

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

**A BENCHMARK STUDY FOR RNA-SEQ BASED SPLICING
ANALYSIS TOOLS: DETECTION OF ABERRANT SPLICING
EVENTS FOR RARE DISEASES**

CANAN KOLAKOĞLU
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING

GEBZE
2023

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES

**A BENCHMARK STUDY FOR RNA-SEQ
BASED SPLICING ANALYSIS TOOLS:
DETECTION OF ABERRANT SPLICING
EVENTS FOR RARE DISEASES**

CANAN KOLAKOĞLU
A THESIS SUBMITTED FOR THE DEGREE OF
MASTER OF SCIENCE
DEPARTMENT OF BIOENGINEERING
BIOINFORMATICS AND SYSTEMS BIOLOGY
PROGRAMME

THESIS SUPERVISOR
PROF. DR. TUNAHAN ÇAKIR

GEBZE
2023

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

RNA-SEQ TABANLI UÇ BİRLŞETİRME
ANALİZİ ARAÇLARI İÇİN BİR
PERFORMANS ÇALIŞMASI: NADİR
HASTALIKLAR İÇİN ANORMAL UÇ
BİRLŞETİRME OLAYLARININ TESPİTİ

CANAN KOLAKOĞLU
YÜKSEK LİSANS TEZİ
BİYOMÜHENDİSLİK ANABİLİM DALI
BİYOİNFORMATİK VE SİSTEM BİYOLOJİSİ
PROGRAMI

DANIŞMANI
PROF. DR. TUNAHAN ÇAKIR

GEBZE

2023



YÜKSEK LİSANS JÜRİ ONAY FORMU

GTÜ Fen Bilimleri Enstitüsü Yönetim Kurulu'nun 19/07/2023 tarih ve 2023/40 sayılı kararıyla oluşturulan jüri tarafından 16/08/2023 tarihinde tez savunma sınavı yapılan Canan KOLAKOĞLU'nun tez çalışması Biyomühendislik Anabilim Dalı Biyoinformatik ve Sistem Biyolojisi Programında YÜKSEK LİSANS tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. Tunahan ÇAKIR

ÜYE

: Dr. Öğr. Üyesi Pınar PİR

ÜYE

: Dr. Öğr. Üyesi Süleyman AKTUNA

ONAY

Gebze Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulu'nun

...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

SUMMARY

Rare diseases affect the general population in a ratio of 1/2000. Nearly 8000 distinct rare diseases exist today, and there are approximately 400 million patients worldwide suffering from rare diseases (Global Genes, 2021). Diagnosis and treatment of a rare disease are quite challenging in terms of time and cost since it involves differentiating a rare event from others. These challenges have led the research in this field to the Next Generation Sequencing (NGS) approach. Especially, RNA sequencing has recently emerged as a complementary approach to genomic approaches. It is known that pathogenic variants affect expression through aberrant splicing. Additionally, it is reported that 15-30% of the variants related to Mendelian diseases cause aberrant splicing. Therefore, RNA-Seq could serve as a key approach thanks to its ability to represent the functional effect of genetic abnormalities by reporting gene expression and splicing.

In this study, a benchmark study was applied using recently developed RNA-Seq specific tools to detect outlier splicing events - FRASER, dasper and SCISSOR. To this end, a publicly available RNA-Seq dataset was utilized, which was derived from the blood samples of people suffering from Cystic Fibrosis (CF). Samples were analyzed to unveil whether the unresponsiveness of some CF patients to lumacaftor/ivacaftor treatment was caused by any abnormal splicing event. As a result, TNFRSF18, CCL3, CD300A, RAB11FIP3, ABCB4, ELMO3 and RAPGEF1 genes were detected to have splicing abnormalities when the samples before and after drug treatment were compared. Also, these genes were shown to interact with cystic fibrosis transmembrane conductance regulator (CFTR) protein. The three tools mentioned above were used to identify abnormal splicing events, and their performances were compared through the methods of splice junction-based outlier detection and base coverage-based outlier detection. The benchmarking study showed that the FRASER tool outperformed the other tools.

Key Words: Rare Diseases, Cystic Fibrosis, RNA-Seq, Aberrant Splicing, Bioinformatic Tools

ÖZET

Nadir hastalıklar günümüzde dünya nüfusunun yaklaşık 2000’de birini etkilemekte ve 8000 farklı nadir hastalık bulunmaktadır. Global Genes’e göre, dünya çapında 400 milyon kişi nadir hastalık taşımaktadır. Nadir bir hastalığın teşhisi ve tedavisi zaman ve maliyet açısından oldukça zorlayıcıdır. Bu zorluklar, bu alandaki çalışmaları Yeni Nesil Dizileme (YND) yaklaşımına yönlendirmiştir. Alandaki yeni gelişmeler ile, RNA dizilemenin (RNA-Seq) genomik yaklaşımların tamamlayıcısı olarak kullanıldığı çalışmaların sayısı artmaktadır. Literatürde, patojenik varyantların anormal uç birleştirme olaylarına sebep olarak gen ifadesini etkilediği belirtilmiştir. Ayrıca, Mendeliyen hastalıklar ile ilişkili varyantların %15-30’unun anormal uç birleştirmeye neden olduğu belirtilmektedir. Bu bilgilere göre, RNA-Seq verisi, gen ifadesi ve uç birleştirme analizleri aracılığıyla genetik anormalliklerin işlevsel etkisini gösterme potansiyeli sayesinde önemli bir yaklaşım olarak değerlendirilmektedir.

Bu çalışmada, son zamanlarda geliştirilen ve RNA-Seq verisine özgü aykırı uç birleştirme olaylarını tespit eden üç biyoinformatik araç (FRASER, dasper ve SCISSOR) kullanılarak bir performans karşılaştırması çalışması gerçekleştirilmiştir. Bu amaçla, Kistik Fibrozis (CF) hastalığına sahip insanların kan RNA-Seq verilerini içeren halka açık bir veriseti kullanılmıştır. Örnekler, verisetindeki bazı CF hastalarının lumakaftor/ivakaftor tedavisine yanıt vermemesinin herhangi bir anormal uç birleştirme olayından kaynaklanıp kaynaklanmadığını ortaya çıkarmak için analiz edilmiştir. Sonuçta, TNFRSF18, CCL3, CD300A, RAB11FIP3, ABCB4, ELMO3 ve RAPGEF1 genlerinin uç birleştirme düzeyinde anormalliklere sahip olduğu ve ilaç tedavisinden önce ve sonra hasta örnekleri arasında farklılaştığı saptanmıştır. Ayrıca, bu genler sistik fibrozis transmembran iletkenlik düzenleyici (CFTR) proteini ile etkileşim göstermektedir. Yukarıda bahsedilen üç biyoinformatik araç kullanılarak anormal uç birleştirme analizleri gerçekleştirilmiş ve bu araçların performansları karşılaştırılmıştır. FRASER aracının diğer araçlara göre daha iyi bir performans sergilediği gösterilmiştir.

Anahtar Kelimeler: Nadir Hastalıklar, Kistik Fibroz, RNA-Seq, Anormal Uç Birleştirme, Biyoinformatik Araçlar

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my thesis supervisor, Prof. Dr. Tunahan ÇAKIR, who not only shared his profound scientific knowledge with me but also encouraged me to learn from failures on the path to success. His support, suggestions and invaluable guidance provided me to keep my enthusiasm alive and gave me the will to complete this work.

I am deeply grateful to PHITECH Bioinformatics for providing me with the great opportunity to conduct my research and for all the resources and support they provided. All the simulations in this study were performed by using the infrastructure provided by PHITECH Bioinformatics. I would like to extend special thanks to my colleagues; Ünzile Güven GÜLHAN, Özgenur AKDAŞ and Betül CEYLAN, who shared with me their experiences and insights.

Finally, I am deeply thankful to my family and my grandparents for their unconditional love and support.

TABLE of CONTENTS

	<u>Page</u>
SUMMARY	v
ÖZET	vi
ACKNOWLEDMENTS	vii
TABLE of CONTENTS	viii
LIST of ABBREVIATIONS and ACRONYMS	x
LIST of FIGURES	xii
LIST of TABLES	xiii
1. INTRODUCTION	1
2. BACKGROUND ASPECTS	3
2.1. Rare Diseases	3
2.2. Next Generation Sequencing	5
2.3. Alternative Splicing	7
2.4. Bioinformatic Solutions to Discover Aberrant Splicing Events from RNA-Seq Data	11
2.4.1. FRASER Tool	12
2.4.2. dasper Tool	15
2.4.3. SCISSOR Tool	20
2.5. Cystic Fibrosis	23
3. MATERIAL AND METHODS	25
3.1. Samples	25
3.2. Bioinformatic Analysis	27
4. RESULTS	31
4.1. QC Analysis	31
4.2. Aberrant Splicing Analysis through FRASER and dasper tools	31
4.3. Outlier Sample Removal Prior to Differential Expression and Differential Splicing Analysis	32
4.4. Detecting the Genes with Abnormal Splicing Events in Non- Responders through Differential Splicing Analysis	38
4.5. Aberrant Splicing Analysis through SCISSOR	42

4.6. Detection of DEGs between Drug Responders and Drug Non-Responders and Identification of PPIs	42
4.7. Comparison of the Splicing and Differential Expression Analysis Results	43
4.8. Benchmarking: Performance Comparison Between Three Recently Developed Splicing Tools	47
5. CONCLUSION	49
REFERENCES	51
BIOGRAPHY	56



LIST OF ABBREVIATIONS AND ACRONYMS

<u>Abbreviations</u> <u>and Acronyms</u>	<u>Explanations</u>
---	---------------------

θ	: Splicing efficiency metrics
ψ_3	: Alternative donor usage metric
ψ_5	: Alternative acceptor usage metric
ASE	: Allele-specific expression
BMI	: Body mass index
CF	: Cystic fibrosis
CFTR	: Cystic fibrosis transmembrane conductance regulator
CNVs	: Copy number variants
DE	: Differential expression
DEG	: Differentially expressed gene
FC	: Fold changes
FDR	: False discovery rate
GTEx	: Genotype-Tissue Expression
kbp	: Kilobase pairs
LCC	: Large control cohort
NGS	: Next generation sequencing
OMIM	: Online Mendelian Inheritance in Man
PCA	: Principle component analysis
PPI	: Protein-protein interaction
QC	: Quality control
RDs	: Rare diseases
RGD	: Rare genetic diseases
RIN	: RNA integrity number
RNA-Seq	: RNA sequencing
RNPs	: Ribonuclear-protein complexes
SCC	: Small control cohort
snoRNAs	: Small nucleolar RNAs
snRNAs	: Small nuclear ribonucleoprotein RNAs

SRA	: Sequence Read Archive
UTRs	: Untranslated regions
VUS	: Variants of unknown significance
WES	: Whole exome sequencing
WGS	: Whole-genome sequencing



LIST OF FIGURES

<u>Figure No:</u>	<u>Page</u>
1: Basic elements of alternative splicing	9
2: The process of alternative splicing	10
3: The workflow of FRASER tool	13
4: Categorization of junction annotation through dasper	17
5: The workflow of dasper tool	19
6: Main steps in the SCISSOR pipeline	21
7: Sample cohort setup of PRJNA512474 project with 60 samples from 40 people	25
8: PCA Graph applied prior to the differential expression analysis	34
9: PCA graphs of the pre-drug group applied prior to the differential splicing analysis	35
10: PCA graphs of the post-drug group applied prior to the differential splicing analysis	36
11: PCA graph to determine the separation of CF samples and control samples in terms of expression profiles	37
12: Venn diagram for the result of abnormal splicing and differential expression analysis for the pre-drug group	44
13: Venn diagram for the result of abnormal splicing and differential expression analysis for the post-drug group	45
14: PPI graph of CFTR gene with the genes detected from differential expression results and differential splicing results	46

LIST OF TABLES

<u>Table No:</u>	<u>Page</u>
1: Rare disease definitions according to different regions.	3
2: Sample metadata	26
3: Drug response of the CF patients	26
4: The genes with significant splicing abnormalities for non-responders but not responders	39

1. INTRODUCTION

Rare diseases (RDs) are disorders affecting a very small percentage of the population. The European Union defines a disease as rare if it affects 1 in 2,000 people in the population. The rarity of RDs makes their diagnosis complicated in terms of time and cost. 80% of the RDs are known to have genetic causes. Therefore, next generation sequencing (NGS) techniques are used as current diagnostic tests in routine clinical practice to identify causative genes. However, despite their current applications, the success in the molecular diagnostic rates achieved with whole exome sequencing (WES) and whole-genome sequencing (WGS) are 35-50%, being far from complete. One of the most important reasons of incomplete success of DNA-based diagnostics through WES and WGS is the lack of interpretation of the functional outcomes of the genetic abnormalities at protein level. In order to fill this gap, transcriptomic-integrated genomic approaches have evolved where RNA sequencing (RNA-Seq) can provide information at functional level by identifying transcriptional abnormalities in the causative gene. RNA-Seq can annotate the pathogenicity of variants using measurements beyond gene expression, such as alternative splicing. According to available reports, a significant proportion of genetic abnormalities associated with Mendelian diseases (15% to 30%) is due to aberrant splicing events, which are currently not possible to detect by DNA-based diagnostics (eg. WES, WGS).

Along with the emerging awareness of the advantages of transcriptomic applications to diagnose Mendelian RDs, recent studies have offered bioinformatic solutions to detect rare splicing abnormalities from RNA-Seq data. For instance, recently developed RNA-Seq based bioinformatic tools; FRASER [Mertes et al., 2021], dasper [Zhang et al., 2021] and SCISSOR [Choi et al., 2021] detect abnormal splicing events that occur with a low frequency in the cohort. This thesis study will be a benchmark study for these recently developed bioinformatics tools to evaluate the performance of these tools in an RNA-Seq dataset collected from rare disease patients.

In this study, we will analyze an RNA-Seq dataset obtained from publicly available databases to investigate splicing abnormalities. The dataset includes samples from patients diagnosed with cystic fibrosis, a rare lung disease. Cystic fibrosis can result from 2000 different variants, some of which lead to the creation of cryptic splice sites. Furthermore, there may be certain factors contributing to disease that can

influence drug responsiveness. Among patients who carry the same variants, it was reported that some individuals responded positively to drug treatment while others did not. Here, in order to highlight the capability of RNA-Seq in identifying outlier splicing events, the cystic fibrosis samples within the dataset were analyzed using bioinformatics tools for aberrant splicing detection: FRASER, dasper, and SCISSOR. The analysis aimed to investigate whether any abnormal splicing events were responsible for the unresponsiveness of certain cystic fibrosis patients to lumacaftor/ivacaftor treatment.

This thesis report was organized as follows: Section 2 describes the background aspects about rare diseases, next generation sequencing, alternative splicing, bioinformatic solutions to discover aberrant splicing events from RNA-Seq data and cystic fibrosis. Section 3 describes materials and methods. Section 4 reports the results from the computational analyses applied on samples and from the comparison of methodologies. Finally, Section 5 provides the conclusions of the thesis study.

2. BACKGROUND ASPECTS

2.1. Rare Diseases

RDs are medical conditions that affect a small number of people compared to the general population. The definition of rarity can vary by region, with some diseases being rare in one area but common in another, as it is summarized in Table 1 [Wikipedia, 2023]. In Europe, the definition of a rare disease is the disease that affects the general population in a ratio of 1/2000. There are thousands of RDs identified so far, and new ones are regularly discovered. Nearly 8000 distinct RDs exist today. The number of RDs can vary depending on the level of specificity used in their classification. While most genetic diseases are rare, not all RDs are genetic. According to Globalgenes records, 80% of RDs have genetic origins and most of these diseases are monogenic/Mendelian [Global Genes]. RDs can manifest from birth or appear in adulthood and are often severe, chronic and even life-threatening. Although RDs are individually rare, there are approximately 400 million patients worldwide suffering from RDs. The ratio of people affected by rare disease is 1/10, and 1/2 of them are children. Moreover, 3/10 of these children die before the age of 5 [Global Genes] .

Table 1: Rare disease definitions in different regions.

Country	Patient ratio as defined	Patient ratio standardized for comparison
Brazil	65 in 100,000	1 in 1,538
United States	<200,000 in population	1 in 1,659
Argentina	1 in 2,000	1 in 2,000
Australia	5 in 10,000	1 in 2,000
European Union	5 in 10,000	1 in 2,000
Mexico	5 in 10,000	1 in 2,000
Singapore	1 in 2,000	1 in 2,000
Switzerland	5 in 10,000	1 in 2,000
United Kingdom	1 in 2,000	1 in 2,000
Japan	<50,000 in population	1 in 2,507
Russian Federation	10 in 100,000	1 in 10,000
Peru	1 in 100,000	1 in 100,000

The field of RDs faces challenges due to limited medical and scientific knowledge, leading to difficulties in diagnosis, access to quality healthcare, and support for patients. There is still no cure for most RDs, however appropriate treatment and medical care help to improve the quality of life and life expectancy. The unique nature of RDs makes their diagnosis complicated in terms of time and cost. Based on the recorded facts by Globalgenes, the average time to get an accurate diagnosis for a rare disease is 6-8 years. These factors lead to underdiagnosis/misdiagnosis or delayed diagnosis of RDs. Still, there is ongoing progress in understanding and treating RDs.

People with RDs face challenges in their quest for a diagnosis, relevant information, access to quality healthcare, and social integration. They are often psychologically, socially, economically, and culturally vulnerable. These difficulties could be addressed through the implementation of appropriate policies [Orphanet, 2012], [Global Genes], [Rahit and Tarailo-Graovac, 2020].

Even when the disease-causing gene is identified, diagnosing and managing these diseases can be challenging due to the variable phenotypes observed, even among individuals with the same causal mutation, such as siblings. Establishing a direct link between genotypes and phenotypes is not always straightforward. This complexity arises from genetic interactions that govern the functional dependency between genes that leads to phenotypic variations. Lots of factors can influence the genetic pathway and give rise to unexpected phenotypes. One of these factors is the presence of genetic modifiers, which are genetic variants that can modify the phenotypic effects of the primary disease-causing variant. These modifiers can either enhance or suppress the severity of the disease condition without being direct causes of the disease. The impact of these modifiers can vary, resulting in significant phenotypic variability and changes in the course of disease. In order to gain a comprehensive understanding of the diverse range of phenotypic outcomes observed in rare disorders, it is crucial to analyze the genomic level differences for each individual [Rahit and Tarailo-Graovac, 2020].

As it is highlighted, the main challenge identified is the difficulty in achieving an accurate diagnosis for patients with RDs. Despite all diagnostic approaches, these patients often face two common issues: delayed diagnosis or a complete lack of diagnosis. In the diagnosis of RDs, traditional techniques such as molecular genetic testing, array-based chromosomal techniques and biochemical tests have commonly

been used. Although these approaches are crucial for the diagnosis of numerous rare genetic conditions, the initial steps still rely on conventional clinical approaches. These include conducting physical examinations, reviewing personal and family medical histories, conducting various laboratory tests, and utilizing imaging studies. It is crucial to gather a comprehensive and detailed family pedigree as it often aids in understanding the inheritance pattern [Ramos-Fuentes et al., 2020].

2.2. Next Generation Sequencing

The era of Next Generation Sequencing (NGS), also known as massively parallel sequencing, has been initiated with the development of Sanger sequencing (also known as capillary or first-generation sequencing). By this technology, the age of genomic sequencing started with a high analytical accuracy in 1977 [Sanger et al., 1977]. However, despite the high sensitivity of Sanger sequencing in mutation identification, it is time-consuming and not cost-effective when performed sequentially for all known genes in the case of genetically heterogeneous complex disorders. NGS technologies have revolutionized genetic testing by allowing a single test for multiple genes. NGS processes millions of sequencing reactions simultaneously, enabling the detection of various genetic variants, from single-nucleotide changes to larger structural variations. Different NGS approaches exist, ranging from targeted gene panels specific to a disease, to whole-exome sequencing (WES) covering all protein-coding genes, and whole-genome sequencing (WGS) analyzing the entire DNA sequence.

With the aim of diagnosis, studies firstly focus on gene panel testing due to its higher throughput and lower costs, but gene panels can detect the causative gene only if it is included in the panel. WES is a broader and more inclusive strategy that is frequently employed when the initial gene panel testing results in a negative outcome, and it has the potential to uncover previously unidentified causative genes. The development of NGS has enabled us to extend our capacity to dig up the abnormalities for Mendelian disorders in both known and unknown disease-causing genes [Cummings et al., 2017]. Especially in the area of genetic RDs, the use of WES and WGS have risen exponentially and have eliminated the need for prior knowledge of specific candidate genes, allowing for the identification of potential disease-causing variants. As a result, WES and WGS have transitioned from being primarily utilized

in research settings to becoming recognized diagnostic tests in routine clinical practice [Ramos-Fuentes et al., 2020].

However, despite their current applications, the success in the molecular diagnostic rates for the patients offered WES or WGS is far from complete, and the success rate is between 25%-50% [Seaby and Ennis, 2020]. More specifically the overall diagnostic rate of WES ranges from 22 to 26%; still, the undefined clinical indications spread the rate from 13 to 44% [Retterer et al., 2016]. Similarly, the overall diagnostic rate of WGS is 34%, while this rate varied by several specifications like disease types or inheritance pattern between 5 to 73% [French et al, 2022], [Meienberg et al, 2016]. A significant portion of patients still face the challenge of not receiving a prognosis or having treatment options available. This incompleteness in success is due to several technical incapacities including the genetic abnormalities in non-coding gene regions -intronic and regulatory regions- and some other challenges such as the variants of unknown significance (VUS). Still, the most challenging restriction of genomic sequencing is that the NGS can currently only interpret the functional and clinical impact of known disease variants. Moreover, although WGS is able to provide information on structural rearrangements, such as copy number variants (CNVs), and tandem-repeat expansions which cannot be described by WES, there are challenges to data interpretation in WGS. For example, it is required to validate the non-coding variants at transcriptional level [Gonorazky et al., 2019]. The challenge lies in analyzing and interpreting the vast amount of data generated by NGS. Additionally, while sequencing costs are decreasing, data analysis and storage are becoming the dominant cost factors due to the expanding volume of sequence data [Seaby and Ennis, 2020].

According to latest estimates, there are approximately 19,000 protein-coding genes in the human genome [Amaral et al., 2023]. In September 2021, in the Online Mendelian Inheritance in Man (OMIM) database, which catalogs information on genetic disorders and their associated genes, there were approximately 7,000 genes linked to RDs [OMIM]. Moreover, although it is hard to specify the exact number of mutations related to RDs, available evidence indicates that an important proportion of disease-causing variants have an impact on protein production and/or processing, even when these variants are situated within non-coding regions and splice site regions [Gonorazky et al., 2019]. Genomic sequencing analyses are capable of detection of variants that differ from reference genome through protein-coding regions and even

non-coding regions. However, the understanding of consequences and the prioritization of a variant from others requires some additional information about function and clinical impacts of the related variant on the gene.

Integrating functional genomic information, such as RNA sequencing (RNA-Seq), is a valuable strategy to improve the interpretation of genetic variants. RNA-Seq allows for the direct observation of changes in gene expression resulting from genetic alterations, thus providing valuable insights into transcriptional perturbations such as splicing anomalies. Previous studies have successfully utilized RNA-Seq to investigate the impact of pathogenic variants identified through DNA sequencing, enabling a deeper comprehension of their functional consequences [Cummings et al., 2017], [Gonorazky et al., 2019]. Recent studies have claimed that RNA-Seq can be used as an integrative approach to genomic sequencing analysis, and they have shown diagnostic utility in specific tissues -even in blood- and in diseases where WES was uninformative for diagnosis for the patients suffering from RDs [Frésard et al., 2019]. As a result, currently RNA-Seq holds a promise as an effective transcriptional method to annotate variants in the coding and non-coding gene regions and to discover novel disease-related genes using the measurements beyond gene expression level, including allele-specific expression (ASE) and alternative splicing. Although the exact number of mutations residing in the non-coding gene regions are unknown, scientific studies have given some evidence to suggest that about 9%-30% of pathogenic variants are found in non-coding regions and affect the expression through aberrant splicing, gene fusions, duplication, short insertions/deletions (indels) or over-expression that may cause loss or gain of functions [Choi et al., 2021]. Accordingly, the systematic detection of such mRNA aberrations, which can reside even in deep-intronic regions, is crucial and can be achieved through RNA-Seq functional analysis [Mertes et al., 2021].

2.3. Alternative Splicing

Almost all protein-coding genes undergo a process called splicing, and the majority of them are alternatively spliced. Alternative splicing plays a crucial role in the expression of genes in organisms with complex cells, enhances the capacity of the human genome to code for proteins and generates a wide range of isoforms from a single gene, effectively maximizing the utilization of genetic information. Various

protein isoforms produced through alternative splicing exhibit distinct biological characteristics, such as protein:protein interaction, subcellular localization, or catalytic ability. The regulation of alternative pre-mRNA splicing governs the functioning of most protein-coding genes. On average, a human protein-coding gene consists of approximately 8.8 exons, with each exon spanning around 145 nucleotides [Tazi et al, 2009] . In contrast, introns, non-coding regions, have an average length of 3365 nucleotides. Additionally, the 5' and 3' untranslated regions (UTRs) measure approximately 770 and 300 nucleotides, respectively. Consequently, this "standard" gene covers a distance of approximately 27 kilobase pairs (kbp). Following the processing of pre-mRNA, the resulting mRNA exported into the cytosol typically comprises 1340 nucleotides of coding sequence, 1070 nucleotides of untranslated regions, and a poly(A) tail [Tazi et al, 2009] . These findings highlight the fact that over 90% of the initial pre-mRNA is eliminated as introns, while only about 10% is retained and spliced together to form exonic sequences. Introns are present in nearly all protein-coding genes, and they are removed through RNA splicing within the cell nucleus during pre-mRNA processing. Moreover, the selection of exons is frequently subject to alternative usage, meaning the cell has the ability to decide whether to exclude a particular segment of pre-mRNA as an intron or include it as an alternative exon in the mature mRNA molecule [Tazi et al, 2009].

The selection of splice sites is finely controlled by a delicate interplay of various factors. The mechanism of pre-mRNA splicing is exceedingly intricate, involving multiple interactions among pre-mRNA molecules, small nuclear ribonucleoproteins generating spliceosome complex, and splicing factor proteins. Given that splice sites exhibit only loosely defined consensus sequences, it becomes crucial to understand how these sites are recognized in the vast genomic sequence background and how their regulation differs. This intricate process is regulated through the interaction of trans-factors, which are proteins and ribonuclear proteins, with cis-factors, which are specific sequence elements present on the pre-mRNA. Additionally, cellular responses to changing environmental conditions play a role in governing this regulation.

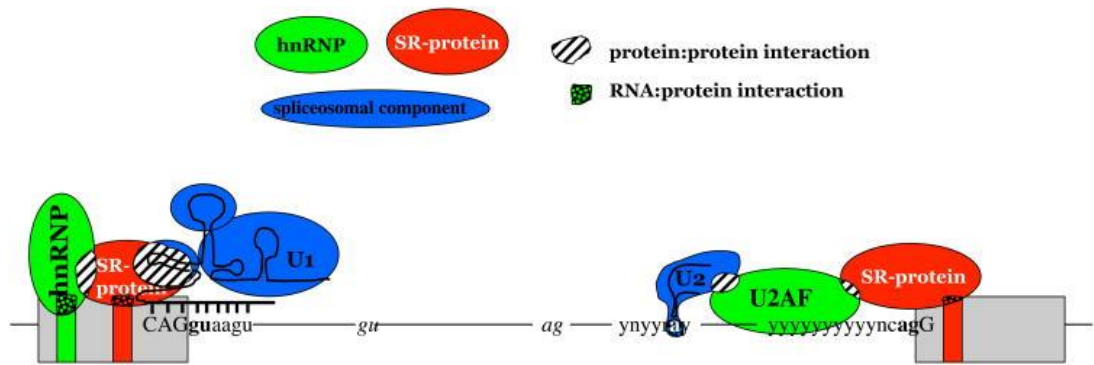


Figure 1: Basic elements of alternative splicing.

The spliceosome, a complex comprising numerous interacting proteins and five small nuclear ribonucleoprotein RNAs (snRNAs) known as U1, U2, U4, U5, and U6, is the primary mediator of the splicing process (Figure 1) [Tazi et al, 2009]. To ensure precise splicing, the spliceosome must identify the boundaries between exons and introns. This recognition is facilitated by consensus sequence elements present at the 5'- and 3'-splice sites of introns, as well as a branch point sequence located near the 3'-end of an intron. During the early stages of spliceosome assembly, U1 forms a base-pairing interaction with the 5'-splice site, while U2 base-pairs with the branch point. Subsequently, a tri-snRNP complex consisting of U4, U5, and U6 associates with the developing spliceosome, leading to the removal of U4 from the complex. This removal enables U6 to replace U1 at the 5'-splice site, resulting in a U6-U2 interaction that brings the 5'-splice site and the branch point in close proximity, facilitating a transesterification step. Through the formation of non-canonical interactions, U5 brings the two exons into close proximity, enabling the second step of splicing, during which the two exons are joined together (Figure 2) [Douglas and Wood, 2011], [Tazi et al, 2009]. Despite this complex recognition pattern provided by the spliceosome complex, the information in splice sites is still not sufficient for regulation, and there are additional regulatory elements increasing the low RNA-binding specificity; (i) Formation of ribonuclear-protein complexes (RNPs: hnRNPs and SR-proteins) (ii) Influence of secondary structures in the pre-mRNA, which can impact splice site selection. These structures play a role in organizing the spatial arrangement of the RNA molecule, affecting the accessibility of splice sites and influencing the splicing outcome. (iii) Regulation of splice site selection by small nucleolar RNAs (snoRNAs), which participate in modifying various RNAs, including pre-mRNAs. They contribute

to the control of splicing decisions. (iv) Dynamic regulation of alternative splicing events by cells, often in response to external signals or stimuli [Tazi et al, 2009].

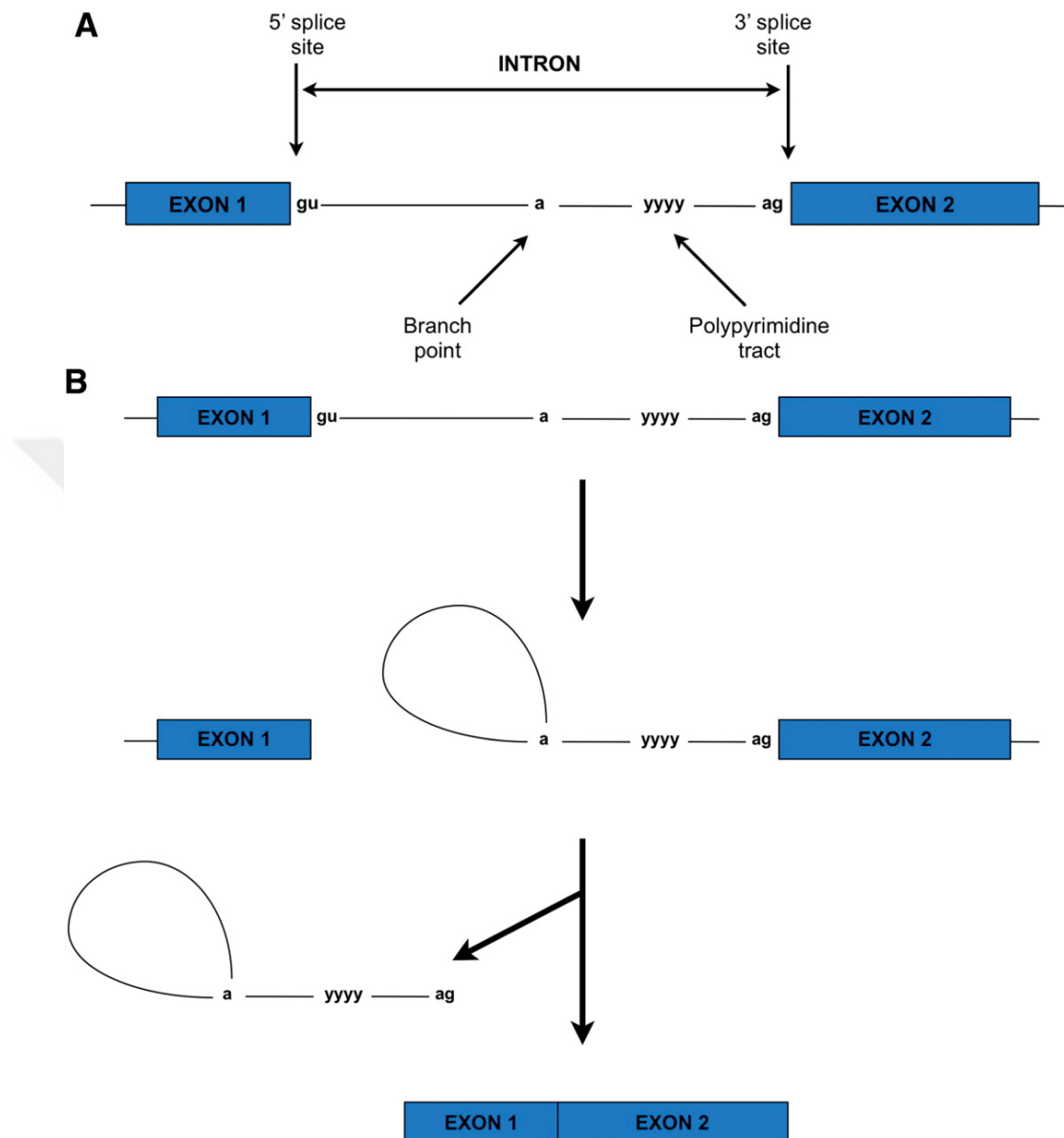


Figure 2: The process of alternative splicing.

Defects in the essential splicing machinery can disrupt vital cellular processes. Each of the regulatory principles governing splicing can be altered in ways that lead to human diseases. Alterations in alternative splicing are frequently associated with human diseases. The effects of mutations on splicing regulation are challenging to predict due to its complexity, but their consequences become evident when they result in disease. Abnormal splicing events can give rise to diseases through various mechanisms, including point mutation/mutations in core splicing elements and

erroneous splice site selection. These mutations disrupt the final protein structure or functionality in several ways, such as altering the ratio of protein isoforms generated by abnormal alternative splicing, activating cryptic splice sites through intronic mutations, generating receptor splicing variants due to single nucleotide polymorphisms, causing exon skipping through mutations in the 5' splice site, affecting exon usage through multiple mutations, or leading to the complete loss of specific splicing factors [Tazi et al, 2009].

Thus, targeting abnormal RNA molecules in diagnostic strategies offers a promising avenue to address defective splicing and potentially treat numerous genetic disorders [Douglas and Wood, 2011]. Aberrant alternative splicing mechanisms, which include skipping, truncation, and elongation of exons as well as intron retention, have also been shown to play an important role in a wide range of pathogenic processes underlying many different diseases. It is reported that 15-30% of the variants related with Mendelian diseases cause outlier splicing [Mertes et.al., 2021]. In addition, there is a pool of unannotated aberrant splice sites -in particular- caused by deep intronic variants. RNA-Seq approach provides the detection of differential splicing and splice outliers by enabling the analysis of intronic regions. Therefore, the processing of RNA-Seq data to identify functional abnormalities is one of the hot topics in recent studies. This approach has a great potential to identify and validate the causal rare-disease genes and functional large-effect of rare variants.

2.4. Bioinformatic Solutions to Discover Aberrant Splicing Events from RNA-Seq Data

NGS technologies enable us to obtain the sequence of desired polynucleotide chains and to analyze the sequence in the light of current information. Still, NGS technologies generate a vast amount of data that is impractical to handle through manual analysis. As a result, the vast quantity of generated data necessitates computational and bioinformatics expertise to effectively handle, analyze, and interpret the extensive NGS datasets. This has led to significant advancements in NGS bioinformatics, made possible by the continuous improvements in computational capacities (hardware), as well as the development of algorithms and software applications/tools (software). These advancements support the entire process, starting

from the initial processing of raw data to the comprehensive analysis and interpretation of specific results within a clinical context. A plethora of bioinformatic applications exists, employing various algorithms to handle NGS data for diverse analytical purposes, such as quality control (QC) and filtering, alignment to reference genome, *de novo* assembly, variant calling, annotation and many other specific analyses. Additionally, RNA-Seq data also offers the analysis of expression and alternative splicing through the utilization of specific bioinformatic tools. The majority of transcriptomic-based bioinformatic analyses in the literature concentrate on differential analysis, wherein two or more distinct sample groups are compared to identify variations in gene expression profiles or alternative splicing patterns between the groups. However, emerging RNA-Seq based tools developed for rare disease studies brought a new methodology to transcriptome studies, including detection of aberrant splicing events such as intron retention and exon skipping [Pereira et al, 2020].

Recently developed RNA-Seq data based bioinformatic tools, -*FRASER*, *SCISSOR* and *dasper*-, offer the detection of aberrant alternative splicing events. Basically, (i) FRASER provides count-based statistical tests for aberrant splicing detection considering latent confounders and deep intronic events [Mertes et.al., 2021], (ii) SCISSOR uses base-resolution read coverage data that provide a rich and comprehensive landscape beyond gene/exon expression values or splice junctions [Choi et al., 2021] and (iii) dasper annotates the regions of junction reads of mRNAs according to the definition of introns determined by Ensembl and detects the exon-exon junction abnormalities based on its machine learning algorithm [Zhang et al., 2021].

2.4.1. FRASER Tool

FRASER is an algorithm designed to identify abnormal splicing patterns in RNA sequencing data [Mertes et.al., 2021] . It accomplishes this by quantifying split reads, which are reads that align to two different genomic locations on the same chromosome strand. FRASER was developed to overcome the limitations of five other methods that also utilize RNA-Seq split reads and are first-line methods for rare condition detection [Cummings et al., 2017], [Kremer et al., 2017], [Frésard et al., 2019], [Jenkinson et al., 2020], [Ferraro et al., 2020]. The shortcomings of these existing methods include:

(i) lack of statistical significance assessment: Some of these methods use a z-score cut-off to detect aberrant splicing events, but this approach does not adequately control for the false discovery rate (FDR), (ii) non-data-driven cut-offs: Setting a z-score cut-off assumes a Gaussian data distribution, but this assumption may be inaccurate when dealing with low split read counts, as seen in splice sites with limited coverage and repressed splice sites, (iii) widespread covariations in split-read-based splicing metrics: Within RNA-Seq data collections, strong covariations in splicing metrics based on split reads can be influenced by various factors, including sex, population structure, technical biases like batch effects, or variations in RNA integrity. These factors are so called confounding factors.

Based on this information, the FRASER algorithm, developed by Mertes et al., addresses these issues and aims to detect aberrant splicing events in RNA-Seq samples through a count-based statistical test. It considers alternative splicing and intron retention by analyzing non-split reads overlapping donor and acceptor splice sites. The algorithm automatically accounts for confounding factors that can impact sensitivity significantly. To optimize its performance, the algorithm fine-tunes its parameters through the use of a denoising autoencoder, enabling it to accurately recall simulated outliers. In contrast to common z-score cut-off methods, FRASER relies on a count distribution and multiple testing correction, leading to a remarkable reduction in the number of false positive calls by two orders of magnitude while only experiencing a minor decrease in sensitivity.

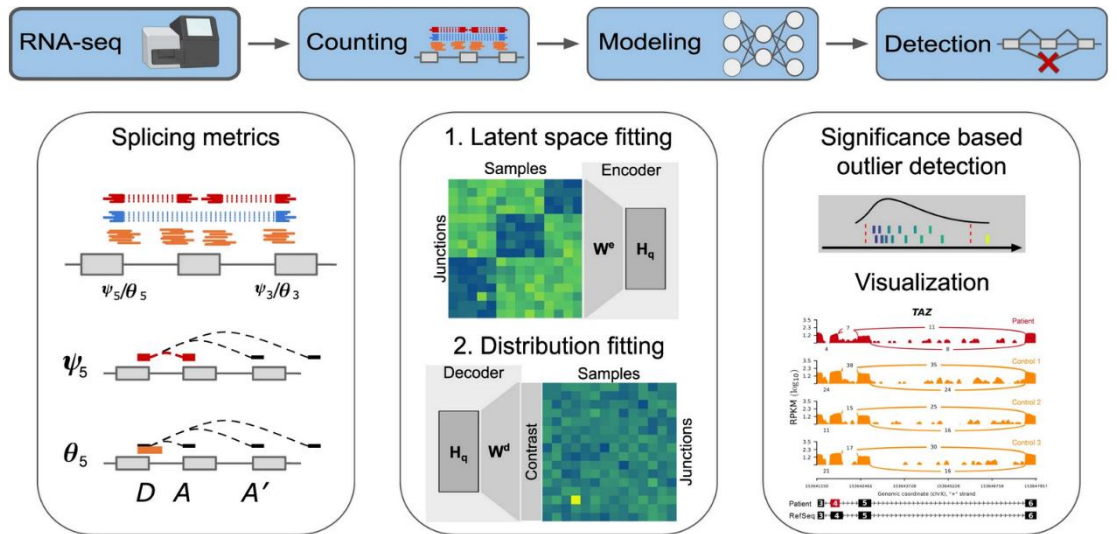


Figure 3: The workflow of the FRASER tool.

The workflow of FRASER starting with the RNA-Seq reads alignment process can be overviewed through a three-step procedure as visualized in Figure 3 [Mertes et.al., 2021].

- 1- Splice Site Map Creation: FRASER generates a splice site map by identifying de novo introns in a gene annotation-free manner, leveraging information from RNA-Seq split reads. Each intron consists of a donor (5' splice site) and an acceptor (3' splice site). For every intron, FRASER calculates important metrics. The first one, denoted as ψ_5 , measures alternative acceptor usage and represents the fraction of split reads originating from a specific intron of interest, compared to all split reads that share the same donor as that particular intron. Similarly, the ψ_3 metric quantifies alternative donor usage for the acceptor. FRASER also considers the donor splicing efficiency metric θ_5 and the acceptor splicing efficiency metric θ_3 , both indicating the fraction of split reads among split and unsplit reads that overlap with a given donor or acceptor. Splicing efficiency metrics (collectively denoted as θ) have lower values in cases of intron retention or impaired splicing.
- 2- Statistical Model Fitting: FRASER fits a statistical model using GTEx project dataset V7 [Lonsdale et al., 2013]. for each splicing metric to account for sample covariations and overdispersed count ratios. In order to estimate latent space, principal component analysis (PCA) is applied to log-transformed splicing metrics. The optimal dimension for the latent space is determined by maximizing the area under the precision-recall curve. This is achieved by independently injecting artificially aberrant values for each splicing metric, a process facilitated by a denoising autoencoder.
- 3- Outlier Detection: Outliers are identified as data points that significantly deviate from the fitted model. To classify observations that notably deviate from their expected values as outliers, developers modeled random deviations using the beta-binomial (BB) distribution. The BB distribution is a distribution for count fractions, parameterized by the expected count ratio and an intra-class correlation parameter, which accounts for variations beyond sampling noise.

The most effective model was achieved through a hybrid approach, wherein the fitting of the latent space and the estimation of expected values were carried out using

a least-squared loss. Simultaneously, the significance of outliers was assessed using the BB distribution.

FRASER stands out due to its distinctive combination of three key features: Firstly, it takes into account non-split reads that overlap with splice sites, enabling the detection of intron retention events. Secondly, it automatically addresses latent confounding factors, thus enhancing its sensitivity. Thirdly, FRASER evaluates statistical significance using a count distribution approach. The implementation of FRASER is available as an R/Bioconductor package. Additionally, FRASER is an integral part of the Detection of RNA-Seq Outlier Pipeline (DROP), which can be accessed on Github [Mertes et.al., 2021]. (<https://github.com/gagneurlab/drop>)

2.4.2. Dasper Tool

Dasper is an innovative method that employs machine learning to integrate disruptions in both exon-exon junction and base pair level coverage data, allowing for the detection of aberrant splicing events [Zhang et al.,2021]. Primarily designed for diagnostics, dasper provides a rank-based report, offering insights into the degree of abnormality for each splicing event. Moreover, it offers a useful feature to visualize candidate genes using sashimi plots for easier interpretation. According to the developers, dasper outperforms existing methods like leafcutter and z-score approaches in accurately detecting pathogenic splicing events. Notably, dasper can identify splicing variants not only at donor and acceptor sites but also deep intronic splicing variants, with enhanced accuracy. By effectively utilizing publicly available GTEx control RNA-Seq datasets, dasper contributes to making RNA-Seq a more cost-effective and standardized solution for the diagnosis of RDs. As an R package on Bioconductor, dasper is readily accessible. (<http://www.bioconductor.org/packages/dasper>)

The methodology employed by dasper for detecting pathogenic splicing events, described in Figure 5 [Zhang et al.,2021], is outlined below:

- 1- Input Data Processing: dasper takes junction read files and BigWig files as inputs. BigWig files are compressed, indexed, binary format files of coverage data. Junction reads are annotated based on whether their start and/or end positions precisely match annotated exon boundaries and whether they match an

existing intron definition from the Ensembl database. Junctions are categorized into various types, including annotated, novel acceptor, novel donor, novel combination, novel exon skip, ambiguous gene, and unannotated, based on their characteristics. Annotated junctions match existing intron definitions, while novel acceptor and novel donor junctions are characterized by having one end overlapping a known exon boundary. On the other hand, novel combination, novel exon skip, and ambiguous gene junctions, are characterized by exhibiting both ends overlapping to known exon boundaries. However, resulting introns for ambiguous gene did not conform to any existing intron definition within Ensembl (Figure 4) [Zhang et al.,2021]. Novel combination junctions are connected to exons associated with multiple transcripts, while novel exon skip junctions are associated with a single transcript only. Ambiguous gene junctions were found to connect exons originating from two different genes. Unannotated junctions lacked any overlap with known exon boundaries at either end.

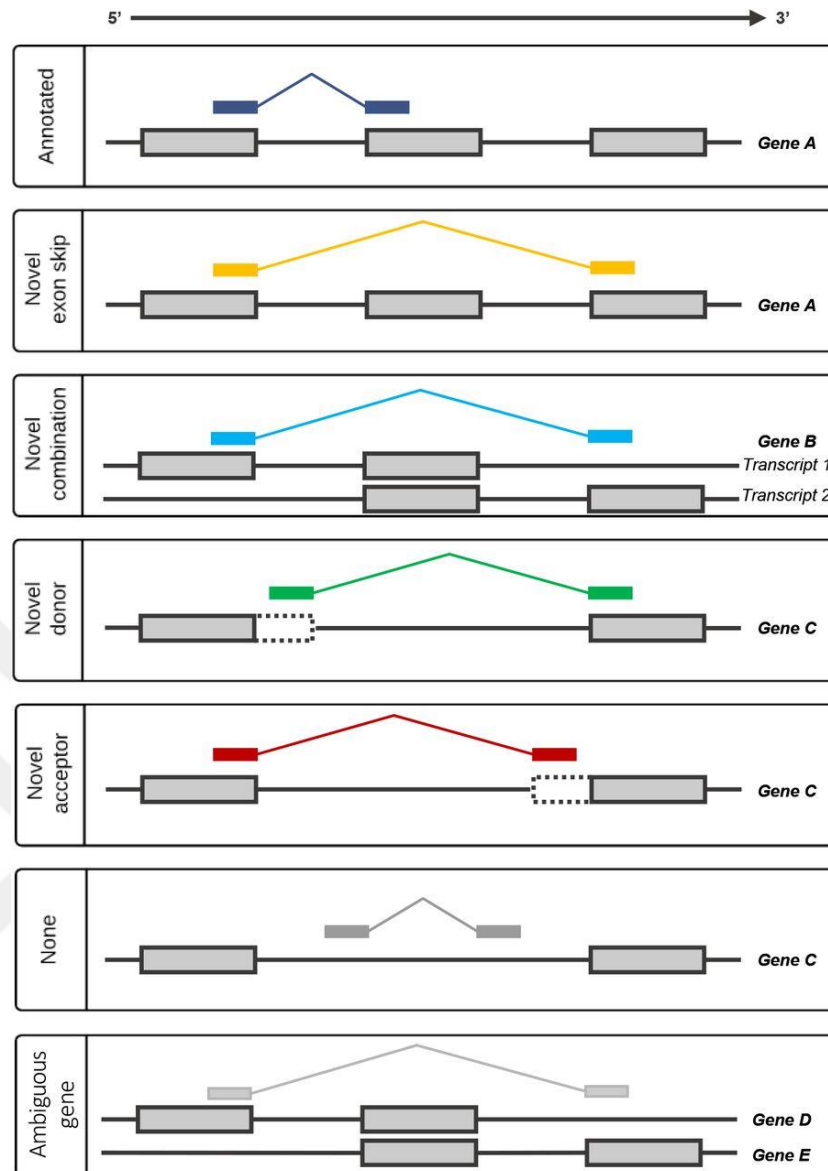


Figure 4: Categorization of junction annotation through dasper.

- 2- Junction Cluster Formation and Normalization: Each junction is associated with a cluster of other junctions that share the same acceptor or donor site. To enable comparison between samples, junction counts are normalized by dividing the counts for each junction by the total counts associated with its corresponding cluster.
- 3- Coverage Information and Normalization: For each junction, three regions of interest are defined: the intron and the two flanking exons. Coverage information is obtained from BigWig files for these regions. The coverage is normalized by dividing the mean coverage across each of the three regions by the total coverage across the exons of the associated gene. z-scores are then calculated to assess

how the junctions and coverage in each sample deviate from the corresponding distribution in control samples. Junctions are split into those with positive z-scores (up-regulated) and negative z-scores (down-regulated). An isolation forest model is fitted separately on the up-regulated and down-regulated junctions using the two z-scores as inputs, resulting in an outlier score per junction.

- 4- Cluster-Level Rank Aggregation and Outlier Detection: Junction-level outlier scores are aggregated to obtain a cluster-level rank in three steps: clusters without at least one up-regulated and one downregulated junction are omitted, then the mean of the up-regulated and downregulated junctions with the highest outlier scores in each cluster is calculated to form the cluster-level outlier score. Finally, within each sample, clusters are ranked based on their cluster-level outlier score, with rank 1 indicating the cluster predicted to be the most aberrant.
- 5- Sashimi Plots: dasper further incorporates coverage information alongside junction counts to enhance the detection of pathogenic splicing events, visualizing candidate genes in the form of sashimi plots.

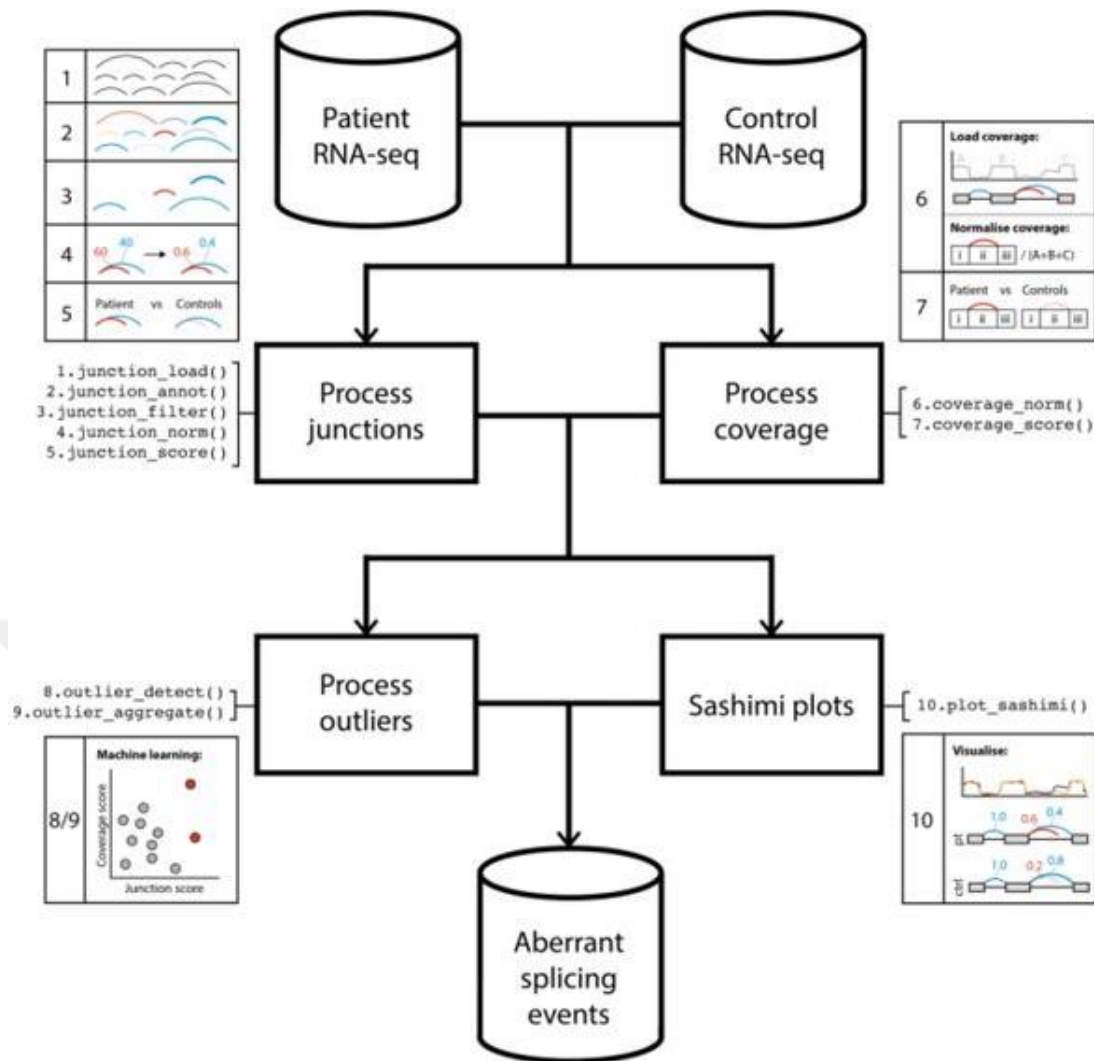


Figure 5: The workflow of dasper tool.

The developers of the dasper tool asserted that all pathogenic splicing events exhibit three distinct characteristics: i) up-regulated novel junctions (referred to as uJs), ii) down-regulated annotated junctions (referred to as DJs), and iii) alterations in coverage within the related exonic or intronic regions. dasper requires each splicing event to possess these features, effectively narrowing down the search space for pathogenic events.

Moreover, dasper incorporates a feature that allows users to download GTEx control data for various clinically accessible tissues, such as fibroblasts, skeletal muscle, whole blood, adipose tissue, and lymphocytes. This functionality enables the running of dasper with just a single patient's RNA-Seq analysis sourced from any of these sample types. Utilizing publicly available control data presents a practical and cost-effective alternative to time-consuming in-house data generation although it may

introduce technical variability between patient and control samples, which could reduce the ability to detect pathogenic splicing events. Indeed, dasper's performance is enhanced when using in-house samples due to reduced technical differences between patient and control data. Nonetheless, this approach may involve increased costs and reduced flexibility, posing barriers to the implementation of RNA-Seq pipelines in diagnostic laboratories. Despite this, dasper is still capable of detecting pathogenic splicing events using the GTEx samples. Furthermore, increasing the number of controls significantly improves the detection of pathogenic splicing events, whether GTEx or in-house control data is employed [Zhang et al.,2021].

2.4.3. SCISSOR Tool

SCISSOR is an abbreviation of “shape changes in selecting sample outliers in RNA-Seq”, and it is a novel framework designed to discover aberrant RNA transcript structures using short read RNA sequencing data. This approach relies on analyzing shape changes in RNA-Seq coverage profiles, offering an unbiased and robust method for detecting changes in mRNA expression, such as alternative splicing, intron retention, de novo splice sites, intra-/intergenic deletions, and alternative transcription start/termination sites (ATS/ATT). Unlike existing methods that focus on gene/exon expression values or splice junctions, SCISSOR leverages base-resolution read coverage data to provide a comprehensive landscape. It transforms and normalizes base-level RNA sequencing coverage data in a transcript-independent manner, followed by a statistical framework for analysis. SCISSOR is capable of detecting structural variations or differential coverage patterns across RNA-Seq cohorts without assuming any specific mechanism driving the coverage variation.

The core principle of SCISSOR is based on the alignment-based representation of the base-level coverage at each specific genomic locus of interest. The base-level read-depth offers a straightforward visualization of the coverage shape of genes, with the y-axis reflecting per-base expression and the x-axis indicating the genomic position (Figure 6) [Choi et al., 2021]. The fundamental assumption underlying SCISSOR is that samples from the same tissue type should exhibit consistent and uniform coverage. Any deviations from this uniform pattern, observed as diverse shape changes in the coverage profile, may indicate abnormal expression patterns resulting from various mutational mechanisms, such as exon skipping, intron retention, gene fusion, deletion,

or internal tandem duplication. As a systematic discovery tool for a wide range of genetic aberrations, SCISSOR identifies shape changes where the aligned coverage significantly differs from the majority of samples, enabling the detection of potentially impactful structural alterations in RNA transcripts.

The normalized coverage accentuates the aberrant features by eliminating the common structure shared by the majority of samples. The scores obtained by projecting the normalized coverage matrix onto each of the most outlying directions corresponding to each example are shown with the kernel density estimates. The colored point in each scatter plot indicates the outlyingness score corresponding to each example.

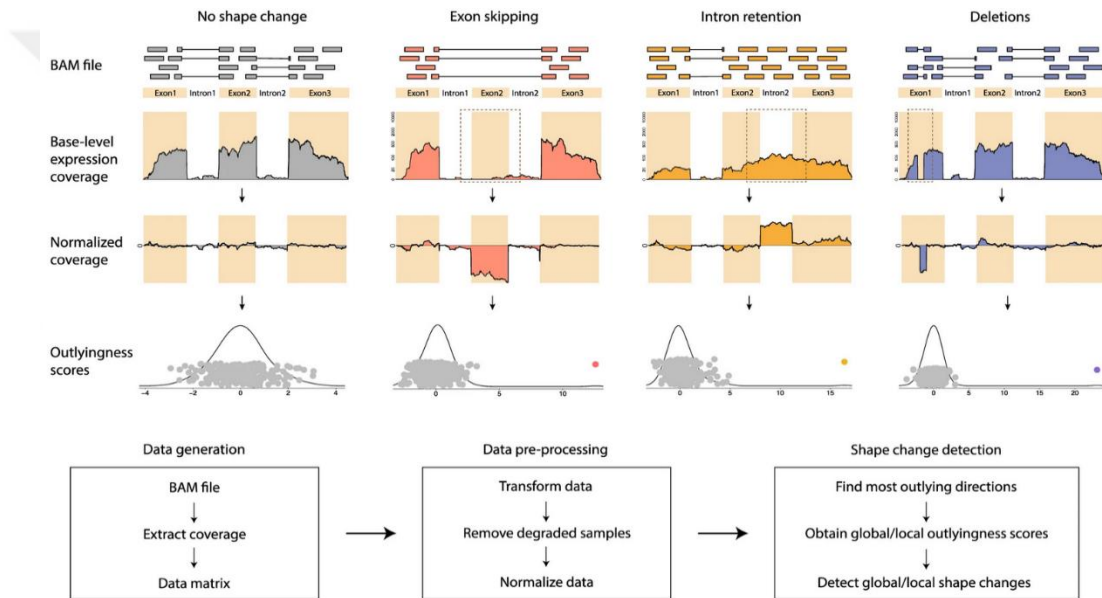


Figure 6: Main steps in the SCISSOR pipeline.

The primary steps in the SCISSOR pipeline are as follows:

- 1- Extraction of RNA-Seq Coverage and Data Matrix Construction: For each subject, RNA-Seq coverage data is extracted from a BAM file for a specific gene. These individual coverage datasets are combined to form a data matrix containing coverage information from all subjects.
- 2- Preprocessing of Data Matrix: The data matrix undergoes pre-processing, which includes transformation, exclusion of degraded samples, and normalization. The normalization and variance stabilization steps are employed to generate more symmetrical distributions of the SCISSOR test statistic. A further preprocessing

step generates the residual per-base expression for each sample after subtracting a model representing the gene's shape across all samples. This residual structure is referred to as "latent gene expression" since it is inferred after subtracting the model from the normalized data.

- 3- Application of Statistical Procedure: The statistical procedure is performed on the normalized data, resulting in the identification of the most outlying directions, outlyingness statistics, and detected shape outliers. The curve following normalization can be envisioned as a single point in a high-dimensional space, where each base position serves as a dimension and the height of latent gene expression represents the value at that position. A latent variable is employed to represent an underlying abnormal gene expression trajectory, which corresponds to an outlying direction in a high-dimensional space. This trajectory is then examined to identify potential outliers. An outlier case with shape changes can manifest as a data point that strongly aligns with one or multiple abnormal trajectories, allowing for the modeling of complex structural variations. Following the extraction of latent gene expression associated with abnormal sequencing coverage, SCISSOR utilizes a robust approach to quantify the level of abnormality in order to identify cases with shape changes (Figure 6) [Choi et al., 2021]. Since the focus is on identifying outlying/abnormal structures, SCISSOR searches for the optimal one-dimensional projection in which each sample achieves the maximum deviation from all other samples, resulting in the determination of the "most outlying direction" (MOD). For each gene being considered, the resulting statistic represents the outlyingness score for each sample, with higher values indicating more significant deviations from other samples in the dataset (Figure 6) [Choi et al., 2021]. The distribution of these outlyingness scores is then observed and can be used to identify and rank outliers, as well as to model statistical significance. For each identified outlier, SCISSOR produces the MOD that characterizes the abnormalities of the corresponding outlier. This MOD information can be used to reconstruct the latent space of underlying outlier directions, providing valuable insights into the structural variations detected.

The output of SCISSOR yields a high-dimensional object that is suitable for unsupervised screening of structural alterations in RNA-Seq cohorts, without making

strong assumptions about the underlying mutational mechanisms responsible for these abnormalities. This unique capability allows SCISSOR to independently rediscover known variants, such as splice site mutations in tumor suppressor genes. Additionally, SCISSOR has the potential to unveil novel variants that may have been previously overlooked or challenging to detect using existing methods, including recurrent alternate transcription start sites and recurrent complex deletions in 3' UTRs. As an open-source R package, SCISSOR is freely available on Github [Choi et al., 2021]. (<https://github.com/hyochoi/SCISSOR>)

2.5. Cystic Fibrosis

Cystic fibrosis (CF) is a hereditary disorder that affects multiple organs and is inherited in an autosomal recessive manner. It arises from mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) protein, which is responsible for chloride ion transport and is activated by cyclic adenosine monophosphate (cAMP). This protein is expressed in various tissues, including the lungs, pancreas, and others [Phuan et al., 2021]. Over 2000 different variants of the CFTR gene have been documented in scientific literature, with more than 300 of them associated with CF of varying severity. While the F508del mutation is the most prevalent in Europe and the US, the distribution of CFTR mutations exhibits significant regional variability [Phuan et al., 2021]. Roughly 70% of Caucasian individuals with CF in the United States carry the F508del mutation, which is associated with severe clinical symptoms when both copies of the CFTR gene bear this mutation. Globally, this genetic variant makes up approximately 60% of all CF genetic profiles, although its prevalence varies significantly across different populations and geographical regions, ranging from 27% to 87.5% [Gallati, 2014]. Nevertheless, the CFTR gene displays a wide range of mutations. These mutations can be categorized into six distinct classes, each corresponding to specific types of CFTR dysfunction, such as problems with protein translation, cellular processing, or the functioning of the CFTR channel. Missense mutations, involving the substitution of a single amino acid, make up 38.74% of CFTR mutations, while frameshift mutations, which involve insertions or deletions, account for 16.25%. Splicing mutations, which result in incorrect intron splicing, contribute to 10.93% of the cases, and nonsense mutations, leading to premature termination codons, are responsible for 8.41% of all documented CFTR mutations globally [Chen

et al., 2021]. Notably, there is a significant number of rare CFTR mutations, some of which give rise to the creation of cryptic splice sites. While CF is classified as a single gene disorder, there has been substantial evidence representing a wide range of phenotypic diversity associated with the disease. The theory suggesting that the diversification in symptoms might be attributed to diverse genetic variations, known as allelic heterogeneity, has prompted extensive research into the CFTR gene. This research has unveiled numerous pathways and their associated components that have a strong biological basis for influencing the course of CF. The identification of additional gene alleles, often referred to as "modifier genes," has played a significant role in explanation of why CF patients with the same mutation can display distinct clinical phenotype [Guillot et al., 2014]. Cystic fibrosis, classified as a rare lung disease, typically manifests in early childhood and is characterized by an average life expectancy of around 40 years [Phuan et al., 2021] , [Schaaf et al., 2020]. Clinical features commonly include exocrine pancreatic insufficiency and bronchiectasis, accompanied by chronic airway infections that lead to respiratory failure and premature death. In the pursuit of effective disease management, a class of drugs known as CFTR modulators has been developed to restore proper CFTR protein function. These modulators have demonstrated benefits in approximately 80% of European patients with cystic fibrosis [Fajac and Sermet, 2021]. However, for many rare CFTR mutations, there are currently no approved CFTR modulators available for treatment [Phuan et al., 2021].

Here, a publicly available dataset containing RNA-Seq blood data from people suffering from CF was used. These samples are prepared and sequenced in a study of *Kopp, et.al* [Kopp et al., 2020]. This dataset contains the samples of patients with homozygous p. Phe508del (F508del) mutation obtained before and after the treatment with the CFTR modulator lumacaftor/ivacaftor. *Kopp, et.al* have shown that while some patients responded to the treatment of CFTR modulator lumacaftor/ivacaftor, some others did not respond. They performed differentially expressed gene (DEG) analysis and reported DEG's between responders and non-responders. In this study it was analyzed to unveil whether the unresponsiveness of some CF patients to lumacaftor/ivacaftor treatment was caused by any abnormal splicing event. The detection of abnormal splicing analysis was performed using three outlier splicing detection tools developed recently: FRASER [Mertes et al., 2021], dasper [Zhang et al., 2021] and SCISSOR [Choi et al., 2021].

3. MATERIAL AND METHODS

3.1. Samples

Here, the total RNA-Seq data of CF patients were obtained from NCBI Gene Expression Omnibus platform as a publicly available dataset. The dataset consists of 60 samples from 40 people: 20 non-CF individuals, 20 CF patients with pre-lumacaftor/ivacaftor treatment, which are referred to as “pre-drug samples”, and 20 CF patients with post-lumacaftor/ivacaftor treatment, which are referred to as “post-drug samples”. The sample cohort setup is visualized in Figure 7. The metadata of the samples is summarized in Table 2 [Kopp et al., 2020]. 20 CF patients were previously categorized in the study of *Kopp, et.al* into 2 groups according to their responsiveness to lumacaftor/ivacaftor medication: drug-responders and drug non-responders. The categorization of the data obtained from the CF people based on the drug response is represented in Table 3. All samples were obtained from Sequence Read Archive (SRA) using PRJNA512474 accession number.

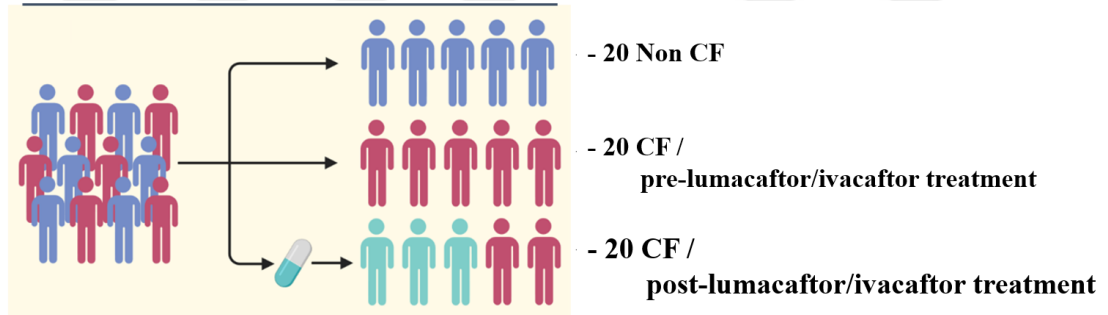


Figure 7: Sample cohort setup of PRJNA512474 project with 60 samples from 40 people.

Additionally, in order to create a control cohort with a large number of blood samples, the data from the Genotype-Tissue Expression (GTEx) Consortium project were included. These samples were filtered using the sample metadata features of Kopp, [et.al](#) study, such as; tissue type = ‘blood’, RNA integrity number (RIN) ≥ 7 , age < 50 and body mass index (BMI) < 30 . Additionally, GTEx samples that were reported to contain outlier events in the study of Ferraro et al. [Ferraro et al., 2020] were removed from the dataset. Finally, 136 GTEx blood samples remained as the control

samples. Consequently, the control cohort used in this thesis study contained 20 non-CF samples from PRJNA512474 dataset and 136 GTEx samples.

Table 2: Sample metadata.

	CF	Non-CF	P value
Subjects (n)	20	20	
Female	55%	65%	0.53
Age (years)	21.6 ± 9.5	37.9 ± 12.1	0.001
Caucasian	100 %	95 %	0.33
Phe508del	100%	-	
BMI (kg/m ²)	21.2 ± 3.4	not available	

Table 3: Drug response of the CF patients.

Patient ID	Pre-Drug Samples	Post-Drug Samples	Drug Response
001	SAMN10678840	SAMN10678871	Responder
002	SAMN10678866	SAMN10678865	Non-Responder
004,019	SAMN10678863	SAMN10678852	Non-Responder
006	SAMN10678849	SAMN10678827	Responder
007	SAMN10678846	SAMN10678848	Responder
009	SAMN10678843	SAMN10678845	Responder
010	SAMN10678839	SAMN10678842	Non-Responder
012	SAMN10678877	SAMN10678879	Non-Responder
013	SAMN10678874	SAMN10678876	Responder
014	SAMN10678870	SAMN10678873	Responder
015	SAMN10678860	SAMN10678859	Non-Responder
016	SAMN10678858	SAMN10678857	Responder
017	SAMN10678856	SAMN10678855	Responder
018	SAMN10678854	SAMN10678853	Responder
020	SAMN10678851	SAMN10678836	Responder
021	SAMN10678835	SAMN10678834	Responder
022	SAMN10678833	SAMN10678832	Non-Responder

The continuation of Table 4

024	SAMN10678831	SAMN10678830	Non-Responder
025	SAMN10678829	SAMN10678828	Responder
026	SAMN10678838	SAMN10678837	Non-Responder

3.2. Bioinformatic Analysis

All the fastq files of the samples from project PRJNA512474 were subjected to the QC analysis using the FastQC (v0.11.7) [Andrews, 2010] and Trimmomatic (v0.39) [Bolger et al., 2014] tools. Using the FastQC tool, a QC check was applied on the fastq files, which contain sequences from total transcriptome data. Then, low-quality sequences were filtered out such that the sequences with average Phred score below 20 were trimmed and the sequences with length shorter than 20 bp were removed through the Trimmomatic tool. After the pre-alignment QC analysis, clean fastq files were aligned to the GRCh38 genome fasta file, which was obtained from the GENCODE (release v24) database, using the STAR aligner (v2.7.10a) [Dobin et al., 2013] in two-pass mode. The genomic features were identified from the GENCODE GRCh38 v24 annotation .gft file. STAR aligner provides fast and accurate alignment for RNA-Seq data in a splice aware manner using the .gtf file against genomic reference sequence. The output of the alignment process is a .bam (Binary Alignment Map) file, a SJ.out.tab file and files representing alignment statistics. Additionally, bigwig files were generated from .bam files using bam_to_bigwig function of the megadept R package. Also, .bam files containing aligned sequences were analyzed with Qualimap2 (v2.2.2) [Okonechnikov et al., 2016] tool to apply post-alignment QC and calculate the proportions of reads that were aligned to different genomic regions of genomic reference sequence. Resulting .bam files were then used as input data in expression and splicing analysis. Also, (i) .SJ.out.tab files containing splice junction reads detected during the alignment step and (ii) bigwig files containing compressed, indexed, binary format files of coverage data, which was derived from .bam files, were used as inputs to identify abnormal splicing events.

In order to determine the differentially expressed genes (DEGs) between the drug-responder samples and drug non-responder samples, an expression analysis was set up for the pre-drug samples and post-drug samples separately. Prior to differential

expression (DE) analysis, reads were quantified at transcript level using Kallisto (v0.46.1) [Bray et al., 2016]. To this end, reference transcriptome index file was generated through Kallisto-index using GRCh38 transcriptome fasta file and gtf file, which was obtained from the GENCODE (release v24) database. Then, the reads in trimmed fastq files were quantified at transcript level in raw counts format by performing pseudoalignment against transcriptomic reference sequence. Transcript level counts were converted to gene level expression counts using the tximport R package (v1.26.0) [Soneson et al., 2015]. Following the quantification step, a table containing gene level raw counts of all samples in drug responder and drug non-responder groups was generated and used in differential expression analysis as the input file. Genes with a mean expression value of 0 in all samples were removed from the table. Differential expression analysis was performed on the data from CF patients using DeSeq2 R package (v1.38.0) [Love et al., 2014]. DeSeq2 normalizes the raw counts using the median-of-ratios method before performing differential analysis. In the median-of-ratios method, which serves as a between-sample normalization technique, sequencing depth and RNA composition are taken into account, but gene length is not considered. After normalization and scaling steps, principal component analysis (PCA) was performed on all CF samples in order to identify the presence of outliers. Detected outliers were removed from the count table and raw counts of remaining samples were again normalized and scaled. With the differential expression analysis step, p-values, Benjamini-Hochberg corrected p-values (adjusted p-values) and fold changes (FC) were calculated for each gene using DeSeq2. The DEGs having statistical significance were filtered using the p-value cut-off of ≤ 0.01 and FC cutoff of $\leq 1/1.2$ or ≥ 1.2 . Protein-protein interactions (PPI) between the DEGs were identified using the STRING database (v12.0), by also including neighbor proteins of the DEGs up to third neighbors [Szklarczyk et al., 2023]. Active interaction sources were chosen as text mining, experiments, databases, co-expression, neighborhood, gene fusion, and co-occurrence.

Splicing analyses were performed using recently developed three splicing tools for pre-drug samples and post-drug samples separately: FRASER v1.10.0 [Mertes et al., 2021], dasper v1.5.2 [Zhang et al., 2021] and SCISSOR v1.0.0 [Choi et al., 2021]. While the dasper and SCISSOR tools were implemented using *Bioconductor* v3.16 R package, the FRASER tool was executed using the DROP workflow (v1.3.3) [Yépez et al., 2021].

Prior to the analysis, FRASER counts the junction reads using the HTSeq tool and saves the count data as an external output. From the FRASER results, the genes with $p\text{-value} < 0.05$ were selected for downstream analysis. In the dasper analysis, bigwig files and splice junction files were used as the input files, and the dasper analysis was executed. The genes with mean outlier score < 0 were filtered out from the dasper results.

Fisher's Exact Test was applied to the results of the FRASER and dasper tools in order to determine the genes with abnormal splicing between drug responder and drug non-responder groups, and this was repeated for the pre-drug samples and post-drug samples separately. In this study this step was termed differential splicing detection analysis. Prior to comparing the results with Fisher's Exact test, PCA was performed to filter out the samples reflecting a very distinct abnormally spliced gene profile compared to the rest of the samples in a given group. A binary matrix was constructed for PCA where columns represent samples, rows represent genes, and each element signifies the absence or presence of an abnormal splicing event for the respective gene and sample combination, based on FRASER or dasper analysis results. After removing outlier samples detected from the PCA plot, Fisher's Exact Test was applied for each gene using the number of samples with abnormal splicing events in Drug Responder and Drug Non-Responder groups. The genes displaying abnormal splicing events were selected based on a $p\text{-value}$ cut-off of ≤ 0.05 and odds-ratio cut-off of < 1 between drug responder and drug non-responder groups. Odds-ratio filtration was applied to determine the genes with abnormal events that were found in Drug Non-Responder groups but not in Drug Responder. The genes obeying those thresholds were then analyzed with the SCISSOR tool.

Aberrant splicing analysis detection using the SCISSOR tool was performed individually for the genes that were identified to have abnormal splicing events by the dasper and FRASER tools. SCISSOR identifies samples exhibiting splicing abnormalities for the specified gene. Therefore, in this study, the SCISSOR tool was iteratively executed within a loop, conducting analyses for each gene on the list across the cohort's samples. Subsequently, Fisher's Exact Test was applied in a similar manner to that of FRASER and dasper with $p\text{-value}$ cut-off of ≤ 0.05 and odds-ratio cut-off of < 1 separately for the pre-drug and post-drug groups.

Ultimately, a comparative analysis was performed on all generated lists from the differential expression analysis, which included information about genes exhibiting aberrant splicing events identified through FRASER, dasper, and SCISSOR.

There were some technical and methodological problems during the reproduction of three tools. In FRASER tool execution, the definition of analysis cohort set up was unclear. And DROP workflow configuration was not clearly explained in documentation. By contacting the developers of the FRASER tool, required information was gained through the github pages of the tools (<https://github.com/c-mertes/FRASER/issues/26>), <https://github.com/gagneurlab/drop/issues/426>). Similarly, one of the important function commands was broken in SCISSOR pipeline and reported to corresponding authors via github page (<https://github.com/hyochoi/SCISSOR/issues/3>). Improvements were made by developers, and these enhancements were incorporated into pipelines to finalize the analyses.

4. RESULTS

4.1. QC Analysis

In the process of QC, pre-alignment metrics of raw total RNA-Seq reads were calculated through FastQC tool from .fastq files, and post-alignment metrics of aligned total RNA-Seq reads were generated through STAR and Qualimap2 tools from .bam files. All QC metrics calculated by all QC tools across all samples were summarized using MultiQC tool. The results showed that (I) the median of the number of total reads was 49,605,135 ($\approx 50\text{M}$), (ii) the median of the rate of uniquely aligned reads was 75%, and (iii) the median of the rate of reads aligned to exonic regions was 80% across all samples. This result indicated that (i) Alignment step was achieved successfully for all samples, (ii) Despite the fact that RNA-Seq data contained intronic reads since it was sequenced as total RNA-Seq, most of the reads were aligned to exonic regions, indicating that most of the reads were coming from mature RNA regions, and (iii) Although the total number of reads was sufficient to identify the differentially expressed genes, it was slightly lower than the suggested value, 75M, in order to determine abnormal splicing events. (Ref, Illumina, Lexogen).

4.2. Aberrant Splicing Analysis through FRASER and dasper tools

Splicing analyses were performed using three recently developed tools; FRASER, dasper and SCISSOR. The developers of the dasper tool asserted that conducting analyses using samples generated through different pipelines does not significantly impact the results. However, the developers of SCISSOR and FRASER did not address this aspect in their discussions. From this point, an analysis was carried out by using FRASER to assess how the utilization of two distinct cohorts, comprised of control samples aligned with two different pipelines could potentially lead to different results. This analysis was performed by using the FRASER tool since it has the ability of removing confounding factors. Two analysis groups were established for this purpose. In the first group, the control samples were available in the aligned

format, where GENCODE v26 was used for the alignment. In the second group, on the other hand, they were aligned using GENCODE v38 since this is the latest version. These groups were respectively labeled as fraser_v26 and fraser_v38. As a result, when applying a p-value cut-off of <0.05 , FRASER identified 95,726 outlier splice junctions in 9,681 genes for fraser_v38 and reported 118,680 outlier splice junctions in 11,255 genes for fraser_v26 when the results of all the pre- and post-drug samples are combined. Among the analysis groups, 56,734 outlier splice junctions in 8,362 genes were found to be common, accounting for 59% of fraser_v38 and 47% of fraser_v26 results at the junction level, and 86% of fraser_v38 and 75% of fraser_v26 results at the gene level. However, when the p-value cut-off was set at 0.01, the ratio of commonly detected abnormal junctions between the groups decreased. Specifically, for fraser_v38, 13,910 junctions in 4,898 genes were detected, while fraser_v26 showed 19,140 junctions in 5,642 genes. Of these, 6,882 junctions in 3,201 genes were found to be common, representing 49% of fraser_v38 and 35% of fraser_v26 results at the junction level, and 65% of fraser_v38 and 56% of fraser_v26 results at the gene level. The results show that the utilization of the samples processed through the different pipelines significantly diversifies the outcomes. Therefore, in the downstream steps, FRASER analysis results from fraser_v26 group were used since the data from the cystic fibrosis dataset (PRJNA512474) were aligned to GENCODE_v24, a version much similar to v26 than v38. On the other hand, dasper analysis with GENCODE_v24 resulted in a remarkable discovery of 505,754 outlier splice junctions for a total of 10,730 genes in the pre- and post-drug samples.

4.3. Outlier Sample Removal Prior to Differential Expression and Differential Splicing Analysis

According to the results of PCA applied to the normalized expression data prior to the differential expression analysis, samples with the ID of 016_SAMN10678858, 004_SAMN10678863, 026_SAMN10678838 were detected as outliers in the pre-drug group (Figure 8a) while 018_SAMN10678853, 001_SAMN10678871, 002_SAMN10678865 were detected as outliers in the post-drug group (Figure 8b).

On the other hand, PCA was applied to the results of dasper and FRASER analyses separately, by representing the results as a binary matrix that shows the

absence or presence of an abnormal splicing event for the respective gene and sample combination PCA detected the sample with the ID of 026_SAMN10678838 as an outlier in the pre-drug group in both dasper and FRASER analyses (Figure 9) while the sample with the ID of 018_SAMN10678853 was detected as an outlier in the post-drug group in both analyses (Figure 10).

At the same time, PCA analysis was performed to examine the expression profiles between CF samples -as pre and post drug groups- and control samples. This analysis also indicated that the samples with the ID of 026_SAMN10678838, 024_SAMN10678831 and 016_SAMN10678858 in the pre-drug group and the samples with the ID of 002_SAMN10678865 in the post-drug group were outliers (Figure 11).

As a result, samples with the ID of 026_SAMN10678838 and 018_SAMN10678853 were detected as outliers by all three PCA analyses. These samples were removed from both the pre-drug group and the post-drug group. Removal of the samples decreased the detected number of outlier splice junctions to 104,865 for a total of 11,057 genes in the FRASER results and 453,505 for a total of 10,428 genes in the dasper results.

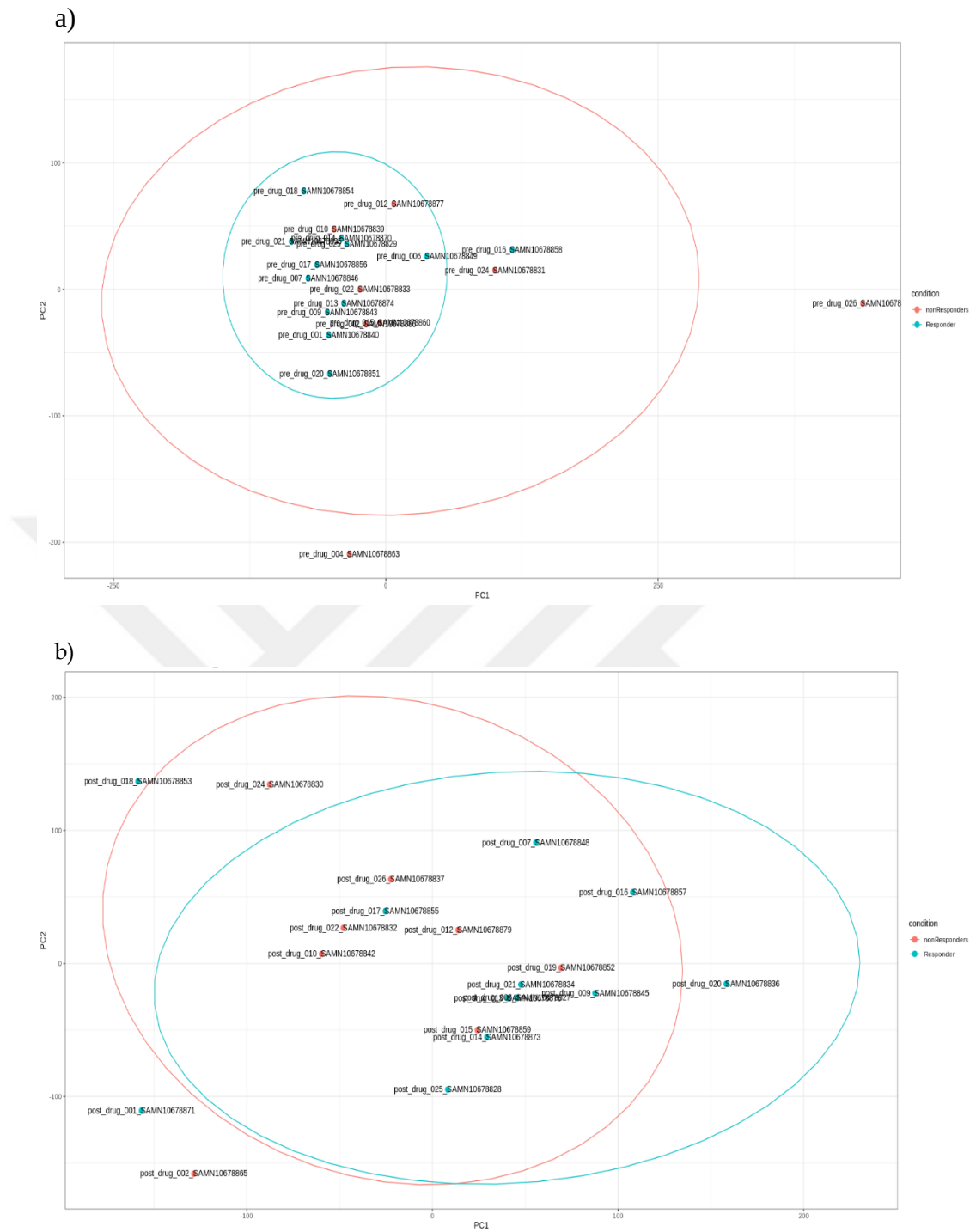


Figure 8: PCA Graph applied prior to the differential expression analysis; a) PCA graph for the pre-drug group, b) PCA graph for the post-drug group.

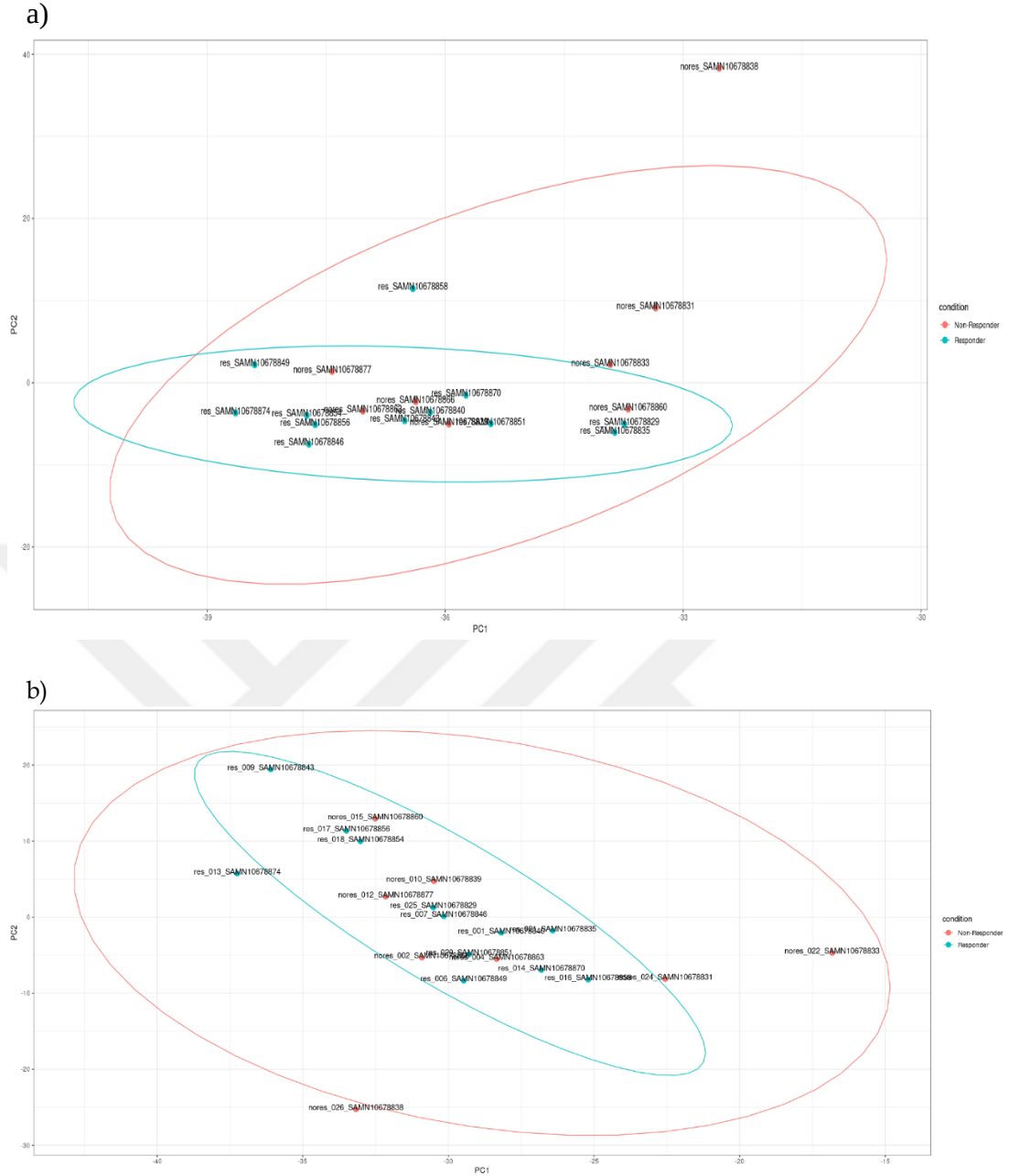


Figure 9: PCA graphs of the pre-drug group applied prior to the differential splicing analysis; a) PCA graph for the dasper results, b) PCA graph for the FRASER results.

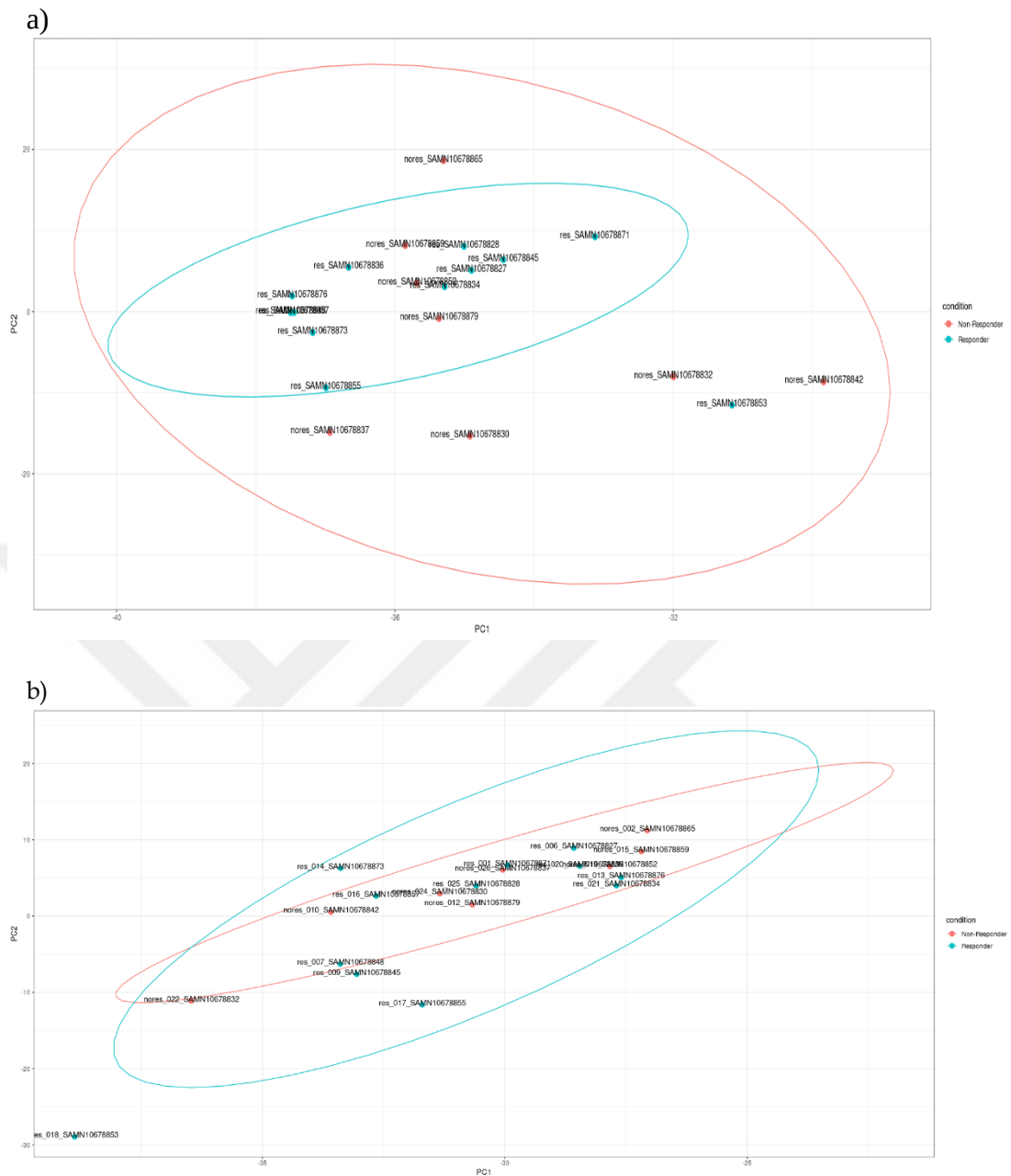


Figure 10: PCA graphs of the post-drug group applied prior to the differential splicing analysis; a) PCA graph for the dasper results, b) PCA graph for the FRASER results.

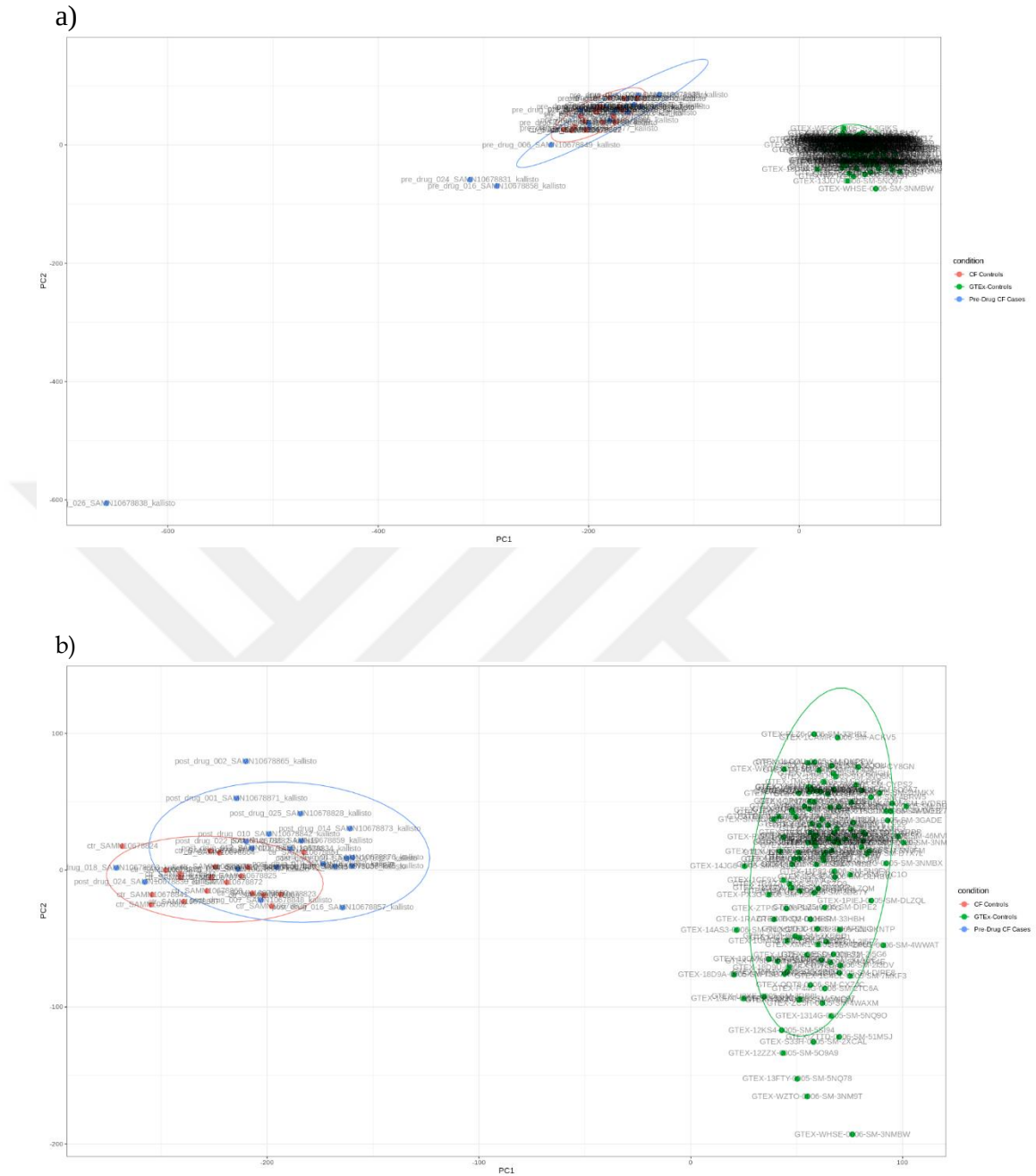


Figure 11: PCA graph to determine the separation of CF samples and control samples in terms of expression profiles. Pink points indicate the CF samples, green points indicate GTEx control samples and blue points indicate CF control samples. a) PCA graph for the pre-drug group, b) PCA graph for the post-drug group.

4.4. Detecting the Genes with Abnormal Splicing Events in Non-Responders through Differential Splicing Analysis

The focus in this thesis study is the splicing abnormality that drives resistance to the drug effect. Therefore, we identified the genes that possess splicing abnormalities in the drug non-responders but no such abnormality in the drug responders. To this end, the Fisher's Exact Test was applied on dasper and FRASER results separately for pre-drug and post-drug groups to identify the correlation of genes with abnormal splicing between the drug responder and the drug non-responder groups. After applying the p-value and Odds ratio cut-offs to the Fisher's Exact Test results, the number of genes with detected splicing abnormality was:

- (i) 100 for the pre-drug samples analyzed with dasper
- (ii) 101 for the post-drug samples analyzed with dasper
- (iii) 88 for the pre-drug samples analyzed with FRASER and
- (iv) 170 for the post-drug samples analyzed with FRASER.

Table 4 gives the gene symbols identified for each of these four comparisons.

Table 4: The genes with significant splicing abnormalities for non-responders but not responders.

dasper analysis for pre-drug group	ABCC3, ACHE, ADGRE3, ARHGDIB, yBCAP29, BCAT1, BSCL2, C9orf78, CCDC149, CCDC15, CCL3, CD247, CD37, CDA, CEP63, CHCHD6, CSNK2A2, DXO, EIF2D, ERP29, EXD3, FCRL2, FUCA2, GCM1, GMEB2, GMPPA, HERC2P10, HEXA, HGST, HLA-DMA, HOTAIRM1, IGF2BP2, IGKV1-5, IVNS1ABP, KCNN4, KIFC2, KLHL7, LAIR1, LARP7, LMBR1, LPP, LY6G5C, MEI1, MLLT1, MRPL2, NBP19, NCOA4, NKG7, NPC2, NUDT2, NUP210, ORC3, OXLD1, PCSK6, PI4KAP2, PIGG, PIK3R5, PIP4K2A, PPOX, PPP2R5E, PPP6R1, PRR5, PRR5-ARHGAP8, PTGES3, PYGM, RAB11FIP3, RANGAP1, RASSF7, RECQL5, RNH1, RPH3A, RPL24, SFSWAP, SKA3, SLC12A7, SLC25A39, SLC28A3, SLC35F6, SMIM1, STRAP, TAGLN2, TAOK3, TBC1D22A, TFEC, TMEM120A, TMEM134, TMEM53, TNFRSF18, TRABD, TRAJ56, TRBV5-5, TRPV2, TVP23C, TVP23C-CDRT4, UCK1, VAC14, VAMP8, ZCWPW1, ZNF80, ZZZ3
dasper analysis for post-drug group	ABCB4, ACACA, ACACB, ADAM22, AHI1, AK6, APEX2, ARPC1B, AS3MT, AUH, BAIAP2, BORCS7-ASMT, BRD2, C1orf159, CCDC12, CCNI, CDC42, CLC, CLTCL1, CPSF6, CTSS, CYP19A1, DDB2, DEPDC5, DJA4, DUS1L, DVL3, EIF2D, EPOR, ERCC8, EXD3, FAM89B, FDFT1, FUOM, GHRL, GMPPA, HCFC1, HLA-P, HORMAD1, HPN, IFR2, IGHV4-34, IGKV3D-11, IGKV3D-20, IRF5, ITFG2, KDELR1, LINC00211, LZIC, MGAM, MILR1, MNT, MVD, MYBL2, NIT1, NMT1, NOP14-AS1, OAZ1, PCM1, PCYT1A, PKM, PRL, PRR3, PSTPIP1, REEP4, RNF11, RRP1, SAFB2, SCAI, SCARB1, SGTA, SH3PXD2B, SIGLEC10, SIRPB3P, SMNDC1, SND1, SNX6, SPDYA, ST14, STRADB, STX3, STX8, SYNRG, TAB1, TAF9, TARM1, TBXAS1, TCF3, TCL6, TIMM44, TMCO6, TMEM134, TMEM182, TNFRSF1A, USP14, VPS29, WDR13, XRCC4, ZMYM3, ZNF350, ZNF615

The continuation of Table 4

FRASER analysis for pre-drug group	AC006129.2, ACADM, ADGRE1, AKAP9, ANP32B, APC15, ARHGAP26, ATAD1, ATP6V0D1, BOP1, BRAT1, C1orf162, CD300A, CEACAM21, CEP57L1, CPPED1, CTC-435M10.3, CXorf65, DERL2, DSE, EDEM1, EIF6, ELMOD3, EPN1, FAS, FUBP3, G6PD, GAB, GALK1, GAS7, GNMT, GNS, HMG20A, HOTAIRM1_3, HPGD, IGLV2-14, KHNYN, KP5, LIMD2, LMF1, LRRC41, MEI1, MKL1, MSL1, NEMP1, NPIPP1, NT5C, OPLAH, PAXIP1, PI4KA, PLEKHM2, PMPCA, PSME2, PTOV1-AS2, QKI, RAB11FIP3, RASA4, RASA4B, RASSF7, RC3H2, RHOBTB3, RP11-197N18.2, RP11-864I4.1, RP3-475N16.1, RPS6KB1, RSE6, SAFB, SEC13, SECISBP2, SECISBP2L, SH2D1A, SLC2A1, SNW1, SRA1, TBC1D1, TMEM175, TMEM259, TMEM91, TNK2, TUT1, TXNDC17, U2AF1, UBL5, WASHC3, WDR55, WDR73, ZFAND1, ZNF569
------------------------------------	--

The continuation of Table 4

FRASER analysis for post-drug group	AC003006.7, AC009133.23, AC068533.7, ACOX1, ACP1, ACTR1B, ADAM15, ADCY7, ADIPOR1, AGO1, ALDH3A2, ARHGDIA, ASPH, ASTE1, ATG4C, ATG9A, ATP6V0D1, AUP1, BAP1, BROX, BTBD10, CALM3, CAPN3, CAPRIN2, CARD17, CBWD1, CBWD4P, CBWD7, CCNY, CD151, CD3D, CDS2, CENPT, CIC, CLASRP, CLEC10A, CLK2, CNOT2, CNOT8, COPB2, CRBN, CSNK1A1, CTB-50L17.14, CTD-2540L5.5, CTDNEP1, DAB1, DCXR, DIAPH3, DJB12, EIF4E3, EIF4G2, ELMO3, EMG1, EMG1_2, EPHB6, FAM69B, FEM1C, FMR1, GMIP, GMPPB, GOLGA3, GPR132, GSK3B, HDHD2, HERC2, HIVEP2, HLX, HMG20B, HOTAIRM1, HOTAIRM1_3, HTRA2, IFITM1, IGHD, IGHM, IGKV3-20, IKZF3, IL15RA, IL6R, IMMT, IMPDH1, ITCH, KIAA2026, KIF5B, KLHL24, KLRK1, KP5, LINC00174, LINC00893, LYSMD3, MARCH5, MED6, MEI1, MFSD2B, MINPP1, MKRN1, MLST8, NDUFA1, NINJ2, NPC1, NPEPL1, NUDT4, OCIAD1, OMA1, PADI2, PAGR1, PCID2, PCMT1, PGAP2, PGLYRP1, PIK3R5, PLD3, PLIN5, PLP2, PLXNB2, POLDIP3, POLI, PPM1B, PPP1R18, PPP2R1A, PRKACA, PRKAG1, PRT, PUS7, RAC2, RAPGEF1, RASSF2, RCE1, RGS19, RNF34, RP11-164J13.1, RP11-318A15.7, RP11-508N22.2, RPP30, SCAP, SCPEP1, SCYL1, SDR39U1, SLC6A8, SLCO4C1, SNHG4, SNRNP35, SOCS7, SPATA20, SPC5, SUN1, TBC1D10B, TBCE, TBXAS1, TCEAL3, TFE3, THEMIS2, TIMMDC1, TMEM173, TMEM8A, TNK2, TRRAP, TSG101, TUBA4A, UBLCP1, ULK3, VPS52, WDFY2, XXbac-BPG181M17.5, YME1L1, ZBTB17, ZDHHC16, ZDHHC7, ZMAT2, ZMYM4, ZNF37A
-------------------------------------	--

4.5. Aberrant Splicing Analysis through SCISSION

Genes identified to be aberrantly spliced based on the FRASER or dasper tools were used as inputs to SCISSION to achieve splicing detection analysis with a third tool. The list had 185 genes for the pre-drug group and 271 genes for the post-drug group. SCISSION identified the state of splicing anomaly for each of these genes in each sample. Then, the responder group and non-responder group were compared using The Fisher's Exact Test. Only 2 aberrantly spliced genes were identified by SCISSION for both analysis groups: GAS7 and PYGM in the SCISSION analysis for pre-drug group and ELMO3 and RAPGEF1 in the SCISSION analysis for post-drug group. This result was surprising since these genes were found to be differentially spliced only in the non-responder group by FRASER or dasper. SCISSION, on the other hand, identified abnormal splicing events for these genes also in the responder group. This can be attributed to the use of total RNA-Seq data since this type of data contains small RNA reads and intronic reads having non-uniform coverage [Choi et al., 2021]. The presence of any read that was aligned to an intronic region or even a little increase in exonic region with the presence of any small RNA reads could be detected as outlier by the SCISSION's outlier detection algorithm. As a result, SCISSION reports false positive abnormal splicing event for many genes, which overshadows the true positive outlier events when the two groups are compared. Therefore, even if SCISSION achieved to catch splicing abnormalities, it was not able to prioritize the genes with abnormal splicing events from the samples of drug-responders and drug non-responders using total RNA-Seq data.

4.6. Detection of DEGs between Drug Responders and Drug Non-Responders and Identification of PPIs

Differential expression (DE) analysis was performed for the pre-drug and post-drug groups separately using reads which were quantified using Kallisto tool and normalized using DeSeq2 . As a result, a total of 326 genes were identified as differentially expressed between the drug-responder and drug non-responder samples in differentially expressed genes analysis for the pre-drug group, where 145 genes were detected as upregulated while 181 genes were detected as downregulated in the

Drug Non-Responder samples with a *p-value* cut-off of ≤ 0.01 and *FC* cutoff of $\leq 1/1.2$ or ≥ 1.2 . Additionally, within the differentially expressed genes for the post-drug group, 735 genes were found to display differential expression in which 500 genes were detected as upregulated while 235 genes were detected as downregulated.

Moreover, the genes exhibiting protein-protein interactions with the detected genes mentioned above were also checked. The goal was to identify genes that had abnormal splicing events but were not differentially expressed. Such genes may be connected to each other through a DEG gene. To achieve this, the DEGs lists were used, and the analysis was performed on the STRING database. The neighbor-state was checked up to third degree relations. For the differentially expressed genes of the pre-drug group, 143 PPIs were detected, and for the differentially expressed genes of the post-drug group 274 PPIs were identified.

4.7. Comparison of the Splicing and Differential Expression Analysis Results

The common genes in the intersection of the results of the differential expression analysis and the abnormal splicing analysis were checked. In the pre-drug group, no common gene was detected between the list of abnormally spliced genes and the list of differentially expressed genes. The union of genes detected by any three splicing abnormality detection tools was used in this comparison. Similarly, there was no common gene in the intersection of the list of abnormally spliced genes and the list of genes being neighbor to the differentially expressed genes through PPI up to third neighbors (Figure 12). We observed exactly the same pattern also for the post-drug group in terms of common genes (Figure 13).

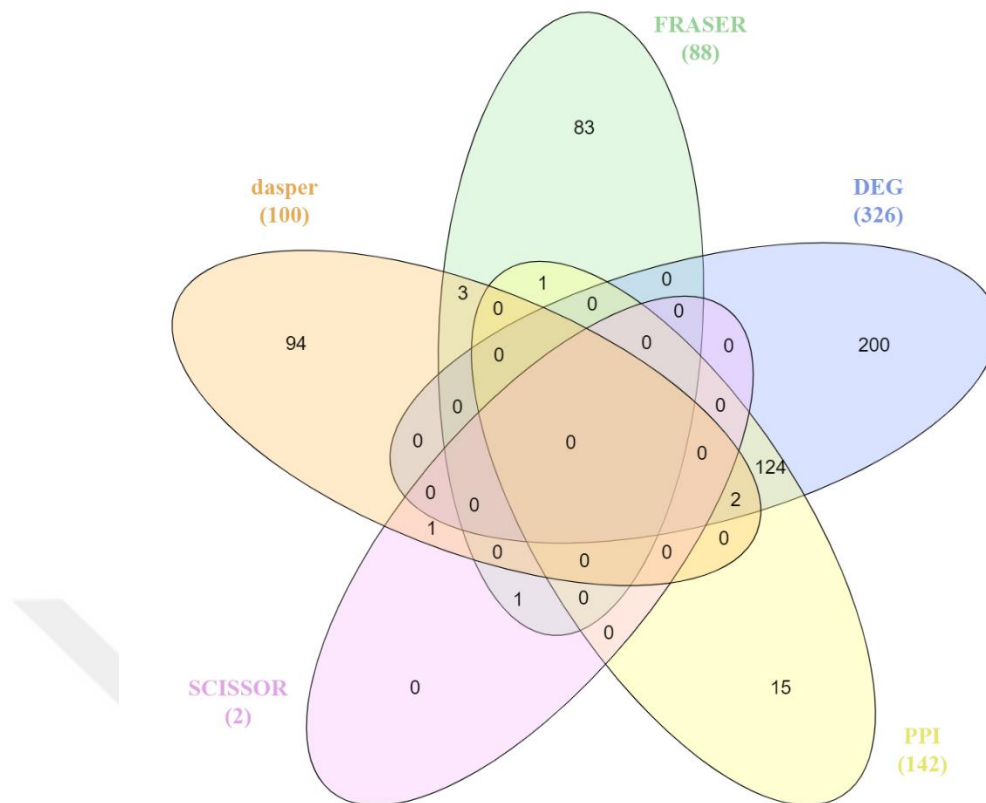


Figure 12: Venn diagram for the result of abnormal splicing and differential expression analysis for the pre-drug group. DEG; Differentially expressed genes, PPI; Protein-protein interactions.

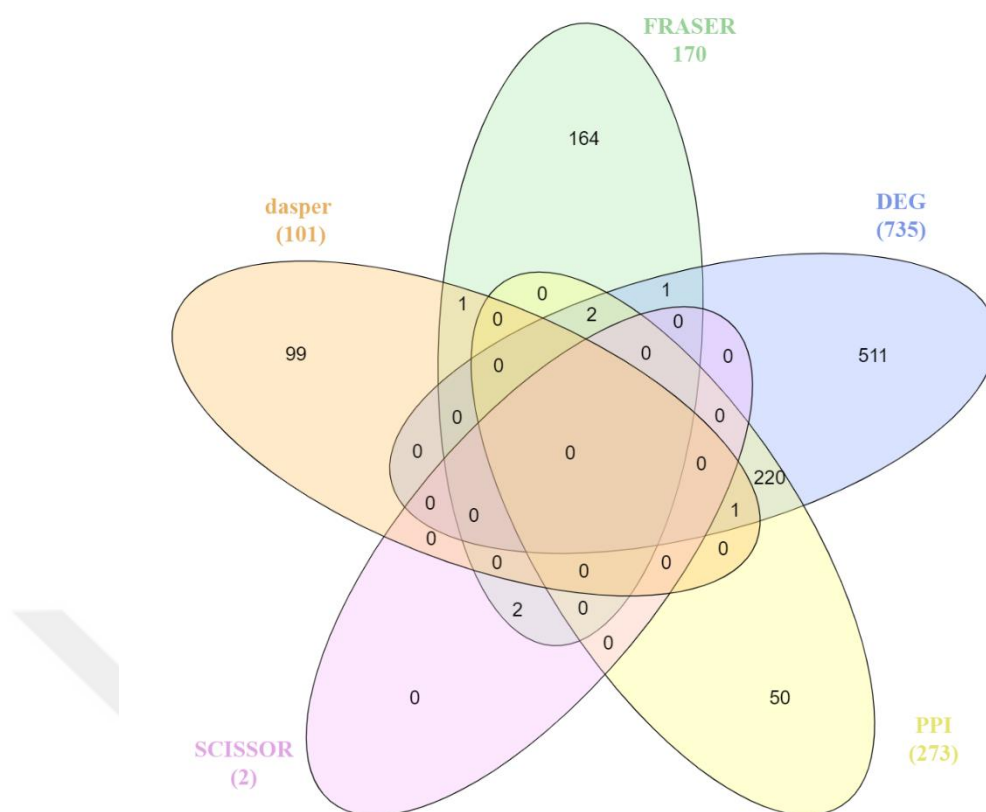


Figure 13: Venn diagram for the result of abnormal splicing and differential expression analysis for the post-drug group. DEG; Differentially expressed genes, PPI; Protein-protein interactions.

Later, the results of differential expression analysis, PPI analysis, dasper analysis, FRASER analysis and SCISSOR analysis were compared separately in a pairwise manner. In the pre-drug group, two genes -TNFRSF18 and CCL3- were found to be common between the results of dasper analysis and differentially expressed genes analysis. On the other hand, CD300A was common in both FRASER analysis results and protein-protein interaction analysis results. there was a single common gene between the SCISSOR analysis results and the FRASER analysis results, which was GAS7. And PYGM was common in the results of both SCISSOR analysis and dasper analysis. Moreover, MEI1, RASSF7, and RAB11FIP3 genes (Figure 12) were detected in both dasper analysis results and FRASER analysis results.

When it comes to the post-drug group, ABCB4 gene was identified in the results of both dasper analysis and differentially expressed genes analysis while TIMMDC1 and KPNA5 genes were common in both FRASER analysis results and differentially

expressed genes analysis results. ELMO3 and RAPGEF1 from the results of SCISSOR analysis overlapped with the FRASER analysis results. Additionally, LINC00893 was commonly detected in both FRASER analysis results and differentially expressed genes analysis results. In the same way, TBXAS1 gene was detected as an outlier in both dasper analysis results and FRASER analysis results (Figure 13). According to the literature, there was no reported relation between these commonly detected genes and CF disease. However, 7 of these 16 genes interacted with the CFTR gene directly or indirectly (Figure 14). So, it may be hypothesized that, any of these genes with splicing abnormality could exacerbate the outcome of the variation detected in CFTR of CF patients in the Drug Non-Responder group.

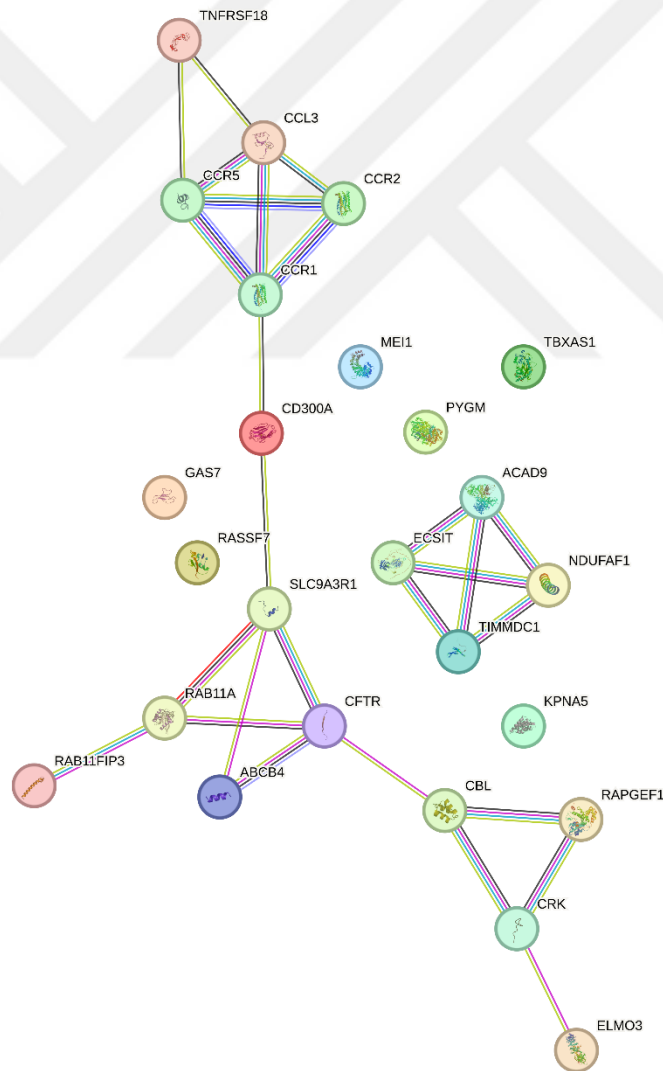


Figure 14: PPI graph of CFTR gene with the genes detected from differential expression results and differential splicing results.

4.8. Benchmarking: Performance Comparison Between Three Recently Developed Splicing Tools

From the results of splicing analysis, FRASER, dasper and SCISSOR tools can be compared in three aspects. Firstly, input file format is an important parameter during the analysis since the outlier detection analysis requires a plenty amount of data to dig up the outliers from control data. The need for intensive data for the input is an obstacle since it necessitates a high-level computational setup such as high core and memory, extend storage environment. dasper eases these concerns by taking .bigwig and splice junction files as input data, which leads to a significant reduction in the file sizes, and, hence, in computational time. However, SCISSOR requires .bam files in every analysis. Moreover, since it does not perform analysis for all the genes in a sample, but it needs to be run separately for each gene, event all .bam files of a big cohort should be provided to SCISSOR analysis as an input for every target gene that may possess outlier splicing. When it comes to the FRASER tool, while it requires .bam files, it also accepts external splice junction count data as input. Also, it could extract the splice junction data calculated during the analysis so that this data could be used as input in following FRASER analysis. On the other hand, input file type is critical in terms of the compatibility of the data content with the algorithm of the tools. SCISSOR results are able to show that this tool fails to detect abnormalities using total RNA-Seq data since it determines any unbalanced read depth as outlier. dasper also calculates read coverage together with splice junction coverage. Like SCISSOR, dasper results have the potential to be populated with false positive outliers. However, dasper reports an abnormal read coverage when it is related with an abnormal splice junction coverage. FRASER also operates the information of splice junction coverage similar to dasper and applies statistical analysis on the splice junction read counts. Although junction read coverage does not greatly differ according to the type of RNA data, additional injections of junction read from total RNA reads other than mRNA reads can generate false positive outlier events. Therefore, mRNA sequencing would be preferable over total RNA sequencing for such analyses.

Secondly, the algorithmic approaches differ in the outlier interpretation. RNA-Seq information contains highly dynamic and sensitive content, the obtained results may be influenced by confounding factors such as age, gender, race, the sequencing

platform, the applied bioinformatic pipelines, and even the version of databases utilized in bioinformatic analysis. Moreover, considering the rare event detection, the presence of confounding factors could easily pollute sample data. Despite FRASER being able to eliminate confounder effects using an autoencoder model algorithm, dasper and SCISSOR tools are far from batch effect awareness. On the other hand, in outlier detection, dasper prioritizes the junctions that are abnormally upregulated or downregulated and determines them as outliers only when they are correlated with another abnormally upregulated or downregulated junction. Although this approach is able to catch differentiated junction motif, it is possible to miss abnormal junctions that were upregulated or downregulated independent of any other abnormal junction. As it is stated above, SCISSOR focuses on the abnormal read depth detection between samples, and the outlier detection steps of total RNA-Seq data could be fluctuated by the presence of non-uniformly expressed reads. FRASER outperforms dasper and SCISSOR tools in outlier detection step on total RNA-Seq data, yet it requires at least ~30 samples to carry out a high accuracy outlier detection.

Finally, dasper and FRASER result in a table output containing all outlier junctions detected for each sample with the position information. FRASER provides a p-value to evaluate the probability assumption between detected junctions while dasper provides a rank-based prioritization. The dasper tool developers do not determine a standardized interval value for rank-based prioritization and the specification of the significant threshold of rank score is evaluated subjectively. So, in practice, the rank-based prioritization was not effective to assess the significant outlier junctions detected from total RNA-Seq, which results in a plenty amount of outlier events. dasper also outputs a plot describing junction map and base coverage across given gene. This plot facilitates the interpretation of the effect of a detected abnormal junction on protein level. SCISSOR outputs a sample ID list, which indicates the samples detected as having outlier event for a given gene. Then, it plots the base coverage profile along the gene by representing the abnormal expression profile for samples with outlier. There is non-standardized transcript annotation information available for each gene. For example, while some genes consist of a single transcript, others can have more than 100 transcripts. Furthermore, the number, description, and numbering of exons in each transcript can vary. Dasper and SCISSOR work based on the transcript information and the evaluation of the results of dasper, and SCISSOR plots require the observation of transcript level knowledge.

5.CONCLUSION

RNA-Seq data plays a key role when the variant level information generated by genomic sequencing analysis is uninformative. Recently, integrative RNA-Seq approach merged as a hot topic in rare Mendelian disease research since it facilitates the prediction of the outcomes of variants in terms of gene expression and alternative splicing. An abnormal splicing event can affect the protein folding and disrupt the expression of the gene or other genes it interacts with.

In this study, the research hypothesis was as follows: the CF patients that are unresponsive to lumacaftor/ivacaftor medication could have a gene with abnormal splicing event affecting gene expression profiles directly or indirectly. Outlier splicing analysis was performed using three recently developed RNA-Seq-specific splicing tools; FRASER, dasper and SCISSOR. In order to identify the differentially expressed genes having abnormal events in unresponsive samples but not in responsive samples, DEG analysis was applied to list the genes showing different expression profiles. Also, genes having protein-protein interaction with the DEGs were analyzed to identify indirect relations with the abnormally spliced genes. dasper detected 100 genes having abnormal splicing events in the pre-drug group and 101 in the post-drug group. FRASER reported 88 and 170 genes with splicing outliers in the pre-drug and post-drug groups, respectively. Unexpectedly, SCISSOR identified only two of the genes detected by FRASER or dasper. From the intersection of the results of outlier splicing, differential expression and protein-protein interaction analyses, TNFRSF18, CCL3, CD300A, RAB11FIP3, ABCB4, ELMO3 and RAPGEF1 genes were detected as they may have a potential effect in drug responsiveness.

As a conclusion, the benchmarking results show that FRASER outperformed the other tools in terms of many features, such as (I) feasible and accessible input data requirement, (ii) data type-sensitive and -specific analysis through confounder detection and removal, (iii) informative output data. A future study can be performed to test the performance of these tools based on polyA enriched mRNA sequencing data rather than total RNA sequencing data. Also, as a further study it could be meaningful to analyze the splicing abnormalities with high amount of CF samples residing Phe508del (F508del) mutation, by focusing on TNFRSF18, CCL3, CD300A, RAB11FIP3, ABCB4, ELMO3 and RAPGEF1 genes. Furthermore, the identification

of variants within CF samples and the assessment of these variant regions using splicing analysis results can help identify rare events that might have been overlooked. Therefore, performing variant analysis on CF samples can complement the analyses conducted in this study. Moreover, exploring other types of CF-causing variants, such as rare deep intronic variants, alongside the frequently observed F508del variant, could provide insights into the underlying reasons for differential drug responsiveness and potentially lead to the creation of new subcategories within the CF population. To create a target variant list, it may be beneficial to prioritize CF mutations that have been reported in genome-wide association studies (GWAS). Combining the variant analysis with the RNA-Seq based functional analysis may facilitate the detection of CF modifier genes.

Additionally, it is known that the transcriptomic profile considerably varies in whole blood samples. Since the splicing tools used in this study are based on expression-based data, it is crucial to ensure that the study samples exhibit a consistent general expression profile. Therefore, the usage of whole blood data in the expression-based analysis could create bias in the recognition of disrupted expression resulting from any functional corruption since whole blood data has a very variable expression profile. One potential solution may be the utilization of the RNA-Seq data obtained from tissues showing more stable expression profile, such as lung tissue, which may represent the gene level changes caused by CF disease better than whole blood tissue.

Besides, it is important to recognize that different procedures for processing raw data and the use of distinct reference genomic data can introduce significant batch effects, and not all RNA-Seq analysis tools have the capability to effectively eliminate these factors. As a potential avenue for future research, all the samples in this study could be subjected to the same RNA-Seq pipeline and reference genomic data to mitigate batch effects and ensure consistency in the analysis.

REFERENCES

Amaral, P., Carbonell-Sala, S., De La Vega, F. M., Faial, T., Frankish, A., Gingeras, T., ... & Salzberg, S. L., (2023), "The status of the human gene catalogue", arXiv preprint arXiv:2303.13996.

Andrews, S., (2010), "FastQC: a quality control tool for high throughput sequence data".

Bolger, A. M., Lohse, M., & Usadel, B., (2014), "Trimmomatic: a flexible trimmer for Illumina sequence data", *Bioinformatics*, 30(15), 2114-2120.

Bray, N. L., Pimentel, H., Melsted, P., & Pachter, L., (2016), "Near-optimal probabilistic RNA-seq quantification.", *Nature biotechnology*, 34(5), 525-527.

Chen, Q., Shen, Y., & Zheng, J., (2021), "A review of cystic fibrosis: Basic and clinical aspects.", *Animal Models and Experimental Medicine*, 4(3), 220-232.

Choi, H. Y., Jo, H., Zhao, X., Hoadley, K. A., Newman, S., Holt, J., ... & Hayes, D. N., (2021), "SCISSOR: a framework for identifying structural changes in RNA transcripts", *Nature communications*, 12(1), 286.

Cummings, B. B., Marshall, J. L., Tukiainen, T., Lek, M., Donkervoort, S., Foley, A. R., ... & MacArthur, D. G., (2017), "Improving genetic diagnosis in Mendelian disease with transcriptome sequencing", *Science translational medicine*, 9(386), eaal5209.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., ... & Gingeras, T. R., (2013), "STAR: ultrafast universal RNA-seq aligner", *Bioinformatics*, 29(1), 15-21.

Douglas, A. G., Wood, M. J., (2011), "RNA splicing: disease and therapy", *Briefings in functional genomics*, 10(3), 151-164.

Ewels, P., Magnusson, M., Lundin, S., & Käller, M., (2016), "MultiQC: summarize analysis results for multiple tools and samples in a single report.", *Bioinformatics*, 32(19), 3047-3048.

Fajac, I., Sermet, I., (2021), "Therapeutic approaches for patients with cystic fibrosis not eligible for current CFTR modulators", *Cells*, 10(10), 2793.

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., ... & Battle, A., (2020), Transcriptomic signatures across human tissues identify functional rare genetic variation. *Science*, 369(6509), eaaz5900.

Ferraro, N. M., Strober, B. J., Einson, J., Abell, N. S., Aguet, F., Barbeira, A. N., ... & Battle, A., (2020), "Transcriptomic signatures across human tissues identify functional rare genetic variation.", *Science*, 369(6509), eaaz5900.

French, C. E., Dolling, H., Mégy, K., Sanchis-Juan, A., Kumar, A., Delon, I., ... & Raymond, F. L., (2022), "Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood.", *Human Genetics and Genomics Advances*, 3(3).

Frésard, L., Smail, C., Ferraro, N. M., Teran, N. A., Li, X., Smith, K. S., ... & Montgomery, S. B., (2019), "Identification of rare-disease genes using blood transcriptome sequencing and large control cohorts.", *Nature medicine*, 25(6), 911-919.

Gallati, S., (2014), "Disease-modifying genes and monogenic disorders: experience in cystic fibrosis.", *The application of clinical genetics*, 133-146.

Global Genes, "Rare Disease Facts", Available at: <https://globalgenes.org/learn/rare-disease-facts/> (Accessed: 2023, July 13)

Gonorazky, H. D., Naumenko, S., Ramani, A. K., Nelakuditi, V., Mashouri, P., Wang, P., ... & Dowling, J. J., (2019), "Expanding the boundaries of RNA sequencing as a diagnostic tool for rare Mendelian disease", *The American Journal of Human Genetics*, 104(3), 466-483.

Guillot, L., Beucher, J., Tabary, O., Le Rouzic, P., Clement, A., & Corvol, H., (2014), "Lung disease modifier genes in cystic fibrosis.", *The international journal of biochemistry & cell biology*, 52, 83-93.

Jenkinson, G., Li, Y. I., Basu, S., Cousin, M. A., Oliver, G. R., & Klee, E. W. (2020). LeafCutterMD: an algorithm for outlier splicing detection in rare diseases. *Bioinformatics*, 36(17), 4609-4615.

Kopp, B. T., Fitch, J., Jaramillo, L., Shrestha, C. L., Robledo-Avila, F., Zhang, S., ... & Mejias, A., (2020), "Whole-blood transcriptomic responses to lumacaftor/ivacaftor therapy in cystic fibrosis", *Journal of Cystic Fibrosis*, 19(2), 245-254.

Kremer, L. S., Bader, D. M., Mertes, C., Kopajtich, R., Pichler, G., Iuso, A., ... & Prokisch, H., (2017), "Genetic diagnosis of Mendelian disorders via RNA sequencing", *Nature communications*, 8(1), 15824.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., ... & Moore, H. F., (2013), "The genotype-tissue expression (GTEx) project", *Nature genetics*, 45(6), 580-585.

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), 1-21.

Meienberg, J., Bruggmann, R., Oexle, K., & Matyas, G., (2016), "Clinical sequencing: is WGS the better WES?.", *Human genetics*, 135, 359-362.

Mertes, C., Scheller, I. F., Yépez, V. A., Çelik, M. H., Liang, Y., Kremer, L. S., ... & Gagneur, J., (2021), "Detection of aberrant splicing events in RNA-seq data using FRASER", *Nature communications*, 12(1), 529.

Ogilvie, V., Passmore, M., Hyndman, L., Jones, L., Stevenson, B., Wilson, A., ... & Boyd, A. C., (2011), "Differential global gene expression in cystic fibrosis nasal and bronchial epithelium", *Genomics*, 98(5), 327-336.

Okonechnikov, K., Conesa, A., & García-Alcalde, F., (2016), "Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data", *Bioinformatics*, 32(2), 292-294.

Online Mendelian Inheritance in Man, OMIM[®], McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), (2023, July 13). World Wide Web URL: <https://omim.org/>

Orphanet, (2012), "Education about rare disease", Available at: https://www.orpha.net/consor/cgi-bin/Education_AboutRareDiseases.php?lng=EN (Accessed: 2023, July 13)

Pereira, R., Oliveira, J., & Sousa, M., (2020), "Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics", *Journal of clinical medicine*, 9(1), 132.

Phuan, P. W., Haggie, P. M., Tan, J. A., Rivera, A. A., Finkbeiner, W. E., Nielson, D. W., ... & Verkman, A. S., (2021), "CFTR modulator therapy for cystic fibrosis caused by the rare c. 3700A> G mutation", *Journal of Cystic Fibrosis*, 20(3), 452-459.

Rahit, K. T. H., Tarailo-Graovac, M., (2020), “Genetic modifiers and rare mendelian disease”, *Genes*, 11(3), 239.

Ramos-Fuentes, F., González-Meneses, A., Ars, E., & Hernández-Jaras, J., (2020), “Genetic diagnosis of rare diseases: past and present.”, *Advances in therapy*, 37, 29-37.

Retterer, K., Juusola, J., Cho, M. T., Vitazka, P., Millan, F., Gibellini, F., ... & Bale, S., (2016), “Clinical application of whole-exome sequencing across clinical indications.”, *Genetics in Medicine*, 18(7), 696-704.

Sanger, F., Nicklen, S., & Coulson, A. R., (1977), “DNA sequencing with chain-terminating inhibitors”, *Proceedings of the national academy of sciences*, 74(12), 5463-5467.

Schaaf, J., Sedlmayr, M., Schaefer, J., & Storf, H., (2020), “Diagnosis of Rare Diseases: a scoping review of clinical decision support systems”, *Orphanet journal of rare diseases*, 15(1), 1-14.

Seaby, E. G., Ennis, S., (2020), “Challenges in the diagnosis and discovery of rare genetic disorders using contemporary sequencing technologies”, *Briefings in Functional Genomics*, 19(4), 243-258.

Soneson, C., Love, M. I., & Robinson, M. D., (2015), “Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences.”, *F1000Research*, 4.

Szkarczyk, D., Kirsch, R., Koutrouli, M., Nastou, K., Mehryary, F., Hachilif, R., ... & von Mering, C., (2023), “The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest.”, *Nucleic acids research*, 51(D1), D638-D646.

Tazi, J., Bakkour, N., & Stamm, S., (2009), “Alternative splicing and disease”, *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease*, 1792(1), 14-26.

Wikipedia, (2023), “Rare disease”, Available at: https://en.wikipedia.org/wiki/Rare_disease (Accessed: 2023, July 13)

Wright, J. M., Merlo, C. A., Reynolds, J. B., Zeitlin, P. L., Garcia, J. G., Guggino, W. B., & Boyle, M. P., (2006), “Respiratory epithelial gene expression in patients with

mild and severe cystic fibrosis lung disease”, American journal of respiratory cell and molecular biology, 35(3), 327-336.

Yépez, V. A., Mertes, C., Müller, M. F., Klaproth-Andrade, D., Wachutka, L., Frésard, L., ... & Gagneur, J., (2021), “Detection of aberrant gene expression events in RNA sequencing data.”, Nature Protocols, 16(2), 1276-1296.

Zhang, D., Reynolds, R. H., Garcia-Ruiz, S., Gustavsson, E. K., Sethi, S., Aguti, S., ... & Ryten, M., (2021), “Detection of pathogenic splicing events from RNA-sequencing data using dasper”, BioRxiv, 2021-03.



BIOGRAPHY

Canan KOLAKOĞLU completed her undergraduate education in 2018 at Yıldız Technical University, Department of Bioengineering. In 2020, she started her master's degree in the Department of Bioengineering, Systems Biology and Bioinformatics Program at Gebze Technical University. Since 2020, she has also been working as a bioinformatician at PHITECH.

