# Ladder-seq partitions RNA-seq reads by length to improve transcriptome quantification and assembly

Shounak Chakraborty

München 2021

Dissertation zur Erlangung des Doktorgrades
der Fakultät für Chemie und Pharmazie
der Ludwig–Maximilians–Universität München

# Ladder-seq partitions RNA-seq reads by length to improve transcriptome quantification and assembly

Shounak Chakraborty

aus

Kolkata, India

2021

**Erklärung:**
Diese Dissertation wurde im Sinne von §7 der Promotionsordnung vom 28. November 2011 von Herrn Dr. Stefan Canzar betreut.

**Eidesstattliche Versicherung:**
Diese Dissertation wurde eigenständig und ohne unerlaubte Hilfe erarbeitet.

München, 30.08.2021

Shounak Chakraborty

Dissertation eingereicht am 31.08.2021
1. Gutachter: Dr. Stefan Canzar
2. Gutachter: Prof. Dr. Julien Gagneur
Mündliche Prüfung am 04.10.2021

# Contents

# Acknowledgments

First and foremost I would like to express my deepest gratitude to my supervisor Dr. Stefan Canzar for giving me the opportunity to work in his lab and with the exciting project Ladder-seq . I am grateful for the support that I have got during my PhD and also for the valuable things related to science and bioinformatics that I have learned while working with him. Dr. Canzar convincingly guided and encouraged me to be professional and do the right thing even when the road got tough. Without his persistent help, the goal of this project would not have been realized.

I would like to express my appreciation to my thesis advisory committee members, Prof. Dr. Julien Gagneur and Prof. Dr. Veit Hornung for their valuable feedback and insightful suggestions. I would like to pay my special regards to Prof. Dr. Julien Gagneur, Prof. Dr. Caroline Friedel, Prof. Dr. Klaus Förstemann, Prof. Dr. Johannes Stigler and Prof. Dr. Veit Hornung for being in my thesis defence committee.

I want to acknowledge the collaboration and help that I got from Francisca with my research with Ladder-seq. Additionally, I want to thank all the members of the Canzar lab, especially Hoan, Pablo and Parastu for the great conversations we have had over journal clubs, lab meetings over lunch or while taking a break from work at the kicker table. You all made the it fun to come to work and kept me going through the four years and through the pandemic.

In addition I would like to thank the Graduate school of Quantitative Biosciences Munich for creating an excellent atmosphere for scientific education. The constant support and help from the coordinators, especially Markus helped me keep afloat at difficult times. The courses organized by QBM helped me gain invaluable scientific knowledge as well as soft skills.

I would like to express my deep and sincere gratitude towards my family, especially my parents, brother and grandparents for their continuous and unparalleled love, help and support. I am forever indebted to them for giving me the opportunities and experiences that have made me who I am. Moreover I am grateful to all my friends in Munich and elsewhere in the world who have always been a major source of support and happiness and have been there especially when things would get a bit discouraging.

Finally I would like to put a special note of gratitude to Sam for being a constant companion through thick and thin and for all the love and support.

# Summary

Transcript level inference methods, such as quantification and assembly methods often lose accuracy due to the lack of long range information from short NGS reads. In this thesis we introduce a novel experimental-computational RNA-seq protocol called Ladder-seq that introduces an additional layer of information based on the length of transcripts. Ladder-seq separates transcripts based on their length prior to sequencing. As a result of this separation, the reads obtained from sequencing contain information about the length of its originating transcript.

We validate the quality of separation of transcripts computationally and we also confirm that the length separation does not add any bias by comparing the Ladder-seq data with publicly available long and short read RNA-seq data set. Additionally we model migration patterns of mRNA through a gel during the process of separation which we use to extend and tailor state-of-the-art RNA-seq methods for quantification, reference based assembly and and *de novo* assembly.

We demonstrate on simulated data that the accuracy of quantification of our extension of kallisto for Ladder-seq for complex genes expressing up to 10 isoforms is the same as the accuracy of quantification of conventional kallisto for genes expressing merely 2 transcripts. Our reference-based assembly scheme based on StringTie2 achieves a 30.8% higher precision in reconstructing the single transcript expressed by a gene when compared to its conventional counterpart and is more than 30% more sensitive on complex genes. Our *de novo* assembly approach using Trinity correctly assembles 78% more transcripts than conventional Trinity and at the same time improves precision equally by 78%.

In a real data set, the comparison of more accurate transcript reconstructions by Ladder-seq reveals 40% more genes harboring isoform switches compared to conventional RNA-seq approaches. We demonstrate that the distribution of reads to transcripts is more accurate in case of Ladder-seq based Kallisto as compared to its conventional counterpart. We utilize our novel approach to study the role of m$^6$A methylation as a regulator of splicing in mouse neural progenitor cells (NPCs). Ladder-seq unveils widespread changes in isoform usage in mouse NPCs upon m$^6$A depletion by *Mettl14* knock-out. Genes harboring isoform switches are enriched for m$^6$A methylated genes and m$^6$A tends to be close to differentially spliced exonic regions. Lastly we verify a selection of novel transcripts exhibiting isoform switches which were only identified by Ladder-seq based pipelines in our own long read sequencing data.

*Chapter 1*

# Introduction

The intricacies of life and the mechanisms behind the processes that govern life have been confounding humans for centuries. For years philosophers have tried to decode the meaning and purpose of life and out came various independent theories ranging from distant places like Greece, India and China to name a few. The concept of biology existed in ancient medical traditions like Ayurveda, Egyptian and Greco-Roman medicine, but the foundations of modern medicine can be attributed to the immense development in science and technology during the European renaissance and early modern period. The discovery of cells by Robert Hooke and the subsequent invention of the first compound microscope by Antoine van Leeuwenhoek in the seventeenth century, opened up a previously unknown world, the world of micro-organisms and kick started the field of biology known as cell biology.

## 1.1 Cell biology

Cell biology is the study of cells and it revolves around the concept of the cell being the fundamental building block of life. The first concrete definition of the cell was stated by scientists Schleiden and Schwann who stated that all living creatures, simple or complex were made up of one or more cells and that the cell is the structural and functional unit of life [1]. This definition of the cell came to be known as the *cell theory*. With the improvement in microscopy, scientists were able to observe the intricacies of the cell in unprecedented details. From the early days of discovering components of cells and their functions, cell biology has evolved a lot to attain its most recent embodiment, the ”Modern cell theory” [2] which includes among others the following main ideas:

- All living things are composed of cells.
- All living things arise from pre-existing cells by division.
- The cell is the fundamental unit of structure and function in all living organisms.
- Cells contain certain molecules called DNA which contain hereditary information which is inherited from parent cells to children cells.

There are various sub-fields within cell biology, for instance the study of biochemical reactions and the metabolism of the cell, the structure of cell components, cell communication etc. Genetics is one such sub-field of cell biology which is the study of heredity and the storage and expression of hereditary information in cells. In the following sections we are going to dive deeper into some of the important concepts of genetics.

### 1.1.1 DNA and the genome

Cells store their hereditary information in double stranded molecules called Deoxyribonucleic acid or DNA. Though the DNA was first isolated as early as 1871 by Friedrich Miescher [3], it was not until the middle and later half of the 20th century that the famous double helical structure of the DNA was identified by the combined efforts of scientists Rosalind Franklin, Maurice Wilkins, James Watson and Francis Crick [3].

The DNA is composed of chains of molecules called nucelotides which are twisted into a double helix shape. Nucleotides in DNA can contain four types of nitrogen bases, adenine, cytosine, guanine and thymine represented as A, C, G and T. The two strands of the DNA comprise two polypeptide chains which contain complementary nucleotides, meaning that an adenine molecule on one of the strands will always be aligned to a thymine molecule in the other strand and a guanine aligned to a cytosine (Fig. 1.1a). This is known as *base complementarity* and the molecular forces between the nucleotides are responsible for the stable double helix structure.

The DNA molecule is further compacted and organized into several parts known as *chromosomes*. The number of chromosomes varies between organisms for example humans have their DNA divided into 23 chromosomes while mice have 20.

Similar to the way of storing information in a computer file using only 1s and 0s, information about protein synthesis is stored using the four letter alphabet A,C,T and G for the four nucleotide bases. A *gene* is a small section of the DNA which can be represented as a sequence of nucleotide pairs and is the basic unit of heredity where information about one or more types of molecules can be stored. All the genetic information in a cell is collectively called the *genome*.

### 1.1.2 Central dogma of molecular biology

Information stored in the DNA needs to be *expressed* in order to guide the synthesis of a myriad of molecules which make the various physiological processes possible in cells. The mechanism of translating the coded information in DNA to synthesize proteins is the same in all cells and is referred to as the ”Central dogma of molecular biology” [5] (Fig. 1.2). According to the central dogma, the DNA is converted to a single stranded molecule called Ribonucleic acid or RNA (Fig. 1.1b) through a process called *transcription*. RNA and DNA have three common nucleotide bases, adenine, guanine and cytosine. In place of thymine, RNA contains uracil which is an unmethylated form of thymine [6].The RNA molecules which encode proteins are called messenger RNA or mRNA. During transcription the double stranded DNA is first converted to an intermediate molecule called *pre-mRNA* or *primary transcript*. In eukaryotic organisms, the pre-mRNA undergoes further processing in order to

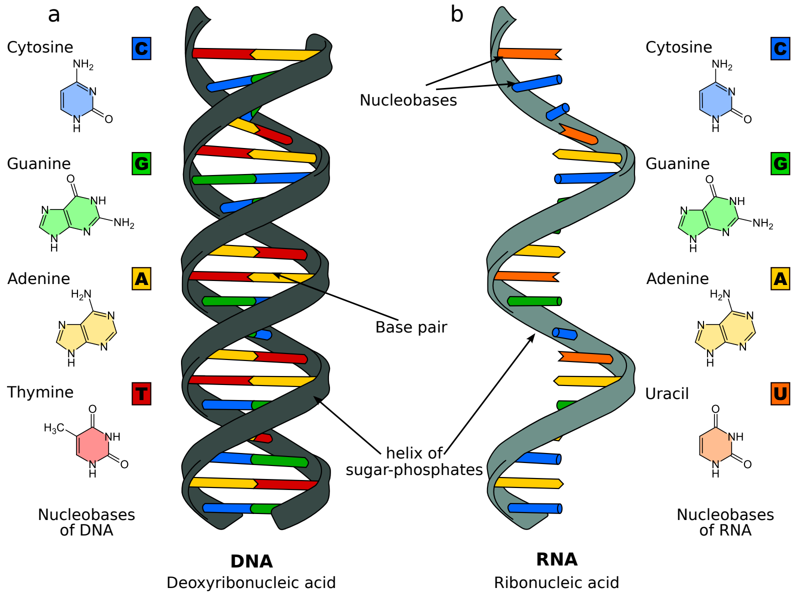

Figure 1.1: Structures of DNA and RNA. (a) Double helical structure of the DNA comprising two polynucleotide chains coiled around each other. The monomeric nucleotides on each strand are joined to each other by hydrogen bonds follwing base pairing rules, (A with T and C with G). (b) Single stranded RNA molecule containing the same nucleotides as DNA except for Thymine which is Uracil in RNA. [4]

form mature mRNA, whereas in prokaryotes the mature mRNA is mostly produced directly from the DNA. Transcription is usually carried out by protein complexes called *RNA polymerases* which bind to specific parts of the DNA called *promoter* regions to start the process of transcription.

The second part of the central dogma states that mature mRNA is converted to proteins through the process called *translation*. Ribosomes are macromolecular protein complexes which carry out the process of translation by assembling amino acids to peptide chains based on the sequence specified by the mature mRNA.

### 1.1.3 RNA and transcriptome

Each cell (or population of cells) in a particular organism contains the same copy of the genome but has a different functionality. Different cells need different sets of proteins to be synthesized from the same genome. A single gene can encode multiple mRNA molecules which in turn can produce a number of proteins. Genes are divided into two types of segments/sequences, exons and introns (Fig. 1.3a). Exons are segments that encode a part of the mature mRNA while introns don't. Pre-mRNA, which is the first form of RNA that is produced from DNA after transcription, contains both the introns and exons. The introns are removed and the exons are joined together to form a mature mRNA molecule, broadly referred to as a *transcript* or *isoform*, through a process called *splicing*.

During the splicing reaction, different exons can be selected to be included or excluded in the resulting mRNA allowing one gene to code for multiple proteins. The variation in transcripts is brought forward by a combinatorial selection of ex-

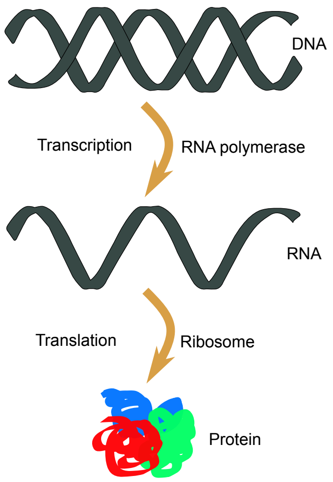

Figure 1.2: The central dogma of molecular biology. Double stranded DNA is converted to RNA through a process called transcription and using a protein called RNA polymerase. Single stranded RNA molecules are converted into proteins with the aid of another protein called ribosome and the process is known as translation.

ons through a process called *alternative splicing* [7]. This is a major contributor to protein diversity in multi-cellular organisms. For instance it allows the human genome, consisting of 20,000 genes to encode for many more proteins. Fig. 1.3 shows an example gene with four exons and its isoforms and proteins which are obtained by combinations of exons.

Other mechanisms that contribute to the diversity of mRNA encoded by genes include *intron retention* where parts of an intron remains in the final transcript and *alternative transcription start and end sites* where there are multiple transcription start and end sites for a single exon.

### 1.1.4 Regulation of gene expression

We have briefly seen how information in cells are stored in DNA. Genes, which are sub-divisions of DNA encode for multiple RNA molecules called transcripts. The wide range of mechanisms that are used to increase or decrease gene expression are known as gene regulation mechanisms. Gene regulation has been observed in every step starting from regulation of transcription in order to control the amount of mRNA produced, post transcriptional modifications of mRNA to post translational modifications of proteins.

*Transcriptional regulation* is an umbrella term representing all the processes that control the transcription of genes to produce mRNA. Though the basic methods of transcriptional regulation are similar in prokaryotes and eukaryotes, the latter contains genes which are more complex and typically encode and express multiple RNA transcripts, hence requiring more complex control mechanisms. Transcription factors (TFs) are a family of proteins which bind to certain sequences in the DNA to activate or deactivate a particular gene. TFs, need to bind to certain parts of the gene called transcription factor binding sites (TFBS) and then the RNA polymerase

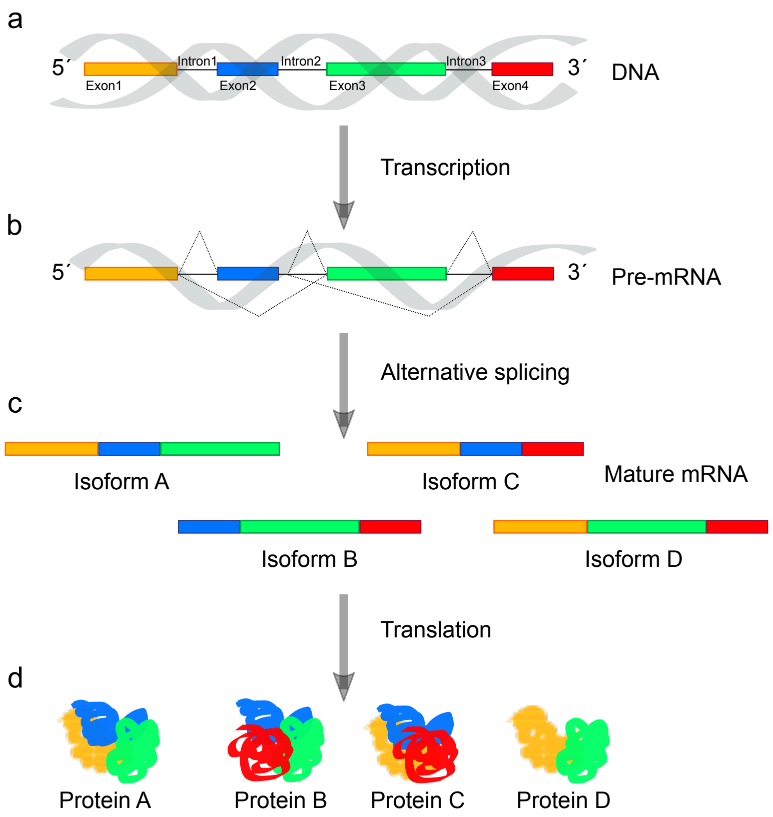

Figure 1.3: Alternative splicing to yield different proteins from the same gene. (a) An example gene consisting of 4 exons and 3 introns. (b) Pre-mRNA, the initial RNA product after transcription containing introns. (c) During the process of alternative splicing, different combinations of exons yields four different isoforms. (d) Four different protein products, one from each isoform.

can bind to the promoter region to start the transcription. TFs can work alone or with other proteins to increase the transcription of genes in which case they are called activators. They can also reduce the expression by blocking the RNA ploymerase from binding to the gene and then are termed as repressors.

There are various other factors that contribute to transcriptional regulation. The structure and folding of the DNA can make it either facilitate or hinder the binding of the RNA-polymerase to the DNA thus regulating gene expression. Epigenetics, which is the study of heritable changes in phenotype that does not involve changes in the DNA sequence, is another very common method of regulation of gene expression. Epigenetic modifications include DNA methyation, histone modifications, chromatin remodelling and non-coding RNA modification [8]. DNA methylation, a common epigenetic modification is a heritable epigenetic mark where a methyl group is covalently transferred to the DNA molecule [9]. When located in the promoter region of a gene, methylation typically acts as a repressor of gene transcription.

Another category of critical regulators of transcript expression are the post transcriptional modifications of RNA molecules. These are modification which the RNA undergoes in between the transcription and translation phases. One of the common post transcriptional modifications is splicing. Polyadenylation, which is another very common modification involves the process of addition of a segment containing only adenine molecules (also called the ploy-A tail) to the end of the mRNA. One of the more recently discovered post transcriptional modifications of RNA is known as *epitranscriptomics.* Similar to DNA molecules, RNA also undergoes chemical modifications.

### 1.1.5 Epitranscriptomic regulation and m$^6$A

The advent of high throughput sequencing and the development of specific antibodies, has revealed N6-methyladenosine (m$^6$A) as the most abundant internal modification of mRNA in eukaryotic cells [10] and is involved in multiple aspects of mRNA biology including alternative splicing [11, 12, 13, 14]. m$^6$A modification plays an important role in the regulation of cell fate, proliferation, and metabolism and the biogenesis of tumours [15]. Similar to DNA methylation, m$^6$A modification adds a methyl group to the adenine in an RNA molecule. Fig. 1.4 shows the difference between unmethylated and methylated adenosine.

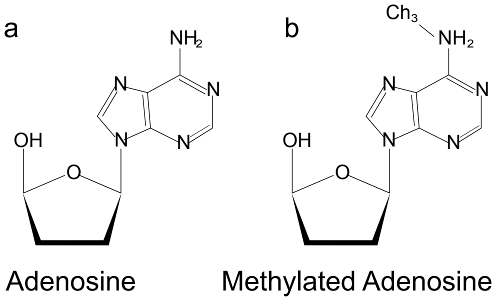

Figure 1.4: Differences between the chemical structure of nonmethylated and methylated adenosine. [16]

m$^6$A modifications are implemented by RNA methyltransferases (writers), RNA demethylase (erasers) and m6a binding proteins (readers). These are proteins which either add remove or simply read methylation modifications from an RNA molecule [17]. *Mettl3* and *Mettl14* are two of the most common m$^6$A writers and overexpression of genes that code for these proteins leads to a increase in m$^6$A [18]. Similarly RNA demethylase *ALKBH5* is an eraser which if overexpressed in cells results in a depletion of m$^6$A and lastly N6-methyladenosine RNA binding protein 1, *YTHDF1* is a reader which promotes the translation of m6A-modified mRNA [15]. m$^6$A is involved in multiple aspects of mRNA biology including alternative splicing [11, 12, 13, 14].

## 1.2 Sequencing

Sequencing refers to various techniques that are used to determine the linear sequence of the four necleotides, adenine, guanine, cytosine and thymine as they occur in the DNA or RNA. Frederick Sanger in 1977 developed a method for sequencing DNA and that opened the door to an enormous amount of hereditary information stored in the double stranded molecules. In case of the DNA, sequencing reveals the order of neucleotides of a given fragment containing both exons and introns while in the case of RNA, the order of nucleotides contain only exons which were included in the particular mature mRNA. Sequencing mRNA molecules is somewhat different from sequencing DNA owing to the fact that mRNA is single stranded. Generally the first step of RNA-seq involves the conversion of mRNA molecules to complementary DNA (cDNA) molecules. The cDNA library can then be sequenced using the same principles as genomic sequencing.

The first sequencing methods, also known as first generation sequencing were introduced in the 1970s and included methods like Maxim Gilbert and Sanger sequencing with the later being the more commonly used of the two [19]. Though these methods were quite accurate in determining the sequence of neucleotides, the low throughput and high cost of sequencing slowly led to the advent of the second generation sequencing (also known as high throughput or massively parallel sequencing). The utility of the newer technologies were further enhanced by the tremendous improvements in computational power and as a result it was possible to sequence and store very large datasets of DNA and RNA.

Second generation sequencing breaks DNA (or cDNA in case of RNA-seq) into shorter fragments by an enzymatic reaction and produces a very high number of short sequences or reads with lengths ranging from 75 to 500 base pairs which represent the nucleotide sequence from each fragment. The extremely high depth (the number of reads from a particular genomic or transcriptomic location) makes NGS an extremely valuable technology for downstream analyses that need a lot of statistical power, for example the estimation of abundance of transcripts. Some of the most common platforms for second generation sequencing are Roche 454, Illumina MiSeq and Illumina HiSeq. Conventional sequencing procedures, which is also called bulk sequencing, involve the extraction of genomic or transcriptomic content from an entire tissue or population of cells and then the subsequent sequencing and downstream analysis. A much newer variant of sequencing is single cell sequencing which focuses on the isolation and sequencing of DNA or mRNA from individual cells.

Despite being a very powerful tool in the analysis of the underlying genomic and transcriptomic landscape, second generation sequencing has its shortcomings. One of the most important limitations is the relatively short length (approx. 75 - 500 base pairs) of reads which do not span more than two exons making the characterization of alternative splicing events difficult. Moreover, genomes often have repeated sequences causing individual reads to map to multiple places in the genome (multi mapping reads) leading to difficulties to analysis such as assembling transcripts. The drawbacks of short reads are discussed in more details in Section 1.4.

The latest sequencing technologies, which are also referred to as third generation sequencing, like Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT) produce reads of length greater than 10000 base pairs and can in principle cover entire transcripts. Third generation sequencing poses a critical advantage over short reads because of their extremely long reads, however the error rates are typically higher in case of long reads which can negatively affect analysis. The depth achieved by long reads is also quite low compared to short reads thus decreasing the power of statistical inference with them.

We have touched upon the most prevalent types of sequencing. The rest of the introduction will focus on some details of short read RNA-seq and its applications.

### Second generation RNA-seq : a technical overview

RNA sequencing (RNA-seq) is a sequencing method for determining sequnce of transcripts that are expressed in cells. Fig. 1.5 is an overview of the general protocol of second generation RNA-seq. The first step is the extraction and selection of RNA from cell. Ribosomal RNA (rRNA) is the most abundant form of RNA in cell, and they do not encode for proteins, hence not so relavant in gene/transcript expression analyses. Thus mRNA, which is translated to proteins needs to be separated from the rRNA before sequencing. One common post transcriptional modification of mRNA is the addition of a 3' poly-A tail [20]. A prevalent way of isolation of mRNA from other forms of RNA is by the use of oligo(dT) primers, which are magnetic beads attached to a chain of deoxy thimidine residues which bind to the poly-A tail of mRNAs [21].

The isolated mRNA is fragmented by chemical hydrolysis or enzymatic digestion to a size appropriate for the chosen sequencing platform [22]. The fragmented mRNA is then converted into complementary DNA (cDNA) by a reverse transcriptase using random primers [22]. Adapter oligonucleotides are ligated to the cDNA to allow amplification and enable sequencing [22]. Alternatively the extracted mRNA can be first converted to a cDNA library by reverse transcription and fragmented and sequenced afterwards [23]. The fragments can be amplified by polymerase chain reaction (PCR) and is sequenced in a high throughput sequencer to obtain reads from either one end (single end reads) or from both ends (paired end reads) of the fragment [23].

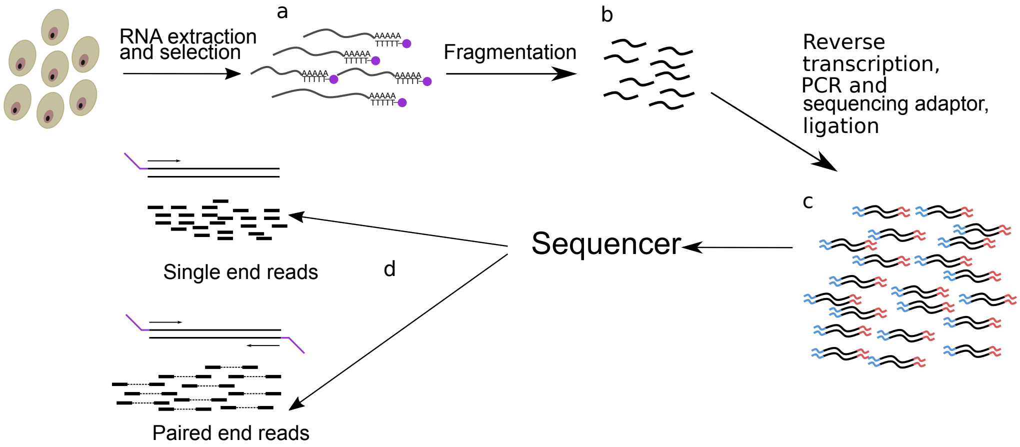
RNA extraction and selection
a
AAAAA
TTTTT
Fragmentation
b
Reverse transcription, PCR and sequencing adaptor, ligation
c
Sequencer
d
Single end reads
Paired end reads

Figure 1.5: Schematic showing the general process of second generation RNA-seq. (a) Cellular RNA is extracted and ploy A tailed mRNA is isolated using oligo dT beads. (b) The isolated mRNA is fragmented. (c) The fragmented mRNA is reverse transcribed using polymerase chain reaction (PCA) to create cDNA libraries to which sequencing adaptors are attached. (d) The sequencer reads fragments either from one side or from both to produce single end or paired end reads.

## 1.3 Analysis of RNA-seq data

Over the past decade, RNA-seq has become a very powerful tool in the analysis of the transcriptome. Common applications involve the quantification and identification of transcripts from the produced sequence read data for which a plethora of computational methods have been developed. Transcript quantification methods assign reads to known species-specific transcripts to obtain a quantitative measurement for their relative expression, while the assembly of transcript sequences can reveal novel mRNA molecules. In contrast to the reference-based assembly that builds full-length transcripts from reads ordered by a prior alignment to a reference genome, the de novo assembly approach reconstructs transcripts based on the sequence overlap of reads alone and can be applied to species for which no or just a highly fragmented reference genome is available. Another important inference from RNA-seq data is to find genes that are differentially expressed across groups of samples [24]. There are many methods for analysing differential expression of genes and transcripts with DESeq2 [24] being one of the most used ones.

### 1.3.1 Expression analysis of transcripts using RNA-seq

Expression analysis or quantification of transcripts involves assigning reads to a previously identified set of transcripts in order to determine their relative abundances. The first step in this process involves mapping reads to a known set of transcripts using splice aware alignment methods such as STAR [25] and HISAT [26]. The task of assigning reads to individual transcripts gets complicated because of the presence of overlapping transcripts from the same gene which results in reads mapping to multiple isoforms. Due to this read mapping ambiguity, quantification methods need to resort to statistical models in order to determine transcript abundances. There are various methods for quantifying transcripts. A popular way of assigning reads is by using a statistical method which maximizes the likelihood of transcript expressions given the observed read data.

#### 1.3.1.1 Estimating transcript abundance

Mapping of RNA-seq reads to a reference produces two categories of mappings, uniquely mapped reads, i.e. reads which map to only one transcript and reads which map to overlapping exons and thus mapping to more than one transcript. Estimating the abundance of transcripts is essentially a counting problem where we want to count the number of reads that map to a particular transcript. While it is easy to determine the originating transcript of a uniquely mapping read, we have to probabilistically assign the non-uniquely mapping reads to transcripts. A naive approach is to assign reads with equal probability to the transcripts that they map to. However this is not correct since the probability of a read originating from a particular transcript would be proportional to the number of copies of the transcript originally present in the sample.

One of the most popular approaches to fractionally assign ambiguously mapping fragments is by using the EM algorithm. The algorithm initializes all transcripts with equal abundances. The expected value of the abundances are calculated based on the observed reads and their estimates are refined in every iteration using the read counts and the abundances calculated in the previous iteration. This process is continued till the values of the abundances converge. Fractional assignment of reads and the EM algorithm are explained in details in Section 4.1.1.

#### 1.3.1.2 Pseudo alignment based quantification

Mapping RNA-seq reads to a genome using conventional algorithms such as STAR [25] and HiSat [26] can be time consuming and can be a deterrant to analysis [27]. A recent alternative to the time consuming step of read mapping involves extracting short sequences of length k (k-mers) from reads and using the counts of these k-mers to estimate transcript abundance [28]. However breaking reads down into k-mers can lead to a loss in accuracy due to the k-mers aligning to many more transcripts than reads, thus making the problem of assigning these reads even more difficult. Kallisto [27] attempts to solve the problem of low accuracy using a method based on *pseudoalignment* of reads and fragments. As the name suggests, there is no real alignment of reads to transcripts in this method. Instead the *pseudoalignment* of a read refers to a subset $S \subset T$, where $T$ is the set containing all the transcripts that need to be quantified. Unlike conventional read mapping, the pseudoalignment does not specify the exact coordinates to which the read maps to, but rather only the set of transcripts that the read maps to.

Kallisto first creates an index using the transcriptome. K-mers are extracted from the annotated set of transcripts and a colored deBruijn graph is constructed where the colors represent different transcripts and nodes represent k-mers. The set of transcripts that overlap a particular k-mer is called the *k-compatibility* class of that k-mer. *Pseudoalignment* of a read is the intersection of all the k-compatibility classes of the k-mers in present in that read. The output of the pseudoalignment step is a set of transcripts for each read called *equivalence class*. The equivalence class of a read represents the possible transcripts from which that read could have originated.

The relative abundances of transcripts are calculated from the equivalence classes of the reads using the following likelihood function.

$$L(\alpha) \propto \prod_{e \in E} \left( \sum_{t \in e} \frac{\alpha_t}{l_t} \right)^{c_e} \tag{1.1}$$

E is the set of equivalence classes observed from the data and $c_e$ represents the number of reads (counts) that have the equivalence class $e$. $\alpha_t$ represents the abundance of a transcript $t$ and $l_t$ is its effective length. The likelihood is optimized using the EM algorithm similar to conventional methods.

### 1.3.2 Assembly using RNA-seq

Short read RNA-seq produces reads which do not span for more than two to three exons in an mRNA. Assembly methods use these short sequences to reconstruct full length transcripts. Alternative splicing is an extremely common phenomenon in eurakryotes and more than 90% genes in humans undergo alternative splicing to produce different mRNA products [29]. Assembling transcripts can help identify novel genes and transcripts as well as confirm the presence of the already annotated ones.

There are two broad categories of transcript assembly methods. Reference based assemblers such as Stringtie [30], Scallop [31], CIDANE [32], CLASS [33], maps reads to a reference genome and then assembles transcripts using the alignments of the reads to the genome. Denovo assemblers such as ABySS [34], SOAPdenovo-trans [35], Oases [36] and Trinity [37], on the other hand infers transcripts without the use of a reference genome and is typically used for situations when a high quality reference genome is not available, for instance for profiling cancer tissues or for organisms which lack an accurately annotated reference genome.

#### 1.3.2.1 Reference based assembly using StringTie

In this part we look at methods for reference based assembly in details and introduce the algorithm used by StringTie. The first step of reference based assembly is similar to the first step of quantification of transcripts. RNA-seq reads are mapped to a reference genome using a splice aware aligner such as such as STAR [25] and HISAT [26]. An alternative splice graph (ASG) is created using reads mapped at each gene locus which contains information about all the transcripts that are expressed in the sample (Fig. 1.6). Each path through the ASG represents an isoform. StringTie uses a maximum-flow problem to assign reads to paths in the ASG in order to determine expressed transcripts and also to identify novel ones.

##### Alternative splice graph

A graph is defined as a pair $G = (V, E)$ where $V$ is a set of vertices or nodes and $E$ is a set of edges connecting the vertices. A graph is called a *directed* graph if its edges have directions and its called a *directed acyclic graph* (DAG), if the edges can never form a closed loop.

An alternative splice graph is a DAG where the nodes represent contiguous genomic segments spanned by reads and edges denote reads which align in between two exons indicating the presence of both the exons in one transcript. The ASG used by StringTie contains two additional vertices, the source $s$ and sink $t$ vertices. The addition of the source and sink vertices allows every transcript to be represented by an $s - t$ path in the ASG. The edges are weighted by the number of spliced alignments spanning them and the vertices are weighted by the average number of reads (per base coverage) that map to the genomic sequence represented by that vertex divided by the length of the segment.

##### Transcript finding algorithm used by StringTie

StringTie attempts to find the transcript whose corresponding $s-t$ path in the ASG has the largest per-base coverage and assigns the maximum number of reads possible to this transcript. The reads supporting the transcript with the largest coverage is removed from the ASG and this process is repeated till there are no more transcripts left in the ASG which are supported by reads.

The heaviest $s - t$ path in the ASG is found using a heuristic algorithm which starts at the vertex with the highest per base coverage and extends the path to first to the source and then to the sink. The adjacent node is chosen based on the highest number of paired end reads that is compatible with the path chosen so far.

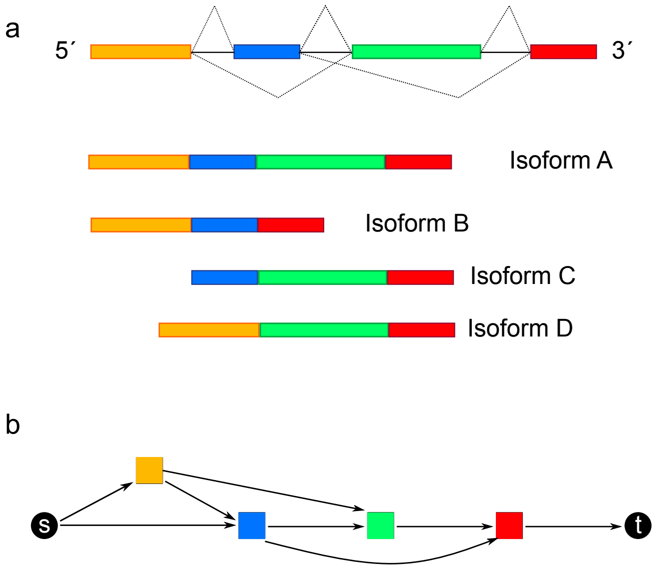

Figure 1.6: Splice graph : The four transcripts in (a) can be represented by the directed acyclic graph in (b). The nodes of the graph represent the exons of the transcripts. Two additional nodes, the start $s$ and sink $t$, colored in black are added to the graph in order to represent each transcript as an $s-t$ path in the graph.

Next a flow network (check [30] for details) is created using the identified heaviest path and the reads that contribute to it. The network represents the transcript with the highest coverage and the reads which map to this transcript contribute to the capacities of this network. It is important to note that all reads contributing to the heaviest transcript might not have originated from this transcript and thus should not be included while calculating its relative abundance. They might map because of shared exonic regions with other isoforms of the same gene. StringTie tries to explain the maximal number of reads that might contribute to the transcript inferred by the flow network and it does so by calculating the maximum flow the network, which gives the relative abundance of the transcript. The reads which contribute to the max-flow are removed from the initial ASG and the process of finding the heaviest path and assigning a max flow to it is repeated till the maximum weight of paths through the ASG falls below a threshold.

#### 1.3.2.2 Denovo assembly using Trinity

In the absence of a high quality reference genome, it is not possible or useful to map transcripts prior to assembly. In these cases transcripts need to be assembled *de novo*. In this section, we discuss some of the methods used for *de novo* assembly and take a look at the strategies used by Trinity [37] for assembling transcripts. The inability to map reads to a reference genome makes the problem of assembling transcripts more complicated. A common approach taken by *de novo* assemblers is to extract sequences of length $k$ (k-mers) from reads and construct a de Bruijn graph using these k-mers. Nodes in the graph represent individual k-mers and two nodes are connected if their sequences overlap by $k-1$ nucleotides. A graph like this can be used to enumerate all the splice variants for a certain gene. The specific algorithms used for creating the *de Bruijn* graph and enumerating transcripts differs between methods. Trinity contains three modules which are invoked in a sequential manner in order to assemble transcripts from RNA-Seq reads.

The first module, **inchworm**, assembles linear transcripts (without considering alternative splicing) by first extracting k-mers and their frequency of occurrence and storing them in a dictionary. It selects the most frequent k-mer in the dictionary and extends to the left and right by choosing the next frequent k-mer with an overlap of $k - 1$ to the current one. The extension process is continued till no overlapping k-mers are left and the resulting sequence represents a *contig*, a complete or incomplete transcript representing at most one complete splice variant for a gene. The next module, **chrysalis** uses contigs built by *inchworm* to create a de Bruijn graph. The contigs are pooled together into individual components if these sequences overlap by a certain length. Each component is used to generate a de Bruijn graph which is then used by the next module to enumerate transcripts. Reads are assigned to the components by choosing the component which shares the most $k - 1$ mers with the reads. Finally **butterfly**, the third module compacts the graph by certain operations such as collapsing multiple nodes in a linear path into one node, trimming spurious edges etc.. Finally it uses the reads pairings from *chrysalis* in a dynamic programming based algorithm to enumerate paths with the most evidence of contiguity which represent the assembled transcripts.

#### 1.3.2.3 Identifying transcripts from long reads

Third generation sequencing technologies have the potential to outcome the drawbacks of short reads, especially because the long reads span entire transcripts and thus eliminating the error prone probabilistic assignment of reads to transcripts. However long reads also come with some limitations such as high error rates, degradation of RNA prior to it being captured for sequencing, long molecules breaking during library preparation and in case of cDNA sequencing the reverse transcriptase failing to capture the full RNA molecule [29]. These shortcomings hamper the transcript detection rate for long reads and methods are forced to discard a number of reads from third generation sequencing technologies which do not cover entire transcripts [29].

Various methods have been developed to correct errors and to extract full length transcripts from error prone and/or fragmented long reads. Some of the most widely used tools are IsoSeq3 (included in smrtlink v5.1, PacificBioscience, Menlo Park, CA, USA) (previously ToFU [38]), TAPIS [39], SQANTI [40], StringTie2[29] and FLAIR [41].

StringTie [30], a method originally designed only for short read data was updated to be used for both short and long reads in a more recent release of the method called StringTie2 [29]. The general algorithm used by both the releases are quite similar as explained in Section 1.3.2.1. For the context of this thesis, we use StringTie2 for assembling transcripts from both short and long reads.

We also use the FLAIR pipeline to identify full length transcript isoforms from ONT reads. It first maps long reads to a reference genome using a spliced aligner for long reads such as minimap2 [42]. A known set of splice sites obtained from previous annotation or assembly using short reads is used to *correct* the mapped long reads where reads containing splice sites that are not included in the known set are discarded. The corrected reads are grouped according to splice junctions and transcription start and end sites and the FLAIR *collapse* module is used to summarize reads from groups to representative isoforms which are also called the *first-pass* isoform set. A further step involves mapping the raw reads to the *first-pass* isoform set and keeping isoforms which are supported by at least 3 reads with sufficiently high mapping quality ($MAPQ \geq= 1$).

## 1.4 Length based separation of mRNA to improve transcriptome quantification and reconstruction

The average length of an mRNA transcript is in the range of a several thousand neucleotides whereas the reads produced by second generation RNA-seq are only a few hundred neucleotides long. Despite many methodological advances, the accuracy of transcript-level inference methods developed over the last decade is severely limited due to the lack of long-range information contained in each individual short read. They perform particularly poor in the quantification of lowly expressed transcripts and transcripts from complex genes [43, 44, 45] that share large parts of their sequences due to alternative splicing which increases assignment ambiguity of short reads. Even more so when seeking to assemble novel transcripts, reference-based assembly methods typically miss several thousands of true transcripts and similarly construct thousands of incorrect sequences [30, 46]. Again, this applies in particular to complex genes expressing multiple isoforms [47, 48], which are highly prevalent in humans [49, 50] and frequently involved in disease pathogenesis [51, 52]. Multi-sample approaches such as the recently introduced PsiCLASS [46] try to address these limitations by assembling transcripts simultaneously across multiple RNA-seq samples. On the other hand, third-generation technologies such as those marketed by PacBio or Oxford Nanopore are able to read full-length transcripts but at a lower throughput, a higher error rate, and a higher cost per base [53].

This thesis introduces a new variant of the RNA-seq protocol that effectively breaks gene complexity by separating mRNAs according to their length into a small number of *bands* prior to their fragmentation. We define gene complexity as the number of expressed transcripts from a particular gene. The difficulty in quantification or assembly increases with an increase in gene complexity. The experimental deconvolution can aid in the computational reconstruction of a transcriptome by providing two types of long-range information. First, reads obtained in different bands must originate from different transcript species (of different length). This can reduce the phasing ambiguity of distant exons imposed especially by complex genes expressing many overlapping transcripts. Second, each band contains by design reads from transcripts of a certain length range. This can guide an algorithm to assemble or assign reads to transcripts only of a correct length. We extend and tailor state-of-the-art RNA-seq analysis methods for quantification, reference-based assembly, and de novo assembly to utilize the extra layer of information introduced in Ladder-seq in order to detect and quantify transcripts at an unprecedented level of accuracy and reveal transcripts that are invisible to conventional RNA-seq approaches.

More accurate transcript-level estimates from Ladder-seq will facilitate downstream differential analysis. Recent studies have highlighted the role of chemical modifications of mRNA as a heretofore unknown layer of regulation of gene expression. m$^6$A is the most abundant internal modification of mRNA in eukaryotic cells [10] and is involved in multiple aspects of mRNA biology including alternative splicing [11, 12, 13, 14]. We utilize the accurate analysis using Ladder-seq in a study of epitranscriptomic regulation of splicing in mouse neural progenitor cells (NPCs).

*Chapter 2*

# An experimental-computational approach to improve RNA-seq analysis

## 2.1 Ladder-seq

The difficulty in assembling and estimating abundance of transcripts from short read data is exacerbated by the fact that reads map to exonic regions which are common among overlapping transcripts. Computational methods usually assign such ambiguously mapping reads to transcripts using statistical models. This statistical assignment of reads to overlapping transcripts becomes harder with the increase in gene complexity due to the increase in the number of possible transcripts that a particular read can be attributed to. However all transcripts from a particular gene are not of the same length and they can be separated by their length into a small number of *bands* prior to their fragmentation (Fig. 2.1). In this chapter we introduce the principles of Ladder-seq and compare the data generated using Ladder-seq with publicly available RNA-seq datasets (of the same cell types as Ladder-seq ) in order to validate whether any bias was introduced due to the length separation. Parts of this chapter are taken from [54].

Ladder-seq is a new variant of the RNA-seq protocol that can be defined as an experimental-computational approach which effectively reduces gene complexity by separating mRNAs according to their length thus generating reads colored by the length range of the transcript that it originates from. This experimental deconvolution simplifies the analysis of RNA-seq data by providing extra layers of information on top of the sequence read out of fragments. Reads in Ladder-seq contain information about the length of the the set of originating transcripts from which a prior probability distribution of observing a random read from a transcript of a certain length observed in a certain band can be calculated. This prior can be leveraged by methods to fine tune analyses such as quantification, assembly and differential expression of transcripts.

We define *effective complexity* of a transcript as the maximum number of transcripts from the same gene that is left behind in a single band after the mRNA has been separated by length. Gene complexity and effective complexity are the same in case of conventional RNA-seq. The premise of Ladder-seq is to make transcript assembly and quantification easier by reducing the effective complexity of transcripts.

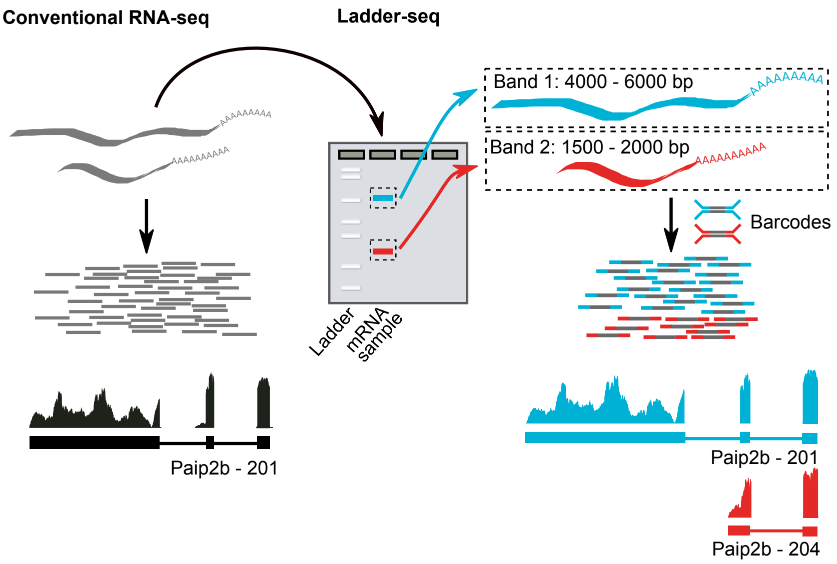

Figure 2.1: Ladder-seq uses a denaturing agarose gel to separate mRNA by length into discrete bands prior to library preparation and sequencing. Each band contains transcripts of a certain length range that depends on the location of cuts through the gel. The originating band of the resulting reads is tracked using barcodes. In our dataset of mouse neural progenitor cells, Ladder-seq reveals transcript Paip2b-204 that contains intronic sequence of transcript Paip2b-201.

## 2.2 Generation of Ladder-seq libraries of mouse neural progenitor cells

To achieve mRNA separation by transcript length, we performed denaturing gel electrophoresis. All samples were run on the same denaturing agarose gel. After electrophoresis, each sample was cut into 7 bands, each containing transcripts of a given length range. Slicing of the gel into 7 bands was guided by a single-stranded RNA ladder running on the same gel and the location of cuts were at approximately 1000 bp, 1500 bp, 2000 bp, 3000 bp, 4000 bp and 6000 bp.

These cuts effectively reduce gene complexity in our data set (Fig. 2.2)a by partitioning transcripts expressed per gene into different subgroups. We denote the size of each subgroup as its effective complexity. For each gene the maximum effective complexity is the largest number of its expressed transcripts contained in the same band.

In theory, the largest reduction in effective complexity would be achieved by separating each transcript of a particular length into a *band* of its own. On the other hand, fewer cuts might be sufficient to achieve a similar improvement over conventional RNA-seq for species with a less complex transcriptome. We examined the reduction of the mean and maximum effective complexity using a R script which we provide in our repository. It visualizes (see 2.2 for an example) and summarizes the distributions of original gene complexities and resulting effective complexities using descriptive statistics either genome-wide or for a given set of genes of interest, based on a related RNA-seq data set of a given species.

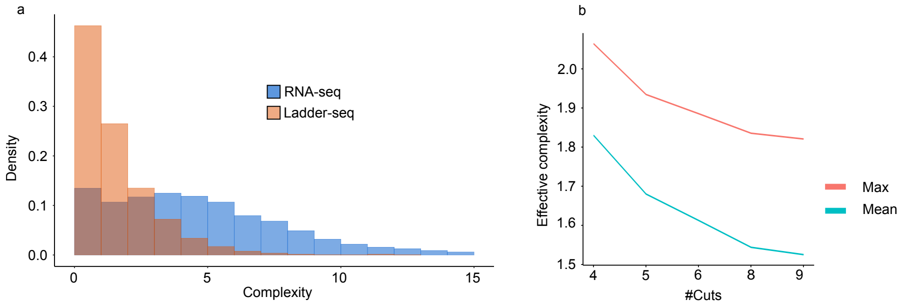

Figure 2.2: (a) Reduced (effective) gene complexity in Ladder-seq. We estimate transcript expression in *Mettl14* KO sample 1 using kallisto on Ladder-seq reads pooled across bands and show the histogram of gene complexity measured as the number of expressed transcripts. In Ladder-seq, we partition the set of expressed transcripts into 7 bands and count the number of transcripts contained in each band according to their annotated length (plus 200 nt average poly(A) tail size [55]), assuming cuts at 1000 bp, 1500 bp, 2000 bp, 3000 bp, 4000 bp and 6000 bp. The resulting histogram of "effective" gene complexity shows an increased fraction of gene bands with low complexity. (b) Reduction in maximum and average effective complexity with an increase in the number of cuts. The set of 6 cuts is identical to the cuts used in our experiments (1000 bp, 1500 bp, 2000 bp, 3000 bp, 4000 bp, 6000 bp). We removed the cut at 1500 bp and in addition the cut at 2000 bp to simulate sets of 5 and 4 cuts, respectively. We added cuts at 2500 bp and 3500 bp to simulate a set of 8 cuts and additional a cut at 5000 bp to simulate 9 cuts.

After electrophoresis, mRNA from each band of each sample was extracted from the agarose gel and equal volumes per band were used for cDNA library construction. Libraries were multiplexed 1:1 for sequencing in the Illumina HiSeq 2500, yielding approximately 100 million 2×76-bp paired-end reads per sample. mRNA from each band of each sample was extracted from the agarose gel and equal volumes per band were used for cDNA library construction. Each band from each sample was given a unique barcode to track the originating band (per sample) of each read.

We generated Ladder-seq datasets from wild-type (WT) and *Mettl14* knock-out (cKO) mouse neural progenitor cells (NPCs). *Mettl14* encodes for a methyltransferase necessary for m$^6$A methylation of mRNA. KO mice have a targeted deletion of *Mettl14* exons 7, 8 and 9 that is only present in NPCs and their progeny [56]. Four independent replicates were prepared per genotype.

## 2.3 Validation of experimental quality

To ensure that our electrophoresis protocol effectively separates denatured mRNAs, we performed a trial electrophoresis run, after which mRNA from each band was run again on a denaturing agarose gel with each band loaded into a separate lane.

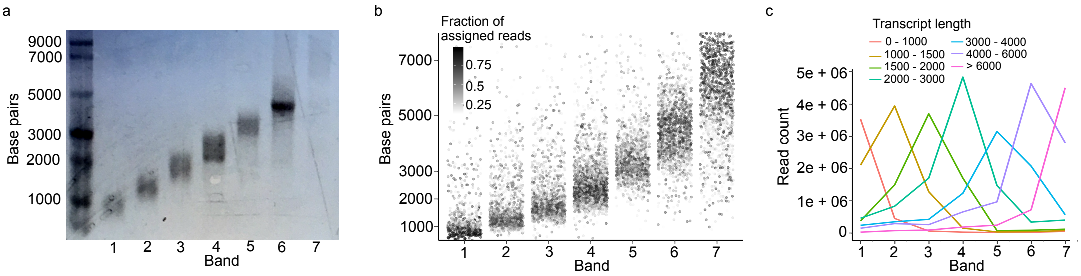

Figure 2.3: (a) Assessment of length separation by denaturing gel electrophoresis. Length separated mRNA was run on a new denaturing agarose gel with each band loaded into a separate lane. (b) In silico gel: For every annotated transcript, the intensity of a point with y-coordinate equal to its annotated length (plus 200 nt average poly(A) tail size [55]) shows the fraction of reads obtained from each band (x-axis) that can be assigned unambiguously to it. (c) Distribution of reads across bands that can be unambiguously assigned to annotated transcripts.

Fig. 2.3a and table table B.1 shows that mRNA was effectively separated into 7 distinct length ranges with a certain degree of overlap between consecutive bands.

We created an *in silico gel* by using the *pseudoalignment* feature in conventional kallisto to get a set of reads that map unambiguously to annotated transcripts. Fig. 2.3b shows the percentage of reads from each band that map uniquely to transcripts. The intensity of the points show that higher bands have more reads that map to longer transcripts and vice versa rendering an image very similar to the actual picture of the agarose gel (Fig. 2.3a). This serves as a further confirmation of the separation mRNA into different bands depending on their length.

### 2.3.1 Estimating mRNA migration

The experimental and in silico gel in Fig. 2.3(a) and Fig. 2.3(b) show that reads usually end up in corresponding bands based on the length of transcripts from which they came from. However we observe a smear indicating the existence of a distribution according to which reads migrate across the gel. Fig. 2.3.1(c) is another representation of the *in silico gel* which shows the uniquely mapping read count from each band with the colors representing the different length ranges. We observe peaks of particular colors representing shorter transcript lengths in smaller bands and vice versa. This gives a better view of the smear as we see that uniquely mapping reads from a certain length range have a peak in the expected band but also counts in the neighbouring bands.

In order to estimate the migration pattern of a transcript of length $\ell$ through the agarose gel across $k$ bands, we introduce probability mass function $f(x)$ over discrete random variable $x \in [k] := \{1, \ldots, k\}$, which indicates the band to which transcripts of length $\ell$ migrate. If we observe reads sampled from transcripts of length $\ell$ in bands $X_1, \ldots, X_n \in [k]$, then we simply count how often reads are obtained in a given band

and take the relative frequency as density estimate:

$$\hat{f}(i) = \frac{\sum_{j=1}^{n} \mathbb{1}(X_j = i)}{n},$$

where $\mathbb{1}$ is the indicator function that takes value 1 if its argument evaluates to true and 0 otherwise. To obtain reads for which we can infer the originating transcript with high confidence, we select reads that uniquely map to a single annotated transcript. More precisely, we run the kallisto pseudoalignment step and select all reads that are compatible with only a single transcript according to the *NH* tag.

In addition, we account for potentially incomplete transcript annotations that may cause reads sampled from unannotated transcripts (of different length) to negatively affect our migration estimate of a transcript (length) it was wrongly assigned to. To this end, we assemble transcripts using StringTie2 from reads pooled across bands and aligned using STAR. We augment the transcript annotation with novel transcripts before running the kallisto pseudoalignment to obtain a more conservative selection of uniquely mapping reads. We do not consider reads mapping (uniquely) to newly assembled transcripts. We further restrict observations to reads that uniquely map to protein-coding transcripts (Ensembl release 95), which are typically annotated more accurately, and which we were able to confirm to be expressed through the StringTie2 assembly on the intron chain level. We require a minimum number of 50 reads to uniquely map to a transcript of length at most 8000 bp to be considered in our estimation. The resulting set of reads along with their band of origin identified by the barcode constitute observations $X_1, \ldots, X_n$ for the length of the transcript they uniquely align to.

If no (high-quality) transcript catalog is available based on which uniquely mapping reads can be identified, e.g. in *de novo* assembly or in the case of poorly studied species, synthetic RNA spike-in controls of varying lengths [57] can be used to similarly estimate transcript migration error from reads mapping to spike-in controls.

Since transcripts of similar length show similar migration patterns through the gel [58], we combine reads uniquely mapping to transcripts of a certain length range to more reliably estimate $f(x)$ based on a larger number of reads. Starting from the shortest transcripts, we greedily define transcript length ranges as the shortest possible length intervals longer than 100 bp which contain at least 50 different transcript species to which at least a total number of $700,000$ reads map uniquely. For each of these length ranges, we estimate one probability mass function $f(x)$ as described above. The resulting length ranges are listed in Table B.3.

### 2.3.2 Comparison with publicly available datasets

#### 2.3.2.1 Correlation

In order to examine if the process of separating transcripts by their length prior to sequencing introduced any systematic bias or any shift in correlation or transcript detection rate, we compared transcript abundances calculated using pooled reads from all bands in Ladder-seq data to publicly available conventional RNA-seq data. The Pearson correlation coefficients of log2 transformed transcript per million (TPM) values demonstrate high technical reproducibility of our Ladder-seq protocol (r=0.96-0.98; Fig. 2.4). We added a pseudo count of 0.1 to the tpm values prior to the log2 transformation in order to avoid very large negative numbers. Furthermore, transcript expression levels were well-correlated between each of the four WT Ladder-seq samples and three conventional RNA-seq reference data sets (without length separation) from WT NPCs rep1, rep2 and rep3 (r=0.81-0.82; Fig. 2.5 and Appendix Table B.5), despite using different experimental batches. Pearson correlation coefficients of our Ladder-seq samples were similar to those of 5 public RNA-seq samples of mouse NPCs [40, 59] (Appendix Tables B.4 and B.5), which holds also

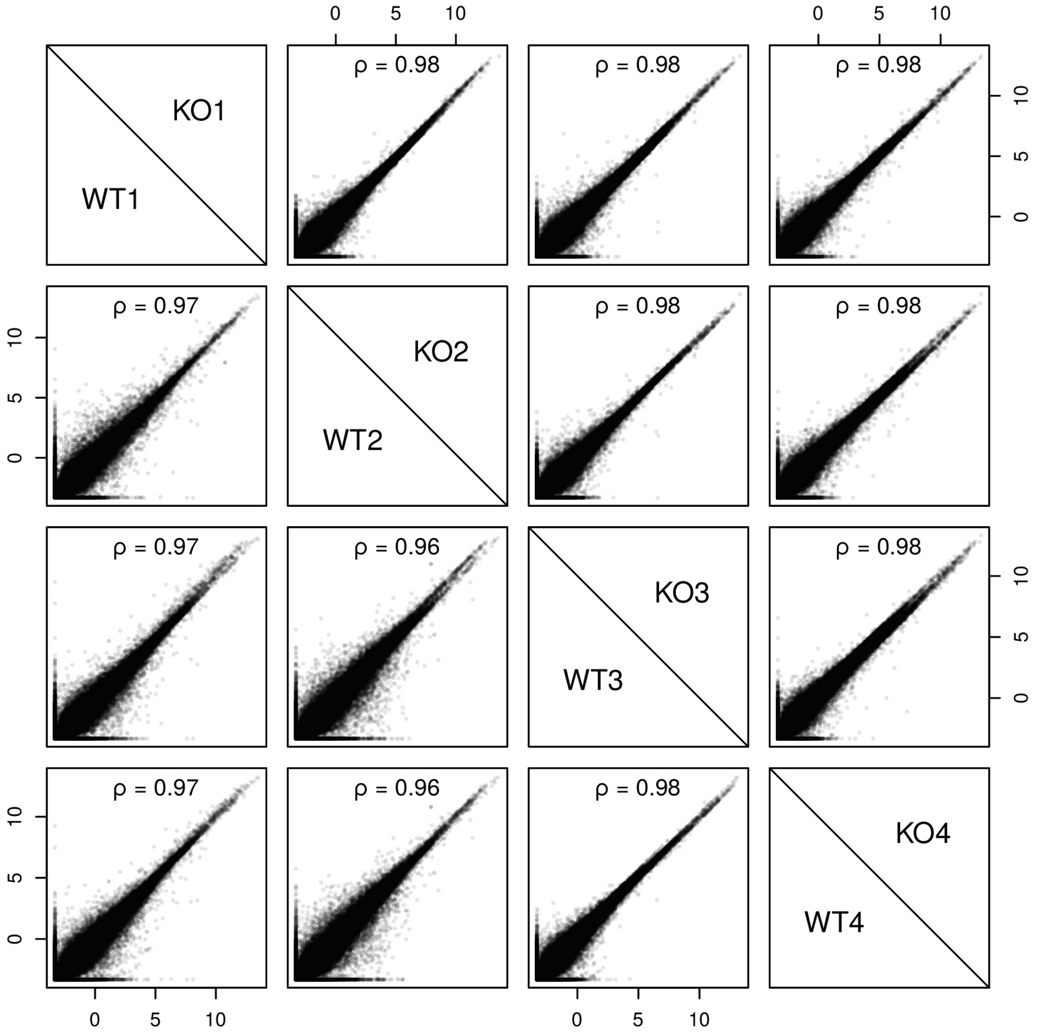

Figure 2.4: Scatter plot of $\log_2$-transformed transcript expression values (TPM) estimated from 4 WT and 4 KO NPC Ladder-seq samples. Pairwise comparisons between WT samples are shown in the lower triangular, between KO samples in the upper triangular of the matrix. Pearson correlation coefficients are shown for all pairwise comparisons. Transcript expression was estimated by kallisto from pooled reads ignoring their separation into bands.

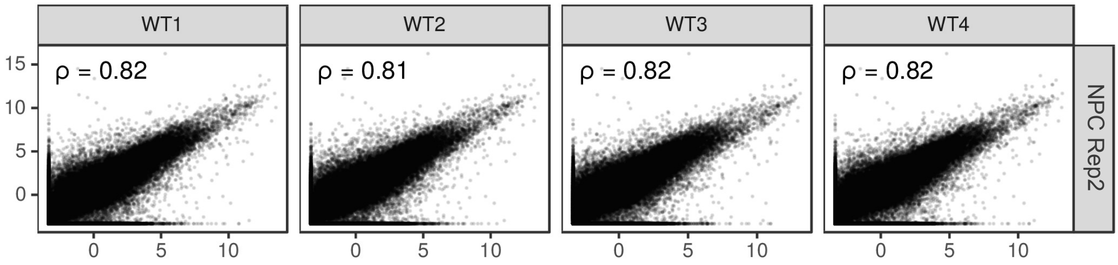

Figure 2.5: Scatter plot of $\log_2$-transformed transcript expression values (TPM) estimated from 4 Ladder-seq samples (90 mio paired-end reads per sample) and a conventional RNA-seq sample (no length separation, 30 mio single-end reads) from WT NPCs. Transcript expression in the 4 Ladder-seq samples was estimated by kallisto from pooled reads ignoring their separation into bands. Pearson correlation coefficients are shown for all pairwise comparisons of the 4 Ladder-seq samples with reference RNA-seq sample NPC Rep2. Correlation with two other reference RNA-seq samples are given in Appendix Table B.5.

when correlation was stratified by transcript length ranges that follow the location of cuts used in our experiments (Fig. 2.6).

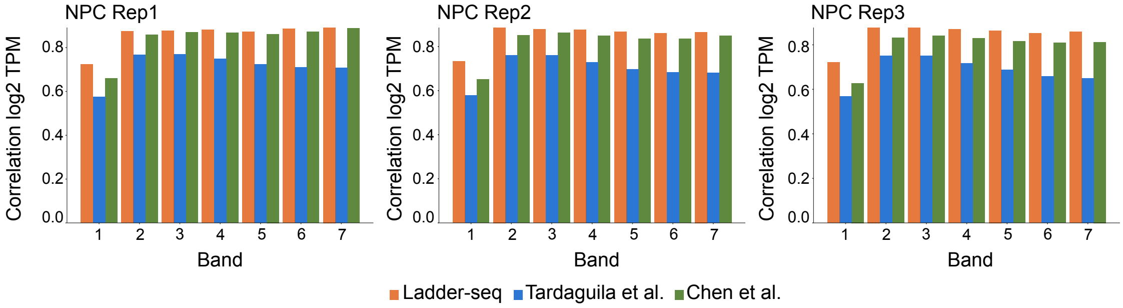

Figure 2.6: Pearson correlation of transcript expression stratified by band. The values reported are mean correlation coefficients across 4 WT Ladder-seq NPC samples and across 2 and 3 regular RNA-seq samples of mouse WT NPCs by Tardaguila et al. [40] and Chen et al. [59], respectively (variance not visible). Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. For each band, $\log_2$-transformed expression values (TPM) of transcripts with annotated length falling in the corresponding range were compared to the 3 reference RNA-seq samples from WT NPCs.

#### 2.3.2.2 Checking for length bias based on residuals

Transcripts with low correlation did not differ significantly in length from highly correlated tanscripts (Fig. 2.7). Using a linear regression model to predict abundance calculated from NPC Rep2 by abundance calculated from Ladder-seq WT data, we divided transcripts into two sets, one with a $|\text{residual}| > 1$ and the other $|\text{residual}| \leq 1$. We compared the distribution of lengths of these two groups of transcripts and using the non-parametric Wilcoxon's test we found no evidence of the means of the distributions being different, signified by extremely high p-values ($> 0.94$). This is further confirmation that there is not systematic bias by length of transcripts.

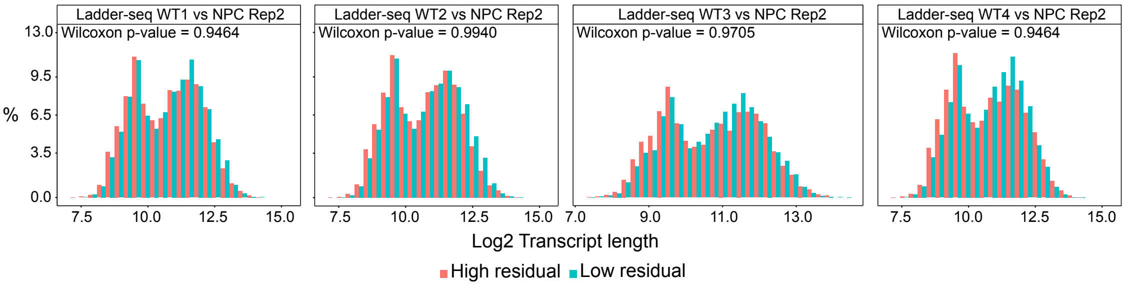

Figure 2.7: Comparison of transcript length between high and low residual transcripts. After fitting a linear regression model between estimated transcript expressions ($\geq$ 1TPM, $\log_2$-transformed) in WT Ladder-seq NPC samples and reference RNA-seq sample NPC Rep2, high residual transcripts were defined as those with $|\text{residual}| > 1$, low residual transcripts as the remaining transcripts ($|\text{residual}| \leq 1$). Transcript expression was estimated by kallisto, pooling reads from all bands in Ladder-seq samples. The two groups of transcripts were compared using Wilcoxon's test.

#### 2.3.2.3 Transcript detection

The total number of detected annotated transcripts is highly similar between Ladder-seq and conventional RNA-seq (Fig. 2.8 and 2.9), and the detection rate increased with transcript length as previously reported [60] (Fig. 2.10 and 2.11).

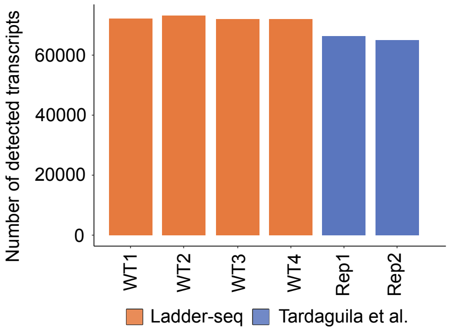

Figure 2.8: Number of detected transcripts by Ladder-seq and conventional RNA-seq. An annotated transcript is considered detected if the estimated count is at least 1. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. The 4 WT Ladder-seq NPC samples are compared to the 2 RNA-seq samples of WT NPCs from Tardaguila et al. [40]. Ladder-seq data sets were randomly sampled to an identical read depth as the data sets by Tardaguila et al. (94 mio single-end reads).

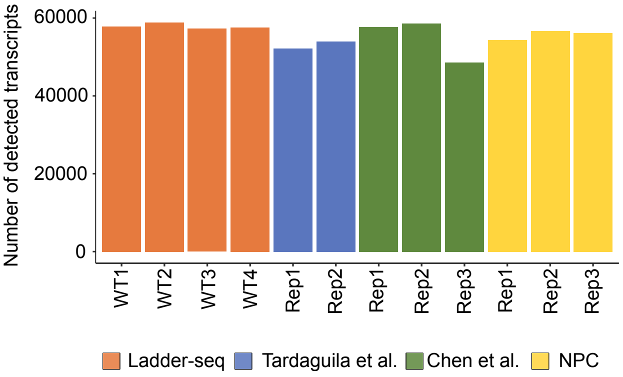

Figure 2.9: Number of detected transcripts by Ladder-seq and conventional RNA-seq. An annotated transcript is considered detected if the estimated count is at least 1. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. The 4 WT Ladder-seq NPC samples are compared to RNA-seq samples of WT NPCs from Tardaguila et al. [40], Chen et al. [59], and the 3 reference samples (NPC). All data sets were randomly sampled to an identical read depth (20 mio single-end reads) as the reference NPC data sets.

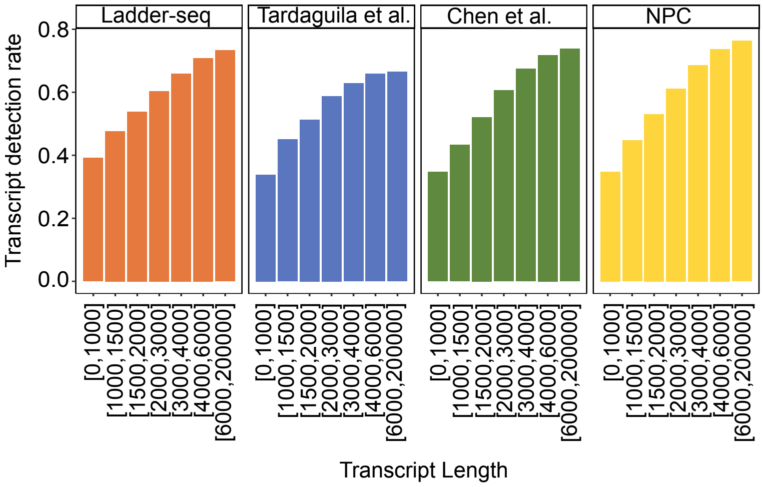

Figure 2.10: Transcript detection rate of Ladder-seq and RNA-seq. The fraction of transcripts with estimated count at least 1 is stratified by transcript length, using ranges that follow the location of cuts used in our experiments. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. The 4 WT Ladder-seq NPC samples are compared to RNA-seq samples of WT NPCs from Tardaguila et al. [40], Chen et al. [59], and the 3 reference samples (NPC). All data sets were randomly sampled to an identical read depth as the reference NPC data sets (20 mio single-end reads).

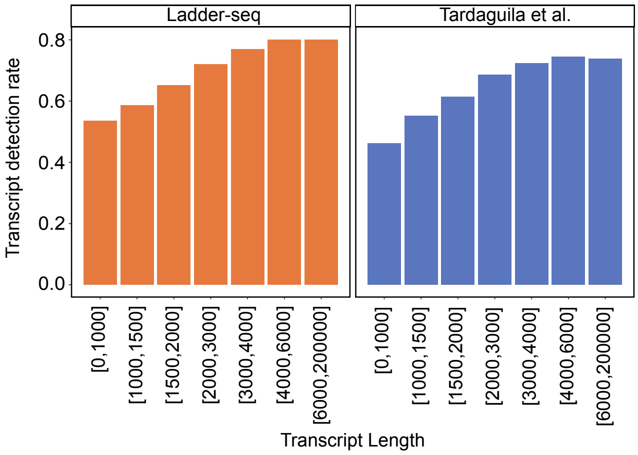

Figure 2.11: Transcript detection rate of Ladder-seq and RNA-seq. The fraction of transcripts with estimated count at least 1 is stratified by transcript length, using ranges that follow the location of cuts used in our experiments. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. The 4 WT Ladder-seq NPC samples are compared to the 2 RNA-seq samples of WT NPCs from Tardaguila et al. [40]. Ladder-seq data sets were randomly sampled to an identical read depth as the data sets by Tardaguila et al. (94 mio single-end reads).

*Chapter 3*

# Benchmark framework

Methods for assembly and quantification need to be tested for their accuracy, meaning that the genes or transcripts or their relative abundances that are predicted or estimated by these methods need to be verified against a *ground truth*, or in other words, information that is known to be true. Ground truth for RNA-seq data does not exist since we don't precisely know the actual set of transcripts and their levels of expression in a sample [61]. As a result, benchmarking for analysis methods for RNA-seq data is often done using simulated data [61]. Some of the most widely used simulators are RSEM [62], Polyester [63] and FluxSimulator [64]. Most of these simulators train some statistical model parameters from *real* RNA-seq data and then use these parameters to simulate reads.
In this chapter we describe in details the methods that we used to create simulated Ladder-seq data from conventional RNA-seq data and also the benchmarking strategies that we used to evaluate the methods that we have extended to be used with Ladder-seq data. Parts of this chapter are taken from [54].

## 3.1 Simulation

We extend RSEM [62] by an additional in silico length separation step that includes the introduction of migration errors to simulate data with similar characteristics as data generated by our novel Ladder-seq protocol. Since the effectiveness of the experimental deconvolution of reads into different bands by Ladder-seq depends on the differences in lengths of expressed, overlapping transcripts, we simulated reads from a transcriptome using abundances and error profiles learned from a real data set. Following the approach in [27], we simulated 30 million and 75 million 2×75-bp paired-end reads from transcripts whose abundances were estimated by RSEM from sample NA12716_7 of the Genetic European Variation in Health and Disease (GEUVADIS) [65].

### 3.1.1 Simulating Ladder separation

Given the RNA-seq reads produced by the simulator, we generate a matching Ladder-seq sample by assigning each read randomly to one of a fixed number of bands (here 7) to introduce in silico length separation. This random assignment follows the probability mass function estimated from our NPC Ladder-seq sample KO 1, given the length of the transcript the read originates from (provided by the simulator). Fig. 3.1 gives an overview of the benchmark strategy.

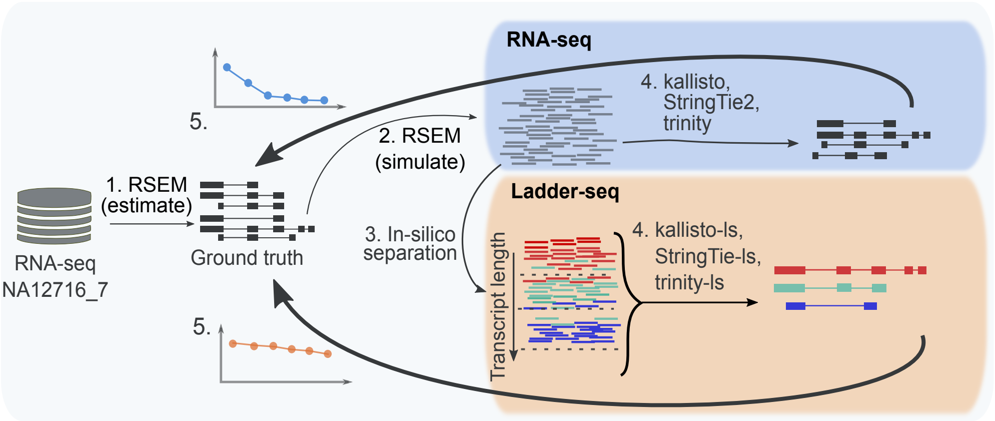

Figure 3.1: Overview of the benchmark strategy. 1. The ground truth transcriptome including abundances and error profile is calculated by RSEM from GEUVADIS sample NA12716_7. 2. Reads are simulated from the ground truth transcriptome by RSEM to obtain RNA-seq samples of different sequencing depths. 3. A matching Ladder-seq sample is obtained by separating reads in silico according to probability mass functions estimated from our NPC Ladder-seq sample (and variants thereof). 4. Transcripts are quantified and assembled by our Ladder-seq tailored transcript analysis methods kallisto-ls, StringTie-ls, and Trinity-ls from the Ladder-seq sample, while their conventional counterparts are run on the corresponding RNA-seq sample. 5. The results are compared to the ground truth to evaluate and compare their accuracy.

### 3.1.2 Different levels of separation

To demonstrate how a more accurate experimental separation of transcripts by length can benefit transcript-level inference from Ladder-seq, we additionally simulated three Ladder-seq experiments that introduce gradually decreasing levels of migration errors. For every transcript length range for which we have estimated probability mass function $f(x)$ from our NPC Ladder-seq sample, we halve the relative frequency of reads in every band as we move further away from its mode and normalize all values to sum up to 1. More precisely, for bands numbered consecutively from 1 to $k$, let $m$ denote the band that contains the mode of $\hat{f}(x)$ estimated for a given length range. Then

$$f^1(i) = \frac{\hat{f}(i)/2^{|i-m|}}{\sum_{j=1}^{k} \hat{f}(j)/2^{|i-m|}} \tag{3.1}$$

Similarly, $f^2(x)$ and $f^3(x)$ are obtained by replacing $\hat{f}$ in (3.1) by $f^1(x)$ and $f^2(x)$, respectively. By randomly assigning simulated reads according to probability mass functions $f^i(x)$, $i = 1, 2, 3$, instead of $\hat{f}(x)$, we obtain three additional Ladder-seq datasets with reduced levels of migration errors.

Finally, we simulated a most optimistic Ladder-seq experiments that is able to perfectly separate transcripts by length, without introducing any migration error.

This leads to a degenerate probability mass function for each length range implied by the 7 in silico cuts in which the read band is a constant random variable that takes only a single value, the correct band corresponding to that length range.

## 3.2 Benchmarking

We used the same metrics as in a benchmark of transcript quantification methods [43] to measure the accuracy of conventional RNA-seq and Ladder-seq based estimates of transcript expression. MARD denotes the arithmetic mean of absolute relative differences, calculated as $|i-j|/(i+j)$ for estimated and ground truth counts $i$ and $j$, respectively. We excluded transcripts with zero estimates by both methods, that is, if $i+j=0$. Pearson correlation was calculated between $\log_2$ transformed TPM values, after adding 0.1 TPM. The advantage of Ladder-seq seq to conventional RNA-seq is due to the reduction of effective complexity achieved by separating transcripts based on their length. Quantification of genes expressing multiple transcripts benefit more from the reduction of effective complexity by Ladder-seq as compared to genes expressing a single or a few transcripts. In order to examine the improvement of accuracy with an increase in gene complexity, we compare the correlation and MARD values of Ladder-seq to conventional quantification based on the complexity of their originating gene. More precisely, we group transcripts into 10 groups, starting from genes with two expressed transcript to genes with 10 expressed transcripts.

Consistent with previous studies [66, 29], the accuracy of reference-based and *de novo* assemblies is evaluated using sensitivity defined as TP/(TP+FN) and precision defined as TP/(TP+FP), where true positives (TP) denote correctly assembled transcripts, false negatives (FN) true transcripts missing in the assembly, and false positives (FP) wrongly assembled transcripts. We considered a transcript truly expressed if reads sampled by RSEM in the 30 million reads data set fully cover the transcript, and if it was estimated by RSEM to be expressed in GEUVADIS sample NA12716_7 with at least 0.1 TPM. An identical ground truth transcriptome facilitates comparison of sensitivity and precision values between different sequencing depths and so we used the ground truth set of transcripts from the 30 million reads data set to evaluate the assembly from the 75 million reads data set. We used the same transcriptome for comparing the results of denovo assemblies between the methods. The next section gives a detailed overview of process of generating the ground truth.

### 3.2.1 Generation of the ground truth transcriptome for benchmarking assembly methods

The reads simulated using RSEM contain information about the transcript and the exact position in the transcript that they were simulated from. Using this information we create a *bed* file containing transcript names and locations of contiguous segments of a transcript which are covered by individual reads. We use *bedtools* [67], specifically its *merge* and *intersect* utilities to merge the locations covered by individual reads and retain transcripts which are covered completely from one end to the other. Additionally, in order to get a correct view of the complexity of assembling transcripts we group overlapping transcripts into contiguous loci and calculate gene complexity on these groups instead of calculating complexity based on annotation. The grouping is achieved using the program *groupGenes* [32] which substitutes the gene-id of a transcript by the locus-id of its group based on shared sub-exons [32]. Finally we divide the set of fully covered and grouped transcripts by gene complexity using the utility *gtfFilter* [32] to complexities of 1 to 10 and 10+ transcripts expressed per gene locus and keep transcripts with an expression level more than 0.1 transcripts per million (TPM).

### 3.2.2 Gffcompare

As in [30, 31], we used GffCompare [68] to compare transcripts assembled by StringTie2 or StringTie-ls to truly expressed transcripts. GffCompare defines an assembled transcript as correct if it shares the exact same sequence of introns with a true transcript. In the *de novo* assembly benchmark, correct assemblies by Trinity and Trinity-ls need to be identified through an alignment of their sequences which we computed using BLAT [69]. Applying commonly used criteria [35, 70], we require the sequences to align with 95% identity and at most 1% insertion and deletion rate, and apply transcript coverage cut-offs of 80%, 85%, 90%, and 95%.

Chapter 4

# Transcript Quantification

Reads that map to a unique genomic position often cannot be assigned unambiguously to one of a gene's transcripts, since alternatively spliced isoforms may overlap in genomic coordinates. Transcript quantification methods therefore use a statistical model of RNA-seq to probabilistically assign reads to transcripts such that estimated transcript abundances can best explain the observed reads. In this chapter we introduce our extension of the statistical model of RNA-seq to our new protocol Ladder-seq and our implementation of an Expectation-Maximization (EM) algorithm that infers maximum likelihood (ML) estimates of transcript abundances in this model. Parts of this chapter are taken from [54].

## 4.1 kallisto-ls

### 4.1.1 Fractional assignment of reads

There are two primary measures of transcript expression, the fraction of transcripts and the fraction of nucleotides of the transcriptome that is made up of a given gene or transcript [71]. We denote these quantities as $\rho_t$ and $\alpha_t$ respectively (equations 4.1 and 4.2).

$$\alpha_t = \frac{\rho_t \times l_t}{\sum_{t' \epsilon T} \rho_{t'} \times l_{t'}} \tag{4.1}$$

$$\rho_t = \frac{\frac{\alpha_t}{l_t}}{\sum_{t' \epsilon T} \frac{\alpha_{t'}}{l_{t'}}} \tag{4.2}$$

The fundamental assumption of RNA-seq is that the fraction of reads derived from a transcript $t$ is a function of $\alpha_t$ and that $N_t/N$ approaches $\alpha_t$ as $N \to \infty$ where $N_t$ is the number of fragments that map to transcript $t$ and $N$ is the total number of fragments [71]. Thus $\hat{\alpha_t}$, an estimator for $\alpha_t$ can be calculated using equation 4.3.

$$\hat{\alpha_t} = \frac{N_t}{N} \tag{4.3}$$

Transcript expression is estimated using a generative model of the RNA-seq read sequencing process. The model parameters are $\theta = \theta_1, ...\theta_T$ which are the expression levels of transcripts $1, ...T$. The observed data of the model consists of $\mathbf{n} = 1...N$ read sequences. The goal is to estimate the model parameters which maximize the probability of the observed data which we denote as $Pr(\mathbf{n}|\theta)$.

The exact mapping of reads to transcripts are unknown and this makes the estimation of the model parameters that maximize the probability of the observed data difficult. In order to simplify this process, we introduce some hidden binary variables which define the mapping of reads to transcripts. The hidden variable for a read $n$ originating from transcript $t$ is defined by $Z_{nt}$ such that

$$Z_{nt} = \begin{cases} 1, & \text{if read } n \text{ comes from transcript } t \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

The complete likelihood of the data along with the hidden variables is given by equation 4.5.

$$Pr(\mathbf{n}, Z|\theta) \tag{4.5}$$

Since the distribution of the hidden variables is also unknown, the EM algorithm is used to estimate the model parameters which maximize equation 4.5. EM alternates between estimating the hidden variables based on current estimates of the model parameters, $\hat{\theta}$ and then recalculating the parameters from these estimates. The E step of EM calculates the expected value of the posterior distribution of the hidden variables, $Pr\left(Z|\mathbf{n}, \hat{\theta}\right)$ and the M step then maximizes the likelihood of the data $Pr\left(\mathbf{n}, Z|\theta\right)$ w.r.t. $Pr\left(Z|\mathbf{n}, \hat{\theta}\right)$ and updates the estimates of the model parameters $\hat{\theta}$. The steps are repeated with the updated values of the model parameters till they converge.

The steps of the EM algorithm are as follows:

1. Initialization : All transcripts are initialized with the same abundance, $\hat{\rho_t}^{(0)} = \frac{1}{T}$, where $T$ is the total number of transcripts.

2. For $m = 1, 2, ...$ repeat

   - E step: If a read $R_n$ maps to a set of transcripts $e$ then the then the expected value of the hidden variable $Z_{nt}$ is calculated based on the current values of the model parameters $\hat{\rho_t}$ (Equation 4.6).

   $$E_{Z|n,\hat{\rho_t}}\left(Z_{nt}^{(m)}\right) = \begin{cases} \frac{\rho_t^{(m)}}{\sum_{t' \epsilon e} \rho_{t'}^{(m)}}, & \text{if } t \epsilon e \\ 0, & \text{otherwise} \end{cases} \tag{4.6}$$

   - M step: This step maximizes the likelihood of observing the data along with the expected values of the hidden variables to yield the maximum likelihood estimators of the model parameters. The values of $\hat{\alpha_t}$ and $\hat{\rho_t}$ are updated by the equations 4.8 and 4.7.

$$\hat{\alpha_t}^{(m+1)} = \frac{1}{N}\left(\sum_{n=1}^{N} E_{Z|n,\hat{\rho_t}}\left(Z_{nt}^{(m)}\right)\right) \tag{4.7}$$

$$\hat{\rho_t}^{(m+1)} = \frac{\frac{\hat{\alpha_t}^{(m+1)}}{l_t}}{\sum_{t'\epsilon T}\frac{\hat{\alpha_t'}^{(m+1)}}{l_{t'}}} \tag{4.8}$$

3. The E and M steps are repeated with the updated values of $\hat{\rho_t}$ until they converge.

### 4.1.2 Conventional kallisto

Kallisto is a pseudo-alignment based RNA-seq quantification method which uses the EM algorithm as described above to estimate transcript quantification levels. Kallisto is based on the following likelihood function [27] of RNA-seq:

$$L(\alpha) \propto \prod_{e\in E}\left(\sum_{t\in e}\frac{\alpha_t}{l_t}\right)^{c_e} \tag{4.9}$$

It counts the number of fragments $c_e$ that cannot be distinguished by the set of transcripts $e$ they are compatible with and are thus considered equivalent. $l_t$ denotes the effective length [72] of transcript $t$ and parameters $\alpha_t$ the probability of obtaining a fragment from a transcript $t$.

### 4.1.3 Adapting the conventional likelihood function to Ladder-seq

Most statistical models assume that reads are sampled proportionally to transcript abundances and their lengths and include parameters for the length distribution of fragments if sequenced from both ends (paired-end reads), for sequence bias and positional bias, and for sequencing errors. The main difference in the Ladder-seq protocol is that the barcode of each read indicates the band from which its originating transcript was extracted. The read's band contains transcripts of a specific length range and thus provides valuable information when trying to probabilistically resolve its assignment ambiguity between transcripts of different length.

In a perfect scenario, all the transcripts of a certain length would migrate to the same band and the abundance estimation could be performed on each band separately. However in reality, all transcripts don't migrate to the band that they are supposed to. The migration of transcripts follows a distribution which can be measured from read mapping data as we have shown in section 2.3.1. As a result of the probabilistic migration of transcripts across bands, reads from many bands can map to a set of transcripts of particular length.

After estimating migration patterns in a Ladder-seq sample, kallisto-ls uses an EM algorithm similar to kallisto to infer maximum likelihood estimates of transcript abundances in our statistical model of Ladder-seq.

In Ladder-seq, we observe fragments that originate from transcripts in different bands. The probability of obtaining a fragment from a transcript $t$ in band $b$ then is $\alpha_t\beta_{tb}$, where $\beta_{tb}$ denotes the fraction of transcript $t$ in band $b$ which we precompute in $\hat{f}(b)$ for each range of transcript lengths as described above. If we split equivalence class counts $c_e$ between $k$ different bands, i.e.

$$c_e = \sum_{b=1}^{k} c_{eb}\,,$$

then the likelihood function for Ladder-seq becomes

$$L(\alpha) \propto \prod_{e\in E}\prod_{b=1}^{k}\left(\sum_{t\in e}\frac{\alpha_t\beta_{tb}}{l_t}\right)^{c_{eb}} \tag{4.10}$$

We extend the EM algorithm implemented in kallisto to find the values of $\alpha$ that maximize likelihood (4.10). Based on the modified likelihood, the hidden variable for a read $n$ from band $b$ originating from transcript $t$ is defined by $Z_{ntb}$ such that

$$Z_{ntb} = \begin{cases} 1, & \text{if } R_n \text{ comes from band } b \text{ and transcript } t \\ 0, & \text{otherwise} \end{cases} \tag{4.11}$$

The E and M steps of the EM algorithm are modified as follows:

1. Initialization : All transcripts are initialized with the same abundance, $\hat{\rho_t}^{(0)} = \frac{1}{T}$, where $T$ is the total number of transcripts.
2. For $m = 1, 2, ...$ repeat
   - E step: If a read $R_n$ from a band $b$ maps to a set of transcripts $e$ then the then the expected value of the hidden variable $Z_{ntb}$ is calculated based on the current values of the model parameters $\hat{\rho_t}$ (Equation 4.12).

$$E_{Z|n,\hat{\rho}_t}\left(Z_{ntb}^{(m)}\right) = \begin{cases} \frac{\rho_t^{(m)}\times\beta_{tb}}{\sum_{t'\epsilon e}\rho_{t'}^{(m)}\times\beta_{t'b}}, & \text{if } t\epsilon e \\ 0, & \text{otherwise} \end{cases} \tag{4.12}$$

   - M step: This step maximizes the likelihood of observing the data along with the expected values of the hidden variables to yield the maximum likelihood estimators of the model parameters $\theta$. Thus the values of $\hat{\alpha_t}$ and $\hat{\rho_t}$ are updated by the equations 4.14 and 4.13.

$$\hat{\alpha_t}^{(m+1)} = \frac{1}{N}\left(\sum_{b=1}^{k}\sum_{n=1}^{N} E_{Z|n,\hat{\rho}_t}\left(Z_{ntb}^{(m)}\right)\right) \tag{4.13}$$

$$\hat{\rho_t}^{(m+1)} = \frac{\frac{\hat{\alpha_t}^{(m+1)}}{l_t}}{\sum_{t'\epsilon T}\frac{\hat{\alpha_t'}^{(m+1)}}{l_{t'}}} \tag{4.14}$$

3. The E and M steps are repeated with the updated values of $\hat{\rho}_t$ until they converge.

Consistent with the original kallisto implementation, the EM algorithm terminates if $\alpha_t N$ has changed by less than 1% compared to the previous iteration for every transcript $t$ with $\alpha_t N > 0.01$, where $N$ is the total number of fragments.

The observed data likelihood remains a concave function under this adjustment, provided we precompute the extent of migration errors as shown by section 2.3.1. We can thus compute maximum likelihood values of transcript abundances using an EM algorithm.

#### 4.1.3.1 Proof of concavity of Ladder-seq likelihood

The log-likelihood function of Ladder-seq is:

$$\ln(L(\alpha)) = \sum_{e \in E} \sum_{b=1}^{k} c_{eb} \ln \left( \sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t} \right). \tag{4.15}$$

For arbitrary but fixed $e \in E$ and $b \in [k]$ we define

$$f(\alpha) = c_{eb} \ln \left( \sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t} \right). \tag{4.16}$$

Analog to [71] we prove in the following that $f(\alpha)$ is concave, from which it follows that $\ln(L(\alpha))$ is concave too. Let $H(\alpha)$ represent the Hessian matrix of function $f(\alpha)$:

$$\begin{aligned} H_{jk}(\alpha) &= \frac{\partial^2 c_{eb} \ln \left( \sum_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t} \right)}{\partial \alpha_j \partial \alpha_k} && (4.17) \\ &= -c_{eb} \frac{\beta_{jb} \beta_{kb}}{l_j l_k} \frac{1}{\left( \sum\limits_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t} \right)^2} && (4.18) \end{aligned}$$

Then we can rewrite $H(\alpha) = -z(\alpha) x^T x$, where

$$\begin{aligned} z(\alpha) &= \frac{c_{eb}}{\left( \sum\limits_{t \in e} \frac{\alpha_t \beta_{tb}}{l_t} \right)^2} \quad \text{and} && (4.19) \\ x &= \left( \frac{\beta_{1b}}{l_1}, \frac{\beta_{2b}}{l_2}, \frac{\beta_{3b}}{l_3}, \ldots, \frac{\beta_{|e|b}}{l_{|e|}} \right). && (4.20) \end{aligned}$$

Since $z(\alpha) > 0$, we have for all $y = \left( y_1, y_2, \ldots, y_{|e|} \right)$:

$$\begin{aligned} y H(\alpha) y^T &= y \left( -z(\alpha) x^T x \right) y^T && (4.21) \\ &= -z(\alpha) (y x^T)(x y^T) && (4.22) \\ &= -z(\alpha) (y x^T)^2 && (4.23) \\ &\leq 0 && (4.24) \end{aligned}$$

Thus $H(\alpha)$ is negative semi-definite and $f(\alpha)$ is concave.

We extend the EM implementation in sotware tool kallisto [27] to quantify transcripts based on compatibilities of reads with transcripts rather than precise alignments. These compatibilities can be computed much faster through a pseudoalignment and provide a sufficient statistic for the abundances [27]. According to recent benchmarks [43, 73], alignment-free methods such as kallisto quantify transcripts much faster yet compare favourably in terms of accuracy to alignment-based methods like RSEM [62].

### 4.1.4 Additional length information improves the assignment of ambiguous reads to transcripts

In order to get a detailed picture of the difference between the fractional distribution of reads by conventional RNA-seq and Ladder-seq , we selected a set of transcripts from the simulated data consisting of 30 million reads and show the difference in distribution of reads by conventional kallisto and kallisto-ls. Reads mapping to overlapping regions of transcripts are conventionally assigned equally to the corresponding transcripts during the first iteration of the EM algorithm. Abundances in the following iterations are estimated depending on the assignment of reads in the first iteration. Ladder-seq can deconvolute the signal, starting from the first iteration, if the transcripts in question have lengths sufficiently different in order to migrate in different bands. Reads in Ladder-seq are not only assigned based on the number of transcripts that they map to, but also on the band that the read originated from, which in turn depends on transcript length.

Fig. 4.1 shows the distribution of reads in conventional-RNA and the Ladder-seq. While reads are assigned to transcripts most always with equal probability by conventional RNA-seq, we observe that Ladder-seq assigns reads from smaller bands to transcripts of smaller lengths with a higher probability and vice versa. This leads to more accurate quantification.

In this particular example conventional kallisto assigns 0.000013 counts to the transcript T1 (ENST00000357668), 334 counts to transcript T2 (ENST00000524124), and 4.11 counts to the transcript T3 (ENST00000519483) while in the ground truth they have 83.04, 250 and 15.30 counts respectively. The overestimation of T2 by conventional kallisto is at the cost of underestimating T1 and T3. This is due fact that transcripts T1 and T3 have huge overlapping regions T2 with very few reads mapping to their short unique regions. This results in most of the overlapping reads being assigned to T3 and hence the erroneous estimation. The assignment of the same set of reads by kallisto-ls leads to a better quantification since the reads are assigned probabilistically based on the band that they originate from. As a result, reads from lower bands are assigned almost exclusively to T3 because of its short length and reads from the higher bands are assigned to T1 and T2 also based on their length. kallisto-ls assigns 67, 256.74 and 17.16 counts to the transcripts T1, T2 and T3 respectively which is much closer to the ground truth counts than conventional kallisto.

Based on the estimation of the degree to which migration errors cause transcripts to end up in the "wrong" band (section 2.3.1) we adjust the probability of obtaining a read in a given band from a specific transcript by the probability of seeing a transcript of the same length in the corresponding band.

## 4.2 Evaluation

To assess the advantages of our Ladder-seq tailored EM implementation, kallisto-ls, over conventional kallisto we compared their performance on simulated Ladder-seq samples and matching RNA-seq sample, respectively (Fig. 3.1). As in the original benchmark in [27], we used RSEM to simulate reads from a transcriptome with abundances and error profiles estimated from sample NA12716_7 of the Genetic European Variation in Health and Disease (GEUVADIS) data set. We simulated 20 datasets of 30 million and 20 datasets containing 75 million RNA-seq reads and from each of these samples we derived a matching Ladder-seq sample using our in silico length separation. The RNA-seq and corresponding Ladder-seq samples differ only in the random assignment of reads to bands but are otherwise identical. The details of the simulation process is explained in Section 3.1.

Prior to running the EM algorithm on the Ladder-seq samples, kallisto-ls estimates migration patterns using the procedure described in the 2.3.1.

We measure quantification accuracy by mean absolute relative difference (MARD) and Pearson correlation (section 3.2), the same metrics used in a benchmark of transcript quantification methods [43]. Even though conventional kallisto provides highly accurate abundance estimates on this simulated RNA-seq sample, the additional length information contained in the corresponding Ladder-seq sample is employed by kallisto-ls to even more accurately quantify transcripts (Fig. 4.2 and Appendix Fig. C.1). In fact, kallisto-ls is able to quantify transcripts of genes expressing 10 isoforms as accurately (in terms of MARD) as conventional kallisto is able to quantify merely two expressed isoforms. A larger fraction of genes expressing a single transcript were estimated to be lowly expressed in sample NA12716_7 by RSEM (Table C.1), making their quantification less accurate by both kallisto-ls and its conventional counterpart. Even the slightly higher fraction among genes expressing 2 transcripts has a small negative impact on quantification accuracy. Nevertheless, kallisto-ls achieved better MARD and correlation for these sets of transcripts (Tables C.2 and C.3). To evaluate the impact that a more precise length separation has on the accuracy of Ladder-seq, we mimic an idealised version of the Ladder-seq protocol which perfectly separates transcripts by length without any migration errors. To this end, the same set of reads is partitioned into the same number of bands deterministically according to the length of the originating transcript. Fig. 4.2 (and Appendix Fig. C.1) shows that a more accurate length separation can in principle improve quantification accuracy even further, yielding a reduction in MARD of more than 31% for genes expressing 4 transcripts, for example. All results are listed in Appendix Tables C.2 and C.3.

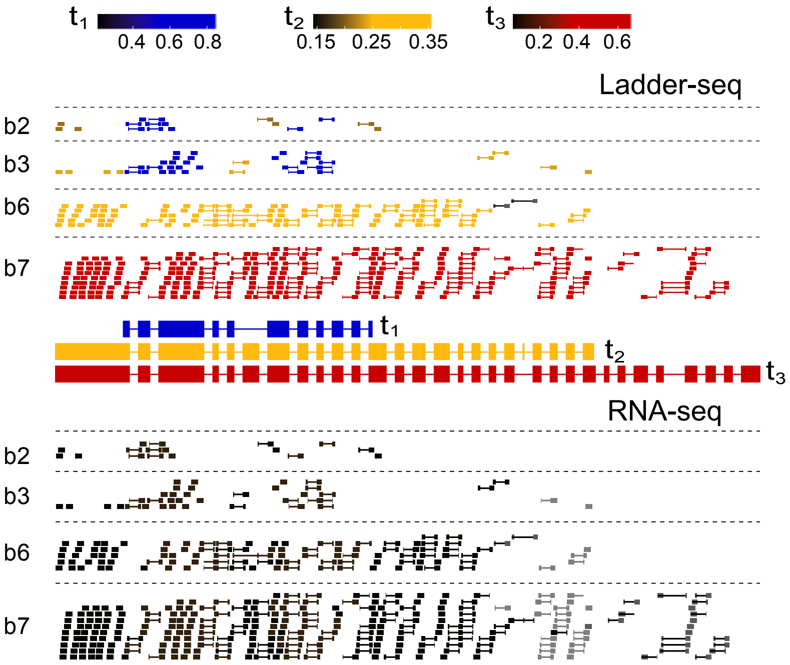

Figure 4.1: Reduced read assignment ambiguity in Ladder-seq.(a) This illustrative example shows reads that were sampled in bands 2,3,6 and 7 in our genome-wide simulation study from 3 transcripts ($t_1$ = ENST00000519483, $t_2$ = ENST00000524124, $t_3$ = ENST00000357668) that largely overlap (not all transcripts shown). The color of each read indicates the transcript to which the read is dominantly assigned after the first E-step of the EM algorithm in the original kallisto implementation based on conventional RNA-seq data (bottom) and in our extension of the algorithm to Ladder-seq (top). More precisely, we color every read according to the additional fraction that is assigned to the transcript of maximal assignment. The original algorithm fractionally assigns each read equally to every transcript it overlaps (normalized by length), leading to indistinguishable black reads. Our adaptation of the algorithm utilizes the partitioning of reads into bands to hint at the read's originating transcript, demonstrated by matching read and transcript colors. Based on the migration patterns estimated from the length of the 3 transcripts, our EM algorithm assigns larger read fractions to transcripts that are expected to occur more abundantly in the read's band.. This length-based deconvolution allows the EM algorithm to ultimately quantify transcript abundances more accurately. In this example, our Ladder-seq specific EM algorithm estimates 67, 257, and 17 counts (rounded) for transcripts $t_1$, $t_2$, and $t_3$ respectively, which closely match their true expression of 83, 250, and 15 counts, respectively. In contrast, original kallisto fails to detect expression of $t_1$ (0 counts) and overestimates expression of $t_2$ (334 counts) from highly ambiguous RNA-seq reads. It estimates 4 counts for $t_3$.

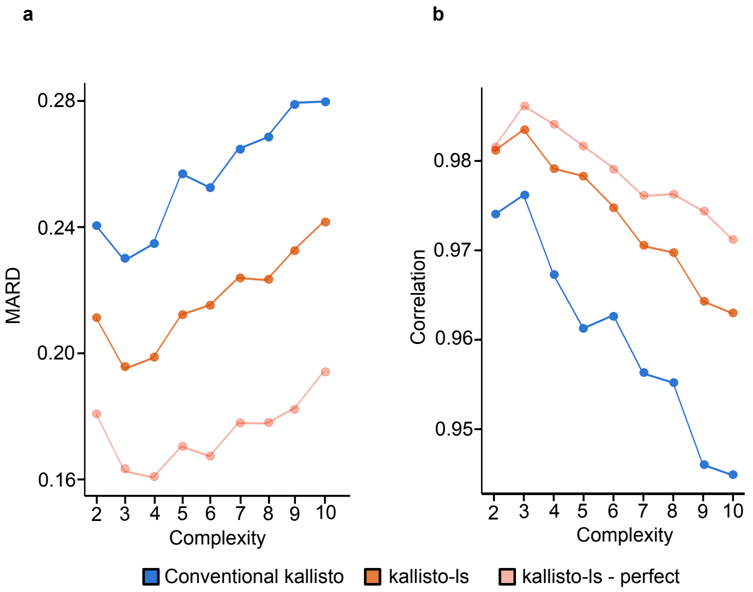

Figure 4.2: Reduced read assignment ambiguity in Ladder-seq improves transcript quantification. Quantification accuracy of kallisto-ls compared to conventional kallisto. 30 million 75 bp paired-end reads were simulated by RSEM from transcripts with abundances estimated from GEUVADIS sample NA12716_7. Results for 75 million simulated reads are shown in Fig. C.1. RSEM simulations were repeated 20 times, and mean values are reported. Pearson correlation of estimated and ground truth abundance in $\log_2$ transformed transcripts per million (TPM) and mean absolute relative difference (MARD) are shown as a function of gene complexity, i.e. the number of transcripts expressed by a gene. For the ease of visualization, we omit genes expressing a single transcript many of which are estimated to be lowly expressed in this sample by RSEM. Nevertheless, kallisto-ls achieves slightly better MARD and correlation for this set of transcripts (Appendix Table C.2).

*Chapter 5*

# Assembly

## 5.1 Reference-based transcript assembly - StringTie-ls - general framework

Fragments in a RNA-seq library are much shorter than their originating transcripts, and therefore the phasing of distant exons into full-length transcripts must rely on local connectivity information provided by reads connecting neighboring exons and on local read coverage. Current methods represent this connectivity by a graph structure such as the splicing graph [74], and infer transcripts as paths through this graph, using the coverage along transcripts as an additional selection criteria. In contrast to the problem of quantifying the abundance of a small set of transcripts annotated for a given gene, the space of possible candidate transcripts that can be obtained by combining locally connected exons in paths through the graph can grow exponentially, and smoothing the local coverage along transcripts cannot unambiguously point to a single best subset of transcripts [75]. Here, we propose a computational framework (see Fig. 5.1) that enables conventional RNA-seq assembly methods to exploit the extra layer of information provided by Ladder-seq to reduce the ambiguity in combining distant splicing events into transcript isoforms. Parts of this chapter are taken from [54].

In this scheme, a separate splicing graph is built from reads in each band, which can help to dramatically shrink the combinatorial space of possible transcripts and facilitate their computational deconvolution into estimates of relative abundance. Furthermore, reads in a given band originate from transcripts of a certain length range which we use to further inform the selection of paths in individual splicing graphs. Note that we do not rely on the location of cuts through the gel as indicated by the RNA ladder to define transcript length constraints but derive constraints from distributions of transcript lengths across bands (Section 2.3.1 and Table B.3). These length constraints can aid in breaking (too long) erroneous fusions and in eliminating (too short) transcript fragments. We then integrate individual assemblies into a non-redundant set of transcripts. Finally, we use kallisto-ls to assign reads to transcripts assembled by the above procedure according to our statistical model of Ladder-seq. In contrast to the hard length constraints applied during the assembly, kallisto-ls assigns reads probabilistically taking into account the computationally estimated migration error and can thus refine our final reconstruction.

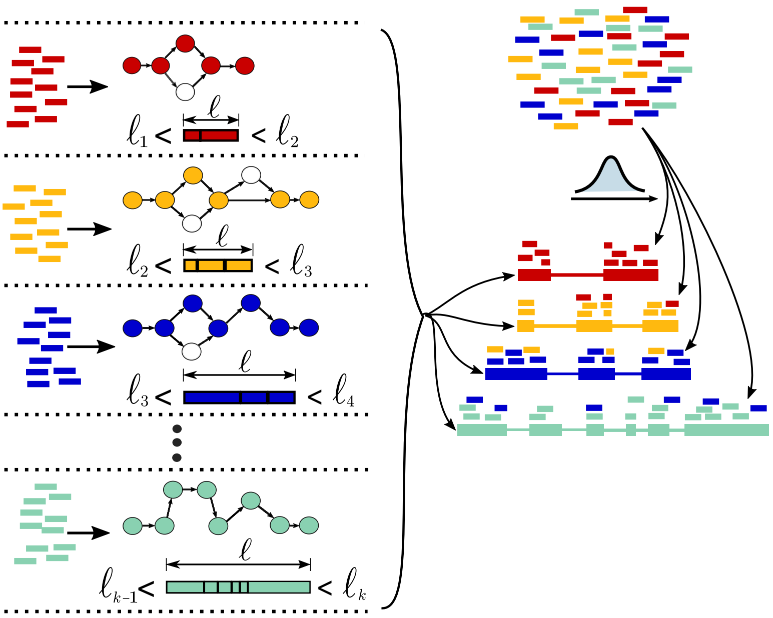

Figure 5.1: Ladder-seq based transcript assembly. Overview of the proposed computational framework. For each band, a graph is constructed that captures connectivity information contained in reads or their alignments. Reference-based assembly methods such as StringTie2 use variants of splicing graphs to capture connectivity of exonic segments in expressed transcripts evidenced by spliced alignments of reads. Transcript sequences are then assembled by traversing paths through these graphs according to some optimization criteria, a maximum flow in the case of StringTie2. In contrast to conventional RNA-seq, where truly expressed transcripts need to be identified among a large number of possible paths through a single graph per locus, Ladder-seq limits the search for expressed transcripts to paths in smaller graphs that are constructed for each band separately. In addition, reads in different bands are obtained from transcripts of a certain length range, imposing length constraints that can further direct the search for the correct paths. After having inferred the best possible set of transcripts satisfying given length constraints in each band independently, we integrate them to a refined set of transcripts by assigning reads to them according to our statistical model of Ladder-seq, which relies on previously estimated migration patterns through the gel. This last step takes into account the dependence between bands introduced by the imperfect experimental separation of transcripts.

### 5.1.1 Detailed method

We chose StringTie2 [29] as the presumably most accurate RNA-seq assembly method [76, 29] to illustrate the benefit of our Ladder-seq tailored assembly approach over its conventional RNA-seq counterpart. Reads from all bands are aligned to the reference genome sequence using a short read aligner such as STAR [25] in the *1passmode*. Prior to mapping, STAR creates an index using the genome and it can be provided with a known set of splice junctions for a more splice informed mapping. We decided to index the genome without using splice junctions in order to find more novel splicing events. We assemble transcripts from every band using StringTie2 with default options. We additionally pool reads from neighboring bands and assemble transcripts using them in order to recover potentially low-expressed transcripts that migrated close to the boundary between two bands.

StringTie-ls estimates migration patterns in a Ladder-seq sample using the histogram based approach described in section 2.3.1 and uses these estimates to identify too short transcript fragments and too long transcript fusions. More precisely, for a transcript $t$ of length $\ell$ assembled in the $j$th band we look up the probability mass function $f(x)$ corresponding to the length range that contains $\ell$ to determine the most likely band $b_i$ to which a transcript of length $\ell$ would have migrated to. If $j \neq i$ and $j \neq i+1$, we remove $t$. Note that band $b_{i+1}$ corresponds to the next longer range of transcripts but can contain also slightly shorter transcripts from band $b_i$ due to secondary structure effects. Similarly, if $t$ was assembled in the combination of bands $j$ and $j+1$, we discard $t$ if $j < i$ or $j > i+2$. To account for potential overlap with longer UTRs, we do not remove too long transcripts assembled in a band $i+2 \ldots 7$ if they are sufficiently high expressed ($> 1$ TPM), contain a unique intron, and if their first or last exon is longer than 500 bp.

The individual assemblies are subsequently merged using the GffCompare tool which computes the union of all intron chains. In other words, transcripts that imply the exact same sequence of introns as a transcript assembled in a different band are discarded. We further eliminate single-exon transcripts that are identified as redundant by the *merge* mode of StringTie2 as well as transcript fragments that are fully contained in other transcripts with compatible intron chains. These transcripts most likely constitute transcript fragments that were only partially assembled from reads obtained from transcripts that migrated to a different band. We retain, however, transcripts with identical (partial) intron chain if they start or end within an intron of the containing transcript, unless a very small overhang of at most 2 bases indicates noisy read alignments. Finally, we quantify assembled transcripts using our statistical model of Ladder-seq implemented in kallisto-ls, and report all transcripts estimated to be expressed with at least 0.1 TPM.

## 5.2 De novo transcript assembly - Trinity-ls

To study the transcriptome of species for which no or just a highly fragmented reference genome is available or in samples with a substantially altered genomic sequence, transcripts need to be assembled *de novo*. However, omitting the read mapping step that arranges reads in order leaves the sequence overlap of reads as the only source of information to be utilized by methods in this most challenging setting of transcript-level inference.

Most methods, including one of the most widely used methods Trinity [37], stitch together $k$-mers, subsequences of $k$ nucleotides, to transcript sequences by traversing paths in so called *de Bruijn* graphs. Alternative paths in complex graphs can in part be resolved to individual isoforms or paralogous transcripts using the pairing information of reads. No part of the data connects subpaths at longer distances, which can cause erroneous fusions of isoforms or paralogs, especially in complex genes with a large number of alternative splicing events [77].

Here, we follow a similar strategy as in the reference-based assembly (Fig. 5.1) to access the additional layer of information provided by Ladder-seq to guide the *de novo* assembly of full-length transcripts by Trinity. We run Trinity on the reads from each band separately using default parameters. In contrast to the reference-based assembly, we do not pool reads from neighboring bands since the absence of a reference genome makes it harder to subsequently detect and remove false positive transcripts. After estimating migration patterns from Ladder-seq data using the histogram-based method (Section 2.3.1), Trinity-ls applies length constraints to assembled transcripts following the same strategy as in the reference-based approach. It then concatenates the individual assemblies, since the absence of a reference genome does not allow to detect potential redundancy with respect to the exon-intron structure of transcripts. Again, Trinity-ls quantifies assembled transcripts using our statistical model of Ladder-seq implemented in kallisto-ls and applies an expression threshold of 0.1 TPM.

We benchmark Trinity-ls, our Trinity based *de novo* assembly approach for Ladder-seq, and conventional Trinity on the same simulated 30 million and 75 million Ladder-seq reads and matching RNA-seq samples used in the experiments on quantification and reference-based assembly. Trinity-ls estimates migration patterns from simulated Ladder-seq samples using the histogram-based approach described above. Applying similar criteria as for example in [35], we consider a transcript correctly assembled if BLAT [69] aligns its sequence to a true transcript with 95% sequence identity and at most 1% insertion and deletion rate. In the most strict setting, we require the reconstructed transcript to cover at least 95% of the full transcript length and additionally evaluate the performance when applying a 80%, 85%, and 90% length threshold.

## 5.3 Evaluation

### 5.3.1 StringTie-ls

We used the same simulated Ladder-seq samples and its matched RNA-seq samples as in the kallisto-ls benchmark (section 4.2) to evaluate the performance of StringTie-ls, our StringTie2 based assembly approach for Ladder-seq, in comparison to conventional StringTie2 ran on the corresponding RNA-seq samples (Fig. 3.1). The larger data set contains the same number of reads (75 million pairs) as used in the original StringTie2 benchmark. Starting from the migration error estimated from our real data, we created additional Ladder-seq samples that mimic an improved length separation step by gradually reducing the degree of migration errors as shown in section 3.1.2. RNA-seq and Ladder-seq reads were aligned identically to the reference genome (GRCh38) using STAR [25], and again StringTie-ls estimated migration patterns from the simulated Ladder-seq samples using the histogram-based method. We calculate sensitivity as the fraction of truly expressed transcripts that precisely match the sequence of introns of an assembled transcript, and precision as the fraction of assembled transcripts that match an expressed transcript (section 3.2).

Fig. 5.2 shows that StringTie-ls is able to correctly reconstruct a much larger fraction of expressed transcripts than conventional StringTie2, and as expected this improvement in sensitivity increases with gene complexity. For genes expressing 4 transcripts, StringTie-ls detects 16% more transcripts than conventional StringTie2, and this improvement increases to 31.1% and 35.2% for complex genes expressing 7 and 10 transcripts, respectively. The gap between these two technologies widens with a more accurate length separation of transcripts, reaching an improvement of 25.2% for genes expressing 4 transcripts, and 49.2% and 58.7% for genes of complexity 7 and 10, respectively, in the most optimistic scenario. In this setting, StringTie-ls is able to reconstruct transcripts of complex genes expressing 7 transcripts with a higher sensitivity than conventional StringTie2 is able to detect just three isoforms expressed by a gene. At the same time, StringTie-ls assembles transcripts with higher precision across all complexity classes. Especially on genes that express only a small number of transcripts, StringTie-ls benefits enormously from the additional length information that allows it to detect too short transcript fragments. For genes expressing a single transcript, for example, StringTie-ls recognizes 699 out of 824 false positive assemblies by conventional StringTie2 as being too short and eliminates them, improving precision by 30.8% compared to its conventional counterpart. For complex genes expressing multiple transcripts a better length separation is required to yield a marked improvement in precision. We observed a similar improvement on the larger data set (Appendix Fig. D.1), with only marginally higher sensitivity and precision compared to the lower sequencing depth. All results are listed in the Appendix Tables D.1-D.4.

### 5.3.2 Trinity-ls

Fig. 5.3 shows an enormous performance gain of Trinity-ls over conventional Trinity, both in terms of sensitivity and precision when assembling transcripts *de novo* from 75 million reads with a 90% transcript length cut-off. For genes expressing 5 transcript isoforms, for example, Trinity-ls achieves a similar sensitivity than Trinity does for genes that express just a single transcript. Again, the low expression of some genes expressing a single transcript (C.1) makes them more difficult to assemble than transcripts of genes with higher complexity. In the most optimistic scenario with perfect length separation, Trinity-ls is able to distinguish 10 isoforms expressed by the same gene more accurately than conventional Trinity recovers the single transcript expressed by a gene. In total, Trinity-ls correctly recovers an additional 4072 (78%) transcripts compared to Trinity, while at the same time increasing precision equally by 78%. A more accurate separation of transcripts by length further boosts the performance of Trinity-ls, approaching an additional 163% of correctly discovered transcripts and a 3.9-fold increase in precision in the most optimistic scenario

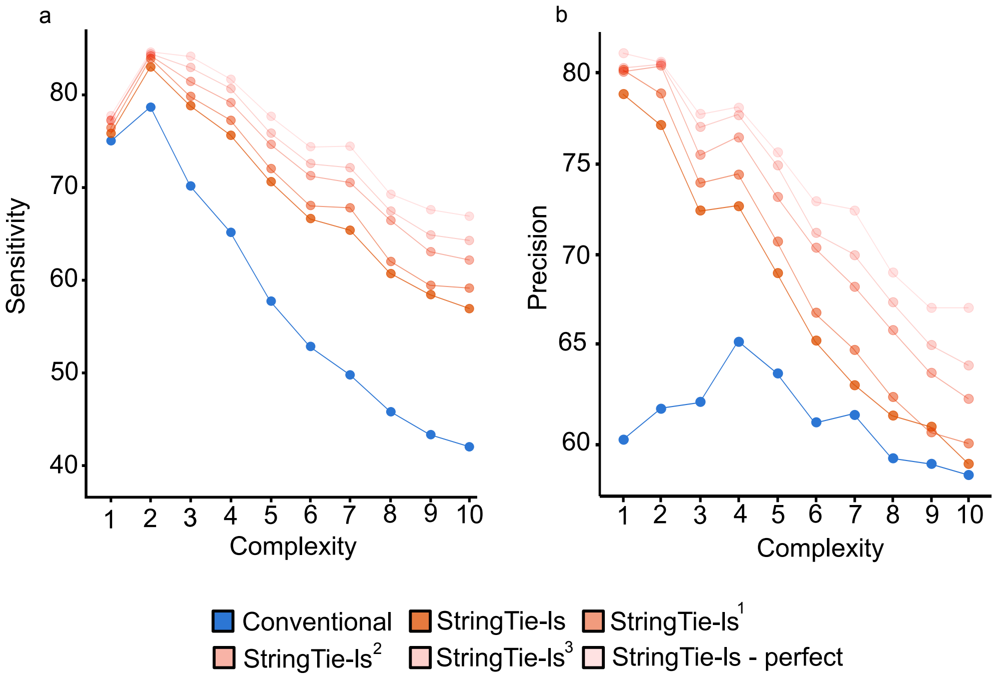

Figure 5.2: Ladder-seq based transcript assembly. Accuracy of transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq reads. Sensitivity (a) and precision (b) of StringTie-ls and its conventional counterpart StringTie2 are shown as a function of gene complexity measured as the number of expressed transcripts. StringTie-ls$^i$ denotes the result of StringTie-ls on the simulated Ladder-seq data set to which $i$-fold error reduction was appliced (see Methods) starting from the migration error estimated from the NPC sample (StringTie-ls). StringTie-ls - perfect represents the results of StringTie-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration error.

in which transcripts can be perfectly separated by their length.

We observed a similar improvement when applying different length thresholds or when analysing a smaller data set containing 30 million reads (Appendix Fig. D.2). As expected, fewer transcripts are correctly reconstructed from the smaller set of reads, yet with slightly higher precision. All results are listed in Appendix Tables D.7-D.6.

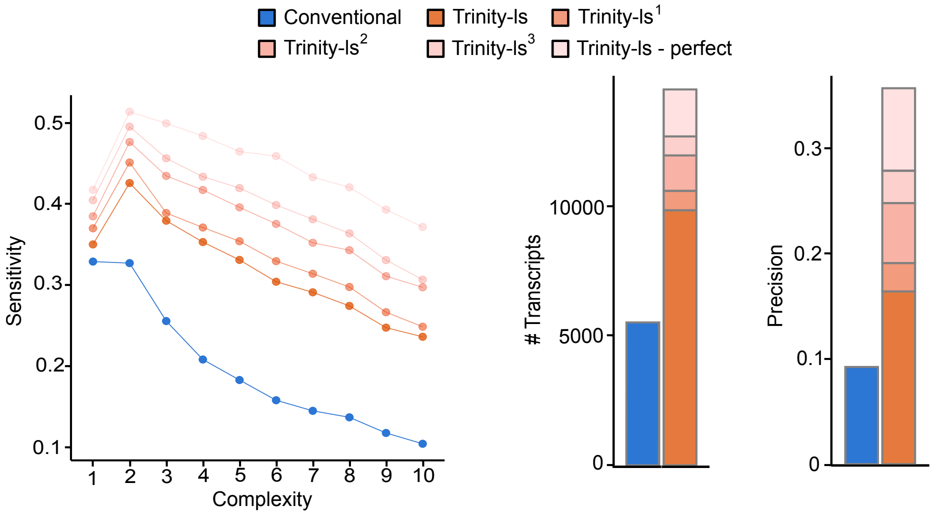

Figure 5.3: Accuracy of *de novo* transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq reads. Trinity-ls$^i$ denotes the results of Trinity-ls on the simulated Ladder-seq data set to which $i$-fold error reduction was applied (see Methods) starting from the migration error estimated from the NPC sample (Trinity-ls). Trinity-ls - perfect represents the results of Trinity-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration erorr. (left) Sensitivity of Trinity-ls and its conventional counterpart Trinity at 90% transcript length cut-off is shown as a function of gene complexity measured as the number of expressed transcripts. (middle) Total number of correctly assembled transcripts at 90% transcript length cut-off. (right) Precision in *de novo* assembly at 90% transcript length cut-off. Overall precision is shown since assembled transcript fragments cannot be assigned unambiguously to individual genes.

*Chapter 6*

# Real data analysis using Ladder-seq

## 6.1 Ladder-seq improves differential analysis of reconstructed transcriptomes

In addition to benchmarking the accuracy of transcript inference from simulated Ladder-seq data, we evaluated its impact on the differential analysis of reconstructed transcriptomes between two biological conditions. We used Ladder-seq to profile the transcriptome of wild type (WT) and *Mettl14* knock-out (KO) mouse neural progenitor cells (NPCs) (4 independent replicates per genotype). To assess transcript usage under these conditions we first assembled transcripts using StringTie-ls on each sample to identify novel transcripts that are expressed consistently across replicates of the same genotype. We quantified annotated (Ensembl release 95) and newly reconstructed transcripts using kallisto-ls and compared their expression between conditions to detect their differential usage. For comparison with conventional RNA-seq, we ran the same computational pipeline replacing the Ladder-seq tailored methods kallisto-ls and StringTie-ls by their conventional counterparts which ignore the separation of reads into bands (Fig. 6.1).
Since no ground truth is available for the real Ladder-seq data sets, we used characteristics of genes and experimental data to provide indirect evidence on the correctness of isoform switches only detected by Ladder-seq and evidence that switches identified only by the conventional pipeline are likely a consequence of inaccurate transcript assembly and quantification. Results presented in this chapter are taken from [54].

### 6.1.1 Isoform switches

Isoform switch is a phenomenon where the expression levels of two isoforms will switch depending on the conditional treatment of the samples. The R Bioconductor package IsoformSwitchAnalyzeR [78] was used for differential isoform usage (DIU) analysis. Identification of differentially used isoforms across all genes with IsoformSwitchAnalyzeR is done through DEXseq [79], a statistical method originally developed for differential exon usage which has since been shown to adequately control for false discovery rate in the setting of DIU. Analysis of consequences of isoform switches was performed through IsoformSwitchAnalyzeR with the function

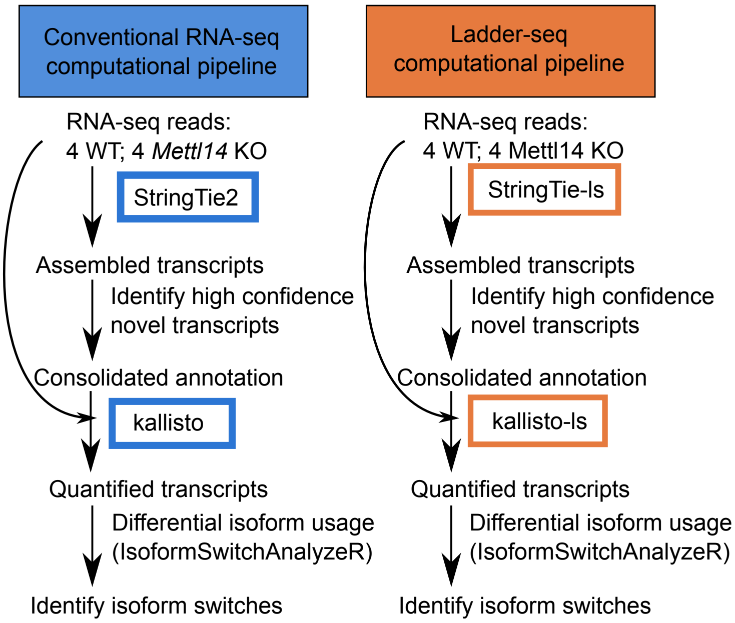
Conventional RNA-seq computational pipeline
RNA-seq reads: 4 WT; 4 *Mettl14* KO
StringTie2
Assembled transcripts
Identify high confidence novel transcripts
Consolidated annotation
kallisto
Quantified transcripts
Differential isoform usage (IsoformSwitchAnalyzeR)
Identify isoform switches
Ladder-seq computational pipeline
RNA-seq reads: 4 WT; 4 Mettl14 KO
StringTie-ls
Assembled transcripts
Identify high confidence novel transcripts
Consolidated annotation
kallisto-ls
Quantified transcripts
Differential isoform usage (IsoformSwitchAnalyzeR)
Identify isoform switches

Figure 6.1: Computational pipeline for differential isoform usage analysis with conventional RNA-seq and Ladder-seq . Reads were aligned using STAR aligner prior to transcript assembly for both pipelines.

analyzeSwitchConsequences. This function allows to add input data from CPAT [80] for analysis of coding potential and from PfamScan [81] for protein domain annotation.
Ladder-seq identified 40% more genes harboring switching isoforms in *Mettl14* KO compared to conventional RNA-seq (Fig. 6.2). While the overlap between the two methods is high (1,114 genes), there is a substantial number of genes that was reported only by conventional RNA-seq (763 genes) and even more identified only by Ladder-seq (1,520 genes).

#### 6.1.1.1 Gene complexity

In order to assess how difficult it is to accurately assemble and quantify those transcripts identified as switching, only by Ladder-seq or the conventional pipeline or by both, we examined gene complexity, which we define as the number of expressed transcripts per gene. Genes identified as switching exclusively by Ladder-seq appear to be particularly hard to reconstruct by the conventional pipeline without the additional length separation (Fig. 6.3). In contrast, Ladder-seq breaks down gene complexity by separating transcripts into bands according to their length, effectively reducing the number of transcripts that need to be reconstructed in an individual band. This *effective complexity* is considerably lower in all three categories of genes identified as switching (Fig. 6.3), including genes identified as switching only by the

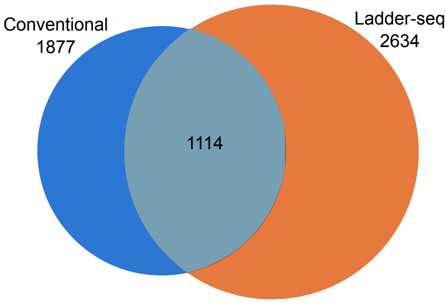

Figure 6.2: Venn diagram showing overlap between switching genes identified by Ladder-seq and conventional RNA-seq.

conventional pipeline which thus did not pose a particular challenge to the Ladder-seq protocol.

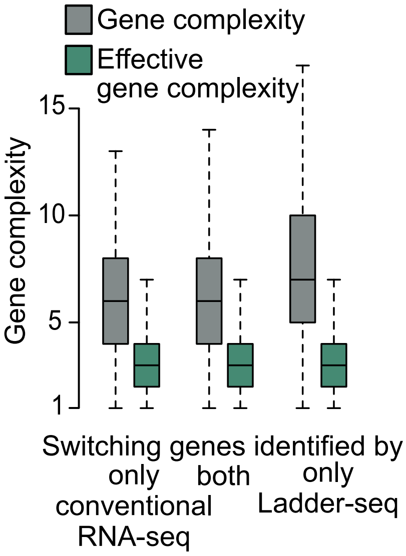

Figure 6.3: Gene complexity and effective gene complexity for switching genes identified by Ladder-seq and conventional RNA-seq. The effective complexity is defined as the number of transcripts in a single band after separating the mRNA by length.

#### 6.1.1.2 Identified isoform switches

By separating reads coming from transcripts of different lengths, Ladder-seq uncovers otherwise buried transcripts that are not identified by conventional RNA-seq. This is exemplified by the isoform switch in gene *Pi4k2a* which is only identified by our method (Fig. 6.4a). *Pi4k2a* expresses mostly the annotated ENSMUST00000066778 transcript in WT, while KO also expresses a shorter un-annotated transcript (TCONS_00005143) in which a normally m$^6$A tagged exonic region is spliced out. The separation of reads from the shorter transcript TCONS_00005143 and reads from the longer transcript ENSMUST00000066778 into bands 4 and 5, respectively, allowed StringTie-ls to detect this novel transcript whose usage switches between conditions (Fig. 6.4b). Interestingly, the shorter transcript which is absent from the mm10 Ensembl release 95, does exist in the later release 98 version (ENSMUST00000235932), confirming that what Ladder-seq assembled is indeed accurate. In addition we confirmed this isoform switch with reverse transcription quantitative PCR (RT-qPCR) on WT and *Mettl14* KO mouse NPCs (Fig. 6.7). Additional illustrative examples of isoform switches uncovered only by Ladder-seq are shown in Figures 6.5 and 6.6.

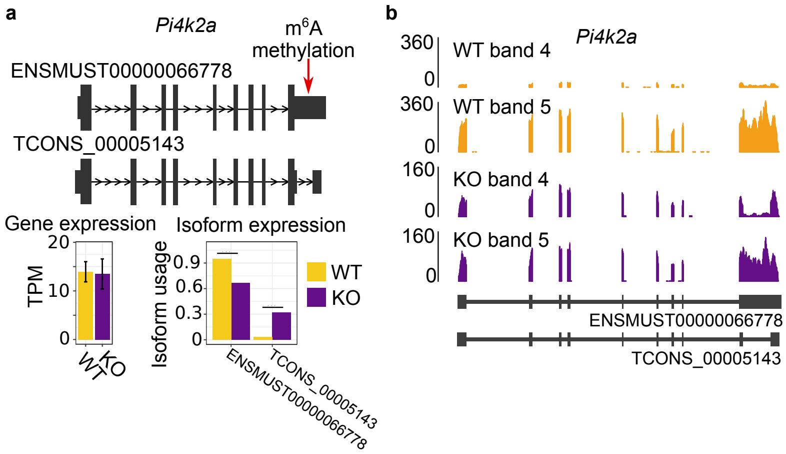

Figure 6.4: (a) Isoform switch identified by Ladder-seq in *Pi4k2a*. Red arrow shows location of m$^6$A methylation. (b) Coverage plot shows how reads from the shorter un-annotated TCONS_00005143 are separated from reads belonging to the longer ENSMUST00000066778.

### 6.1.2 Accuracy of read assignment by Ladder-seq based methods as compared to conventional methods

Ladder-seq makes use of estimated probability distributions describing how reads from a transcript of a given length are expected to be distributed among the different bands of the gel, i.e. how a mRNA molecule migrates through a denaturing agarose gel. We can use these distributions to assess whether the originating bands of reads assigned during quantification to a transcript of a given length follow the estimated distribution. We used Jensen Shannon divergence (JSD), a measure of similarity between two probability distributions, to compare estimated to assigned read band distributions for transcripts as quantified by conventional kallisto or by kallisto-ls. JSD values for kallisto-ls were consistently low for all identified switching genes, which is to be expected given that kallisto-ls makes explicit use of these distributions to guide the assignment of reads. On the other hand, JSD values for conventional kallisto were highest for those genes identified as switching only by conventional

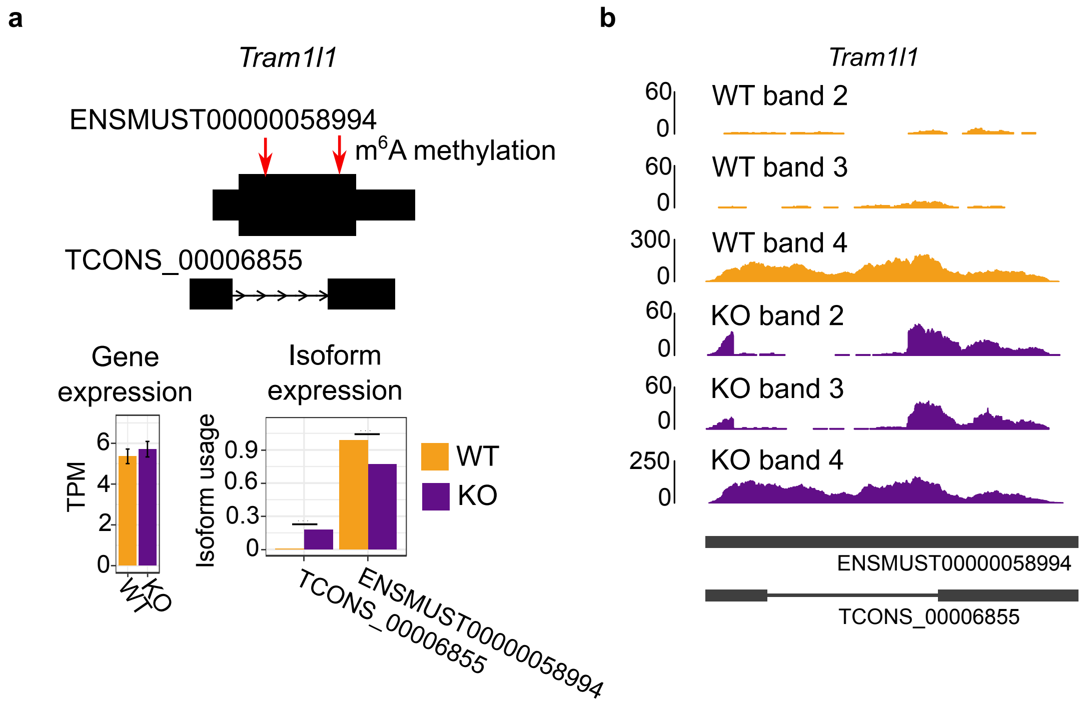

Figure 6.5: Isoform switche identified only by Ladder-seq in gene Tram1l1. (a) Red arrows show location of m$^6$A methylation. TCONS00006855 is a novel isoform of Tram1l1 that was assembled by both methods, but conventional RNA-seq failed to identify the isoform switch. Without length information, conventional RNA- seq reads in KO bands 2 and 3 were predominantly assigned to the annotated transcript in band 4. Error bars indicate 95% confidence intervals. (b) Coverage plots for switching gene Tram1l1 showing separation of reads from transcripts of different lengths.

RNA-seq (Fig. 6.8a). In fact, more generally we observed that the more conventional kallisto differs from kallisto-ls in its transcript quantification, the more its assigned read band distribution deviates from the estimated distribution, resulting in larger JSD values (Fig. 6.8c). At the same time, the conventional pipeline leads to larger JSD values if fewer reads are available that can be uniquely mapped to individual transcripts and thus direct the correct assignment of ambiguous reads (Fig. 6.8b). This makes larger JSD values likely an indication of erroneous assignments of reads by conventional kallisto.

## 6.2 Mettl14 KO leads to isoform switches in m$^6$A methylated genes

Having identified a large set of genes with isoform switches using Ladder-seq, we next set out to delineate the characteristics of these events and their relationship to m$^6$A methylation. We identified m$^6$A tagged genes in a public m$^6$A RIP-seq dataset from mouse NPCs [82] and built a set of high confidence m$^6$A peaks. BED files

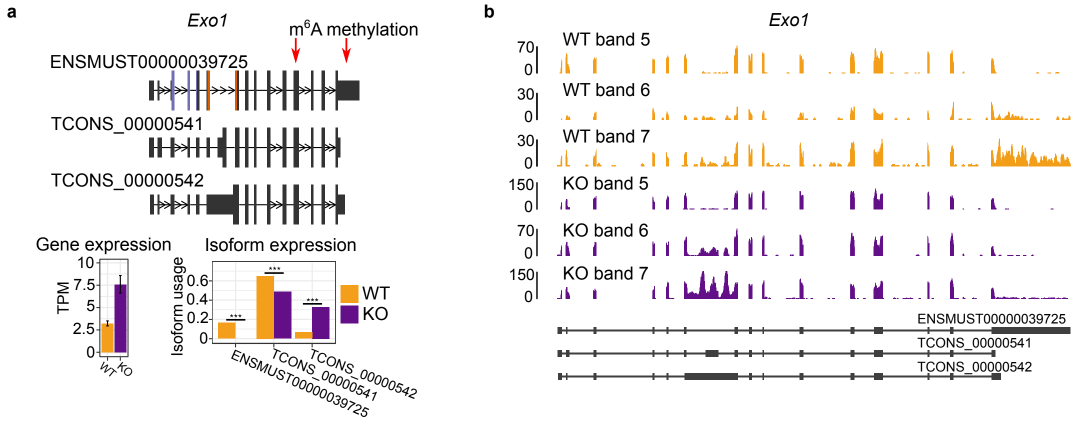

Figure 6.6: Isoform switches identified only by Ladder-seq in gene Exo1. Red arrows show location of m$^6$A methylation. TCONS 00000541 and TCONS 00000542 are novel isoforms of Exo1 detected only by Ladder-seq. Error bars indicate 95% confidence intervals. (b) Coverage plots for switching gene Exo1 separation of reads from transcripts of different lengths.

containing peaks called by MACS2 [83] from two replicates with two input samples each, were downloaded from NCBI's Gene Expression Omnibus (GSE104686). Using Bedtools intersect [84] we identified peaks that were reproducible in both replicates with both input controls. We then annotated these high confidence peaks using the annotatePeaks.pl program from the Homer suite [85] to identify the genes harboring m$^6$A methylation. We also performed Gene Ontology enrichment analyses using the R Bioconductor package TopGO [86]. Only genes passing the pre-filtering step for differential isoform usage (TPM >1) were considered for the gene universe.
We found that switching genes are significantly enriched for m$^6$A methylated genes (fisher's exact test p-value = 2.36 e-19), with 1,141 out of 2,634 switching genes containing m$^6$A (Fig. 6.10a). These genes are enriched for Gene Ontology terms related to transcriptional regulation, neurogenesis and synaptic signaling (Fig. 6.9(a)).

### 6.2.1 Spatial proximity between m$^6$A and alternative splicing

To investigate the involvement of m$^6$A methylation in isoform switching, we explored a potential spatial proximity between m$^6$A and alternative splicing. We assessed whether exonic segments [32] bounding differentially spliced regions are enriched for m$^6$A methylation. Pairs of switching isoforms from m$^6$A methylated genes were partitioned into minimal exonic segments that are bounded by splice sites, transcription start, or transcription end sites of the two involved transcripts. These segments represent the largest exonic fragments that are entirely contained in one or both of the two transcripts. A segment bounds a differentially splice region if it is part of only one of the two transcripts, if it is not the first or last segment of that transcript, and if it is adjacent to a segment that is contained in both transcripts. We take into account the length of segments in the Fisher exact test by

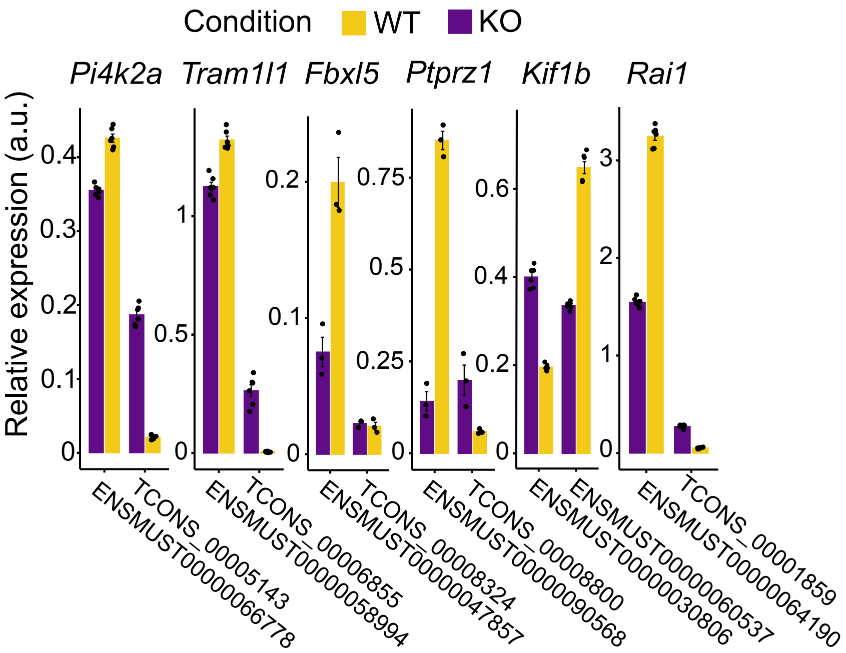

Figure 6.7: Relative quantification of isoform expression with RT-qPCR. Three biological replicates were tested per genotype. Each sample was tested in triplicate and normalized to B-Actin. Expression levels of each differentially expressed isoform were normalized to the expression of a common isoform identified in both WT and Mettl14 KO, which consistently showed no significant difference between WT and KO NPCs.

distinguishing individual bases that can lie within or outside of bounding segments and that can be methylated or not.

We analysed alternative splicing events using the IsoformSwitchAnalyzeR R Bioconductor package [87] with the functions extractSplicingSummary, which summarizes the types of alternative splicing occurring in each isoform switch, and extractSplicingEnrichment, which identifies the uneven usage of a particular alternative splicing type in one of the conditions assayed. We found a significant enrichment of m$^6$A within these segments (fisher's exact test: pvalue = 8.6 e-39), with 32.2% overlapping at least one m$^6$A peak, compared to 20% of all remaining exonic segments (Fig. 6.10b). This enrichment persists when normlizing for segment length (pvalue = 1.09 e-5) to account for a possible bias towards longer exons [88, 89]. In this test of association m$^6$A tags within spliced out introns are not accounted for, given the nature of m$^6$A RIP-seq data. Illustrative examples of m$^6$A methylation within a differentially spliced exonic segment are shown for neurogenesis related genes *Fbxl5* [90] and *Ptprz1* [91] (Fig. 6.9b and Fig. 6.10c).

### 6.2.2 Consequences of isoform switches on functional protein domains

We then studied the consequences of isoform switches on functional protein domains. We found 295 genes with loss of functional domains in the upregulated isoform in the KO. Gene Ontology enrichment analysis of these genes shows enrichment

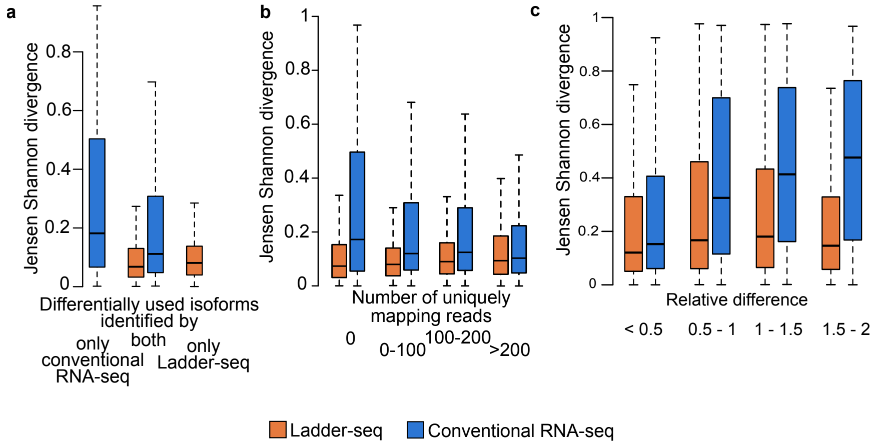

Figure 6.8: (a) Jensen Shannon divergence between estimated and assigned read band distributions for differentially used isoforms identified only by Ladder-seq and conventional RNA-seq. (b) Jensen Shannon divergence between estimated and assigned read band distributions for all identified transcripts by Ladder-seq and conventional RNA-seq grouped by number of available uniquely mapping reads. (c) Jensen Shannon divergence for Ladder-seq and conventional RNA-seq for all identified transcripts grouped by relative difference in abundance estimation by both methods. Relative difference is defined as the absolute difference in estimated transcript abundance (in TPM) divided by the average of the two. Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ±1.5× interquartile range.

for neuronal function related terms such as glutamatergic synaptic transmission, synapse organization and GABA secretion (Fig. 6.9c). For example, the isoform switch in *Kif1b* leads to upregulation of the shorter Kif1b-alpha isoform compared to WT where the most prominent expressed isoform is the longer Kif1b-beta isoform with no significant change in the overall gene expression (Fig. 6.10d). Kif1b-alpha lacks multiple domains contained in the longer beta isoform and is expressed in non-neuronal tissues, while Kif1b-beta is the neuronal isoform and is responsible for the transport of synaptic vesicle precursors [92].

To delineate the role of m$^6$A in different types of alternative splicing we categorized the splicing events occurring within each switching pair. Most splicing events were balanced between WT and *Mettl14* KO, meaning that the number of gains and losses of a certain type of splicing event (e.g. exon skipping) in the upregulated isoform were roughly the same. Intron retention events, however, were imbalanced with upregulated isoforms in KO having significantly more intron retention losses than gains (Fig. 6.9d and Fig. 6.10e). Again, these genes were enriched for m$^6$A methylated genes (pvalue = 1.6 e-06). Gene Ontology enrichment analysis revealed enrichment for terms unrelated to neuronal functions but rather associated with pluripotency, such as DNA repair, DNA recombination and gamete generation (Fig. 6.9e). We explored the consequences of intron retention losses in our dataset and found an enrichment for non-sense mediated decay (NMD) insensitive isoforms as well as for shorter 3'UTR (Fig. 6.10f), both hallmarks of decreased regulation of gene expression [93, 94]. Finally, we validated a selection of identified isoform switches by performing qPCR in WT and *Mettl14* KO mouse NPCs (Fig. 6.7) with qPCR primers specific for each pair of switching isoforms and a primer common to both isoforms for normalization (Appendix Table B.8). Together, these results indicate that *Mettl14* KO in mouse NPCs leads to widespread changes in isoform usage in genes that are normally tagged with m$^6$A methylation, and that m$^6$A tends to be close to differentially spliced regions of switching genes. Isoform switches lead to loss of functional protein domains in neuronal genes and loss of intron retentions in non-neuronal and pluripotency related genes.

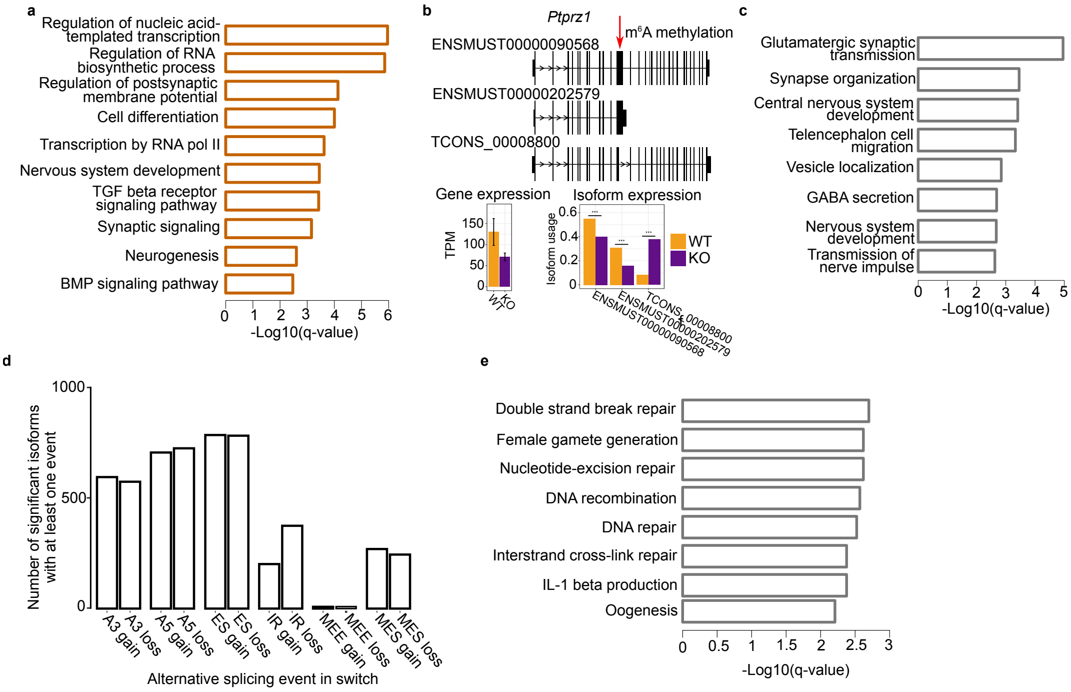

Figure 6.9: *Mettl14* KO leads to isoform switches in m$^6$A methylated genes. (a) Gene Ontology for m$^6$A methylated genes containing isoform switches. (b) Isoform switch in *Ptprz1*. Red arrow shows location of m$^6$A methylation. (c) Gene Ontology analysis for genes with loss of protein domains in KO NPCs. (d) Splicing analysis: Number of gains and losses of each splicing event in KO NPCs. A3: Alternative 3' acceptor site; A5: Alternative 5' acceptor site; ES: Exon skipping; IR: Intron retention; MEE: Mutually exclusive exon; MES: Multiple exon skipping. (e) Gene Ontology enrichment analysis of genes with intron retention loss in KO NPCs.

## 6.3 Long-read sequencing confirms many Ladder-seq transcripts in mouse NPCs

Third-generation sequencing technologies such as those from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) can produce reads longer than 10,000 bp which in principle allow to cover full-length transcripts. We therefore per-

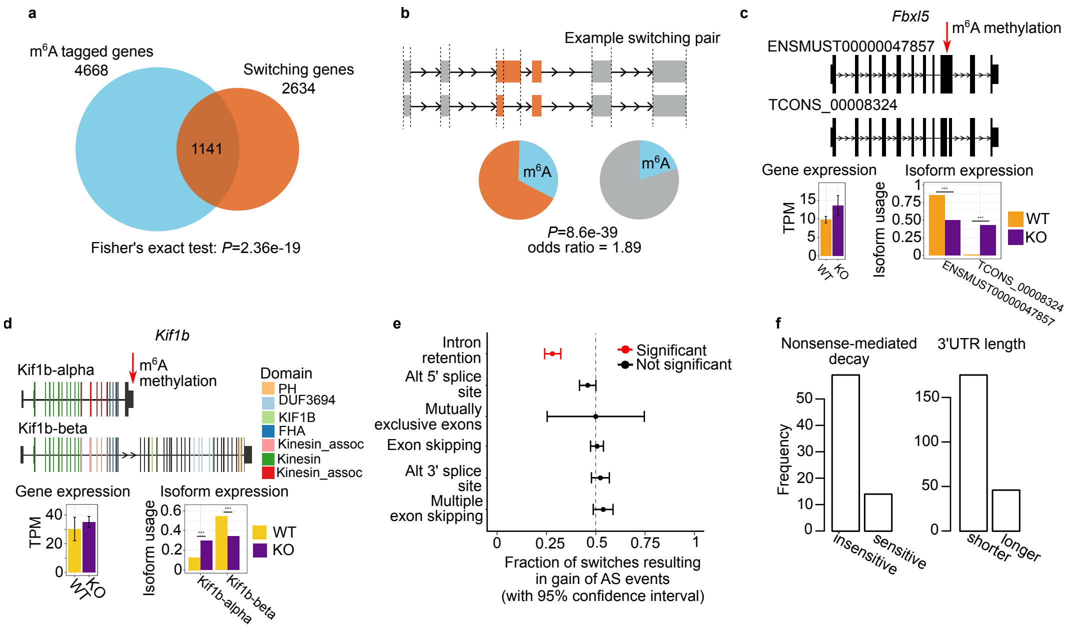

Figure 6.10: *Mettl14* KO leads to isoform switches in m$^6$A methylated genes and leads to loss of protein domains and loss of intron retentions. (a) Venn diagram showing overlap between switching genes and m$^6$A methylated genes. (b) Enrichment of m$^6$A methylation within exonic segments bounding differentially spliced regions. In this example, both the differentially spliced exonic region and the two segments flanking it (in orange) are considered. Pie charts show percentage of exonic segments overlapping m$^6$A peaks. P-value and odds ratio from Fisher's exact test. (c) Isoform switch in *Fbxl5*. Red arrow shows location of m$^6$A methylation. (d) Isoform switch in gene *Kif1b* leads to upregulation of shorter Kif1b-alpha isoform with loss of several protein domains. Red arrow indicates location of m$^6$A methylation. (e) Splicing enrichment analysis: Proportion of isoform switches resulting in gain of each splicing event in KO NPCs. Test of equal proportions was used to identify proportions significantly different from 0.5. Error bars represent 95% confidence interval. (f) Number of intron retention losses resulting in (left) Nonsense mediated decay sensitive or insensitive isoforms and (right) shorter or longer 3'UTR length.

formed ONT long-read native RNA (ONT-RNA) and direct cDNA (ONT-cDNA) sequencing of wild-type and *Mettl14* knock-out mouse NPCs and obtained 2.1 and 1.8 million high-quality reads in WT and KO, respectively, from the native RNA library, and 5.8 and 4.9 million reads in WT and KO, respectively, from the cDNA library. Gene and transcript expression levels were well-correlated between ONT and Ladder-seq samples (Fig. 6.11 and Tables B.6,B.7) and consistent with previously reported correlations between ONT and conventional RNA-seq data [95, 60].

### 6.3.1 Processing of ONT long-read libraries

Though ONT reads are theoretically supposed to span entire transcripts many of the long reads capture only partial transcripts due to premature degradation of the

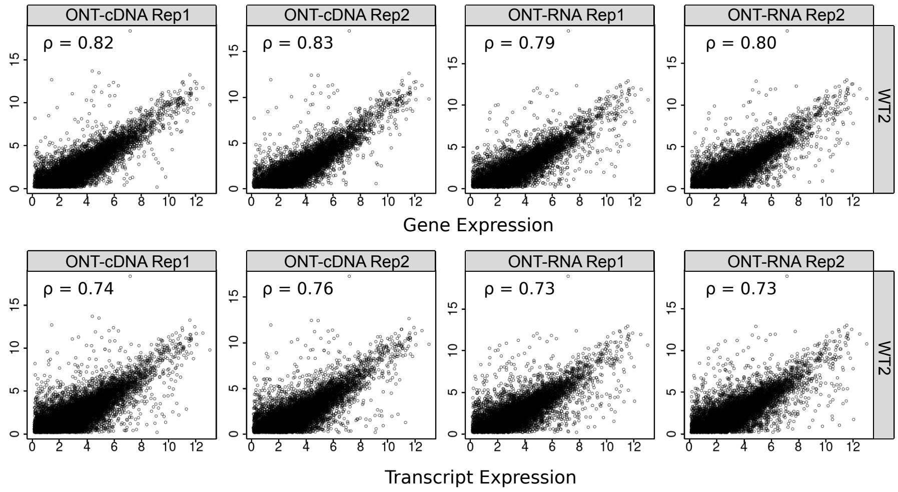

Figure 6.11: Scatter plot of $\log_2$-transformed gene (top row) and transcript (bottom row) expression values (TPM) estimated from Ladder-seq sample WT2 and the four ONT long read samples from WT NPCs. Expression in the Ladder-seq samples was estimated by kallisto from pooled reads ignoring their separation into bands. Only protein coding genes with expression higher than 1TPM in both compared samples were included in the analysis. Pearson correlation coefficients are shown for all pairwise comparisons. Correlations with three other Ladder-seq samples are listed in Tables B.6 and B.7.

mRNA, long molecules breaking during library preparation and failure of the reverse transcriptase to capture the complete molecule in case of cDNA sequencing [29]. We used FLAIR v1.5.1 [41] to identify and StringTie2 [29] to assemble transcripts from ONT reads.

ONT reads were aligned to Ensembl mouse genome assembly GRCm38 using minimap2 v2.17-r941. Following recommendations at `https://github.com/lh3/minimap2`, we used option `-ax splice` to allow spliced alignments and provided splice junctions extracted from the corresponding Ensembl release 95 transcriptome annotation with parameter `--junc-bed`. In the alignment of native RNA reads, we additionally used options `-k14 -uf` as recommended. We ran FLAIR with default settings on pooled reads from both WT and KO replicates and extracted condition-specific transcripts that had an estimated count of at least 1 in at least one of the 2 replicates per condition. FLAIR uses minimap2 internally to align reads using options `-ax splice -t 8 --secondary=no` and corrects misaligned splice sites using the Ensemble 95 annotation. It groups corrected reads with identical intron chains while comparing TSS/TSE with a window size of 100 bp, collapsing them to representative transcripts. It retains transcripts with at least 3 aligned reads with minimum MAPQ of 1. StringTie2 was run with the `-L` option (for long reads) on each of 2 bam files generated respectively from pooled replicates of 2 conditions. GffCompare v0.10.4 was used to compare transcripts between ONT data sets and with transcripts assembled in Ladder-seq. Transcripts were considered identical if they shared the exact same sequence of introns. To quantify expression and compute the number and rate of detected annotated transcripts (Ensemble release 95) in an ONT data set, we followed the strategy proposed in [60]. We aligned reads to the mouse cDNA sequences from Ensembl GRCm38.95 using minimap2 with options `-ax map-ont` and quantified their expression using salmon v1.2.1 with options `-l A` and `--noErrorModel`. A transcript was considered detected if its estimated count was at least 1.

### 6.3.2 Identification of transcripts assembled from short reads by StringTie-ls in ONT data

The lower sequencing depth and the higher error rate of long reads as compared to short reads result in an incomplete transcriptome reconstruction from long reads that also includes false transcripts. Nevertheless, a transcript assembled by StringTie-ls from the Ladder-seq data or by StringTie2 from the corresponding RNA-seq sample (ignoring the separation of reads into bands) is likely to be truly expressed if it can be independently identified in the long read data. We therefore provide in Tables B.11 and B.12 the number of transcripts identified from long reads by FLAIR or assembled by StringTie2 that were also assembled by StringTie-ls or its conventional counterpart in at least one of the four short read replicate samples. Conventional StringTie2 missed many long read transcripts successfully recovered by StringTie-ls, in both conditions and compared to both native RNA and cDNA libraries. The large number of transcripts assembled from short reads that were not identified in the long read data but that were contained in the Ensembl gene annotation can be attributed to the incompleteness of the long read transcriptomes.

We compared the Ladder-seq inferred WT transcriptome of mouse NPCs (Fig. 6.1) used to study isoform switches in m$^6$A methylated genes with transcripts identified by FLAIR [41] from ONT long reads. 63.3% of ONT-cDNA transcripts were contained in at least one WT Ladder-seq transcriptome with relative expression at least 0.1 TPM. Among those, 79.3% were independently assembled by StringTie2 from the ONT-cDNA data, whereas only 24.7% of the remaining transcripts (those only reported by ONT-cDNA) were also found by StringTie2 in the long-read data (Fig 6.12a). Similarly, we found 68.7% of transcripts that occur in both our long-read and the Ladder-seq transcriptome to also be contained in a recently published (Dong et al. [96]) ONT long-read mouse NPC transcriptome (Fig. 6.12a). In contrast, this independently generated set of transcripts contained only 11.4% of ONT-cDNA transcripts that were not reported by Ladder-seq. The substantially lower validation rate in an independent data set or by the StringTie2 assembly suggests that a larger fraction of transcripts missing in the Ladder-seq transcriptomes were falsely inferred by FLAIR from ONT-cDNA reads, and similarly from our ONT-RNA data (Fig. 6.13a). In contrast, 32.5% of transcripts uniquely identified by Ladder-seq (average TPM $\geq$ 1) were also identified in the previously published NPC dataset by Dong et al. This almost three times higher validation rate suggests high confidence for this subset of transcripts, which is further supported by a larger fraction of annotated transcripts amongst them. While 69% of Ladder-seq-only transcripts (TPM $\geq$ 1) were annotated, this was true for only 2.7% of FLAIR-only transcripts (TPM $\geq$ 1). Furthermore, a large fraction of FLAIR-only transcripts (18.1% compared to 0.5% for Ladder-seq-only transcripts) matched an annotated sequence of introns only partially, which may reflect a failure of ONT-reads to cover full-length transcript sequences [60].

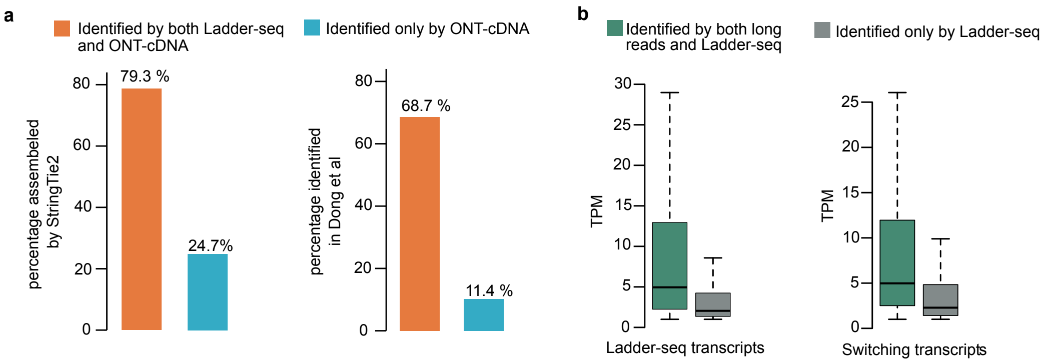

Figure 6.12: Comparison of Ladder-seq and ONT direct cDNA long-read sequencing (ONT-cDNA) on mouse NPCs. (a) Orange bars show validation by StringTie2 (left panel) or by an independent ONT dataset (Dong et al. [96]) (right panel) of transcripts found by both Ladder-seq and ONT-cDNA while light blue bars show validation values for transcripts reported only by ONT-cDNA. (b) Boxplots showing expression levels (TPM) for transcripts identified both by long-reads and Ladder-seq (green boxes) and for transcripts identified only by Ladder-seq (grey boxes). Left panel shows values for all Ladder-seq transcripts with TPM higher than 1. Right panel shows values for Ladder-seq switching transcripts with TPM higher than 1. Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ±1.5× interquartile range.

As expected [60], Ladder-seq detected more annotated genes and transcripts than could be mapped from the ONT libraries (Fig. 6.14 and 6.15). Nevertheless, 71.1% of transcripts reconstructed by Ladder-seq with relative abundance at least 1 TPM were identified by FLAIR or assembled by StringTie2 in the ONT-cDNA data set, or were contained in Dong et al. (Fig. 6.16). This overlapping set of transcripts showed higher expression levels than the remaining set of transcripts that were uniquely identified by Ladder-seq (Figure 6.12b), suggesting the limited sequencing depth of the ONT data set as one possible explanation for their absence in the long-read transcriptome [60]. This was consistently observed in the ONT-RNA data (Fig. 6.16 and 6.13b). A more likely explanation for the low abundance of transcripts reported only by FLAIR (Fig. 6.17) is a higher rate of incorrectly inferred sequences among them as suggested by their low validation rate and low fraction of annotated transcripts (see above). A similar fraction (57.8%) of transcripts upregulated in WT or KO as part of an isoform switch in our Ladder-seq analysis were identified by FLAIR or assembled by StringTie2 in our WT and KO ONT-cDNA data sets. We observed a similar shift in relative transcript expression between overlapping and uniquely identified switching transcripts (Fig. 6.12b and Fig. 6.13b).

### 6.3.3 Validation of isoform switches inferred by Ladder-seq in ONT read data

For 5 out of the 6 isoform switches validated by RT-qPCR (Fig. 6.7), the two participating isoforms were identified by at least one of the two methods (StringTie2 or

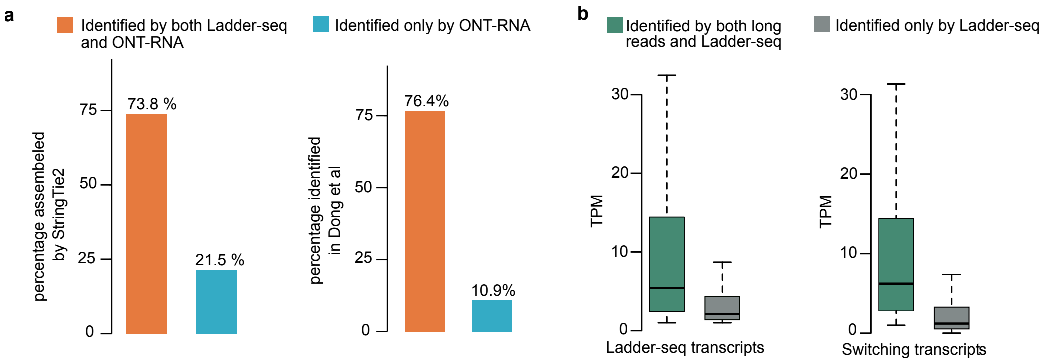

Figure 6.13: Comparison of Ladder-seq and ONT native RNA long-read sequencing (ONT-RNA) on mouse NPCs. (a) Orange bars show validation by StringTie2 (left panel) or by an independent ONT dataset (Dong et al. [96]) (right panel) of transcripts found by both Ladder-seq and ONT-RNA while light blue bars show validation values for transcripts reported only by ONT-RNA. (b) Boxplots showing expression levels (TPM) for transcripts identified both by long-reads and Ladder-seq (green boxes) and for transcripts identified only by Ladder-seq (grey boxes). Left panel shows values for all Ladder-seq transcripts with TPM higher than 1. Right panel shows values for Ladder-seq switching transcripts with TPM higher than 1. Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ±1.5× interquartile range.

FLAIR) in the ONT-cDNA data set (Appendix Table B.9). StringTie2 alone assembled both involved transcripts for 4 isoform switches, while FLAIR identified them only in 2 switches. The single switch for which both methods independently detected both isoforms was formed by the two highest expressed transcripts. In contrast, the only isoform missed by both methods was the lowest expressed among all 12 transcripts. Additionally, the shorter transcript which is absent from the mm10 Ensembl release 95 and exists in the later release 98 version (ENSMUST00000235932) (Section 6.1.1.2), is also present in the ONT data (Appendix Table B.9) and (Appendix Table B.10. Overall, the two methods disagreed on the presence of 6 out of 12 validated switching isoforms, which underlines the nontrivial nature of the computational task of inferring high confidence transcripts from long reads. As expected, the lower sequencing depth in the ONT-RNA data set resulted in a smaller number of confirmed isoforms (Appendix Table B.10).

Finally, we used ONT long reads of WT and KO NPCs to validate novel transcripts involved in isoform switches reported only by conventional RNA-seq or switches identified only by Ladder-seq. Among all 499 novel switching isoforms detected exclusively by Ladder-seq, 206 (41.3%) were identified from ONT-cDNA or ONT-RNA long read data by FLAIR or assembled by StringTie2, or were contained in a recently published [96] ONT long-read mouse NPC transcriptome. While the validation rate among novel switching isoforms identified by both Ladder-seq and conventional RNA-seq is slightly higher (56.9%), only 18 out of 97 (18.6%) novel switching isoforms reported only by conventional RNA-seq were confirmed by long-read sequencing.

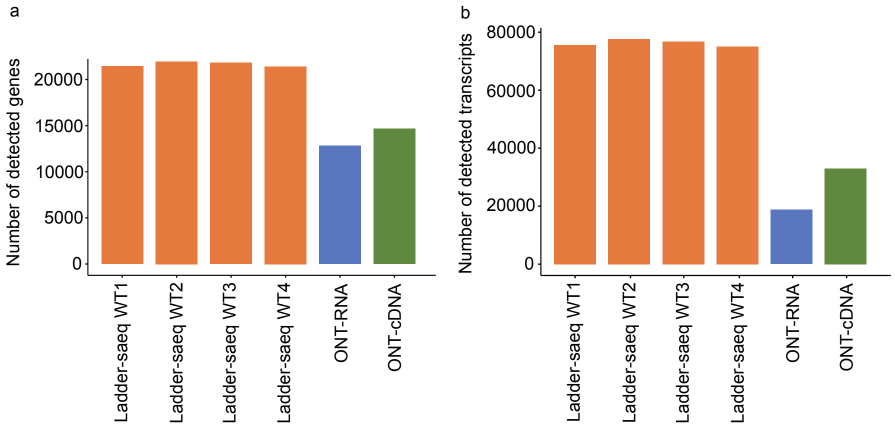

Figure 6.14: Number of detected genes (a) and transcripts (b). An annotated gene or transcript is considered detected if the estimated count is at least 1. Both replicates were pooled for the native RNA and direct cDNA ONT samples of WT NPCs.

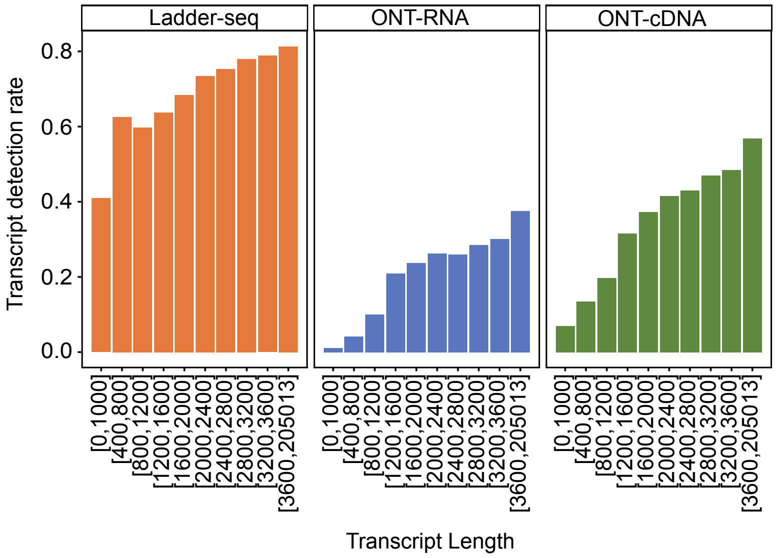

Figure 6.15: Transcript detection rate of Ladder-seq and ONT long-read sequencing. The fraction of transcripts with estimated count at least 1 is stratified by transcript length, using identical ranges as in [60]. For Ladder-seq, mean transcript fractions across the 4 WT NPC samples are reported (variance not visible). Both replicates were pooled for the native RNA and direct cDNA ONT samples of WT NPCs.

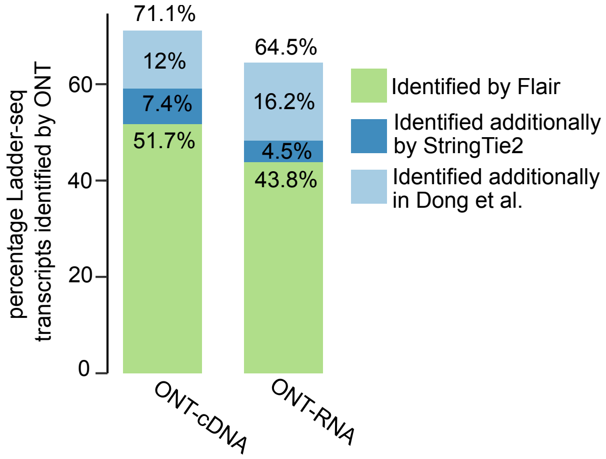

Figure 6.16: Cumulative percentage of Ladder-seq transcripts identified by long-read sequencing. Bars show percentage of Ladder-seq transcripts identified by FLAIR (green), plus those additionally identified by StringTie2 (blue), plus transcripts additionally found in a recently published long-read mouse NPC transcriptome (light blue) (Dong et al. [96]).

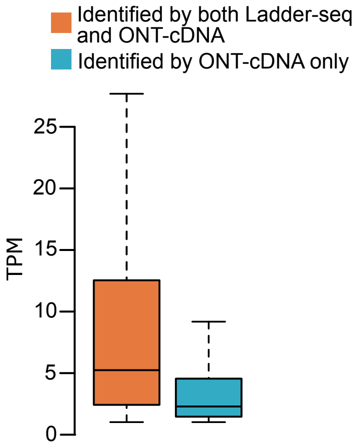

Figure 6.17: Comparison of expression between common and FLAIR-only transcripts. Boxplot shows expression levels (TPM) for transcripts identified both by long-reads (using FLAIR) and Ladder-seq (TPM $\geq 1$) and for transcripts identified only by FLAIR from ONT-cDNA reads. Boxplot definition: Bottom and top of the box correspond to lower and upper quartiles of the data, bar is the median and whiskers are median ±1.5× interquartile range.

*Chapter 7*

# Discussion and Outlook

## 7.1 Discussion

In this thesis we introduced Ladder-seq, a combined experimental-computational approach that dramatically improves the accuracy with which the set of expressed transcripts can be inferred from short RNA-seq reads, incorporating varying degrees of prior knowledge of a species' genome sequence or its transcriptome. The experimental separation of transcripts by their lengths provides an additional layer of information that can be utilized by computational analysis methods to detect and quantify transcripts that cannot be distinguished based on short read data alone. In contrast, a higher read depth alone cannot mitigate identifiability issues of conventional RNA-seq [97, 98]. We demonstrated that a more accurate reconstruction of the transcriptome benefits its subsequent comparison and in our experiments revealed isoform switches of differentially methylated transcript isoforms that are invisible to conventional RNA-seq approaches.

### 7.1.1 Computational framework

Our computational framework for reference-based and *de novo* assembly of transcripts from Ladder-seq reads employs previously developed methods StringTie2 and Trinity without any internal modifications. We therefore provide a Snakemake-based [99] workflow template that allows users to implement the same framework based on other methods that have originally been developed for the analysis of conventional RNA-seq data. This will make a plethora of computational methods that have been developed over the last decade instantly available for the analysis of Ladder-seq data sets.

On the other hand, we expect algorithms that are tailored to the specifics of Ladder-seq to even further improve the accuracy of reconstructed transcriptomes. Our modification of kallisto explicitly models the length information contained in Ladder-seq reads as well as the errors made in the experimental separation of transcripts, and considers reads from all bands at once. StringTie-ls and Trinity-ls, however, use their conventional counterparts as black boxes separately on each band and impose discrete length constraints on their output. Algorithms that borrow information across bands when analyzing splicing graphs or *de Bruijn* graphs and take length information into account already during graph traversal would make better use of Ladder-seq data and could help to push the boundaries of this technology even further.

### 7.1.2 Experimental protocol

On the experimental side, the Ladder-seq protocol involves a denaturing gel electrophoresis to achieve length separation of mRNAs. In our proof of principle experiment we separated transcripts into 7 bands. In principle, a larger number of cuts could further reduce the effective complexity transcriptome-wide (Fig. 2.2(b)) or of a subset of genes of interest, and thus simplify the computational task of inferring their expressed transcripts. On the other hand, fewer cuts might be sufficient to achieve a similar improvement over conventional RNA-seq for species with a less complex transcriptome. In our repository we therefore provide R code that can guide the selection of the number and approximate location of cuts. It visualizes (see Fig. 2.2 for an example) and summarizes the distributions of original gene complexities and resulting effective complexities using descriptive statistics either genome-wide or for a given set of genes of interest, based on a related RNA-seq data set of a given species. In practice, the separation accuracy of the gel will ultimately limit the number of cuts that will benefit the computational analysis, but other separation strategies might imply different trade-offs.

We used a gel based approach to separate transcripts because of its relative simplicity and low cost. Most laboratories have access to this technology, making it easy for groups working routinely with RNA sequencing to implement our novel Ladder-seq protocol. However, the separation of mRNAs by their lengths could be achieved using other technologies including solid phase reversible immobilization beads [100], capillary electrophoresis [101], and ion-pair reversed-phase HPLC [102]. These methods will vary in degrees of accuracy in separating mRNAs, costs, and level of involvement for the experimentalist. As we demonstrated with our simulated data experiments, a higher accuracy in the separation step will yield a greater advantage in transcriptome reconstruction.

High accuracy of Ladder-seq transcriptomes of mouse NPCs was confirmed by comparison with transcripts inferred from ONT long reads. While the overlap between the two technologies was large, many transcripts were uniquely inferred from long reads. Their substantially lower validation rate, however, suggests the presence of a larger fraction of false transcripts. Alternatively, the low expression of transcripts uniquely identified by Ladder-seq indicates the limited sequencing depth of ONT as a possible reason for their absence in the long-read data set. Both differences between long-read sequencing and Ladder-seq are expected. Even though long-read technology greatly simplifies many analytical challenges that occur in short-read assembly, experimental challenges and higher error rate of long reads motivated the development of different computational strategies to extract high-confidence, full-length transcripts. Different approaches and filtering criteria can yield substantially different results [96], as observed in our own experiments using StringTie2 and FLAIR. In addition, long-read sequencers have much lower throughput and thus detect a much smaller fraction of genes and transcripts as contained in short-read libraries. The lower sequencing depth renders the statistical comparison of transcript abundances between conditions as performed in our study infeasible. Current studies therefore combine long reads with high-throughput short-read (Ilumina) sequencing [103], and limit the differential analysis to fold change calculations [104]. Ladder-seq improves this limitation by combining the high throughput of short-read RNA-seq with the ability to reveal transcript isoforms that are invisible to conventional RNA-seq. However, if a large number of overlapping transcripts expressed by a complex gene have similar lengths, Ladder-seq will not offer any benefit over conventional RNA-seq in resolving such intrinsically difficult expression patterns from short reads.

### 7.1.3 Biological findings

In our Ladder-seq experiment on mouse NPCs, we explored the consequences of the deletion of m$^6$A writer protein *Mettl14* on isoform usage. Ladder-seq identified a large number of genes with isoform switches. We showed that differentially spliced exonic segments of a transcript tend to lie close to a methylation site. This result suggests a direct involvement of m$^6$A in alternative splicing in NPCs, possibly through interaction of m$^6$A readers with the splicing machinery as it has been reported for other cell types and organisms [11, 12, 13, 14]. Which nuclear m$^6$A reader is active in NPCs remains to be determined. An intriguing finding of our study is the enrichment for intron retention losses in *Mettl14* KO NPCs in non-neuronal genes related to DNA repair and gamete generation. Intron retentions are known to act as regulators of gene expression during normal development [105], and previous work reported progressive intron retention gains in genes related to cell cycle, pluripotency and DNA repair during the process of differentiation from mouse embryonic stem cells to neurons [106]. Expression of these genes is under tighter control as differentiation progresses. Intron retention losses in *Mettl14* KO NPCs suggest that they are in a lesser state of differentiation compared to WT NPCs, which fits with the previous finding of delayed differentiation of radial glial cells (RGC) in *Mettl14* KO mice [56]. This is the first in-depth analysis of m$^6$A-mediated alternative splicing in NPCs, and highlights the diversity of m$^6$A function within a single cell type. It further extends the role of m$^6$A in NPCs from mediating mRNA degradation [56] to regulating isoform usage, which is known to be especially important in the brain.

## 7.2 Outlook

The identification and quantification of RNA molecules in biological samples have come a long way starting from the days of microarrays to third generation and single cell sequencing. The per base cost of bulk RNA sequencing has decreased drastically over the last decades to a point now that an entire genome can be sequenced under 1000 dollars. The throughput on the other hand has exponentially increased and nowadays a standard sequencing experiment would be able to produce approximately 30 to 100 million reads which would have been unimaginable twenty years ago when the protocol was first introduced. The extremely high depth of short read RNA-seq enables researchers to assemble transcripts and estimate abundance quite accurately. In the last two decades, a multitude of computational methods have been developed for the purpose of down stream analysis of RNA-seq data. We have discussed a few of the most important methods and analyses in the introduction. Nevertheless there are some drawbacks of short read RNA-seq data for example limited read length and multi-mapping reads posing difficulty in assembling repeat regions in the genome.

The transcriptomic space is very complex and the characterization of expressed transcripts can be quite difficult, mostly because of the ambiguity in recognizing the originating transcript from RNA-seq reads. We set out with the goal of deconvolving the transcriptomic puzzle by providing an extra layer of information to the methods for assembly and quantification. To this end, we have worked on a new protocol for RNA-seq called Ladder-seq which aims to separate transcripts by their length prior to their fragmentation using a gel-electrophoresis technique. We have extended one the most used methods for quantification and developed pipelines for both reference based and denovo assembly using state-of-the-art tools. Our benchmarking point to a significant improvement of Ladder-seq based methods compared to conventional RNA-seq based methods.

Ladder-seq, the concerted advancement of the RNA-seq protocol and its computational methods, will allow research facilities to study the composition and dynamics of the transcriptome at an unprecedented level of accuracy based on a technology that has been established for over more than a decade.

*Appendix A*

# Software and data availability

## Data availability

Ladder-seq raw sequencing data from WT and *Mettl14* KO mouse NPCs, conventional RNA-seq data from WT mouse NPCs and ONT long-read sequencing from WT and Mettl14 KO mouse NPCs are available in GEO (GSE158985).

## Code availability

The kallisto-ls, StringTie-ls, and Trinity-ls programs and workflows are available at

- `https://github.com/canzarlab/kallisto-ls`,
- `https://github.com/canzarlab/LadderSeq-Assembly`, and
- `https://github.com/canzarlab/LadderSeq-DeNovo`,

respectively. The results of our benchmark studies can be reproduced via a Snakefile [99] available at `https://github.com/canzarlab/ladder_benchmark`.

*Appendix B*

# Comparison of Ladder-seq with other datasets

| Sample | Band | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
| KO1 | 4116 | 5178 | 5377 | 6916 | 7483 | 11245 | 27274 |
| KO2 | 3494 | 3881 | 6393 | 7415 | 8638 | 12206 | 24566 |
| KO3 | 3895 | 3534 | 5581 | 7918 | 9927 | 11697 | 24203 |
| KO4 | 3916 | 5328 | 5967 | 7711 | 8647 | 9518 | 25819 |
| WT1 | 3825 | 4249 | 5716 | 8906 | 7856 | 8786 | 28274 |
| WT2 | 4044 | 4908 | 5375 | 8394 | 5797 | 9925 | 32397 |
| WT3 | 4717 | 3622 | 6125 | 8422 | 10350 | 8140 | 28191 |
| WT4 | 4015 | 3677 | 5851 | 9518 | 8748 | 12355 | 22991 |

Table B.1: Approximate number of transcripts per band in WT and KO Ladder-seq NPC samples. Values reported denote the number of transcripts whose majority of uniquely assigned reads were obtained from the corresponding band.

| Reference sample | Ladder-Seq | Tardaguila et al. | Chen et al. |
|---|---|---|---|
| NPC Rep1 | 0.82 (0.0016) | 0.68 (0.0009) | 0.79 (0.0095) |
| NPC Rep2 | 0.82 (0.0031) | 0.67 (0.0043) | 0.78 (0.0090) |
| NPC Rep3 | 0.81 (0.0054) | 0.66 (0.0048) | 0.76 (0.0085) |

Table B.2: Pearson correlation of transcript expressions. The values reported are mean correlation coefficients and standard deviations across 4 WT Ladder-seq NPC samples and across 2 and 3 regular RNA-seq samples of mouse WT NPCs by Tardaguila et al. [40] and Chen et al. [59], respectively. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. $\log_2$-transformed values (TPM) were compared to the 3 reference RNA-seq samples from WT NPCs.

| Number | Length interval (bp) |
|---|---|
| 1 | [0,909] |
| 2 | [910,1082] |
| 3 | [1083,1267] |
| 4 | [1268,1439] |
| 5 | [1440,1589] |
| 6 | [1590,1756] |
| 7 | [1757,1919] |
| 8 | [1920,2069] |
| 9 | [2070,2272] |
| 10 | [2273,2461] |
| 11 | [2462,2700] |
| 12 | [2701,2931] |
| 13 | [2932,3229] |
| 14 | [3230,3553] |
| 15 | [3554,3892] |
| 16 | [3893,4524] |
| 17 | [4525,5323] |
| 18 | [5324,6777] |
| 19 | [6378, - ) |

Table B.3: Transcript length intervals used in the analysis of neural progenitor cells. For each length range listed, we estimate a probability mass function that models the migration pattern of transcripts whose lengths fall within that range.

| Reference sample | Ladder-Seq | Tardaguila et al. | Chen et al. |
|---|---|---|---|
| NPC Rep1 | 0.95 (0.0020) | 0.89 (0.0029) | 0.95 (0.0008) |
| NPC Rep2 | 0.94 (0.0021) | 0.89 (0.0013) | 0.94 (0.0017) |
| NPC Rep3 | 0.94 (0.0016) | 0.88 (0.0013) | 0.93 (0.0025) |

Table B.4: Pearson correlation of gene expressions. The values reported are mean correlation coefficients and standard deviations across 4 WT Ladder-seq NPC samples and across 2 and 3 regular RNA-seq samples of mouse WT NPCs by Tardaguila et al. [40] and Chen et al. [59], respectively. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. $\log_2$-transformed values (TPM) were compared to the 3 reference RNA-seq samples from WT NPCs.

| Reference sample | Ladder-Seq | Tardaguila et al. | Chen et al. |
|---|---|---|---|
| NPC Rep1 | 0.82 (0.0016) | 0.68 (0.0009) | 0.79 (0.0095) |
| NPC Rep2 | 0.82 (0.0031) | 0.67 (0.0043) | 0.78 (0.0090) |
| NPC Rep3 | 0.81 (0.0054) | 0.66 (0.0048) | 0.76 (0.0085) |

Table B.5: Pearson correlation of transcript expressions. The values reported are mean correlation coefficients and standard deviations across 4 WT Ladder-seq NPC samples and across 2 and 3 regular RNA-seq samples of mouse WT NPCs by Tardaguila et al. [40] and Chen et al. [59], respectively. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq. $\log_2$-transformed values (TPM) were compared to the 3 reference RNA-seq samples from WT NPCs.

| Reference sample | WT1 | WT2 | WT3 | WT4 |
|---|---|---|---|---|
| ONT-cDNA Rep1 | 0.82 | 0.82 | 0.83 | 0.82 |
| ONT-cDNA Rep2 | 0.82 | 0.83 | 0.84 | 0.83 |
| ONT-RNA Rep1 | 0.78 | 0.79 | 0.80 | 0.79 |
| ONT-RNA Rep2 | 0.78 | 0.80 | 0.80 | 0.79 |

Table B.6: Pearson correlation of gene expressions between Ladder-seq and ONT long read samples. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq samples WT1-4. $\log_2$-transformed values (TPM) were compared to the 2 biological replicates of ONT cDNA and direct RNA samples from WT NPCs. Only protein coding genes with expression higher than 1TPM in both compared samples were included in the analysis.

| Reference sample | WT1 | WT2 | WT3 | WT4 |
|---|---|---|---|---|
| ONT-cDNA Rep1 | 0.74 | 0.74 | 0.75 | 0.75 |
| ONT-cDNA Rep2 | 0.76 | 0.76 | 0.77 | 0.77 |
| ONT-RNA Rep1 | 0.71 | 0.73 | 0.73 | 0.73 |
| ONT-RNA Rep2 | 0.71 | 0.73 | 0.73 | 0.73 |

Table B.7: Pearson correlation of transcript expressions between Ladder-seq and ONT long read samples. Expression was estimated by kallisto using pooled reads across bands in Ladder-seq samples WT1-4. $\log_2$-transformed values (TPM) were compared to the 2 biological replicates of ONT cDNA and direct RNA samples from WT NPCs. Only protein coding genes with expression higher than 1TPM in both compared samples were included in the analysis.

|  |  | Primer pair |  |
|---|---|---|---|
| Gene | Target isoform | forward | reverse |
| *Pi4k2a* | ENSMUST00000066778 | CTGTCATGAGAGGCCAGATCCTA | CCTGTCACCTGCAGGATTTCT |
|  | TCONS_00005143 | ACAATAAGAGCCCCCTGCAC | GACCCCTGCTGGCTCCT |
|  | common | TCAGGGGAGAATCGTTGCTG | CCCTGGTTGAGAACAAGGCA |
| *Tram1l1* | TCONS_00006857 | AGCGGTACCAGAAAGGGTTG | AGAGTGCATTGCCATTCCGA |
|  | TCONS_00006855 | GCCGGTGACTACTGTATCC | GGACCGTCTCTTCCTTCCAC |
|  | common | TGTGGAAGGAAGAGACGGTC | GCACAGAGACACCACATAGC |
| *Fbxl5* | ENSMUST00000047857 | AGGACTAGTGTCTGTTGGCAG | GAAGTCGCTGGGAGTGTAGTC |
|  | TCONS_00008324 | ACCATGGTCTCAGTTGGTCTTG | AGCCTTGCCTGCACTTTTCAG |
|  | common | AAGTGGTCTCAGCTGGCAAA | ATACCAGTCACCTCTTGCCCA |
| *Ptprz1* | ENSMUST00000090568 | ATGACACAGGCATAGCTCCG | GGCTACTATTACTGGCCTCTGC |
|  | TCONS_00008800 | AACCAGTATACAATGAGGCCAGT | CAGACACGATCACAAGGGGT |
|  | common | GATTGTTCACGATGAGCACGG | GACTCCCGGCCTCATCAAAT |
| *Kif1b* | ENSMUST00000030806 | TCCTTTACAAAAAGGAGAAGGAGGA | ATCAGAATCCGCGTCCAGTC |
|  | TCONS_00007441 | GCAGCAGAGACTGGACTACG | CTCTGCAGCCAGAGATCGAG |
|  | common | TTGCCATACGGGAAGATGGG | GCCTGGCCAACCCTTGTAAT |
| *Rai1* | TCONS_00001859 | TTGCCTTCCTCTCTCTCCAG | ACGGCAGCCTCTTATGTTTG |
|  | ENSMUST00000171108 | TAGCTGTGGACATGCCGTGTA | CATTGGCACATGGGTAGTGG |
|  | common | ACATAAGAGGCTGCCGTTGT | CTGGATGGGATCAAGGACCG |

Table B.8: List of primers used for quantitative RT PCR analysis.

| Transcript | Gene of Origin | Novel (yes/no) | Upregulated in: (Genotype) | Identified by Flair | Identified by StringTie2 | in Dong et al. | Transcript TPM |
|---|---|---|---|---|---|---|---|
| TCONS_00006857 | *Tram1l1* | yes | WT | no | yes | no | 5.182658 |
| TCONS_00006855 | *Tram1l1* | yes | KO | no | yes | no | 1.019803 |
| TCONS_00007441 | *Kif1b* | yes | WT | yes | no | no | 15.8415 |
| ENSMUST00000030806 | *Kif1b* | no | KO | yes | yes | yes | 10.74802 |
| ENSMUST00000047857 | *Fbxl5* | no | WT | yes | yes | yes | 8.185762 |
| TCONS_00008324 | *Fbxl5* | yes | KO | no | yes | no | 5.860712 |
| ENSMUST00000171108 | *Rai1* | no | WT | no | no | no | 0.9653895 |
| TCONS_00001859 | *Rai1* | yes | KO | no | yes | no | 2.400327 |
| ENSMUST00000066778 | *Pi4k2a* | no | WT | yes | yes | yes | 12.9707 |
| TCONS_00005143 | *Pi4k2a* | yes | KO | no | yes | yes | 4.405247 |
| ENSMUST00000090568 | *Ptprz1* | no | WT | yes | yes | no | 69.8157 |
| TCONS_00008800 | *Ptprz1* | yes | KO | yes | yes | yes | 27.55367 |

Table B.9: Identification by ONT direct cDNA long-read sequencing (ONT-cDNA) of qPCR-validated switching transcripts in Mettl14 KO mouse NPCs. Table describes whether each differentially used transcript is identified by FLAIR or StringTie2 in our ONT-cDNA dataset or is contained in a recently published long-read mouse NPC transcriptome (Dong et al. [96]).

| Transcript | Gene of Origin | Novel (yes/no) | Upregulated in: (Genotype) | Identified by Flair | Identified by StringTie2 | in Dong et al. | Transcript TPM |
|---|---|---|---|---|---|---|---|
| TCONS_00006857 | *Tram1l1* | yes | WT | yes | no | no | 5.182658 |
| TCONS_00006855 | *Tram1l1* | yes | KO | no | yes | no | 1.019803 |
| TCONS_00007441 | *Kif1b* | yes | WT | no | no | no | 15.8415 |
| ENSMUST00000030806 | *Kif1b* | no | KO | no | no | yes | 10.74802 |
| ENSMUST00000047857 | *Fbxl5* | no | WT | yes | yes | yes | 8.185762 |
| TCONS_00008324 | *Fbxl5* | yes | KO | no | yes | no | 5.860712 |
| ENSMUST00000171108 | *Rai1* | no | WT | no | no | no | 0.9653895 |
| TCONS_00001859 | *Rai1* | yes | KO | no | no | no | 2.400327 |
| ENSMUST00000066778 | *Pi4k2a* | no | WT | yes | yes | yes | 12.9707 |
| TCONS_00005143 | *Pi4k2a* | yes | KO | no | yes | yes | 4.405247 |
| ENSMUST00000090568 | *Ptprz1* | no | WT | yes | yes | no | 69.8157 |
| TCONS_00008800 | *Ptprz1* | yes | KO | yes | yes | yes | 27.55367 |

Table B.10: Identification by ONT native RNA long-read sequencing (ONT-RNA) of qPCR-validated switching transcripts in Mettl14 KO mouse NPCs. Table describes whether each differentially used transcript is identified by FLAIR or StringTie2 in our ONT-RNA dataset or is contained in a recently published long-read mouse NPC transcriptome (Dong et al. [96]).

| | Conventional | StringTie-ls |
|---|---|---|
| FLAIR | 15643 | 17731 |
| FLAIR + Ensembl | 22285 | 26869 |
| StringTie2 | 16438 | 19184 |
| StringTie2 + Ensembl | 23415 | 28588 |
| FLAIR + StringTie2 | 19050 | 22797 |
| FLAIR + StringTie2 + Ensembl | 24539 | 30416 |
| All | 20215 | 25319 |
| All + Ensembl | 24910 | 31654 |

Table B.11: Number of transcripts identified by conventional StringTie2 or StringTie-ls that were independently identified in ONT long reads in WT samples. The set of transcripts assembled in any of the four WT Ladder-seq samples was compared to transcripts identified by FLAIR, assembled by StringTie2, or inferred from either of the two methods (FLAIR + StringTie2). Set *All* additionally contains transcripts from a recently published [96] ONT long-read mouse NPC transcriptome. Each long read transcriptome was alternatively augmented by transcripts annotated in Ensembl release 95 (+Ensembl).

| | Conventional | StringTie-ls |
|---|---|---|
| FLAIR | 15005 | 16888 |
| FLAIR + Ensembl | 22089 | 26128 |
| StringTie2 | 15395 | 17931 |
| StringTie2 + Ensembl | 23252 | 27848 |
| FLAIR + StringTie2 | 18119 | 21554 |
| FLAIR + StringTie2 + Ensembl | 24300 | 29538 |

Table B.12: Number of transcripts identified by conventional StringTie2 or StringTie-ls that were independently identified in ONT long reads in *Mettl14* KO samples. The set of transcripts assembled in any of the four KO Ladder-seq samples was compared to transcripts identified by FLAIR, assembled by StringTie2, or inferred from either of the two methods (FLAIR + StringTie2). Each long read transcriptome was alternatively augmented by transcripts annotated in Ensembl release 95 (+Ensembl).

*Appendix C*

# Quantification results

| Gene complexity | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 23.39 | 12.10 | 9.5 | 8.6 | 8.5 | 8.3 | 9.2 | 8.7 | 7.9 | 8.3 |

Table C.1: Fraction of low expressed transcripts. For each gene complexity, we show the fraction of transcripts with ground truth expression less than 0.5 TPM as estimated by RSEM in sample NA12716_7.

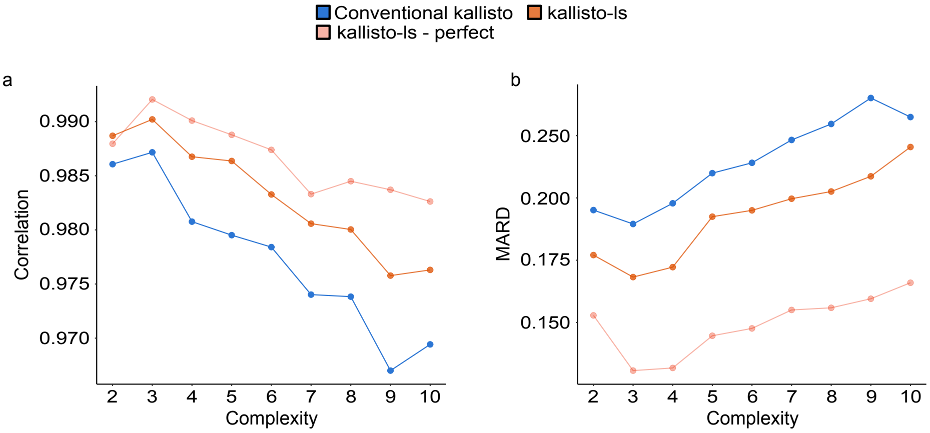

Figure C.1: Quantification accuracy of kallisto-ls compared to conventional kallisto on 75 million paired-end reads simulated by RSEM from GEUVADIS sample NA12716_7. Mean values across 20 simulations are reported. Pearson correlation of estimated and ground truth abundance in $\log_2$ transformed transcripts per million (TPM) and mean absolute relative difference (MARD) are shown as a function of gene complexity, i.e. the number of transcripts expressed by a gene. For ease of visualization, we omit genes expressing a single transcript, many of which are estimated to be lowly expressed in this sample by RSEM. Nevertheless, kallisto-ls achieves slightly better MARD and correlation for this set of transcripts (Table C.3).

| Complexity | Conventional kallisto | | kallisto-ls | | kallisto-ls perfect | |
|---|---|---|---|---|---|---|
| | MARD | Correlation | MARD | Correlation | MARD | Correlation |
| 1 | 0.288 | 0.924 | 0.272 | 0.932 | 0.261 | 0.935 |
| 2 | 0.240 | 0.974 | 0.211 | 0.981 | 0.180 | 0.981 |
| 3 | 0.229 | 0.976 | 0.194 | 0.983 | 0.162 | 0.986 |
| 4 | 0.234 | 0.967 | 0.198 | 0.979 | 0.160 | 0.984 |
| 5 | 0.256 | 0.961 | 0.212 | 0.978 | 0.170 | 0.981 |
| 6 | 0.252 | 0.962 | 0.215 | 0.974 | 0.167 | 0.979 |
| 7 | 0.264 | 0.956 | 0.224 | 0.970 | 0.177 | 0.976 |
| 8 | 0.268 | 0.955 | 0.223 | 0.969 | 0.177 | 0.976 |
| 9 | 0.279 | 0.946 | 0.232 | 0.964 | 0.182 | 0.974 |
| 10 | 0.280 | 0.944 | 0.241 | 0.962 | 0.194 | 0.971 |
| 11+ | 0.315 | 0.922 | 0.268 | 0.950 | 0.212 | 0.958 |

Table C.2: Quantification accuracy of kallisto-ls compared to conventional kallisto on 30 million paired-end reads simulated by RSEM from GEUVADIS sample NA12716_7. Mean values across 20 simulations are reported. Pearson correlation of estimated and ground truth abundance in $\log_2$ transformed transcripts per million (TPM) and mean absolute relative difference (MARD) are shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene. In the calculation of MARD, the absolute difference between estimated counts and the ground truth is divided by the sum of the two, where transcripts with zero estimates by both methods were excluded. kallisto-ls perfect refers to the results of kallisto-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration error.

| Complexity | Conventional kallisto | | kallisto-ls | | kallisto-ls perfect | |
|---|---|---|---|---|---|---|
| | MARD | Correlation | MARD | Correlation | MARD | Correlation |
| 1 | 0.265 | 0.936 | 0.254 | 0.942 | 0.243 | 0.944 |
| 2 | 0.195 | 0.986 | 0.177 | 0.988 | 0.152 | 0.987 |
| 3 | 0.189 | 0.987 | 0.168 | 0.990 | 0.130 | 0.992 |
| 4 | 0.197 | 0.980 | 0.172 | 0.986 | 0.131 | 0.990 |
| 5 | 0.209 | 0.979 | 0.192 | 0.986 | 0.144 | 0.988 |
| 6 | 0.214 | 0.978 | 0.195 | 0.983 | 0.147 | 0.987 |
| 7 | 0.223 | 0.974 | 0.199 | 0.980 | 0.155 | 0.983 |
| 8 | 0.229 | 0.973 | 0.202 | 0.980 | 0.155 | 0.984 |
| 9 | 0.240 | 0.966 | 0.208 | 0.975 | 0.159 | 0.983 |
| 10 | 0.232 | 0.969 | 0.220 | 0.976 | 0.165 | 0.982 |
| 11+ | 0.273 | 0.954 | 0.242 | 0.967 | 0.188 | 0.973 |

Table C.3: Quantification accuracy of kallisto-ls compared to conventional kallisto on 75 million paired-end reads simulated by RSEM from GEUVADIS sample NA12716_7. Mean values across 20 simulations are reported. Pearson correlation of estimated and ground truth abundance in $\log_2$ transformed transcripts per million (TPM) and mean absolute relative difference (MARD) are shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene. In the calculation of MARD, the absolute difference between estimated counts and the ground truth is divided by the sum of the two, where transcripts with zero estimates by both methods were excluded. kallisto-ls perfect refers to the results of kallisto-ls on the most optimistic Ladder-seq experiment in which transcripts perfectly separate by length, without any migration error.

*Appendix D*

# Assembly results

## D.1 Ref-based assembly results

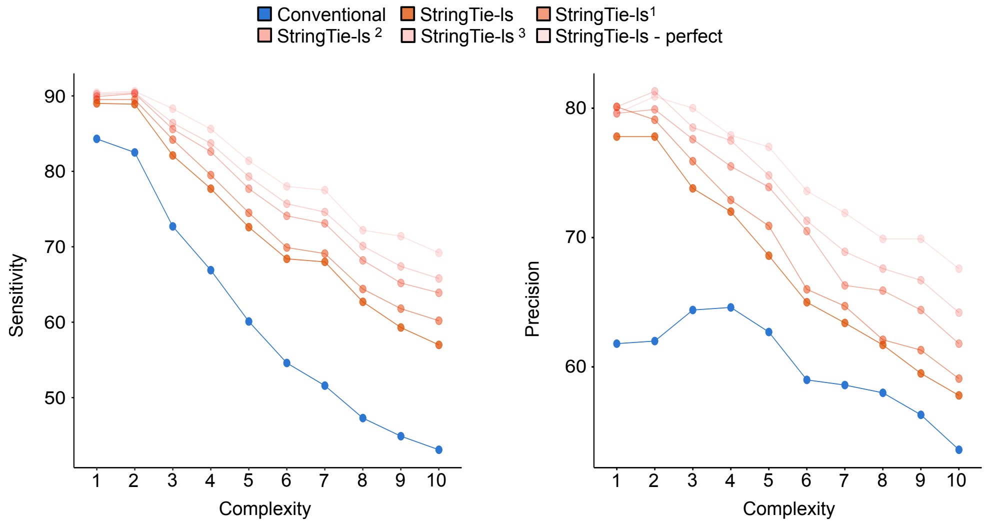

Figure D.1: Accuracy of transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq paired-end reads. RNA-seq and Ladder-seq reads were aligned identically to the reference genome (GRCh38) using STAR [25]. Sensitivity and precision of StringTie-ls and its conventional counterpart StringTie2 are shown as a function of gene complexity measured as the number of expressed transcripts. Sensitivity and precision are calculated with respect to the same set of ground truth transcripts as in the smaller 30 million read pairs data set. All results are listed in Appendix Tables D.3 and D.4.

| Complexity | Sensitivity | | | | | |
|---|---|---|---|---|---|---|
| | Conventional | StringTie-ls | StringTie-ls [1] | StringTie-ls [2] | StringTie-ls [3] | StringTie-ls perfect |
| 1 | 74.8 | 75.6 | 76.2 | 77.0 | 77.0 | 77.5 |
| 2 | 78.5 | 82.8 | 83.7 | 84.0 | 84.2 | 84.4 |
| 3 | 70.0 | 78.6 | 79.6 | 81.2 | 82.7 | 83.9 |
| 4 | 65.1 | 75.5 | 77.1 | 79.0 | 80.5 | 81.5 |
| 5 | 57.7 | 70.5 | 71.9 | 74.5 | 75.7 | 77.5 |
| 6 | 52.8 | 66.5 | 67.9 | 71.1 | 72.4 | 74.2 |
| 7 | 49.8 | 65.3 | 67.7 | 70.4 | 72.0 | 74.3 |
| 8 | 45.9 | 60.7 | 62.0 | 66.4 | 67.4 | 69.2 |
| 9 | 43.4 | 58.4 | 59.4 | 63.0 | 64.8 | 67.5 |
| 10 | 42.1 | 56.9 | 59.1 | 62.1 | 64.2 | 66.8 |
| 11+ | 32.8 | 48.8 | 50.4 | 54.2 | 56.2 | 58.5 |

Table D.1: Sensitivity of transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq paired-end reads. Sensitivity of StringTie-ls and its conventional counterpart StringTie2 is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

| Complexity | Precision | | | | | |
|---|---|---|---|---|---|---|
| | Conventional | StringTie-ls | StringTie-ls [1] | StringTie-ls [2] | StringTie-ls [3] | StringTie-ls perfect |
| 1 | 60.3 | 78.9 | 80.2 | 80.1 | 80.3 | 81.1 |
| 2 | 62.0 | 77.2 | 78.9 | 80.4 | 80.5 | 80.6 |
| 3 | 62.3 | 72.6 | 74.1 | 75.6 | 77.1 | 77.8 |
| 4 | 65.6 | 72.9 | 74.6 | 76.6 | 77.8 | 78.2 |
| 5 | 63.9 | 69.3 | 71.0 | 73.4 | 75.1 | 75.8 |
| 6 | 61.2 | 65.6 | 67.1 | 70.6 | 71.4 | 73.1 |
| 7 | 61.7 | 63.3 | 65.2 | 68.6 | 70.3 | 72.7 |
| 8 | 59.3 | 61.6 | 62.6 | 66.2 | 67.7 | 69.3 |
| 9 | 59.0 | 61.0 | 60.7 | 63.9 | 65.4 | 67.4 |
| 10 | 58.4 | 59.0 | 50.1 | 62.5 | 64.3 | 67.4 |
| 11+ | 52.3 | 53.9 | 54.4 | 57.1 | 58.8 | 60.9 |

Table D.2: Precision of transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq paired-end reads. Precision of StringTie-ls and its conventional counterpart StringTie2 is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

| Complexity | Sensitivity | | | | | |
|---|---|---|---|---|---|---|
| | Conventional | StringTie-ls | StringTie-ls [1] | StringTie-ls [2] | StringTie-ls [3] | StringTie-ls perfect |
| 1 | 84.3 | 89.0 | 89.5 | 89.9 | 90.2 | 90.4 |
| 2 | 82.5 | 88.9 | 89.5 | 90.3 | 90.4 | 90.6 |
| 3 | 72.7 | 82.1 | 84.2 | 85.6 | 86.4 | 88.3 |
| 4 | 66.9 | 77.7 | 79.5 | 82.6 | 83.7 | 85.6 |
| 5 | 60.1 | 72.6 | 74.5 | 77.7 | 79.3 | 81.4 |
| 6 | 54.6 | 68.4 | 69.9 | 74.1 | 75.7 | 78.0 |
| 7 | 51.6 | 68.0 | 69.1 | 73.1 | 74.6 | 77.5 |
| 8 | 47.3 | 62.7 | 64.4 | 68.2 | 70.1 | 72.2 |
| 9 | 44.9 | 59.3 | 61.8 | 65.2 | 67.4 | 71.4 |
| 10 | 43.1 | 57.0 | 60.2 | 63.9 | 65.8 | 69.2 |
| 11+ | 34.0 | 49.9 | 51.0 | 55.6 | 58.1 | 60.9 |

Table D.3: Sensitivity of transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq paired-end reads. Sensitivity of StringTie-ls and its conventional counterpart StringTie2 is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

| Complexity | Precision | | | | | |
|---|---|---|---|---|---|---|
| | Conventional | StringTie-ls | StringTie-ls $^1$ | StringTie-ls $^2$ | StringTie-ls $^3$ | StringTie-ls perfect |
| 1 | 61.8 | 77.8 | 80.1 | 79.6 | 80.1 | 79.6 |
| 2 | 62.0 | 77.8 | 79.1 | 79.9 | 81.3 | 80.9 |
| 3 | 64.4 | 73.8 | 75.9 | 77.6 | 78.5 | 80.0 |
| 4 | 64.6 | 72.0 | 72.9 | 75.5 | 77.5 | 77.9 |
| 5 | 62.7 | 68.6 | 70.9 | 73.9 | 74.8 | 77.0 |
| 6 | 59.0 | 65.0 | 66.0 | 70.5 | 71.3 | 73.6 |
| 7 | 58.6 | 63.4 | 64.7 | 66.3 | 68.9 | 71.9 |
| 8 | 58.0 | 61.7 | 62.1 | 65.9 | 67.6 | 69.9 |
| 9 | 56.3 | 59.5 | 61.3 | 64.4 | 66.7 | 69.9 |
| 10 | 53.6 | 57.8 | 59.1 | 61.8 | 64.2 | 67.6 |
| 11+ | 49.7 | 53.7 | 53.8 | 56.8 | 58.9 | 61.8 |

Table D.4: Precision of transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq paired-end reads. Precision of StringTie-ls and its conventional counterpart StringTie2 is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

## D.2 De novo assembly results

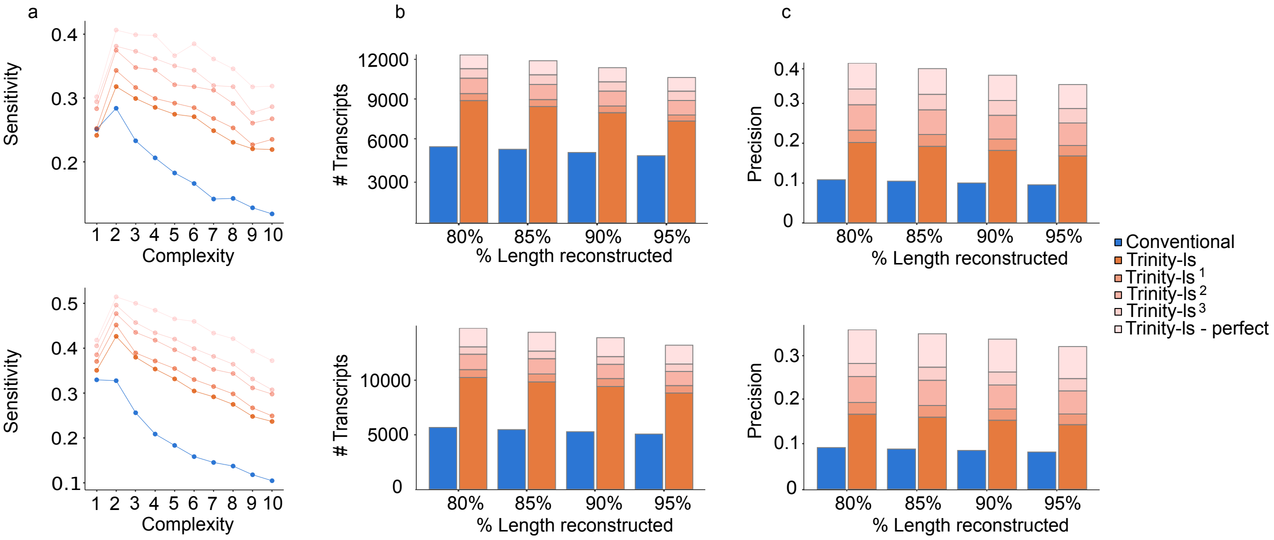

Figure D.2: Accuracy of *de novo* transcript assembly from 30 million (top row) and 75 million (bottom row) simulated RNA-seq and matching Ladder-seq paired-end reads. (a) Sensitivity of Trinity-ls and its conventional counterpart Trinity at 90% transcript length cut-off is shown as a function of gene complexity measured as the number of expressed transcripts. (b) Total number of correctly assembled transcripts at different transcript length cut-offs. (c) Precision at different transcript length cut-offs. All results are listed in Appendix Tables D.7-D.6.

| Method | # of Transcripts | | | | Precision | | | |
|---|---|---|---|---|---|---|---|---|
| | 80% | 85% | 90% | 95% | 80% | 85% | 90% | 95% |
| Conventional Trinity | 5571 | 5377 | 5189 | 4977 | 0.093 | 0.089 | 0.086 | 0.083 |
| Trinity-ls | 10070 | 9671 | 9261 | 8665 | 0.167 | 0.160 | 0.153 | 0.144 |
| Trinity-ls [1] | 10772 | 10388 | 9966 | 9334 | 0.193 | 0.186 | 0.179 | 0.167 |
| Trinity-ls [2] | 12160 | 11756 | 11253 | 10614 | 0.251 | 0.243 | 0.232 | 0.219 |
| Trinity-ls [3] | 12825 | 12438 | 11944 | 11274 | 0.280 | 0.272 | 0.261 | 0.247 |
| Trinity-ls perfect | 14514 | 14138 | 13652 | 12976 | 0.356 | 0.347 | 0.335 | 0.318 |

Table D.5: Accuracy of *de novo* transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq paired-end reads. Total number of correctly assembled transcripts and precision are reported at different transcript length cut-offs.

| Complexity | Sensitivity | | | | | |
|---|---|---|---|---|---|---|
| | Conventional Trinity | Trinity-ls | Trinity-ls [1] | Trinity-ls [2] | Trinity-ls [3] | Trinity-ls perfect |
| 1 | 0.329 | 0.350 | 0.370 | 0.385 | 0.405 | 0.417 |
| 2 | 0.327 | 0.426 | 0.451 | 0.476 | 0.495 | 0.514 |
| 3 | 0.256 | 0.379 | 0.389 | 0.435 | 0.456 | 0.499 |
| 4 | 0.208 | 0.353 | 0.371 | 0.417 | 0.434 | 0.484 |
| 5 | 0.183 | 0.331 | 0.354 | 0.396 | 0.420 | 0.465 |
| 6 | 0.158 | 0.304 | 0.330 | 0.375 | 0.399 | 0.459 |
| 7 | 0.145 | 0.291 | 0.314 | 0.352 | 0.381 | 0.433 |
| 8 | 0.137 | 0.274 | 0.298 | 0.343 | 0.364 | 0.421 |
| 9 | 0.118 | 0.248 | 0.267 | 0.311 | 0.331 | 0.393 |
| 10 | 0.105 | 0.236 | 0.249 | 0.297 | 0.307 | 0.372 |
| 11+ | 0.079 | 0.184 | 0.207 | 0.239 | 0.259 | 0.314 |

Table D.6: Sensitivity of *de novo* transcript assembly from 75 million simulated RNA-seq and matching Ladder-seq paired-end reads. Sensitivity of Trinity-ls and its conventional counterpart Trinity at 90% transcript length cut-off is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

| Method | # of Transcripts | | | | Precision | | | |
|---|---|---|---|---|---|---|---|---|
| | 80% | 85% | 90% | 95% | 80% | 85% | 90% | 95% |
| Conventional Trinity | 5468 | 5283 | 5052 | 4823 | 0.111 | 0.107 | 0.102 | 0.097 |
| Trinity-ls | 8778 | 8351 | 7908 | 7307 | 0.207 | 0.197 | 0.186 | 0.172 |
| Trinity-ls [1] | 9279 | 8849 | 8388 | 7743 | 0.238 | 0.227 | 0.215 | 0.199 |
| Trinity-ls [2] | 10386 | 9941 | 9457 | 8786 | 0.304 | 0.291 | 0.277 | 0.257 |
| Trinity-ls [3] | 11071 | 10632 | 10127 | 9454 | 0.345 | 0.331 | 0.315 | 0.294 |
| Trinity-ls perfect | 12065 | 11644 | 11145 | 10440 | 0.412 | 0.397 | 0.380 | 0.356 |

Table D.7: Accuracy of *de novo* transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq paired-end reads. Total number of correctly assembled transcripts and precision are reported at different transcript length cut-offs.

| Complexity | Sensitivity | | | | | |
|---|---|---|---|---|---|---|
| | Conventional Trinity | Trinity-ls | Trinity-ls [1] | Trinity-ls [2] | Trinity-ls [3] | Trinity-ls perfect |
| 1 | 0.251 | 0.241 | 0.252 | 0.283 | 0.294 | 0.301 |
| 2 | 0.283 | 0.317 | 0.343 | 0.374 | 0.380 | 0.406 |
| 3 | 0.233 | 0.299 | 0.316 | 0.347 | 0.373 | 0.398 |
| 4 | 0.206 | 0.285 | 0.299 | 0.343 | 0.361 | 0.397 |
| 5 | 0.182 | 0.274 | 0.291 | 0.320 | 0.350 | 0.366 |
| 6 | 0.166 | 0.270 | 0.284 | 0.317 | 0.343 | 0.384 |
| 7 | 0.142 | 0.248 | 0.268 | 0.312 | 0.319 | 0.361 |
| 8 | 0.143 | 0.230 | 0.253 | 0.291 | 0.317 | 0.345 |
| 9 | 0.128 | 0.220 | 0.226 | 0.260 | 0.277 | 0.317 |
| 10 | 0.118 | 0.219 | 0.235 | 0.267 | 0.286 | 0.318 |
| 11+ | 0.083 | 0.177 | 0.187 | 0.211 | 0.233 | 0.262 |

Table D.8: Sensitivity of *de novo* transcript assembly from 30 million simulated RNA-seq and matching Ladder-seq paired-end reads. Sensitivity of Trinity-ls and its conventional counterpart Trinity at 90% transcript length cut-off is shown for gene complexities 1-10 and larger than 10 (11+). Complexity denotes the number of transcripts expressed by a gene.

# Bibliography

[1] C. M. O'Connor and J. U. Adams. *Essentials of Cell Biology.* Cambridge, MA: NPG Education, 2010.

[2] S.L. Wolfe. *Biology of the Cell.* Wadsworth Publishing Company, 1972.

[3] L. Pray. Discovery of dna structure and function: Watson and crick. *Nature Education 1(1):100*, 2008.

[4] Sponk Wikimedia Commons. Comparison of a single-stranded rna and a double-stranded dna with their corresponding nucleobases, 2010.

[5] Francis Crick. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, aug 1970.

[6] Jeremy M. Berg, John L. Tymoczko, and Lubert. Stryer. *Biochemistry.* Freeman, New York, 2003.

[7] D. L. Black. Mechanisms of alternative pre-messenger rna splicing. *Annual Reviews of Biochemistry, 72, 291-336.*, 2003.

[8] Xiulin Jiang, Baiyang Liu, Zhi Nie, Lincan Duan, Qiuxia Xiong, Zhixian Jin, Cuiping Yang, and Yongbin Chen. The role of m6a modification in the biological functions and diseases. *Signal Transduction and Targeted Therapy*, 6(1):74, Feb 2021.

[9] Bilian Jin, Yajun Li, and Keith D. Robertson. Dna methylation: superior or subordinate in the epigenetic hierarchy? *Genes & cancer*, 2(6):607–617, Jun 2011. 21941617[pmid].

[10] Ronald C. Desrosiers, Karen H. Friderici, and Fritz M. Rottman. Characterization of novikoff hepatoma mrna methylation and heterogeneity in the methylated 5' terminus. *Biochemistry*, 14(20):4367–4374, Oct 1975.

[11] Irmgard U Haussmann, Zsuzsanna Bodi, Eugenio Sanchez-Moran, Nigel P Mongan, Nathan Archer, Rupert G Fray, and Matthias Soller. m 6 a potentiates sxl alternative pre-mrna splicing for robust drosophila sex determination. *Nature*, 540(7632):301–304, 2016.

[12] Marek Bartosovic, Helena Covelo Molares, Pavlina Gregorova, Dominika Hrossova, Grzegorz Kudla, and Stepanka Vanacova. N6-methyladenosine demethylase fto targets pre-mrnas and regulates alternative splicing and 3'-end processing. *Nucleic acids research*, 45(19):11356–11370, 2017.

[13] Wen Xiao, Samir Adhikari, Ujwal Dahal, Yu-Sheng Chen, Ya-Juan Hao, Bao-Fa Sun, Hui-Ying Sun, Ang Li, Xiao-Li Ping, Wei-Yi Lai, et al. Nuclear m6a reader ythdc1 regulates mrna splicing. *Molecular cell*, 61(4):507–519, 2016.

[14] Katherine I Zhou, Hailing Shi, Ruitu Lyu, Adam C Wylder, Żaneta Matuszek, Jessica N Pan, Chuan He, Marc Parisien, and Tao Pan. Regulation of co-transcriptional pre-mrna splicing by m6a through the low-complexity protein hnrnpg. *Molecular cell*, 76(1):70–81, 2019.

[15] Yuhao Zhang, Xiuchao Geng, Qiang Li, Jianglong Xu, Yanli Tan, Menglin Xiao, Jia Song, Fulin Liu, Chuan Fang, and Hong Wang. m6a modification in rna: biogenesis, functions and roles in gliomas. *Journal of Experimental & Clinical Cancer Research*, 39(1):192, Sep 2020.

[16] R. Karthiya and Piyush Khandelia. m6a rna methylation: Ramifications for gene expression and human health. *Molecular Biotechnology*, 62(10):467–484, Oct 2020.

[17] Kate D. Meyer and Samie R. Jaffrey. Rethinking m(6)a readers, writers, and erasers. *Annual review of cell and developmental biology*, 33:319–342, Oct 2017. 28759256[pmid].

[18] Ben Yue, Chenlong Song, Linxi Yang, Ran Cui, Xingwang Cheng, Zizhen Zhang, and Gang Zhao. Mettl3-mediated n6-methyladenosine modification is critical for epithelial-mesenchymal transition and metastasis of gastric cancer. *Molecular Cancer*, 18(1):142, Oct 2019.

[19] James M. Heather and Benjamin Chain. The sequence of sequencers: The history of sequencing dna. *Genomics*, 107(1):1–8, Jan 2016. 26554401[pmid].

[20] Nick J. Proudfoot, Andre Furger, and Michael J. Dye. Integrating mrna processing with transcription. *Cell*, 108(4):501–512, 2002.

[21] Shanrong Zhao, Ying Zhang, Ramya Gamini, Baohong Zhang, and David von Schack. Evaluation of two main rna-seq approaches for gene quantification in clinical rna sequencing: polya+ selection versus rrna depletion. *Scientific Reports*, 8(1):4781, Mar 2018.

[22] Yongjun Chu and David R. Corey. Rna sequencing: platform selection, experimental design, and data interpretation. *Nucleic acid therapeutics*, 22(4):271–274, Aug 2012. 22830413[pmid].

[23] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan 2009.

[24] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biology*, 15(12):550, Dec 2014.

[25] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. Star: ultrafast universal rna-seq aligner. *Bioinformatics*, 29(1):15–21, Jan 2013.

[26] Daehwan Kim, Ben Langmead, and Steven L. Salzberg. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, 12(4):357–360, March 2015.

[27] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnology*, 34(5):525, 2016.

[28] Rob Patro, Stephen M. Mount, and Carl Kingsford. Sailfish enables alignment-free isoform quantification from rna-seq reads using lightweight algorithms. *Nature Biotechnology*, 32(5):462–464, May 2014.

[29] Sam Kovaka, Aleksey V. Zimin, Geo M. Pertea, Roham Razaghi, Steven L. Salzberg, and Mihaela Pertea. Transcriptome assembly from long-read rna-seq alignments with stringtie2. *Genome Biology*, 20(1):278, 2019.

[30] Mihaela Pertea, Geo M Pertea, Corina M Antonescu, Tsung-Cheng Chang, Joshua T Mendell, and Steven L Salzberg. StringTie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat Biotech*, 33(3):290–295, March 2015.

[31] Mingfu Shao and Carl Kingsford. Accurate assembly of transcripts through phase-preserving graph decomposition. *Nature Biotechnology*, 35(12):1167–1169, Dec 2017.

[32] Stefan Canzar, Sandro Andreotti, David Weese, Knut Reinert, and Gunnar W. Klau. CIDANE: comprehensive isoform discovery and abundance estimation. *Genome Biology*, 17(1):16, Jan 2016.

[33] Li Song and Liliana Florea. Class: constrained transcript assembly of rna-seq reads. *BMC Bioinformatics*, 14(5):S14, Apr 2013.

[34] Jared T. Simpson, Kim Wong, Shaun D. Jackman, Jacqueline E. Schein, Steven J. M. Jones, and Inanç Birol. Abyss: a parallel assembler for short read sequence data. *Genome research*, 19(6):1117–1123, Jun 2009. 19251739[pmid].

[35] Yinlong Xie, Gengxiong Wu, Jingbo Tang, Ruibang Luo, Jordan Patterson, Shanlin Liu, Weihua Huang, Guangzhu He, Shengchang Gu, Shengkang Li, Xin Zhou, Tak-Wah Lam, Yingrui Li, Xun Xu, Gane Ka-Shu Wong, and Jun Wang. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, 30(12):1660–1666, 02 2014.

[36] Marcel H. Schulz, Daniel R. Zerbino, Martin Vingron, and Ewan Birney. Oases: robust de novo rna-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8):1086–1092, Apr 2012. 22368243[pmid].

[37] Manfred G. Grabherr, Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. Birren, Chad Nusbaum, Kerstin Lindblad-Toh, Nir Friedman, and Aviv Regev. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nat Biotechnol*, 29(7):644–652, May 2011.

[38] Sean P. Gordon, Elizabeth Tseng, Asaf Salamov, Jiwei Zhang, Xiandong Meng, Zhiying Zhao, Dongwan Kang, Jason Underwood, Igor V. Grigoriev, Melania Figueroa, Jonathan S. Schilling, Feng Chen, and Zhong Wang. Widespread polycistronic transcripts in fungi revealed by single-molecule mrna sequencing. *PLOS ONE*, 10(7):1–15, 07 2015.

[39] Salah E. Abdel-Ghany, Michael Hamilton, Jennifer L. Jacobi, Peter Ngam, Nicholas Devitt, Faye Schilkey, Asa Ben-Hur, and Anireddy S. N. Reddy. A survey of the sorghum transcriptome using single-molecule long reads. *Nature Communications*, 7(1):11706, Jun 2016.

[40] Manuel Tardaguila, Lorena de la Fuente, Cristina Marti, Cécile Pereira, Francisco Jose Pardo-Palacios, Hector del Risco, Marc Ferrell, Maravillas Mellado, Marissa Macchietto, Kenneth Verheggen, Mariola Edelmann, Iakes Ezkurdia, Jesus Vazquez, Michael Tress, Ali Mortazavi, Lennart Martens, Susana Rodriguez-Navarro, Victoria Moreno-Manzano, and Ana Conesa. Sqanti: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Research*, 28(3):396–411, 2018.

[41] Alison D. Tang, Cameron M. Soulette, Marijke J. van Baren, Kevyn Hart, Eva Hrabeta-Robinson, Catherine J. Wu, and Angela N. Brooks. Full-length transcript characterization of sf3b1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns. *Nature Communications*, 11(1):1438, Mar 2020.

[42] Heng Li. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, 34(18):3094–3100, 05 2018.

[43] Chi Zhang, Baohong Zhang, Lih-Ling Lin, and Shanrong Zhao. Evaluation and comparison of computational tools for rna-seq isoform quantification. *BMC Genomics*, 18(1):583, 2017.

[44] Mingxiang Teng, Michael I. Love, Carrie A. Davis, Sarah Djebali, Alexander Dobin, Brenton R. Graveley, Sheng Li, Christopher E. Mason, Sara Olson, Dmitri Pervouchine, Cricket A. Sloan, Xintao Wei, Lijun Zhan, and Rafael A.

Irizarry. A benchmark for RNA-seq quantification pipelines. *Genome Biology*, 17(1):74, Apr 2016.

[45] Douglas C. Wu, Jun Yao, Kevin S. Ho, Alan M. Lambowitz, and Claus O. Wilke. Limitations of alignment-free tools in total RNA-seq quantification. *BMC Genomics*, 19(1):510, Jul 2018.

[46] Li Song, Sarven Sabunciyan, Guangyu Yang, and Liliana Florea. A multi-sample approach increases the accuracy of transcript assembly. *Nature Communications*, 10(1):5000, Nov 2019.

[47] Derek Aguiar, Li-Fang Cheng, Bianca Dumitrascu, Fantine Mordelet, Athma A. Pai, and Barbara E. Engelhardt. Bayesian nonparametric discovery of isoforms and individual specific quantification. *Nature Communications*, 9(1):1681, 2018.

[48] Katharina E. Hayer, Gregory R. Grant, Angel Pizarro, John B. Hogenesch, and Nicholas F. Lahens. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics*, 31(24):3938–3945, 09 2015.

[49] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40:1413, Nov 2008.

[50] Eddie Park, Zhicheng Pan, Zijun Zhang, Lan Lin, and Yi Xing. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics*, 102(1):11–26, 2018.

[51] Marina M. Scotti and Maurice S. Swanson. Rna mis-splicing in disease. *Nature Reviews Genetics*, 17:19, Nov 2015. Review Article.

[52] Anabella Srebrow and Alberto R. Kornblihtt. The connection between splicing and cancer. *Journal of Cell Science*, 119(13):2635–2641, 2006.

[53] Wei Vivian Li, Shan Li, Xin Tong, Ling Deng, Hubing Shi, and Jingyi Jessica Li. AIDE: annotation-assisted isoform discovery with high precision. *Genome Research*, 2019.

[54] Ringeling Francisca, Rojas, Chakraborty Shounak, Vissers Caroline, Reiman 3 Derek, Patel Akshay, M, Lee Ki-Heon, Hong Ari, Park Chan-Woo, Reska Tim, Gagneur Julien, Chang Hyeshik, Spletter Maria, Yoon Ki-Jun, Ming Guo-li, Song Hongjun, , and Canzar Stefan. Ladder-seq partitions rna-seq reads to improve transcriptome reconstruction and reveals a critical role of m6a as a regulator of alternative splicing in neural progenitor cells. (under review). 2021.

[55] Christian R. Eckmann, Christiane Rammelt, and Elmar Wahle. Control of poly(a) tail length. *Wiley Interdisciplinary Reviews: RNA*, 2(3):348–361, 2011.

[56] Ki-Jun Yoon, Francisca Rojas Ringeling, Caroline Vissers, Fadi Jacob, Michael Pokrass, Dennisse Jimenez-Cyrus, Yijing Su, Nam-Shik Kim, Yunhua Zhu, Lily Zheng, et al. Temporal control of mammalian cortical neurogenesis by m6a methylation. *Cell*, 171(4):877–889, 2017.

[57] Lichun Jiang, Felix Schlesinger, Carrie A. Davis, Yu Zhang, Renhua Li, Marc Salit, Thomas R. Gingeras, and Brian Oliver. Synthetic spike-in standards for rna-seq experiments. *Genome research*, 21(9):1543–1551, Sep 2011.

[58] Evan H. Hurowitz and Patrick O. Brown. Genome-wide analysis of mrna lengths in saccharomyces cerevisiae. *Genome Biology*, 5(1):R2, Dec 2003.

[59] Kelan Chen, Jiang Hu, Darcy L. Moore, Ruijie Liu, Sarah A. Kessans, Kelsey Breslin, Isabelle S. Lucet, Andrew Keniry, Huei San Leong, Clare L. Parish, Douglas J. Hilton, Richard J. L. F. Lemmers, Silvère M. van der Maarel, Peter E. Czabotar, Renwick C. J. Dobson, Matthew E. Ritchie, Graham F. Kay, James M. Murphy, and Marnie E. Blewitt. Genome-wide binding and mechanistic analyses of smchd1-mediated epigenetic regulation. *Proceedings of the National Academy of Sciences*, 112(27):E3535–E3544, 2015.

[60] Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D. Robinson, and Shobbir Hussain. A comprehensive examination of nanopore native rna sequencing for characterization of complex transcriptomes. *Nature Communications*, 10(1):3359, Jul 2019.

[61] Ales Varabyou, Steven L. Salzberg, and Mihaela Pertea. Effects of transcriptional noise on estimates of gene and transcript expression in rna sequencing experiments. *Genome Research*, 31(2):301–308, 2021.

[62] Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. *BMC Bioinformatics*, 12(1):323, 2011.

[63] Andrew E. Jaffe, Alyssa C. Frazee, Jeffrey T. Leek, and Ben Langmead. Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, 31(17):2778–2784, 04 2015.

[64] Thasso Griebel, Benedikt Zacher, Paolo Ribeca, Emanuele Raineri, Vincent Lacroix, Roderic Guigó, and Michael Sammeth. Modelling and simulating generic RNA-seq experiments with the flux simulator. *Nucleic Acids Research*, 40(20):10073–10083, September 2012.

[65] Tuuli Lappalainen, Michael Sammeth, Marc R. Friedländer, Peter A. C. 't Hoen, Jean Monlong, Manuel A. Rivas, Mar Gonzàlez-Porta, Natalja Kurbatova, Thasso Griebel, Pedro G. Ferreira, Matthias Barann, Thomas Wieland, Liliana Greger, Maarten van Iterson, Jonas Almlöf, Paolo Ribeca, Irina Pulyakhina, Daniela Esser, Thomas Giger, Andrew Tikhonov, Marc Sultan, Gabrielle Bertier, Daniel G. MacArthur, Monkol Lek, Esther Lizano, Henk P. J. Buermans, Ismael Padioleau, Thomas Schwarzmayr, Olof Karlberg,

Halit Ongen, Helena Kilpinen, Sergi Beltran, Marta Gut, Katja Kahlem, Vyacheslav Amstislavskiy, Oliver Stegle, Matti Pirinen, Stephen B. Montgomery, Peter Donnelly, Mark I. McCarthy, Paul Flicek, Tim M. Strom, The Geuvadis Consortium, Hans Lehrach, Stefan Schreiber, Ralf Sudbrak, Ángel Carracedo, Stylianos E. Antonarakis, Robert Häsler, Ann-Christine Syvänen, Gert-Jan van Ommen, Alvis Brazma, Thomas Meitinger, Philip Rosenstiel, Roderic Guigó, Ivo G. Gut, Xavier Estivill, and Emmanouil T. Dermitzakis. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501:506, Sep 2013. Article.

[66] Zheng Chang, Guojun Li, Juntao Liu, Yu Zhang, Cody Ashby, Deli Liu, Carole L. Cramer, and Xiuzhen Huang. Bridger: a new framework for de novo transcriptome assembly using rna-seq data. *Genome Biology*, 16(1):30, Feb 2015.

[67] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 01 2010.

[68] Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek, and Steven L Salzberg. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, 11(9):1650, 2016.

[69] W. James Kent. Blat–the blast-like alignment tool. *Genome research*, 12(4):656–664, Apr 2002.

[70] Juntao Liu, Ting Yu, Zengchao Mu, and Guojun Li. TransLiG: a de novo transcriptome assembler that uses line graph iteration. *Genome Biology*, 20(1):81, Apr 2019.

[71] Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson, and Colin N. Dewey. RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, 26(4):493–500, 12 2009.

[72] Adam Roberts and Lior Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature methods*, 10(1):71–73, 1 2013.

[73] Pierre-Luc Germain, Alessandro Vitriolo, Antonio Adamo, Pasquale Laise, Vivek Das, and Giuseppe Testa. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. *Nucleic Acids Research*, 44(11):5054–5067, Jun 2016.

[74] Steffen Heber, Max Alekseyev, Sing-Hoi Sze, Haixu Tang, and Pavel A. Pevzner. Splicing graphs and EST assembly problem. *Bioinformatics*, 18:S181–S188, 07 2002.

[75] Lior Pachter. Models for transcript quantification from RNA-seq. *arXiv*, 2011.

[76] Sayed Mohammad Ebrahim Sahraeian, Marghoob Mohiyuddin, Robert Sebra, Hagen Tilgner, Pegah T. Afshar, Kin Fai Au, Narges Bani Asadi, Mark B. Gerstein, Wing Hung Wong, Michael P. Snyder, Eric Schadt, and Hugo Y. K. Lam. Gaining comprehensive biological insight into the transcriptome

by performing a broad-spectrum rna-seq analysis. *Nature Communications*, 8, 12 2017.

[77] Zheng Chang, Zhenjia Wang, and Guojun Li. The impacts of read length and transcriptome complexity for de novo assembly: A simulation study. *PLOS ONE*, 9(4):1–8, 04 2014.

[78] Kristoffer Vitting-Seerup and Albin Sandelin. IsoformSwitchAnalyzeR: analysis of changes in genome-wide patterns of alternative splicing and its functional consequences. *Bioinformatics*, 35(21):4469–4471, April 2019.

[79] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Res*, 22(10):2008–2017, Oct 2012.

[80] Hyun Jung Park, Liguo Wang, Shengqin Wang, Surendra Dasari, Jean-Pierre Kocher, and Wei Li. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Research*, 41(6):e74–e74, 01 2013.

[81] Robert D Finn, Alex Bateman, Jody Clements, Penelope Coggill, Ruth Y Eberhardt, Sean R Eddy, Andreas Heger, Kirstie Hetherington, Liisa Holm, Jaina Mistry, et al. Pfam: the protein families database. *Nucleic acids research*, 42(D1):D222–D230, 2014.

[82] Yang Wang, Yue Li, Minghui Yue, Jun Wang, Sandeep Kumar, Robert J Wechsler-Reya, Zhaolei Zhang, Yuya Ogawa, Manolis Kellis, Gregg Duester, et al. N 6-methyladenosine rna modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nature neuroscience*, 21(2):195, 2018.

[83] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, et al. Model-based analysis of chip-seq (macs). *Genome biology*, 9(9):1–9, 2008.

[84] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 01 2010.

[85] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, May 2010.

[86] Adrian Alexa, Jörg Rahnenführer, and Thomas Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 04 2006.

[87] Kristoffer Vitting-Seerup and Albin Sandelin. The landscape of isoform switches in human cancers. *Molecular Cancer Research*, 15(9):1206–1220, 2017.

[88] Pedro J Batista, Benoit Molinie, Jinkai Wang, Kun Qu, Jiajing Zhang, Lingjie Li, Donna M Bouley, Ernesto Lujan, Bahareh Haddad, Kaveh Daneshvar, et al. m6a rna modification controls cell fate transition in mammalian embryonic stem cells. *Cell stem cell*, 15(6):707–719, 2014.

[89] Shengdong Ke, Endalkachew A Alemu, Claudia Mertens, Emily Conn Gantman, John J Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Scharff, Michael J Moore, Christopher Y Park, et al. A majority of m6a residues are in the last exons, allowing the potential for 3' utr regulation. *Genes & development*, 29(19):2037–2053, 2015.

[90] Takayoshi Yamauchi, Masaaki Nishiyama, Toshiro Moroishi, Atsuki Kawamura, and Keiichi I Nakayama. Fbxl5 inactivation in mouse brain induces aberrant proliferation of neural stem progenitor cells. *Molecular and cellular biology*, 37(8), 2017.

[91] Kazuya Kuboyama, Akihiro Fujikawa, Ryoko Suzuki, and Masaharu Noda. Inactivation of protein tyrosine phosphatase receptor type z by pleiotrophin promotes remyelination through activation of differentiation of oligodendrocyte precursor cells. *Journal of Neuroscience*, 35(35):12162–12171, 2015.

[92] Laura Conforti, Carlotta Dell'Agnello, Novella Calvaresi, Massimo Tortarolo, Andrea Giorgini, Michael P Coleman, and Caterina Bendotti. Kif1b$\beta$ isoform is enriched in motor neurons but does not change in a mouse model of amyotrophic lateral sclerosis. *Journal of neuroscience research*, 71(5):732–739, 2003.

[93] Tatsuaki Kurosaki, Maximilian W Popp, and Lynne E Maquat. Quality and quantity control of gene expression by nonsense-mediated mrna decay. *Nature reviews Molecular cell biology*, 20(7):406–420, 2019.

[94] Steve Lianoglou, Vidur Garg, Julie L Yang, Christina S Leslie, and Christine Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, 27(21):2380–2396, 2013.

[95] Dafni A Glinos, Garrett Garborcauskas, Paul Hoffman, Nava Ehsan, Lihua Jiang, Alper Gokden, Xiaoguang Dai, Francois Aguet, Kathleen L. Brown, Kiran Garimella, Tera Bowers, Maura Costello, Kristin Ardlie, Ruiqi Jian, Nathan R Tucker, Patrick T Ellinor, Eoghan D Harrington, Hua Tang, Michael Snyder, Sissel Juul, Pejman Mohammadi, Daniel G MacArthur, Tuuli Lappalainen, and Beryl Cummings. Transcriptome variation in human tissues revealed by long-read sequencing. *bioRxiv*, 2021.

[96] Xueyi Dong, Luyi Tian, Quentin Gouil, Hasaru Kariyawasam, Shian Su, Ricardo De Paoli-Iseppi, Yair David Joseph Prawer, Michael B Clark, Kelsey Breslin, Megan Iminitoff, Marnie E Blewitt, Charity W Law, and Matthew E Ritchie. The long and the short of it: unlocking nanopore long-read RNA sequencing data with short-read differential expression analysis tools. *NAR Genomics and Bioinformatics*, 3(2), 04 2021. lqab028.

[97] Vincent Lacroix, Michael Sammeth, Roderic Guigo, and Anne Bergeron. Exact transcriptome reconstruction from short sequence reads. In *International Workshop on Algorithms in Bioinformatics*, pages 50–63. Springer, 2008.

[98] Yan Huang, Yin Hu, Corbin D. Jones, James N. MacLeod, Derek Y. Chiang, Yufeng Liu, Jan F. Prins, and Jinze Liu. A robust method for transcript quantification with rna-seq data. *Journal of Computational Biology*, 20(3):167–187, 2013. PMID: 23461570.

[99] Johannes Köster and Sven Rahmann. Snakemake - a scalable bioinformatics workflow engine. *Bioinformatics*, 28(19):2520–2522, 08 2012.

[100] Margaret M. DeAngelis, David G. Wang, and Trevor L. Hawkins. Solid-phase reversible immobilization for the isolation of PCR products. *Nucleic Acids Research*, 23(22):4742–4743, 01 1995.

[101] Krzysztof Sobczak and Wlodzimierz J. Krzyzosiak. RNA structure analysis assisted by capillary electrophoresis. *Nucleic Acids Research*, 30(22):e124–e124, 11 2002.

[102] Arezou Azarani and Karl H. Hecker. RNA analysis by ion-pair reversed-phase high performance liquid chromatography. *Nucleic Acids Research*, 29(2):e7–e7, 01 2001.

[103] Yuancong Wang, Jinyan Xu, Min Ge, Lihua Ning, Mengmei Hu, and Han Zhao. High-resolution profile of transcriptomes reveals a role of alternative splicing for modulating response to nitrogen in maize. *BMC Genomics*, 21(1):353, May 2020.

[104] Runsheng Li, Xiaoliang Ren, Qiutao Ding, Yu Bi, Dongying Xie, and Zhongying Zhao. Direct full-length rna sequencing reveals unexpected transcriptome complexity during caenorhabditis elegans development. *Genome Research*, 2020.

[105] Aishwarya G Jacob and Christopher WJ Smith. Intron retention as a component of regulated gene expression programs. *Human genetics*, 136(9):1043–1057, 2017.

[106] Ulrich Braunschweig, Nuno L Barbosa-Morais, Qun Pan, Emil N Nachman, Babak Alipanahi, Thomas Gonatopoulos-Pournatzis, Brendan Frey, Manuel Irimia, and Benjamin J Blencowe. Widespread intron retention in mammals functionally tunes transcriptomes. *Genome research*, 24(11):1774–1786, 2014.