TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

TOOLS AND STRATEGIES FOR RNA-SEQUENCING DATA ANALYSIS

Arfa Mehmood

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1719 | MEDICA – ODONTOLOGICA | TURKU 2023

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

TOOLS AND STRATEGIES FOR RNA-SEQUENCING DATA ANALYSIS

Arfa Mehmood

TURUN YLIOPISTON JULKAISUJA – ANNALES UNIVERSITATIS TURKUENSIS
SARJA – SER. D OSA – TOM. 1719 | MEDICA – ODONTOLOGICA | TURKU 2023

University of Turku

Faculty of Medicine
Institute of Biomedicine
Physiology
Drug Research Doctoral Programme (DRDP)

Supervised by

Professor Laura Elo
Turku Bioscience Centre
University of Turku and Åbo Akademi
Turku, Finland

Professor Matti Poutanen
Institute of Biomedicine
University of Turku
Turku, Finland

Dr. Asta Laiho
Turku Bioscience Centre
University of Turku and Åbo Akademi
Turku, Finland

Reviewed by

Docent Päivi Saavalainen
University Researcher, Translational Immunity Research Program
University of Helsinki
Helsinki, Finland

Associate Professor Valerio Izzi
Faculty of Biochemistry and Molecular Medicine
University of Oulu
Oulu, Finland

Opponent

Professor Inge Jonassen
Head of Department of Informatics
University of Bergen
Bergen, Norway

ISBN 978-951-29-9317-8 (PRINT)
ISBN 978-951-29-9318-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)
Painosalama, Turku, Finland 2023

To My Parents

UNIVERSITY OF TURKU
Faculty of Medicine
Institute of Biomedicine
Physiology
ARFA MEHMOOD: Tools and strategies for RNA-sequencing data analysis
Doctoral Dissertation, 98 pp.
Drug Research Doctoral Programme (DRDP)
August 2023

ABSTRACT

RNA-Sequencing (RNA-seq) has enabled the in-depth study of the transcriptome, becoming the primary research method in the field of molecular biology. The typical aim of RNA-seq is to quantify and detect differentially expressed (DE) and differentially spliced (DS) genes. Numerous methodologies and tools have been developed in recent years to assist in analyzing RNA-seq data. However, it is difficult for researchers to decide which methods or strategies they should adopt to optimize the analysis of their datasets.

In this Thesis, in Study I, we applied the gene-level DE analysis approach to detect the androgen-regulated genes between cancerous and benign samples in 48 primary prostate cancer patients. Combined with other measurements from the same samples, our analysis indicated that patients having TMPRSS-ERG gene fusion had distinct intratumoral androgen profiles compared to TMPRSS-ERG negative tumors. However, the DE can remain undetected when the expression varies across the gene due to reasons such as alternative splicing. Hence, to account for this problem, an alternate analysis approach has been suggested in which the statistical testing of lower feature levels (e.g. transcripts, transcript compatibility counts, or exons) is performed initially, followed by aggregating the results to the gene level. In Study II, we tested this alternate approach on these lower features and compared the results to those from the conventional gene-level approach. In the alternate approach, two methods (Lancaster method and empirical brown method (ebm)) were tested for aggregating the feature-level results to gene-level results. Our results suggest that the exon-level estimates improve the detection of the DE genes when the ebm method is used for aggregating the results. Accordingly, R/Bioconductor package EBSEA was developed using the winning approach.

RNA-seq data can also be used to find DS events between conditions. However, the detection of DS is more challenging than the detection of DE. In Study III, a comprehensive comparison of ten DS tools was performed. We concluded that exon-based and event-based methods (rMATS and MAJIQ) performed overall best across the different evaluation metrics considered. Furthermore, we observed overall low concordance between the results reported by the different tools, making it recommendable to use more than one tool when performing DS analysis, and to concentrate on the overlapping results.

KEYWORDS: Differential gene expression, Alternative splicing, Differential splicing, Splicing events, RNA-seq

TURUN YLIOPISTO
Lääketieteellinen tiedekunta
Biolääketieteen laitos
Fysiologia
ARFA MEHMOOD: Työkaluja ja strategioita RNA-sekvensointidatan analyysiin
Väitöskirja, 98 s.
Lääketutkimuksen tohtoriohjelma (DRDP)
Elokuu 2023

TIIVISTELMÄ

RNA-sekvensointi (RNA-seq) on mahdollistanut transkriptomin yksityiskohtaisen tarkastelun ja siitä on tullut hyvin suosittu työkalu molekyylibiologian tutkimuksessa. RNA-sekvensointitutkimusten tyypillinen tarkoitus on selvittää näyteryhmien välillä eriävästi ilmentyviä ja silmukoituvia geenejä. RNA-sekvensointidatojen analyysiin on kehitetty runsaasti työkaluja ja usein on haastavaa valita näiden joukosta optimaaliset välineet tietyn aineiston käsittelyyn.

Tässä väitöstyössä osajulkaisussa I tunnistettiin androgeenihormonien säätelemiä eriävästi ilmentyviä geenejä syöpäkudoksen ja terveen kudoksen välillä 48 eturauhassyöpäpotilaalla. Kun nämä tulokset yhdistettiin muihin samojen potilaiden käytettävissä oleviin mittausarvoihin, havaittiin, että TMPRSS-ERG-geenifuusion omaavien potilaiden syöpäkudoksen androgeenihormonigeenien ilmentymisprofiili poikkesi verrattuna niihin potilaisiin, joilta ei löytynyt vastaavaa geenifuusiota. On kuitenkin mahdollista, että tällä lähestymistavalla eriävä ilmentyminen jää joidenkin geenien osalta havaitsematta, jos ilmentymistaso vaihtelee geenin eri osissa, esimerkiksi vaihtoehtoisen silmukoinnin vaikutuksen vuoksi. Ratkaisuksi tähän on esitetty uudenlaista lähestymistapaa, jossa tilastollinen testaus näyteryhmien välillä suoritetaan geenin rakenteen osalta hienojakoisemmalla tasolla (esimerkiksi transkriptien, transkriptiyhteensopivien mittausyksiköiden tai eksonien tasolla) ja vasta näin saadut osatulokset yhdistetään geenitason kokonaistulokseksi. Julkaisussa II verrattiin tätä lähestymistapaa perinteiseen geenitason analyysiin testaamalla kahta eri menetelmää tulosten yhdistämiseen takaisin geenitasolle: 1) Lancaster-menetelmää ja 2) empiiristä Brown-menetelmää (ebm). Tulosten perusteella eksonitason mittausarvojen käyttö yhdistettynä ebm-menetelmään paransi eriävästi ilmentyvien geenien tunnistusta. Tämä lähestymistapa on sisällytetty väitöstyössä kehitettyyn geenien eriävää ilmentymistä analysoivaan R/Bioconductor -analyysipakettiin EBSEA.

RNA-sekvensointidataa voidaan käyttää myös eriävien silmukointitapahtumien tunnistamiseen näyteryhmien välillä. Tämä on kuitenkin haastavampaa kuin geenien eriävän ilmentymisen analyysi. Julkaisussa III vertailtiin kymmentä eriävien silmukointitapahtumien tunnistamiseen kehitettyä työkalua. Näistä työkaluista eksoniperustaiset ja silmukointitapahtumaperustaiset työkalut (erityisesti rMATS ja MAJIQ) tuottivat parhaat kokonaistulokset käytetyillä vertailukriteereillä. Työkalujen tuottamien tulosten välillä havaittiin kuitenkin merkittäviä eroja, minkä johdosta tulosten jatkotarkastelussa on hyödyllistä keskittyä niihin tuloksiin, jotka ovat löydettävissä useammalla kuin yhdellä työkalulla.

AVAINSANAT: Eriävä geenien ilmentyminen, vahtoehtoinen silmukointi, eriävä silmukointi, silmukointitapahtumat, RNA-sekvensointi.

Table of Contents

Abbreviations

A	Adenine
A3SS	Alternative 3' Splice Site
A5SS	Alternative 5' Splice Site
AF	Alternative First Exon
AL	Alternative Last Exon
C	Cytosine
cDNA	complementary Deoxyribonucleic Acid
DE	Differentially Expressed
DHT	Dihydrotestosterone
DNA	Deoxyribonucleic Acid
dNTP	Deoxynucleoside triphosphate
DS	Differentially Spliced / Differential Splicing
DTU	Differential Transcript Usage
ebm	Empirical Brown's Method
ER	Exclusion Reads
FDR	False Discovery Rate
FP	False Positives
G	Guanine
GEO	Gene Expression Omnibus
HCa	Hepatocellular Carcinoma
HTS	High Throughput Sequencing
HVS	Human Validated Dataset
IR	Inclusion Reads
log2 FC	Logarithm 2-Fold Change
LSV	Local Splicing Variation
mRNA	messenger Ribonucleic Acid
MAQC	Microarray Quality Control
MVS	Mouse Validated Dataset
MXE	Mutually Exclusive Events
pAUC	partial Area Under the Curve
PCa	Prostate Cancer Dataset

PCR	Polymerase Chain Reaction
PSI	Percentage Spliced In
RI	Retained Intron
RLE	Relative Log Expression
RNA	Ribonucleic Acid
RNA-seq	RNA Sequencing
ROTS	Reproducibility-Optimized Test Statistics
RSEM	RNA-seq by Expectation-Maximization
SE	Skipped Exon
SNP	Small Nucleotide Polymorphism
SRA	Sequence Read Archive
TCCs	Transcript Compatibility Counts
TFs	Transcription Factors
TMM	Trimmed Mean of M Values
T	Thymine
U	Uracil

List of Original Publications

This dissertation is based on the following original publications, which are referred to in the text by their Roman numerals:

I Knuuttila M, **Mehmood A**, Mäki-Jouppila J, Ryberg H, Taimen P, Knaapila J, Ettala O, Boström PJ, Ohlsson C, Venäläinen MS, Laiho A, Elo LL, Sipilä P, Mäkelä SI, Poutanen M. Intratumoral androgen levels are linked to TMPRSS2-ERG fusion in prostate cancer. *Endocrine-Related Cancer*, 2018; 25(9): 807-819.

II **Mehmood A**, Laiho A, Elo LL. Exon-level estimates improve the detection of differentially expressed genes using RNA-seq studies. *RNA Biology,* 2020; 18(11):1739-1746.

III **Mehmood A**, Laiho A, Venäläinen MS, McGlinchey AJ, Wang N, Elo LL. Systematic evaluation of differential splicing tools for RNA-seq studies. *Briefings in Bioinformatics*, 2019; 21(6):2052-2065.

1 Introduction

Since the emergence of RNA-sequencing (RNA-seq), the method has been extensively used to study the transcriptome at an unprecedented level. Transcriptome is the complete set of transcripts in a cell, expressed in a particular physiological condition or developmental stage. Compared to Sanger sequencing and microarrays, RNA-seq allows for transcript analysis at higher accuracy, enables single base-level resolution, and provides an extended dynamic range of expression and lower background signal (Garber et al., 2011; Z. Wang et al., 2009). The technology provides quantification for all genes simultaneously (Mortazavi et al., 2008) compared to older techniques such as measuring the expression of a single gene at a time using the polymerase chain reaction (PCR). The typical aim of RNA-seq is to find the differentially expressed (DE) genes between different physiological conditions, developmental stages, tissue types, or normal and diseased samples (Han et al., 2015). Besides finding DE genes, other possible downstream analyses include identifying and quantifying spliced genes, detecting differentially spliced (DS) genes between conditions, finding gene fusions, analyzing allele-specific expression, and detecting variants (SNPs) (Conesa et al., 2016; Han et al., 2015). These analyses help to identify and interpret the functional elements of the genome (Z. Wang et al., 2009). In the past decade, numerous strategies, methods, and tools emerged and have been refined for analyzing the RNA-seq data for different purposes, but still, consensus has not been reached yet on the optimal pipelines (Conesa et al., 2016).

RNA-seq analysis involves sequencing followed by bioinformatics analysis. Sequencing generates sequence reads from the samples and the sample preparation step involves messenger (mRNA) extraction, cDNA library construction, and sequencing cluster generation (Han et al., 2015). For bioinformatic analysis, the first step is to determine the quality of the sequencing reads before they are passed through different tools for pre-processing and downstream analyses. For identifying DE genes, the reads are mapped to the reference genome or transcriptome. The gene-level read counts are typically quantified using the exon-union method, which sums the exon read counts across each gene. The gene counts are then statistically tested to find DE genes between the different conditions. This approach to finding DE genes will be referred to as the conventional gene-level analysis approach in this

Thesis. We applied this widely used conventional gene-analysis approach in Study I to study the expression of androgen-regulated genes in tissue specimens of primary prostate cancer. In prostate cancer, the TMPRSS-ERG fusion gene is a commonly found gene fusion and may be causing the activation of the testosterone-independent dihydrotestosterone (DHT) biosynthesis via the alternative pathway. We further studied RNA-seq and androgen concentration data to detect the differences in expression of androgen target genes in TMPRSS-ERG positive (TMPRSS-ERG+) tumors compared to TMPRSS-ERG negative (TMPRSS-ERG-) tumors.

The computational methodologies for gene-level RNA-seq analysis are being actively refined and several independent studies (Kanitz et al., 2015; Laiho & Elo, 2014; Yi et al., 2018) have been published, suggesting that using the lower feature-level data increases the accuracy and power of DE analysis compared to the conventional gene-level approach. With these approaches, the statistical testing is performed at the lower feature level (e.g. transcript, transcript compatibility count (TCC) or exon), followed by aggregating the feature level p-values to gene-level p-values (Laiho & Elo, 2014; Yi et al., 2018). In Study II, we systematically tested this approach by using the different lower feature levels. We also investigated the effect of considering the dependence of lower-level features during statistical testing, which to our knowledge, has not been systematically done before.

Another vital aspect of RNA-seq analysis is to find DS changes between different sample groups. Various approaches of DS have been published and they can be categorized into isoform-based and count-based methods (Chen, 2013; Hooper, 2014; Liu et al., 2014; Trapnell, Roberts, et al., 2012). The count-based methods are further classified into event-based and exon-based methods. Isoform-based methods identify DS by statistically testing the relative abundances of the transcripts. In contrast, the count-based approach uses either the exon counts or quantified splicing events between the conditions. We performed an independent comparative analysis of ten DS tools on four real RNA-seq datasets (Study III).

In conclusion, this work focuses on the enhanced analysis of RNA-seq data, providing further insights and evaluating the performance of the different approaches and methodologies developed for detecting DE and DS genes. The findings of the included studies will help researchers to carry out more optimal data analysis by choosing approaches or tools appropriate for their RNA-seq datasets and purpose. The EBSEA data analysis package, developed in Study II, uses the alternate DE analysis approach by performing the statistical testing at the exon level and aggregating these results to gene level using the ebm aggregation method. The package is available as a Bioconductor module and provides intuitive visualization options for the users. EBSEA can be applied for other -omic data analyses as well.

2 Review of the Literature

2.1 Gene Expression

The genetic information in a cell is stored in the deoxyribonucleic acid (DNA), packaged around the nucleosome, and organized into the chromatin structure. The genome size and the number of protein-coding genes and transcripts vary in different species. For example, human genome consists of around 3.0 billion base pairs with estimated 20,000 protein-coding genes and about 198,093 transcripts while the mouse genome has around 2.7 billion base pairs, 22,000 protein-coding genes, and 118,925 transcripts (Breschi et al., 2017). However, there is no direct correlation between genome size and the organism's complexity (Leslie A. Pray, 2008).

DNA is a double-stranded helix, which is held together by hydrogen bonds formed between the complementary bases: Adenine (A) with Thymine (T) and Cytosine (C) with Guanine (G). DNA contains sequence information of protein-coding genes and non-coding regulatory elements that help to regulate gene expression. The messenger RNA molecules (mRNA) transcribed from the DNA are further translated into proteins responsible for many cellular functions in a living organism. A vast effort has been committed to understanding the flow of information from DNA to protein. Dysfunction in gene expression can lead to different diseases such as developmental disorders, diabetes, cardiovascular diseases, and cancers (Lee & Young, 2013).

According to the central dogma of molecular biology, DNA is first transcribed into mRNA, which is used as a template to further translate it into protein (Fig. 1). However, it has later been learned that there are also other molecules transcribed from DNA, such as functional RNAs (Palazzo & Lee, 2015), retroviruses, and prions (Ryu, 2016). Modern high-throughput sequencing (HTS) technologies have enabled the study of both protein-coding and non-protein coding transcripts at unprecedented levels. HTS technologies have made it possible to sequence millions of DNA fragments simultaneously, providing comprehensive insight into the cell's genomic and transcriptomic landscape (Churko et al., 2013).

During gene expression, the DNA sequence is first transcribed into an mRNA molecule (Fig. 1) which is a linear polymer of the four different nucleotides linked together by a phosphodiester bond. In contrast to the DNA molecule, the mRNA

molecule is chemically unstable and consists of ribose as the sugar instead of deoxyribose present in the DNA. Another difference between DNA and RNA molecules is the presence of uracil base (U) in mRNA rather than T used in DNA. Further, mRNA molecule undergoes post-transcriptional modifications in the nucleus, including capping to add G base at the 5'-end of the mRNA, poly-A tailing to add multiple A bases at the 3' end of the mRNA, and splicing to remove the non-coding intron sequences between the exons. The modified mature mRNA is transported out of the nucleus and translated into protein, followed by post-translational modification.

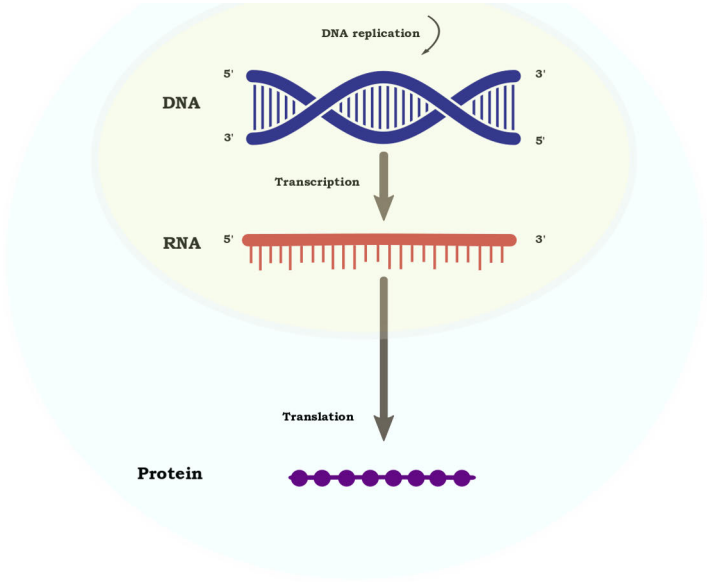

Figure 1. During gene expression the DNA sequence of the protein coding genes is first transcribed into an mRNA molecule and then further translated into protein.

2.2 RNA Splicing

RNA splicing is a vital modification of pre-mRNA to a mature mRNA. It is estimated that approximately 95% of the eukaryote genes undergo splicing (Pan et al., 2008) resulting in the expansion of the transcriptome and functional proteome vastly beyond the number of available genes. Splicing is a complex process in which the non-coding intervening sequences (introns) are excised, and protein-coding sequences (exons) are ligated together to form a mature mRNA. In fact, exons only form a small fraction of the pre-mRNA compared to the introns. The length of the introns may range from 10 to 100,000 bases, whereas the size of the exons is more

uniform, with humans having a median exon length of 120 bases (El Marabti & Younis, 2018).

Splicing can be divided into constitutive splicing and alternative splicing. In constitutive splicing, the introns are removed, and the exons are ligated in the order they appear in the pre-mRNA. In contrast, alternative splicing (AS) is a process that directs mRNA precursors to form different transcripts by selecting various splice sites, and thus, different combinations of exons (Fig. 2). These different transcripts are used to produce proteins that differ in their cellular function and the processes they participate.

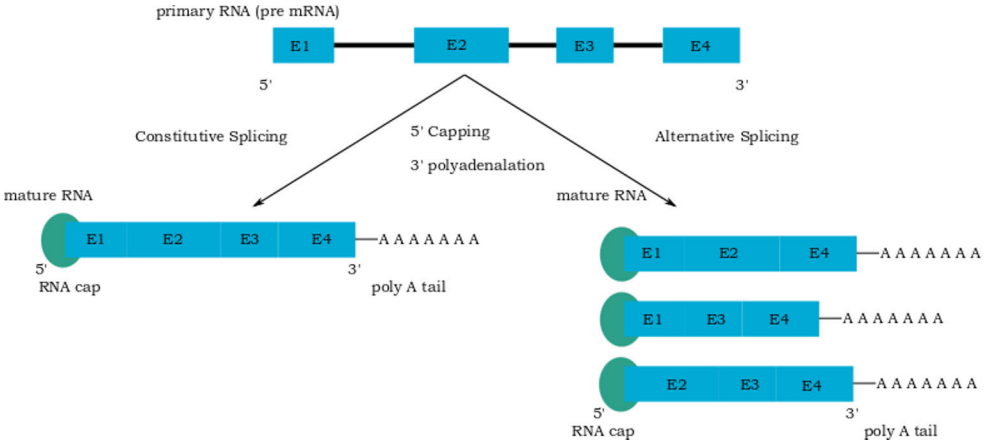

Figure 2. The primary mRNA undergoes post-transcriptional modifications, including RNA capping, polyadenylation, and splicing. The mRNA can be spliced into either constitutive or alternatively spliced mRNA. In constitutive splicing, the exons appear in the same sequence as in the pre-mRNA, whereas in alternative splicing, the exons appear in varying patterns in the resulting mature RNA.

2.2.1 Alternative Splicing Events

There are many different types of AS events which can be divided into four basic classes: 1) skipped exons (SE), 2) alternative 5' (donor) splice sites (A5SS), 3) alternative 3' (acceptor) splice sites (A3SS) and 4) retained introns (RI). Other less common classes are mutually exclusive events (MXE), alternative first exons (AF), and alternative last exons (AL) (Fig. 3). In higher eukaryotes, the SE event is the most prevalent splicing event constituting around 30 - 40 % of all splicing events (E. Kim et al., 2007, 2008). This is followed by A3SS (~ 18.4%), A5SS (~7.9 %), and RI (< 5%). In contrast, plants exhibit a high level of RI (~30%) and a low level of SE (< 5%) (E. Kim et al., 2008).

In SE, the exon and its flanking introns are excised out of the alternative transcript (Fig. 3). This splicing event can lead to various human diseases and is

considered a therapeutic target in Duchenne muscular dystrophy treatment (Aartsma-Rus & Van Ommen, 2007). In A5SS and A3SS, two or more splice sites are recognized either upstream or downstream of the exon. In RI, the intron is unspliced in the final transcript and is found to control the post-transcriptional expression of the gene. In humans, high levels of RI characterize cancer cells of all types except in breast cancer (Dvinge & Bradley, 2015). Another study showed that RI is a widespread regulatory mechanism, and it aids the transcriptome's functional tuning in mammals (Braunschweig et al., 2014). High levels of RI are found in transcripts expressed at relatively low level, negatively regulating cytoplasmic transcript levels (Braunschweig et al., 2014).

Figure 3. Four basic classes of splicing events: 1) skipped exons, 2) alternative 5' splice sites, 3) alternative 3' splice sites, and 4) retained introns. Other less common events include mutually exclusive events, alternative first exons and alternative last exons.

2.3 Gene Fusion

Gene fusion happens when the two independent genes or parts of them fuse to form a chimera due to DNA rearrangement. Different mechanisms, such as insertions, deletions, inversions, and translocations, can lead to these gene fusions (Fig. 4A). Besides this, continuous splicing of a gene or trans- or cis-splicing of pre-mRNA can lead to fusion transcripts (Fig. 4B). Gene fusion may result in a new protein with new functionality compared to its parental genes. Gene fusions were first discovered

in hematologic malignancies, and they have been found in several solid tumors (Stengel et al., 2018). HTS technology has recently enabled the efficient identification of gene fusions, and there are now approximately 10,000 known fusion genes (Latysheva & Babu, 2016). Nowadays fusion genes are used in many research areas, such as development of biomarkers and diagnostic and therapeutic agents (Gao et al., 2018; Parker & Zhang, 2013).

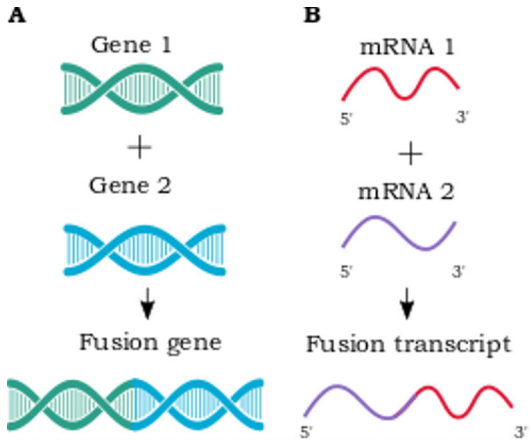

Figure 4. **A)** The rearrangement of DNA due to insertion, translocation, inversion, or deletion in two different genes leads to a gene fusion. **B)** The fusion of two mRNAs can lead to a fusion transcript.

2.4 RNA-Sequencing Technology

RNA-seq is a HTS approach widely used to study the transcriptome, the primary goal being to find DE genes between samples/groups/conditions. In addition, RNA-seq is also used to detect novel genes and isoforms, fusion genes, DS, and allele-specific expression (Conesa et al., 2016; E. T. Wang et al., 2008). The technology is under active development, and it offers many advantages over older technologies such as microarrays. RNA-seq has been shown to detect lowly expressed transcripts and does not require the use of predetermined interrogation sequences (Kukurba & Montgomery, 2015).

The different HTS platforms use varying protocols. In this Thesis, I will discuss the RNA-seq workflow of the most popular of them – Illumina sequencing technology. A typical Illumina RNA-seq library construction workflow includes the steps of RNA extraction, mRNA enrichment, RNA fragmentation, cDNA synthesis, adaptor ligation, and cluster generation including amplification, and sequencing (Fig. 5) (Stark et al., 2019). Many details must be considered before performing an RNA-seq experiment, depending on the study objective. These details include the choice of the number of biological replicates, the desired sequencing depth,

sequencing type (single- or paired-end) and read length. In single-end sequencing data, the reads are sequenced from one end of the sequence fragment, whereas in paired-end sequencing the reads are sequenced from both ends of the fragment, enabling alignment to reference genome with increased accuracy. The depth of the coverage is the measure of the average number of times that a specific genomic site (base) is sequenced. At higher coverage, more sequencing reads are produced and thus the resolution of the analysis is increased as each base is covered by a higher number of reads. With the goal to detect DE, the read length of 50-75 base pairs, single-end sequencing, and 20 million sequencing reads per sample are recommended. However, for detecting alternative splicing, allele-specific expression, or sequence variants, it is recommended to use a read length greater than 75 base pairs, paired-end sequencing, and 40-100 million reads per sample.

2.4.1 RNA Extraction

To sequence the samples, their RNA first needs to be extracted (Fig. 5A). RNA is typically extracted either by phenol-Chloroform (e.g. TRIzol) or silica gel-based column method (e.g. Qiagen) (Sultan et al., 2014). It is to note that DNA contamination can negatively influence the analysis results. Thus, it is essential to check the quality of the extracted RNA by measuring the RNA integrity (Schroeder et al., 2006). RNA quality can substantially impact the success of sequencing experiments. In some cases, high-quality samples are not available, for example when samples have been stored in paraffin or when human autopsy samples are used where RNA is typically partially degraded. For these samples, special sequencing library preparation protocols are available. The level of RNA degradation can also be taken into consideration during data analysis (Kukurba & Montgomery, 2015).

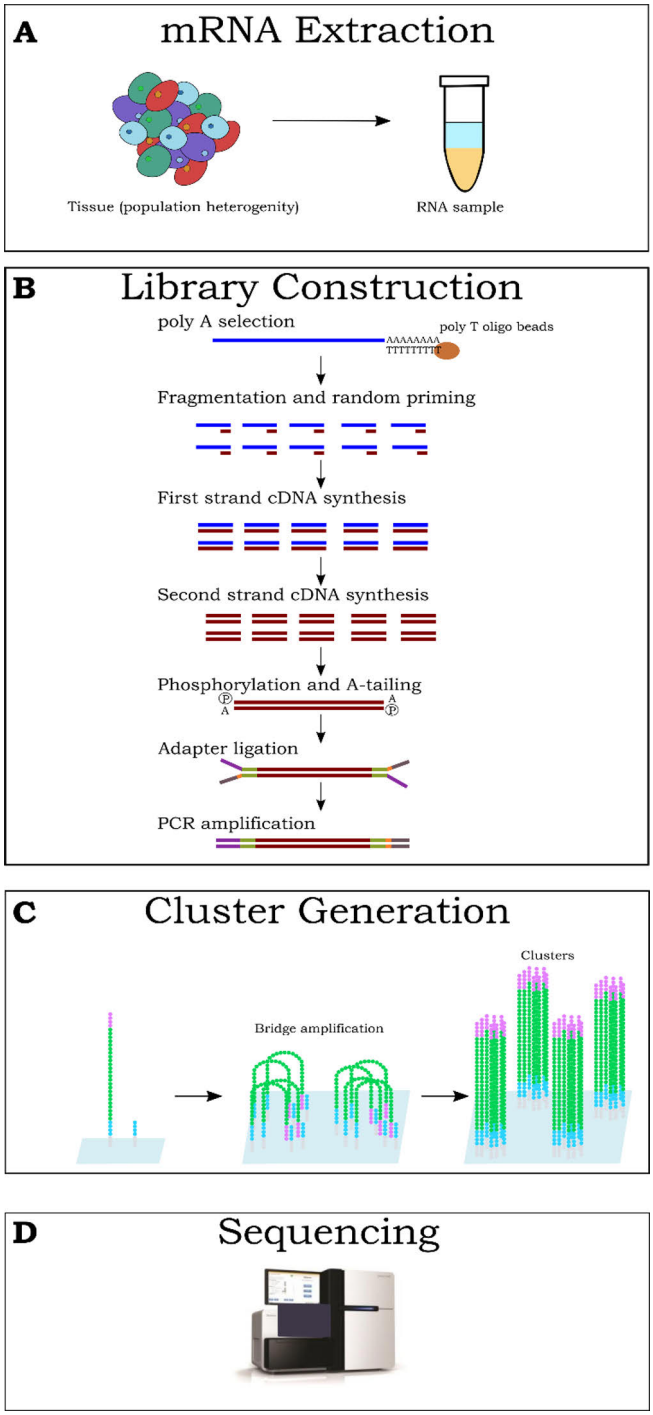

Figure 5. RNA-Sequencing involves A) mRNA extraction, B) library construction, C) cluster generation, and D) sequencing.

2.4.2 Library Construction

When the RNA has been extracted from the samples, the next step in the sequencing workflow is library construction. The protocol varies depending on whether total RNA or mRNA sequencing is performed. In total RNA-seq, only highly abundant ribosomal RNA is depleted from the samples in a separate step. Typically, complementary sequences available in commercial kits such as RiboMinus or RiboZero are used (Petrova et al., 2017). With only ribosomal sequences depleted, many different RNA species including pre-mRNA, mRNA, transfer RNA, microRNA, and long non-coding RNAs are preserved in the sample. However, the focus of RNA-seq typically is to sequence only the mRNA coding regions. In this case, prior to library construction, poly-(A) containing mRNA is enriched using poly-T oligos attached to magnetic beads (Rio et al., 2010).

Next, the long RNA molecules are typically fragmented to a length ranging from 200 - 500 base pairs via RNA hydrolysis or nebulization. After this, RNA molecules are converted to cDNA molecules. Other possible library preparation methods include amplification of cDNA SMART-PCR to generate full-length cDNA from RNA samples followed by Nextera tagmentation in which the DNA molecules are fragmented and tagged for preparing DNA libraries for Illumina sequencing. This approach allows the library preparation from samples with a low amount of RNA.

Initially, Illumina-based RNA-seq used hexamer priming or short sequences of Ts complementary to the poly-A tails for reverse transcribing the mRNA to cDNA. In this process, RNA is removed after reverse transcription, and the second strand is synthesized to form double-stranded cDNA. However, this way the expressed DNA strand information is lost. To avoid the loss of strand information, a dUTP method is widely used nowadays, incorporating deoxy-UTP during the synthesis of the second cDNA strand that allows subsequent destruction of the uridine-containing strand (Parkhomchuk et al., 2009). Other alternative approaches of strand-specific RNA-seq include 3' end or 5' end-based library preparation protocols which selectively label either end of the RNA strand. This approach is cost-efficient for quick DE analysis of large number of samples - however, it is not optimal for alternative splicing analysis where full-length RNA needs to be analyzed.

The cDNA fragments are then ligated with adapter sequences. The adapters have different functional elements known as motifs, such as sequences required for attachment to the flow cell oligos and clonal amplification, a sequence for priming, and a barcode sequence for multiplexing (L. Wang et al., 2011). In multiplexing, the cDNA is barcoded which allows combining multiple samples into a single sequencing lane which reduces sequencing costs and multiplexes up to 96 samples in one lane (Hou et al., 2015). Furthermore, cDNA fragments of optimal size for sequencing (typically 300-500 base pairs) are selected using gel electrophoresis (Fig. 5B).

2.4.3 Cluster Generation

In cluster generation, the cDNA fragments are isothermally amplified to form clusters on the flow cell (J. Kim & Easley, 2011). A flow cell is a thick glass slide with several lanes, which are covered in oligos that are complementary to library adapters on the cDNA fragments. The single-strand cDNA fragments are passed over the flow cell containing two types of oligos attached to its cell surface. The fragment hybridizes with oligos, and the polymerase molecule complex moves along the strand to produce the complementary strand. The double strand is then denatured, and the original strand is washed away. The remaining strand folds over and hybridizes with the second type of oligo, which is extended by a polymerase to form a double bridge. The strands are then again denatured to form single strands. This process is called bridge amplification, resulting in thousands of sequence clusters all over the flow cell (Fig. 5C). After cluster generation, the sequence fragments are ready for sequencing (Fig. 5D).

2.4.4 Sequencing

Illumina applies an ensemble-based sequencing by synthesis approach in which tens of millions of sequence fragment clusters are sequenced in parallel. The sequencing primers are added so that the fragments start to get reversibly transcribed. In each cycle of sequencing, labeled deoxynucleoside triphosphates (dNTPs) are added. After the addition of each dNTP, the fluorescent dye is imaged to identify the added base and enzymatically cleaved to allow incorporation of the next dNTP. The process is repeated number of times depending on the desired read length. The sequencing produces a set of images converted to readable sequences using base-calling software. Output is generated in fastq file format which also includes quality values for each sequenced base.

2.5 Bioinformatic Analysis

The bioinformatics analysis of RNA-seq data to detect DE genes starts with pre-processing and is then followed by downstream analysis. The pre-processing of RNA-seq data typically involves quality control, read alignment, and quantification steps while the most important downstream analysis steps contain normalization and DE testing. This conventional gene-level approach is summarized in this section and in Fig. 6.

2.5.1 Pre-processing of RNA-seq Data

2.5.1.1 Quality Control

The raw data in the format of fastq files contain the sequence reads and the associated base quality scores. FASTQC tool (Andrews, 2010), developed by Babraham Institute, is commonly used to analyse the quality of the reads. With the tool, the quality of the reads is assessed regarding different metrics such as average quality, GC content, PCR duplicates, duplicated reads and the presence of sequencing adapters.

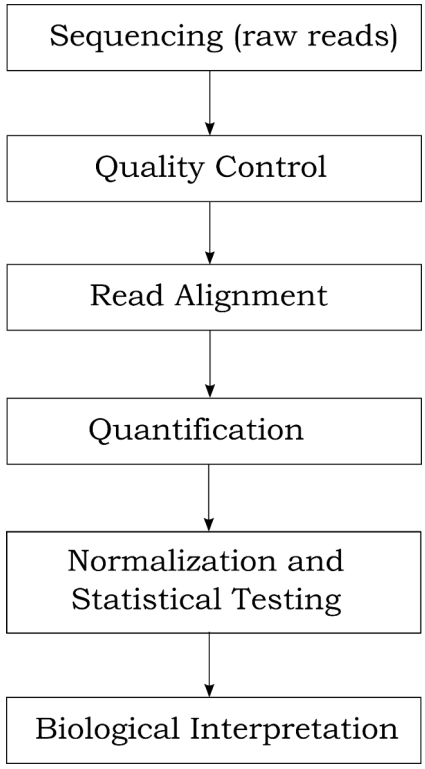

Figure 6. In conventional gene-level analysis, the reads are aligned after checking the quality of the reads. The gene-level read counts are quantified, normalized, and statistical testing between sample groups is performed.

2.5.1.2 Read Alignment

With read alignment the short sequencing reads are mapped to their genomic location of origin according to the reference genome used (Conesa et al., 2016). Many different alignment tools have been developed in the past years such as Tophat2 (D.

Kim et al., 2013), STAR (Dobin et al., 2013) , HISAT2 (D. Kim et al., 2019), RSEM (B. Li & Dewey, 2011) and Kallisto (Bray et al., 2016). These tools align the reads using different strategies such as splice-aware alignment and pseudo alignment. The splice-aware aligners such as Tophat2, STAR, and HISAT2 align the exonic reads to the reference genome. The spliced-read mapping is challenging due to the need to correctly determine the exon-intron boundaries where one part of the read will map to one exon and the other part to another exon. Pseudo aligners such as RSEM and Kallisto quantify transcript abundances by determining which transcripts the reads are compatible with rather than first aligning reads to the genome or transcript. In our studies, we have used STAR, RSEM and Kallisto.

2.5.1.3 Quantification

Quantification is the analysis step where the gene or transcript expression levels are estimated. The most widely used tools to quantify gene-level expression signals are HTSeq (Anders et al., 2015) and featureCount (Liao et al., 2014). These tools aggregate the raw counts of mapped reads on the genomic features of interest, such as genes. With HTSeq and featureCount the gene-level expression values are produced by counting the number of reads that overlap any of the gene's exons. Multimapping reads, which map to several different genome locations are typically excluded as for them the actual target gene cannot be confirmed (Liao et al., 2014). However, these tools do not specifically consider the different isoforms (transcripts) of the genes. Transcript-level expression abundance values can be estimated using sophisticated algorithms such as Cufflink (Trapnell et al., 2010), RSEM (B. Li & Dewey, 2011), Kallisto (Bray et al., 2016), Salmon (Patro et al., 2017) or BitSeq (Glaus et al., 2012). In this thesis, we have used RSEM and Kallisto to produce transcript-level expression abundance values.

In addition to gene counts, transcript counts, and transcript compatibility counts (TCCs) can also be calculated using for example Kallisto. TCCs are the number of reads compatible with the same set of transcripts.

2.5.2 Differential Gene Expression Analysis

Typical aim of RNA-seq analysis is to find DE genes or transcripts between the different sample groups/conditions/tissues. The DE genes are considered either upregulated in which case the gene expression is increased or downregulated in which case the expression is decreased in the treatment sample group compared to the normal/control sample group.

2.5.2.1 Normalization

The raw read counts are affected by different factors such as between-sample differences in library composition (sequencing depth) and within-sample differences in gene length. Hence, the counts are normalized to correct for systematic technical biases. To remove these biases, reads per kilobase per million mapped reads (RPKM) (Mortazavi et al., 2008), fragment per kilobase per million mapped reads (FPKM), and transcript per million (TPM) (Wagner et al., 2012) values were introduced. RPKM and FPKM are analogous, where RPKM is used for single-end sequencing data, and FPKM is used for paired-end sequencing data. RPKM re-scales read counts to correct the sequencing depth bias and gene length differences. RPKM is calculated as:

$$\text{RPKM} = 10^9 * \frac{Reads\ mapped\ to\ the\ transcript}{Total\ reads * Transcript\ length}$$

TPM was introduced later, slightly modified from RPKM. TPM is the measurement of the proportion of transcripts in the pool of RNA:

$$\text{TPM} = 10^6 * \frac{Reads\ mapped\ to\ a\ transcript\ /Transcript\ length}{Sum(Reads\ mapped\ to\ the\ transcript/Transcript\ length)}$$

When calculating TPM, the gene is first normalized for the gene length, followed by the sequencing depth. Therefore, the sum of all TPMs is the same in all samples.

The normalization methods of the popular edgeR (Robinson et al., 2010) and DESeq2 (Love et al., 2014) analysis packages do not consider the varying length of genes. The gene length-normalization has been omitted in them as correcting for this is not necessary for performing statistical testing to detect DE genes between sample groups (Dillies et al., 2013; Oshlack & Wakefield, 2009).

The Relative Log Expression (RLE) normalization method assumes that most of the genes are not DE. It scales the gene-wise read counts according to the gene's geometric mean across all samples. TMM normalization method estimates the relative gene-wise expression levels by computing absolute expression levels for each sample relative to the chosen reference sample and using these as scaling factors.

TMM normalization approach was used in Study I to normalize the raw count data. In Study II, RLE normalization was used with DESeq2 for normalizing the gene-, transcript-, and exon-level count data. FPKM values were used with the cuffdiff2 tool. In Study III, TPM values were produced using RSEM tool.

2.5.2.2 Statistical Testing

The goal of the statistical testing in the context of RNA-seq is to find DE (up- and downregulated) genes between the sample groups. A large set of tools and packages have been developed to perform statistical testing to detect DE genes. These methods can be divided into three different categories:

1. Methods based on negative binomial models, e.g. DESeq2, edgeR and baySeq (Hardcastle & Kelly, 2010).
2. Methods based on log-normal distribution, e.g. limma.
3. Non-parametric methods, e.g. SAMSeq (J. Li & Tibshirani, 2013)and Reproducibility-Optimized Test statistics (ROTS).

DE analysis methods typically provide their results in a table format where the rows represent the genes (or other features) and the columns represent different result values such as average expression, log2 fold-change (FC), p-value and adjusted p-value. The DE genes are typically selected based on log2FC and adjusted p-value. The FC, which is mostly given as log2, describes the size of the change in expression values between the two compared groups.

As thousands of tests are carried out, this can result in many false positive (FPs) findings. Multiple testing correction is thus carried out to adjust the statistical confidence measures based on the number of tests performed. A widely used approach for multiple testing correction is Benjamini-Hochberg (BH) correction (Benjamini & Yekutieli, 2001). Benjamini Hochberg ranks the p-values from the smallest to the largest, so that the smallest p-value is assigned the highest rank (1), and the largest p-value is assigned the rank n (n being the number of p-values). The Benjamini-Hochberg's critical value is compared to each p-value, the critical value calculated by (i/m) Q, where i is the rank, m is the total number of tests, and Q is the false discovery rate selected. The p-values smaller or equal to (i/m) Q are then regarded significant.

2.5.3 Differential Splicing Analysis

2.5.3.1 Differential Splicing (DS)

Various tools have been developed to find DS genes from the RNA-seq datasets, as shown in Fig. 7. The detection of DS genes is a challenging analysis task especially due to the short-read sequencing (usually 100-150 base pairs). Further, the accurate estimation of the transcript abundances is hindered due to the reads mapping to different transcripts of the same gene. In addition, it is difficult to account for the complexity of the splicing mechanism and, hence, the attempt has been to simplify

strategies to find DS genes. Another shortcoming of the current DS tools is that they are not necessarily maintained and updated regularly, and as a result, they can be difficult to install and operate. In many cases available documentation is also very limited. Further, it can be difficult to interpret and compare the results of the DS tools as various metrics are used to characterize the findings depending on the tool. In this Thesis, adjusted p-value (< 0.05) was used to select the DS genes. The first generation of DS tools, including MISO (Katz et al., 2010), MATS (Shen et al., 2012), ALEXA-Seq (Griffith et al., 2010), rSeqDiff (Y. Shi & Jiang, 2013) and SpliceSeq (Ryan et al., 2012), was limited to analyzing only one sample in each condition. Thus, these methods do not consider the biological variability between the samples within a sample group and hence cannot be used with experiments having replicates.

2.5.3.2 Differential Splicing Methodologies

Two major strategies have been designed to study DS using RNA-seq data: 1) isoform-based and 2) count-based methods (Fig. 7). The isoform-based methods estimate the expression of the full-length transcripts based on the sequencing reads. These methods detect the DS by revealing changes for each gene in the relative transcript abundances between two or more conditions. This is also known as the detection of differential transcript usage. Rather than estimating the transcript abundances, count-based methods detect DS between sample groups by comparing the distribution of reads on counting units, e.g. exons (and junctions) and splicing events.

The count-based methods are further classified into exon-based and event-based methods. The exon-based strategy assumes that the DS can be traced based on the exon and junction signals. These methods such as DEXSeq (Anders et al., 2012), JunctionSeq (Hartley & Mullikin, 2016), DSGseq (W. Wang et al., 2013), edgeR, and limma use the exon-level (and junction-level) count data to find DS genes by comparing the read counts on the exons or exons and junctions of the gene. The event-based methods include rMATS (Shen et al., 2014), SUPPA2 (Trincado et al., 2018), and dSpliceType (Zhu et al., 2015) and they compare the percentage spliced in values (PSI) between sample condition groups.

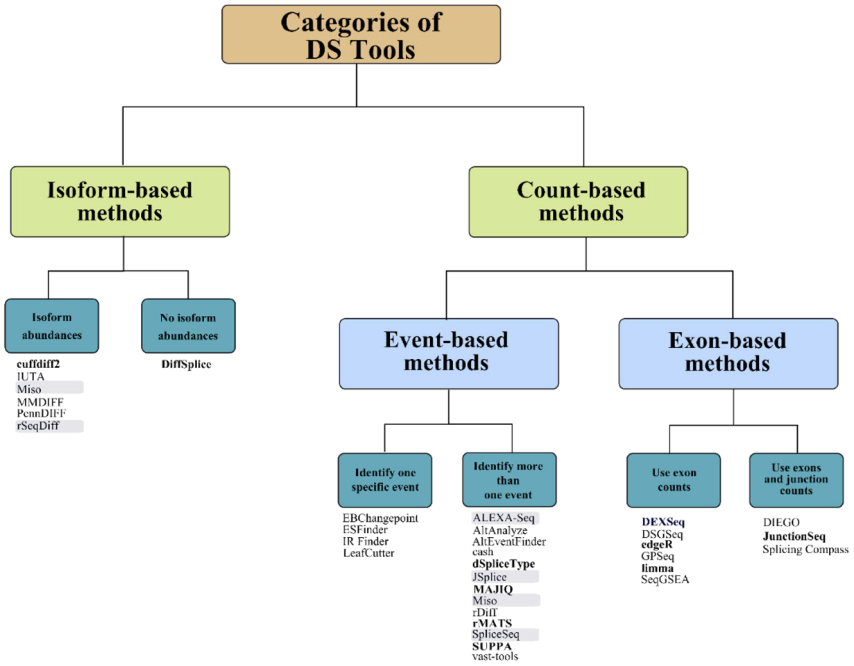

Figure 7. Schematic illustration of the methodologies developed for the DS analysis. The methods highlighted in grey do not support testing with replicated samples. The methods represented in bold were included in the method comparison conducted in this thesis (Adapted with permission from Publication III: SFigure 1).

2.5.3.3 Percentage Spliced In (PSI)

PSI is the percentage or proportion of the isoform's relative abundances, and the DS is then calculated in terms of the differences between these relative abundances. Initially, PSI was defined as the ratio of the density of inclusion reads to the sum of the densities of inclusion and exclusion reads:

$$\mathrm{PSI} = \frac{IR}{IR + ER},$$

where IR indicates the inclusion reads and ER the exclusion reads (Fig. 8). A PSI value of 1 indicates that the constitutive exons are included in all the gene's transcripts, and a value less than 1 means reduced inclusion of the alternative exon. However, this PSI approach fails to capture the complexity of splicing as the tools cannot identify in which transcripts the splicing event changes have taken place. To address this issue, MAJIQ (Vaquero-Garcia et al., 2016) introduced the concept of local splicing variations (LSVs). LSVs are defined as the splits in the graphs where several edges come into (or leave from) a single exon. MAJIQ detects, quantifies,

tests, and visualizes the LSVs and it also specifies the type of splicing events in the LSVs.

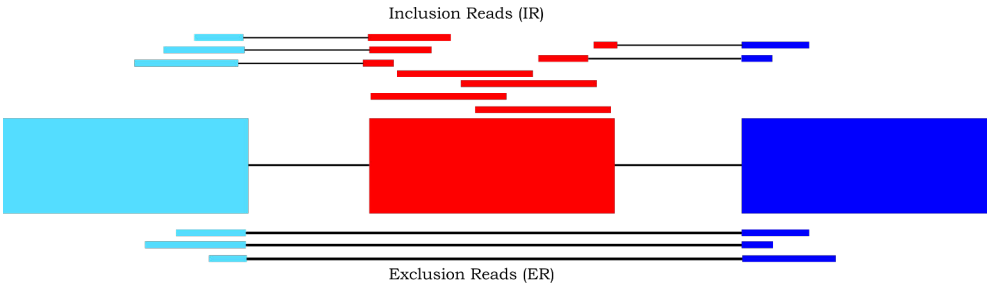

Figure 8. Skipped exon event in which the reads on the top represent the inclusion reads (IR), and the reads at the bottom represent the exclusion reads (ER).

The early methods (such as rMATS) calculated the PSI estimates based on the junctions, exons, or both. More sophisticated approaches (e.g. SUPPA and MAJIQ) also consider quantification uncertainty and lengths of isoform-specific segments to provide more robust estimates improving the predictive power and thus the resolution of detecting the DS genes.

SUPPA characterizes the PSI value of a splicing event through the set of transcripts and denotes them as F1 (event included) and F2 (no event). The PSI of a splicing event is the ratio of the abundance of transcripts that includes one form of the event over the abundance of transcripts with either form of the event (F1 ∪ F2).

$$\text{PSI} = \frac{\sum_{k \in F1} TPM_k}{\sum_{j \in F1 \cup F2} TPM_j},$$

where TPM_k means the transcripts that include the event and TPM_j means all transcripts of a gene.

3 Aims

This Thesis's main objective was to apply, improve and compare the existing methods developed for RNA-sequencing data analysis, especially focusing on differential gene expression and differential splicing of genes.

The specific aims carried out during the studies are as follows:

1. To apply the gene-level approach for detecting DE genes between tumor and normal tissue from primary prostate cancer specimens and to study how androgen concentration and androgen-regulated genes differ in tumors with or without a TMPRSS-ERG fusion gene.

2. To study whether the detection of DE genes can be improved if the initial statistical testing is carried out on a lower feature level (transcript counts, exon counts, or transcript compatibility counts (TCCs)).

3. To compare the DS tools' performance based on different evaluation metrics such as precision, recall, FDR, as well as time and memory usage and biological relevance of results.

4 Materials and Methods

4.1 Datasets

The dataset used in Study I was in-house data from patients diagnosed with localized adenocarcinoma of the prostate. The datasets for Studies II and III were publicly available and downloaded from Array Express (Parkinson et al., 2007), Gene Expression Omnibus (GEO) (Edgar et al., 2002), and the Sequence Read Archive (SRA) (Leinonen et al., 2011).

The Study I dataset consisted of tumor and benign prostate samples from 48 patients suffering from primary prostate cancer. The Ethics committee approved the study protocol of the Hospital District of Southwest Finland. Written consent was obtained from all patients participating in the study, conducted according to the Declaration of Helsinki principles. The RNA quality was confirmed using the Fragment Analyzer, and the RIN quality number of all the samples was sufficiently high (>5.5). The RNA-seq was performed at Finnish Functional Genomic Centre (FFGC), Turku, Finland.

In Study II, we used two publicly available datasets. The first dataset, Microarray Quality Control dataset (MAQC) (L. Shi et al., 2006), was downloaded from SRA, having the accession number SRA010153. The dataset consists of two samples from Ambion's human brain and Stratagene's human universal reference. The dataset was selected as it has a large number of corresponding qRT-PCR measurements, considered as the gold standard reference for gene expression. The second dataset consisted of 28 tumor and normal samples from 14 prostate cancer patients (Ren et al., 2013), and it was downloaded from ArrayExpress, having accession number E-MAT-567.

In Study III, we used four different datasets to evaluate the DS tools' robustness, running time, and memory usage. We selected two datasets for the relatively large number of samples available and the other two as they had approximately 30 qPCR-validated genes. The prostate cancer dataset used in Study II was used in Study III as well and was referred to as the Prostate Cancer dataset (PCa) in Study III. The Hepatocellular Carcinoma (HCa) dataset was downloaded from GEO, having accession number GSE77314 (Liu G, Hou G, Li L, Li Y, Zhou W, 2014). This dataset includes 100 tumor and normal samples from metastasis of hepatocellular

carcinoma. The Mouse dataset (MVS) was downloaded from GEO, having accession number GSE64357 (GSM1569076-77, GSM1569083-84) (Bebee et al., 2015). In MVS we compared samples of double knockouts of Esrps (Esrp1 and Esrp2 gene) and wild-type mice, which had corresponding 28 qPCR-validated DS genes. The Human data set (HVS) was downloaded from SRA, having accession number SRS354082 (Shen et al., 2014). It contains six samples from GS689 and PC3E prostate cancer cell lines, having 32 corresponding qPCR-validated genes.

4.2 Methodology and Analysis Tools

4.2.1 Gene-level Differential Expression (Study I)

We used a conventional gene-level approach to identify the DE genes between cancer and benign samples of patients with primary prostate cancer, and between TMPRSS-ERG+ and TMPRSS-ERG- tumors. The RNA-seq data quality was examined using the FastQC (v0.11.3) (Andrews, 2010). The sequencing reads were aligned to human reference genome (hg19) available at UCSC (downloaded from Illumina iGenome website) using STAR aligner (v2.5.0.c) (Dobin et al., 2013). The alignment bam files from different sequencing lanes were merged using the Picard tool (v1.77) (Broad Institute, 2009). The subread tool (v1.5.0) (Liao et al., 2014) was used to quantify the uniquely mapped reads associated with each gene using RefSeq gene annotations.

During the downstream analysis, the gene-level counts were normalized for library size using the TMM (Law et al., 2014) approach in the edgeR package (Robinson et al., 2010). The normalized counts were further transformed using the voom (Law et al., 2014) approach in the limma package (Ritchie et al., 2015). The DE genes between the sample groups were detected using the ROTS analysis package (Suomi et al., 2017). ROTS optimizes the reproducibility among a family of modified test statistics. The fusion genes in each sample were additionally identified with the FusionCatcher tool (v.0.99.6a) (Nicorici et al., 2014).

4.2.2 Exon-level Differential Expression (Study II)

In Study II, Kallisto (v 0.44.0) (Bray et al., 2016) was used to produce the TCCs, and the transcript abundances from the raw RNA-seq reads using the pseudo alignment and quantification algorithm. The human Ensemble GRCh38 (release 80) was used for genome and transcriptome annotation. tximport package (Soneson, Love, et al., 2016) was used to import the transcript and gene counts from the Kallisto result files. The alignment bam files produced with the Kallisto quant method were summarized at the exon level using the subread tool (v.1.6.2) (Liao et al., 2014). The

read count matrix was produced at four different feature levels (gene, exon, transcript, and TCCs). The data were normalized and filtered before performing the statistical testing with the DESeq2 (Love et al., 2014). Additionally, gene- and exon-level counts produced using STAR (v2.6.1b) and the subread tool (v.1.6.2) (Liao et al., 2014) were analyzed.

When using the alternate lower feature-level approach, feature-level p-values were aggregated either using the Lancaster method (Lancaster, 1961) or the empirical Brown's Method (ebm) (Poole et al., 2016). Lancaster method is the generalization of Fisher's method (Fisher, 1992) that uses weights for aggregating the p-values. The independent p-values are converted to chi-variables with w_i degrees of freedom. The test statistic for the Lancaster method becomes:

$$T = \sum_{i=1}^{K} \Phi_{wi}^{-1} \ (p_i)$$

under the null hypothesis, and Φ_{wi}^{-1} represents the inverse cumulative distribution function of the chi-square distribution with w_i degrees of freedom.

The ebm method is the empirical adaptation of Brown's approach (Brown, 1975). The method extended Fisher's method by considering the dependence between the p-values. In our study, this approach allows the consideration of multiple exons, transcripts, or TCCs per gene. For example, exons that belong to the same gene are not independent but often show similar expression levels. Brown developed an approximation to the Fisher test's null distribution when the p-values are derived from data with a multivariate normal distribution with a specified covariance matrix. The test statistic of the method is based on a re-scaled chi-square distribution χ_{2f}^2 where c is the constant scale factor and f is the re-scaled number of the degrees of freedom. Brown showed that the covariance could be calculated using numerical integration although it was computationally expensive, especially for large datasets. ebm provides a more efficient solution by approximating the covariance empirically based on the data.

The p-values provided by Lancaster and the ebm methods are corrected for multiple testing using the Benjamin Hochberg method.

4.2.3 Differential Splicing (Study III)

The files downloaded from SRA and GEO were converted to fastq format using sratoolkit (v.2.8.0) (Ncbi, 2011). The quality of the reads was checked for all the datasets using the FastQC tool (v0.11.3) (Andrews, 2010), and if needed, the low-quality reads were trimmed using trimgalore (v0.4.1) (Andrews, 2015). STAR (v2.6.1b) (Dobin et al., 2013) was used to align the reads to the Ensembl reference genome (*Homo sapiens*: GRCh37, *Mus musculus:* NCMIM37).

In Study III, we compared ten different DS tools. The tools represent both the isoform-based and count-based (exon-based, and event-based) approaches. Included isoform-based tools were cufflinks/cuffdiff2 (Trapnell et al., 2010; Trapnell, Hendrickson, et al., 2012), DiffSplice (Hu et al., 2013), exon-based tools were DEXSeq (Anders et al., 2012), edgeR (Robinson et al., 2010) , JunctionSeq (Hartley & Mullikin, 2016) and limma (Ritchie et al., 2015) and event-based tools were dSpliceType (Zhu et al., 2015), MAJIQ (Vaquero-Garcia et al., 2016), rMATS (Shen et al., 2014) and SUPPA/SUPPA2 (Alamancos et al., 2014; Trincado et al., 2018). The input files for the DS tools were prepared according to the descriptions provided by the tools.

4.3 Evaluation Of Results

In biological experiments, as the ground truth is generally missing, we used different evaluation metrics to test the performance of the different approaches and tools.

4.3.1 Partial Area Under the Curve (pAUC)

In study II, we evaluated the different counting scheme's accuracy using the MAQC dataset as it has corresponding qRT-PCR measurements available for 840 genes. pROC package (Robin et al., 2011) was used to calculate the pAUC with specificity above 0.8 at various $\log_2$FC values ranging from 0.5 to 5. pAUC summarizes the portion of the receiver operating curve (ROC) over a specified interval of interest (Ma et al., 2013). The ROC curve is a plot of sensitivity against 1-specificity for the varying value threshold (Hajian-Tilaki, 2013).

4.3.2 False Discovery Rate (FDR)

In Study II and Study III, the FDR was calculated for prostate cancer (PCa) and hepatocellular carcinoma (HCa) datasets by performing mock comparisons by randomly subsampling samples into two groups using only the samples from the normal group. We hypothesized that the genes detected in these comparisons would represent false positives (FP) as the difference between these normal samples is minor compared to the real comparisons. Further, FDR was calculated by scaling the median number of FP found in the mock comparison to the number of genes found in the corresponding real comparison.

4.3.3 Precision and Recall

In Study III, we performed real comparisons by randomly subsampling from the tumor and normal samples into subsets. We repeated each subsampling ten times for subsets and performed the analysis once for the whole dataset. We selected the DS genes having FDR < 0.05. However, we first aggregated the results obtained from the DS tools in the form of isoform-, exon-, and event- level to the gene-level results.

In Study III, we also evaluated the consistency and reproducibility of the DS tools by calculating the precision and recall. They were calculated by comparing the DS genes found in the subsample to those detected in the complete dataset. The precision is defined as:

$$\text{Precision } (DS_{full}, DS_{subset}) = \frac{|\, DS_{full} \cap DS_{subset} \,|}{|\, DS_{subset}|}$$

The recall is defined as:

$$\text{Recall } (DS_{full}, DS_{subset}) = \frac{\left|\, DS_{full} \cap DS_{subset} \,\right|}{\left|\, DS_{full}\right|}$$

4.3.4 Functional Enrichment Analysis

Functional enrichment analysis, also known as gene set enrichment analysis, helps to determine whether some biological functions are enriched in the list of query genes. In Study III, functional enrichment analysis was carried out using topGO package (Alexa & Rahnenfuhrer, 2016) to find the most enriched gene ontologies (GO) by considering the top 500 genes detected by each DS tool. For summarization, we selected the top ten most enriched terms in at least one of the methods. dSpliceType was not considered in this analysis, as it only provided the DS genes under FDR < 0.05, which were less than 500. The p-values of the GO terms were used to perform hierarchical clustering and visualized using heatmaps. Heatmap is a data visualization technique in which the data matrix is represented graphically in two dimensions, and the values/magnitudes are represented by color intensity.

5 Results

5.1 Conventional gene-level analysis- Study I

TMPRSS2-ERG gene fusion is commonly found in prostate cancer and is potentially associated with androgen concentration (DHT, DHEA, testosterone, and A-dione) and changes in the androgen metabolizing enzymes. In this study, gene expression levels and androgen concentrations were measured using RNA-seq and gas chromatography-tandem mass spectrometry (GC-MS/MS) technologies, respectively. DHT concentration was significantly higher in the cancerous samples ($P < 0.001$) than in the benign samples, while no significant differences were found between the DHEA, testosterone, and A-dione concentration. The samples which did not have any hormonal therapies prior to the surgery were classified as TMPRSS2-ERG+ ($n = 23$) or TMPRSS2-ERG- ($n = 15$) based on the ERG expression measured from RNA-seq data. The RNA-seq results were validated for five TMPRSS-ERG+ patients and four TMPRSS-ERG- patients using immunohistochemistry, and a full match was observed.

The unsupervised hierarchical clustering was performed for the log2 transformed cancer/benign ratios of the androgens in TMPRSS-ERG+ and TMPRSS-ERG- patients. This hierarchical clustering produced three clusters, out of which two were enriched with TMPRSS-ERG+ tumors (Fig. 9). These results show that androgen biosynthesis and metabolism are altered in TMPRSS-ERG+ and TMPRSS-ERG- tumors compared to the benign samples. Further, the DHT/ testosterone ratios were higher in TMPRSS-ERG+ tumors than in TMPRSS-ERG- tumors, and no difference was found between the benign and serum levels. RNA-seq analysis was carried out to study the expression of the 5α-reductase (SRD5A) enzymes which convert testosterone to DHT. SRD5A3 expression was higher in the TMPRSS-ERG+ cancer specimen and suggested a testosterone independent DHT biosynthesis via an alternative pathway. Furthermore, 31 androgen-regulated genes were DE between TMPRSS-ERG+ and TMPRSS-ERG- tumors, including ERG and other well-characterized androgen-dependent genes such as NKX3.1, STEAP4, and SPOCK1. Altogether, this study suggests altered androgen response due to different androgen concentrations between TMPRSS-ERG+ and TMPRSS-ERG- tumors.

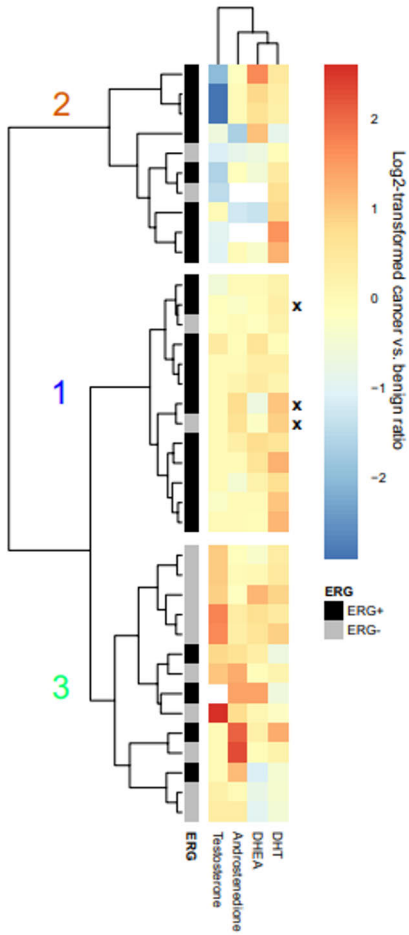

Figure 9. Heatmap of the hierarchical clustering of the androgen concentration normalized to the benign tissue concentrations in patients without hormonal treatment and marked with TMPRSS-ERG status. The samples showing biological recurrence marked with x (Adapted with permission from Publication I: Figure 3A).

5.2 Exon-level estimates- Study II

In Study II, we compared the performance of the conventional gene-level approach to the alternate approach using different count data (transcript counts, exon counts, TCCs) in detecting the DE genes. The accuracy and robustness of the approaches were tested using two publicly available RNA-Seq datasets (MAQC and prostate cancer). In the alternate approach, the lower-feature level data was tested using DESeq2, and the p-values were then aggregated to gene-level p-values using either the Lancaster or ebm method.

In the MAQC dataset, we found that the alternate approach of testing the lower feature levels, especially exon and TCCs with the ebm aggregation method, outperformed the accuracy of the conventional gene-level approach (Fig. 10). With the Lancaster aggregation method, the performance of exon and TCCs based approaches was poor. This clearly indicates that with exon and TCCs, it is important to choose an aggregation method that is able to take the dependence of the features into consideration. The results based on the exon counts from either STAR or Kallisto were similar while in contrast, with the gene-level counts Kallisto provided clearly better results.

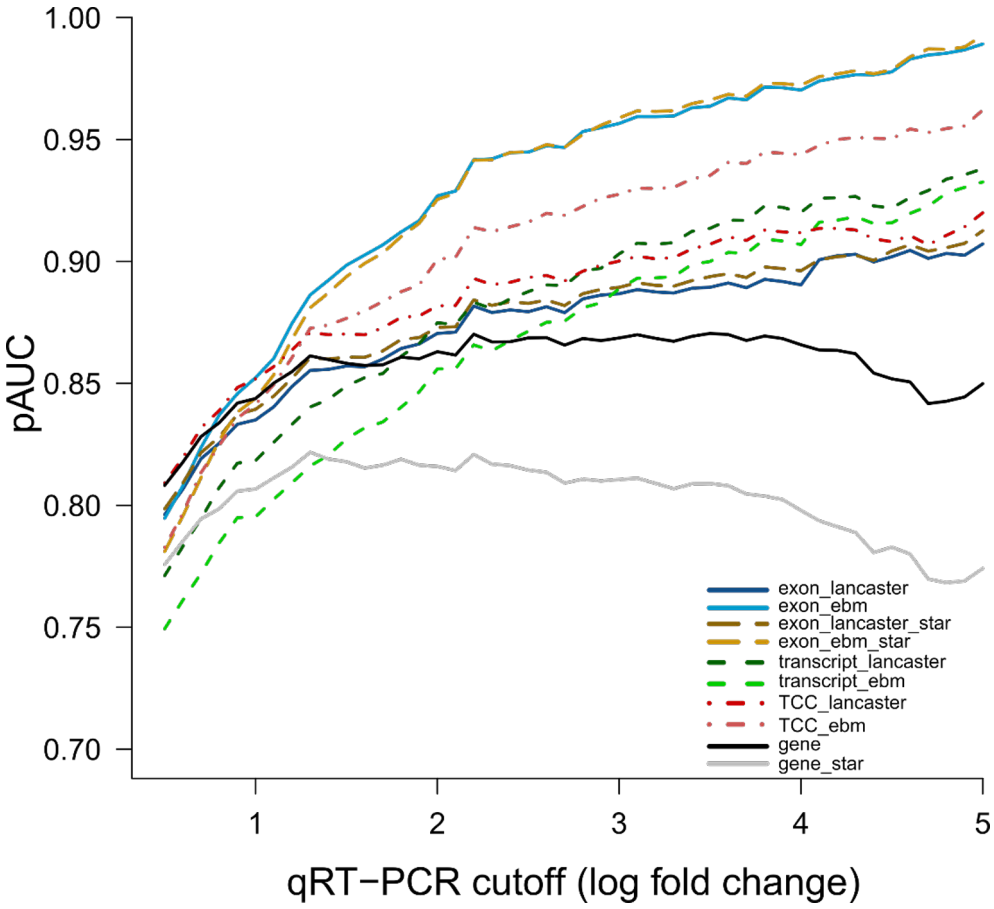

Figure 10. pAUC values at a specificity 0.8 at varying qRT-PCR cut-offs ranging from 0.5 to 5 with an increment of 0.1 across different levels of count data in MAQC dataset (Adapted with permission from Publication II: Figure 2).

Furthermore, we studied the robustness of the approaches by detecting the number of DE genes and FDR in the prostate cancer dataset. Here we found that with the

lower feature-level approaches more DE genes were detected than with the conventional gene-level approach. Overall, the different analysis schemes showed low FDR, except for the exon count and TCC based approaches when used with the Lancaster aggregation method (Fig. 11). The analysis scheme using gene counts from STAR produced fewer DE genes compared to Kallisto - however, the FDRs remained the same.

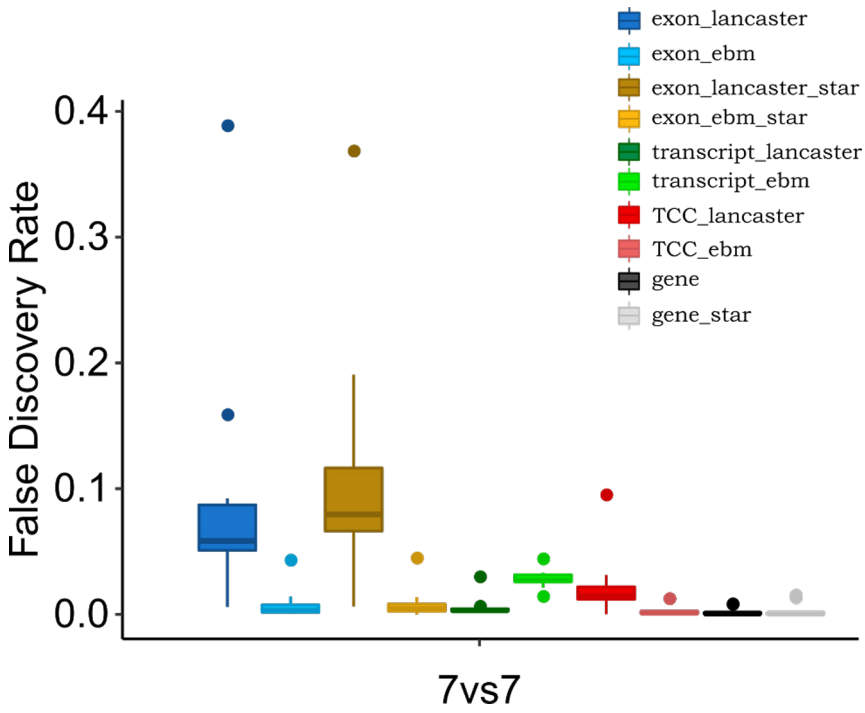

Figure 11. FDR across the different analysis schemes based on the mock and real comparisons in the prostate cancer dataset (Adapted with permission from Publication II: Figure 3B).

5.3 Differential Splicing Comparison – Study III

In addition to detecting DE genes, RNA-seq data is often used to identify DS genes. We performed a comprehensive comparison of ten DS tools, where the chosen tools represented different analysis strategies: isoform-based methods (cuffdiff2, DiffSplice), exon-based methods (edgeR, DEXSeq, JunctionSeq and limma) and event-based methods (dSpliceType, MAJIQ and rMATS (rMATS_3.2.2 and rMATS_3.2.5) and SUPPA (SUPPA and SUPPA2)). Study III compared the tools' reproducibility and consistency using PCa and HCa datasets by performing real and mock comparisons. The evaluation was based on the number of DS gene identifications, precision, recall, and FDR. In addition, we evaluated the tools by the

similarity between the detected DS genes and the top 500 ranked DS genes, functional enrichment analysis results, runtime, and memory usage. Furthermore, the MVS and HVS datasets were used to evaluate the tools' ability to identify genes, previously validated using qPCR. The HVS dataset was further used to study the effect of sequencing depth on the DS tools.

The number of DS genes detected was highly variable between the tools in PCa and HCa datasets (Fig. 12A and B). In general, cuffdiff2 and SUPPA2 detected the lowest number of DS genes compared to other tools. A robust tool is expected to detect more DS genes when the number of available replicate samples increases. However, SUPPA/SUPPA2, cuffdiff2 and dSpliceType identified fewer DS genes with the increasing sample number. In addition, with DiffSplice the number of detected DS genes did not correlate with the number of available samples and with MAJIQ roughly the same number of DS genes was detected irrespective of the sample size.

We calculated the tool's precision and recall (Fig. 12C, D, E, and F) by comparing the number of DS genes detected in the randomly sampled subsets to the number of detections in the complete PCa and HCa datasets. We found that, overall, both precision and recall increased with the increasing number of samples except for cuffdiff2, which did not detect any DS genes in most subsets.

FDR and its variability decreased in the random subsets with the increase in the number of samples (Fig. 12G and H). Moreover, DiffSplice showed the lowest FDR in both datasets, followed by the exon-based methods DEXSeq, JunctionSeq and limma. The event-based methods rMATS and MAJIQ performed better than other event-based methods.

Furthermore, we inspected the overlap between DS genes identified by different tools. Fig. 13A shows the overlap of the DS genes in the HCa dataset. The DS gene lists of different tools in general showed high overlap with limma results which produced the longest DS gene list. Most DS genes detected by limma on the other hand, were not identified by other tools.

To overcome the huge difference in the length of the DS gene lists produced by the different tools, we further compared the top 500 ranked genes from each tool. dSpliceType was not considered for this as it provided less than 500 DS genes. We found generally a low overlap between the top 500 ranked genes across the tools. However, the highest overlap between the top-ranked genes was detected by the two versions of rMATS and SUPPA, respectively. The highest overlap between different tools was found between the exon-based tools, and the lowest overlap was found between the isoform-based tool DiffSplice and other methods.

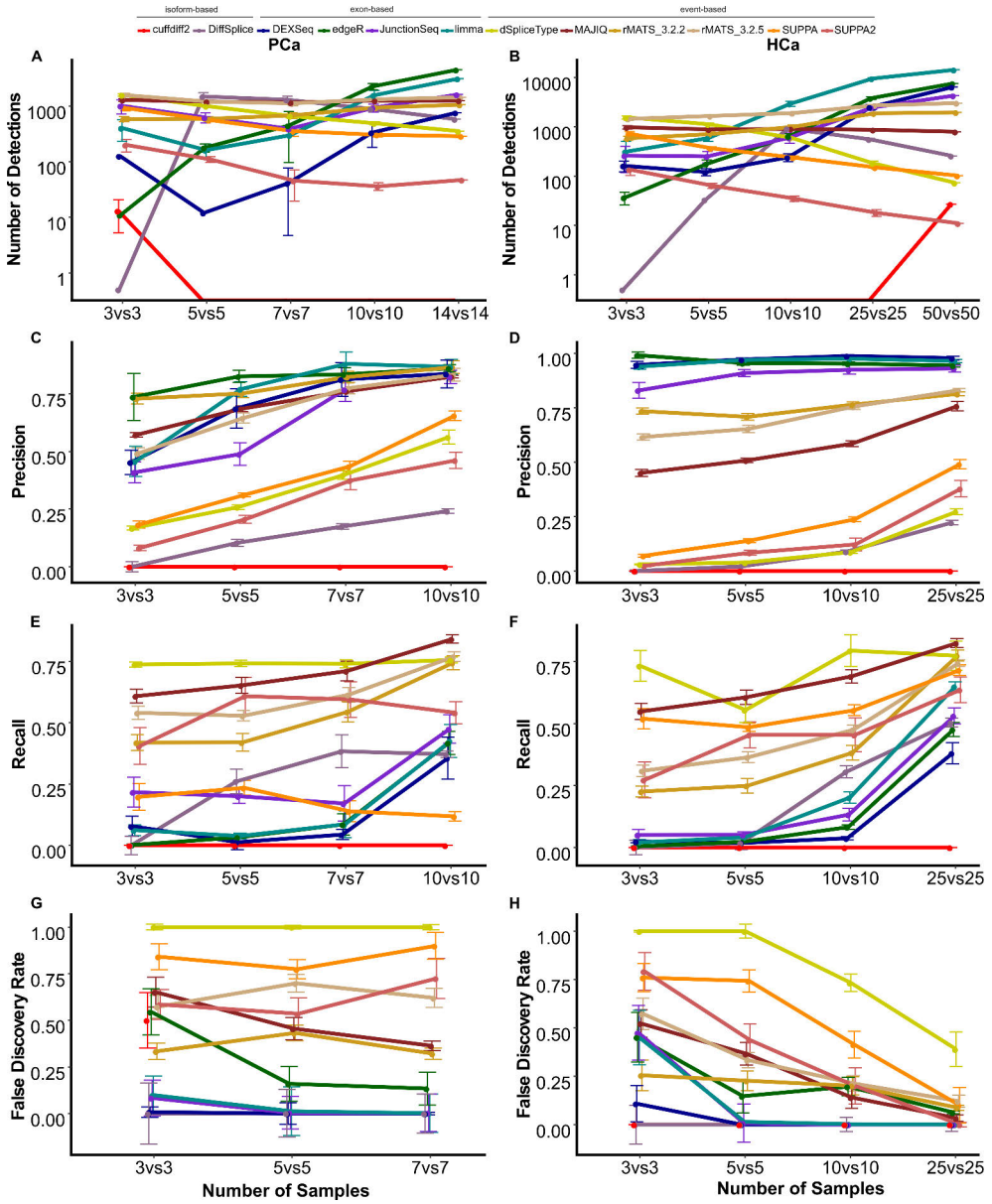

Figure 12. Median and standard error of the number of detections, precision, recall and FDR of the ten repetitions with the different number of samples in the PCa and HCa datasets. Number of DS genes detected in **A)** PCa and **B)** HCa dataset. Precision in the **C)** PCa and **D)** HCa dataset. Recall in **E)** PCa and **F)** HCa dataset. FDR in the **G)** PCa and **H)** HCa dataset (Adapted with permission from Publication III: Figure 2).

Additionally, we ran the gene ontology enrichment analysis of the top 500 ranked DS genes for both the PCa and HCa datasets. Fig. 13B shows the heatmap of the top gene ontology enrichment p-values in the HCa dataset. rMATS and MAJIQ detected the most significantly enriched GO terms.

We chose MVS and HVS datasets for further validation as they have a number of qPCR-validated DS genes available (28 in MVS and 32 in HVS datasets). The event-based tools; MAJIQ and SUPPA, detected the highest number of validated DS genes in both datasets (MAJIQ 88% and SUPPA 71% in the MVS datasets and MAJIQ 94% and SUPPA 97% in the HVS dataset). All tools detected a high percentage of validated DS genes in the HVS dataset whereas the tools in general detected a varying number of DS genes in the MVS dataset. dSpliceType did not provide any results for the MVS dataset due to an unknown technical error.

The HVS dataset was further used to investigate the effect of sequencing depth as it had more than 100 million reads per sample. We calculated the number of detections, precision, and recall at various down-sampling levels of the sequencing data (20 - 100 million reads), using DS genes in the full HVS dataset as the truth set. We observed that the precision and number of validated qPCR DS genes were generally stable above 40 to 50 million reads.

The total running time and memory usage were measured for each tool on the PCa and HCa datasets. The tools were run on a computer cluster, managed by free, open-source Simple Linux Utility for resource management (SLURM). Overall, limma and edgeR took the shortest time to run, and MAJIQ used the least maximum memory. We observed large differences in the running times: the fastest tools limma and edgeR took an hour to run, whereas cuffdiff2, DEXSeq, DiffSplice, JunctionSeq and rMATS took days to run for the complete dataset.

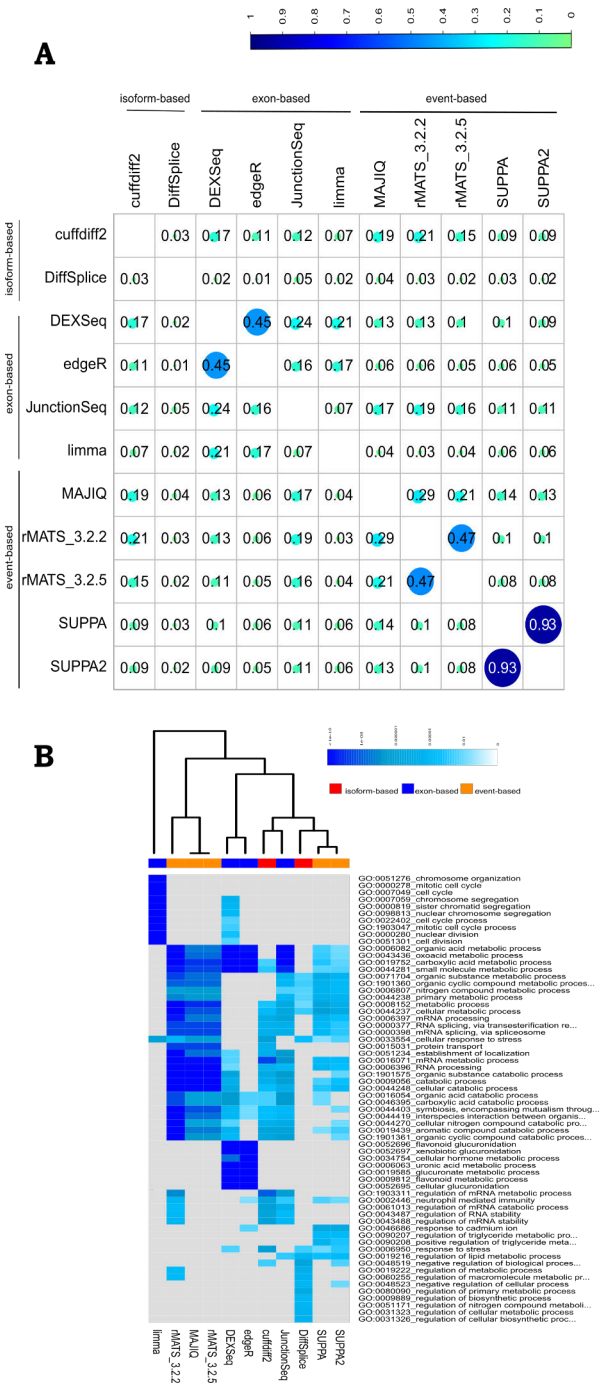

Figure 13. Similarity between the methods in the complete HCa dataset within the top 500 genes from each method. **B)** Heatmap of the p-values of the top enriched GO biological processes across the tools in the HCa dataset. The grey color represents missing values (Adapted with permission from Publication III: Figure 3B and 4B).

6 Discussion

RNA-seq technology is nowadays extensively used in transcriptome studies and many methodologies and tools have been developed for preprocessing and the downstream analysis of the data. Possible analysis applications include detecting differentially expressed genes, differentially spliced genes, allele-specific expression, fusion genes, and variants (SNPs). However, the influx of tools and methodologies has made it difficult for researchers to decide which method or strategy they should apply to analyze their data in order to perform an efficient analysis and produce consistent and reliable results. When new methods/tools are published, their superiority is typically shown by benchmarking them to existing state-of-the-art methods. Unfortunately, these comparisons may be biased, and there exists a need for independent, comprehensive comparisons of these methods/tools in order to help the researchers to make their choice of the approach to use. In this thesis, I tackled two different aspects of RNA-seq analysis, particularly focusing on analyzing differential expression and differential splicing. I applied the conventional gene-level approach to prostate adenocarcinoma samples and further demonstrated how the choice of the feature level for statistical analysis combined with different aggregation methods may impact the results. I compared the performance of ten differential splicing detection tools in terms of robustness, similarity, and consistency on biological data from human and mouse samples, with varying sample sizes and sequencing depth.

In Study I, I applied the conventional gene-level approach of differential expression analysis in primary prostate cancer data to further find an association between androgen concentrations and androgen-regulated genes. Moreover, the patients who did not receive any hormonal therapies were divided into TMPRSS-ERG fusion gene positive and negative tumors based on the expression of the ERG gene. The results indicate that TMPRSS-ERG positive tumor samples have distinct intratumoral androgen profiles compared to TMPRSS-ERG negative tumor samples, which potentially leads to testosterone-independent DHT production via an alternative pathway and induces androgen target gene expression. In the alternative pathway, A-dione is converted to androstanedione followed by its conversion to DHT by HSD17B activity (Chang et al., 2011). Hence, novel drugs inhibiting

testosterone-independent androgen biosynthesis could potentially be a treatment of choice for patients diagnosed with TMPRSS-ERG positive gene fusion.

Differential expression detection approaches have largely matured, and many comparisons have been performed to find out the best methods (Quinn et al., 2018; Seyednasrollah et al., 2013; Soneson & Delorenzi, 2013). However, the previous work has concentrated on gene-level count data, leaving the question open whether using lower feature-level data would improve the detection of differentially expressed genes. Some previous studies indicate that using the lower feature-level data could improve the statistical power and, thus, the accuracy of detecting differentially expressed genes. Yi et. al used transcript counts, and transcript compatibility counts whereas Laiho et. al used exon counts in their studies. However, Study II is to our knowledge the first one that systematically compared several feature levels and also further investigated the effect of taking the dependence of the features into consideration during statistical testing.

For aggregating the feature p-values to gene-level results, I firstly selected the Lancaster method as it outperformed Fisher and Sidak methods in an earlier comparison (Yi et al., 2018). Secondly, I selected the ebm aggregation method as it takes the dependence of the features into account. In Study II, we evaluated the accuracy and robustness of the different lower feature-level schemes and compared them to the conventional gene-level approach using the MAQC and prostate cancer datasets. My results were consistent with the earlier studies (Laiho et. al and Yi et. al) in that the statistical power of detecting the differentially expressed genes increased with the presence of multiple measurements per gene in the initial statistical testing. In my analysis, I further found out that the analysis scheme based on exon-level count data outperformed the other feature-level schemes when combined with ebm aggregation method. This exon-level scheme also showed the largest improvement with the ebm aggregation method out of the different feature-level schemes. The success of the ebm aggregation method is not surprising, considering the strong dependence of exon level expression values across each gene that this method is able to take into consideration. Noticeably, the transcript and TCC based analysis schemes also improved the results over the conventional gene-level approach. However, estimating the transcript abundances remains a challenging task especially in the situation when only the 3'-ends of the transcripts are sequenced. Another shortcoming of using TCCs is that their biological interpretation is difficult as they do not have a direct biological element associated with them, contrary to exons and transcripts. There are also no established annotations for the TCCs associated with a gene, although they depend on the reads that are compatible with a specific set of transcripts.

Although many tools have been developed for detecting differential splicing, only very few of them are widely used. In addition, these tools have not been

extensively compared for their performance and making such comparisons is challenging for several reasons. The tools developed differ greatly in their strategy and the whole workflow typically needs to be set up separately for each tool. For example, the preparation of annotation files even for one workflow is a considerable effort, requiring the user to overcome numerous technical issues which can be challenging to tackle due to incomplete documentation. Most of the tools are also not routinely updated, and thus, it can be very hard to get them running in the latest computing environments. Another challenge for benchmarking DS tools arises from the lack of reliable references or spike in datasets.

Despite the present challenges complicating the DS tool evaluation and comparison, in Study III, I set out to compare the performance of ten different DS analysis tools, covering isoform-based and count-based (exon-based and event-based) approaches. For the comparisons, I selected four RNA-seq datasets: two out of them for the relatively large number of replicates available, deep enough sequencing, and appropriate read length, and the other two datasets for their availability of qPCR-validated splicing events. As the comprehensive ground truth information was missing, I based the evaluation on the assumption that the tools that constantly perform robustly across different evaluation metrics would most likely perform well also in other studies. The consistency and robustness of the tools were evaluated by performing real and mock comparisons by randomly subsampling samples from the full dataset into subsets.

In our DS tool comparison, the exon-based methods (DEXseq, JunctionSeq, limma, edgeR) and the event-based methods (MAJIQ, rMATS) overall showed low FDR, high precision, and moderate recall, thus, showing robust performance in the PCa and HCa datasets. In the HVS and MVS datasets, event-based methods MAJIQ and SUPPA identified the highest proportion of RT-qPCR validated genes. The different tools in general showed low similarity between the top-ranking DS genes, with the highest similarity observed between the exon-based tools. The event-based methods MAJIQ and rMATS showed the most enriched gene ontology terms. Limma and edgeR took the least running time and average memory usage, whereas MAJIQ took the least maximum memory. Comparison of the tools showed that the results largely varied between them, and no single tool outperformed the others in all the evaluation metrics. These results are largely consistent with Liu's and Soneson's previous work in which they evaluated the DS methods on real and simulated plant RNA-seq data (Liu et al., 2014; Soneson, Matthes, et al., 2016).

The DS tools were run with the default settings as this is what most users are likely to do. In the future, it would be interesting to compare the performance of these tools using different parameter settings and to see how they influence the results.

The alignment files for the comparisons were produced using STAR, which was selected for its robust performance in previous aligner comparisons, even with default settings. The focus of my study also was not on the influence of the aligner on the DS, although it is possible that some tools perform better with a specific aligner. This is an issue that could be considered in later studies.

Another limitation of my DS tool comparison study is that the focus was on making the group comparisons between two sample groups, rather than incorporating more complex experimental setups and including batch effects or confounding variables. However, all exon-based tools included in the comparison also support complex experimental design, whereas cuffdiff2, dSpliceType, and MAJIQ only support unpaired two-group comparisons. rMATS and SUPPA/SUPPA2 in addition support paired two-group comparisons.

Although few methods (JunctionSeq, rMATS, and MAJIQ) can find novel unannotated splicing events, my DS comparison study was limited by running the tools with the complete annotation containing the gene structure (except with DiffSplice which does not require any). The comparison of tools using incomplete annotation was not attempted as generally low overlap and a decrease in performance were seen between the top-ranking DS genes already when using the DS tools with complete annotations. In the future, when the tools mature, it will be interesting to evaluate their performance also on incomplete annotation. Also repeating the current study with complete annotation would be recommended when new high-quality ground truth datasets eventually become available.

Besides methodological development there has been an evolution in sequencing technologies from short-read sequencing to long-read sequencing that will increase the accuracy of alternative splicing pattern detection. Recently, long-read sequencing methods such as PacBio or Oxford nanopore, have gained popularity for RNA-seq experiments. In addition, Illumina has also introduced a new long-read protocol, called Ultra-Long Read sequencing for DNA that can be adapted to RNA-seq in the future. The long-read methods allow the detection of full-length transcripts including splice variants and they provide greater accuracy in transcript quantification. Further, emerging technologies in full-length single cell RNA-seq such as Smart-seq2 and STRT-seq provide information about splice variants in individual cells, providing unprecedented resolution into cell-to-cell variability in alternative splicing events.

7 Summary/Conclusions

The studies carried out in this Thesis aim to identify and develop robust methodologies and tools for the reliable detection of differentially expressed and differentially spliced genes in RNA-seq datasets. The choice of the analysis method or strategy has a huge impact on the downstream analysis and the end results. Hence, this work provides guidelines for the researchers regarding the optimal selection of tools for their work.

The Thesis is divided into two subcategories. The first part concentrates on differential expression analysis, specifically by applying conventional gene-level analysis to cancer data and further testing the performance of different lower feature-level analysis schemes in comparison to the conventional gene-level approach (Study I, II). Based on the performance evaluation, my suggestion is to use the alternate exon-level approach for the detection of differentially expressed genes and combine this approach with the ebm method for aggregating the p-values to the gene level. Following this recommendation, I have implemented EBSEA package for differential expression analysis, which I am actively maintaining in the R/Bioconductor analysis platform.

In the second part of the Thesis, I evaluated the performance of ten differential splicing analysis tools. The tools were chosen based on their popularity among users and in addition considering some very recently developed promising tools. Based on the large differences between the results of the tools observed, I would currently recommend using more than one tool in any given study and then concentrating on the overlapping findings produced by the different tools.

The studies presented in this Thesis provide a valuable resource for researchers, aiming to optimize their differential expression and differential splicing analysis workflows, by enabling them to make informed decisions on the methods and strategies to incorporate.

Acknowledgements

This research work was carried out at the Medical Bioinformatics Centre in Turku Bioscience, University of Turku and Åbo Akademi during 2016-2020.

I would like to express my sincere gratitude to all those who have supported and encouraged me during my doctoral studies. Firstly, I would like to thank my supervisors Professor Laura Elo, Dr. Asta Laiho, and Professor Matti Poutanen for their guidance throughout the Ph.D. process. Their expertise and insights have been invaluable and I am deeply grateful for their mentorship and support that have kept me focused and motivated during my studies. I would also thank my follow-up committee member Daniel Nicorici for his constructive feedback on the progress of my research.

I would also like to extend my appreciation to my co-authors who contributed to the articles included in my thesis Aidan J. McGlinchey, Matias Knuutttila, Mikko Venäläinen, and Ning Wang. I am thankful to my colleagues at Medical Bioinformatic Centre for their support. In particular, I want to acknowledge Deepankar Chakraborty, Fatemeh Seyednasrollah, Maria Jaakkola, Mehrad Mahmoudian, Teemu Daniel Laajala, Tomi Suomi, Veronika Suni, and Ye Hong. I would also thank Dr. Jukka Lehtonen and Esko Pakarinen for their technical support and advice on technical issues. I would also like to thank the friends I made in my doctoral school Gabriela Martinez Chacon, Hanna Heikelä, Srikar Nagelli and Syeda Afshan.

I am also grateful to my family for their unwavering support and encouragement, particularly during the more challenging moments in my life. Their belief in me has been a constant source of motivation, and I cannot express my gratitude enough. Thanks to my parents and siblings (Shehzad Mahmood and Mariam Rukh) for everything you have done for me in this life and for the emotional support. Last but not the least, I am grateful to my best friend and husband, Asim Alvie, for the support, especially during the tough times. My kids, Arish and Zoya, are the best thing that ever happened to me. Your smiles, laughter, and hugs have always brightened up my days. I am very fortunate to have you in my life.

Thank you to all those who have contributed to my journey, in big and small ways, and for making unforgettable memories. I am truly grateful for your support.

April 2023
Arfa Mehmood

References

Aartsma-Rus, A., & Van Ommen, G. J. B. (2007). Antisense-mediated exon skipping: A versatile tool with therapeutic and research applications. In *RNA*. https://doi.org/10.1261/rna.653607

Alamancos, G. P., Pagès, A., Trincado, J. L., Bellora, N., & Eyras, E. (2014). SUPPA: a super-fast pipeline for alternative splicing analysis from RNA-Seq. *BioRxiv*, 008763. https://doi.org/10.1101/008763

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

Alexa, A., & Rahnenfuhrer, J. (2016). topGO: enrichment analysis for Gene Ontology. R Packag. version 2.26.0. *R Package Version 2.26.0*. https://doi.org/10.1038/ncomms8832

Anders, S., & Huber, W. (2010). Differential expression analysis for sequence count data. *Genome Biology*, *11*(10). https://doi.org/10.1186/gb-2010-11-10-r106

Anders, S., Pyl, P. T., & Huber, W. (2015). HTSeq-A Python framework to work with high-throughput sequencing data. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btu638

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

Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data.* Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/. https://doi.org/citeulike-article-id:11583827

Andrews, S. (2015). *Babraham Bioinformatics - Trim Galore!* Trim Galore! Wrapper Script for Automated Quality and Adapter Trimming and Quality Control. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Bebee, T. W., Park, J. W., Sheridan, K. I., Warzecha, C. C., Cieply, B. W., Rohacek, A. M., Xing, Y., & Carstens, R. P. (2015). The splicing regulators Esrp1 and Esrp2 direct an epithelial splicing program essential for mammalian development. *ELife*, *4*(September2015). https://doi.org/10.7554/eLife.08954

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, *29*(4), 1165–1188. https://doi.org/10.1214/aos/1013699998

Braunschweig, U., Barbosa-Morais, N. L., Pan, Q., Nachman, E. N., Alipanahi, B., Gonatopoulos-Pournatzis, T., Frey, B., Irimia, M., & Blencowe, B. J. (2014). Widespread intron retention in mammals functionally tunes transcriptomes. *Genome Research*. https://doi.org/10.1101/gr.177790.114

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

Breschi, A., Gingeras, T. R., & Guigó, R. (2017). Comparative transcriptomics in human and mouse. In *Nature Reviews Genetics* (Vol. 18, Issue 7, pp. 425–440). Nature Publishing Group. https://doi.org/10.1038/nrg.2017.19

Broad Institute. (2009). *Picard Tools - By Broad Institute*. Github.

Brown, M. B. (1975). 400: A Method for Combining Non-Independent, One-Sided Tests of Significance. *Biometrics*. https://doi.org/10.2307/2529826

Chang, K. H., Li, R., Papari-Zareei, M., Watumull, L., Zhao, Y. D., Auchus, R. J., & Sharifi, N. (2011). Dihydrotestosterone synthesis bypasses testosterone to drive castration-resistant prostate cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(33), 13728–13733. https://doi.org/10.1073/pnas.1107898108

Chen, L. (2013). Statistical and Computational Methods for High-Throughput Sequencing Data Analysis of Alternative Splicing. *Statistics in Biosciences*, *5*(1), 138–155. https://doi.org/10.1007/s12561-012-9064-7

Churko, J. M., Mantalas, G. L., Snyder, M. P., & Wu, J. C. (2013). Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. In *Circulation Research*. https://doi.org/10.1161/CIRCRESAHA.113.300939

Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M. W., Gaffney, D. J., Elo, L. L., Zhang, X., & Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. In *Genome Biology* (Vol. 17, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13059-016-0881-8

Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., Keime, C., Marot, N. S., Castel, D., Estelle, J., Guernec, G., Jagla, B., Jouneau, L., Laloë, D., Le Gall, C., Schaëffer, B., Le Crom, S., Guedj, M., & Jaffrézic, F. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbs046

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

Dvinge, H., & Bradley, R. K. (2015). Widespread intron retention diversifies most cancer transcriptomes. *Genome Medicine*, *7*(1). https://doi.org/10.1186/s13073-015-0168-9

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*. https://doi.org/10.1093/nar/30.1.207

El Marabti, E., & Younis, I. (2018). The cancer spliceome: Reprograming of alternative splicing in cancer. In *Frontiers in Molecular Biosciences*. https://doi.org/10.3389/fmolb.2018.00080

Fisher, R. A. (1992). *Statistical Methods for Research Workers*. https://doi.org/10.1007/978-1-4612-4380-9_6

Gao, Q., Liang, W. W., Foltz, S. M., Mutharasu, G., Jayasinghe, R. G., Cao, S., Liao, W. W., Reynolds, S. M., Wyczalkowski, M. A., Yao, L., Yu, L., Sun, S. Q., Caesar-Johnson, S. J., Demchok, J. A., Felau, I., Kasapi, M., Ferguson, M. L., Hutter, C. M., Sofia, H. J., … Ding, L. (2018). Driver Fusions and Their Implications in the Development and Treatment of Human Cancers. *Cell Reports*. https://doi.org/10.1016/j.celrep.2018.03.050

Garber, M., Grabherr, M. G., Guttman, M., & Trapnell, C. (2011). Computational methods for transcriptome annotation and quantification using RNA-seq. In *Nature Methods*. https://doi.org/10.1038/nmeth.1613

Glaus, P., Honkela, A., & Rattray, M. (2012). Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bts260

Griffith, M., Griffith, O. L., Mwenifumbo, J., Goya, R., Morrissy, A. S., Morin, R. D., Corbett, R., Tang, M. J., Hou, Y.-C., Pugh, T. J., Robertson, G., Chittaranjan, S., Ally, A., Asano, J. K., Chan, S. Y., Li, H. I., McDonald, H., Teague, K., Zhao, Y., … Marra, M. A. (2010). Alternative expression analysis by RNA sequencing. *Nature Methods*, *7*(10), 843–847. https://doi.org/10.1038/nmeth.1503

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. In *Caspian J Intern Med* (Vol. 4, Issue 2).

Han, Y., Gao, S., Muegge, K., Zhang, W., & Zhou, B. (2015). Advanced applications of RNA sequencing and challenges. *Bioinformatics and Biology Insights*, *9*, 29–46. https://doi.org/10.4137/BBI.S28991

Hardcastle, T. J., & Kelly, K. A. (2010). BaySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, *11*. https://doi.org/10.1186/1471-2105-11-422

Hartley, S. W., & Mullikin, J. C. (2016). Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq. *Nucleic Acids Research*, *44*(15), e127. https://doi.org/10.1093/nar/gkw501

Hooper, J. E. (2014). A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Human Genomics*, *8*(1), 3. https://doi.org/10.1186/1479-7364-8-3

Hou, Z., Jiang, P., Swanson, S. A., Elwell, A. L., Nguyen, B. K. S., Bolin, J. M., Stewart, R., & Thomson, J. A. (2015). A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Scientific Reports*. https://doi.org/10.1038/srep09570

Hu, Y., Huang, Y., Du, Y., Orellana, C. F., Singh, D., Johnson, A. R., Monroy, A., Kuan, P. F., Hammond, S. M., Makowski, L., Randell, S. H., Chiang, D. Y., Hayes, D. N., Jones, C., Liu, Y., Prins, J. F., & Liu, J. (2013). DiffSplice: The genome-wide detection of differential splicing events with RNA-seq. *Nucleic Acids Research*, *41*(2). https://doi.org/10.1093/nar/gks1026

Kanitz, A., Gypas, F., Gruber, A. J., Gruber, A. R., Martin, G., & Zavolan, M. (2015). Comparative assessment of methof for the computational inference of transcript isoform abundance from RNA-seq data. *Genome Biology*, *16*(1). https://doi.org/10.1186/s13059-015-0702-5

Katz, Y., Wang, E. T., Airoldi, E. M., & Burge, C. B. (2010). Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, *7*(12), 1009–1015. https://doi.org/10.1038/nmeth.1528

Kim, D., Paggi, J. M., Park, C., Bennett, C., & Salzberg, S. L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology*. https://doi.org/10.1038/s41587-019-0201-4

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., & Salzberg, S. L. (2013). TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology*. https://doi.org/10.1186/gb-2013-14-4-r36

Kim, E., Goren, A., & Ast, G. (2008). Alternative splicing: Current perspectives. In *BioEssays* (Vol. 30, Issue 1, pp. 38–47). https://doi.org/10.1002/bies.20692

Kim, E., Magen, A., & Ast, G. (2007). Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl924

Kim, J., & Easley, C. J. (2011). Isothermal DNA amplification in bioanalysis: Strategies and applications. In *Bioanalysis*. https://doi.org/10.4155/bio.10.172

Kukurba, K. R., & Montgomery, S. B. (2015). RNA sequencing and analysis. *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.top084970

Laiho, A., & Elo, L. L. (2014). A note on an exon-based strategy to identify differentially expressed genes in RNA-seq experiments. *PLoS ONE*, *9*(12). https://doi.org/10.1371/journal.pone.0115964

Lancaster, H. O. (1961). THE COMBINATION OF PROBABILITIES: AN APPLICATION OF ORTHONORMAL FUNCTIONS. *Australian Journal of Statistics*. https://doi.org/10.1111/j.1467-842X.1961.tb00058.x

Latysheva, N. S., & Babu, M. M. (2016). Discovering and understanding oncogenic gene fusions through data intensive computational approaches. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw282

Law, C. W., Chen, Y., Shi, W., & Smyth, G. K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biology*, *15*, R29. https://doi.org/10.1186/gb-2014-15-2-r29

Lee, T. I., & Young, R. A. (2013). Transcriptional regulation and its misregulation in disease. In *Cell*. https://doi.org/10.1016/j.cell.2013.02.014

Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkq1019

Leslie A. Pray. (2008). Eukaryotic genome complexity. *Nature Education*, *96*, 1–1.

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1), 323. https://doi.org/10.1186/1471-2105-12-323

Li, J., & Tibshirani, R. (2013). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical Methods in Medical Research*, *22*(5), 519–536. https://doi.org/10.1177/0962280211428386

Liao, Y., Smyth, G. K., & Shi, W. (2014). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. https://doi.org/10.1093/bioinformatics/btt656

Liu G, Hou G, Li L, Li Y, Zhou W, L. L. (2014). Potential diagnostic and prognostic marker dimethylglycine dehydrogenase (DMGDH) suppresses hepatocellular carcinoma metastasis in vitro and in vivo. *Oncotarget*, *7*, 32607–32616.

Liu, R., Loraine, A. E., & Dickerson, J. A. (2014). Comparisons of computational methods for differential alternative splicing detection using RNA-seq in plant systems. *BMC Bioinformatics*, *15*(1), 364. https://doi.org/10.1186/s12859-014-0364-4

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Ma, H., Bandos, A. I., Rockette, H. E., & Gur, D. (2013). On use of partial area under the ROC curve for evaluation of diagnostic performance. *Statistics in Medicine*. https://doi.org/10.1002/sim.5777

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*. https://doi.org/10.1038/nmeth.1226

Ncbi. (2011). SRA Handbook. *Archives*, 1–14.

Nicorici, D., Satalan, M., Edgren, H., Kangaspeska, S., Murumagi, A., Kallioniemi, O., Virtanen, S., & Kilkku, O. (2014). FusionCatcher - a tool for finding somatic fusion genes in paired-end RNA-sequencing data. In *bioRxiv*. https://doi.org/10.1101/011650

Oshlack, A., & Wakefield, M. J. (2009). Transcript length bias in RNA-seq data confounds systems biology. *Biology Direct*. https://doi.org/10.1186/1745-6150-4-14

Palazzo, A. F., & Lee, E. S. (2015). Non-coding RNA: What is functional and what is junk? *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2015.00002

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., & Blencowe, B. J. (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, *40*(12), 1413–1415. https://doi.org/10.1038/ng.259

Parker, B. C., & Zhang, W. (2013). Fusion genes in solid tumors: An emerging target for cancer diagnosis and treatment. In *Chinese Journal of Cancer*. https://doi.org/10.5732/cjc.013.10178

Parkhomchuk, D., Borodina, T., Amstislavskiy, V., Banaru, M., Hallen, L., Krobitsch, S., Lehrach, H., & Soldatov, A. (2009). Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkp596

Parkinson, H., Kapushesky, M., Shojatalab, M., Abeygunawardena, N., Coulson, R., Farne, A., Holloway, E., Kolesnykov, N., Lilja, P., Lukk, M., Mani, R., Rayner, T., Sharma, A., William, E., Sarkans, U., & Brazma, A. (2007). ArrayExpress - A public database of microarray experiments and gene expression profiles. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkl995

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

Petrova, O. E., Garcia-Alcalde, F., Zampaloni, C., & Sauer, K. (2017). Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. *Scientific Reports*. https://doi.org/10.1038/srep41114

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., & Knijnenburg, T. A. (2016). Combining dependent P-values with an empirical adaptation of Brown's method. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw438

Quinn, T. P., Crowley, T. M., & Richardson, M. F. (2018). Benchmarking differential expression analysis tools for RNA-Seq: Normalization-based vs. log-ratio transformation-based methods. *BMC Bioinformatics*. https://doi.org/10.1186/s12859-018-2261-8

Ren, S., Peng, Z., Mao, J.-H., Yu, Y., Yin, C., Gao, X., Cui, Z., Zhang, J., Yi, K., Xu, W., Chen, C., Wang, F., Guo, X., Lu, J., Yang, J., Wei, M., Tian, Z., Guan, Y., Tang, L., … Sun, Y. (2013). RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. *Cell Research*, *23*(5), 732–732. https://doi.org/10.1038/cr.2013.61

Rio, D. C., Ares, M., Hannon, G. J., & Nilsen, T. W. (2010). Enrichment of poly(A)+ mRNA using immobilized oligo(dT). *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.prot5454

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Müller, M. (2011). pROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-12-77

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., & Weinstein, J. N. (2012). SpliceSeq: A resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. *Bioinformatics*, *28*(18), 2385–2387. https://doi.org/10.1093/bioinformatics/bts452

Ryu, W. S. (2016). Molecular Virology of Human Pathogenic Viruses. In *Molecular Virology of Human Pathogenic Viruses*. https://doi.org/10.1016/c2013-0-15172-0

Schroeder, A., Mueller, O., Stocker, S., Salowsky, R., Leiber, M., Gassmann, M., Lightfoot, S., Menzel, W., Granzow, M., & Ragg, T. (2006). The RIN: An RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*. https://doi.org/10.1186/1471-2199-7-3

Seyednasrollah, F., Laiho, A., & Elo, L. L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, *16*(1), 59–70. https://doi.org/10.1093/bib/bbt086

Shen, S., Park, J. W., Huang, J., Dittmar, K. A., Lu, Z. X., Zhou, Q., Carstens, R. P., & Xing, Y. (2012). MATS: A Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data. *Nucleic Acids Research*, *40*(8). https://doi.org/10.1093/nar/gkr1291

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

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., De Longueville, F., Kawasaki, E. S., Lee, K. Y., Luo, Y., Sun, Y. A., Willey, J. C., Setterquist, R. A., Fischer, G. M., Tong, W., Dragan, Y. P., Dix, D. J., Frueh, F. W., … Zong, Y. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*. https://doi.org/10.1038/nbt1239

Shi, Y., & Jiang, H. (2013). rSeqDiff: detecting differential isoform expression from RNA-Seq data using hierarchical likelihood ratio test. *PLoS One*, *8*(11), e79448. https://doi.org/10.1371/journal.pone.0079448

Soneson, C., & Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-14-91

Soneson, C., Love, M. I., & Robinson, M. D. (2016). Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences [version 2; referees: 2 approved]. *F1000Research*. https://doi.org/10.12688/F1000RESEARCH.7563.2

Soneson, C., Matthes, K. L., Nowicka, M., Law, C. W., & Robinson, M. D. (2016). Isoform prefiltering improves performance of count-based methods for analysis of differential transcript usage. *Genome Biology*. https://doi.org/10.1186/s13059-015-0862-3

Stark, R., Grzelak, M., & Hadfield, J. (2019). RNA sequencing: the teenage years. In *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0150-2

Stengel, A., Nadarajah, N., Haferlach, T., Dicker, F., Kern, W., Meggendorfer, M., & Haferlach, C. (2018). Detection of recurrent and of novel fusion transcripts in myeloid malignancies by targeted RNA sequencing. *Leukemia*. https://doi.org/10.1038/s41375-017-0002-z

Sultan, M., Amstislavskiy, V., Risch, T., Schuette, M., Dökel, S., Ralser, M., Balzereit, D., Lehrach, H., & Yaspo, M. L. (2014). Influence of RNA extraction methods and library selection schemes on RNA-seq data. *BMC Genomics*. https://doi.org/10.1186/1471-2164-15-675

Suomi, T., Seyednasrollah, F., Jaakkola, M. K., Faux, T., & Elo, L. L. (2017). ROTS: An R package for reproducibility-optimized statistical testing. *PLoS Computational Biology*, *13*(5). https://doi.org/10.1371/journal.pcbi.1005562

Trapnell, C., Hendrickson, D. G., Sauvageau, M., Goff, L., Rinn, J. L., & Pachter, L. (2012). Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, *31*(1), 46–53. https://doi.org/10.1038/nbt.2450

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., Pimentel, H., Salzberg, S. L., Rinn, J. L., & Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, *7*(3), 562–578. https://doi.org/10.1038/nprot.2012.016

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., & Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology*, *28*(5), 511–515. https://doi.org/10.1038/nbt.1621

Trincado, J. L., Entizne, J. C., Hysenaj, G., Singh, B., Skalic, M., Elliott, D. J., & Eyras, E. (2018). SUPPA2: Fast, accurate, and uncertainty-aware differential splicing analysis across multiple conditions. *Genome Biology*. https://doi.org/10.1186/s13059-018-1417-1

Vaquero-Garcia, J., Barrera, A., Gazzara, M. R., Gonzalez-Vallinas, J., Lahens, N. F., Hogenesch, J. B., Lynch, K. W., & Barash, Y. (2016). A new view of transcriptome complexity and regulation through the lens of local splicing variations. *ELife*. https://doi.org/10.7554/eLife.11752

Wagner, G. P., Kin, K., & Lynch, V. J. (2012). Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory in Biosciences*. https://doi.org/10.1007/s12064-012-0162-3

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., & Burge, C. B. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. https://doi.org/10.1038/nature07509

Wang, L., Si, Y., Dedow, L. K., Shao, Y., Liu, P., & Brutnell, T. P. (2011). A low-cost library construction protocol and data analysis pipeline for illumina-based strand-specific multiplex RNA-seq. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0026426

Wang, W., Qin, Z., Feng, Z., Wang, X., & Zhang, X. (2013). Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*. https://doi.org/10.1016/j.gene.2012.11.045

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. In *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2484

Yi, L., Pimentel, H., Bray, N. L., & Pachter, L. (2018). Gene-level differential analysis at transcript-level resolution. *Genome Biology*, *19*(1), 53. https://doi.org/10.1186/s13059-018-1419-z

Zhu, D., Deng, N., & Bai, C. (2015). A Generalized dSpliceType Framework to Detect Differential Splicing and Differential Expression Events Using RNA-Seq. *IEEE Transactions on Nanobioscience*, *14*(2), 192–202. https://doi.org/10.1109/TNB.2015.2388593

TURUN
YLIOPISTO
UNIVERSITY
OF TURKU

Painosalama, Turku, Finland 2023

ISBN 978-951-29-9317-8 (PRINT)
ISBN 978-951-29-9318-5 (PDF)
ISSN 0355-9483 (Print)
ISSN 2343-3213 (Online)