double transgenic mouse model for Alzheimer's disease targets TGF-β type II receptor. *Brain Res* **1357**, 166-174.

[40] Tesseur I, Zou K, Esposito L, Bard F, Berber E, Can JV, Lin AH, Crews L, Tremblay P, Mathews P, Mucke L, Masliah E, Wyss-Coray T (2006) Deficiency in neuronal TGF-β signaling promotes neurodegeneration and Alzheimer's pathology. *J Clin Invest* **16**, 3060-3069.

[41] Fessel J (2019) Ineffective levels of transforming growth factors and their receptor account for old age being a risk factor for Alzheimer's disease. *Alzheimers Dement* **5**, 899-905.

[42] Chao CC, Hu S, Frey WH, 2nd, Ala TA, Tourtellotte WW, Peterson PK (1994) Transforming growth factor beta in Alzheimer's disease. *Clin Diagn Lab Immunol* **1**, 109-110.

[43] Long JM, Maloney B, Rogers JT, Lahiri DK (2019) Novel upregulation of amyloid-beta precursor protein (APP) by microRNA-346 via targeting of APP mRNA 5'-untranslated region: Implications in Alzheimer's disease. *Mol Psychiatr* **24**, 345-363.

[44] Liu J-L, Fan Y-G, Yang Z-S, Wang Z-Y, Guo C (2018) Iron and Alzheimer's disease: From pathogenesis to therapeutic implications. *Front Neurosci* **12**, 632-632.

[45] Schonrock N, Ke YD, Humphreys D, Staufenbiel M, Ittner LM, Preiss T, Götz J (2010) Neuronal MicroRNA deregulation in response to Alzheimer's disease amyloid-β. *PLoS One* **5**, e11070.

[46] Indrieri A, Carrella S, Carotenuto P, Banfi S, Franco B (2020) The pervasive role of the miR-181 family in development, neurodegeneration, and cancer. *Int J Mol Sci* **21**, 2092.

[47] Rodriguez-Ortiz CJ, Prieto GA, Martini AC, Forner S, Trujillo-Estrada L, LaFerla FM, Baglietto-Vargas D, Cotman CW, Kitazawa M (2020) miR-181a negatively modulates synaptic plasticity in hippocampal cultures and its inhibition rescues memory deficits in a mouse model of Alzheimer's disease. *Aging Cell* **19**, e13118.

[48] Babiarz JE, Hsu R, Melton C, Thomas M, Ullian EM, Blelloch R (2011) A role for noncanonical microRNAs in the mammalian brain revealed by phenotypic differences in Dgcr8 versus Dicer1 knockouts and small RNA sequencing. *RNA* **17**, 1489-1501.

[49] Ureña-Peralta JR, Alfonso-Loeches S, Cuesta-Diaz CM, García-García F, Guerri C (2018) Deep sequencing and miRNA profiles in alcohol-induced neuroinflammation and the TLR4 response in mice cerebral cortex. *Sci Rep* **8**, 15913.

[50] McGeer PL, McGeer EG, Yasojima K (2000) Alzheimer disease and neuroinflammation. In *Advances in Demen-*

*tia Research*, eds. Jellinger K, Schmidt R, Windisch M, Springer Vienna, Vienna, pp. 53-57.

[51] Heneka MT, Carson MJ, Khoury JE, Landreth GE, Brosseron F, Feinstein DL, Jacobs AH, Wyss-Coray T, Vitorica J, Ransohoff RM, Herrup K, Frautschy SA, Finsen B, Brown GC, Verkhratsky A, Yamanaka K, Koistinaho J, Latz E, Halle A, Petzold GC, Town T, Morgan D, Shinohara ML, Perry VH, Holmes C, Bazan NG, Brooks DJ, Hunot S, Joseph B, Deigendesch N, Garaschuk O, Boddeke E, Dinarello CA, Breitner JC, Cole GM, Golenbock DT, Kummer MP (2015) Neuroinflammation in Alzheimer's disease. *Lancet Neurol* **14**, 388-405.

[52] Calsolaro V, Edison P (2016) Neuroinflammation in Alzheimer's disease: Current evidence and future directions. *Alzheimers Dement* **12**, 719-732.

[53] Guo X, Qiu W, Garcia-Milian R, Lin X, Zhang Y, Cao Y, Tan Y, Wang Z, Shi J, Wang J, Liu D, Song L, Xu Y, Wang X, Liu N, Sun T, Zheng J, Luo J, Zhang H, Xu J, Kang L, Ma C, Wang K, Luo X (2017) Genome-wide significant, replicated and functional risk variants for Alzheimer's disease. *J Neural Transm* **124**, 1455-1471.

[54] Chang L, Yuan Y, Li C, Guo T, Qi H, Xiao Y, Dong X, Liu Z, Liu Q (2016) Upregulation of SNHG6 regulates ZEB1 expression by competitively binding miR-101-3p and interacting with UPF1 in hepatocellular carcinoma. *Cancer Lett* **383**, 183-194.

[55] Dong X, Song X, Ding S, Yu M, Shang X, Wang K, Chang M, Xie L, Song X (2021) Tumor-educated platelet SNORD55 as a potential biomarker for the early diagnosis of non-small cell lung cancer. *Thorac Cancer* **12**, 659-666.

[56] Yi Q, Zou WJ (2019) A novel four-snoRNA signature for predicting the survival of patients with uveal melanoma. *Mol Med Rep* **19**, 1294-1301.

[57] Lanni C, Masi M, Racchi M, Govoni S (2021) Cancer and Alzheimer's disease inverse relationship: An age-associated diverging derailment of shared pathways. *Mol Psychiatr* **26**, 280-295.

[58] Majd S, Power J, Majd Z (2019) Alzheimer's disease and cancer: When two monsters cannot be together. *Front Neurosci* **13**, 155.

[59] Tosar JP, Garcia-Silva MR, Cayota A (2021) Circulating SNORD57 rather than piR-54265 is a promising biomarker for colorectal cancer: Common pitfalls in the study of somatic piRNAs in cancer. *RNA* **27**, 403-410.

[60] Murphy C (2019) Olfactory and other sensory impairments in Alzheimer disease. *Nat Rev Neurol* **15**, 11-24.

[61] Albers MW, Gilmore GC, Kaye J, Murphy C, Wingfield A, Bennett DA, Boxer AL, Buchman AS, Cruickshanks KJ, Devanand DP, Duffy CJ, Gall CM, Gates GA, Granholm A-C, Hensch T, Holtzer R, Hyman BT, Lin FR, McKee AC, Morris JC, Petersen RC, Silbert LC, Struble RG, Trojanowski JQ, Verghese J, Wilson DA, Xu S, Zhang LI (2015) At the interface of sensory and motor dysfunctions and Alzheimer's disease. *Alzheimers Dement* **11**, 70-98.

[62] Cooper JM, Lathuiliere A, Migliorini M, Arai AL, Wani MM, Dujardin S, Muratoglu SC, Hyman BT, Strickland DK (2021) Regulation of tau internalization, degradation, and seeding by LRP1 reveals multiple pathways for tau catabolism. *J Biol Chem* **296**, 100715.

[63] Aloi MS, Prater KE, Sopher B, Davidson S, Jayadev S, Garden GA (2021) The pro-inflammatory microRNA miR-155 influences fibrillar β-Amyloid1-42 catabolism by microglia. *Glia* **69**, 1736-1748.

[64] Qian C, Yang C, Lu M, Bao J, Shen H, Deng B, Li S, Li W, Zhang M, Cao C (2021) Activating AhR alleviates cognitive deficits of Alzheimer's disease model mice by upregulating endogenous Aβ catabolic enzyme Neprilysin. *Theranostics* **11**, 8797-8812.

[65] Gui T, Burgering BMT (2021) FOXOs: Masters of the equilibrium. *FEBS J*, doi: 10.1111/febs.16221.

[66] Krafczyk N, Klotz L-O (2022) FOXO transcription factors in antioxidant defense. *IUBMB Life* **74**, 53-61.

[67] Manolopoulos KN, Klotz LO, Korsten P, Bornstein SR, Barthel A (2010) Linking Alzheimer's disease to insulin resistance: The FoxO response to oxidative stress. *Mol Psychiatr* **15**, 1046-1052.

[68] Du S, Zheng H (2021) Role of FoxO transcription factors in aging and age-related metabolic and neurodegenerative diseases. *Cell Biosci* **11**, 188-188.

[69] Sohn H-Y, Kim S-I, Park J-Y, Park S-H, Koh YH, Kim J, Jo C (2021) ApoE4 attenuates autophagy via FoxO3a repression in the brain. *Sci Rep* **11**, 17604.

[70] Maiese K (2021) Targeting the core of neurodegeneration: FoxO, mTOR, and SIRT1. *Neural Regen Res* **16**, 448-455.

[71] Maiese K (2021) Neurodegeneration, memory loss, and dementia: The impact of biological clocks and circadian rhythm. *Front Biosci (Landmark Ed)* **26**, 614-627.

[72] Harper SJ, Wilkie N (2003) MAPKs: New targets for neurodegeneration. *Expert Opin Ther Targets* **7**, 187-200.

[73] D'Mello SR (2021) When good kinases go rogue: GSK3, p38 MAPK and CDKs as therapeutic targets for Alzheimer's and Huntington's disease. *Int J Mol Sci* **22**, 5911.

[74] Sun A, Liu M, Nguyen XV, Bing G (2003) P38 MAP kinase is activated at early stages in Alzheimer's disease brain. *Exp Neurol* **183**, 394-405.

[75] Xu H, Liu X, Li W, Xi Y, Su P, Meng B, Shao X, Tang B, Yang Q, Mao Z (2021) p38 MAPK-mediated loss of nuclear RNase III enzyme Drosha underlies amyloid beta-induced neuronal stress in Alzheimer's disease. *Aging Cell* **20**, e13434.

[76] Ariga T, Miyatake T, Yu RK (2010) Role of proteoglycans and glycosaminoglycans in the pathogenesis of Alzheimer's disease and related disorders: Amyloidogenesis and therapeutic strategies—A review. *J Neurosci Res* **88**, 2303-2315.

[77] Cheng L, Doecke JD, Sharples RA, Villemagne VL, Fowler CJ, Rembach A, Martins RN, Rowe CC, Macaulay SL, Masters CL, Hill AF, Australian Imaging, Biomarkers and Lifestyle (AIBL) Research Group (2015) Prognostic serum miRNA biomarkers associated with Alzheimer's disease shows concordance with neuropsychological and neuroimaging assessment. *Mol Psychiatry* **20**, 1188-1196.

## 4.2 Systematic analysis of alternative splicing in time course data using Spycone

### Citation

The article titled "Systematic analysis of alternative splicing in time course data using Spycone" has been published online at Bioinformatics on 29 December 2022.

### Full citation:

Chit Tong Lio, Gordon Grabert, Zakaria Louadi, Amit Fenn, Jan Baumbach, Tim Kacprowski, Markus List, Olga Tsoy, Systematic analysis of alternative splicing in time course data using Spycone, Bioinformatics, Volume 39, Issue 1, January 2023, btac846, https://doi.org/10.1093/bioinformatics/btac846

### Summary

Alternative splicing is crucial for protein diversity and function in diseases and developments. Most of these biological processes are dynamic, and crucial changes occur temporarily. Time series data with multiple time points is often employed to detect these changes. This paper addressed the challenges in analyzing alternative splicing (AS) across different time points during biological processes like disease progression. Traditional methods often overlook the dynamic nature of AS, and no method provides a framework for a systematic approach to isoform switch analysis for time course data.

I introduce Spycone, a novel framework for systematic time course transcriptomics data analysis, focusing on isoform switches (IS) relevant to biological functions. Spycone has implemented a novel algorithm to identify significant IS events. This algorithm highlights a new metric for isoform switch isoforms filtering: the event importance metric. This metric indicates the expression level of the switching isoforms since higher expressed isoforms will have more biological impact. For downstream analysis, Spycone utilizes total isoform usage to cluster genes with similar AS patterns, performs splicing-aware functional analysis using NEASE and active modules identification using DOMINO. In addition, the detected IS events can be applied to splicing factors analysis to extract potential regulators. I used a new simulation method to generate synthetic time course data for evaluation based on the hidden Markov model. The simulated data contains isoform-switching events that act as ground truth. Spycone outperforms the only competitor, TSIS, regarding precision and recall.

Spycone's utility is also demonstrated through real-world data, particularly its application to SARS-CoV-2 infection data. First, Spycone found more isoform-switching events with high event importance than TSIS. With clustering analysis on the level of total isoform usage, Spycone found four clusters grouped with genes with similar AS patterns, each represented by a time series trend. After applying NEASE on these clusters, we found enrichment of SARS-Cov-2 associated pathways, particularly Toll-like receptors (TLR) 7/8 cascades and some immune-related pathways. Some pathways were not previously associated with SARS-Cov-2 infection, such as kinesins, signaling by NTRK and degradation of AXIN. Spycone then utilizes DOMINO to extract network modules for each cluster. These modules are enriched by isoform-switching genes, and the resulting module also shows affected domains by the IS event. This analysis allows the extraction of interesting gene candidates whose functionality is affected by the IS event. For

example, I found three kinases and a protein chaperone affected by IS events, particularly HSP90AA1, which is associated with the endoplasmic reticulum stress caused by the infection.
The study highlights Spycone's superior precision and recall compared to existing tools, suggesting its potential to reveal intricate mechanisms of disease and development.

### Availability

The Spycone package is available as a PyPI package. The source code of Spycone is available under the GPLv3 license at https://github.com/yollct/spycone and the documentation at https://spycone.readthedocs.io/en/latest/.

### Contribution

I had the leading role in algorithmic framework design, wrote all the code for the algorithm and experiments, obtained and processed the data and ran and adjusted competing methods for a fair comparison. I wrote the first draft of the manuscript and generated all the figures. All project-related activities were supervised by Prof. T. Kacprowski, Prof. J. Baumbach, and Dr. M. List. Prof. T. Kacprowski, Dr. M. List and Dr. O. Tsoy revised the manuscript. All authors provided their feedback on the final manuscript.

### Rights and permissions

The original article is embedded with permission of Oxford Academic Press. All rights belong to Oxford Academic Press.

### Additional supplementary material

Supplementary data are available at Bioinformatics online https://academic.oup.com/bioinformatics/article/39/1/btac846

*Bioinformatics*, 39(1), 2023, btac846
https://doi.org/10.1093/bioinformatics/btac846
Advance Access Publication Date: 29 December 2022
Original Paper

Systems biology

# Systematic analysis of alternative splicing in time course data using Spycone

**Chit Tong Lio [1,2], Gordon Grabert[3,4], Zakaria Louadi[1,2], Amit Fenn[1,2], Jan Baumbach [1,5], Tim Kacprowski [3,4], Markus List [2,†] and Olga Tsoy[1,*,†]**

[1]Institute for Computational Systems Biology, University of Hamburg, Notkestrasse 9, Hamburg 22607, Germany, [2]Chair of Experimental Bioinformatics, TUM School of Life Sciences, Technical University of Munich, Freising 85354, Germany, [3]Division Data Science in Biomedicine, Peter L. Reichertz Institute for Medical Informatics of Technische Universität Braunschweig and Hannover Medical School, Braunschweig 38106, Germany, [4]Braunschweig Integrated Centre of Systems Biology (BRICS), TU Braunschweig, Braunschweig 38106, Germany and [5]Institute of Mathematics and Computer Science, University of Southern Denmark, Odense 5000, Denmark

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the last two authors should be regarded as Joint Last Authors.

Associate Editor: Valentina Boeva

Received on May 30, 2022; revised on November 16, 2022; editorial decision on December 26, 2022; accepted on December 28, 2022

## Abstract

**Motivation:** During disease progression or organism development, alternative splicing may lead to isoform switches that demonstrate similar temporal patterns and reflect the alternative splicing co-regulation of such genes. Tools for dynamic process analysis usually neglect alternative splicing.

**Results:** Here, we propose Spycone, a splicing-aware framework for time course data analysis. Spycone exploits a novel IS detection algorithm and offers downstream analysis such as network and gene set enrichment. We demonstrate the performance of Spycone using simulated and real-world data of SARS-CoV-2 infection.

**Availability and implementation:** The Spycone package is available as a PyPI package. The source code of Spycone is available under the GPLv3 license at https://github.com/yollct/spycone and the documentation at https://spycone.readthedocs.io/en/latest/.

**Contact:** olga.tsoy@uni-hamburg.de

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

## 1 Introduction

Changes in alternative splicing (AS) lead to a differential abundance of gene isoforms between experimental conditions or time points. If the relative abundance of two isoforms of a gene changes between two conditions or time points, this behavior is called isoform switching (IS). While differential isoform expression focus on the change in the expression value of one isoform, IS detects switches of predominantly expressed isoforms between conditions. A change of the predominant isoform appears as an intersection in time course data. However, existing methods for time course change points detection are applied to detect abrupt change between states while IS events are usually slow and gradual changes of isoform expression (Aminikhanghahi and Cook, 2017). IS has a functional impact on the gene when the two switching isoforms perform different functions or when they have different interaction partners. Vitting-Seerup and Sandelin (2017) showed that IS changes the functions of 19% ($N = 2352$) of genes with multiple isoforms in cancer, most of them leading to a protein domain loss. In cardiovascular disease, the IS of Titin causes clinical symptoms of dilated cardiomyopathy (Makarenko et al., 2004). Therefore, the detection and functional interpretation of IS events is a promising strategy to reveal the mechanism of disease development.

However, the above examples refer to molecular snapshots of dynamic processes. In order to study such dynamic processes, like disease progression, we need time course data. By identifying groups of genes with similar temporal expression or AS/IS patterns, we can dissect the disease progression into mechanistic details. A study of mouse retinal development has shown that genes with similar temporal exon usage patterns shared similar biological functions and cell type specificity (Wan *et al.*, 2011). However, existing tools for AS analysis mostly focus on a single condition or two conditions from snapshot experiments. Tools developed for time course data analysis, e.g. TiCoNE (Wiwie et al., 2019), moanin (Varoquaux and Purdom, 2020), TimesVector-web (Jang *et al.*, 2021) focus on gene expression level and neglect splicing. Thus, systematic time course AS analysis is usually done manually. Common approaches are semi-automatic clustering of temporal patterns of percent-spliced-in (PSI) value (Trincado *et al.* 2017; Xing *et al.* 2020) or differential splicing analysis between pairs of time points (Hooper *et al.* 2020). PSI values indicate the fraction of transcripts carrying an AS event and thus do not directly reflect isoform switches which are crucial for interpreting functional consequences of AS. Iso-MaSigPro uses a generalized linear model to detect differential expression changes along time courses between two experimental groups (Nueda *et al.*, 2018). However, Iso-MaSigPro is limited in time series data with two conditions and it does not provide information like switching points. TSIS, the only available tool to perform AS time course analysis in one condition, detects IS events whose effect lasts across several time points (Guo *et al.*, 2017). However, TSIS treats all IS events similarly, independent of their expression level. As a result, TSIS emphasizes isoforms with low expression while isoforms with comparably high expression levels are expected to be more involved in biological processes.

We introduce Spycone, a splicing-aware framework for systematic time course transcriptomics data analysis. It employs a novel IS detection method that prioritizes isoform switches between highly expressed isoforms over those with minor expression levels, thus focusing on biologically relevant changes rather than transcriptional noise. Spycone operates on both gene and isoform levels. For isoform-level data, the total isoform usage is quantified across time points. We have incorporated clustering methods for grouping genes and isoforms with similar time course patterns, as well as network and gene set enrichment methods for the functional interpretation of clusters. The IS genes within the same clusters are expected to interact cooperatively with other functionally related genes. Thus, we hypothesize that disease mechanisms or developmental changes can be identified with network and functional enrichment methods. We compare the performance of Spycone and TSIS on a simulated and real-world dataset. On the latter, we demonstrate how Spycone identifies network modules that are potentially affected by alternatively spliced genes during SARS-CoV-2 infection.

# 2 Materials and methods

## 2.1 Data preprocessing

We demonstrated the performance of Spycone on RNA-seq data from SARS-CoV-2 infected human lung cells (Calu-3) with eight time points and four replicates for each time point (de la Fuente *et al.*, 2020).

For the SARS-CoV-2 dataset, we used Trimmomatic v0.39 (Bolger *et al.*, 2014) to remove Illumina adapter sequences and low-quality bases (Phred score < 30) followed by Salmon v1.5.1 (Patro *et al.*, 2017) for isoform quantification with a mapping-based model, the human genome version 38 and an Ensembl genome annotation version 104.

## 2.2 Protein–protein interaction network and domain–domain interaction

A protein–protein interaction (PPI) network is obtained from BioGRID (v.4.4.208) (Oughtred *et al.*, 2021) and a domain–domain interaction network from 3did (v2019_01) (Mosca *et al.*, 2014). The edges of the PPI network are weighted according to the number of interactions found between the domains of the protein (nodes of PPI), given by the domain–domain interaction. Weighting PPIs with domain-based information can result in a functionally more interpretable network in diseases and pathways (Shim and Lee, 2016).

## 2.3 Simulation

We used the SARS-CoV-2 dataset described above as a reference for setting the parameters of a negative binomial distribution of gene expression counts, as well as the parameters of the Poisson distribution of the number of isoforms for each gene.

### 2.3.1 First-order Markov chain

A first-order Markov chain is used for the simulation of the gene states at each time point. In the simplest form, we defined two gene states: switched or unswitched. Change of the states along the time course depends on the transition probabilities.

We used a Dirichlet distribution to simulate relative abundance for each isoform of a gene. The relative abundance of an isoform is the ratio of the isoform expression to the total gene expression. The outcome of the Dirichlet distribution is $k$-dimensional vectors $x$ with real numbers between 0 and 1 such that the sum of the elements in $x$ is 1. This is suitable to simulate probability distribution of $k$ categories. The Dirichlet distribution is defined as:

$$f(M_t|\alpha_0, \alpha_1, \ldots, \alpha_k) = \frac{1}{\mathrm{beta}(\alpha)} \prod_{i=1}^{k} M_i^{\alpha_i - 1}, \tag{1}$$

$$\mathrm{beta}(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i)}, \tag{2}$$

where the parameter is a $k$-dimensional vector governing the distribution of the probabilities. In our case, $k$ equals the number of isoforms in a gene, where each isoform will be assigned an $i$ value. The higher $i$, the higher the probability of the isoform $i$.

In Model 1, where we assumed that switching isoforms are highly expressed, the $\alpha$ for switching isoforms are $\alpha = \{1, 2, \ldots, s\}{*}10$, $s$ is the number of switching isoforms, while $\alpha$ for the remaining isoforms are 1. To introduce switching events, the isoform probabilities of two highly expressed isoforms are swapped. For instance, if the isoform probabilities of the unswitched state for gene $g$ with five isoforms are $\{0.03, 0.07, 0.1, 0.3, 0.5\}$. Then, the isoform probabilities for the switched state is $\{0.03, 0.07, 0.1, 0.5, 0.3\}$.

In Model 2, where we assumed that isoforms with abundance higher than 0.3 have equal chances to switch, the vector is $\alpha = \{1, 2, \ldots, k\}{*}10$, $k$ is the number of all isoforms. To introduce switching events, the probabilities of two random isoforms will be swapped.

After we simulated abundances for each isoform, we multiplied it to a gene expression mean selected based on real-life dataset to obtain the transcript expression mean ($\mu_{\mathrm{i}}$). The gene expression means are randomly picked among the genes with the same number of isoforms from the real-world dataset.

We simulated time course data with 10 time points and 3 replicates using 10 000 genes. The transcript expression of replicates is sampled from normal distribution with a given transcript mean ($\mu_i$), and the variance is sampled from a gamma distribution as the following:

$$\theta_i \sim \mathrm{gamma}\left(a = \frac{\mu_i + \mathrm{noise}}{\mu_i * \mathrm{noise}}\right) \tag{3}$$

$$\mathrm{Count}t_i \sim \mathrm{normal}(\mu_i, \theta_i). \tag{4}$$

In order to simulate the differences generated for each individual experiment in real life, we tested on noise levels 1, 5 and 10. This setting can ensure isoforms with higher abundance will have a higher variance compared to those with lower abundance. The simulated dataset can be downloaded from doi: 10.5281/zenodo.7228475. The code for generating the benchmarking figures is stored in https://github.com/yollct/spycone_benchmark.

The data were analyzed with Spycone's detect_isoform_switch function and TSIS's iso.switch function. TSIS was further tested in two modes—TSIS and major_TSIS—major_TSIS uses the max.ratio = TRUE parameter.

## 2.4 Detection of isoform switch events

### 2.4.1 Spycone

The first step of IS detection is to filter out transcripts that have an average transcripts per million (TPM) <1 over all time points.

Spycone then detects IS events based on the relative abundance of the isoforms. The IS events are defined with the following metrics:

*2.4.1.1 Switching points.* Switch points refer to the points where two time courses intersect in at least 60% of the replicates. For every pair of isoforms in a gene, Spycone detects all possible switch points for further analysis. For a dataset that has only one replicate, Spycone checks the intersection between isoform pairs in one replicate.

*2.4.1.2 Switching probability.* As TSIS, Spycone calculates a switching probability for each IS event. A switching probability is the average of the ratio of samples where the relative abundance $I$ of isoform $i$ is higher than isoform $j$ before switch ($T_1$), and vice versa, the ratio of samples where the relative abundance $I$ isoform $i$ is lower than of isoform $j$ after switch ($T_2$). If two isoforms switched between time interval $T_1$and $T_2$ the switching probability between isoform $i$ and isoform $j$ is:

$$P(\text{switch}) = [P_{T_1}(\mathrm{I}_{i,t} > \mathrm{I}_{j,t}) + P_{T_2}(\mathrm{I}_{i,t} < \mathrm{I}_{j,t})]/2, \tag{5}$$

where $P$ denotes the frequency of the respective condition between relative abundance ($I$) of isoform $i$ and $j$ at each time point $t$ within the time interval $T$.

*2.4.1.3 Significance of switch points.* If replicates are available, Spycone calculates the significance of a switch point by performing a two-sided Mann–Whitney $U$-test between relative abundance before and after the switch point similar to TSIS. For a dataset that has only one replicate, a permutation test is performed, where the time points within a time course are permuted. An empirical $P$-value is calculated to indicate the probability for the two switching isoforms to have a higher dissimilarity coefficient and higher difference of relative abundance before and after switch. Since the goal here is to select genes that have significant IS, Spycone takes the best switch point for further analysis with the smallest $P$-value. Other significant switch points will be reported as part of the result for users to investigate.

*2.4.1.4 Difference of relative abundance.* To quantify the magnitude of changes during IS, Spycone calculates the average difference of relative abundance before and after a switch point. If replicates are available, Spycone calculates the average change of relative abundance. We selected a cutoff of 0.1, where the changes in the relative abundance accounts for at least 10% of the total gene expression. Difference of relative abundance $I$ between switching isoforms $i$ and $j$ is defined as:

$$\mathrm{Diff}_{i,j,s} = \left[\sum_{r=1}^{R}(I^r_{i,s+1} - I^r_{i,s})/R + \sum_{r=1}^{R}(I^r_{j,s+1} - I^r_{j,s})/R\right]/2\,, \tag{6}$$

where $s$ is a switch point of Isoform $i$ and $j$; $R$ is the number of replicates.

*2.4.1.5 Event importance.* Event importance is a novel metric that accounts for the expression level of switching isoforms. We defined event importance of a switch occurs between time point $t$ and $t+1$ as:

$$\text{Event importance} = \sum_{r=1}^{R}\left[\left(\frac{I^r_{aGt}}{\max(I^r_{Gt})} + \frac{I^r_{aGt+1}}{\max(I^r_{Gt+1})} + \frac{I^r_{bGt}}{\max(I^r_{Gt})} + \frac{I^r_{bGt+1}}{\max(I^r_{Gt+1})}\right)/4\right]/R, \tag{7}$$

where $I^r_{aGt}$is the relative abundance of isoform $a$ of a gene $G$ at time point $t$; and $R$ is the total number of replicates. Each $I$ is normalized to the highest relative abundance $\max(I^r_{Gt})$ at the corresponding time point. The metric takes the average of the relative abundance of isoforms $i$ and $j$ before and after switch.

For the analysis, we used IS events with differences of relative abundance higher than 0.2 and event importance higher than 0.3.

*2.1.4.6 Dissimilarity coefficient.* Dissimilarity coefficients $d_{i,j}$ assess the dissimilarity of the time course between isoforms. It is calculated based on the Pearson correlation $r_{i,j}$ between time course $I$ and $J$:

$$r_{i,j} = \frac{\mathrm{cov}(I,J)}{\sigma_i, \sigma_j} \tag{8}$$

$$d = \frac{1-r}{2}. \tag{9}$$

The higher coefficient, the less similar are the time courses.

*2.4.1.7 Domain inclusion or exclusion.* We used the Pfam database v.35.0 (Mistry *et al.*, 2021) to map domains to isoforms. Spycone compares isoforms in the IS event with each other to define if there is a loss/gain of domain.

*2.4.1.8 Multiple testing correction.* Finally, we implemented multiple testing corrections for IS detection. Available corrections are Bonferroni, Holm–Bonferroni and Benjamini–Hochberg false discovery rate. We use the Benjamini–Hochberg method as default.

### 2.4.2 TSIS

To detect IS in TSIS, we used the following parameters: (i) the switching probability $> 0.5$; (ii) difference before and after switch $> 10$; (iii) interval lasting before and after at minimum one time point; (iv) $P$-value $< 0.05$ and (v) Pearson correlation $< 0$. More detailed descriptions of parameters are found in Guo *et al.* (2017). The above parameters are set with defaults suggested by TSIS, except parameter (iii), since we have a larger interval between time points (12 h at maximum).

## 2.5 Change of total isoform usage

Isoform usage measures the relative abundance of an isoform. Isoform usage of all isoforms from one gene are summed up to obtain the total isoform usage. We defined the change of total isoform usage as between two consecutive time points:

$$\Delta\text{total isoform usage} = \sum_{A=0}^{n}\left|\left(\frac{I_{AGt1}}{\sum_{A=0}^{n}(I_{AGt1})} - \frac{I_{AGt0}}{\sum_{A=0}^{n}(I_{AGt0})}\right)\right|, \tag{10}$$

where $I$ is the expression of isoform $A$ of gene $G$ at time points $t_1$ and $t_0$; and $n$ is the total number of all isoforms for gene $G$.

## 2.6 Clustering analysis

The clustering algorithms are implemented using the scikit-learn machine learning package in python (v0.23.2) (Pedregosa *et al.*, 2011) and tslearn (v0.5.1.0) time course machine learning package in python (Tavenard *et al.*, 2020). The available algorithms are K-means, K-medoids, agglomerative clustering, DBSCAN and OPTICS.

The number of clusters is chosen manually by visually checking the Ward distance dendrogram (Supplementary Fig. S6).

## 2.7 Gene set enrichment and network analysis

For enrichment analysis, Spycone uses g:Profiler and NEASE. g:Profiler is a functional enrichment toolkit for GO terms and pathways (Raudvere *et al.*, 2019). Gene set enrichment method is performed using Fisher's exact test. NEASE (Louadi *et al.*, 2021) is an enrichment method for co-regulated alternative exons. We used NEASE with KEGG and Reactome pathways. For a seamless analysis, the newest version of the NEASE's Python package (v1.1.9) is integrated with Spycone.

Spycone employs DOMINO (0.1.0) (Levi *et al.*, 2021) for active module identification in PPI networks using default parameters.

### 2.8 Splicing factor co-expression and motif enrichment analysis

List of splicing factors and their position-specific scoring matrices (PSSMs) are obtained from the mCross database (downloaded in 2022), currently only available for *Homo sapiens* (Feng *et al.*, 2019). First, we filtered splicing factors with TPM > 1 in all time points. Next, we calculated the correlation between the relative abundance of each isoform and the expression of splicing factors. We filtered the pairs with correlation >0.7 or <−0.7 and adjusted *P*-value <0.05.

Finally, we performed motif enrichment analysis using the motifs module from the Biopython library (Cock *et al.*, 2009). The motifs module computes the log-odd probability of a specific region in the genome to match the binding motif using the PSSM (Henikoff and Henikoff, 1996). Hence, the higher the log-odd score, the more likely the binding. We compared these scores obtained from the lost, gained and unregulated exons from the same clusters. A Mann–Whitney *U*-test is performed on the sets of scores. Each motif threshold is selected using the distribution of the PSSM score over the frequency of nucleotides (background). The threshold is set at a false positive rate <0.01, meaning the probability of finding the motif in the background is <0.01.

## 3 Results

### 3.1 Spycone overview

Spycone is available as a python package that provides systematic analysis of time course transcriptomics data. Figure 1 shows the workflow of Spycone. It uses gene or isoform expression and a biological network as an input. It employs the sum of changes of all isoforms relative abundance (total isoform usage) (de la Fuente *et al.*, 2020) (see Section 2), i.e. the sums of pairwise changes in relative isoform abundance, across time points to detect IS events. It further provides downstream analysis such as clustering by total isoform usage, gene set enrichment analysis, network enrichment and splicing factors analysis. Visualization functions are provided for IS events, cluster prototypes, network modules and gene set enrichment results.

*IS detection.* We propose novel metrics for the detection and selection of significant IS across time. IS events are described as a change of the isoform distribution between two conditions (time points). To detect an IS, our algorithm first searches for switch points, i.e. a specific time point where two isoform expression time courses intersect.

The main challenges to detect time course IS are: (i) most genes have multiple isoforms, the changes of the relative abundance can be due to factors other than AS, e.g. RNA degradation. (ii) Most IS have multiple switch points, with different magnitudes of change in abundance; we need to consider how prominent the changes in abundance are to be recognized as an IS event. (iii) Most genes have multiple lowly expressed isoforms that constitute noise and might not be biologically relevant. An ideal IS detection tool, therefore, should prioritize IS events according to their expression level (Supplementary Fig. S1).

Spycone overcomes these challenges by using a novel approach to detect IS events. Spycone uses two metrics: a *P*-value and event importance. The *P*-value is calculated by performing a two-sided Mann–Whitney *U*-test between relative abundance before and after the switch point among the replicates. Event importance is the average of the ratio of the relative abundance of the two switching isoforms to the relative abundance of the isoform with the highest expression between the switching time points (see Section 2). Examples of high and low event importance are illustrated in Figure 2. The event importance will be highest when an IS includes the highest expressed isoform. Similarly, event importance will be low if an IS occurs between two lowly expressed isoforms. We also provide different metrics to comprehensively assess features of the IS events including switching probability, difference of abundance before and after switching and a dissimilarity coefficient (see Section 2).

*Clustering analysis for identifying co-spliced genes.* Similar to how transcription factors co-regulate sets of genes, in the context of AS, the splicing events of a subset of genes are co-regulated by splicing factors (Barberan-Soler *et al.*, 2011). For genes with important IS events (identified as described above), we want to quantify the impact of splicing regulation between two time points. To this end, Spycone clusters genes by changes in total isoform usage over time to identify co-spliced genes. A previous study showed that clustering performance is highly dependent on the dataset and the clustering method (Javed *et al.*, 2020). Therefore, Spycone offers various clustering techniques, including agglomerative clustering (hierarchical clustering) (Johnson, 1967), K-Means clustering (Hartigan and Wong, 1979), K-Medoids clustering (Park and Jun, 2009), DBSCAN (Ester *et al.*, 1996), OPTICS (Ankerst *et al.*, 1999) and various distance metrics such as euclidean distance, Pearson distance, as well as tslearn (Tavenard *et al.*, 2020) for calculating the dynamic time warping distance measure.

With temporal patterns of the clusters, Spycone dissects context-specific processes in terms of AS. In order to gain functional knowledge of the clusters, Spycone offers g:Profiler (Raudvere *et al.*,

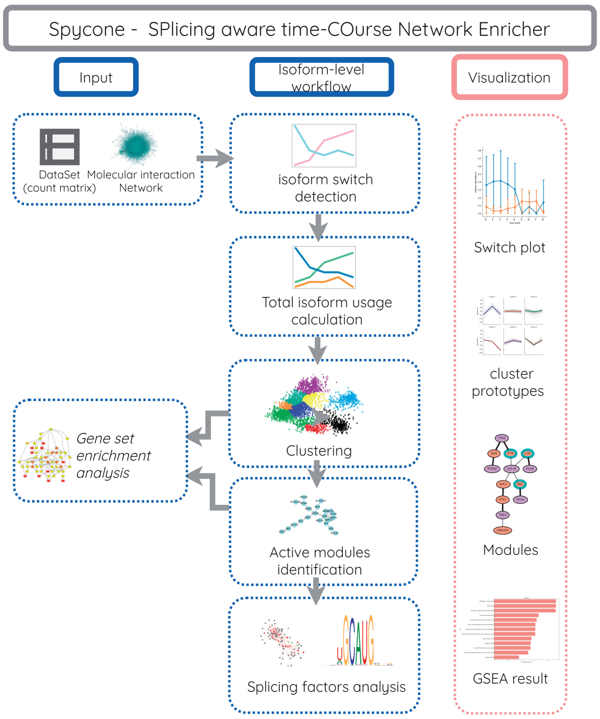

Spycone - SPlicing aware time-COurse Network Enricher
Input
Isoform-level workflow
Visualization
DataSet (count matrix)
Molecular interaction Network
isoform switch detection
Total isoform usage calculation
Clustering
Active modules identification
Splicing factors analysis
Gene set enrichment analysis
Switch plot
cluster prototypes
Modules
GSEA result

**Fig. 1.** Overview of the Spycone architecture. Spycone takes count matrices and biological networks as input. We provide isoform-level functions such as isoform switch detection and total isoform usage calculation. Users could also cluster the gene count matrix directly. For downstream analysis, we integrated multiple clustering algorithms and an active modules identification algorithm (DOMINO). We also implemented splicing factors analysis for isoform-level data. Finally, visualizations are provided to better evaluate and interpret the results

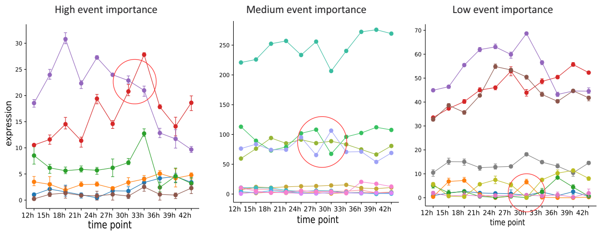


**Fig. 2.** Plots showing the examples of three levels of event importance. Each plot contains all isoforms of a gene. The circle indicates the IS events with the corresponding level of event importance

2019) and NEASE (Louadi *et al.*, 2021) for gene set enrichment analysis. The former conducts classical enrichment analysis for multiple ontologies and pathway databases. The latter combines information from PPI and domain–domain interaction networks and allows to predict functional consequences of AS events caused by a set of IS genes.

*Active modules identification.* Genes with consistent temporal patterns are thought to be functionally related in terms of co-regulation, molecular interactions or participation in the same cellular processes. To uncover the underlying mechanism that is represented by a temporal pattern, Spycone projects the results of the clustering analysis on a molecular interaction network for active modules identification, i.e. for detection of subnetworks enriched in genes affected by IS. We incorporated DOMINO (Levi *et al.*, 2021) as it has been previously demonstrated the best performance for this task (Lazareva *et al.*, 2021). To elucidate the functional impact of IS events, we further leveraged domain–domain interaction information from the 3did database (Mosca *et al.*, 2014). Spycone identifies domains lost/gained during IS, which might indicate a functional switch, and affected edges in the PPI network. This provides additional insights about the functional consequences of time course IS.

*Splicing factor analysis.* Spycone also provides splicing factor analysis using co-expression and RNA-binding protein motif search. Splicing factors are a group of RNA-binding proteins that regulate the splicing of genes. We assume that the expression of splicing factors that are responsible for an IS event correlates with the relative abundance of participating isoforms. Spycone calculates the correlation between the expression value of a list of RNA-binding proteins derived from ENCODE eCLIP data (Feng *et al.*, 2019) and the relative abundance of isoforms involved in IS. We implemented PSSM of RNA-binding protein motifs to calculate and detect the potential binding sites along the sequence of the targeted isoforms (see Section 2).

## 3.2 Evaluation using simulated data

To evaluate the performance of Spycone, we compared its performance (precision and recall) to TSIS using simulated data. TSIS provides an option to filter for IS events that involve only the highest abundance isoform—we refer to the result after filtering as major_TSIS. We aimed to investigate whether the performance of TSIS improves when applying this option.

We use a hidden Markov model to simulate the switching state of the genes at each time point (see Section 2). We simulated two models (Supplementary Fig. S2): Model 1 allows only major isoforms, i.e. those with the highest abundance per gene, to be involved in IS events across time points; Model 2 allows IS to occur between isoforms with relative abundance higher than 0.3. We used Model 2 to show that neither tool is biased towards events that involve only major isoforms.

For both tools, we varied their parameters (difference of relative abundance), to investigate how this affects their precision and recall. We also considered different levels of variance of gene expression, namely 1, 5 and 10, across replicates to mimic the noise (Fig. 3).

In Model 1, Spycone achieved high precision and recall. The precision of TSIS dropped drastically with increasing recall. After filtering major events, TSIS's recall reached 0.5. Spycone performs better in the setting with the highest noise level as it maintains high precision (0.95) and acceptable recall (0.75). In Model 2, Spycone achieved higher precision and recall than TSIS; however, they dropped as the model allows more IS events. We applied spline regression to detect switch points and calculated precision and recall as above (Supplementary Fig. S3). Results showed that spline regression does not improve precision and recall in both tools. Moreover, TSIS has a higher algorithmic complexity of $O(n^*\log(n))$ than Spycone with a complexity of $O(n)$, leading to a drastically lower runtime for Spycone in the range of a few minutes rather than hours (Supplementary Fig. S4). In summary, Spycone outperforms TSIS in detecting IS events.

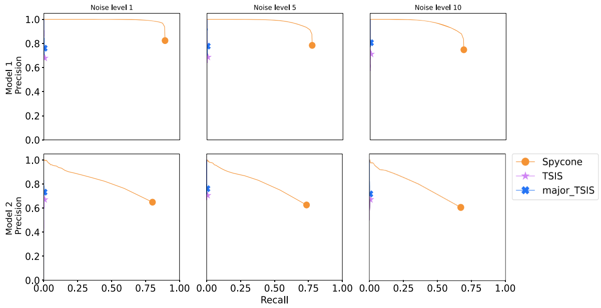


**Fig. 3.** Precision and recall curves for Spycone, TSIS and major-isoforms filtered TSIS from simulated data of two models (rows) and three noise levels (columns)

## 3.3 Application to SARS-Cov2 infection data

We applied Spycone to an RNA-seq time course dataset of SARS-CoV-2-infected human lung cells (Kim *et al.*, 2021). The dataset contains eight time points: 0, 1, 2, 4, 12, 16, 24 and 36 h post-infection. We kept isoforms with TPM > 1 across all time points resulting in 36 062 isoforms for IS event detection with Spycone and TSIS. To call an IS significant, we used the following criteria: for Spycone, (i) switching probability > 0.5; (ii) difference of relative abundance > 0.2 before and after the switch; (iii) dissimilarity coefficient > 0.5; and (iv) adjusted *P*-value < 0.05. For TSIS, we used (i) switching probability > 0.5; (ii) difference of expression before and after switch > 10; (iii) correlation coefficient < 0; and (iv) adjusted *P*-value < 0.05. The dissimilarity coefficient from Spycone and the correlation coefficient from TSIS are used to filter for IS events with negatively correlated isoforms. The values are chosen according to the performance on Model 2 simulated data with noise level 10 that showed the best precision. Spycone reported 915 IS events, of which 418 affected at least 1 protein domain. TSIS reported 985 events, of which 417 affected at least one protein domain. On gene level, Spycone reported 745 genes with IS events, TSIS reported 858 genes where 225 genes were found by both Spycone and TSIS (Fig. 4A).

We then used the event importance metric to assess the ability of each method to detect IS events from higher abundance isoforms. We calculated event importance for IS events identified by Spycone, TSIS and major_TSIS (Fig. 4B). Spycone results include mostly events with high importance, while in TSIS events with low importance prevail. Supplementary Table S1 shows the result for the SARS-CoV-2 dataset from the Spycone IS detection. Event importance has no clear prevalence towards overall gene expression and adjusted *P*-value (Supplementary Figs S5 and S6).

To exclude IS events with lowly expressed isoforms, we applied a filter of event importance higher than 0.3 to both Spycone and TSIS results. We calculated the change of total isoform usage of the IS genes across time points and employed Ward linkage hierarchical clustering. This led to four clusters with similar temporal patterns of changes in total isoform usage for Spycone (Fig. 4C, Supplementary Fig. S7, Supplementary Table S2) and four clusters for TSIS (Supplementary Fig. S10A). Each cluster is represented by a cluster prototype, which is the median change of total isoform usage per pair of time points.

IS events that lead to domain gain or loss might break the interactions, hence rewiring the PPI network. Moreover, if the IS events belong to the same cluster, it indicates the synchronized gain or loss of interactions with particular pathways. Our goal is therefore to assess if IS events within clusters rewire interactions with particular pathways during SARS-CoV-2 infection. We performed AS-aware pathway enrichment analysis using NEASE with KEGG (Kanehisa *et al.*, 2016) and Reactome (Jassal *et al.*, 2020) pathway databases for results from Spycone (Supplementary Fig. S8, Supplementary Table S3) and TSIS (Supplementary Table S4, Supplementary Fig S10B). In addition, we performed classical gene set enrichment analysis using g:Profiler. The results are not informative since only five terms are found in Cluster 3 and zero in others.

Overall, clusters with similar prototypes from both tools are enriched in distinct pathway terms. For example, TSIS's Cluster 1

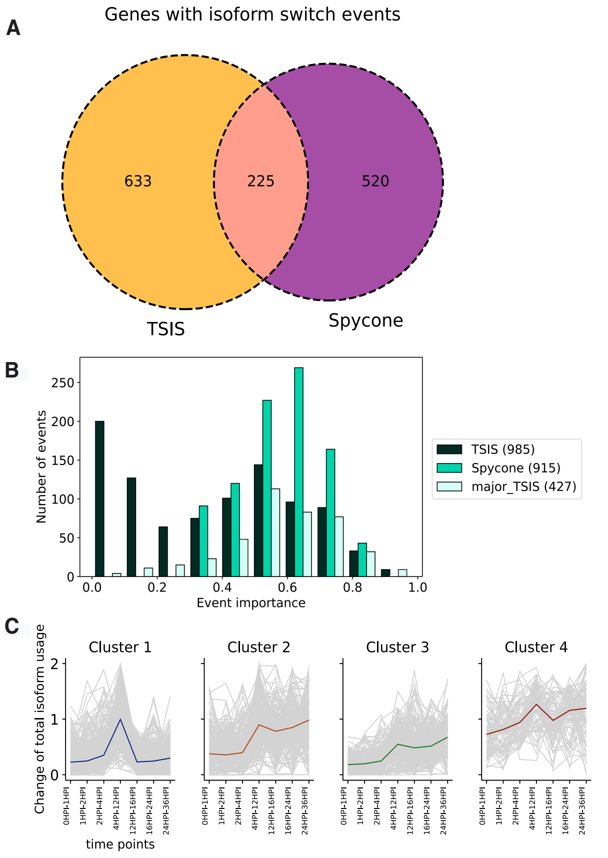


**Fig. 4.** Comparing IS detection results from Spycone and TSIS (**A**) Venn diagram showing the number of genes detected with isoform switch events by Spycone and TSIS (all). (**B**) The distribution of the detected events in Spycone, TSIS and major_TSIS based on the event importance metric of Spycone. The number of events for each tool is indicated in brackets in the legend. (**C**) Cluster prototypes and all objects show the pattern of the change of total isoform usage across time points

and Spycone's Cluster 1 have a strong peak between 4 and 12 h post-infection. Only transforming growth factor (TGF)-beta signaling is commonly found in both tools. MAPK pathway and DNA damage checkpoint are enriched uniquely in Spycone. TSIS's Cluster 2 and Spycone's Cluster 3 have lower changes of total isoform usage overall. Spycone's clusters showed more unique and relevant terms: 70 enriched Reactome terms in Spycone's clusters and only 7 terms in TSIS's clusters. TSIS's Cluster 3 and Spycone Cluster 2 show an increase of change of total isoform usage after 12 h post-infection. Spycone's cluster is enriched uniquely in protein folding chaperonin complex TriC/CCT and NOTCH signaling pathway. Finally, TSIS's Cluster 4 and Spycone's Cluster 4 have increasing changes of total isoform usage overall. TSIS's cluster is enriched in mitosis-related pathways, cell cycle and tubulin folding. Whereas in Spycone's Cluster 4 is found with signaling by PTK6, interferon, metabolism of proteins, pentose phosphate pathway, etc.

Next, we detected active modules that show over-representation of IS genes from the same cluster based on DOMINO using a PPI network from BioGRID (Oughtred *et al.*, 2021) (see Section 2). Detected active modules suggest the impact of splicing on regulatory cascades and cellular trafficking (Table 1, Fig. 5, Supplementary Fig. S7).

### 3.3.1 Splicing factor Anaysis

Assuming that multiple IS events occurring between the same time points are co-regulated by the same splicing factor, we perform co-expression and motif analysis. The co-expression analysis yields thirteen significant RNA-binding proteins that are positively or negatively correlated with at least two isoforms of the same gene: in cluster 1 - FUBP3, HLTF, IGF2BP3, ILF3, RBFOX2, RBM22, SF3B1 and TAF15; in cluster 3 - IGF2BP3, RBM22, RPS6, SRSF7 and SUGP2; ( $|r| > 0.6$ and adjusted p-value $< 0.05$) (Table S5). To investigate whether the regulated exons, i.e. the lost or gained exons after IS events, show higher PSSM scores to a certain RNA-binding protein motif than the unregulated exons in a cluster, we applied motif enrichment analysis. We calculated PSSM scores along the flanking regions of the exons 5' and 3' boundaries and excluded the first and last exons in an isoform since these are often regulated by 5'-cap binding proteins and polyadenylation regulating proteins (Zheng, 2004). All exons in the switched isoforms within a cluster are categorized to 1) lost exons, 2) gained exons, and 3) unregulated exons for the analysis (Fig.6, Table S6). RNA-binding proteins with multiple motifs are numbered with an underscore. Each motif is selected with a threshold where the false-positive rate is below 0.01. Position-specific log-odd scores higher than the corresponding threshold are obtained after calculating the PSSM scores of each motif for all exons (see Methods section). The ILF3_9 and ILF3_14 motifs show higher log-odd scores at the 5' end of the lost/gain exons than of the unregulated exons in cluster 1 (one-sided Mann-Whitney U test p-value $< 0.05$) (Fig. 6A). HLTF_7 and SRSF7_1 motifs show higher log-odd scores at the 3' end (Fig. 6B).

## 4 Discussion

AS regulates dynamic processes such as development and disease progression. However, AS analysis tools typically compare only two conditions and neglect how AS changes dynamically over time. Currently, the only existing tool for time course data analysis that accounts for splicing is TSIS. TSIS detects temporal IS events but is biased towards IS events between lowly expressed isoforms and does not offer features for downstream analysis which is important for interpreting the functional consequences of IS events.

Spycone, a framework for analysis of time course transcriptomics data, features a new approach for detecting temporal IS events and a new event importance metric to filter out lowly expressed isoforms. We demonstrate that Spycone's IS detection method outperforms TSIS in terms of precision and recall based on simulated data. A key advantage of Spycone is that it explicitly considers how well IS events agree across replicates while TSIS considers averaged expression values among replicates and/or by natural spline-curves fitting. More specifically, Spycone uses a non-parametric Mann–Whitney *U*-test to test for significant IS and performs multiple testing correction to reduce type I error.

We have demonstrated the usability of Spycone by analyzing time course transcriptomics data for SARS-CoV-2 infection where we found affected signaling cascades. We performed NEASE enrichment on the clusters and compared the results from Spycone and TSIS. Spycone results are enriched in relevant terms such as mitogen-activated protein kinase (MAPK) pathway (Cluster 1), NOTCH signaling (Cluster 2), fibroblast growth factor receptor (FGFRs) and toll-like receptor (TLR) pathways (Cluster 3) and pentose phosphate pathway (Cluster 4). NOTCH signaling pathways are found upregulated in the lungs of infected macaques (Rosa *et al.*, 2021).

The MAPK pathway has a pro-inflammatory effect by interacting with SARS-CoV-2 downstream pathogenesis, especially in patients suffering from cardiovascular disease (Weckbach *et al.*, 2022). TLR 7/8 cascades are related to ssRNA, and there is a study supporting the association of TLR 7/8 with SARS-CoV-2 infection (Salvi *et al.*, 2021). The pentose phosphate pathway is an alternative pathway of glycolysis that produces more reduced NADP (NADPH) oxidase. It is activated during SARS-CoV-2 infection in response to oxidative stress and the activation of the immune response (Yang *et al.*, 2021). Spycone also detected the enrichment of pathways which association with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) infection has not been characterized yet: kinesins, signaling by NTRKs, degradation of AXIN, signaling by Hedgehog and 5-phosphoribose 1-diphosphate biosynthesis.

The active modules extracted from the clusters highlight mechanisms involved in the host cell response to infection. In Cluster 2

**Table 1.** Related biological processes and pathways of the respective modules found in clusters

| Clusters | Module | Related biological processes |
|---|---|---|
| 1 (Fig. 5, Supplementary Fig. S9A) | 1 | RNA splicing and mRNA processing |
| | 2 | Cellular protein modification genes, signal transduction in the VEGF signaling pathway |
| | 3 | Positive regulation of protein ubiquitination |
| | 4 | Protein and intracellular trafficking genes |
| | 5 | Protein and intracellular trafficking genes |
| 2 (Supplementary Fig. S9B) | 1 | Transcription and mRNA splicing |
| | 2 | Ras and Rho protein signaling transduction |
| | 3 | Ubiquitination |
| | 4 | Protein import into the nucleus |
| | 5 | Transition of cell cycle to G2/M phase |
| 3 (Supplementary Fig. S9C) | 1 | Regulation of transcription, cell cycle arrest and protein catabolic process |
| | 2 | Transmembrane receptor protein tyrosine kinase signaling pathway, in particular MAPK cascade and ERK cascade |
| | 3 | Protein ubiquitination |
| | 4 | Histone acetylation |
| | 5 | Organelle membrane fusion |
| 4 (Supplementary Fig. S9D) | 1 | Transcription and apoptotic processes |

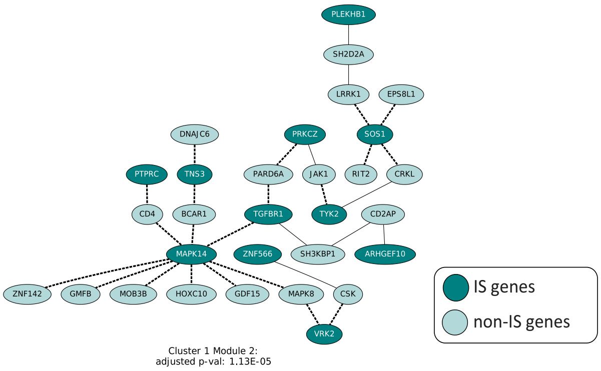


**Fig. 5.** Spycone results in modules of the PPI network and their corresponding gene set enrichment results. Active network modules are identified using DOMINO. Each node represents a domain of a gene. Darker nodes are the isoform switched genes and lighter nodes are non-IS genes from the PPI. Dashed edges are the affected interactions between the genes due to the loss/gain of domains during the IS events

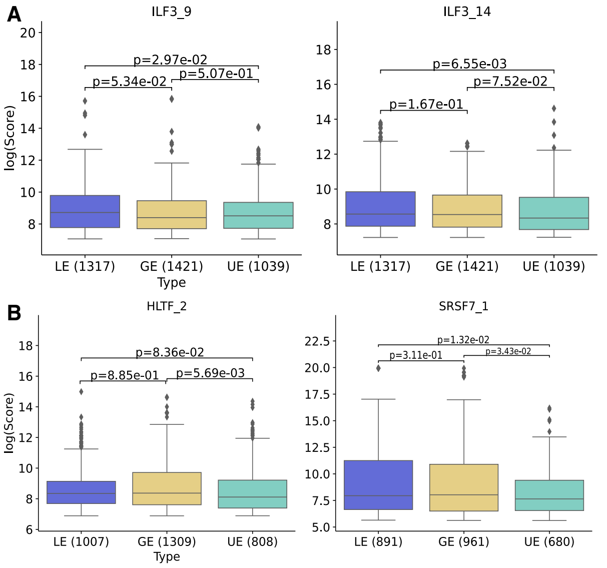


**Fig. 6.** (**A**) Boxplots showing the PSSM score difference between lost/gained exons and unregulated exons at the exon 5′ boundaries in logarithmic scale (one-sided Mann–Whitney *U*-test *P*-value < 0.05). (**B**) Boxplots showing the PSSM scores difference between lost/gained exons and unregulated exons at the exon 3′ boundaries in logarithmic scale (one-sided Mann–Whitney *U*-test *P*-value < 0.05). LE, lost exons; GE, gained exons; UE, unregulated exons

Module 1 (Fig. 5B) revealed that interactions between three kinases (MAPK39, AURKC and DCLK2) and a protein chaperone, HSP90AA1 is affected by IS. HSP90 is expressed under the endoplasmic reticulum (ER) stress caused by SARS-CoV-2 and its inhibitor is identified as a therapeutic inhibition target (Wyler *et al.*, 2021). A previous study found that knock down of MAP3K9 reduced SARS-CoV-2 virus replication (Higgins *et al.*, 2021). DCLK2 is differentially expressed in SARS-CoV-2 patients (Alqutami *et al.*, 2021). AURKC would be an interesting candidate to investigate for its role in SARS-CoV-2 infection.

Besides these three kinases, network enrichment analysis highlighted the general importance of kinases in infection development, e.g. JAK1, LYN, TYK2 and PRKCZ (Fig. 5). JAK1 is responsible for interferon signaling (Yan *et al.*, 2021). Inhibition of LYN reduces the efficiency of SARS-CoV-2 virus replication (Meyer *et al.*, 2021). TYK2, which is a key player for IFN signaling, has been associated with cytokine storms in SARS-CoV-2 patients (Solimani *et al.*, 2021). IS events of kinases might cause major rewiring of the transduction cascade, which could lead to altered immune response, cell cycle control and promote viral replication.

Our analysis also suggests an important role of growth factor receptors (FGFR, epidermal growth factor receptor (EGFR) and vascular endothelial growth factor (VEGF)) and their downstream kinases. They are essential for viral infection since they modulate cellular processes like migration, adhesion, differentiation and survival. One example is that activation of EGFR in SARS-CoV-2 can suppress the IFN response and aid viral replication (Klann *et al.*, 2020).

Another key finding is that E3 ubiquitin ligases are affected by IS. They are known to mediate host immune response by removing virus particles. Various virus species hijack the host E3 ubiquitin ligases in favor of viral protein production (Dubey *et al.*, 2021). They are also involved in maintaining TMPRSS2 stabilization during virus entry to the host cells (Chen *et al.*, 2021).

In splicing factor analysis, ILF3 and SRSF7 are identified as a splicing factor affecting the splicing of exons. ILF3 plays a role in antiviral response by inducing the expression of interferon-stimulated genes (Watson *et al.*, 2020). In another computational analysis, SRSF7 is also predicted to have binding potential with SARS-CoV-2 RNA (Horlacher *et al.*, 2021).

Lastly, in order to get confident time course analysis results, one will need high-resolution data in terms of number of time points and sample replicates. Consequently, at least three time points and three replicates are recommended in Spycone analysis. However, this criterion is rather met due to technical and economical restraints. Thus, Spycone also provides an option for a permutation test with only one replicate for the dataset under investigation. We demonstrated this usage in a tumor development dataset with one replicate (see Supplementary information).

*Limitations*. Spycone achieves high precision and considerably higher recall than the only competing tool TSIS. Nevertheless, the moderate recall we observe in particular in the presence of noise shows that there is further room for method improvement. In our simulation Model 2, where we allowed for isoform switches between minor isoforms, we observed a reduction in both precision and recall. Spycone identifies only two isoforms that switch per event, but in reality, an event could involve more than two isoforms. In the future, we should consider multiple-isoforms switches to handle more complex scenarios. In addition, the usage of weighted PPI network might introduce selection bias. However, the higher weight gives higher confidence to an interaction, meaning more domains between the proteins are interacting. Therefore, using weighted PPI helps prioritizing interactions with higher confidence. We believe this advantage outweighs the potential bias. Nevertheless, the usage of weighted PPI is optional.

Spycone uniquely offers features for detailed downstream analysis and allows for detecting the rewiring of network modules in a time course as a result of coordinated domain gain/loss. This type of analysis is limited by the availability of the structural annotation. However, the current developments in computational structural biology that could expand the information about domains and domain–domain interactions e.g. AlphaFold2 (Jumper *et al.*, 2021), will greatly strengthen our tool. Lastly, our PSSM-based approach for splicing factor analysis does not allow us to investigate splicing factors that bind indirectly through other adaptor proteins, requiring further experiments that establish binding sites for such proteins. In our future work, we plan to optimize the algorithm and include introns in the analysis.

Spycone was thus far applied exclusively to bulk RNA-seq data. When considering tissue samples, IS switches between time points could also be attributed to changes in cellular composition. An attractive future prospect is thus to apply Spycone for studying IS in single-cell RNA-seq data where dynamic IS events could be traced across cellular differentiation using the concept of pseudotime. However, the current single-cell RNA-seq technologies are limited in their ability to discern isoforms (Arzalluz-Luque and Conesa, 2018).

## 5 Conclusion

With declining costs in next-generation sequencing, time course RNA-seq experiments are growing in popularity. Although AS is an important and dynamic mechanism it is currently rarely studied in a time course manner due to the lack of suitable tools. Spycone closes this gap by offering robust and comprehensive analysis of time course IS. Going beyond individual IS events, Spycone clusters genes with similar IS behavior in time course data and offers insights into the functional interpretation as well as putative mechanisms and co-regulation. The latter is achieved by RNA-binding protein motif analysis and highlights splice factors that could serve as potential drug targets for diseases. Using simulated and real data, we showed that Spycone has better precision and recall than its only competitor, TSIS and that Spycone is able to identify disease-related pathways in the real-world data, as we demonstrated for SARS-CoV-2 infection. In summary, Spycone brings mechanistic insights about the role of temporal changes in AS and thus perfectly complements RNA-seq time course analysis.

## Funding

This work was developed as part of the ASPIRE project and is funded by the German Federal Ministry of Education and Research (BMBF) under grant number 031L0287B. This work was also supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the *e:Med* research and funding concept (*grants 01ZX1908A, 01ZX2208A and 01ZX1910D). JB was partially funded by his VILLUM Young Investigator Grant nr.13154.

*Conflict of Interest*: none declared.

## Data availability

The SARS-CoV-2 infection RNA-sequencing data are obtained from the GEO database (accession ID GSE157490). The Spycone package is available as a PyPI package. The source code of Spycone is available under the GPLv3 license at https://github.com/yollct/spycone. The code used to produce the result shown in this manuscript is compiled into the Google colab notebook (https://colab.research.google.com/drive/13CjfzZizPlmxzsT-zm6zEgfdFce1fzSC?usp=sharing). This workflow is documented in the AIMe registry: https://aime.report/DXKacH (Matschinske *et al.*, 2021).

## References

Aminikhanghahi,S. and Cook,D.J. (2017) A survey of methods for time series change point detection. *Knowl. Inf. Syst.*, **51**, 339–367.

Alqutami,F. *et al.* (2021) COVID-19 transcriptomic atlas: a comprehensive analysis of COVID-19 related transcriptomics datasets in different tissues and clinical settings. *Front. Genet.*, **12**, 755222.

Ankerst,M. *et al.* (1999) OPTICS: ordering points to identify the clustering structure. *SIGMOD Rec.*, **28**, 49–60.

Arzalluz-Luque,Á. and Conesa,A. (2018) Single-cell RNAseq for the study of isoforms—how is that possible? *Genome Biol.*, **19**, 1–19.

Barberan-Soler,S. *et al.* (2011) Co-regulation of alternative splicing by diverse splicing factors in *Caenorhabditis elegans*. *Nucleic Acids Res.*, **39**, 666–674.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Chen,Y. *et al.* (2021) A high-throughput screen for TMPRSS2 expression identifies FDA-approved compounds that can limit SARS-CoV-2 entry. *Nat. Commun.*, **12**, 3907.

Cock,P.J.A. *et al.* (2009) Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

de la Fuente,L. *et al.* (2020) tappAS: a comprehensive computational framework for the analysis of the functional impact of differential splicing. *Genome Biol.*, **21**, 119.

Dubey,A.R. *et al.* (2021) Biochemical strategies of E3 ubiquitin ligases target viruses in critical diseases. *J. Cell. Biochem.*, **123**, 161–182.

Ester,M. *et al.* (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231.

Feng,H. *et al.* (2019) Modeling RNA-Binding protein specificity in vivo by precisely registering protein-RNA crosslink sites. *Mol. Cell*, **74**, 1189–1204.e6.

Guo,W. *et al.* (2017) TSIS: an R package to infer alternative splicing isoform switches for time-series data. *Bioinformatics*, **33**, 3308–3310.

Hartigan,J.A. and Wong,M.A. (1979) Algorithm as 136: a K-Means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.*, **28**, 100–108.

Henikoff,J.G. and Henikoff,S. (1996) Using substitution probabilities to improve position-specific scoring matrices. *Comput. Appl. Biosci.*, **12**, 135–143.

Higgins,C.A. *et al.* (2021) SARS-CoV-2 hijacks p38ß/MAPK11 to promote viral protein translation. bioRxiv, 2021.08.20.457146.

Hooper,J.E. *et al.* (2020) An alternative splicing program for mouse craniofacial development. *Front. Physiol.*, **11**, 1099.

Horlacher,M. *et al.* (2021) Computational mapping of the human-SARS-CoV-2 protein-RNA interactome. bioRxiv, 2021.12.22.472458.

Jang,J. *et al.* (2021) TimesVector-Web: a web service for analysing time course transcriptome data with multiple conditions. *Genes*, **13**, 73.

Jassal,B. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.

Javed,A. *et al.* (2020) A benchmark study on time series clustering. *Mach. Learn. Appl.*, **1**, 100001.

Johnson,S.C. (1967) Hierarchical clustering schemes. *Psychometrika*, **32**, 241–254.

Jumper,J. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.

Kanehisa,M. *et al.* (2016) KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.*, **44**, D457–D462.

Kim,D. *et al.* (2021) A high-resolution temporal atlas of the SARS-CoV-2 translatome and transcriptome. *Nat. Commun.*, **12**, 5120.

Klann,K. *et al.* (2020) Growth factor receptor signaling inhibition prevents SARS-CoV-2 replication. *Mol. Cell.*, **80**, 164–174.e4.

Lazareva,O. *et al.* (2021) On the limits of active module identification. *Brief. Bioinform.*, **22**, bbab066.

Levi,H. *et al.* (2021) DOMINO: a network-based active module identification algorithm with reduced rate of false calls. *Mol. Syst. Biol.*, **17**, e9593.

Louadi,Z. *et al.* (2021) Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases. *Genome Biol.*, **22**, 327.

Makarenko,I. *et al.* (2004) Passive stiffness changes caused by upregulation of compliant titin isoforms in human dilated cardiomyopathy hearts. *Circ. Res.*, **95**, 708–716.

Matschinske,J. *et al.* (2021) The AIMe registry for artificial intelligence in biomedical research. *Nat. Methods*, **18**, 1128–1131.

Meyer,B. *et al.* (2021) Characterising proteolysis during SARS-CoV-2 infection identifies viral cleavage sites and cellular targets with therapeutic potential. *Nat. Commun.*, **12**, 5553.

Mistry,J. *et al.* (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.

Mosca,R. *et al.* (2014) 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic Acids Res.*, **42**, D374–D379.

Nueda,M.J. *et al.* (2018) Identification and visualization of differential isoform expression in RNA-seq time series. *Bioinformatics*, **34**, 524–526.

Oughtred,R. *et al.* (2021) The BioGRID database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.*, **30**, 187–200.

Park,H.-S. and Jun,C.-H. (2009) A simple and fast algorithm for K-medoids clustering. *Expert Syst. Appl.*, **36**, 3336–3341.

Patro,R. *et al.* (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods*, **14**, 417–419.

Pedregosa,F. *et al.* (2011) Scikit-learn: machine learning in python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Raudvere,U. *et al.* (2019) g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.*, **47**, W191–W198.

Rosa,B.A. *et al.* (2021) IFN signaling and neutrophil degranulation transcriptional signatures are induced during SARS-CoV-2 infection. *Commun. Biol.*, **4**, 290.

Salvi,V. *et al.* (2021) SARS-CoV-2-associated ssRNAs activate inflammation and immunity via TLR7/8. *JCI Insight*, **6**(18), e150542.

Shim,J.E. and Lee,I. (2016) Weighted mutual information analysis substantially improves domain-based functional network models. *Bioinformatics*, **32**, 2824–2830.

Solimani,F. *et al.* (2021) Janus kinase signaling as risk factor and therapeutic target for severe SARS-CoV-2 infection. *Eur. J. Immunol.*, **51**, 1071–1075.

Tavenard,R. *et al.* (2020) Tslearn, a machine learning toolkit for time series data. *J. Mach. Learn. Res*, **21**, 1–6.

Trincado,J.L. *et al.* (2017) SUPPA2 provides fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biol.*, **19**, 40.

Varoquaux,N. and Purdom,E. (2020) A pipeline to analyse time-course gene expression data. *F1000Res.*, **9**, 1447.

Vitting-Seerup,K. and Sandelin,A. (2017) The landscape of isoform switches in human cancers. *Mol. Cancer Res.*, **15**, 1206–1220.

Wan,J. *et al.* (2011) Dynamic usage of alternative splicing exons during mouse retina development. *Nucleic Acids Res.*, **39**, 7920–7930.

Watson,S.F. *et al.* (2020) ILF3 contributes to the establishment of the antiviral type I interferon program. *Nucleic Acids Res*, **48**, 116–129.

Weckbach,L.T. *et al.*; EMB Study Group. (2022) Association of complement and MAPK activation with SARS-CoV-2-Associated myocardial inflammation. *JAMA Cardiol.*, **7**, 286–297.

Wiwie,C. *et al.* (2019) Time-resolved systems medicine reveals viral infection-modulating host targets. *Syst. Med. (New Rochelle)*, **2**, 1–9.

Wyler,E. *et al.* (2021) Transcriptomic profiling of SARS-CoV-2 infected human cell lines identifies HSP90 as target for COVID-19 therapy. *iScience*, **24**, 102151.

Xing,Y. *et al.* (2020) Dynamic alternative splicing during mouse preimplantation embryo development. *Front. Bioeng. Biotechnol.*, **8**, 35.

Yan,B. *et al.* (2021) SARS-CoV-2 drives JAK1/2-dependent local complement hyperactivation. *Sci. Immunol.*, **6**, eabg0833.

Yang,H.-C. *et al.* (2021) G6PD deficiency, redox homeostasis, and viral infections: implications for SARS-CoV-2 (COVID-19). *Free Radic. Res.*, **55**, 364–374.

Zheng,Z.-M. (2004) Regulation of alternative RNA splicing by exon definition and exon sequences in viral and mammalian gene expression. *J. Biomed. Sci.*, **11**, 278–294.

# 5 Unpublished Results

## 5.1 Comprehensive benchmark of differential transcript usage analysis for static and dynamic conditions

### Citation

The article titled "Comprehensive benchmark of differential transcript usage analysis for static and dynamic conditions" has been submitted and available online at bioRxiv on 15 Jan 2024.

### Full citation:

Lio CT, Düz T, Hoffmann MD, Willruth LL, Baumbach J, Tsoy O, List M. Comprehensive benchmark of differential transcript usage analysis for static and dynamic conditions. bioRxiv. 2024:2024-01.

### Summary

Differential transcript usage (DTU) provides a great means for investigating the effect of alternative splicing changes. This is done by observing and modeling the changes in transcript abundance within a gene among conditions. DTU changes indicate the change in the underlying splicing patterns between conditions. Several tools allow this type of analysis, and several benchmarking analyses are performed. However, not all published DTU tools are benchmarked, and some additional experimental settings should be investigated. This paper aimed to provide an updated view of DTU analysis with never-benchmarked tools. I evaluated the DTU tools under different experimental conditions, such as bulk RNA sequencing with fewer and more replicates in both single-end and paired-end data. I compared the qualitative difference (i.e., functional enrichment) between pairwise comparison DTU tools and time series DTU tools. I simulated single-cell balanced and unbalanced data with DTU events to evaluate single-cell DTU tools.
I used RSEM simulator for both bulk RNA-seq and single-cell RNA-seq data. For bulk RNA-seq, single-end and paired-end data with 50 and 100 million reads are simulated, each with four and eight replicates. For single-cell data, only balanced dataset with two cell types are simulated. Each group in each dataset contains ranging from 10 to 500 cells. These data contain DTU events as ground truth for the evaluation. With simulated data, I provided guidelines for using DTU tools for both paired-end and single-end data with fewer or more replicates processed with different quantification methods. Our result provided insights into the biological interpretation of time series analysis using time-series-specific tools. We evaluated single-cell DTU tools on simulated single-cell data. Our results provided a comprehensive current DTU analysis under various experimental conditions.

### Availablitiy

The RNA-seq information for the prostate tumor dataset can be accessed via GSE222260, while data concerning patients infected with SARS-Cov-2 is available from GSE162562 and GSE190680. Additionally, the time series RNA-seq data related to SARS-Cov-2 infection was sourced from GSE157490.

### Contribution

I did the literature research, designed and performed the analysis, from the exploratory data analysis which I tried different approaches to finalizing the analysis workflow. I generated all the figures and wrote the manuscript. I prepared the manuscript for submission to peer-reviewed journals.

### Rights and permissions

### Additional supplementary material

Supplementary material can be found on the bioRiv website: https://www.biorxiv.org/content/10.1101/2024.01.14.57554
material.

# Comprehensive benchmark of differential transcript usage analysis for static and dynamic conditions

Chit Tong Lio[2], Tolga Düz[1], Markus Hoffmann[2,3,4], Lina-Liv Willruth[2], Jan Baumbach[1,5], Markus List[2], Olga Tsoy[1]

1 Chair of Computational Systems Biology, University of Hamburg, Notkestrasse 9, 22607 Hamburg, Germany
2 Data Science in Systems Biology, Technical University of Munich, 85354 Freising, Germany
3 Institute for Advanced Study, Technical University of Munich, Garching D-85748, Germany
4 National Institute of Diabetes, Digestive, and Kidney Diseases, National Institutes of Health, Bethesda, MD 20892, USA

5 Institute of Mathematics and Computer Science, University of Southern Denmark, Campusvej 55, 5000 Odense, Denmark

## Abstract

RNA sequencing offers unique insights into transcriptome diversity, and a plethora of tools have been developed to analyze alternative splicing. One important task is to detect changes in the relative transcript abundance in differential transcript usage (DTU) analysis. The choice of the right analysis tool is non-trivial and depends on experimental factors such as the availability of single- or paired-end and bulk or single-cell data. To help users select the most promising tool for their task, we performed a comprehensive benchmark of DTU detection tools. We cover a wide array of experimental settings, using simulated bulk and single-cell RNA-seq data as well as real transcriptomics datasets, including time-series data. Our results suggest that DEXSeq, edgeR, and LimmaDS are better choices for paired-end data, while DSGseq and DEXSeq can be used for single-end data. In single-cell simulation settings, we showed that satuRn performs better than DTUrtle. In addition, we showed that Spycone is optimal for time series DTU/IS analysis based on the evidence provided using GO terms enrichment analysis.

## Introduction

In higher eukaryotes, alternative splicing (AS) is an important process contributing to protein diversity. Different splicing events include exon skipping, alternative 3'/5' splice site usage, mutually exclusive exon usage and intron retention. When a gene is spliced differently between two conditions, the relative abundance of a transcript can shift, irrespective of a change in the overall expression of a gene. In differential transcript usage (DTU), the distribution of transcript abundance changes, irrespective of a change in gene expression. DTU can have various functional consequences, e.g., switching from a protein-coding transcript to a non-coding transcript or switching between transcripts with different functions. Isoform switching is a special case of DTU, where we focus on DTU of the most abundant transcript [1].

DTU has been studied in various diseases. For example, Vitting-Seerup and Sandelin et al. showed that 19% of genes with multiple transcripts involve functional isoform switches in cancer [1]. Parkinson disease-related gene candidates were implicated in DTU, but these were not detected as differentially expressed genes, highlighting the importance of this type of analysis [2]. Since many tools have been proposed for DTU analysis, a key question is which tool to choose for a particular analysis.

To address this question, previous benchmark analyses have compared different workflows for DTU detection. Focusing on plant systems, Liu et al. compared differential splicing detection tools in simulated and real datasets [3]. Differential splicing events were simulated based on the changes indicated by relative transcript abundance in each gene using the Flux simulator [4]. DEXSeq and DSGSeq performed well with simulated data with area under the ROC curve (AUC) around 0.8 [5,6]. To evaluate the performance of tools in detecting novel, i.e. not previously annotated, events, the authors also performed DTU analysis with a truncated annotation. Cufflinks performed best with acceptable precision (0.9) and recall (0.7) on the *de novo* splicing events discovery . DICAST, a docker-integrated alternative splicing benchmark tool, allows users to compare splicing-aware mapping tools and splicing event detection tools on simulated and real data sets [7,8]. Similarly, a large-scale study by Jiang et al. focused on event-based tools applied to simulated datasets [9]. However, only tools that detect and quantify splicing events in one condition are included in the pipeline. In Merino et al., differential splicing tools were tested systematically in scenarios with differential splicing and/or differential transcript expression [10]. The authors concluded that DEXSeq [6] and LimmaDS [11] are the best tools for detecting DTU. However, the pipeline used the outdated tool TopHat [12] as aligner, whereas STAR [13] has been shown to perform better [14,15]. In a method paper by Love, Soneson and Patro, DEXSeq [6] and DRIMSeq [16] are used to perform DTU analysis [17]. These two papers only included five DTU tools, while we could currently find twelve tools for detecting DTU. The recent addition of new contenders motivated us to perform a comprehensive benchmark covering various experimental settings. In particular, we acknowledge a growing interest in single-cell DTU analysis, which has thus far not been covered in benchmarking analyses.

We further consider the challenging scenario of time series AS analysis, which was not previously covered in benchmark studies despite the importance of such analysis in recapitulating AS changes during development or in response to environmental changes. For example, time-dependent AS genes were detected in plants after exposure to cold temperatures, suggesting changes in night-to-day conversion and circadian control [18].

We compared twelve DTU detection tools, six of which had not previously been benchmarked. We utilized both simulated datasets and actual human transcriptomic datasets for this comparison. Our simulations covered various settings: in bulk settings, sequencing technology types (either single-end or paired-end), number of replicates (four or eight), and three background levels are considered. The term ‘background'' refers to the likelihood of a gene not exhibiting differentially expressed transcripts, with a higher probability indicating a greater number of genes without DTU events. Our primary focus in the results section is on paired-end data. In single-cell settings, the number of cells and two background levels are considered. While we anticipate that paired-end sequencing excels in transcript detection, some studies still employ cheaper single-end sequencing, e.g. in time-series analysis, to support a larger sample number [19]. Understanding the performance of single-end data in DTU analysis is therefore crucial. In each bulk scenario, we simulated three scenarios contributing to transcript changes. We further categorized the results based on smaller or larger fold changes and the number of isoforms in a gene to understand the impact of different features on DTU detection. We simulated single-cell datasets to evaluate single-cell DTU tools such as DTUrtle and satuRn [20,21]. Additionally, we explored the qualitative differences between static pairwise comparisons and time series DTU analyses.

For paired-end sequencing, edgeR, DEXSeq, and LimmaDS emerged as top-performing tools [6,11,22]. In the context of single-end sequencing, we recommend DEXSeq and DSGseq [5]. LimmaDS was robust in detecting different types of DTU events (Figure 1). For time series data, Spycone was particularly effective in identifying biologically relevant events throughout the progression of a SARS-Cov-2 infected cell line. In single-cell data, satuRn has a better performance than DTUrtle. Taken together, this analysis provides a comprehensive view of the current state of DTU analysis in various scenarios.

## Methods

### Simulation

Our simulation process is similar to the one proposed by Merino et al., i.e. we used RSEM (v1.3.3) to simulate single-end and paired-end data using estimated abundances that are inferred from sequencing model parameters from real datasets and reference transcriptome [10]. The rsem-calculate-expression function estimates model parameters from real datasets [23]. The function collects statistics from the dataset, including the number of reads, the number of reads aligned to multiple and unique loci, the read and fragment length distribution and the quality score distribution.

The single-end data model parameters are estimated from GSE157490 [19], a cell line dataset with SARS-Cov2 infection sequenced at 100M reads. The paired-end data used to learn parameters is GSE162562 and GSE190680, which is also a dataset of patients with SARS-Cov2 infection sequenced at 100M reads [24,25]. Each dataset is simulated with 50 million reads, which is the minimum depth to robustly detect DTU [26] and 100 million reads.

Baseline transcript expression levels are taken from the SARS-CoV2 datasets. Next, we adjust the transcript counts to generate simulated data for three DTU scenarios (Figure 1). Since changes in transcript expression and DTU are confounded by changes in the overall expression of a gene, we consider both effects together. First, for each gene, we consider a random fold change ranging from 2 to 5 between conditions. The transcript ratios are generated using a Dirichlet distribution, which describes the probabilities of k categories given a density distribution with k dimensions. This approach is ideal for simulating transcript ratios as the sum of the vectors is 1. k represents the number of transcripts in a gene, with each transcript being assigned an expression value. The higher the probability associated with transcript i, the higher the expression value. To have a higher statistical power for detecting DTU transcripts, we simulate DTU transcripts with higher expression level. The following formula shows the simulation of expression value for each transcript i in condition j. :

$$transcript\ expression\ (i,j)\ =\ baseline\ gene\ expression\ *\ fold\ change\ *\ transcript\ ratio$$

Note that for the baseline (e.g. a control), the fold change is 1, whereas for the condition of interest we consider the random fold change. For creating replicates with measurement noise, we compute the expression values using a negative binomial distribution, where the dispersion for each transcript is estimated using DESeq2:

$$transcript\ expression\ of\ replicate\ x\ (i,j)\ =\ negative\ binomial\ \sim(\ transcript\ expression,\ 1/dispersion)$$

We split the genes across scenarios to obtain a mixture for the final data set. In theory, we could consider data sets that only consider individual scenarios, but this would not result in a realistic data set for evaluating the tools. In the next step, we must modify the transcript ratios according to the scenarios we consider. For scenario S1, only a single transcript of the gene changes expression. For scenario S2, more than two transcripts are subject to changes in relative abundance. In scenario S3, the relative abundance of two transcripts is swapped, signifying an isoform switch event.

We considered three background levels with an increasing fraction of genes whose expression remains unchanged: 0, 0.1, and 0.5. The modified transcript results are then used for the simulation. The rsem-simulate-reads command is used for the simulation. RSEM reference is generated with the human genome GRCh38, theta0 parameter, noise proportion to the background is set to 0.1.
In our study, we simulated a total of four conditions, incorporating various parameter combinations (Table 1). 100M reads are simulated only with four replicates and background level 0.5.

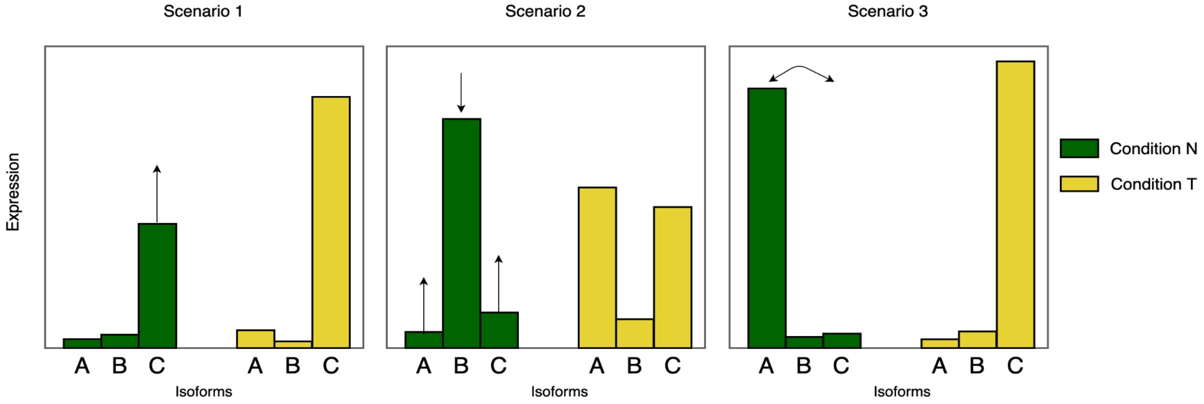


Figure 1. Scenario 1 (S1) are events that have one differentially expressed transcript, which are usually detected as differentially expressed genes as well. S2 are events that change the distribution of the transcript abundance between conditions. This change can involve multiple transcripts. S3 are events when the abundance is redistributed between two transcripts, it is called isoform switch (IS). The arrows indicate the changes represented in each scenario.

Table 1. Different simulated data generated

| Type | replicates | depth | background |
|---|---|---|---|
| paired-end | 4 | 50M, 100M | 0, 0.1, 0.5 |
| paired-end | 8 | 50M | 0, 0.1, 0.5 |
| single-end | 4 | 50M, 100M | 0, 0.1, 0.5 |
| single–end | 8 | 50M | 0, 0.1, 0.5 |

**Single cell simulation**

We used the same simulation workflow for a single cell, except the model parameters for RSEM are learned from a demultiplexed Smart-seq2 dataset derived from human cells [27]. We adapted our simulation method for single-cell transcript counts. In particular, we used the *–single-cell-prior* parameters in RSEM to model sparse matrices. RSEM seems to perform better in simulating Smart-seq2 dataset (Figure S2). To test the methods with simple single-cell data, we used two cell types with a population of 900 to simulate. We simulated balanced datasets with 20, 50, 100, 200, 500 and 1000 cells containing the same amount of cells for each cell type. We simulated DTU events in single-cell with two background levels - 0 and 0.1. Background level here also means the percentage of expressed genes that will stay unchanged between the two cell types.

**Differential transcript usage methods**

We performed an in-depth literature search in databases PubMed and Google Scholar with keywords such as “differential transcript usage”, “differential isoform usage”, “isoform switch”. We selected publications that describe a novel method for DTU analysis for bulk transcriptomics data. Iso-DOT is excluded due to long runtime (>20 days without parallelization) [28], rSeqDiff is excluded due to not supporting replicates [29] and IUTA is excluded due to incompatibility with STAR output [30].

Tools that can detect DTU can be categorised into exon/junction-centric (JunctionSeq [31], seqGSEA [32], DSGSeq [5]) and transcript-centric (DEXSeq [6], DRIM-Seq [16], DTUrtle [20], iso-KTSP [33], satuRn [21], NBSplice [34], edgeR [22], LimmaDS [11]), as well as assembly-based (Cufflinks/cuffdiff [35]).

Exon-centric tools like DEXSeq (v1.24.0) and JunctionSeq (v1.5.4) use a generalized linear model to analyze differences in exon and splice junction usage. DEXSeq, originally designed for exon counts, also works with transcript counts. It uses a formula to compare conditions and identifies significant genes based on adjusted p-values. JunctionSeq also uses adjusted p-values for gene evaluation. DSGSeq (v0.1.0) compares exon counts between conditions using negative binomial statistics. At the same time, seqGSEA (v1.36.0) combines DSGseq and DESeq methods, using exon counts and a rank-based strategy to output p-values for differential transcript usage (DTU) genes. Transcript-centric tools include DRIMSeq (v1.24.), which uses a dirichlet-multinomial model for transcript abundance analysis, and DTUrtle (v1.0.2), which adds extra filtering steps and uses StageR for more accurate gene-level false discovery rate correction. Iso-KTSP (v1.0.3) identifies transcript pairs that differentiate conditions, scoring them based on expression frequency. satuRn (v1.4.2) models transcript counts using a quasi-binomial model, and Cufflinks/cuffdiff (v2.2.1) aligns transcripts de novo and analyzes differential expression. For gene significance, tools using adjusted p-values consider values below 0.05 as significant. Iso-KTSP and DSGseq, which don't provide p-values, use cutoffs of 0.5 and 5 based on author recommendations. These tools are listed in Table 2.

Table 2. All DTU tools published after 2010.

| Tool | Implementation | Year | Reference | Principle idea | Excluded |
|---|---|---|---|---|---|
| DEXSeq | R | 2012 | [6] | Negative binomial generalized linear model to model transcript counts | |
| DRIMSeq | R | 2016 | [16] | Use dirichlet-multinomial model to model relative abundance of transcript | |
| seqGSEA | R | 2014 | [32] | Negative binomial models in DSGSeq and DESeq | |
| DTUrtle | R | 2021 | [20] | Dirichlet-multinomial model from DRIMSeq | |
| JunctionSeq | R | 2016 | [31] | Negative binomial generalized linear model to model junction counts | |
| NBsplice | R | 2020 | [34] | Negative binomial generalized linear model to model transcript counts, and test with a linear hypothesis | |
| satuRn | R | 2021 | [21] | Quasi-binomial generalized linear model to model transcript counts | |
| limmaDS | R | 2013 | [11] | Apply linear model to detect transcript changes | |
| Cuffdiff2 | C++ | 2012 | [35] | Use a poisson model to estimate changes in transcript counts | |
| iso-KTSP | Java | 2014 | [33] | Classify transcripts to condition specific based on the change of transcript abundance | |
| DSGSeq | R | 2013 | [5] | Negative binomial statistics to detect transcript changes | |
| edgeR | R | 2010 | [22] | Negative binomial generalized linear model to model transcript counts | |
| IUTA | R | 2014 | [30] | Estimate transcript usage, followed by testing DTU under Aitchison geometry | Incompatible with STAR |
| IsoDOT | R | 2015 | [28] | Estimate transcript usage with a penalized regression method | Long run time |
| rSeqDiff | R | 2013 | [29] | Apply linear poisson model to estimate transcript counts | Not supporting replicates |
| Time series tools | | | | | |
| TSIS | R | 2017 | [36] | detection and characterization of isoform switches for time series data | |
| Spycone | python | 2023 | [37] | detection and characterization of isoform switches for time series data | |

**Differential transcript usage methods for time series data**

There are two time series tools for detection of isoform switches in time series data: TSIS [36] and Sypcone [37]. TSIS and Spycone are the tools that detect switch points between transcript pairs and

calculate adjusted p-values based on the replicates. Then, it applies filters to select features like switching probabilities (i.e. the ratio of samples that has a higher relative abundance in one transcript than the other) , the difference of transcript expression before and after the switch. In Spycone, additional metrics are calculated such as event importance and domain difference. To detect DTU genes in TSIS, the following filtering metrics are used by default: 1) probability of switching > 0.5, 2) difference of expression before and after switching > 1, 3) p-value < 0.05, 4) correlation coefficient > 0.5. For the usage of Spycone, DTU genes are filtered with default parameters: 1) difference of relative abundance before and after switch > 0.2, 2) adjusted p-value < 0.05, 3) dissimilar correlation > 0.5 and 4) event importance > 0.3. For DEXSeq, we used a log-likelihood test with a reduced model. For LimmaDS, each time point is treated as a factor in multiple conditions.
We used clusterProfiler R package to perform GO term enrichment analysis [38].

**Preprocessing and quantification**

STAR (v2.7.8a) is used to align the reads to the genome (GRCh38 v107). We used STAR, which is currently among the best tools for splice-aware alignment [7]. Salmon (v1.7.0), kallisto (v0.44.0), RSEM and Cufflinks (v2.2.1) are used for transcript quantification. HTSeq (v2.0.1) is used to quantify exon counts for the input of seqGSEA. The workflow is shown in Fig. 2.

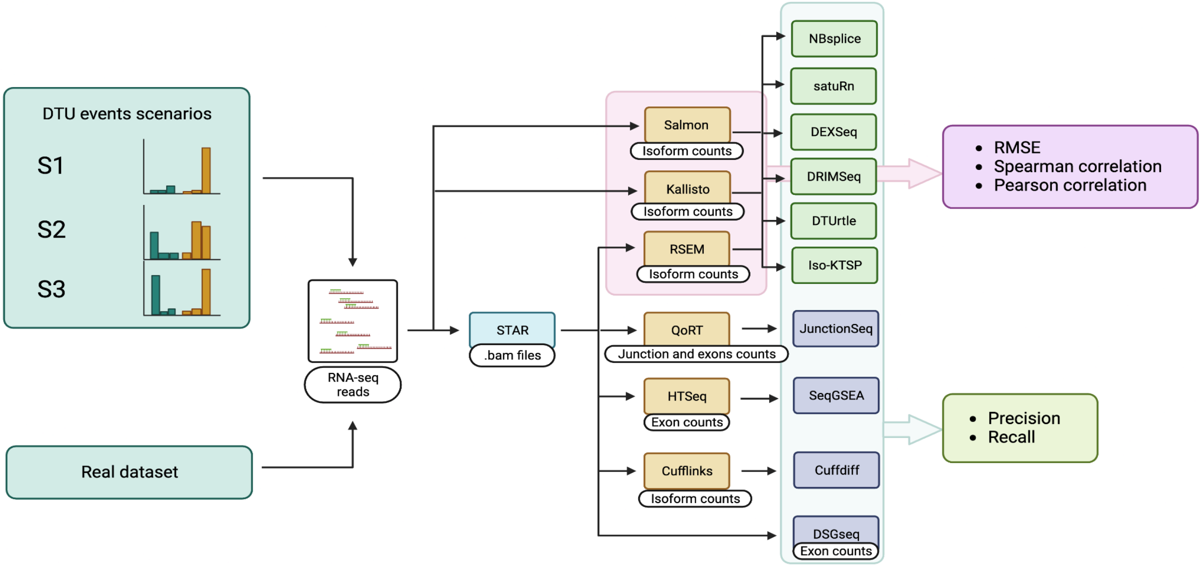


Figure 2. Analysis workflow of the methods.
Different scenarios are simulated to evaluate the performance of the tools. STAR is used to map RNA-seq reads and the resulting bam files are used by RSEM, QoRT, HTSeq and cufflinks to generate transcript counts, junction and exons counts. Salmon and Kallisto directly use RNA-seq reads in fastq format to generate transcript counts. Transcript counts derived from Salmon, Kallisto and RSEM are compared by calculating root mean squared error (RMSE), Spearman and Pearson correlation. We generate DTU detection results from NBsplice, satuRn, DEXSeq, DRIMSeq, DTUrtle, and Iso-KTSP using transcript counts from Salmon, Kallisto and RSEM. Other DTU tools use the corresponding count tables. All DTU detection results are compared by calculating precision and recall compared to the simulated ground truth.

**Evaluation of performance**

To evaluate the effectiveness of the tools on simulated data, we employed precision and recall as our performance metrics. We obtained a list of genes identified as having DTU, based on an adjusted p-value of less than 0.05, or surpassing a specific threshold in the case of iso-KTSP and DSGseq. These genes are positive. True positives (TP) are those among the positive genes that also appear in the simulated ground truth, while the rest are categorized as false positives (FP). Conversely, genes that are simulated with DTU but remain undetected by the tools are labeled as false negatives (FN). For stratification analysis, each event scenario is calculated separately. For different DTU events scenarios, P are genes simulated with the corresponding scenario. The formulas for calculating precision and recall are as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1\,score = \frac{2TP}{2TP + FP + FN}$$

To evaluate the performance in a real dataset, a prostate tumor dataset (GSE222260) is used. The dataset consists of 10 normal tissues and 20 prostate carcinoma tissues.

**Data availability**

The RNA-seq data of prostate tumor dataset is available from GSE222260 [39] and data from patients with SARS-Cov-2 infection is available from GSE162562 and GSE190680 [24,25]. Both datasets were preprocessed and analyzed as mentioned above. The time series RNA-seq data of SARS-Cov-2 infection data was obtained from GSE157490 [19]. The data is processed as described in [40].

**Code availability**

Code for simulation, analysis and plot generation are available at https://github.com/yollct/diffIsoUsage_benchmark under the terms of the GNU General Public License, Version 3.

## Results

### Overall performance of DTU detection

In this study, we evaluated twelve DTU tools using simulated datasets that incorporated various scenarios, including single-end or paired-end data, four or eight replicates, and three distinct background levels.

The datasets included three DTU scenarios for differentially expressed genes. We assessed precision, recall and F1 scores for each scenario based on the significant results obtained, as detailed in Table 1. For tools providing adjusted p-values, a threshold of 0.05 was used to identify positive results, while for iso-KTSP and DSGseq, the thresholds were set at 0.8 and 5, respectively. The transcript counts from each simulated dataset were compared against a ground truth outlined in the supplementary file (Figure S1).

Figure 3 presents the metrics for the DTU detection tools on simulated data with four and eight replicates. Generally, an increase in recall was observed with eight replicates. DRIMSeq, DTUrtle, and JunctionSeq demonstrated comparable performances. NBSplice showed the highest precision overall, even with a background of 0.5 in four replicates, though its recall was low. LimmaDS achieved the highest recall (>0.2) in both sets of replicates. However, iso-KTSP's precision decreased to 0.4 as background increased. In contrast, DEXSeq and satuRn showed less impact on precision at a background of 0.5 (Figure 3A). When evaluating with F1 scores, iso-KTSP appears to be the best performing tool despite low recall (Figure 3B).

Figure S4 illustrates the results for single-end data, where all tools exhibited low recall in four replicates, possibly due to the lack of additional information from the paired-end protocol. This limitation can be mitigated with more replicates (eight). edgeR and DEXSeq achieved the best precision in four and eight replicates, respectively. Notably, DEXSeq combined with Kallisto showed promising results in both sequence types (Figure S5). Similar to paired-end data, DEXSeq with

Kallisto excelled in both four and eight replicates. However, LimmaDS displayed low recall in all cases (close to 0), and its precision significantly dropped to 0.4 when the background was set at 0.5 (Figure S12).

To investigate the effect of increasing sequencing depth, we simulated 100M reads for 4 replicates at 0.1 background. Most tools have increased recall. LimmaDS has the most improvement (+0.1), but precision dropped. DSGseq has increased precision and decreased recall (Figure S19).

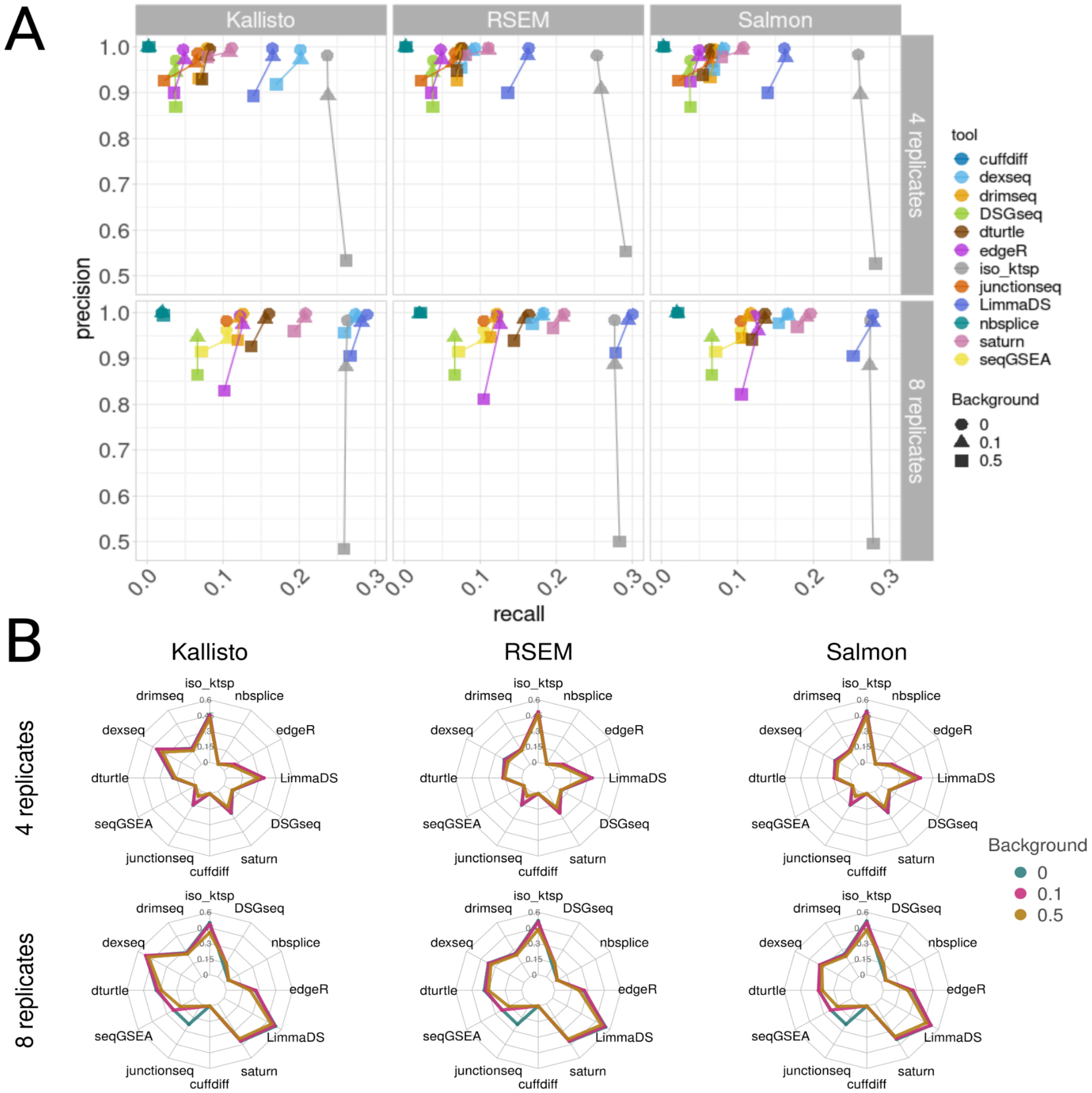


Figure 3. Metrics plot of all combinations of quantification tools and DTU methods from paired-end data with 4 replicates (upper-row) and 8 replicates (lower-row). A) Precision and recall plot. B) Radar plots showing the corresponding F1 scores.

**Performance of tools in different stratifications**

Figure 4 presents the F1 scores for various event types based on data quantified by Salmon. The results from Kallisto and RSEM quantifications are provided in the supplementary figures (Figures S6-8). The ranking of tools according to F1 scores are the same, except DEXSeq with kallisto quantification has the best performance. Additionally, we have stratified the genes according to the number of transcripts (Figure S3-5) and the extent of the fold change (Figure S9-11). F1 scores are determined using the ground truth for each specific category. While iso-KTSP demonstrates improved performance in Figure 4, this is not reflected in Figure 3A, where its tendency towards lower recall needs to be taken into account.

As expected, the performance of all tools improves with eight replicates. LimmaDS, in particular, achieves the highest F1 score and recall (>0.4) for S2 and S3 DTU events (Figure S9, 11). When eight replicates are used, satuRn maintains high precision, even with a high background (Figure S10). When paired with Kallisto, DEXSeq outperforms LimmaDS. Generally, an increase in fold change magnitude correlates with higher detection of events, thereby boosting the F1 scores. Nonetheless, iso-KTSP's F1 score remains consistent regardless of fold change. Moreover, satuRn demonstrates superior precision. Subsequently, we categorized genes into groups based on the number of transcripts, revealing a direct correlation between the number of transcripts and F1 scores. However, this categorization appears to have minimal impact on precision (Figure S4).
ev
In the analysis of single-end data, DEXSeq combined with Kallisto consistently exhibited superior performance in both four and eight replicate scenarios, as shown in Figure S13. Specifically, DEXSeq achieved an F1 score of 0.2 for S1 DTU events and approached 0.6 with eight replicates. iso-KTSP maintained steady F1 scores across all cases, similar to its performance in paired-end data, but it also identified false positives, as indicated in Figure S12. In contrast, LimmaDS showed weaker results in single-end data. For quantifications using RSEM and Salmon, seqGSEA outperformed others with an F1 score of 0.2 in eight replicates, while JunctionSeq led in the four-replicate category. Similar to paired-end data, a positive correlation between fold change and F1 scores was observed (Figure S14). However, the number of isoforms appeared to have a minimal impact on the results (Figure S15).

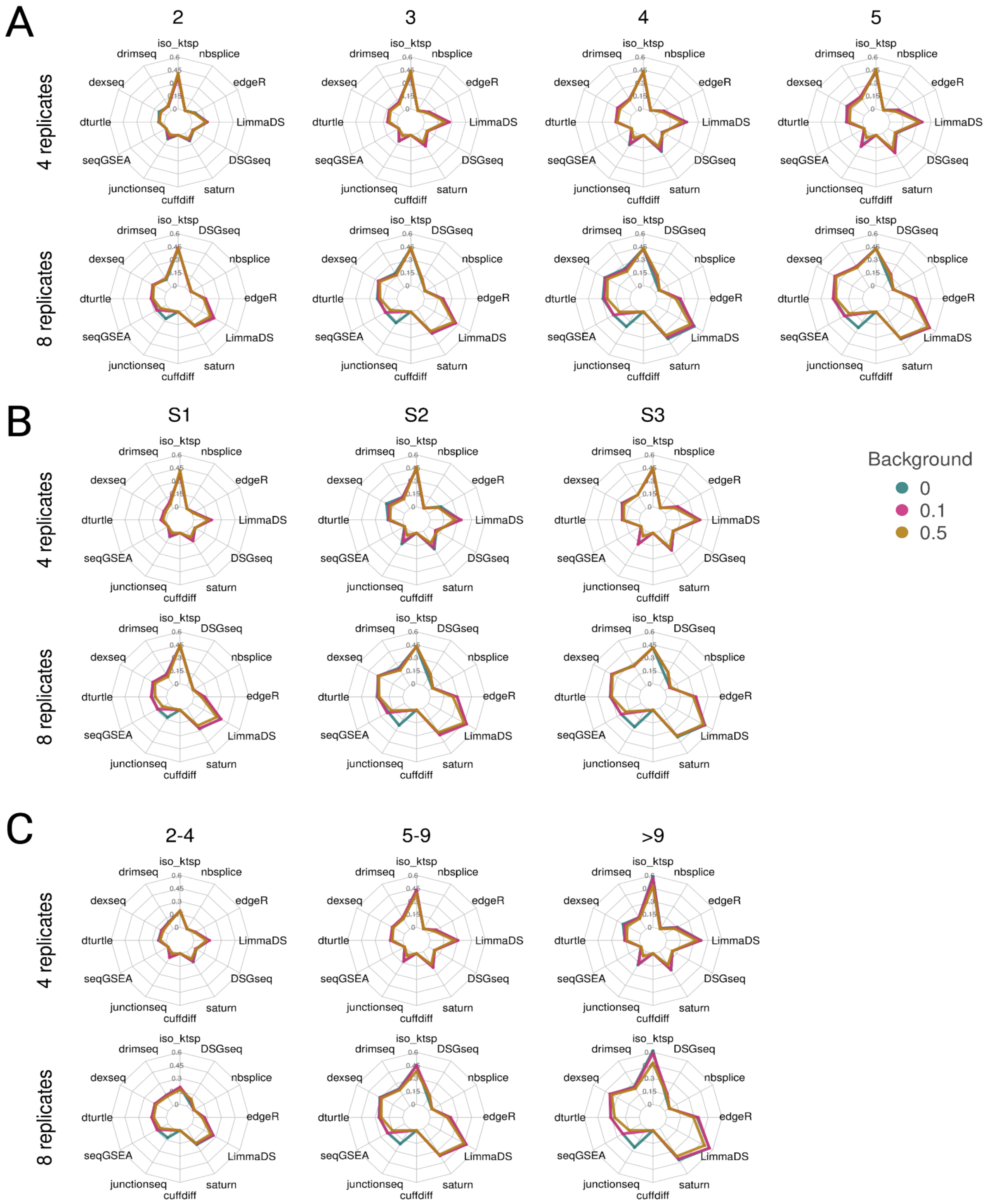


**Figure 4**. Radar plots showing F1 scores calculated after stratification for all combinations of quantification tools and DTU methods in paired-end data with 4 replicates (top row) and 8 replicates (bottom row). This result is derived from Salmon quantification. A) DTU events stratified by different scenarios B) DTU events stratified by fold change. C) DTU events stratified by number of transcripts.

### Benchmarking with a real transcriptome dataset

In our study, we utilized the paired-end prostate cancer dataset, previously used by Merino et al., to evaluate various DTU tools. We quantified sequencing reads using Salmon, Kallisto, and RSEM, applying all DTU tools except JunctionSeq, which was excluded due to its excessive runtime. Figure 5 illustrates the overlap in DTU gene sets detected by different tools, along with the F1 scores derived from our simulation analysis (Figure 5).

Iso-KTSP identified the highest number of DTU genes, but the F1 score was relatively low (0.25), indicating a higher rate of false positives in the simulated data. Conversely, LimmaDS, with a higher F1 score (0.4), detected only a limited number of genes in this dataset. We observed variations in the number of genes detected by different DTU tools depending on the quantification method used. For instance, DEXSeq identified the most genes with Kallisto counts, while edgeR and LimmaDS detected the most with Salmon counts, and satuRn found the most with RSEM counts. Notably, NBSplice did not detect any significant genes in Salmon and Kallisto, and only one gene in RSEM.

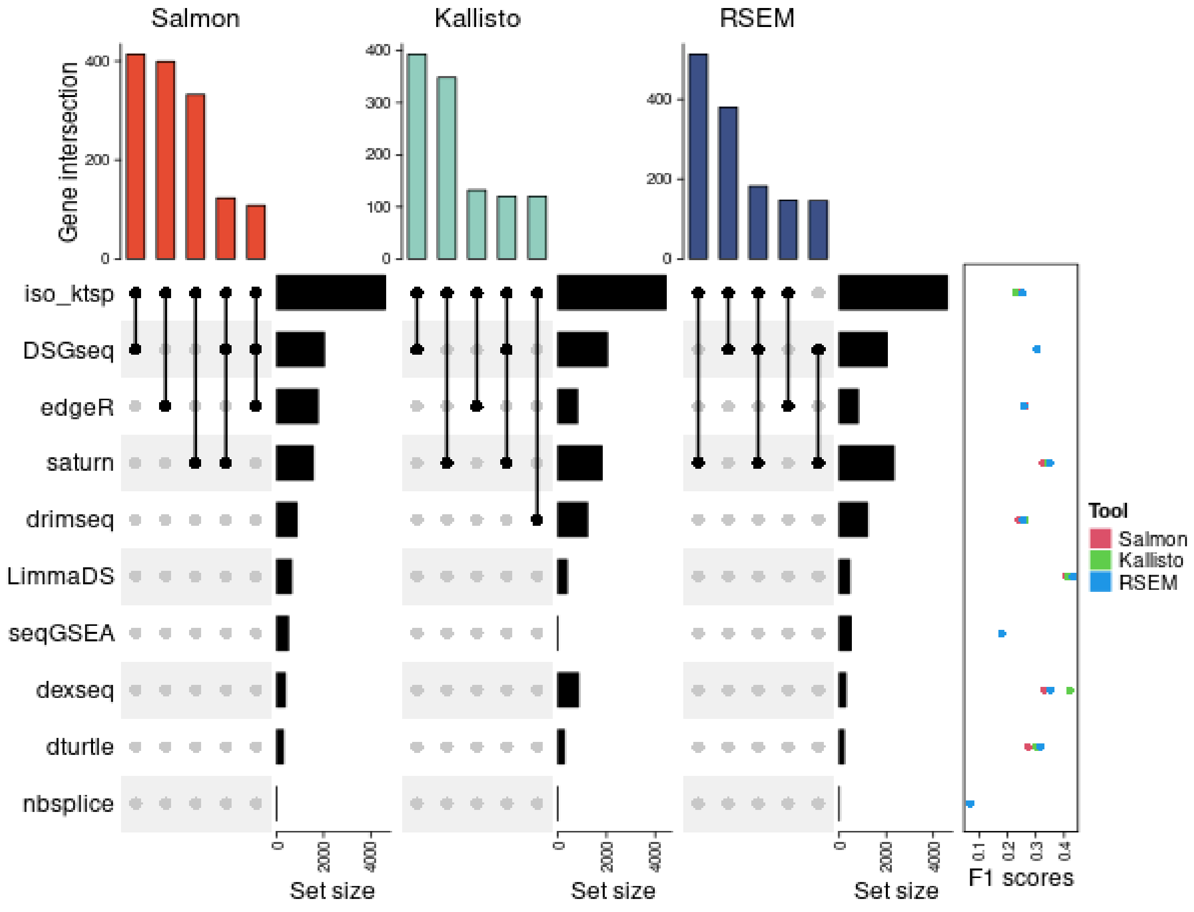


Figure 5. Upset plots showing the overlapping DTU genes found by each tool
Cancer dataset obtained from [39] are analyzed with the workflows. Each upset plot shows the number of DTU genes detected for each tool and the overlaps among them. The right dot plot shows the F1 scores obtained from the simulated dataset, with 8 replicates and a background of 0.5.

**Performance in single-cell data**

We assessed the performance of DTUrtle and satuRn on single-cell data using a simulated dataset created with the specified method. Each dataset comprised two cell types, each with an equal number of cells. We calculated precision and recall based on the simulated ground truth. Figure 6 illustrates the precision and recall as the number of cells in each cell type increases. Generally, we note a high precision (around 0.9) when there are 50 or more cells in each cell type. The recall, however, shows a gradual increase with the rising number of cells. SatuRn demonstrated a higher recall than DTUrtle, reaching a recall of 0.9 when each cell type had 500 cells, compared to DTUrtle's 0.73. As the background level rises, both precision and recall decline. SatuRn registered a precision of 0.83 and a recall of 0.88, whereas DTUrtle posted a precision of 0.93 and a recall of 0.69. However, with a higher background level, an increase in the number of cells resulted in a slight dip in precision (a decrease of 0.06 from 100 to 200 cells and a further decrease of 0.01 thereafter). In addition, we performed a pseudo-bulk analysis where transcript counts are aggregated into meta cells. The meta cells are then analyzed using methods designed for bulk data. The result shows that this approach does not improve precision and recall (Figure S17).

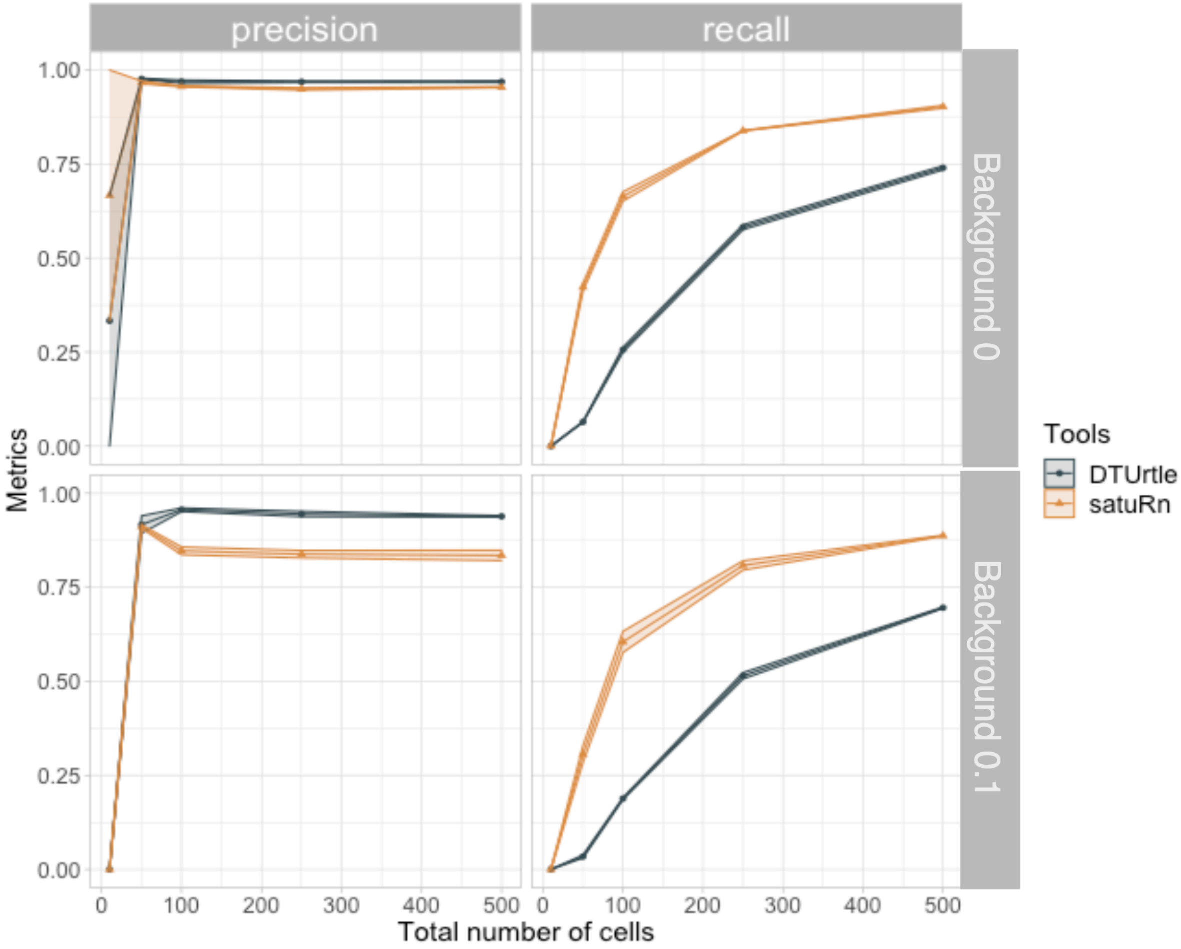


Figure 6. Precision and recall of satuRn and DTUrtle for single-cell simulation. Each plot consists of the result from a different number of cells (x-axis). Precision from simulation with background level 0 (top left). Recall from simulation with background level 0 (top right). Precision from simulation with background level 0 (bottom left). Recall from simulation with background level 0.1 (bottom right).

### Time series isoform switch analysis

Another method for inferring isoform switches is time series isoform switch detection. With time series data, we can extract dynamic changes in transcript usage. Here, we applied DEXSeq, LimmaDS, satuRn, TSIS and Spycone on time series single-end transcriptomic data for SARS-Cov-2 infection with eight time points and four replicates [19]. Figure 7 shows the results of the comparison. In all the comparisons, DEXSeq and LimmaDS have many overlapping DTU genes (1731). edgeR didn’t find any significant genes. Spycone and TSIS, which are specifically designed for time series data, have only a few overlaps (Figure. 7A). In GO terms biological processes enrichment of the significant DTU genes, the terms enriched in each of the sets are different, in which DEXSeq, LimmaDS and TSIS are enriched in generic cellular functions such as cadherin binding, ubiquitin-related terms etc (Figure. 7B). satuRn has similar but less enriched  The Spycone gene set is enriched in MHC protein complex binding, which is essential to adaptive immunity. For the term IgA bindings, IgA are found to be as part of the early humoral immune response to neutralize SARS-Cov-2 virus upon infection [41].

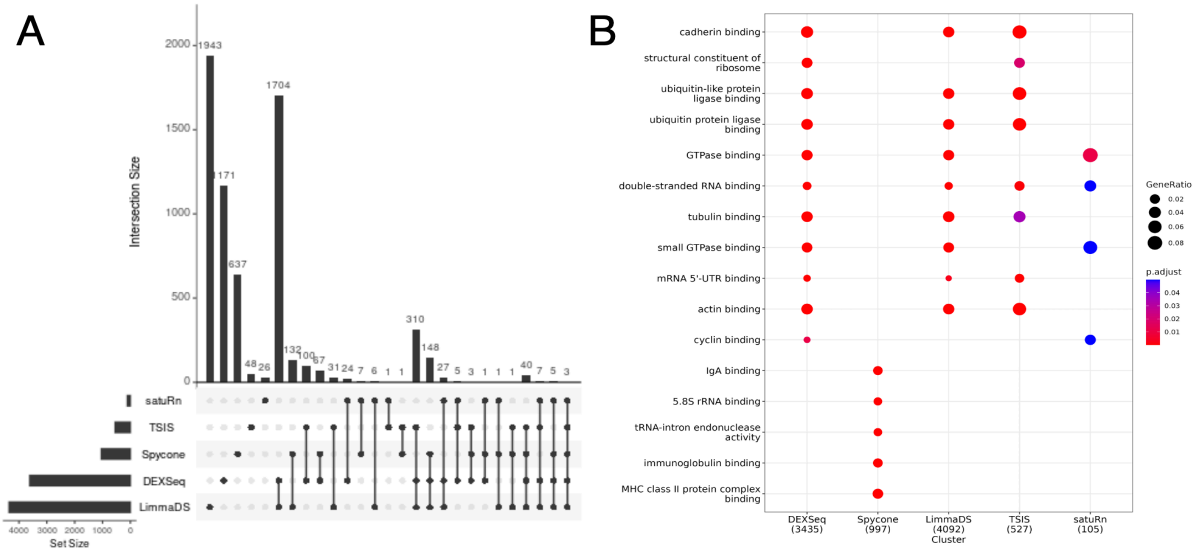


Figure 7. A) An upset plot showing the overlapping DTU genes detected by DEXSeq, TSIS, NBSplice, LimmaDS and Spycone time series isoform switch detection. B) Dotplot showing the enrichment terms from the DTU genes detected by all tools. NBSplice didn’t report significant genes.

## Discussion

In this analysis, we conducted a comprehensive benchmark study using both simulated data and real transcriptomics datasets, considering both static and dynamic conditions. Our simulation approach aimed to mimic real datasets, including the presence of stochastic noise. To achieve this, we simulated three background levels: 0, 0.1, and 0.5, representing the proportion of expressed genes that remain unchanged between two conditions. Additionally, we simulated both paired-end and single-end RNA-Seq data with four and eight replicates.

Our simulation approach contrasts with that of Merino et al. [10], where the authors specifically assigned abundance to each transcript between two conditions to simulate differential transcript abundance. Instead, we adopted a dirichlet-multinomial model to simulate abundance based on transcript probabilities. The Dirichlet-multinomial model effectively manages transcript expression as multivariate count data, accommodating overdispersion to more accurately reflect real-world data scenarios. Furthermore, while Merino et al. focused solely on simulating DTU events where two transcripts change in abundance and evaluated them using DTU tools, we simulated events with more than two transcripts change in abundance (S2) and assessed their performance with DTU tools.

In the simulated scenarios, we observed that single-end data exhibited lower recall (Figure S12). This decrease in recall can be attributed to the limitations of single-end sequencing, which cannot capture both ends of a cDNA fragment, leading to a reduced probability of observing junction reads. Consequently, single-end data has lower power for detecting the correct transcript. In our simulations, we found that single-end data could detect approximately 10,000 transcripts, while paired-end sequencing detected around 30,000 transcripts at a depth of 50M reads (Figure S18). However, despite its higher sensitivity, paired-end sequencing faces challenges in transcript detection due to transcriptional noise stemming from transcriptional stochasticity [42]. Furthermore, short reads cannot accurately detect transcript counts due to the increased likelihood of mapping reads to multiple genomic locations. This limitation can be mitigated by long-read sequencing technology once it becomes more prevalent [43].

Another factor influencing our observations is the simulation methods employed. RSEM served as the simulation engine, and since it is one of the quantification methods used in our analysis, there may be a bias in favor of RSEM. In our simulation approach, we used the mean expression values for each transcript from a real transcriptome dataset. Additionally, our simulation approach utilized a Dirichlet-multinomial model to generate transcript abundances, which might favor tools that also utilize the Dirichlet-multinomial model for modeling, such as DRIMseq and DTUrtle.

In our findings, we noted that tools employing a generalized linear model, such as NBSplice, DEXSeq, satuRn, JunctionSeq, and edgeR, as well as those utilizing a Dirichlet-multinomial model, such as DRIMSeq and DTUrtle, typically exhibited superior precision. Conversely, the only tool employing a linear model, LimmaDS, demonstrated higher recall. As for other approaches employed, they generally exhibited low performance.

For evaluating the performance of different DTU tools, we utilized precision, recall and F1 score (Table 2). For transcript-level analysis, we recommend using paired-end sequencing, as DTU analysis with single-end data captures less transcript information. When working with fewer than four replicates in paired-end data, we suggest using RSEM or Salmon as the quantification tool. Both tools perform similarly, but Salmon offers better runtime efficiency. If recall is a top priority, consider LimmaDS. If there are fewer replicates, DEXSeq could be a better choice. For more replicates, consider edgeR. When single-end sequencing is the only option, we also provided a guideline. For prioritizing recall, DSGseq can be employed. If there are fewer replicates, edgeR could be a better choice. For more replicates, consider DEXSeq. In addition, all tools' performance is affected by the fold change and the number of transcripts. As the fold changes and the number of transcripts increases, recall of the tools generally increases. Most prominently, the precision of the tools decreases due to the greater number of S1 of DTU events found. S1 events are likely to be regulated by transcription factors rather than splicing. In our analysis, the tool with higher recall in S2 and S3 events and lower recall in S1 events is LimmaDS.

In our time series case study, we utilized both pairwise comparison tools that accommodate time series data and dedicated time series tools on a single-end time series transcriptome dataset from SARS-Cov-2 infected human cells. Interestingly, each tool identified different genes and few overlaps were observed. We observed that the pairwise comparison tools identified a substantial number of genes, exceeding 3000, whereas the time series tools reported fewer genes. Among the time series tools, only 48 genes were commonly identified. The Spycone method predominantly highlighted switching transcripts with high expression levels by calculating the event importance metrics. While TSIS found a lot of low expressed transcripts based on previous findings [37]. Furthermore, while Spycone focuses on detecting IS events, our simulation study revealed that both LimmaDS and DEXSeq do not differentiate between DTU scenarios. These observations suggest the importance of distinguishing S1 events, as doing so can yield unique insights. Even though the time series data is derived from single-end sequencing, it is shown that there are quantitative and qualitative differences while applying pairwise comparison tools and time series tools.

Greater effort is required to differentiate between DTU events scenarios, especially since these events are often mixed with varying degrees of change in real-world situations. In future work, we could leverage additional metrics from Spycone and apply them as filtering criteria to the results generated by pairwise comparison tools.

Exploring differential transcript usage (DTU) in single-cell data presents a fascinating avenue of study. Analyzing single cells offers a deeper understanding of transcript usage heterogeneity across various cell types. This can potentially reveal alternative splicing patterns that contribute to the emergence of distinct cell types. Several tools, such as DTUrtle and satuRn, have been developed specifically for detecting DTU in single-cell data, alongside with bulk RNA-seq. Our analysis indicates that while satuRn boasts a higher recall, its precision is marginally lower than that of DTUrtle. For datasets with a larger number of cells (>250 per cell type), DTUrtle is recommended for those prioritizing precision. Conversely, satuRn is the preferred option for datasets with fewer cells (<250 per cell type) and where recall is prioritized.

However, it's worth noting that our analysis was based solely on a simulated dataset with an equal number of cells across two cell types. This balanced distribution does not always reflect real-world scenarios. Future analyses could benefit from simulating unbalanced datasets. Additionally, single-cell datasets often comprise more than just two cell types. As such, tools capable of comparing multiple cell types are more desirable. For instance, Acorde is designed to pinpoint co-DTU across several cell types [44]. Exploring DTU variations across pseudo-time within single-cell data presents a compelling direction for future research. However, single-cell transcript analysis faces technical challenges in obtaining accurate transcript counts. Ongoing developments in single-cell transcript analysis technologies suggest a promising future for understanding alternative splicing at the single-cell level [45,46].

| Type of sequencing | Number of replicates | Quantification | Priority | Recommended tools |
|---|---|---|---|---|
| paired-end | >8 | kallisto | precision | DEXSeq |
| | | | recall | LimmaDS |
| | <8 | | precision | DEXSeq |
| | | | recall | DEXSeq |
| paired-end | >8 | RSEM | precision | satuRn |
| | | | recall | LimmaDS |
| | <8 | | precision | satuRn |
| | | | recall | LimmaDS |
| paired-end | >8 | Salmon | precision | satuRn |
| | | | recall | LimmaDS |
| | <8 | | precision | satuRn |
| | | | recall | LimmaDS |
| single-end | >8 | Kallisto | precision | DEXSeq |
| | | | recall | DEXSeq |
| | <8 | | precision | DEXSeq |
| | | | recall | DEXSeq |
| single-end | >8 | RSEM | precision | seqGSEA |
| | | | recall | seqGSEA |
| | <8 | | precision | LimmaDS |
| | | | recall | LimmaDS |
| single-end | >8 | Salmon | precision | seqGSEA |
| | | | recall | seqGSEA |
| | <8 | | precision | LimmaDS |
| | | | recall | LimmaDS |

Table 2. Recommended workflow for bulk single-end and paired-end data in DTU analysis. This workflow is for counts derived from different transcript quantification methods.

**Conclusion:**

In this comprehensive benchmark study, we have rigorously evaluated various transcriptomics datasets under both static and dynamic conditions, utilizing a blend of simulated and real data. We observed a general trend where tools using generalized linear models or Dirichlet-multinomial models showed superior precision, while LimmaDS, which employs a linear model, demonstrated higher recall. This suggests that the choice of DTU tools should be tailored to the specific needs of the study, considering factors like the number of replicates, sequencing method (single-end or paired-end), and the prioritization of precision or recall. Our analysis of time series data revealed interesting insights into DTU. Tools like Spycone, designed for time series DTU detection, showed differences in functional outcome.

Looking ahead, there is a clear need for further differentiation between DTU events in complex real-world scenarios. Additionally, exploring DTU in single-cell data remains a promising avenue, albeit with technical challenges in obtaining accurate transcript counts. As single-cell transcript analysis technologies continue to evolve, they hold significant promise for advancing our understanding of alternative splicing at the single-cell level. Future research could benefit from simulating more diverse and unbalanced datasets, as well as focusing on tools capable of comparing multiple cell types and analyzing DTU variations across pseudo-time.

**Key points:**

- We performed an analysis involving a comprehensive benchmark study that used both simulated data and real transcriptomics datasets. It considered both static and dynamic conditions. We included the 12 DTU tools for static conditions. Our study suggests that tools using generalized linear models produce better precision and with linear models produce better recall.
- Based on the stratifications of the DTU genes, recall of most tools are positively affected by the fold changes and number of transcripts. Our results show that LimmaDS is better in detecting S2 and S3 events scenarios.
- We provided guidelines for performing DTU analysis for different sequencing types, considering the number of replicates. LimmaDS, edgeR and DEXSeq are better for paired-end sequencing. DSGseq and DEXSeq are better for single-end sequencing.
- We provided evidence that Spycone can detect IS that has a different biological interpretation to the condition of interest.
- For datasets containing more than 250 cells per cell type, DTUrtle is the suggested choice for those valuing precision. On the other hand, for datasets with less than 250 cells per cell type, satuRn is recommended when recall is of greater importance.

**Funding**

This work was supported by the Technical University Munich – Institute for Advanced Study, funded by the German Excellence Initiative. This work was supported in part by the Intramural Research Programs (IRPs) of the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). JB was partially funded by his VILLUM Young Investigator Grant nr.13154. Partly funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 422216132. This work was supported by the German Federal Ministry of Education and Research (BMBF) within the framework of the *e:Med* research and funding concept (*grants 01ZX1908A / 01ZX2208A* and *grants 01ZX1910D / 01ZX2210D*). This project has received funding from the European Union’s Horizon

2020 research and innovation program under grant agreement No 777111. This publication reflects only the author's view, and the European Commission is not responsible for any use that may be made of the information it contains.

## Author contributions

C.T.L. planned and carried out the analysis. T.D. performed the single-cell analysis. M.H. and L.W. preprocessed the SARS-CoV2 datasets. C.T.L., M.L., M.H., O.T., and J.B. wrote and reviewed the manuscript.

## References

1. Vitting-Seerup K, Sandelin A. The Landscape of Isoform Switches in Human Cancers. Mol. Cancer Res. 2017; 15:1206–1220
2. Dick F, Nido GS, Alves GW, et al. Differential transcript usage in the Parkinson's disease brain. PLoS Genet. 2020; 16:e1009182
3. Liu R, Loraine AE, Dickerson JA. Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. BMC Bioinformatics 2014; 15:364
4. Griebel T, Zacher B, Ribeca P, et al. Modelling and simulating generic RNA-Seq experiments with the flux simulator. Nucleic Acids Res. 2012; 40:10073–10083
5. Wang W, Qin Z, Feng Z, et al. Identifying differentially spliced genes from two groups of RNA-seq samples. Gene 2013; 518:164–170
6. Anders S, Reyes A, Huber W. Detecting differential usage of exons from RNA-Seq data. Nature Precedings 2012; 1–1
7. Fenn A, Tsoy O, Faro T, et al. Alternative splicing analysis benchmark with DICAST. bioRxiv 2022; 2022.01.05.475067
8. Manz Q, Tsoy O, Fenn A, et al. ASimulatoR: splice-aware RNA-Seq data simulation. Bioinformatics 2021; 37:3008–3010
9. Jiang M, Zhang S, Yin H, et al. A comprehensive benchmarking of differential splicing tools for RNA-seq analysis at the event level. Brief. Bioinform. 2023; 24:
10. Merino GA, Conesa A, Fernández EA. A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies. Brief. Bioinform. 2019; 20:471–481
11. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015; 43:e47
12. Kim D, Pertea G, Trapnell C, et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14:R36
13. Dobin A, Davis CA, Schlesinger F, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29:15–21
14. Engström PG, Steijger T, Sipos B, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat. Methods 2013; 10:1185–1191
15. Baruzzo G, Hayer KE, Kim EJ, et al. Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat. Methods 2017; 14:135–139
16. Nowicka M, Robinson MD. DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. F1000Res. 2016; 5:1356
17. Love MI, Soneson C, Patro R. Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification. F1000Res. 2018; 7:952
18. Calixto CPG, Guo W, James AB, et al. Rapid and Dynamic Alternative Splicing Impacts the Arabidopsis Cold Response Transcriptome. Plant Cell 2018; 30:1424–1444
19. Kim D, Kim S, Park J, et al. A high-resolution temporal atlas of the SARS-CoV-2 translatome and transcriptome. Nat. Commun. 2021; 12:5120
20. Tekath T, Dugas M. Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle. Bioinformatics 2021;
21. Gilis J, Vitting-Seerup K, Van den Berge K, et al. *satuRn*: Scalable analysis of differential transcript usage for bulk and single-cell RNA-sequencing applications. F1000Res. 2021; 10:374
22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 2010; 26:139–140
23. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011; 12:323
24. Lee HK, Knabl L, Pipperger L, et al. Immune transcriptomes of highly exposed SARS-CoV-2

asymptomatic seropositive versus seronegative individuals from the Ischgl community. Sci. Rep. 2021; 11:4243
25. Lee HK, Knabl L, Knabl L Sr, et al. Immune transcriptome analysis of COVID-19 patients infected with SARS-CoV-2 variants carrying the E484K escape mutation identifies a distinct gene module. Sci. Rep. 2022; 12:2784
26. Liu Y, Ferguson JF, Xue C, et al. Evaluating the impact of sequencing depth on transcriptome profiling in human adipose. PLoS One 2013; 8:e66883
27. Vento-Tormo R, Efremova M, Botting RA, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature 2018; 563:347–353
28. Sun W, Liu Y, Crowley JJ, et al. IsoDOT Detects Differential RNA-isoform Expression/Usage with respect to a Categorical or Continuous Covariate with High Sensitivity and Specificity. J. Am. Stat. Assoc. 2015; 110:975–986
29. Shi Y, Jiang H. rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. PLoS One 2013; 8:e79448
30. Niu L, Huang W, Umbach DM, et al. IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data. BMC Genomics 2014; 15:862
31. Hartley SW, Mullikin JC. Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. Nucleic Acids Res. 2016; 44:e127
32. Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing. Bioinformatics 2014; 30:1777–1779
33. Sebestyén E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 2015; 43:1345–1356
34. Merino GA, Fernández EA. Differential splicing analysis based on isoforms expression with NBSplice. J. Biomed. Inform. 2020; 103:103378
35. Trapnell C, Hendrickson DG, Sauvageau M, et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat. Biotechnol. 2013; 31:46–53
36. Guo W, Calixto CPG, Brown JWS, et al. TSIS: an R package to infer alternative splicing isoform switches for time-series data. Bioinformatics 2017; 33:3308–3310
37. Lio CT, Grabert G, Louadi Z, et al. Systematic analysis of alternative splicing in time course data using Spycone. Bioinformatics 2023; 39:
38. Wu T, Hu E, Xu S, et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation (Camb) 2021; 2:100141
39. Kannan K, Wang L, Wang J, et al. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc. Natl. Acad. Sci. U. S. A. 2011; 108:9172–9177
40. Lio CT, Louadi Z, Fenn A, et al. Systematic analysis of alternative splicing in time course data using Spycone. bioRxiv 2022; 2022.04.28.489857
41. Sterlin D, Mathian A, Miyara M, et al. IgA dominates the early neutralizing antibody response to SARS-CoV-2. Sci. Transl. Med. 2021; 13:
42. Varabyou A, Salzberg SL, Pertea M. Effects of transcriptional noise on estimates of gene and transcript expression in RNA sequencing experiments. Genome Res. 2020; 31:301–308
43. Berbers B, Saltykova A, Garcia-Graells C, et al. Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified Bacillus. Sci. Rep. 2020; 10:4310
44. Arzalluz-Luque A, Salguero P, Tarazona S, et al. Acorde unravels functionally interpretable networks of isoform co-usage from single cell data. Nat. Commun. 2022; 13:1828
45. Hagemann-Jensen M, Ziegenhain C, Chen P, et al. Single-cell RNA counting at allele and isoform resolution using Smart-seq3. Nat. Biotechnol. 2020; 38:708–714
46. Hahaut V, Pavlinic D, Carbone W, et al. Fast and highly sensitive full-length single-cell RNA sequencing using FLASH-seq. Nat. Biotechnol. 2022; 40:1447–1451

# 6 Discussion and Outlook

With the increased usage of large omics datasets, computational methods can facilitate analysis and our understanding of complex biological mechanisms and disease development. In the previous chapters, I showed that differential gene co-expression network analysis could extract potential snoRNA biomarkers involved in the etiology of Alzheimer's disease. On the other hand, to improve isoform switch detection in time series data, I developed the Spycone framework for transcript-level time series data by incorporating prior biological knowledge like a PPI network with computational analysis such as clustering, functional analysis, and splicing factors analysis. The goal is to decipher the underlying effects of the changes in transcriptome due to alternative splicing and the potential mechanism that leads to specific biological conditions. In the following sections, I will discuss the findings of each publication, the pitfalls of the methods used, and the results in the benchmark analysis for differential isoform usage analysis (unpublished). Finally, I will discuss the outlook and perspective of the technology and computational analysis in omics.

## 6.1 Discovery of small RNAs biomarkers in Alzheimer's disease mouse model

Tg4-42 mouse model expresses a wild-type N-truncated A$\beta$, which is highly abundant in patients' CA1 area of the hippocampus [168]. This protein is associated with symptoms like synaptic hyperexcitability, reactive astroglia, and reduction of glucose metabolism. Compared to the 5XFAD mouse model, Tg4-42 exhibits synaptic hyperexcitability and loss of glucose metabolism, accumulation of N-truncated A$\beta$, and neuronal loss without plaque formation. Tg4-42 mouse model is therefore considered to be a better option for studying the etiology of Alzheimer's disease.

In the snoRNA publication, I used a data-driven approach - differential gene co-expression network analysis - to extract potential small RNA biomarkers that contribute to developing the abovementioned phenotypes of the Tg4-42 mouse model. The intuition behind the method is to elucidate the small RNA pairs that change their co-expression behavior between wild-type and Tg4-42 mouse models. Instead of looking for changes for each small RNA independently, I aimed for a systematic view of the changes that indicate the rewiring of the underlying transcriptional regulatory mechanism. Using the centrality measures (see Methods), I extracted miRNAs and snoRNAs that have been associated with Alzheimer's disease: Mir30b [137], Mir598 [169], Mir99b [170], Mir106b [171], Mir346 [172], Mir181 [173] and Snord49a [174]. In contrast, others have not been linked to Alzheimer's disease and are considered potential novel biomarkers: Mir1981, Snord38A, Snord99. With the lack of prior knowledge of the functions of snoRNAs and the miRNAs, I used the databases for miRNA interaction (miRBD [143]) and snoRNA (snoDB [144]) to extract interactors (genes) and performed functional enrichment analysis with those genes. The results further provide evidence that the small RNA markers could be linked to Alzheimer's disease and can be further studied in neurodegenerative disease in humans. However, several limitations should be considered when using this method. The pitfalls of co-expression analysis are discussed in the following section.

## 6.2 Pitfalls of co-expression analysis

In biological studies, there are various approaches to network analysis, which can be broadly classified into two categories: those that utilize protein-protein interaction (PPI) as prior knowledge and those that use network reconstruction. In the publication of small RNA in an Alzheimer's mouse model, I employed a network reconstruction method to construct a co-expression network based on similarities in differential co-expression patterns across multiple samples between different conditions using DGCA. This approach was selected due to the lack of prior knowledge of miRNA and snoRNA, particularly the interaction information amongst the small RNAs. The resulting network consisted of nodes representing small RNAs, with edges weighted with differential z-scores of the correlation coefficients from two conditions, denoting the differential interactions between the wild-type and Tg4-42 mouse model. The DGCA method established edges between two nodes if their differential z-scores exceeded a predetermined threshold. Once the final network was constructed, we investigated connections that were specific to the targeted conditions.

Although the co-expression network is a promising approach for constructing context-specific networks, some limitations must be considered. These include potential confounding factors such as batch effects or technical variation, which can impact the network structure. Additionally, co-expression networks may only capture some functional relationships between genes/proteins, and other interaction types, such as post-translational modifications or protein-protein interactions, may be overlooked. The DGCA approach captures linear relationships due to the nature of the Pearson correlation. Linear methods can select strong relationships that are more likely to have direct interactions [175]. WGCNA (Weighted gene co-expression network analysis) allows the detection of gene modules that are highly interconnected in the network. On the other hand, GENIE3 (a tree-based regression method)is developed to capture non-linear relationships. However, WCGNA and GENIE3 consider only one condition as input and are recommended to use more samples (>20). In our case, using DGCA can directly compare the two conditions. Nevertheless, these network reconstruction methods remain valuable in biological studies and should be investigated in future work. The following are some guidelines for performing co-expression analysis [176, 177]:

- **Parameters :** The threshold of a correlation coefficient affects the topology of the final network. Carefully choosing the threshold, e.g., visualizing the distribution before setting a threshold, can improve the quality of the network.

- **False-positives :** As discussed above, RNA-sequencing data is noisy due to transcriptional noise and sequencing artifacts. Due to this reason, some edges of the co-expression network may arise from this noise. It is challenging to distinguish artifact edges from context-specific edges [178]. This could be improved by having a larger sample size (Ballouz et al. 2015 suggested >20 sample size) and a reasonable filtering threshold for expression level. Another source of false-positives arises from confounded variables: genes that covariate due to technical artifacts like RNA integrity number, mapping covariate, and GC bias. Using the principle component correction method can improve confounding effects that cause false positives [179].

- **Normalization:** Where spike-in Normalization is not available in Johnson et al. 2022, they compared different normalization methods affecting the accuracy of co-expression network reconstruction. In summary, between-sample Normalization (specifically Counts adjusted with TMM Factors (CTU) or Counts adjusted with Upper quartile Factors (CUF)) is recommended. While for comparing samples within a dataset, TPM is recommended [180].

## 6.3 Systematic analysis of alternative splicing with Spycone

In the publication of Spycone, I aimed to develop a novel IS detection algorithm and a framework for systematic analysis of time series data. IS events involve the re-distribution of isoform abundance over time. Current challenges in detecting IS events in RNA-seq data include the fact that most genes have multiple isoforms. For example, detecting IS events in genes with two isoforms is harder than in genes with ten isoforms. These genes also have multiple lowly expressed isoforms that could be noise. The Spycone algorithm aims to overcome these challenges by proposing novel metrics to filter relevant IS events. I proposed using event importance to select IS events that contribute to a higher abundance of overall gene expression.

I tested Spycone in two simulation models. The first model consists of only highly expressed isoforms in the IS events, while lowly expressed isoforms are switched in the second model. Spycone demonstrates high precision and significantly better recall in both models than TSIS, its only direct competitor. However, moderate recall, particularly in noisy conditions, indicates potential areas for methodological enhancement. In the simulation of the second model, we noted a decline in precision and recall. This result indicates that the algorithm needs further improvement to detect IS events involving lowly expressed isoforms. While Spycone identifies switches involving only two isoforms, real-world events may include more than two. Future developments should thus account for multiple-isoform switches to better address complex scenarios. In Spycone, a PPI network is used for network enrichment. Combining prior knowledge from the PPI network and transcriptomics data can facilitate biological interpretations and analysis. However, there are limitations to using the PPI network (see section 6.5). Spycone employs a weighted PPI network using the domain-domain interactions (DDI) (see method section 3.2.2), which is prone to selection bias. Selection bias occurs when genes are often chosen as a study target in many studies; these genes are usually hub genes (nodes that have a greater number of edges (interactions). However, a higher weight in this network implies greater confidence in an interaction, as it indicates more domains interacting between proteins. This approach of using weighted PPI, which prioritizes higher-confidence interactions, is believed to offset the potential selection bias.

Spycone is not limited to transcriptomics data alone; users can apply it to various time-series data like proteomics. Other general time series data analysis tools are TiCoNE [181] and moanin [182]. TiCoNE provides a graphical user interface in Cytoscape [183] that provides clustering and network enrichment using KeyPathwayMiner [184]. However, the tool is limited in reproducibility due to the use of a graphical user interface and lack of biological interpretation like functional enrichment analysis. In moanin, users can perform differential gene expression analysis in time series data, clustering, and functional enrichment analysis. However, k-means is the only clustering algorithm provided. We know from previous studies that the performance of the clustering algorithm is highly dependent on the data [185]. In Spycone, multiple clustering algorithms are incorporated from Scikit-learn and tslearn library. Functional enrichment and network enrichment analysis are incorporated with gprofiler [153] and DOMINO [152].

Time series proteomics analysis is getting popular in studying dynamic processes like osteogenic differentiation [186], COVID-19 progression [187], mouse embryonic development [188]. In proteomics, peptides are mapped to protein groups instead of genes/transcripts in transcriptomics, mainly to one protein isoform. This limits the use of isoform switch detection in Spycone and only allows downstream analysis at the gene level by bypassing the isoform switch detection step. Currently, users are required to provide a processed count matrix. Different proteomics technologies (label-free, TMT tags) require different normalization methods, which may challenge non-computational experts. To address this issue, we should make normalization processes more accessible and user-friendly for non-computational experts. This could involve the development of intuitive software interfaces that guide users through the normalization process, offering automatic detection of the proteomics technology used and suggesting appropriate normalization methods accordingly. There is no best type of data to study biological processes; however, transcriptomics currently has the advantage of looking beyond gene expression to alternative splicing of genes. However, Spycone is prone to noise in transcriptomics data, which indirectly affects the sensitivity of downstream analysis. These limitations are discussed in detail in the following sections.

## 6.4 Noise and stochasticity in transcriptomics

Transcriptomic data is inherently noisy. Gene expression is regulated by the expression of another gene, forming regulatory cascades. The transcription initiation process depends on time and space: it starts when transcription factors signal transcription and when subunits of the transcriptional complex are available. Transcription can be initiated by stochastic intrinsic noise when the above factors are met. As a result of transcription, the expression of the target genes will enhance the effect of the intrinsic noise. These gene expressions might be misunderstood as condition-specific markers if detected under specific conditions [189]. Though these products might not be related to the biological/disease condition of interest, they could have the advantage of the flexibility of handling sudden events like cell stress [190]. This stochasticity could arise from several sources: 1) stochastic molecular processes of biochemical substances and probabilistic collision (e.g., Brownian motion) [191], which drives the binding of transcription factors to gene promoters, and the initiation of mRNA and protein degradation. 2) Other cellular factors include the activity of ribosomes and polymerases, cell size, age, and cell-cycle stage. In addition, promoter noise can result in an observational phenomenon known as transcriptional bursting. The fluctuation of gene expression depends on the stochastic characteristic of the promoter, including the binding of regulatory elements and the affinity of the binding sites in the promoter. In RNA-seq, bursting can be reflected by observing high variability among genes. Transcriptional noise can increase the number of low-count genes and affect the accuracy of downstream analysis, such as differential expression analysis.

Another possible source of noise is technical artifacts. In library preparation, genomic DNA might contaminate the samples, or even during poly(A) enrichment, oligoT primers bind to Adenince-rich regions instead of poly(A) tails. This limitation is then also inherited in the Spycone method. The above events could produce false positives when aligning reads and isoforms expression quantification. During library preparation, genomic DNA contamination in RNA-seq samples must be thoroughly removed. This can be done using ribosomal RNA depletion or poly(A) enrichment and DNAse treatment. After sequencing, several measures can be taken to reduce false positives: 1) perform a thorough quality check of the transcriptomics data before alignment. For raw fastq files, quality reads are filtered based on sequencing quality indicated by phred scores [192], GC contents, overrepresented k-mers, and duplicated reads. After alignments, the unique mapping percentage is expected to be around 70-90% [193]. 2) Filter genes and transcripts with low read counts. 3) Use appropriate normalization methods to correct technical biases and experimental noise, such as normalization by sequencing depth or spike-in controls. 4) Utilize multiple replicates to increase the analysis's statistical power and reduce the effect of biological variability.

In short, using transcriptomics in Spycone requires careful consideration of biological and technical noise to ensure accurate analysis. The inherent stochasticity of gene expression, influenced by complex regulatory networks and cellular conditions, can be misleading if not adequately accounted for. Furthermore, technical artifacts introduced during sample preparation and sequencing can further complicate the analysis. Rigorous preprocessing steps are essential, including quality checks, appropriate normalization, and multiple replicates. Despite paying attention to the abovementioned measures, several best practices of RNA-seq should be followed to get the best out of the RNA-seq data [89, 194, 195].

## 6.5 Pitfalls of the usage of PPI networks

In Spycone, the PPI network is used as a prior network for network enrichment analysis. The aim is to extract functional modules for genes clustered together based on expression or splicing patterns. However, there are limitations to using PPI networks, most notably the presence of selection or study bias in PPI networks. Regarding network theory, most PPI networks exhibit a power-law distribution of the node degrees, resulting in sparse connectivity and a few highly connected hub nodes [196]. Recent studies show that several biases drive the PPI network to have a power-law distribution instead of biological motive [197], indicating that the phenomenon of power-law distribution might not be the result of biological reason. PPI networks are constructed based on experimental validation or prediction and are heavily influenced by research focus [121]. Genes already known to be involved in certain diseases, such as cancer, receive significant attention, resulting in an over-representation of edges in the network and a higher likelihood of being identified as hubs, perpetuating this bias. This selection bias can lead to the under-representation of other genes and pathways that may also be relevant to disease development and progression but have received less attention from the scientific community [198]. Another limitation of PPI networks is the lack of specificity regarding conditions. Interactions in normal conditions might not be present in cancer conditions, and vice versa. In addition, splicing affects the structural outcome of a protein. Protein isoforms might contain different sets of domains, which usually define the function and the interaction partners of the isoforms. PPI networks ignore this aspect entirely. DIGGER was developed to fill this gap [199] by incorporating domain-domain interaction information into PPI.

Despite these limitations, PPI networks remain valuable in biomedical research, particularly for identifying disease-related gene modules. To mitigate selection bias, researchers need to consider using multiple complementary methods for network construction and analysis, including integrating data from different sources, such as gene expression and functional annotation, and using unbiased sampling techniques to identify nodes and edges in the network. By doing so, researchers can broaden their perspective and uncover novel gene interactions and pathways relevant to disease. In addition, computational tools are developed to reduce selection bias and improve the quality of PPI networks. For example, AlphaFold2 [200] and AlphaFold-Multimer [201] are developed to accurately predict protein-protein interaction and protein complex. Constructing condition-specific networks, as in the first publication of small RNA in Alzheimer's mouse model, can also mitigate selection bias and non-context-specific edges.

## 6.6 Usage of annotations

In Spycone, I implemented a framework that requires different sources of annotations: the mapping of transcripts to the genome annotation, mapping of domain information for each transcript, domain-domain interaction, and protein-protein interaction, as well as splicing factor discovery with known splicing factors. Reference annotation is a valuable resource for identifying genes and transcripts in RNA-seq data analysis. However, this approach has limitations as it ignores the dynamic evolution of biological systems. One such limitation is that reference annotations may be derived from a limited set of cell types, leading to an underestimation of transcript diversity in different scenarios [202]. Additionally, reference annotations can overlook the complexity of alternative splicing. Moreover, due to the limitations of short-read sequencing, reconstructing transcripts that were present in the cell is even more challenging [203]. This limitation includes the generally low sequencing depth in most studies to effectively detect transcripts. Nevertheless, reference annotation remains useful for identifying transcripts in biological data. Researchers can use *de novo* assembly methods, such as Cufflinks, as a data-driven approach for transcript reconstruction to discover novel transcripts in short read data [156]. However, *de novo* assembly methods yield a subpar performance than those that use annotations due to lowly expressed RNAs and complex alternative splicing events [10]. Long-read sequencing can potentially facilitate the discovery of novel transcripts. However, the sequencing depth of the current state in long-read technology is too low to provide a high confidence level. In the computational aspects, long-read sequencing data can also be assembled with reference-based or reference-free tools. In reference-based tools (e.g., Bambu, FLAIR) perform well with the reference genome; however, using the reference genome might lose the potential of long-read sequencing to detect novel transcripts. On the other hand, it is challenging to perform reference-free assembly as seen in the benchmark [204].

## 6.7 Limitations and Guidelines for RNA-seq DTU analysis

In my comprehensive benchmark study, I evaluated twelve DTU tools using simulated and real transcriptomics datasets under static and dynamic conditions [91] (see chapter 5). The simulations aimed to replicate real datasets, including stochastic noise, by simulating three background levels (0, 0.1, and 0.5) to represent the proportion of genes unchanged between two conditions. We also simulated single-end and paired-end RNA-Seq data with four and eight replicates.

This simulation approach differed from Merino et al. (2019), who assigned specific abundances to each transcript between two conditions. I used a Dirichlet-multinomial model for simulating abundance based on transcript probabilities, which handles transcript expression as multivariate count data and accounts for overdispersion. Unlike Merino et al., who focused on DTU events involving two transcripts, I simulated events with more than two transcripts changing in abundance and evaluated them using DTU tools. This simulation method, mainly using RSEM as the simulation engine, might introduce a bias towards RSEM quantification. I used mean expression values from a real transcriptome dataset and a Dirichlet-multinomial model to generate transcript abundances, potentially favoring tools like DRIMseq and DTUrtle that use similar models. For the DTU tool performance evaluation, I used precision, recall, and F1 score.

In the simulations, single-end data showed lower recall, attributed to its inability to capture both ends of a cDNA fragment, thus reducing junction read observation. Single-end data detected about 10,000 transcripts, while paired-end sequencing detected around 30,000 transcripts at 50M read depth. However, paired-end sequencing also faces challenges in transcript detection due to transcriptional noise. Long-read sequencing technologies could address short reads' limitations in accurate transcript count detection in the future.

Tools using generalized linear models (e.g., NBSplice, DEXSeq, satuRn, JunctionSeq, edgeR) and Dirichlet-multinomial models (e.g., DRIMSeq, DTUrtle) generally showed higher precision than non-linear models. LimmaDS, using a linear model, demonstrated higher recall. However, further investigation is needed to determine whether this advantage is due to the potential bias from the simulation method. I recommend paired-end sequencing for transcript-level analysis, as it captures more transcript information than single-end data. For fewer replicates in paired-end data, RSEM or Salmon are suggested, with Salmon being time-efficient.

Based on the analysis, I have devised a guideline for conducting Differential Transcript Usage (DTU) analysis, as depicted in figure 6.1. This guideline aims to assist researchers in selecting the appropriate tools and strategies based on their specific experimental setups and analytical goals. Here are the key recommendations: DEXSeq, LimmaDS, and satuRn are better choices for paired-end data. When the data is quantified using Salmon and RSEM, satuRn is recommended when precision is favored, and LimmaDS when recall is favored. Kallisto works the best in general with DEXSeq. In single-end data, Kallisto works the best with DEXSeq as well. For quantification with RSEM and Salmon, LimmaDS is the best choice if there are fewer than eight replicates; otherwise, seqGSEA is used for more than eight replicates.

Each tool identified different genes with minimal overlap in the time series case study on SARS-Cov-2 infected human cells. Pairwise comparison tools (DEXSeq, NBSplice, LimmaDS, satuRn) identified over 3000 genes, while time series tools (Spycone, TSIS) found fewer. Spycone focused on high-expression switching transcripts, and TSIS identified many low-expressed transcripts. Spycone mainly focuses on finding genes with IS events that are highly expressed using the event importance metric. The resulting genes in each tool are performed gene set enrichment analysis. The result distinguished Spycone from other tools, providing a different view of the biological interpretation of the result.

In the single-cell analysis, I used RSEM to generate simulated single-cell data. In the preprint [91], the simulated dataset contains a balanced number of cells comparing two cell types. The number of cells in each cell type ranges from 10 to 500. Two background levels are simulated, indicating a gene's probability to remain unchanged between the two cell types. The simulated datasets are applied with DTUrtle and satuRn, where both tools' papers demonstrated the application to single-cell datasets. Essentially, both tools achieve high precision (close to 1) with 0 background. satuRn have decreased precisions (0.8) with

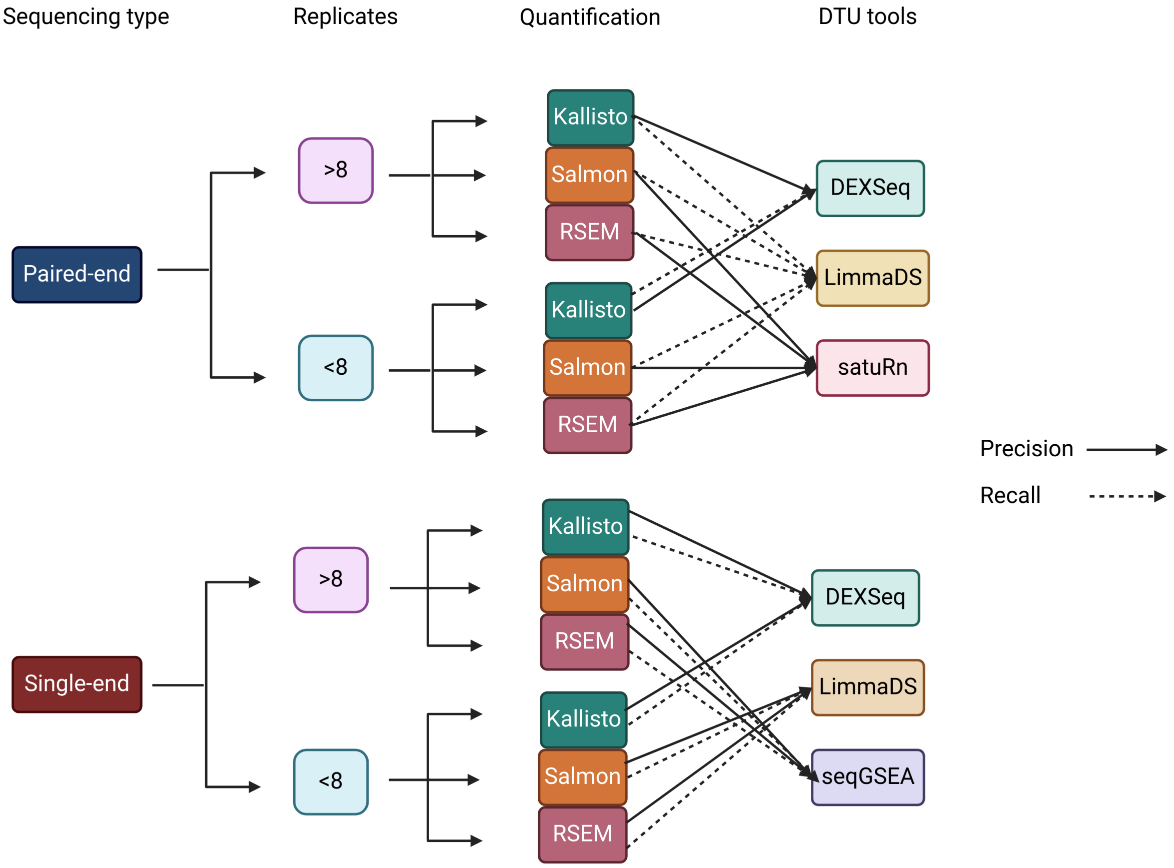


**Figure 6.1** Recommendation of DTU tools

For both paired-end and single-end data, according to the number of replicates and the quantification tool used, are shown with the recommended DTU tools. The solid lines and the dotted lines represent prioritizing over precision and recall, respectively. Source: own work

0.1 background [91]. One of the limitations of this simulation method is the correctness of using RSEM as a single-cell simulator. Even though RSEM provides a single-cell prior parameter to tackle zero-inflated data, it is not a standard tool in this scenario. Moreover, all single-cell datasets contain unbalanced cell types; this simulated dataset only illustrates a perfect scenario, which does not account for real-world data. To address these limitations, I used another tool, scDesign3, dedicated to single-cell data simulation. I simulated balanced and unbalanced datasets with 0.1 background (Figure 6.2. In a balanced dataset, two groups are compared. Each group contains from 50 to 700 cells. In the unbalanced dataset, I compared two to seven groups, each containing a random number of cells. DTUrtle and satuRn show similar precisions across both balanced and unbalanced data. satuRn has a slight decrease in precision in unbalanced datasets. In balanced datasets, satuRn and DTUrtle increase recall with an increased number of cells in each group. In unbalanced data, recall is generally lower than in balanced data. However, the number of groups does not influence the recall, except for seven groups. In this analysis, I have focused on the tools that detect DTU events in bulk and single-cell data. There are more single-cell-only tools that should be investigated in the future, like MARVEL [205], scQUNIT [206], SpliZ [207] and BRIE2 [208]. A limitation of DTUrtle and satuRn is that they can only compare two groups simultaneously. When comparing more groups in the unbalanced dataset, I combined the results from the pairwise comparisons for each combination of cell types. This can limit the findings of cell-type-specific DTU events. Moreover, more conditions should be considered in future simulations, such as library size (per cell), batch effects, and dropout rate, as well as with and without annotation.

## 6.8 Summary of the thesis

In this dissertation, I utilized co-expression analysis to build condition-specific gene-level networks. Through this network analysis, I identified potential miRNA and snoRNA biomarkers for the Tg4-42 Alzheimer's disease mouse model. I employed a differential co-expression network approach to detect small RNA pairs with shifting correlations, suggesting potential targets specific to Alzheimer's disease. To further explore the role of small RNAs in Alzheimer's disease, I conducted gene-set enrichment analysis. I used a small RNA interactors database to determine the potential functions of these dysregulated small RNAs. Notably, four out of five miRNAs with high centrality in the network are already linked to the molecular mechanisms of Alzheimer's disease. This suggests that the identified snoRNAs with high centralities may play a role in the disease's pathogenesis. Additionally, pathway enrichment analysis of snoRNA interactors revealed their involvement in Glycosaminoglycan biosynthesis, supporting the significance of these snoRNAs since Glycosaminoglycan is known to contribute to amyloid fibril formation.

In the second publication, I introduced Spycone. Spycone is a splice-aware time-course network enricher. Spycone addresses a gap in time-series analysis in alternative splicing and systems biology. This novel algorithm can identify genes that undergo isoform switches over a time series, outperforming the existing tool TSIS. Spycone offers downstream analysis such as clustering, functional enrichment, and network enrichment analysis, enabling a comprehensive examination of changes in alternative splicing patterns under different biological conditions. Spycone will be adapted for comparative analysis in pairwise conditions in future developments. I analyzed a time course transcriptomics dataset after SARS-Cov-2 infection as a use case. Clusters of genes are identified according to the patterns of total isoform usage over

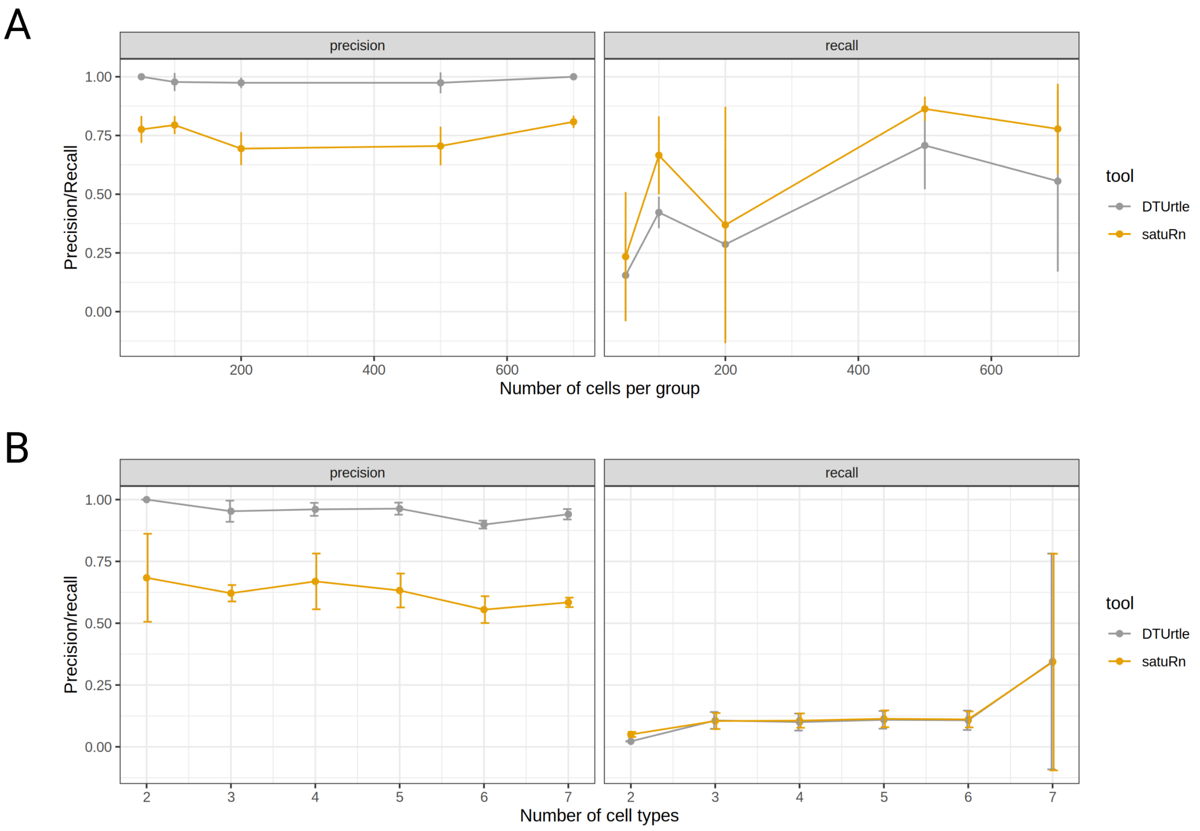


**Figure 6.2** DTU analysis benchmark in simulated single-cell data.

A. Precision and recall plots for balanced data with different number of cells per group (i.e. cell type). B. Precision and recall plots for unbalanced data with different number of groups compared. Source: own work time. These clusters are involved in the mitogen-activated protein kinase (MAPK) pathway, TLR 7/8 pathway, and pentose phosphate pathway. These pathways are known to be associated with SARS-Cov-2 infection. Furthermore, Spycone detected pathways that have yet to be characterized by association with SARS-Cov-2. These include kinesins, signaling by NTRKs, AXIN degradation, Hedgehog signaling, and 5-phosphoribosyl 1-diphosphate biosynthesis. In addition, network enrichment analysis with the DOMINO algorithm identified gene candidates affected by isoform switch events.

## 6.9 Outlook on sequencing technology

This thesis is grounded in short-read sequencing technologies. A notable limitation of these technologies is their inadequate detection of repetitive sequences, which are prevalent in intronic regions. However, detecting changes in these intronic regions is crucial for accurate splicing analysis. In contrast, long-read sequencing technologies like PacBio and Oxford Nanopore offer much longer reads, revolutionizing transcriptomics [209]. For example, PacBio's HiFi reads average around 1000 bp, aligning more closely with the average length of human transcripts than the 150 bp average of short-read technologies like Illumina.

Long-read sequencing offers several advantages for transcriptome analysis. It simplifies aligning and reconstructing original transcripts, as it can map reads spanning entire transcripts directly to the genome or assemble them into full-length transcripts. This is particularly beneficial for alternative splicing analysis, where short-read technologies struggle to differentiate between isoforms. Long-read sequencing data can also be adapted in the Spycone framework in the future. Long-read sequencing also enables the study of genes with highly repetitive regions, like transposable elements or introns, which are challenging to detect with short-read sequencing.

Furthermore, long-read sequencing is advantageous for studying complex genomes, such as those of viruses, by resolving structural variants and other genomic complexities. However, it has limitations. For instance, polymerase limitations in PacBio sequencing can affect read length and accuracy. Additionally, long-read sequencing tends to be more error-prone than short-read sequencing. Not to mention, the cost of long-read sequencing per sample is higher than that of short-read. Despite these challenges, advancements in sequencing technology and bioinformatics are making long-read sequencing a more prevalent choice for transcriptome analysis.

## 6.10 Outlook on splicing analysis

Splicing analysis can be broadly categorized into event-based and exon-/isoform-based analysis, as discussed in the introduction section. Event-based tools identify alternative splicing events and quantify the features (exon or splice site) with PSI values. Isoform-based tools rely on alignments of the sequencing reads to the reference transcriptome and quantifying the expression of the transcripts. Each approach provides a different view of the transcriptome; combining them will give us a fuller understanding of the impact of alternative splicing. Alternative splicing events found by event-based tools can be used to quantify transcripts that incorporate these events. Alternatively, detected splicing events can be cross-validated in the quantified transcript expression to check for consistency.

Another outlook is the time series aspects; performing time series analysis in event-based tools is challenging. There is no event-based method dedicated to time series data. One could independently analyze each time point and represent the changes over time using PSI values.

The advent of single-cell sequencing technologies gives us a whole new dimension in splicing analysis. Event-based and isoform-based methods for single-cell alternative splicing analysis have developed rapidly. One approach is to perform a splicing analysis along pseudo-time in single cells. Pseudo-time inference is a popular method to order single cells according to the developmental state. Investigating splicing events along the pseudo-time trajectory can bring new insights into the mechanistic view of the biological condition.

## 6.11 Outlook on omics analysis

In this dissertation, my approach primarily centered on analyzing transcriptomics data. However, it is essential to acknowledge that gene expression within our cells is influenced by many factors beyond transcription factors. This includes protein content, epigenetic markers, and other molecular components. Recent advances in omics analysis involve integrating multi-omics data, including proteomics, metabolomics, epigenomics, and spatial transcriptomics. Transcriptomics is a rapidly evolving field, particularly with the emergence of next-generation sequencing. It has widespread application in biological and biomedical research. Proteomics is the study of the proteome in a cell, tissue, or organism, but there are still challenges in understanding how transcriptome data can be mapped to proteome data. The challenges include genomic mutation, alternative splicing, post-transcriptional and post-translational modification. Some studies have little evidence that alternative splicing isoforms are found in proteomics [210]. This suggests that there is still a large gap in understanding the conversion from transcriptome to proteome. In addition, proteins can be located in multiple compartments in a cell, implying different functions and having multiple interaction partners [211].

Epigenomics provides several layers of information that can be incorporated, including DNA-protein interaction, chromatin modification, chromatin accessibility, and chromosome conformation. Integrating epigenomics can provide a better understanding of gene regulatory mechanisms. For example, ATAC-seq measures the openness of the chromatin regions, and accessible chromatin regions indicate a possibility of the presence of gene enhancers. Traditionally, chromatin accessibility and transcriptomics are profiled separately in different cells from the same population. The recent development of technologies allows simultaneous profiling of both in the same cells [212, 213]; this dramatically improves the robustness in integrative analysis of RNA-seq and ATAC-seq in single-cells. Simultaneous DNA methylation and transcriptome profiling can also reveal chromatin accessibility through measuring CpG methylation [214].

Spatial transcriptomics is another developing omics layer that allows studying biological systems at a subcellular level. Traditionally, spatial information is obtained through in-situ visualization, such as FISH (fluorescence *in situ* hybridization)-based methods, which use fluorescence-tagged oligonucleotide probes to bind to single mRNA molecules. For instance, smFISH, developed in 2008, uses fluorescence-tagged oligonucleotides probes to bind to single mRNA molecules [215]. Recently developed MERFISH+ can detect many mRNAs (up to 10000 genes) with high accuracy and efficiency [216]. A more direct method to capture spatial transcriptome is to dissect the specimens using cryosection and perform RNA-seq. However, this method requires highly specialized techniques, and the resolution highly depends on the tissue type. It also suffers from low throughput. A higher throughput method includes spatial labeling and in situ RNA capture. 10x Visium from 10x Genomics (originally Spatial Transcriptomics that was acquired by 10x Genomics in 2018) offers spatially barcoded RNA-seq method as well as histological staining and imaging of the tissue slide [217]. Slide-seq uses a different approach. Instead of spatial labeling on the tissue slide, the tissue is placed on top of barcoded beads and is then permeabilized to diffuse out the mRNA for sequencing [218].

Each of the technologies mentioned above is developing at a high speed. More stringent efforts should be applied to produce more reliable, high-quality, and reproducible data to make these applicable in clinical settings.

# Abbreviations

**RNA** Ribonucleic acid  
**DNA** Deoxyribonucleic acid  
**DTU** Differential Transcript Usage  
**tRNA** transfer ribonucleic acid  
**mRNA** messenger ribonucleic acid  
**rRNA** ribosomal ribonucleic acid  
**miRNA** micro ribonucleic acid  
**snoRNA** small nucleolar ribonucleic acid  
**snRNP** small nuclear ribonucleoproteins  
**snRNA** small nuclear ribonucleic acid  
**ESE** Exon splicing enhancers  
**ISE** Intron splicing enhancers  
**ESS** Exon splicing silencers  
**ISS** Intron splicing silencers  
**RRM** RNA-recognition motif  
**hnRNP** heterogeneous nuclear ribonucleoproteins  
**PTB** polypyrimidine-tract-binding protein  
**UTR** Untranslated region  
**NMD** Nonsense-mediated decay  
**ER** endoplasmic reticulum  
**EST** Expressed Sequence Tag  
**NGS** Next-generation sequencing  
**TGS** Third-generation sequencing  
**ddNTP** dideoxyribonucleoside triphosphates  
**ZMW** zero-mode waveguides  
**CCS** circular consensus sequencing  
**ONT** Oxford Nanopore Technologies  
**ASIC** application-specific integrated circuit  
**SAGE** serial analysis of gene expression  
**CAGE** cap analysis of gene expression  
**UMI** unique molecular identifiers  
**RPKM** reads per kilobase per million reads  
**TPM** transcripts per million  
**TMM** trimmed-mean of M-values

**RLE** relative log-expression

**NB** negative binomial

**DTE** differential transcript expression

**DGE** differentially expressed genes

**IS** isoform switch

**GLM** generalized linear models

**PCA** Principal component analysis

**tSNE** t-Distributed Stochastic Neighbor Embedding

**UMAP** Uniform Manifold Approximation and Projection

**DSG** differentially spliced genes

**PCST** ast prize-collecting Steiner tree

**PPI** protein-protein interactions

**DDI** domain-domain interactions

# Bibliography

[1] Félix Vázquez-Chona, Bong K Song, and Eldon E Geisert. “Temporal changes in gene expression after injury in the rat retina”. In: *Investigative ophthalmology & visual science* 45.8 (2004), pp. 2737–2746.

[2] Ray Zhang et al. “A circadian gene expression atlas in mammals: implications for biology and medicine”. In: *Proceedings of the National Academy of Sciences* 111.45 (2014), pp. 16219–16224.

[3] Ricardo N Ramirez et al. “Dynamic gene regulatory networks of human myeloid differentiation”. In: *Cell systems* 4.4 (2017), pp. 416–429.

[4] Charlotte Delay, Wim Mandemakers, and Sébastien S Hébert. “MicroRNAs in Alzheimer’s disease”. In: *Neurobiology of disease* 46.2 (2012), pp. 285–290.

[5] Sébastien S Hébert et al. “Loss of microRNA cluster miR-29a/b-1 in sporadic Alzheimer’s disease correlates with increased BACE1/$\beta$-secretase expression”. In: *Proceedings of the National Academy of Sciences* 105.17 (2008), pp. 6415–6420.

[6] Wang-Xia Wang et al. “The expression of microRNA miR-107 decreases early in Alzheimer’s disease and may accelerate disease progression through regulation of $\beta$-site amyloid precursor protein-cleaving enzyme 1”. In: *Journal of Neuroscience* 28.5 (2008), pp. 1213–1223.

[7] Chunliu Huang et al. “A snoRNA modulates mRNA 3 end processing and regulates the expression of a subset of mRNAs”. In: *Nucleic acids research* 45.15 (2017), pp. 8647–8660.

[8] Wenbin Guo et al. “TSIS: an R package to infer alternative splicing isoform switches for time-series data”. In: *Bioinformatics* 33.20 (2017), pp. 3308–3310.

[9] Ravindra N Singh and Natalia N Singh. “Mechanism of splicing regulation of spinal muscular atrophy genes”. In: *RNA Metabolism in Neurodegenerative Diseases* (2018), pp. 31–61.

[10] Ruolin Liu, Ann E Loraine, and Julie A Dickerson. “Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems”. In: *BMC bioinformatics* 15.1 (2014), pp. 1–16.

[11] Gabriela A Merino, Ana Conesa, and Elmer A Fernández. “A benchmarking of workflows for detecting differential splicing and differential expression at isoform level in human RNA-seq studies”. In: *Briefings in bioinformatics* 20.2 (2019), pp. 471–481.

[12] Ralf Dahm. “Friedrich Miescher and the discovery of DNA”. In: *Developmental Biology* 278.2 (Feb. 2005), pp. 274–288. DOI: 10.1016/j.ydbio.2004.11.028. URL: https://doi.org/10.1016/j.ydbio.2004.11.028.

[13] J. D. WATSON and F. H. C. CRICK. “Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid”. In: *Nature* 171.4356 (Apr. 1953), pp. 737–738. DOI: 10.1038/171737a0. URL: https://doi.org/10.1038/171737a0.

[14] Francis Crick. “Central Dogma of Molecular Biology”. In: *Nature* 227.5258 (Aug. 1970). Number: 5258 Publisher: Nature Publishing Group, pp. 561–563. ISSN: 1476-4687. DOI: 10.1038/227561a0. URL: https://www.nature.com/articles/227561a0 (visited on 08/17/2022).

[15] Matthew Cobb and Nathaniel Comfort. “What Rosalind Franklin truly contributed to the discovery of DNA’s structure”. In: *Nature* 616.7958 (2023), pp. 657–660.

[16] Robert W. Holley et al. “Structure of a Ribonucleic Acid”. en. In: *Science* 147.3664 (Mar. 1965), pp. 1462–1465. ISSN: 0036-8075, 1095-9203. DOI: 10.1126/science.147.3664.1462.

[17] Katrien J. Baeyens, Hendrik L. De Bondt, and Stephen R. Holbrook. “Structure of an RNA double helix including uracil-uracil base pairs in an internal loop”. en. In: *Nature Structural Biology* 2.11 (Jan. 1995), pp. 56–62. ISSN: 1545-9985. DOI: 10.1038/nsb0195-56.

[18] T. CASPERSSON and JACK SCHULTZ. “Pentose Nucleotides in the Cytoplasm of Growing Tissues”. In: *Nature* 143.3623 (Apr. 1939), pp. 602–603. DOI: 10.1038/143602c0. URL: https://doi.org/10.1038/143602c0.

[19] Gerald F. Joyce. “RNA evolution and the origins of life”. In: *Nature* 338.6212 (Mar. 1989), pp. 217–224. DOI: 10.1038/338217a0. URL: https://doi.org/10.1038/338217a0.

[20] Eva Jablonka and Eörs Szathmáry. “The evolution of information storage and heredity”. In: *Trends in ecology & evolution* 10.5 (1995), pp. 206–211.

[21] Harry F. Noller. “Ribosomal Rna and Translation”. In: *Annual Review of Biochemistry* 60.1 (1991), pp. 191–227. DOI: 10.1146/annurev.bi.60.070191.001203.

[22] Scott C. Blanchard et al. “tRNA dynamics on the ribosome during translation”. In: *Proceedings of the National Academy of Sciences* 101.35 (Aug. 2004), pp. 12893–12898. DOI: 10.1073/pnas.0403884101.

[23] Douglas L. Black. “Protein Diversity from Alternative Splicing: A Challenge for Bioinformatics and Post-Genome Biology”. English. In: *Cell* 103.3 (Oct. 2000), pp. 367–370. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/S0092-8674(00)00128-8.

[24] Robin C. Friedman et al. “Most mammalian mRNAs are conserved targets of microRNAs”. eng. In: *Genome Research* 19.1 (Jan. 2009), pp. 92–105. ISSN: 1088-9051. DOI: 10.1101/gr.082701.108.

[25] Mariana Lagos-Quintana et al. “Identification of novel genes coding for small expressed RNAs”. In: *Science* 294.5543 (2001), pp. 853–858.

[26] Zissimos Mourelatos et al. “miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs”. In: *Genes & development* 16.6 (2002), pp. 720–728.

[27] Ahmet M Denli et al. “Processing of primary microRNAs by the Microprocessor complex”. In: *Nature* 432.7014 (2004), pp. 231–235.

[28] Sunny Sharma et al. “Specialized box C/D snoRNPs act as antisense guides to target RNA base acetylation”. In: *PLoS genetics* 13.5 (2017), e1006804.

[29] Paula J. Grabowski, Richard A. Padgett, and Phillip A. Sharp. “Messenger RNA splicing in vitro: An excised intervening sequence and a potential intermediate”. en. In: *Cell* 37.2 (June 1984), pp. 415–427. ISSN: 0092-8674. DOI: 10.1016/0092-8674(84)90372-6.

[30] Britta Wallmen, Monika Schrempp, and Andreas Hecht. “Intrinsic properties of Tcf1 and Tcf4 splice variants determine cell-type-specific Wnt/$\beta$-catenin target gene expression”. In: *Nucleic acids research* 40.19 (2012), pp. 9455–9469.

[31] Mathieu Gabut et al. “An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming”. In: *Cell* 147.1 (2011), pp. 132–146.

[32] Sophie C Bonnal, Irene López-Oreja, and Juan Valcárcel. “Roles and mechanisms of alternative splicing in cancer—implications for care”. In: *Nature reviews Clinical oncology* 17.8 (2020), pp. 457–474.

[33] James Dominic Mills and Michal Janitz. “Alternative splicing of mRNA in the molecular pathology of neurodegenerative diseases”. In: *Neurobiology of aging* 33.5 (2012), 1012–e11.

[34] Natsumi Ohsawa et al. “Alternative splicing of PDLIM3/ALP, for $\alpha$-actinin-associated LIM protein 3, is aberrant in persons with myotonic dystrophy”. In: *Biochemical and biophysical research communications* 409.1 (2011), pp. 64–69.

[35] A. Gregory Matera and Zefeng Wang. “A day in the life of the spliceosome”. en. In: *Nature Reviews Molecular Cell Biology* 15.22 (Feb. 2014), pp. 108–121. ISSN: 1471-0080. DOI: 10.1038/nrm3742.

[36] Tatsuya Suzuki, Hiroto Izumi, and Mutsuhito Ohno. “Cajal body surveillance of U snRNA export complex assembly”. In: *Journal of Cell Biology* 190.4 (2010), pp. 603–612.

[37] Saori Kitao et al. “A compartmentalized phosphorylation/dephosphorylation system that regulates U snRNA export from the nucleus”. In: *Molecular and cellular biology* 28.1 (2008), pp. 487–497.

[38] Clemens Grimm et al. “Structural basis of assembly chaperone-mediated snRNP formation”. In: *Molecular cell* 49.4 (2013), pp. 692–703.

[39] Judith E Sleeman and Angus I Lamond. “Newly assembled snRNPs associate with coiled bodies before speckles, suggesting a nuclear snRNP maturation pathway”. In: *Current Biology* 9.19 (1999), pp. 1065–1074.

[40] Saba Valadkhan. “Role of the snRNAs in spliceosomal active site”. In: *RNA Biol* 7.3 (2010), pp. 345–353.

[41] Hansen Du and Michael Rosbash. “The U1 snRNP protein U1C recognizes the 5 splice site in the absence of base pairing”. In: *Nature* 419.6902 (2002), pp. 86–90.

[42] Kristi L Fox-Walsh et al. “The architecture of pre-mRNAs affects mechanisms of splice-site pairing”. In: *Proceedings of the National Academy of Sciences* 102.45 (2005), pp. 16176–16181.

[43] Janine O Ilagan et al. “Rearrangements within human spliceosomes captured after exon ligation”. In: *Rna* 19.3 (2013), pp. 400–412.

[44] Alberto R Kornblihtt et al. “Multiple links between transcription and splicing”. In: *Rna* 10.10 (2004), pp. 1489–1498.

[45] Benoit Chabot et al. “An intron element modulating 5’splice site selection in the hnRNP A1 pre-mRNA interacts with hnRNP A1”. In: *Molecular and cellular biology* (1997).

[46] Russ P Carstens, Eric J Wagner, and Mariano A Garcia-Blanco. “An intronic splicing silencer causes skipping of the IIIb exon of fibroblast growth factor receptor 2 through involvement of polypyrimidine tract binding protein”. In: *Molecular and cellular biology* 20.19 (2000), pp. 7388–7400.

[47] Harald König, Helmut Ponta, and Peter Herrlich. “Coupling of signal transduction to alternative pre-mRNA splicing by a composite splice regulator”. In: *The EMBO journal* 17.10 (1998), pp. 2904–2913.

[48] Brad A Amendt, Zhi-Hai Si, and C Martin Stoltzfus. “Presence of exon splicing silencers within human immunodeficiency virus type 1 tat exon 2 and tat-rev exon 3: evidence for inhibition mediated by cellular factors”. In: *Molecular and cellular biology* (1995).

[49] The UniProt Consortium. “UniProt: the Universal Protein Knowledgebase in 2023”. In: *Nucleic Acids Research* 51.D1 (Nov. 2022), pp. D523–D531. ISSN: 0305-1048. DOI: 10.1093/nar/gkac1052.

[50] *Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes*. http://genesdev.cshlp.org/content/28/22/2498.short. (Accessed on 01/27/2023).

[51] Frank A Laski et al. “Characterization of tRNA precursor splicing in mammalian extracts.” In: *Journal of Biological Chemistry* 258.19 (1983), pp. 11974–11980.

[52] Johannes Popow et al. “Analysis of orthologous groups reveals archease and DDX1 as tRNA splicing factors”. In: *Nature* 511.7507 (2014), pp. 104–107.

[53] Derek B. Scott et al. “An NMDA Receptor ER Retention Signal Regulated by Phosphorylation and Alternative Splicing”. en. In: *Journal of Neuroscience* 21.9 (May 2001), pp. 3063–3072. ISSN: 0270-6474, 1529-2401. DOI: 10.1523/JNEUROSCI.21-09-03063.2001.

[54] Andrea Thiele et al. “AU-rich elements and alternative splicing in the -catenin 3UTR can influence the human -catenin mRNA stability”. en. In: *Experimental Cell Research* 312.12 (July 2006), pp. 2367–2378. ISSN: 0014-4827. DOI: 10.1016/j.yexcr.2006.03.029.

[55] Mainul Hoque et al. “Analysis of alternative cleavage and polyadenylation by 3 region extraction and deep sequencing”. In: *Nature methods* 10.2 (2013), pp. 133–139.

[56] Shivashankar H. Nagaraj, Robin B. Gasser, and Shoba Ranganathan. “A hitchhiker’s guide to expressed sequence tag (EST) analysis”. In: *Briefings in Bioinformatics* 8.1 (Jan. 2007), pp. 6–21. ISSN: 1467-5463. DOI: 10.1093/bib/bbl015.

[57] Fred Sanger and Alan R Coulson. “A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase”. In: *Journal of molecular biology* 94.3 (1975), pp. 441–448.

[58] Joachim Messing, Roberto Crea, and Peter H. Seeburg. “A system for shotgun DNA sequencing”. In: *Nucleic Acids Research* 9.2 (Jan. 1981), pp. 309–321. ISSN: 0305-1048. DOI: 10.1093/nar/9.2.309.

[59] John Eid et al. “Real-Time DNA Sequencing from Single Polymerase Molecules”. In: *Science* 323.5910 (Jan. 2009), pp. 133–138. DOI: 10.1126/science.1162986.

[60] Daniel R. Garalde et al. “Highly parallel direct RNA sequencing on an array of nanopores”. en. In: *Nature Methods* 15.33 (Mar. 2018), pp. 201–206. ISSN: 1548-7105. DOI: 10.1038/nmeth.4577.

[61] Liu Xu and Masahide Seki. “Recent advances in the detection of base modifications using the Nanopore sequencer”. en. In: *Journal of Human Genetics* 65.11 (Jan. 2020), pp. 25–33. ISSN: 1435-232X. DOI: 10.1038/s10038-019-0679-0.

[62] Valerio Costa et al. “Uncovering the Complexity of Transcriptomes with RNA-Seq”. In: *Journal of Biomedicine and Biotechnology* 2010 (June 2010). Ed. by Momiao Xiong, p. 853916. ISSN: 2314-6133. DOI: 10.1155/2010/853916.

[63] Darren J Day et al. “Identification of non-amplifying CYP21 genes when using PCR-based diagnosis of 21-hydroxylase deficiency in congenital adrenal hyperplasia (CAH) affected pedigrees”. In: *Human Molecular Genetics* 5.12 (1996), pp. 2039–2048.

[64] Lira Mamanova et al. “FRT-seq: amplification-free, strand-specific transcriptome sequencing”. In: *Nature methods* 7.2 (2010), pp. 130–132.

[65] Martin Barron and Jun Li. “Identifying and removing the cell-cycle effect from single-cell RNA-Sequencing data”. In: *Scientific reports* 6.1 (2016), p. 33892.

[66] Carlos F Buen Abad Najar, Nir Yosef, and Liana F Lareau. “Coverage-dependent bias creates the appearance of binary splicing in single cells”. In: *eLife* 9 (June 2020). Ed. by L Stirling Churchman and Patricia J Wittkopp, e54603. ISSN: 2050-084X. DOI: 10.7554/eLife.54603.

[67] Simone Picelli et al. “Full-length RNA-seq from single cells using Smart-seq2”. en. In: *Nature Protocols* 9.11 (Jan. 2014), pp. 171–181. ISSN: 1750-2799. DOI: 10.1038/nprot.2014.006.

[68] Michael Hagemann-Jensen et al. “Single-cell RNA counting at allele and isoform resolution using Smart-seq3”. en. In: *Nature Biotechnology* 38.66 (June 2020), pp. 708–714. ISSN: 1546-1696. DOI: 10.1038/s41587-020-0497-0.

[69] Michael Hagemann-Jensen, Christoph Ziegenhain, and Rickard Sandberg. “Scalable single-cell RNA sequencing from full transcripts with Smart-seq3xpress”. en. In: *Nature Biotechnology* 40.1010 (Oct. 2022), pp. 1452–1457. ISSN: 1546-1696. DOI: 10.1038/s41587-022-01311-4.

[70] Daniel Ramsköld et al. “Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells”. In: *Nature biotechnology* 30.8 (2012), pp. 777–782.

[71] Roger Volden and Christopher Vollmers. “Single-cell isoform analysis in human immune cells”. en. In: *Genome Biology* 23.1 (Dec. 2022), p. 47. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02615-z.

[72] Aziz M. Al'Khafaji et al. "High-throughput RNA isoform sequencing using programmable cDNA concatenation". en. In: (Oct. 2021), p. 2021.10.01.462818. DOI: 10.1101/2021.10.01.462818. URL: https://www.biorxiv.org/content/10.1101/2021.10.01.462818v1.

[73] Kevin P. McCormick, Matthew R. Willmann, and Blake C. Meyers. "Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments". In: *Silence* 2.1 (Feb. 2011), p. 2. ISSN: 1758-907X. DOI: 10.1186/1758-907X-2-2.

[74] Carsten A. Raabe et al. "Biases in small RNA deep sequencing data". In: *Nucleic Acids Research* 42.3 (Feb. 2014), pp. 1414–1426. ISSN: 0305-1048. DOI: 10.1093/nar/gkt1021.

[75] Simon Andrews et al. *FastQC: a quality control tool for high throughput sequence data. 2010.* 2017.

[76] Yoshinori Fukasawa et al. "LongQC: a quality control tool for third generation sequencing long read data". In: *G3: Genes, Genomes, Genetics* 10.4 (2020), pp. 1193–1196.

[77] Liguo Wang, Shengqin Wang, and Wei Li. "RSeQC: quality control of RNA-seq experiments". In: *Bioinformatics* 28.16 (2012), pp. 2184–2185.

[78] Fernando Garcıa-Alcalde et al. "Qualimap: evaluating next-generation sequencing alignment data". In: *Bioinformatics* 28.20 (2012), pp. 2678–2679.

[79] Derek W Barnett et al. "BamTools: a C++ API and toolkit for analyzing and managing BAM files". In: *Bioinformatics* 27.12 (2011), pp. 1691–1692.

[80] Alexander Dobin et al. "STAR: ultrafast universal RNA-seq aligner". In: *Bioinformatics* 29.1 (2013), pp. 15–21.

[81] Amit Fenn et al. "Alternative splicing analysis benchmark with DICAST". en. In: (Jan. 2022). DOI: 10.1101/2022.01.05.475067. URL: http://biorxiv.org/lookup/doi/10.1101/2022.01.05.475067.

[82] Bo Liu et al. "deBGA: read alignment with de Bruijn graph-based seed and extension". In: *Bioinformatics* 32.21 (2016), pp. 3224–3232.

[83] Rob Patro et al. "Salmon provides fast and bias-aware quantification of transcript expression". en. In: *Nature Methods* 14.44 (Apr. 2017), pp. 417–419. ISSN: 1548-7105. DOI: 10.1038/nmeth.4197.

[84] Naoki Nariai et al. "TIGAR2: sensitive and accurate estimation of transcript isoform expression with longer RNA-Seq reads". In: *BMC genomics* 15 (2014), pp. 1–9.

[85] Nicolas L. Bray et al. "Near-optimal probabilistic RNA-seq quantification". en. In: *Nature Biotechnology* 34.55 (May 2016), pp. 525–527. ISSN: 1546-1696. DOI: 10.1038/nbt.3519.

[86] Ryan Tewhey et al. "Microdroplet-based PCR enrichment for large-scale targeted sequencing". en. In: *Nature Biotechnology* 27.1111 (Nov. 2009), pp. 1025–1031. ISSN: 1546-1696. DOI: 10.1038/nbt.1583.

[87] Günter P. Wagner, Koryu Kin, and Vincent J. Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples". en. In: *Theory in Biosciences* 131.4 (Dec. 2012), pp. 281–285. ISSN: 1611-7530. DOI: 10.1007/s12064-012-0162-3.

[88] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *bioinformatics* 26.1 (2010), pp. 139–140.

[89] Ana Conesa et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome biology* 17.1 (2016), pp. 1–19.

[90] Michael I Love, Wolfgang Huber, and Simon Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2”. In: *Genome biology* 15.12 (2014), pp. 1–21.

[91] Chit Tong Lio et al. “Comprehensive benchmark of differential transcript usage analysis for static and dynamic conditions”. In: *bioRxiv* (2024), pp. 2024–01.

[92] Yafang Li et al. “RNA-seq analysis of differential splice junction usage and intron retentions by DEXSeq”. In: *PloS one* 10.9 (2015), e0136653.

[93] Michael I. Love, Charlotte Soneson, and Rob Patro. “Swimming downstream: statistical analysis of differential transcript usage following Salmon quantification”. eng. In: *F1000Research* 7 (2018), p. 952. ISSN: 2046-1402. DOI: 10.12688/f1000research.15398.3.

[94] Stephen W Hartley and James C Mullikin. “Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq”. In: *Nucleic acids research* 44.15 (2016), e127–e127.

[95] Stephen W Hartley and James C Mullikin. “QoRTs: a comprehensive toolset for quality control and data processing of RNA-Seq experiments”. In: *BMC bioinformatics* 16.1 (2015), pp. 1–7.

[96] Weichen Wang et al. “Identifying differentially spliced genes from two groups of RNA-seq samples”. In: *Gene* 518.1 (2013), pp. 164–170.

[97] Xi Wang and Murray J Cairns. “SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-Seq data integrating differential expression and splicing”. In: *Bioinformatics* 30.12 (2014), pp. 1777–1779.

[98] Malgorzata Nowicka and Mark D Robinson. “DRIMSeq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics”. In: *F1000Research* 5 (2016).

[99] Koen Van den Berge et al. “stageR: a general stage-wise method for controlling the gene-level false discovery rate in differential expression and differential transcript usage”. In: *Genome biology* 18.1 (2017), pp. 1–14.

[100] Tobias Tekath and Martin Dugas. “Differential transcript usage analysis of bulk and single-cell RNA-seq data with DTUrtle”. In: *Bioinformatics* 37.21 (2021), pp. 3781–3787.

[101] Endre Sebestyén, Michał Zawisza, and Eduardo Eyras. “Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer”. In: *Nucleic acids research* 43.3 (2015), pp. 1345–1356.

[102] Jeroen Gilis et al. “satuRn: Scalable analysis of differential transcript usagenbsp;for bulk and single-cell RNA-sequencing applications”. en. In: 10:374 (May 2021). DOI: 10.12688/f1000research.51749.1. URL: https://f1000research.com/articles/10-374.

[103] Simon Anders, Alejandro Reyes, and Wolfgang Huber. “Detecting differential usage of exons from RNA-seq data”. In: *Nature Precedings* (2012), pp. 1–1.

[104] Gabriela Alejandra Merino and Elmer Andrés Fernández. "Differential splicing analysis based on isoforms expression with NBSplice". In: *Journal of Biomedical Informatics* 103 (2020), p. 103378.

[105] Matthew E Ritchie et al. "limma powers differential expression analyses for RNA-sequencing and microarray studies". In: *Nucleic acids research* 43.7 (2015), e47–e47.

[106] Sreya Ghosh and Chon-Kit Kenneth Chan. "Analysis of RNA-Seq data using TopHat and Cufflinks". In: *Plant Bioinformatics: Methods and Protocols* (2016), pp. 339–361.

[107] Liang Niu et al. "IUTA: a tool for effectively detecting differential isoform usage from RNA-Seq data". In: *BMC genomics* 15.1 (2014), pp. 1–13.

[108] Wei Sun et al. "IsoDOT detects differential RNA-isoform expression/usage with respect to a categorical or continuous covariate with high sensitivity and specificity". In: *Journal of the American Statistical Association* 110.511 (2015), pp. 975–986.

[109] Yang Shi and Hui Jiang. "rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test". In: *PloS one* 8.11 (2013), e79448.

[110] Kristoffer Vitting-Seerup and Albin Sandelin. "IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences". In: *Bioinformatics* 35.21 (2019), pp. 4469–4471.

[111] Panthadeep Bhattacharjee and Pinaki Mitra. "A survey of density based clustering algorithms". In: *Frontiers of Computer Science* 15 (2021), pp. 1–27.

[112] Zheng Kuang et al. "High-temporal-resolution view of transcription and chromatin states across distinct metabolic states in budding yeast". en. In: *Nature Structural  Molecular Biology* 21.1010 (Oct. 2014), pp. 854–863. ISSN: 1545-9985. DOI: `10.1038/nsmb.2881`.

[113] Therese Sørlie et al. "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications". In: *Proceedings of the National Academy of Sciences* 98.19 (2001), pp. 10869–10874.

[114] Tom Ronan, Zhijie Qi, and Kristen M Naegle. "Avoiding common pitfalls when clustering biological data". In: *Science signaling* 9.432 (2016), re6–re6.

[115] Jun Wan et al. "Dynamic usage of alternative splicing exons during mouse retina development". In: *Nucleic acids research* 39.18 (2011), pp. 7920–7930.

[116] María José Nueda et al. "Identification and visualization of differential isoform expression in RNA-seq time series". In: *Bioinformatics* 34.3 (Feb. 2018), pp. 524–526. ISSN: 1367-4803. DOI: `10.1093/bioinformatics/btx578`.

[117] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists". In: *Nucleic Acids Research* 37.1 (Jan. 2009), pp. 1–13. ISSN: 0305-1048. DOI: `10.1093/nar/gkn923`.

[118] Aravind Subramanian et al. "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles". In: *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550.

[119] Zakaria Louadi et al. “Functional enrichment of alternative splicing events with NEASE reveals insights into tissue identity and diseases”. In: *Genome Biology* 22.1 (Dec. 2021), p. 327. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02538-1.

[120] Damian Szklarczyk et al. “STRING v10: protein–protein interaction networks, integrated over the tree of life”. In: *Nucleic acids research* 43.D1 (2015), pp. D447–D452.

[121] Rose Oughtred et al. “The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions”. In: *Protein Science* 30.1 (2021), pp. 187–200.

[122] Max Kotlyar et al. “Integrated interactions database: tissue-specific view of the human and model organism interactomes”. In: *Nucleic acids research* 44.D1 (2016), pp. D536–D541.

[123] Roberto Mosca et al. “3did: a catalog of domain-based interactions of known three-dimensional structure”. In: *Nucleic acids research* 42.D1 (2014), pp. D374–D379.

[124] Sailu Yellaboina et al. “DOMINE: a comprehensive collection of known and predicted domain-domain interactions”. In: *Nucleic acids research* 39.suppl_1 (2011), pp. D730–D735.

[125] Marija Buljan et al. “Tissue-specific splicing of disordered segments that embed binding motifs rewires protein interaction networks”. In: *Molecular cell* 46.6 (2012), pp. 871–883.

[126] Koyel Mitra et al. “Integrative approaches for finding modular structure in biological networks”. In: *Nature Reviews Genetics* 14.10 (2013), pp. 719–732.

[127] Edwin K Silverman et al. “Molecular networks in Network Medicine: Development and applications”. In: *Wiley Interdisciplinary Reviews: Systems Biology and Medicine* 12.6 (2020), e1489.

[128] Murodzhon Akhmedov et al. “A fast prize-collecting steiner forest algorithm for functional analyses in biological networks”. In: *Integration of AI and OR Techniques in Constraint Programming: 14th International Conference, CPAIOR 2017, Padua, Italy, June 5-8, 2017, Proceedings 14*. Springer. 2017, pp. 263–276.

[129] Olga Lazareva et al. “On the limits of active module identification”. In: *Briefings in Bioinformatics* 22.5 (2021), bbab066.

[130] Lisette JA Kogelman et al. “Identification of co-expression gene networks, regulatory genes and pathways for obesity based on adipose tissue RNA Sequencing in a porcine model”. In: *BMC medical genomics* 7.1 (2014), pp. 1–16.

[131] Pei He et al. “Epigenetically regulated co-expression network of genes significant for rheumatoid arthritis”. In: *Epigenomics* 11.14 (2019), pp. 1601–1612.

[132] Qi Zhang et al. “Integrated proteomics and network analysis identifies protein hubs and network alterations in Alzheimer’s disease”. In: *Acta neuropathologica communications* 6 (2018), pp. 1–19.

[133] Martin Lempp et al. “Systematic identification of metabolites controlling gene expression in E. coli”. In: *Nature communications* 10.1 (2019), p. 4463.

[134] Zaiba Hasan Khan et al. “Co-expression network analysis of protein phosphatase 2A (PP2A) genes with stress-responsive genes in Arabidopsis thaliana reveals 13 key regulators”. In: *Scientific reports* 10.1 (2020), p. 21480.

[135] Valur Emilsson et al. “Genetics of gene expression and its effect on disease”. In: *Nature* 452.7186 (2008), pp. 423–428.

[136] Andrew T McKenzie et al. “DGCA: a comprehensive R package for differential gene correlation analysis”. In: *BMC systems biology* 10 (2016), pp. 1–25.

[137] Yvonne Bouter et al. “miRNA alterations elicit pathways involved in memory decline and synaptic function in the hippocampus of aged Tg4-42 mice”. In: *Frontiers in Neuroscience* 14 (2020), p. 580524.

[138] Wolfgang Huber et al. “Variance stabilization applied to microarray data calibration and to the quantification of differential expression”. In: *Bioinformatics* 18.suppl_1 (2002), S96–S104.

[139] Ronald A Fisher. “Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population”. In: *Biometrika* 10.4 (1915), pp. 507–521.

[140] Yoav Benjamini and Yosef Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing”. In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[141] Gene Ontology Consortium. “The Gene Ontology (GO) database and informatics resource”. In: *Nucleic acids research* 32.suppl_1 (2004), pp. D258–D261.

[142] Minoru Kanehisa and Susumu Goto. “KEGG: kyoto encyclopedia of genes and genomes”. In: *Nucleic acids research* 28.1 (2000), pp. 27–30.

[143] Yuhao Chen and Xiaowei Wang. “miRDB: an online database for prediction of functional microRNA targets”. In: *Nucleic acids research* 48.D1 (2020), pp. D127–D131.

[144] Philia Bouchard-Bourelle et al. “snoDB: an interactive database of human snoRNA sequences, abundance and interactions”. In: *Nucleic acids research* 48.D1 (2020), pp. D220–D225.

[145] Chit Tong Lio et al. “Systematic analysis of alternative splicing in time course data using Spycone”. In: *Bioinformatics* 39.1 (2023), btac846.

[146] Jung Eun Shim and Insuk Lee. “Weighted mutual information analysis substantially improves domain-based functional network models”. In: *Bioinformatics* 32.18 (2016), pp. 2824–2830.

[147] Jaina Mistry et al. “Pfam: The protein families database in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D412–D419.

[148] Jerzy Neyman and Egon S Pearson. “On the use and interpretation of certain test criteria for purposes of statistical inference: Part I”. In: *Biometrika* (1928), pp. 175–240.

[149] Sture Holm. “A simple sequentially rejective multiple test procedure”. In: *Scandinavian journal of statistics* (1979), pp. 65–70.

[150] Fabian Pedregosa et al. “Scikit-learn: Machine learning in Python”. In: *the Journal of machine Learning research* 12 (2011), pp. 2825–2830.

[151] Romain Tavenard et al. “Tslearn, a machine learning toolkit for time series data”. In: *The Journal of Machine Learning Research* 21.1 (2020), pp. 4686–4691.

[152] Hagai Levi, Ran Elkon, and Ron Shamir. “DOMINO: a network-based active module identification algorithm with reduced rate of false calls”. In: *Molecular systems biology* 17.1 (2021), e9593.

[153] Uku Raudvere et al. “g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update)”. In: *Nucleic acids research* 47.W1 (2019), W191–W198.

[154] Huijuan Feng et al. “Modeling RNA-binding protein specificity in vivo by precisely registering protein-RNA crosslink sites”. In: *Molecular cell* 74.6 (2019), pp. 1189–1204.

[155] Peter JA Cock et al. “Biopython: freely available Python tools for computational molecular biology and bioinformatics”. In: *Bioinformatics* 25.11 (2009), p. 1422.

[156] Cole Trapnell et al. “Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation”. In: *Nature biotechnology* 28.5 (2010), pp. 511–515.

[157] Pär G Engström et al. “Systematic evaluation of spliced alignment programs for RNA-seq data”. In: *Nature methods* 10.12 (2013), pp. 1185–1191.

[158] Giacomo Baruzzo et al. “Simulation-based comprehensive benchmarking of RNA-seq aligners”. In: *Nature methods* 14.2 (2017), pp. 135–139.

[159] Minghao Jiang et al. “A comprehensive benchmarking of differential splicing tools for RNA-seq analysis at the event level”. In: *Briefings in Bioinformatics* 24.3 (2023), bbad121.

[160] Kalpana Kannan et al. “Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing”. In: *Proceedings of the National Academy of Sciences* 108.22 (2011), pp. 9172–9177.

[161] Bo Li and Colin N Dewey. “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”. In: *BMC bioinformatics* 12 (2011), pp. 1–16.

[162] Doyeon Kim et al. “A high-resolution temporal atlas of the SARS-CoV-2 translatome and transcriptome”. In: *Nature communications* 12.1 (2021), p. 5120.

[163] Hye Kyung Lee et al. “Immune transcriptomes of highly exposed SARS-CoV-2 asymptomatic seropositive versus seronegative individuals from the Ischgl community”. In: *Scientific reports* 11.1 (2021), p. 4243.

[164] Yichuan Liu et al. “Evaluating the impact of sequencing depth on transcriptome profiling in human adipose”. In: *PloS one* 8.6 (2013), e66883.

[165] Dongyuan Song et al. “scDesign3 generates realistic in silico data for multimodal single-cell and spatial omics”. In: *Nature Biotechnology* 42.2 (2024), pp. 247–252.

[166] Yuhan Hao et al. “Dictionary learning for integrative, multimodal and scalable single-cell analysis”. In: *Nature Biotechnology* (2023). DOI: `10.1038/s41587-023-01767-y`. URL: `https://doi.org/10.1038/s41587-023-01767-y`.

[167] Vincent D Blondel et al. “Fast unfolding of communities in large networks”. In: *Journal of statistical mechanics: theory and experiment* 2008.10 (2008), P10008.

[168] Yvonne Bouter et al. “N-truncated amyloid $\beta$ (A$\beta$) 4-42 forms stable aggregates and induces acute and long-lasting behavioral deficits”. In: *Acta neuropathologica* 126 (2013), pp. 189–205.

[169] Javier Riancho et al. “MicroRNA profile in patients with Alzheimer’s disease: analysis of miR-9-5p and miR-598 in raw and exosome enriched cerebrospinal fluid samples”. In: *Journal of Alzheimer’s Disease* 57.2 (2017), pp. 483–491.

[170] Xiaoyang Ye et al. “MicroRNAs 99b-5p/100-5p regulated by endoplasmic reticulum stress are involved in abeta-induced pathologies”. In: *Frontiers in Aging Neuroscience* 7 (2015), p. 210.

[171] Wei Liu, Jingya Zhao, and Guangxiu Lu. “miR-106b inhibits tau phosphorylation at Tyr18 by targeting Fyn in a model of Alzheimer’s disease”. In: *Biochemical and biophysical research communications* 478.2 (2016), pp. 852–857.

[172] Jun-Lin Liu, Zhan-You Wang, and Chuang Guo. “Iron and Alzheimer’s disease: from pathogenesis to therapeutic implications”. In: *Frontiers in neuroscience* 12 (2018), p. 411985.

[173] Serena Silvestro, Placido Bramanti, and Emanuela Mazzon. “Role of miRNAs in Alzheimer’s disease and possible fields of application”. In: *International journal of molecular sciences* 20.16 (2019), p. 3979.

[174] Xiaoyun Guo et al. “Genome-wide significant, replicated and functional risk variants for Alzheimer’s disease”. In: *Journal of neural transmission* 124 (2017), pp. 1455–1471.

[175] Shengjun Hong et al. “Canonical correlation analysis for RNA-seq co-expression networks”. In: *Nucleic acids research* 41.8 (2013), e95–e95.

[176] Elise AR Serin et al. “Learning from co-expression networks: possibilities and challenges”. In: *Frontiers in plant science* 7 (2016), p. 185898.

[177] Harun Pirim. “Construction of gene networks using expression profiles”. In: *Soft Computing for Biological Systems* (2018), pp. 67–89.

[178] S. Ballouz, W. Verleyen, and J. Gillis. “Guidance for RNA-seq co-expression network construction and analysis: safety in numbers”. In: *Bioinformatics* 31.13 (July 2015), pp. 2123–2130. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btv118.

[179] Princy Parsana et al. “Addressing confounding artifacts in reconstruction of gene co-expression networks”. In: *Genome biology* 20.1 (2019), pp. 1–6.

[180] Kayla A. Johnson and Arjun Krishnan. “Robust normalization and transformation techniques for constructing gene coexpression networks from RNA-seq data”. In: *Genome Biology* 23.1 (Jan. 2022), p. 1. ISSN: 1474-760X. DOI: 10.1186/s13059-021-02568-9.

[181] Christian Wiwie et al. “Elucidation of time-dependent systems biology cell response patterns with time course network enrichment”. In: *arXiv preprint arXiv:1710.10262* (2017).

[182] Nelle Varoquaux and Elizabeth Purdom. “A pipeline to analyse time-course gene expression data”. In: *F1000Research* 9 (2020), p. 1447.

[183] Paul Shannon et al. “Cytoscape: a software environment for integrated models of biomolecular interaction networks”. In: *Genome research* 13.11 (2003), pp. 2498–2504.

[184] Nicolas Alcaraz et al. “KeyPathwayMiner 4.0: condition-specific pathway analysis by combining multiple omics studies and networks with Cytoscape”. In: *BMC systems biology* 8 (2014), pp. 1–6.

[185] Ali Javed, Byung Suk Lee, and Donna M Rizzo. “A benchmark study on time series clustering”. In: *Machine Learning with Applications* 1 (2020), p. 100001.

[186] Jianjia Li et al. “Dynamic proteomic profiling of human periodontal ligament stem cells during osteogenic differentiation”. In: *Stem Cell Research & Therapy* 12 (2021), pp. 1–16.

[187] Jee-Soo Lee et al. “Longitudinal proteomic profiling provides insights into host response and proteome dynamics in COVID-19 progression”. In: *Proteomics* 21.11-12 (2021), p. 2000278.

[188] Linlin Sui et al. “Dynamic proteomic profiles of in vivo-and in vitro-produced mouse postimplantation extraembryonic tissues and placentas”. In: *Biology of Reproduction* 91.6 (2014), pp. 155–1.

[189] Jonathan M Raser and Erin K O’shea. “Noise in gene expression: origins, consequences, and control”. In: *Science* 309.5743 (2005), pp. 2010–2013.

[190] Gábor Balázsi, Alexander Van Oudenaarden, and James J Collins. “Cellular decision making and biological noise: from microbes to mammals”. In: *Cell* 144.6 (2011), pp. 910–925.

[191] Paul C Bressloff. *Stochastic processes in cell biology*. Vol. 41. Springer, 2014.

[192] Brent Ewing et al. “Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment”. In: *Genome research* 8.3 (1998), pp. 175–185.

[193] Ana Conesa et al. “A survey of best practices for RNA-seq data analysis”. In: *Genome Biology* 17.1 (Jan. 2016), p. 13. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0881-8.

[194] Guillermo Marco-Puche et al. “RNA-Seq perspectives to improve clinical diagnosis”. In: *Frontiers in genetics* 10 (2019), p. 492554.

[195] Matthew Chung et al. “Best practices on the differential expression analysis of multi-species RNA-seq”. In: *Genome biology* 22.1 (2021), p. 121.

[196] Albert-László Barabási, Zoltán N Oltvai, and Stefan Wuchty. “Characteristics of biological networks”. In: *Complex networks* (2004), pp. 443–457.

[197] Marta Lucchetta et al. “Emergence of power-law distributions in protein-protein interaction networks through study bias”. In: *bioRxiv* (2023). DOI: 10.1101/2023.03.17.533165. URL: https://www.biorxiv.org/content/early/2023/03/21/2023.03.17.533165.

[198] Leto Peel, Tiago P Peixoto, and Manlio De Domenico. “Statistical inference links data and theory in network science”. In: *Nature Communications* 13.1 (2022), p. 6794.

[199] Zakaria Louadi et al. “DIGGER: exploring the functional role of alternative splicing in protein interactions”. In: *Nucleic acids research* 49.D1 (2021), pp. D309–D318.

[200] Patrick Bryant, Gabriele Pozzati, and Arne Elofsson. “Improved prediction of protein-protein interactions using AlphaFold2”. In: *Nature communications* 13.1 (2022), p. 1265.

[201] Richard Evans et al. “Protein complex prediction with AlphaFold-Multimer”. In: *biorxiv* (2021), pp. 2021–10.

[202] Eric T Wang et al. “Alternative isoform regulation in human tissue transcriptomes”. In: *Nature* 456.7221 (2008), pp. 470–476.

[203] Yarden Katz et al. “Analysis and design of RNA sequencing experiments for identifying isoform regulation”. In: *Nature methods* 7.12 (2010), pp. 1009–1015.

[204] Francisco J Pardo-Palacios et al. “Systematic assessment of long-read RNA-seq methods for transcript identification and quantification”. In: *bioRxiv* (2021).

[205] Wei Xiong Wen, Adam J Mead, and Supat Thongjuea. “MARVEL: an integrated alternative splicing analysis platform for single-cell RNA sequencing data”. In: *Nucleic Acids Research* 51.5 (2023), e29–e29.

[206] Gonzalo Benegas, Jonathan Fischer, and Yun S Song. “Robust and annotation-free analysis of isoform variation using short-read scRNA-seq data”. In: *bioRxiv* (2021).

[207] Julia Eve Olivieri, Roozbeh Dehghannasiri, and Julia Salzman. “The SpliZ generalizes ‘percent spliced in’to reveal regulated splicing at single-cell resolution”. In: *Nature methods* 19.3 (2022), pp. 307–310.

[208] Yuanhua Huang and Guido Sanguinetti. “BRIE2: computational identification of splicing phenotypes from single-cell transcriptomic experiments”. In: *Genome biology* 22.1 (2021), p. 251.

[209] Vivien Marx. “Method of the year: long-read sequencing”. In: *Nature Methods* 20.1 (2023), pp. 6–11.

[210] Michael L. Tress, Federico Abascal, and Alfonso Valencia. “Alternative Splicing May Not Be the Key to Proteome Complexity”. English. In: *Trends in Biochemical Sciences* 42.2 (Feb. 2017), pp. 98–110. ISSN: 0968-0004. DOI: 10.1016/j.tibs.2016.08.008.

[211] Caroline Seydel. “Diving deeper into the proteome”. en. In: *Nature Methods* 19.99 (Sept. 2022), pp. 1036–1040. ISSN: 1548-7105. DOI: 10.1038/s41592-022-01599-9.

[212] Zhana Duren et al. “Regulatory analysis of single cell multiome gene expression and chromatin accessibility data with scREG”. In: *Genome Biology* 23.1 (May 2022), p. 114. ISSN: 1474-760X. DOI: 10.1186/s13059-022-02682-2.

[213] Eleni P. Mimitou et al. “Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells”. en. In: *Nature Biotechnology* 39.1010 (Oct. 2021), pp. 1246–1258. ISSN: 1546-1696. DOI: 10.1038/s41587-021-00927-2.

[214] Sebastian Pott. “Simultaneous measurement of chromatin accessibility, DNA methylation, and nucleosome phasing in single cells”. In: *eLife* 6 (June 2017). Ed. by Bing Ren, e23203. ISSN: 2050-084X. DOI: 10.7554/eLife.23203.

[215] Arjun Raj et al. “Imaging individual mRNA molecules using multiple singly labeled probes”. en. In: *Nature Methods* 5.1010 (Oct. 2008), pp. 877–879. ISSN: 1548-7105. DOI: 10.1038/nmeth.1253.

[216] Chenglong Xia et al. “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression”. eng. In: *Proceedings of the National Academy of Sciences of the United States of America* 116.39 (Sept. 2019), pp. 19490–19499. ISSN: 1091-6490. DOI: 10.1073/pnas.1912459116.

[217] Fredrik Salmén et al. “Barcoded solid-phase RNA capture for Spatial Transcriptomics profiling in mammalian tissue sections”. eng. In: *Nature Protocols* 13.11 (Nov. 2018), pp. 2501–2534. ISSN: 1750-2799. DOI: `10.1038/s41596-018-0045-2`.

[218] Samuel G. Rodriques et al. “Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution”. eng. In: *Science (New York, N.Y.)* 363.6434 (Mar. 2019), pp. 1463–1467. ISSN: 1095-9203. DOI: `10.1126/science.aaw1219`.