Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Advanced data exploration techniques for augmented transcriptional landscape and its better quantification

PhD Thesis

Author: Agata Muszyńska
Supervisor: dr hab. inż. Paweł Łabaj
Co-supervisor: Assoc. Prof. Dr. David Kreil

Gliwice, February 2023

I would like to dedicate my thesis
to my beloved grandparents

Contents

CONTENTS

List of abbreviations

AS alternative splicing. 5, 6, 10, 69, 88, 102

ASE alternative splicing event. 63, 64, 88, 90–92, 106

BH Benjamini- Hochberg. 50

BY Benjamini -Yekutieli. 50

CI Confidence Interval. 50

DEA Differential Expression Analysis. 15, 34, 49

DEG Differentialy Expressed Genes. 31, 58, 61, 62, 88

DGE Differential Gene Expression. 5, 6

EB Empirical Bayes. 47, 48

FDR False Discovery rate. 49

FM Full-text Minute-space. 40, 41

FPKM Fragments Per Kilobase Million. 22

FWER Familywise Error Rate. 49

GO Gene Ontology. 32–34, 51, 52, 69, 73, 77, 96, 132

List of abbreviations

GSEA Gene Set Enrichment Analysi. 5

logFC logarithm of fold change. 50

MM mismatch. 11

MSD Minimum Significant Difference. 50

nASE novel alternatively spliced event. 10, 63

NB Negative Binomial. 24, 48

NGS Next-Generation Sequencing. 35, 95

PEER Probabilistic Estimation of Expression Residuals. 45

PM perfect match. 11

RPKM Reads Per Kilobase Million. 22

SMRT Single Molecule, Real-Time. 18

SVA surrogate variable analysis. 31, 44, 45, 95, 96

T-DBG transcriptome de Bruijn graph. 39

TMM trimmed mean of M-values values. 23, 34, 46, 48

TPM Transcripts Per Kilobase Million. 22

WTS whole transcriptome sequencing. 37

ZMW zero- mode waveguides. 18

List of abbreviations

Chapter 1

Introduction

In recent years, researchers have become more and more aware of a wider reproducibility crisis that the whole scientific community is facing. As highlighted by Baker [6], 70% of the 1,500 researchers interviewed have tried and failed to reproduce another scientist's experiments, and more than half have failed to reproduce their own experiments. Furthermore, Freedman, Cockburn, and Simcoe [41] estimated that alone in the year 2015 approximately $28 billion was spent on preclinical research that was not reproducible.

Despite comprehensive benchmarks and best practice guidelines provided, such as by the MAQC and SEQC consortia and in follow up studies [23, 108, 22], the reproducibility still remains a challenge also for RNA-seq data analysis. RNA-seq is the latest state-of-the-art technology that allows systematic assessment of gene activities in a genome-scale assay

that measures the abundances of RNA molecules encoded by all the genes through Next-Generation Sequencing methods (see Sections 2.3 and 2.4). Fast development of the technology is followed by a flood of new analytical tools and approaches, outpacing the development of best-practice guidelines. The problem is compounded by outdated but simply well-known tools remaining in frequent use long after they have been superseded by better methods, and by this affects the quality of the results. As an example, an article from 2019 [72] shows that many researchers still do not correct for the effects of different gene lengths, leading to significant biases in the analysis results. Reproducibility challenges can also arise from sloppy study design and simply from the characteristics of the measured data, such as low signal levels or confounding factors (see Section 4.1.1). Even with careful application of state-of-the-art methods for detrending and normalization, confounding factors could not always be identified and removed for a meaningful quantitative analysis, as shown in Sections 4.1.1 and 4.3. In Sections 4.1.2-4.1.7 my work shows that despite poor results in quantitative data analysis, we can still extract robust and valuable qualitative results.

Another, more complex reason behind the lack of reproducibility in RNA-seq is the transcriptomic landscape complexity [27]. Although the human genome is the most studied and complete, there are still frequent annotation updates, giving different numbers of genes and transcripts. A summary of changes in each release of the Ensemble/GENCODE annotation is

provided in Zerbino, Frankish, and Flicek [105]. The total number of genes and transcripts in the reference annotations has changed a lot in the years 2003-2019. Moreover, the distribution of particular feature types changed considerably. A closer look at the human and mouse reference models reveals that, although both contain a similar number of genes, the number of mouse transcripts is smaller [40, 39]. Statistics for current GENCODE releases state that there are 252,416 transcripts annotated for 62,696 human genes [91] but only 149,423 transcripts annotated for 56,923 genes in mice [92]. This disproportion might indicate incomplete information for the mouse transcriptome, which can cause misleading alignments and thus reproducibility issues as gene annotation evolves. Therefore, when analyzing RNA-Seq data, it is crucial to also consider the qualitative component of searching for novel alternatively spliced transcripts that extend the reference gene models, so that future experiments can better reproduce the expression profile estimates for the transcripts observed (see Section 3.1).

I examined the data analysis requirements in multiple real-world projects at different stages of RNA-seq data analysis, and despite the project-specific differences, many time-consuming steps were often similar for the different data sets. I then looked for integrated solutions that incorporated state-of-the-art methods and could automate the complex analysis. Interestingly, while available software could be used for individual steps, there was no integrated solution suitable for end-to-end state-of-the-art RNA-seq analysis. Where more comprehensive pipelines were presented, they were con-

structed for a particular data set and did not generalize to other applications, limiting what experiments could be done with them.

1.1 The aims of the thesis research project

The aim of this thesis research work was to stress-test recommended best practices in RNA-seq data analysis developed on benchmarking data sets on real world experiments and, as a result, *to advance the reproducibility and robustness of analysis results*. The solution presented here is a carefully designed end-to-end pipeline for quantitative *and* qualitative analysis of RNA-Seq data, validated on independent real-world data sets collected in new experiments. Based on the diverse experience gained during the analysis of the different data sets, a selection of methods best suited for the range of challenges encountered has been compiled. Three selected projects are described in greater detail in Sections 3.2.1, 3.2.2 and 3.2.3.

Results from all these projects corroborate that there are no simple gold standards when it comes to measurements and analyses in the modern molecular life sciences, with best practice depending on the type of data and the research question, so carefully tailored approaches are necessary. Specifically, this is in line with my research work [38] within the Epigenomics Quality Control (EpiQC) Working Group of the US FDA, where *Power and Limitations* of multiple approaches for interrogating modifica-

tions in DNA were characterized, as well as earlier studies by the US FDA MAQC and SEQC Consortia [22, 23].

1.2 Novelty

Stress-testing on multiple diverse experiments led to the development of a unique analysis pipeline solution that integrates multiple stages of RNA-seq data analysis. Specifically, this is in contrast to existing earlier workflows, which are focused on particular steps, such as alignment. At the same time, my pipeline provides a range of suitable choices for each of the critical stages of analysis, covering quality control, preprocessing, alignment, quantification, differential gene expression (DGE) and Gene Set Enrichment Analysis (GSEA), and allows an easy path to implement alternative options. Moreover, it integrates the novel feature of comprehensive qualitative analysis of RNA-Seq data analysis. The provided alternative splicing (AS) analysis is not limited to just the detection of alternative splicing events, but crucially also provides an overview of consequences at the transcript and protein levels required for a functional interpretation. The whole computational package is complemented by a visualization tool. The whole pipeline is tailored to allow easy data flow between different analysis stages. A general pipeline overview is presented in Figure 3.1, with relevant issues also discussed in a recent review manuscript, in which I

also provide a comprehensive overview of RNA-seq technology and available tools [26].

Sections 4.1.2 - 4.1.7 outline in detail how the real-world RNA-Seq data despite even strong confounding effects still provide a very rich source of information. So even if state-of-the-art approaches do not allow for meaningful quantitative analysis of the RNA-Seq data one can still effecitvely explore the qualitative potential of those, and this should become a standard analysis step. That part, however, is missing from standard analysis pipelines even though it is crucial for a better understanding and annotation of gene models. My pipeline bridges that gap, providing an end-to-end workflow - from raw reads, through alignment, Differential Gene Expression (DGE) and Alternative Splicing AS analysis, to protein-level aftermath. Adding this led me to the novel discovery that not only are mouse genes strongly under-annotated but this especially affects genes relevant to the nervous system. The identified new Alternatively Spliced Events yielded unknown transcripts leading to novel functions in the enrichment analysis of gene activities. In addition, I found that specific functional profiles were associated with different types of splicing events, the first such report ever to my knowledge. Interestingly, although suffering strong batch effects in quantitative analysis, the results of the novel alternative splicing event detection were very stable (4.1.2-4.1.7). With the help of an independent external benchmark *side data set* from the SEQC2 project I could characterize potential confounders (See Section 4.2), what has further implications on

the results interpretations.

Last but not least, while the main focus of my research work was on the analysis of the RNA-seq data, expression profiling by microarrays is still widely used, and there are massive public and private repositories of microarray expression profiling data. Building on observations of the complementary nature of these high-throughput technologies [23], I could demonstrate how data analysis techniques developed for one technology could be adapted in the novel context of another, considering the nature of the data source. This allowed me to build and validate an integrated analysis pipeline for both, sharing components for the similar essential steps of differential gene expression analysis and functional downstream analysis (See Section 4.3).

1.3 Thesis outline

The thesis is structured as follows:

- Chapter 2 provides a brief introduction to transcriptomics and high-throughput analysis methods in general. It also describes current challenges and state-of-the-art solutions in RNA-seq data analysis.
- Chapter 3 introduces the pipeline developed and provides a comprehensive overview of all the methods used in the analysis, explaining

what makes them a good choice to obtain robust and reliable results. This part also gives a summary of the data sets used.

- Chapter 4 presents results for a microarray data set and two RNA-seq data sets. It also highlights which part of the data analysis methods can effectively be shared between the two technologies, and how.
- Chapter 5 provides a summary and conclusion.

Chapter 2

Background

2.1 Basics of transcriptomics

Although humans contain approximately 20,000 protein coding genes [73], only a fraction of them are actively expressed as transcripts at any given time. Transcripts are then processed and translated into proteins that perform a wide variety of functions in all living organisms. The whole process is known as the central dogma of molecular biology and consists of many complex stages [20]. Even though in the past decades our knowledge of those processes has expanded very rapidly, with many new tools and technologies being constantly developed, much remains unknown.

Genes consist of exons and introns. Transcription is the stage where genetic information is being rewritten to new molecules- premRNAs. One of the modifications that premRNA undergoes is cutting out introns and

concatenating exons in a final molecule. During this process, different versions of the transcript might be produced, depending on which exons are bound together and how, this mechanism is known as alternative splicing (AS). Almost all multiexonic genes in mammals undergo alternative splicing [83, 76]. This process is crucial in cell development and differentiation, and its dysfunction is associated with numerous diseases [98]. It also participates in the post-transcriptional regulation of mRNA levels and is the main mechanism that has allowed eukaryotes to produce a repertoire of diverse and highly specific proteins from a limited number of genes and therefore plays an important role in evolution [10]. The more complex an organism, the more widespread AS [59]. Splicing has been shown to vary even more between different tissues than between individuals and has also been found to occur more frequently in functionally complex tissues such as the brain [99, 104]. This is due to the complicated processes that take place in the nervous system [70]. AS contributes to the formation of complex neural networks and also to synaptic plasticity [18]. As recently reported, many novel alternatively spliced events (nASEs), characteristic of a particular cell type, can still be found in different regions of the brain [55].

Transcriptomics denotes techniques used to study RNA molecules and the complex processes they undergo. Among high-throughput technologies, the most popular are those based on hybridization (high-density microarrays) or sequencing (RNA-Seq by Next-Generation Sequencing). Once the data are generated, there are numerous ways and algorithms that can

be used to study them.

2.2 Microarray technologies summary

Microarray approach is based on hybridization between two DNA strands due to the nucleic acids strands property of complementarity. Specific nucleotide base pairs are bound together by two or three hydrogen bonds. Microarrays consist of a predesigned library of synthetic nucleic acid probes that are immobilized and spatially arrayed on a solid matrix. These probes hybridize with complementary mRNA sequences that appear in an examined sample. Thanks to fluorescent labeling of binding sequences, they generate a signal whose strength is dependent on the amount of mRNA bound to the spot. To assess the gene expression level, microarrays are scanned, and the signal obtained must be properly preprocessed [50]. According to Fajriyah [35] the two most used platforms are Affymetrix (recently acquired by Thermo Fisher Scientific) and Illumina.

Affymetrix produces oligonucleotide microarrays, composed of short 25-mer oligonucleotide probes organized in 11- 20 pairs complementary to different regions of the same transcript. Each pair consists of a fully complementary probe- PM (perfect match) and a probe that contains one non-complementary nucleotide in the 13 th position- MM (mismatch). DNA sequences are synthesized directly on the surface of the plate. The probes are carefully chosen and constructed to match parts of the sequence of

known or predicted open reading frames. The technique used to produce these arrays is called photolithography. The array, whose surface is made of silica, is covered with special chemical substances that are used to bind specific sequences. These substances are protected by light-sensitive masking agents. The plate is covered by a single nucleotide solution. Then, the places where this nucleotide should be bound are irradiated and the solution is washed away after the nucleotide is attached. The whole procedure is repeated until the sequences of every probe are fully constructed. To estimate gene expression levels, fluorescent dye is used. During several biochemical reactions, RNA is labeled with biotin. The plate is then placed in this solution for a few hours, so that RNA could hybridize with the oligonucleotide probes. After the solution is washed, the array is exposed to a fluorescent label bound to strepatavidin. Due to the fact that biotin has a strong affinity for streptavidin, it binds to the places of the array, where hybridization occurred [50].

Illumina microarrays, on the other hand, use BeadArray technology. They were designed to overcome some of the limitations of spotted arrays, such as poor data quality. The technique is based on small (3 microns in diameter) silicone beads, randomly positioned across wells on an array. Each bead is covered with 50-mer oligonucleotide sequence, specific to the characteristic position in the genome. These sequences are repeated random number of times (usually about 700,000).There are up to 1,536 different bead types, each of them is replicated on an array for about 30 times. The

location and type of each bead is determined in a sequential decoding process, with complementary dye-labeled oligonucleotides, called decoders [47].

A microarray experiment is performed under the assumption that the intensities of the genes reflect the actual levels of mRNA. However, raw microarray data obtained after scanning contain relevant biological information that is highly influenced by a number of non- biological sources of variation. This so- called technical bias can be caused by many reasons, such as uneven hybridization, batch bias, scanner settings, background fluorescence [103]. Therefore, to achieve biologically meaningful data, correction of technical bias is a crucial step in microarray data analysis. It improves concordance with known biological information. This stage is divided into three steps: background correction, normalization, and summarization [43]. In order to obtain true signal values, the data should be adjusted for non- specific binding and optical noise, which is done in the background correction step. Optical noise is introduced by a scanner, which measures hybridization strengths. Depending on the scanner used, different signal values will be obtained. Nonspecific binding occurs because PM probes, apart from detecting transcripts from the intended gene (specific hybridization), detect also other sequences (nonspecific hybridization) [94]. The normalization step aims at manipulating the data in a way that will make measurements from different arrays comparable, which means achieving a measurement scale that has the same origin (zero) for all spots.

Affymetrix GeneChip uses a set of 10-20 probes to measure expression levels of a gene and on average 4 probes for an exon. After preprocessing, those multiple measurements have to be combined to provide a final measure of gene expression. There are multiple methods available to perform this step [51, 12].

What we are looking for in a microarray experiment are relevant changes in gene expression level between different conditions. The simplest way to asses if a particular gene changes its expression is to evaluate the log ratio between two conditions and set a cut- off value. If log fold change is above the cut-off value, a gene is considered differentially expressed. This method, however, has no statistical support and is not robust to type I and II errors. That is why to decide whether the expression of gene A is different in the treated group than in the control group, the measurement is repeated multiple times and then usually a statistical test is applied [75]. Through this process, we compare how much gene expression has changed between different conditions and within replicates of the same condition. We assume that the gene did not change its expression; that is, the so-called null hypothesis, and it is true for majority of genes. If the true null hypothesis is rejected, a Type I error occurs. Type II error means that the false null hypothesis was accepted. The value that indicates if the result is significant is the p-value. It is the probability of observing a particular result or a more extreme result assuming that the null hypothesis is true. Small p-values give strong evidence against the null hypothesis. Genes with low

p-values are the ones that are referred to as significantly differentially expressed; for those genes, we can reject the null hypothesis and conclude that there are differentially expressed. The typical threshold for p-value is 5% but that cutoff is arbitrary, and one might need to set it, for example, a bit higher when it comes to more noisy data. What p-values inform about is actually the probability of making type I error. If a p- value threshold is set to 5% and 20,000 genes are tested, we should be aware that 1,000 genes will be considered significant although they are actually not. There are two approaches to control these false positives either by controlling Family-Wise Error or the False Discovery Rate [45]. After this adjustment, a corrected p-value is obtained which is then used. `Limma package` is one of the most popular approaches for DEA. It fits a linear model and uses moderate t-statistics to detect differentially expressed genes [84].

2.3 NGS technologies summary

Next-generation sequencing, also known as second-generation sequencing, is derived from the Sanger sequencing technique (first-generation sequencing) with a huge improvement in terms of throughput, that is, the number of sequencing reactions in a single run. The pioneering chain-termination method of DNA sequencing was developed by Frederick Sanger and colleagues in 1977. The process requires, among others, two types of

nucleotides- normal and modified ones, which lack a 3'-OH group and thus prevent two consecutive nucleotides from forming a phosphodiester bond, resulting in termination of DNA strand elongation. In addition, those modified nucleotides are also radioactively or fluorescently labeled, allowing for detection. The process is repeated four times for each of the nucleotides. This results in DNA fragments of different lengths, which are then separated by capillary electrophoresis and visualized by autoradiography or UV light to determine the exact DNA sequence [87].

RNA-seq describes all the experimental and computational methods used to assess the origin and abundance of RNA molecules in a sample studied. The main difference between this approach and microarrays is that randomly sampled fragments are sequenced, and thus, we measure the expression of any alternative transcript irrespective of whether it or its parent gene is known or unknown. We are not bound to only known ones as we do not rely on predefined set of probes, as in microarrays.

Nowadays there are many different vendors that provide RNA-seq platforms and therefore the technology and analysis could differ a lot. There are also different applications of sequencing, such as gene expression profiling, alternative splicing or fusion gene discovery, or determining cell-type abundance. Common and general steps include RNA isolation (from tissue, cell, or bulk RNA), library preparation, which represents all RNA molecules in a given sample, and actual sequencing.

Major providers of RNA-seq solutions include Illumina, Thermo Fisher,

Pacific Biosciences, and Oxford Nanopore Technologies. Each company provides different technologies that are targeted at a wide variety of applications. They differ in sequencing and detection methods, read lengths, throughput, run time, and costs and availability.

Illumina provides sequencing by synthesis system, which produces short (50-500bp, depending on the system), paired-end or single reads. cDNA is passed through a flow cell, which is a glass slide with lanes with oligo adapter sequences on the surface. These sequences are complementary to adapters on cDNA fragments and bind to the surface at both ends, forming a bridge. In a process of cluster generation, the sequence is amplified, and a new molecule hybridizes nearby. The process is repeated many times, simultaneously, for millions of clusters, resulting in many copies of the original fragments. After this step is completed, the reverse strands are cleaved and washed away. The remaining forward strand is sequenced in a process in which fluorescently labeled nucleotides are attached to the growing complementary strand. As the nucleotide is incorporated, the signal is emitted and stored for subsequent analysis [62].

Thermo Fisher's Ion Torrent technology is unique in its approach and allows short reads to be detected. Rather than the fluorescent signal, it detects changes in pH caused by hydrogen ions released during incorporation of a nucleotide into the mix. Ion Torrent uses special semiconductor chips with microwells that contain multiple copies of a template molecule that undergoes sequencing. The benefits of this approach include a lower

cost and faster run time; however, it struggles with homopolymer regions and has a lower throughput [62].

PacBio and Oxfrod Nanopore are examples of methods that produce long reads and are also referred to as third-generation sequencing methods. PacBio is also based on sequencing by synthesis, but introduces SMRT (Single Molecule, Real-Time) technology. Adapters are added to the double-stranded template, forming a circular molecule. The molecules are then immobilized in small wells (zero-mode waveguides) in a SMRT cell. Each cell consists of many ZMW but each ZMW contains only one molecule, no amplification is required. During nucleotide incorporation, a light signal is detected. With this approach, nucleotide incorporation is measured in real time for each molecule separately. This approach provides great improvements in terms of speed and accuracy but has lower throughput and can be very expensive. Nanopore provides a cheaper alternative to obtaining long reads. It is also single molecule technique, where a molecule is guided through a protein pore embedded in a membrane. When passing through a pore, DNA changes its ion current, which is specific to the type of nucleotide passing. This simple design also allows for small device sizes [62].

When the read generation process is completed, data is stored, usually in FASTQ files, providing information about detected reads and reference quality score. Next steps include quality control, preprocessing, and alignment to the reference genome or transcriptome. If the reference is

unknown, it is possible to obtain it using de novo transcriptome assembly.

The quality control step checks for artifacts introduced in the library preparation and sequencing process itself. One of the most popular software used for this purpose is `FastQC` [4], which can accept files in FASTQ format, but also already aligned reads in BAM or SAM format. It provides a variety of plots that summarize potential problems with the samples analyzed. Possible issues include untrimmed adapters, sequence duplication, and sequence length distribution. We should also check the GC content and sequence contamination with other organisms. Acceptable artifacts levels are dependent on experiment, it is advised that outliers with more than 30% disagreement should be discarded. It is typical for read quality to decrease towards the 3' end, but too low quality values might decrease mapping quality, and thus should also be trimmed [21].

Read alignment is a process of finding the place in the reference transcriptome or genome where the read originates. This step is computationally intensive, and the exact time and computational resources depend on the software used. Usually, to accelerate this process, the reference is transformed into an index beforehand. The most popular transformation is the Burrows–Wheeler algorithm. It is a lossless compression method that has many applications, but due to the many repeated patterns in DNA strings, it is particularly useful for genomic data. The idea behind the algorithm is to build an array where rows contain all possible cyclic rotations of the input string, sort them lexicographically, and return the last column

of the obtained array. This column is the output- desired index. It contains chunks of the same characters that can be stored in compact form. It is also possible to easily recreate the original string from the index [16].

The final part of the usual pipeline, just as in the microarray approach, includes the estimation of the abundances of transcripts or genes and the subsequent comparison of expression levels between conditions. Statistical approaches are usually employed to detect expression levels of various genomic features (such as genes, exons, and transcripts) that exhibit significant statistical differences across experimental groups.

2.4 RNA-seq technology challenges

RNA-seq is a powerful and commonly used technique, with many new approaches, both laboratory protocols and algorithms used for further analysis, constantly being developed and improved. However, there are still many issues that need to be addressed to obtain robust and reproducible results.

There are many sources of distortion through many steps leading from collecting a tissue sample to the results obtained by bioinformaticians. The library preparation step itself includes many stages and is a source of huge noise in data. An article by Fu et al. [42] states that at the stage of PCR amplification, only 0.7% of the original signal from a target still persists.

Choosing sequencing technology is crucial for obtaining reliable results, as there is a trade-off between read length, throughput, and accuracy. Detecting changes like single nucleotide polymorphisms requires the highest accuracy, whereas tasks such as annotating novel genes would benefit from longer reads. I would like to focus on the issues related to choosing appropriate approaches in the analysis of generated reads. One of the problems is how to deal with junction-spanning reads, which are reads that span more than one exon. Possible options include using a splice-informed aligner (such as `HiSat2` [61] or `MAGIC-BLAST` [13]) or aligning reads against the transcriptome [9]. The choice is not trivial as technology is rapidly changing and there is no single right answer. However, there are articles benchmarking different tools and providing guidelines [108, 23].

Apart from the changes in technology, the reference annotation also changes over time, as explained in Section 1. The choice of annotation can have a huge impact on the results obtained. This is true for both using outdated annotation, but also deciding between different sources for annotation. The most popular ones are Ensembl [24] and RefSeq [74], however, the SEQC study shows that the AceView annotation is the most accurate [23].

Gene and transcript level abundances estimation is a crucial step for further analysis. Gene-level quantification is the most common approach. The simplest way is to directly count fragments per gene, based on coordinate information from a GTF file, and treating a gene as a union of its

transcripts [21]. Approaches differ in how to treat multi-mapped reads or in how much of a fragment must be assigned to a feature to be counted. There is also the possibility of assigning reads to transcripts and then aggregating the results at the gene level. This approach allows for observing the expression of different isoforms and allows us to properly model multi-mapping reads. On the other hand, due to alternative splicing, the origin of many reads can be unambiguous, and to resolve this issue, probabilistic modeling is needed [9].

However, raw read counts are not enough to properly infer about differential expression. They need to be adjusted for different transcript lengths, total number of reads mapped for a given sample (library size), and also GC- content. To account for this, several normalization methods have been developed. To normalize within the sample for gene length and between samples for library size RPKM (Reads per Kilobase Million) and its extension for paired end reads - FPKM (Fragments per Kilobase Million) was introduced. RPKM divides counts by transcript length and by the total number of reads. However, those measures did not account for the possibility of a different transcript length distribution in another sample. That is why TPM (transcripts per kilobase of base million) is becoming more and more popular. The difference is that instead of division by the total number of reads, it uses the sum of reads normalized for the length of the transcript [62]. As far as RPKM, FPKM and TPM account for library size and gene length biases and allow comparison between samples, they perform

poorly on data that are skewed by highly expressed features [15]. Examples of methods that account for the high variability in the data are `TMM` [85] and `DESeq2` [71].

There is yet another set of factors that can cause biases in RNA-seq data analysis and cannot be addressed with the normalization methods described above. In addition to known confounders, such as library size, the data could also be affected by unobserved factors. This reflects the variation related to, for example, different laboratory, protocol (GC content, evenness of the gene body coverage, nucleotide composition), date of experiment [69]. In molecular biology, the sources of non- biological variation are usually denoted as batch effects; however, the exact definition of this term is a challenging task. As stated in Lazar et al. [65] there are at least five different definitions. To avoid further confusion, I would like to define batch effect as a known, non-biological source of variation resulting from processing samples in different bundles. Other, unknown sources of non-biological variation will be denoted as hidden confounding factors. Popular approaches for detecting (hidden) and correcting (hidden and known) confounders include `PEER` [93] and `SVAseq` [66].

Even after applying all the corrections described before, there could still be a need for additional filters to avoid a high eFDR. For RNA-seq data, it is advised to apply not only filter for small effect size (fold change), similarly to microarrays, but also for expression levels[22, 23, 69].

Due to the complicated nature of RNA-seq experiments, the measures

of expression for the same gene under different conditions cannot be directly compared. There are several reasons behind this. We cannot be certain about all the existing kinds of RNAs in the total DNA, since what we take for an experiment is a statistical sample giving us only relative mRNA levels (relative to other mRNAs present in the current sample). Another ambiguity is introduced by the fact that reads can align to multiple places, and mRNA levels also change over time, so we must ensure that the change we observe is due to change of conditions indeed. Similarly to microarrays, statistical modeling is used to solve this issue. The main difference for differential expression analysis between microarrays and RNA-seq is that the latter generates discrete count values, rather than a continuous signal. That is why the statistical approaches used for microarrays cannot be applied unless a proper transformation of the counts is performed. An example could be using limma algorithm, originally developed for microarrays, with the `voom` transformation [64]. Random sampling of RNA-seq reads causes noise visible in variability between technical replicates, which can be modeled quite well by Poisson distribution. However, the variability can get even higher when the samples are taken from different individuals. Thus, read counts are very often modeled with negative binomial (NB) distribution (overdispersed Poisson distribution) [62]. Both `DESeq2` [71] and `edgeR` [85] algorithms use this approach. There is also a group of approaches that do not make any assumptions about the underlying distribution of the data and perform statistical testing based on ranked gene

lists. An example can be `SAMseq` [68] and `NOIseq` [96] methods. It can be beneficial to consider nonparametric methods for experiments with a sufficient number of biological replicates (at least 5-10) [62, 28].

As we can see, the complex nature of RNA-seq experiments involves many possible challenges that must be considered during data analysis. The amount of available tools and approaches, even though thoroughly described and benchmarked in many scientific articles, can be overwhelming. The researchers claim that the correct combination of methods leads to high robustness and reproducibility of the RNA-seq data analysis results [108, 23]. However, it should be noted that every RNA-seq experiment might potentially have different blend of methods giving optimal results, thus it is not possible to construct an all-purpose approach.

Chapter 3

Methods and data used in analysis

3.1 General pipeline overview

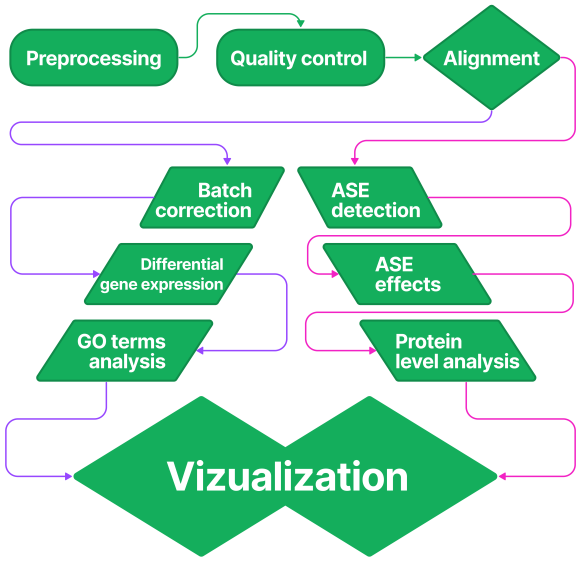


Figure 3.1: General pipeline overview.

The pipeline uses the `Snakemake` workflow management system [34] to enclose commands used for different parts of the analysis and to run every stage automatically for desired samples. It is faster than simply using `bash`, provides better control over the workflow, and comes with a set of additional advantages. In terms of flexibility and exploratory analysis, it is better than workflows like `Galaxy`, which are a great tool for users willing to automate some routine analysis, but with little knowledge in terms of computer sciences.

`Snakemake` comes with three very important features: scalability, reproducibility, and transparency. Scalability refers to the fact that it allows for running tasks on different amount of available resources and different sample sizes. It also automatically decides which jobs can be run in parallel, depending on the resources needed by the job and those available. Reproducibility denotes that results generated are the same between different runs on different systems, given that settings remain the same. Snakemake workflows are written in a way that complex tasks (like alignment) are broken down into particular jobs (such as reference indexing, file decompression, actual alignment, sorting, quality control, and quantification). One jobs outputs are following ones inputs. Such an approach makes all the analysis steps understandable and transparent. It also allows for control over specific parameters via the configuration file, which reduces the possibility of losing information about options used in case of large and often repeated analysis. A very useful option, especially in case

of analysis requiring many tasks run on a huge amount of samples, is the dry run mode, where `Snakemake` does not run the workflow but resolves all the jobs and the order of running them, providing information whether or not the pipeline is correct.

Almost all pipeline's dependencies are installed via the `Conda` package and environment management system [3]. `Conda` is an open source software which allows for installation of packages and their dependencies in separate and independent environments. This enables running several different distributions of desired language along with all required dependencies on the same system, without the need for resolving any possible conflicts. With all requirements defined in the YAML file, `Conda` automatically builds a new environment, resolves conflicts, and downloads all dependencies. What makes `Conda` a great tool is also the fact that it allows non-administrative users to install and manage software within environment isolated from the main operating system. `Snakemake` supports conda environments and even provides the possibility to define separate environments for particular jobs.

Additionally, pipeline consists of an `R Markdown` [2] script for microarray data analysis and a set of `R Markdown` scripts for RNA-seq data analysis. `R Markdown` is a file format that enables creating dynamic documents with `R`. One of the main advantages of using `R Markdown` is once again reproducibility. It explicitly combines text and code pieces into one document. The code in `R Markdow` documents is organized within chunks making anal-

ysis steps transparent and easier to understand. A great property of these solutions is that, with the import of the `reticulate package` [97], `R` and `Python` languages can very easily be used interchangeably, written as separate code chunks within those documents. `R` comes with a variety of visualization and statistical analysis methods. It also uses `Bioconductor` [44], which is a comprehensive repository of software for analyzing data from biological experiments. While it provides great solutions for differential gene expression or Gene Ontology analysis, `Python` can sometimes be an easier and faster way for some general tasks. It can also integrate `PyEnsembl` package which provides interface for Ensembl reference metadata and also enables custom reference metadata analysis.

The pipeline makes use not only of the interchangeability of methods between sequencing and microarrays, but also of different solutions aiming into making analysis more reproducible, transparent, and automatic. The chosen tools are easy to integrate together to be used in different stages of analysis, making data flow between different steps automatic. A unique property of the presented solution is combining different stages of analysis into one pipeline. Workflows available usually combine only a few selected steps presented here. Apart from preprocessing and alignment modules for alternative splicing discovery and analysis of selected events in more detailed way with `InterProScan` and visualization module are incorporated. This step is crucial for exploring and expanding currently known gene models; however, it is not yet present in available pipelines.

The most recent version is available on the GitHub page (https://github.com/aagatam/Pipeline). This page also contains requirements for running the pipeline as well as detailed description of particular stages with outputs.

Microarray data analysis script in R Markdown consists of the following steps:

- **Preprocessing**

  This pipeline accepts files in IDAT format, which contains summarized intensities for each probe-type on an array, that is why summarization step was not necessary here. The first approach included using the BGX file supplied by Illumina as annotation and NEQC normalization implemented in Limma package [84] . The second and third option utilizes illuminaHumanv4.db package [29] as annotation and VSN or quantile normalization provided by beadarray package [30].

- **Quality control**

  To check data quality MA plots, density plots and boxplots are available.

- **Confounding factors correction**

  To account for confounding factors the SVA algorithm was used for all three sets of normalized data.

- **Differential gene expression analysis**

  Bioconductor's Limma package was used for DEGs discovery.

- **GO terms analysis**

  A Parentchild [46] algorithm with Fisher test was used with p-value cut-off of 1%.

- **Visualization**

  Possible visualization include heatmaps, PCA plots, Venn diagrams, and barplots for topGO terms.

RNA-seq analysis is divided into several `R Markdown` scripts and the `snakemake` pipeline. As FASTQ files usually take a lot of disk space, three input options are available, supporting different compression methods. There is a possibility of providing uncompressed files, fastq.gz files and also fastq.dsrc files. The last option is not as popular as the previous ones; however, it is specifically designed for effective FASTQ files compression [25]. The pipeline consists of following steps:

- **Quality control**

  Initial quality control on FASTQ files is performed with `FastQC` [4], then also alignment quality report is produced by `MultiQC` [33].

- **Alignment and quantification**

  Performed either to genome with `HiSat2` or pseudoalignment to transcriptome with `Kallisto`. Each variant consists of index building (if necessary). All intermediate files not necessary for further analysis (uncompressed FASTQ files, SAM files, unsorted BAM files) are temporary files, removed after the job is finished. `HiSat2` workflow con-

tains also quantification wit `StringTie2` [79] and preparation of input files for further analysis with `R`.

- **Alternative splicing discovery with `Spladder`**

  `Spladder` performs an alternative splicing analysis on BAM files obtained for genome alignment.

- **Protein level implications analysis with `Bisbee`**

  `Spladder` output files are prepared for `Bisbee` analysis and then `Bisbee` reports effects, peptides, and FASTA files with changed transcript for all 6 ASE.

- **Joint `Bisbee` and `Spladder` analysis**

  Pipeline automatically runs another `R Markdown` script to analyze both programs output and provides pdf report, csv, and txt files with results for interesting events and associated GO terms, as well as files used in the next step by `InterProScan`.

- **Protein level implications analysis with `InterProScan`**

  The FASTA files from the previous step are grepped for interesting events and fed into `InterProscan` to obtain protein domains information.

- **Visualization**

  Also in a form of `R Markdown` script visualization for changes introduced with the new event is available.

Alternatively, after alignment, differential expression analysis can be performed. `R Markdown` script that performs this step consists of the following steps:

- **Preprocessing**

  Either `DESeq2` or `edgeR's TMM` preprocessing is used. Also genes/transcripts with low number of mapped reads are removed.

- **Confounding factors correction**

  To account for confounding factors the `SVAseq` algorithm was used for all three sets of normalized data.

- **Differential gene expression analysis**

  Three approaches are available for DEA: `limma, edgeR and DESeq2`.

- **GO terms analysis**

  A Parentchild algorithm with Fisher test was used with p-value cut-off of 1%.

- **Visualization**

  Possible visualizations include heatmaps, violinplots, PCA plots, Venn diagrams.

3.2 Experimental study design and data

3.2.1 Real NGS data - main data set

For this study, tissue samples were obtained from the dorsal part of the lumbar spinal cord of c57/BL6 mice. The genome-wide transcriptional profiling (RNA-seq) study was performed on three batches of control (WTP) and three types of gene knockouts mice. Several mouse lines with conditional deletion of the mu (MOR) and the delta (DOR) opioid receptor and proenkephalin (PENK) within specific brain structures have been used in the study. For each group, there was a subgroup with induced neuropathic pain (PNSL) and a respective control subgroup in which a sham operation (SHAM) was performed (see Table 3.1). There were four biological replicates for each condition, so a total of 88 samples were analyzed. This data is described as 'real data' as it was not indented to be a benchmarking data set but rather a way of finding targets for neuropathic pain treatment. Thus those came with some 'real world' issues, potentially affecting downstream analysis. Namely, low signal values and complex batch effects. The following table summarizes the experiment.

	Wildtype	Knockout		
Batch/ Sample type	WTP	DLX	CMV	NAV
PENK (Batch 1)	SHAM/PNSL	SHAM/PNSL	SHAM/PNSL	-
DOR (Batch 2)	SHAM/PNSL	SHAM/PNSL	SHAM/PNSL	SHAM/PNSL
MOR (Batch 3)	SHAM/PNSL	SHAM/PNSL	SHAM/PNSL	SHAM/PNSL

Table 3.1: Study design. PENK- proenkephalin, DOR- delta opioid receptor, MOR- μ opioid receptor, WTP- wildtype, DLX- knockout in the forebrain, CMV- systematic knockout, NAV- knockout in the peripheral nerve, SHAM- sham surgery(control), PNSL-neuropathic pain.

3.2.2 Reference NGS data

For this part reference RNA samples A and B from the SEQC2 consortium [57] were used, where A is a mixture of 10 different cancer cell lines and B- healthy individual. Then, samples A and B were mixed in different ratios, which enabled validation of results based on titration. Figure 3.2 summarizes the experiment.

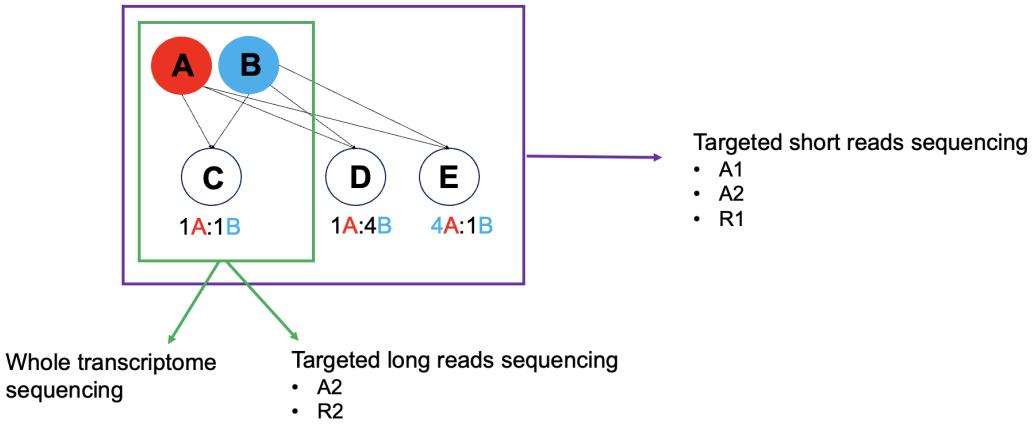


Figure 3.2: SEQC2 study design.

In the project, samples were targeted with multiple commercial and custom panels. For part of the work presented here, we used data obtained with the use of following targeting panels:

- Agilent commercial (A1) - commercial panel targeting 1064 genes,
- Agilent custom (A2) - panel design by the SEQC, combining different targets from commercial panels (eg. A1) + known oncogenes, targeting 2125 genes,
- Roche custom (R1) - panel designed by Roche to target the same genomic regions as A2.

Each sample was targeted with those panels, and then 4 independent libraries have been created for short read sequencing by Illumina. As it was a data set designed for benchmarking studies of the SEQC consortium, it is well described and the signal is designed to be strong.

In the project also complementary: i) long read sequencing (PacBio and ONT) on samples A, B, C targeted by panels A2 and R2 (subset of 564 genes from R1); ii) long read sequencing (PacBio and ONT) of individual cell lines composing sample A targeted by panel R2; and iii) long read (PacBio and ONT) whole transcriptome sequencing (WTS) of samples A and B was performed. These rich long read data sets were then used to predict, with the use of IsoQuant [82] possible new transcripts which were then rigorously filtered based on encoded to study design ground truth to remove possible false positives. We have resulted with about 70k new alternative tran-

scripts, over 8k are ones from genes located on targeting panels. Obtained in this way set of new alternative transcripts is used here as an extension to the comprehensive AceView annotation. The SEQC2 study is ongoing and the results are not yet published, thus more details from the study cannot be provided in this thesis.

3.2.3 Microarray data

Microarray data were obtained from seven patients suffering from Parkinson's disease and also from seven healthy volunteers. Analysis was performed using Illumina HumanHT-12 v4 microarrays. As these arrays are designed to target specific transcripts, whole analysis was done at this level. There were fourteen samples, but as those microarrays consist of twelve lanes, that is why two samples were run on a different array (Healthy 6 and 7).

3.3 Alignment and quantification programs

3.3.1 Kallisto

`Kallisto` [14], along with `Salmon` [78] and `Sailfish` [77] is one of the alignment- free quantification methods. It introduces an idea of pseudoalignment, which assumes that the exact place in the transcript where the read

is coming from is not relevant; what matters is only the transcript itself. With removing the need for alignment, `Kallisto` reduces the time necessary for read processing, which is the major bottleneck in RNA-seq analysis [14].

There are several steps in the `Kallisto` algorithm. Beforehand an index is built with use of the transcriptome de Bruijn graph (T-DBG). Each transcript is represented as a set of k-mers, and the index stores information about their original transcipt(s) and positions in the form of a hash table. Each node in T-DBG is a k-mer and can be associated with more than one transcript, which is referred to as the k-compatibility class. To align reads, each one is also decomposed into k mers, which are used to find a matching path in T-DBG. Another adjustment that improves speed is to skip redundant information. When a read k-mer is matched, `Kallisto` skips neighboring k-mers, using the k-compatibility class of the node as a look-up, because they often belong to the same transcript. Another improvement is introduced by the fact that `Kallisto` assigns reads to transcripts and also quantifies their abundances at the same time [14].

Despite the fact that `Kallisto` does not perform a standard alignment according to the authors and also the follow-up papers, it is still very accurate and fast at the same time [54, 32, 108].

3.3.2 HiSat2 + Stringtie

`HiSat` [60] is an example of a splice- aware aligner that uses the genome as a reference. It provides several improvements to speed up the alignment process.

`HiSat` introduces a new hierarchical indexing strategy based on the Burrows-Wheeler [16] transform and the FM index [37]. Indexing is performed in a very similar way to `Bowtie's` FM index, but with the difference in using two different indexes:

- global FM index for the whole genome,
- many local FM indexes of about 64,000 bp, that together cover the whole genome.

It also provides three categories of exon- spanning reads:

- Long-anchored reads with at least 16 bp aligned in each of the read,
- intermediate- anchored reads with 8-15 bp in one exon,
- short-anchored reads with 1-7 bp in one of the exons.

The latter two categories are those that provide the main challenges in aligning correctly and also according to [60] take up to 30-60% of the total run time for other aligners. Here `HiSat` takes advantage of the different types of indexes and also of splice sites information, either found by previous alignments or already known ones. First, a global index is used

for part of the read to find its possible location in the genome, and then one of the local indexes is used to align the remaining part of the read.

Another improvement was introduced with the `HiSat2` version, which uses a graph-based FM index (GFM) [61]. This explains the fact that the reference genome was built with information from a small number of individuals, 70% of which come from a single person , which does not reflect the genetic diversity between individuals and populations [63]. With the graph approach, it is possible to make use of extensive information available in public databases and expand the reference with additional data that contain information about different genetic variants.

In order to assemble transcripts and genes and assess their abundances from short aligned reads, one of the options is to use `StringTie` software. It allows genome-guided transcriptome assembly combined with concepts of de novo genome assembly and estimation of expression levels for genes and transcripts [79]. According to the authors, 36-60% more transcript than with `Cufflinks` is correctly identified. Furthermore, the expression levels estimated by `StringTie` showed a higher agreement with the true values [80].

`StringTie` assembles transcript fragments and infers about isoforms. It can also leverage annotation files to infer those isoforms with greater confidence. A network flow algorithm borrowed from optimization theory is used to reconstruct and quantify transcripts at the same time. `StringTie` assembles the splice graph and then calculates the abundances of each

annotated transcript by calculating the maximum flow through the network. Next, this isoform is removed and maximum flow is recalculated for the next most common isoform. As a result, we receive concluded annotations and estimated expression levels [80].

3.3.3 Magic-BLAST + Salmon

`Magic-BLAST` is another splice-aware alignment tool that is used for fast and accurate mapping of both short and long reads against a genome or transcriptome. It also allows for accurate mapping of introns, which is a rare trait [13].

What makes it different from many other aligners is that it does not build one index, instead it builds an index for a batch of reads and then runs it against BLAST database to search for matches. At first, it looks for a perfect 16 base match (seed alignment). In order to avoid ambiguous matches, a selective masking technique is used. Original 16-base matches are not indexed in the lookup table if they appear in the reference more than a given number of times (by default, 60). In addition, seeds with more than 15 A's or T's are also masked out. The next step is to expand the match to the length given by the user using a simplified greedy alignment extension procedure. For paired reads, the sum of the quality of the pair is taken to select the best match [13].

Again, to obtain transcripts and genes abundances a specific software

is needed. `Salmon`, as mentioned before, is another of the alignment-free quantification methods. It can work in two ways- either performing quasi-mapping (indexing + quantification) of FASTA/FASTQ files or perform quantification using pre-computed alignments to transcriptome (in this case by `Magic-BLAST`) from BAM/SAM files [78].

Quasi-mapping is based on a concept similar to `Kallisto's` pseudoalignment called lightweight alignment. The difference is that, in fact, it tracks the approximate location and orientation of all mapped reads. According to the authors, this piece of information is crucial for accurate quantification. To find that position, `Salmon` uses chains of maximal exact matches (MEMs) and super maximal exact matches (SMEMs), which can be computed very efficiently [78].

3.3.4 Comparison

There are a couple of issues that one should consider when deciding on alignment and quantification tools. `Kallisto` is a very good choice if what we are looking for is solely quantification. It is very fast, due to the lack of alignment process, but at the same time have performance comparable with standard approaches [108]. An additional benefit is the fact that `Kallisto` does not produce SAM/BAM alignment files and thus requires less resources not only in terms of computational power but also available disk space.

Alignment- free approaches are not enough in the case of studies reaching beyond quantification, such as alternative splicing. `HiSat2` is a good option in this case, as it has the property of identifying splice junctions and is also currently the fastest splice aware tool available. As `HiSat2` maps reads against the genome, it has to detect exon- exon junctions. It is no longer an issue for the third option- `Magic-BLAST` used with alignment to the transcriptome and `Salmon`. Aligning directly to the transcriptome removes this issue and also gives better results for data with weaker signal. For the purpose of this work, `HiSat2` is a sufficient option and thus was incorporated in the pipeline, along with `Kallisto` for lightweight analysis option.

3.4 Methods for confounding factors discovery and removal

3.4.1 SVAseq

`SVA` stands for surrogate variable analysis. The concept was introduced in 2007 to identify and remove unknown sources of variation in genomic data and was initially designed for microarrays. It enables to capture, model, and also remove all possible variables (known, unknown, and latent) affecting the value of interest by looking simultaneously at all expression levels. Surrogate variables estimation is performed using the iteratively re-

weighted least squares approach [67].

SVAseq is an extension of this method that aims to analyze the count data derived from sequencing experiments. To account for the type of data, a moderate log log transform is applied prior to the calculation of surrogate variables [66].

3.4.2 PEER

Probabilistic Estimation of Expression Residuals (PEER) is another tool used for the discovery and removal of unwanted variations. It is a collection of Bayesian approaches combined with factor analysis methods. The assumption is that those latent factors have a global effect and affect a large portion of all genes. PEER first estimates hidden factors from expression data and then incorporates them into the analysis along with known and measured confounding variables [93].

3.4.3 Comparison

PEER and SVA were both top tools for confounding factors discovery and removal according to the SEQC article [69]. Since that article was published, SVA was extended and SVAseq algorithm was introduced, that is tailored to be used for RNA-seq data. Another argument in favour of SVAseq is that it is available as an R package and detected confounders can easily be combined with further algorithms for differential gene expression.

3.5 Methods for differential gene expression analysis

3.5.1 Limma

`Limma` is an `R/Bioconductor` staple package when it comes to statistical genomics. It provides not only methods for differential gene expression discovery but also a variety of approaches for modeling and visualizations for microarrays, RNA-seq, protein arrays, and other types of data [84].

`Limma` was originally developed for microarrays and thus provides many ways to preprocess this kind of data, including reading in and normalization of different types of arrays. However, several improvements have been made over the years, so that after initial steps all downstream analysis methods are now available not only for microarrays but also for other platforms. This includes RNA-seq differential expression and splicing analyses, which will be discussed in the current chapter [84].

There are several statistical principles that `Limma` integrates that make it one of the most effective and frequently used approaches for high- throughput expression studies. Although `Limma` originally applies `Quantile` normalization, it is recommended to use `TMM` approach for RNA-seq data (described in Section 3.5.3). Due to the discrete nature of the RNA-seq data, prior to analysis counts are converted to the log scale, and mean-variance trend is estimated and subsequently converted into precision weights and

incorporated into the analysis. This process is called the `voom` method [64]. `Limma` then fits a linear model for each row (gene or transcript) in the data set but at the same time borrows information between those genes and thus allows for different variability levels between targets and samples. To achieve that, the Empirical Bayes (EB) method is used. The estimated variance of the genes becomes a compromise between the measure obtained for the gene itself and the global variability across all genes. This procedure might be sufficiently influenced by genes with very low or small variances. To avoid this a robust EB procedure was introduced, which incorporates mean-variance trend into global variance estimate. Genes with extremely low or high variances are identified and treated as outliers. This approach makes results more reliable even for small sample sizes and enables measuring possible correlation between samples or genes [81].

3.5.2 DESeq2

`DESeq2` algorithm basic assumption is that majority of genes are not differentially expressed. It uses the "median ratio method" for normalization. Each gene's counts in each sample are divided by its geometric mean across all samples. This corrects for both library sizes and also differences in RNA composition between samples [71].

The counts are modeled by Negative Binomial distribution, where the dispersion is estimated as the maximum of fitted value for each gene and

the gene- wise estimate. Finally, EB is used to shrink the gene-wise dispersion estimates towards the fitted values to obtain the final dispersion values. Wald test is used for differntial expression testing.

`DESeq2` also automatically detects and removes outliers using Cook's distance. In addition, it removes genes with low counts (below the threshold determined by an optimization procedure) [71].

3.5.3 EdgeR

The `edgeR` package uses weighted trimmed mean of the log expression ratios (`trimmed mean of M-values values- TMM`) [86]. It assumes that majority of genes are not differentially expressed and excludes highly expressed or variable genes. Then a weighted average of the remaining genes is used to calculate the normalization factor.

In the next step data is modeled using NB model, which accounts for biological and technical variation. The degree of overdispersion is modeled and then shrunken towards the common or trended dispersion, obtained from information borrowing between genes, with an empirical Bayes method. Differentially expressed genes are detected with exact test, similar to Fisher's exact test adapted for overdispersed data or with generalized linear model likelihood ratio test [85].

3.5.4 Comparison

No statistical modeling can fully capture biological variance present in the data. Each algorithm came with assumptions that may or may not be satisfied, and depending on the data sets one of them might perform better and capture more of the true signal, but different algorithms results are not mutually exclusive. `Limma, edgeR` and `DeSeq2` are among the most popular choices for DEA, they are all proven to provide reliable results for complex designs and smaller number of biological replicates [62]. Depending on the situation one might consider taking an intersection of results for all three methods, or just take the approach giving the biggest number of genes. The first approach would results in the most certain list of genes, whereas the second is useful during initial screening and exploring the results.

3.5.5 Choosing differentially expressed genes

As mentioned in Section 2.2 a corrected p-value is used to decide whether a gene is differentially expressed or not. There are two approaches for p-value corection.

- Familywise Error Rate (FWER) is the probability of at least one type I error among all rejected hypothesis [45]. An example of correction method is Holm approach [52].
- False Discovery Rate (FDR) is the proportion of type I errors among all rejected hypothesis [45]. Here the most popular approaches are

Benjamini- Hochberg (BH) [7] and Benjamini-Yekutieli (BY) [8] corrections.

The most popular approach is to use Benjamini- Hochberg correction with threshold for the p-value set for 5%. BH correction is the least stringent one and also provides good balance between finding truly differential expressed genes and limiting False Positives. On the other hand, it assumes individual tests to be independent of each other, which is not necessarily true for genes and transcripts. That is why BY correction is a more correct approach, as it does not make such assumption [45].

There is no real reason behind setting a p-value thershold for 5%, simply it has to be set somewhere and can be tweaked if we would like to change the number of genes detected. While it is a proper approach to shorten our gene list and reduce the number of False Positives, it is not a correct value to sort this list by. The p-value does not inform about how "big" the effect is, it is only a way of indicating the probability of obtaining the effect of given size by chance. Using only p-value ranking might lead to false conclusions and lower reproducibility. It is the combination of p-value and logarithm of fold change (logFC) that gives most reliable results [22, 89].

Minimum Significant Difference (MSD) is an example of more complex method of sorting gene lists, than simply using logFC. It can be described as the worst possible logFC estimation, within 95% confidence interval (CI). For positive logFC values it is the lower CI boundary, and negative of the upper CI value otherwise [106].

3.6 Functional analysis methods

The Gene Set Enrichment Analysis of differentially expressed genes aims at obtaining the potential biological meaning of the experiment. Exploring Gene Ontology annotations is one of the most popular methods to receive this information.

Gene Ontology is a major bioinformatics initiative to store and unify the vocabulary to describe the roles of genes and gene products in many organisms. There are three independent ontologies available: biological process, molecular function, and cellular component [5].

Genome annotation is the process of attaching biological information to sequences. After obtaining a list of genes that have significantly changed their expression, it is possible to annotate them with associated GO terms. The `Bioconductor package go.DB` provides us with detailed information on the most recent version of Gene Ontologies, while the `topGO package` [1] is designed to enable enrichment analysis of GO terms, as well as interpretation and visualization of results. There are many algorithms and test statistics to extract relevant GO terms provided by the package. For this project, the Fisher test was used with the Parentchild algorithm [46].

GO terms form a direct acyclic graph (DAG). In this graph, the nodes represent individual terms. Direct edges connect nodes in such a way that each term is a more specific child of one or more parents. The graph goes from less to more specific nodes. There are many algorithms to account

for the GO topology. So far, the most common method was term-for-term approach. It assumes that if a gene is assigned to a term, it is also assigned to all parents of this term. This approach suffers from overlapping annotations. Each GO term shares all the annotations of all of its descendants; in addition, individual genes might be associated with multiple unrelated terms that are connected only by the root term. What differs between this approach and the Parentchild algorithm is the definition of the analyzed sets. Parentchild algorithm is applied only to a child and its parent (or parents). It is a better approach because some graphs have a very complex structure, whereas others do not [46].

3.7 Methods for alternative splicing discovery

3.7.1 Spladder

`Spladder` is a software which allows for detection and quantification of novel and existing in current annotation splicing events. It also allows for differential testing of events as well as provides a module for splicing variation visualizations. What is unique for `Spladder` is that rather than focusing on whole transcripts, it focuses solely on alternative soling events [58]. Although it was designed for short read data, our initial tests showed that it might also be used for long reads.

Based on the current annotation, `Splader` builds a splicing graph and then expands it with events detected in provided alignment files. Currently, `Spladder` supports detection of six canonical types of ASE: exon skip, intron retention, alternative 3' and alternative 5' splice sites, multiple exon skips, and mutually exclusive exons [58].

3.7.2 IsoQuant

`IsoQuant` is also a tool for alternative splicing discovery, however it is specifically tailored for long reads. It can either take an annotation file, reference genome and alignment files, similarly to `Spladder`, or can perform alignment using `minimap2`. It also does not only focus on particular events but reports full transcripts[82].

`IsoQuant` starts with assigning reads to already known isoforms from reference. Then there is transcript quantification step, where multi-mapped reads are treated as potential new isoforms and are omitted. Afterwards uniquely mapped reads are corrected with regards to the reference. The last step is transcript model construction and novel isoforms discovery using intron graph, which is based on splice graph approach used in `Spladder` [82].

3.8 Methods for alternative events implications analysis

3.8.1 Bisbee

`Bisbee` is a program specifically designed to handle `Spladder` output files. It enables differential splicing analysis, splicing outlier analysis, and splice isoform protein sequence prediction [48]. `Bisbbe` needed to be customized for this project as similarities rather than differences were studied. `Bisbee` output files were used to determine which isoform was present or not in the reference genome. Also information about predicted effects on the protein levels was used. `Bisbee` also enables us to generate FASTA files containing reference and altered sequences. The latter was used as an input to the `InterProScan` software in further analysis.

3.8.2 InterProScan

`InterPro` is a database of protein sequences built on information provided from a variety of resources. It integrates PFAM, Panther, PROSITE profiles and many other databases. This gives an overview of proteins families, domains and sites. `InterProScan` is a software which provides the possibility to query this enormous collection of protein information. It is available through a website, but also as a standalone software package [56].

3.9 Microarray between samples normalization methods

3.9.1 Quantile

`Quantile` normalization assumes that there is an underlying common distribution of intensities across chips. To check if two data sets come from the same distribution one can use a qqplot. The method give the data sets the same distributions by transforming the quantiles of each to have the same value. The algorithm is very simple and fast [11].

3.9.2 NEQC

`NEQC` method performs non-parametric background correction (using negative control probes) and then `quantile` normalization (using both negative and positive control probes) [90]. Background correction method introduced in this approach is similar to very popular correction method for Affymetrix microarrays- `RMA`. It uses normal-exponential convolution model to fit the negative control probes on the array. There are three different methods for parameter estimation (non-parametric, maximum likelihood and Bayesian), however non- parametric approach is simple and fast, whilst still reliable [102].

3.9.3 VSN

`VSN` method aims at stabilizing the variance of microarray data across the full range of expression. The method is useful when one needs to use traditional statistical methodologies such as ANOVA, which assume the normal distribution of the data with constant variance. The first attempt was to apply log transformation to the data. This approach indeed made the variance constant for large expression values, but it also led to problems when it comes to negative or very small values. `VSN` transformation is the logarithm at the upper end of the intensity scale, approximately linear at the lower end, and smoothly interpolates in between [31].

3.9.4 Comparison

Each normalization have different advantages and works best depending on a data set, that is why all three methods were included in the pipeline. Quantile normalization is proven to work very well, but on the other hand can, along with the technical variability, remove also interesting biological variation if the assumptions are not satisfied [49]. `NEQC`, apart from `quantile` normalization, adds additional step- background correction. According to [88] data that were background normalized tend to better reflect the real fold changes, but this correction might introduce additional variation. `VSN` normalization allows for better precision when it comes to transcripts expressed at lower levels, which tend to have larger variances.

Chapter 4

Results

4.1 Real data

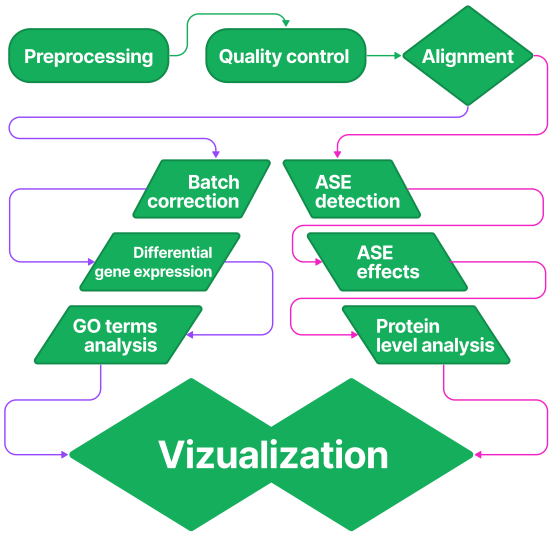


Figure 4.1: Pipeline stages used for real data set analysis.

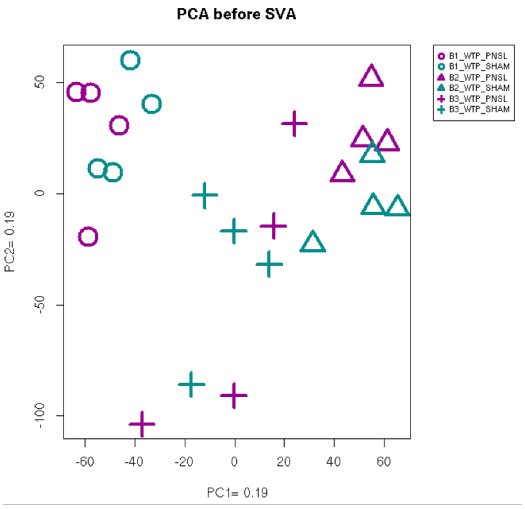
For the neuropathic pain data set the whole pipeline was run. Analysis of differential expression stems greatly for the approach developed for microarray data. The aim was to see how the combination of different guidelines, based mainly on artificially created data sets with strong signals, can improve analysis of a problematic data set.

4.1.1 Differential expression analysis

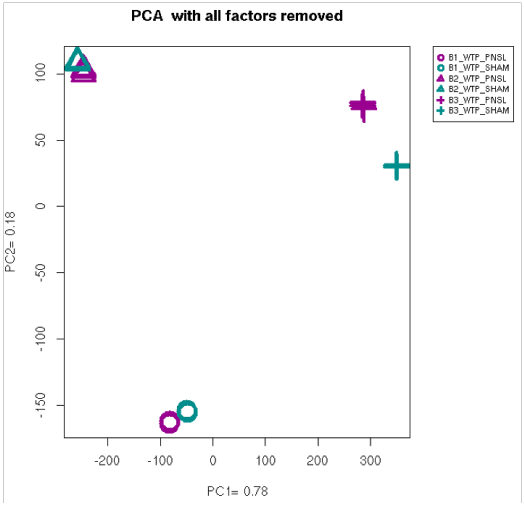
As the study design for the neuropathic pain data was quite complex, my objective was to analyze the simple difference between the control group (WTP_SHAM) and the group with induced neuropathic pain (WTP_-PNSL) for each batch. The idea was to compare lists of differentially expressed genes obtained for each of the 3 batches and proceed with the method that gives the best results in terms of reproducibility. In the first attempt, most of the tools applied for DEG discovery (`Limma, EdgeR, DeSeq2`) did not give any results. Only `Limma` showed up to 33 DEGs, depending on a batch and thus this method was chosen for further analysis.

As the data come in three runs, and also from different mice, it was obvious that it was affected by both known and hidden confounding factors. According to [108] and [69] adjusting for that should result in great improvement in reproducibility across laboratories. For this purpose, the `SVAseq` algorithm was used; however, its proper application requires a thorough rethink.

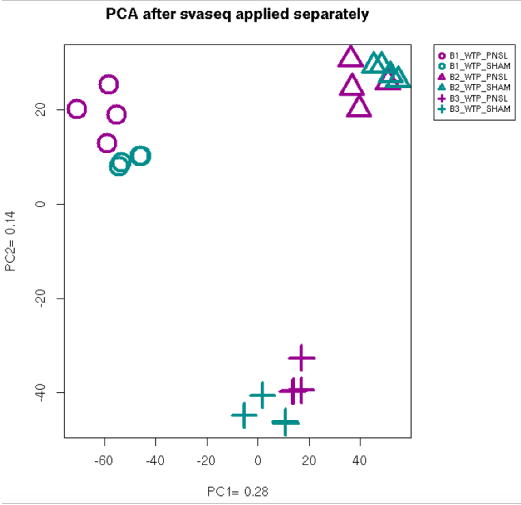
`SVAseq` asses the possible number of hidden factors and then in the next step allows one to either remove them (which should be used only for visualization) or adjust the data for further analysis. It is also possible to include known confounding factors; however, they should also be automatically detected by `SVAseq`. The first factor should account for the batch effect. The challenge is how many of these factors should be removed. If too many, there is a risk of removing not only unwanted variability, but also true changes between conditions. Our indicator for that were PCA plots and also the `Limma` algorithm producing warnings or even errors. As we can see in Figure 4.2a all groups and batches are mixed together before applying `SVAseq`. In Figure 4.2b we can see incorrect `SVAseq` use, where all factors have been removed and there is no variability between technical replicates. Proper approach is shown in Figures 4.2c and 4.2d. We can see that the unwanted variability is removed, but at the same time samples within clusters remain different from each other. The last two figures represent also another question- how to apply correction method to batches themselves. Should we indicate that SHAM and PNSL samples come from different batches (Figure 4.2c) or should we treat them together (Figure 4.2d)? The first approach causes clustering by batch and the second- by group. The answer to this question depends on the type of analysis one would like to perform-whether to compare all PNSL samples versus all SHAM samples or to seek for reproducibility between batches.



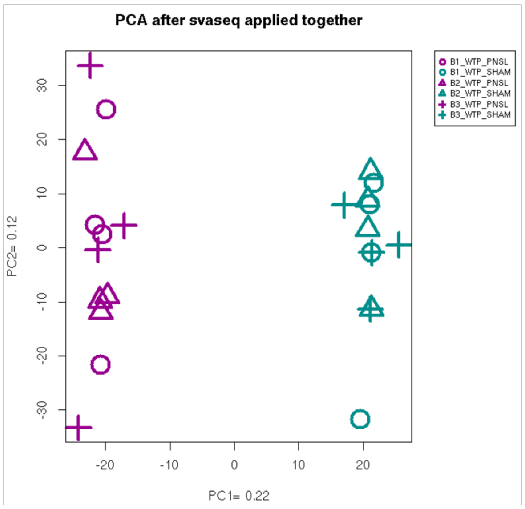
(a) PCA before applying SVAseq



(b) PCA with all detected factors removed (15), applied separately for each batch.



(c) PCA with 7 factors removed, applied separately for each batch.



(d) PCA with 7 factors removed, applied together for all batches.

Figure 4.2: PCA plots before and after different ways of applying SVAseq.

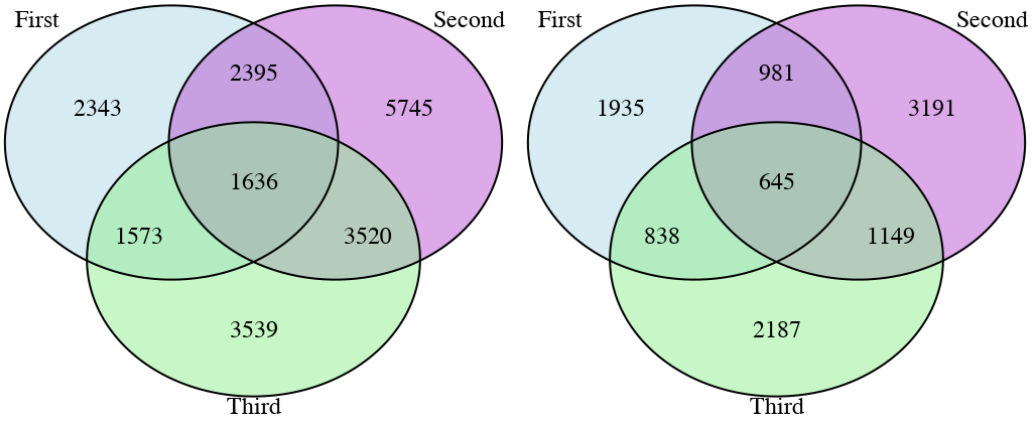
This is reflected in the reproducibility results obtained for the data analyzed. In Figure 4.3 we can see that applying `SVAseq` significantly increased the number of DEGs detected, but the reproducibility still remains at approximately 8% if applied `SVAseq` separately (Fig. 4.3a), but even lower- 6% when applied together(Fig. 4.3b). In addition a higher number of common DEGs is obtained for the separate approach, when comparing two consecutive batches in Table 4.1. A number of differentially expressed genes detected for SHAM groups comparision between two consecutive batches is presented in Table 4.2. The number is higher for the separate approach, which indicates that the other method is correct for this type of analysis. However, this does not apply to the comparison between batches second and third. This can be explained by the observation that batch 2 and batch 3 seem to be more similar than any of them with batch 1. This is visible on the PCA plots (Figures 4.2b and 4.2c) and also in Table 4.3. Later, it was confirmed that batch 2 and batch 3 samples were prepared by one laboratory, while batch 1 was prepared by a different laboratory. This causes the higher False Positive number when samples are analyzed together.

Batch	Analysed separately	Analysed together
1 vs 2	21%	19%
1 vs 3	21%	19%
2 vs 3	28%	20%

Table 4.1: Percentage of common DEGs between batches.

Batch	Analysed separately	Analysed together
1 vs 2	9870	4627
1 vs 3	8470	5978
2 vs 3	1880	5373

Table 4.2: Number of DEGs for control comparison (False Positives).



(a) SVAseq applied separately.

(b) SVAseq applied together.

Figure 4.3: Venn diagrams showing reproducibility between batches for different SVAseq approaches.

Although lists of differentially expressed genes were obtained, the reproducibility results remain low. That is why, another recommendation to add additional filter on logFC (above 1) was used. It resulted in reproducibility of differential expression calls with up to 95% concordance in DEGs according to [23]. However in this particular case the signal change is so small, that applying this filter resulted in removing a huge portion of genes and worsened reproducibility (Figure 4.4).

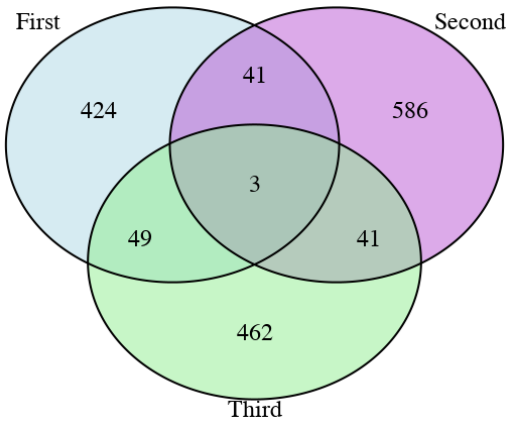


Figure 4.4: Reproducibility results obtained after setting logFC>1.

4.1.2 Global view on alternative splicing

As quantitative analysis for neuropathic pain data did not provide satisfactory biological results, focus was shifted on qualitative analysis of all 88 samples to explore the unseen landscape of the mouse transcrtiptome. By including also knockout samples, we should be able to identify novel ASEs that are specific for the spinal cord, regardless of the stress conditions. Intron retention events were excluded from the analysis because the library preparation protocol was based on ribodepletion and thus a high number of False Positives nASEs of this type could be expected due to the presence of immature mRNA. Spladder reports results for two isoforms, one containing given event, one excluding it, thus, after considering ASE already existing in the annotation, our data are divided into three groups:

- new isoform + known isoform (new+old),

- both new isoforms (new+new),
- both known isforms (old+old).

`Spladder`, among other metrics, reports the PSI (percent spliced in) value for each event and sample. This is the ratio of the signal supporting given event and sum of the signals for both events. Using PSI and appropriate threshold, False Positives number can be reduced. For this purpose plots shown on Fig. 4.5 were made. They show how number of valid events is changing for three groups and all considered types of ASEs, depending on standard deviation threshold. After analyzing it, a set of criteria to choose valid events was chosen. The first one, applied for all three groups, was setting a threshold on PSI standard deviation. Those events that had it above 0.2 in all 4 replicates were treated as valid. The reason for that is that around this value we can see the plot's elbow, where the trend is changing. As further `Bisbee` analysis is currently available only for events with at least one isoform already in annotation, it was conducted only for the group with new+old events. To focus on strong enough events which have a higher chance to not be false positives, a more stringent approach was applied and also a threshold on the PSI value was added- it should be above 0.2 (which essentially means that the glsnase constitutes at least 1/4 of already known ASE).

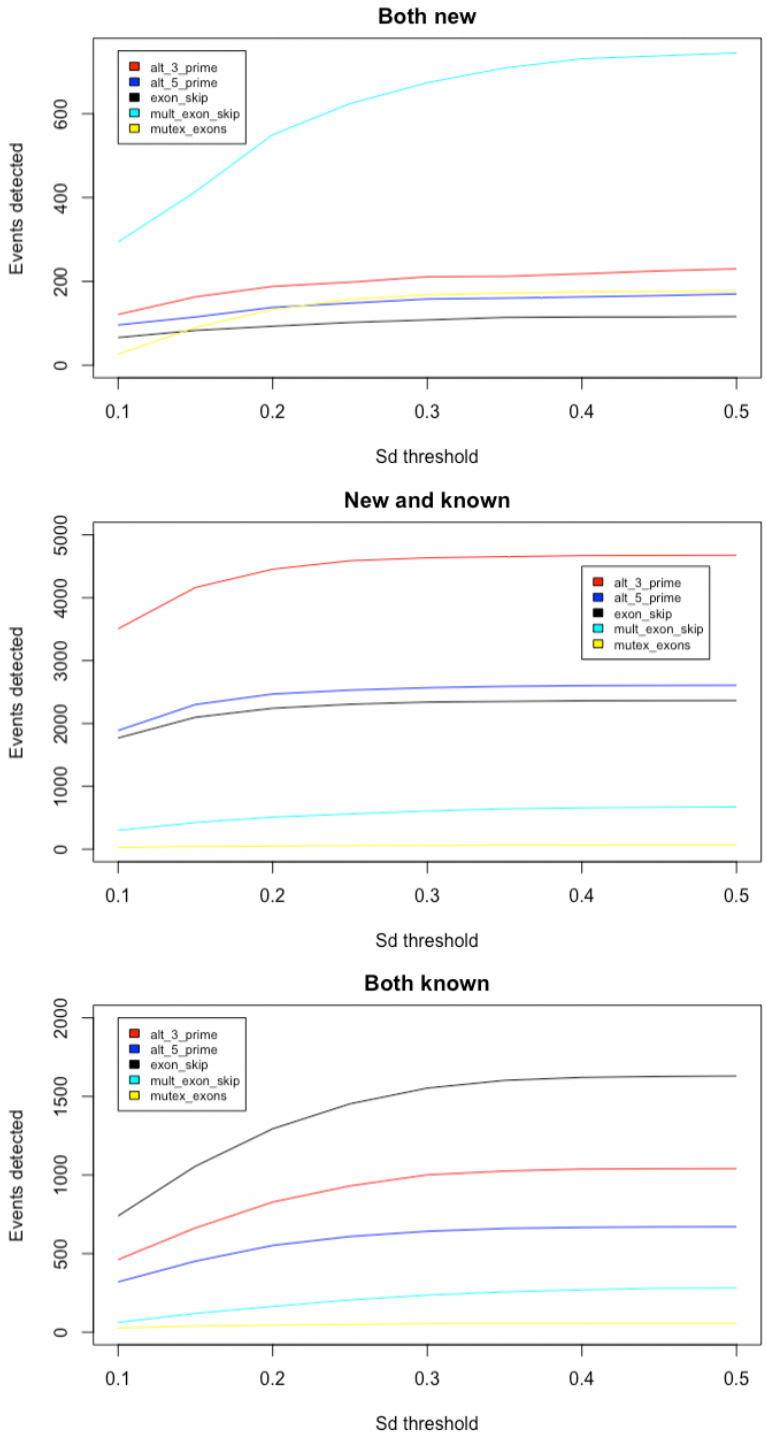


Figure 4.5: Plots showing number of events common for all 88 samples depending on a standard deviation threshold for three groups of events.

Despite the high diversity of the samples analyzed and the low reproducibility at the gene expression level, still common events for all of them were found, with standard deviation for PSI lower than 0.2 in all 22 groups (Table 4.3 and Table 4.4) reaching up to about 60% events for the new+old group are common for all samples. We can see in Table 4.3 that although mutually exclusive exons and multiple exon skip are the smallest groups, noticeably more of them pass the stringent threshold (12.37% and 5.54% respectively in contrast to around 0.16% for more numerous groups).

Although many events can occur at the same gene we can see in Table 4.4 that still a huge number of different genes is affected by alternative splicing. Majority of those genes contain only one event. This observation also applies to genes affected by events selected with an additional filter.

Event	Both iso known			Both iso new			New and old					Total
		sd<0.2			sd<0.2			sd<0.2		PSI>0.2 & sd<0.2		
	All	Common	Percentage	All	Common	Percentage	All	Common	Percentage	Common	Percentage	
Alternative 3 prime	2692	828	33.76	1154	188	16.29	8256	4453	53.49	13	0.16	12102
Alternative 5 prime	2108	552	26.19	915	138	15.08	4996	2468	49.40	8	0.16	8019
Exon skip	5776	1294	22.40	3477	93	2.67	7269	2242	30.84	12	0.17	16522
Mutually exclusive exons	95	45	47.37	305	133	43.61	97	51	52.58	12	12.37	497
Multiple exon skip	383	164	42.82	1006	550	54.67	886	510	58.89	48	5.54	2255

Table 4.3: Table showing number of detected ASEs depending on a type and group and also common number of events.

Event	Both iso known			Both iso new			New and old					Total
		sd<0.2			sd<0.2			sd<0.2		PSI>0.2 & sd<0.2		
	All	Common	Percentage	All	Common	Percentage	All	Common	Percentage	Common	Percentage	
Alternative 3 prime	2075	702	33.83	698	143	20.49	4364	2322	53.21	13	0.30	7137
Alternative 5 prime	1722	498	28.92	648	107	16.51	3295	1614	48.98	7	0.21	5665
Exon skip	3668	929	25.33	1885	61	3.24	4294	1419	33.05	11	0.26	9847
Mutually exclusive exons	86	42	48.84	131	45	34.35	89	46	51.69	11	12.36	306
Multiple exon skip	363	160	44.08	746	449	60.19	762	459	60.24	40	5.25	1871

Table 4.4: Table showing number of genes containing detected ASEs depending on a type and group and also common number of genes.

Looking closer into this for new+old group we can see in Fig. 4.6:

- black line is the total number of detected events,
- green line is the number of common events for all samples,
- the red part of a bar- intersection is the number of all events for particular sample, where all four probes met PSI criteria,
- the blue part are remaining events,
- red line- median for intersection.

This makes us sure that although there are many limitations in studied data with multiple confounding factors resulting in difficulty of proceeding with classical quantitative analysis we see very stable pattern for the existence of large group of new gene isoforms resulting in observation of new ASE for the known isoforms/transcripts. This observation is stable across all three groups (See Supplementary Figures S1, S2).

Figure 4.6: Barplots showing summary statistics for different events.

4.1.3 Known ASE

When examining a group with only events already annotated, we can see in Figure 4.7 that the overlap between genes containing common events for different types of AS is very small. This is also true for the common GO terms presented on UpSet plots in Figure 4.8. In Table 4.5 we can see that up to 61% of genes is unique for a given event. The percentages are also high for unique GO terms in Table 4.5, however, they drop noticeably for CC terms, which is expected as we are studying selected tissue.

It is worth noting that those results, both for genes and GO terms, are greatly limited by the lower number of events reported for multiple exon skip and mutually exclusive exons events. Still among 5 CC terms reported for all 5 types of events, three directly indicate the nervous system (Table 4.7). Excluding those two types gives 29 additional common terms for the remaining types, among which there are terms related with the nervous system (ex. node of Ranvier, neuron projection, myelin sheath) but also with the electron transport chain (ex. respirasome, respiratory chain complex, mitochondrial respirasome).

If we take a more detailed look at the results for GO terms, especially at the Cellular Component, we can see that for every type of event, among the top 10 terms there are many connected with the nervous system (Supplementary Figure S4).

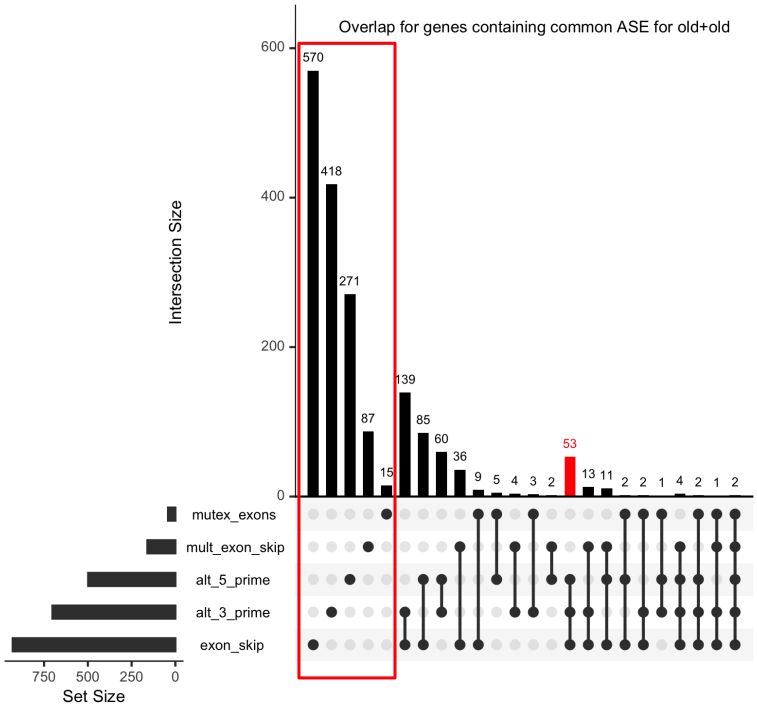
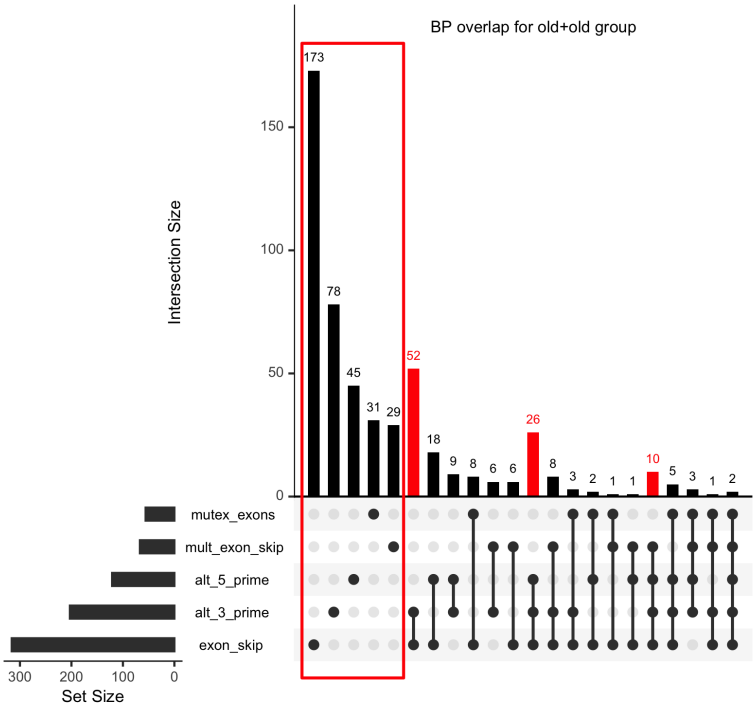


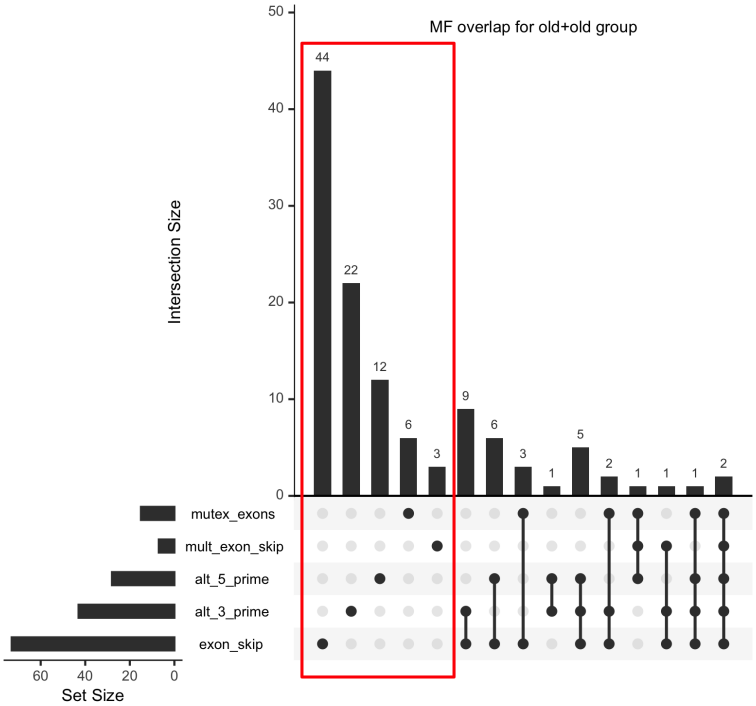
Figure 4.7: UpSet plot for genes containing common nAES for old+old group.

Type	Total	Unique	Percentage
Mutually exclusive exons	42	15	36%
Multiple exon skip	160	87	52%
Alternative 5 prime	498	271	54%
Alternative 3 prime	702	418	60%
Exon skip	929	570	61%

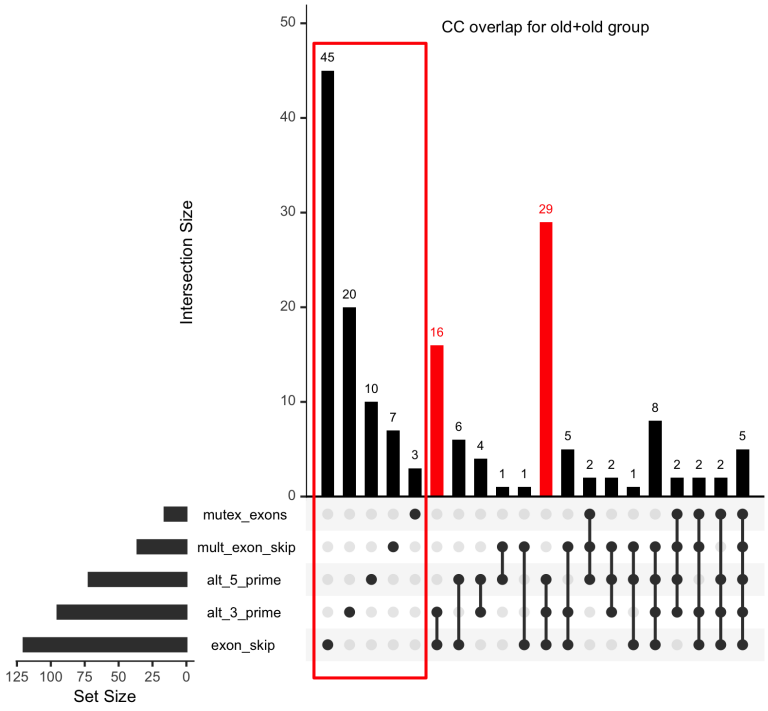
Table 4.5: Table showing percentage of genes unique for a given event.



(a) UpSet plot for Biological Processes



(b) UpSet plot for Molecular Function



(c) UpSet plot for Cellular Component

Figure 4.8: UpSet plots for GO terms for old+old group

Type	BP	MF	CC
Mutually exclusive exons	55%	53%	19%
Multiple exon skip	43%	43%	19%
Alternative 5 prime	37%	43%	14%
Alternative 3 prime	38%	51%	21%
Exon skip	55%	60%	38%

Table 4.6: Table showing percentage of terms unique for a given type of event.

Type	CC	BP	MF
Common terms	cytosol, endomembrane system, cell projection, perikaryon, presynapse	metabolic process, positive regulation of biological process	protein domain specific binding, protein binding

Table 4.7: Table showing common GO terms for different types of events for old+old group.

4.1.4 New ASE for known isoform

When examining group with new event for already existing one, we can see that the overlap of common genes and terms is low; however, here it is even more clear that this observation is limited by low number of mostly mutually exclusive exon events, but also multiple exon skip (Figure 4.9 and Figure 4.10). In every UpSet plot, we can see two peaks showing high number of common genes and different types of GO terms. The larger of those two peaks contain common terms for exons skip, alternative 3 and 5 prime ends, and the other includes also multiple exon skip group for every plot, except the one for CC terms, where the opposite is true. Again, in the top 10 CC terms we can observe terms strongly related with the nervous system (Supplementary Figure S7).

The percentage of unique genes for events, shown in Table 4.8, is lower than in old+old group, reaching only 38%. It is lower as well for unique GO terms (Table 4.9) and once again we can observe the drop in unique events for the CC terms.

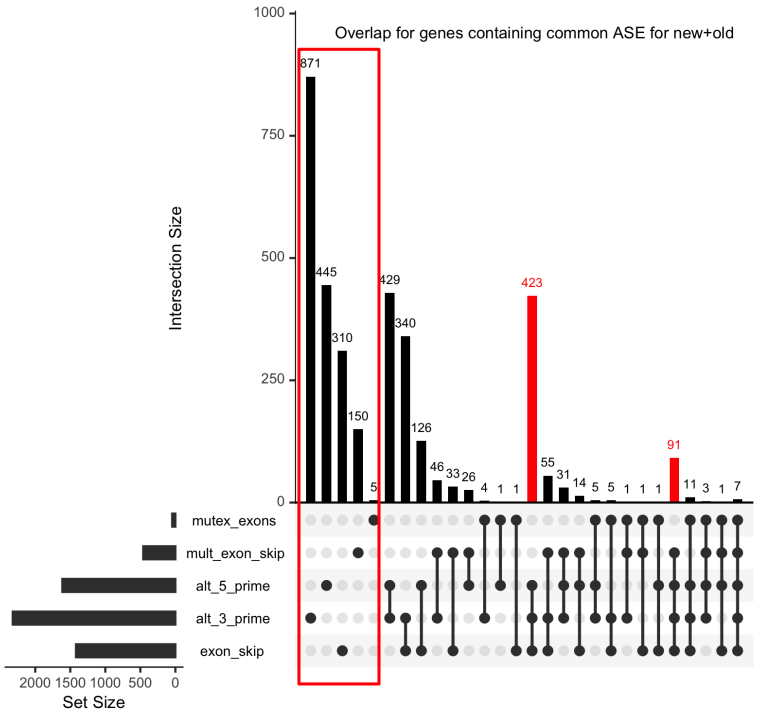
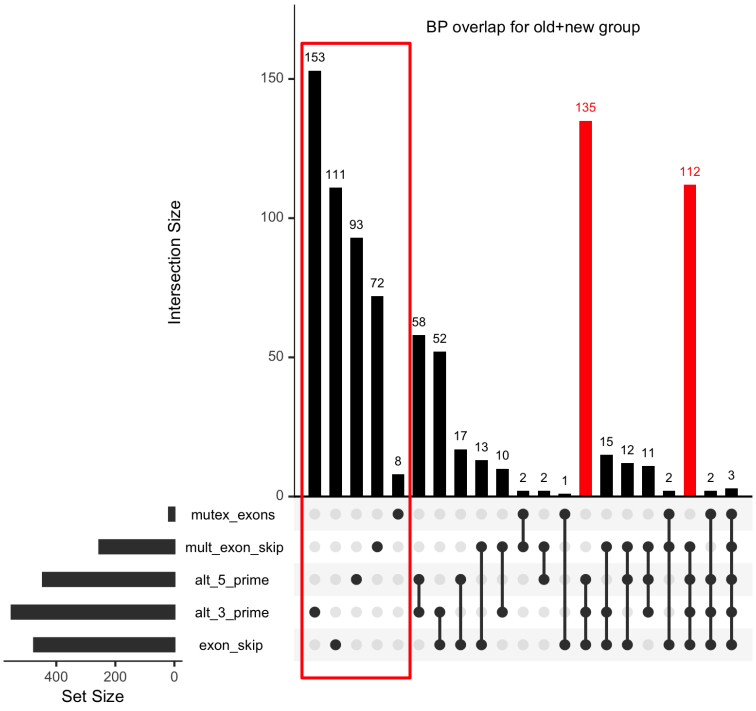


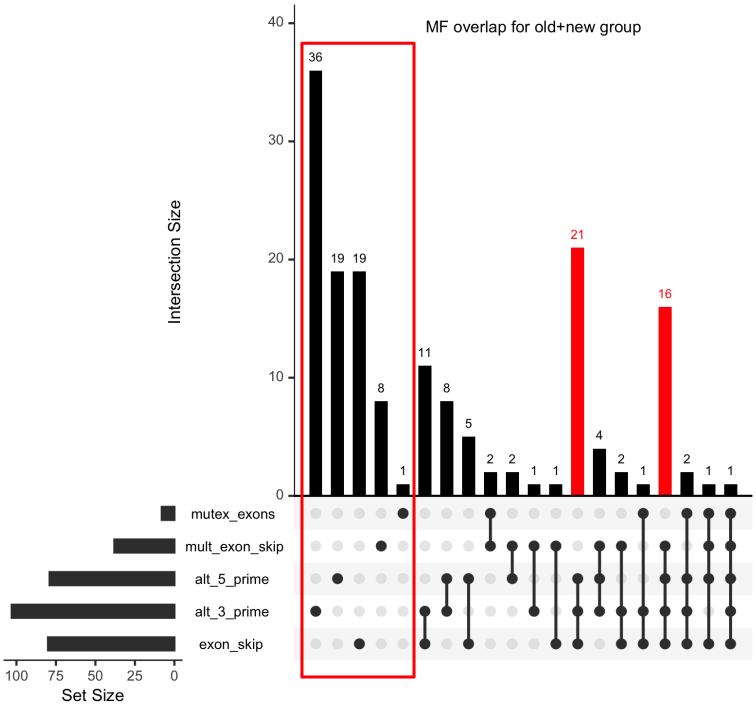
Figure 4.9: UpSet plot for genes containing common nAES for new+old group.

Type	Total	Unique	Percentage
Mutually exclusive exons	46	5	11%
Multiple exon skip	459	150	33%
Alternative 5 prime	1614	445	28%
Alternative 3 prime	2322	871	38%
Exon skip	1419	310	22%

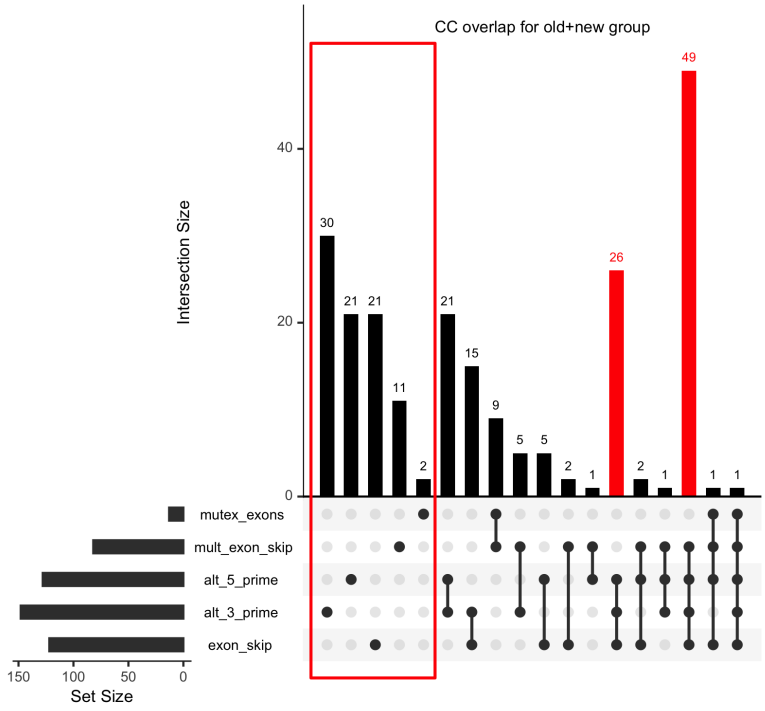
Table 4.8: Table showing percentage of genes unique for a given event.



(a) UpSet plot for Biological Processes



(b) UpSet plot for Molecular Function



(c) UpSet plot for Cellular Component

Figure 4.10: UpSet plots for GO terms for new+old group

Type	BP	MF	CC
Mutually exclusive exons	44%	13%	15%
Multiple exon skip	28%	21%	13%
Alternative 5 prime	21%	24%	16%
Alternative 3 prime	28%	35%	20%
Exon skip	23%	24%	17%

Table 4.9: Table showing percentage of terms unique for a given type of event.

Type	BP	MF	CC
Common terms	integral component of synaptic membrane	regulation of biological quality, neuron differentiation, neurogenesis	molecular adaptor activity

Table 4.10: Table showing common GO terms for different types of events for new+old group.

4.1.5 New ASE with both isoforms new

Astonishingly, for a group that contains only new events, a different pattern is observed. The lack of commonalities is not only limited to genes (Figure 4.11) but affects also different subgroups of GO terms comparisons (Figure 4.12 and Table 4.12). The characteristic peaks for the two visible subgroups for the old + old and new + old groups are no longer present on the UpSet plots. Table 4.11 shows that up to 83% of genes is unique to a given event, multiple exon skip in this case. Interestingly, the common part for the CC terms in all events is the largest of all the comparisons; among the 10 reported terms, we can once again see the connection with the nervous system (Table 4.13). In Table 4.12 we can see that the percentage of unique events noticeably drops when looking at the CC terms.

Type	Total	Unique	Percentage
Mutually exclusive exons	45	13	29%
Multiple exon skip	449	373	83%
Alternative 5 prime	107	64	60%
Alternative 3 prime	143	89	62%
Exon skip	61	20	33%

Table 4.11: Table showing percentage of genes unique for a given event.

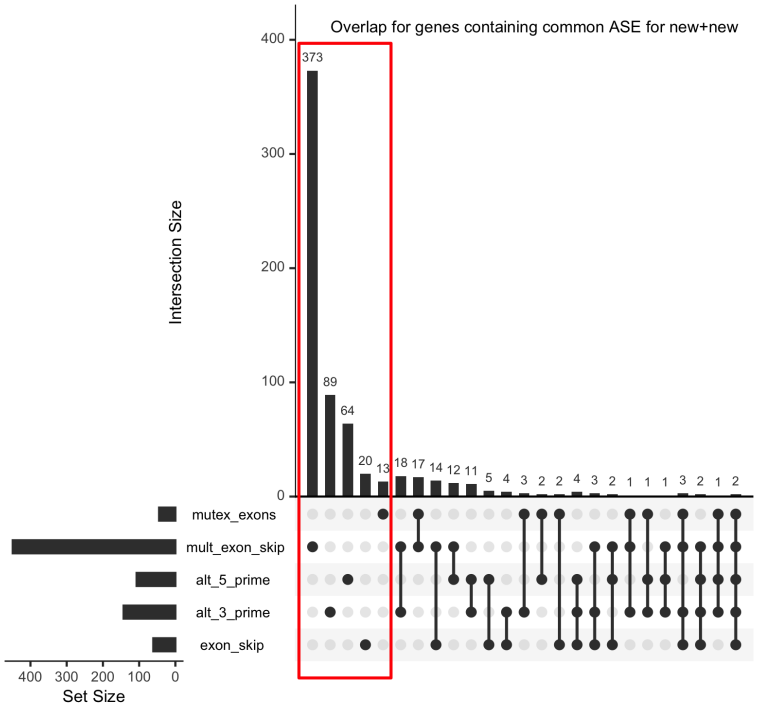
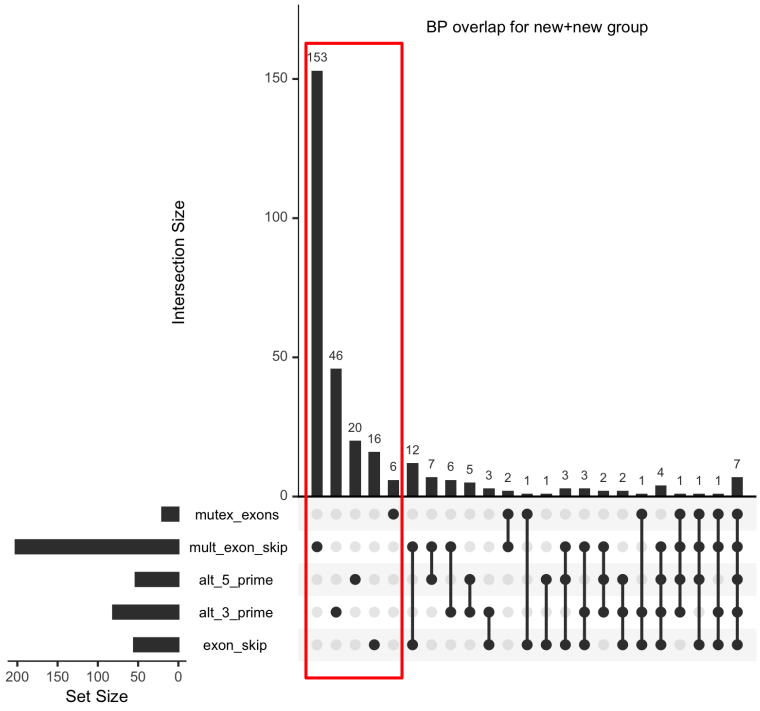
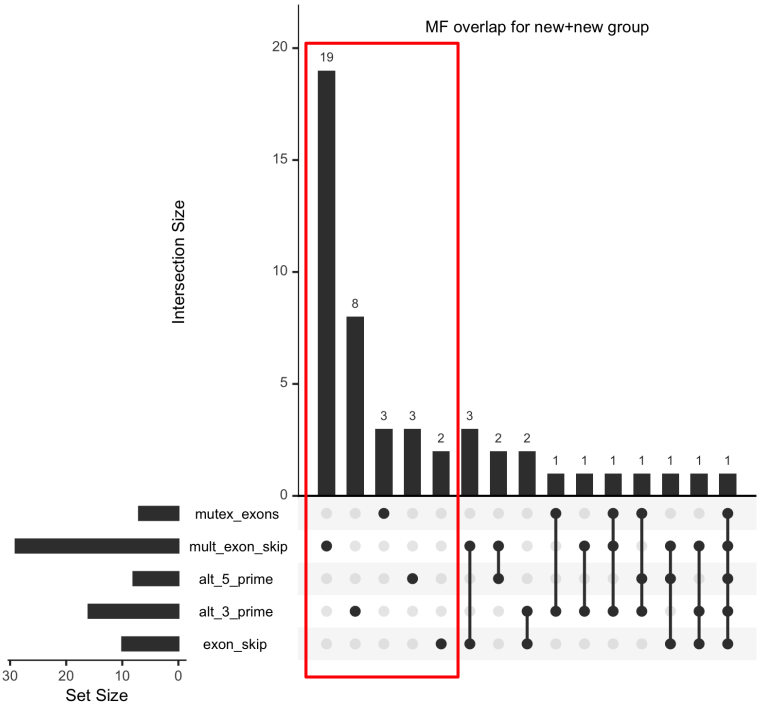


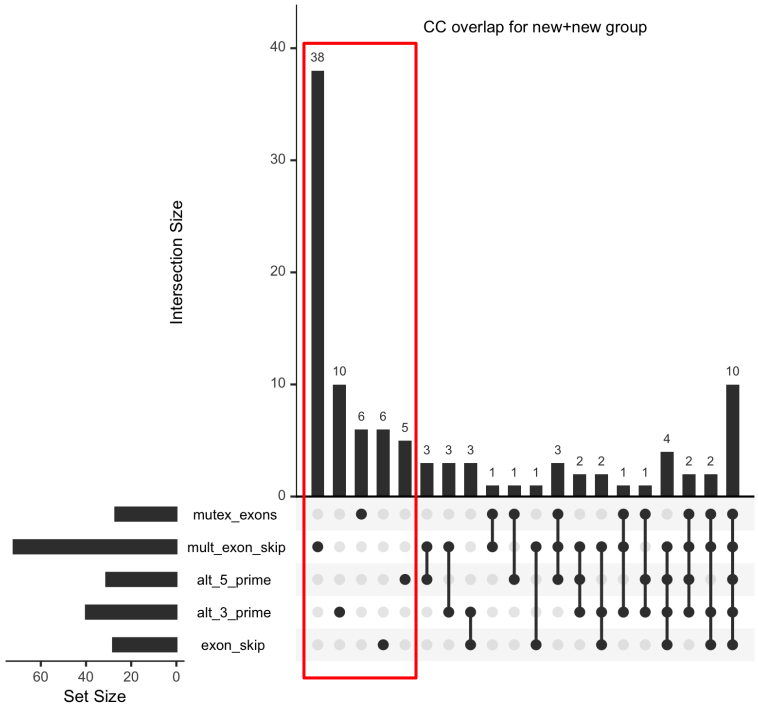
Figure 4.11: UpSet plot for genes containing common nAES for new+new group.



(a) UpSet plot for Biological Processes



(b) UpSet plot for Molecular Function



(c) UpSet plot for Cellular Component

Figure 4.12: UpSet plots for GO terms for new+new group

Type	BP	MF	CC
Mutually exclusive exons	30%	43%	30%
Multiple exon skip	76%	66%	19%
Alternative 5 prime	38%	36%	9%
Alternative 3 prime	57%	50%	12%
Exon skip	29%	20%	11%

Table 4.12: Table showing percentage of terms unique for a given type of event.

Type	CC	BP	MF
Common terms	cell junction, postsynapse, cytoplasm, cellular anatomical entity, axon initial segment, somatodendritic compartment, cell projection, cytosol, organelle, node of Ranvier	cellular process cellular component organization or biogenesis, localization, cellular localization, neuron differentiation, cell development, multicellular organism development	binding

Table 4.13: Table showing common GO terms for different types of events for new+new group.

4.1.6 Functional level analysis implications of nASE

Complementary approach, except for the top 10 most significant terms available as Supplementary Figures S3 - S11, were two kinds of plots demonstrating changes in GO terms, depending on a group analyzed. The first type shows how the top 10 terms from the old+old group change if old+new events are added and then also new+new events. The second type of plot shows the top 10 terms from every group and how relevant are they in other groups. We can observe if the number of genes annotated for that term increases and if the p-value changes. All plots are available as an additional pdf file (GO_changes.pdf).

Figure 4.13 presents top 10 Molecular Function GO terms for alternative 3 prime event and how significant they are among other groups. We can notice three terms which are not relevant in the old+old and new+old group; however, for the new+new group of events, those terms are significantly annotated. These terms are:

- structural constituent of postsynaptic intermediate filament cytoskeleton,
- phosphorylation-dependent protein binding,
- ATPase-coupled transmembrane transporter activity

The actin filaments, which build the cytoskeleton, can dynamically form different structures in response to new stimuli, which is described as experience-dependent plasticity [95]. Protein phosphorylation is a major factor in signal transduction pathways [100]. Transmembrane transporter activity could be related to signaling via neurotransmitters in the nervous system.

Figure 4.13: Top 10 MF GO terms for alternative 3' event shown across three groups.

Figure 4.14 presents the first type of plot for alternative 3 prime CC terms. We can see that the terms relevant to old+old group are still valid after adding genes for new+old and new+new groups, the difference is that the number of genes annotated to that term becomes bigger. This is another indicator that the reference annotation is incomplete.

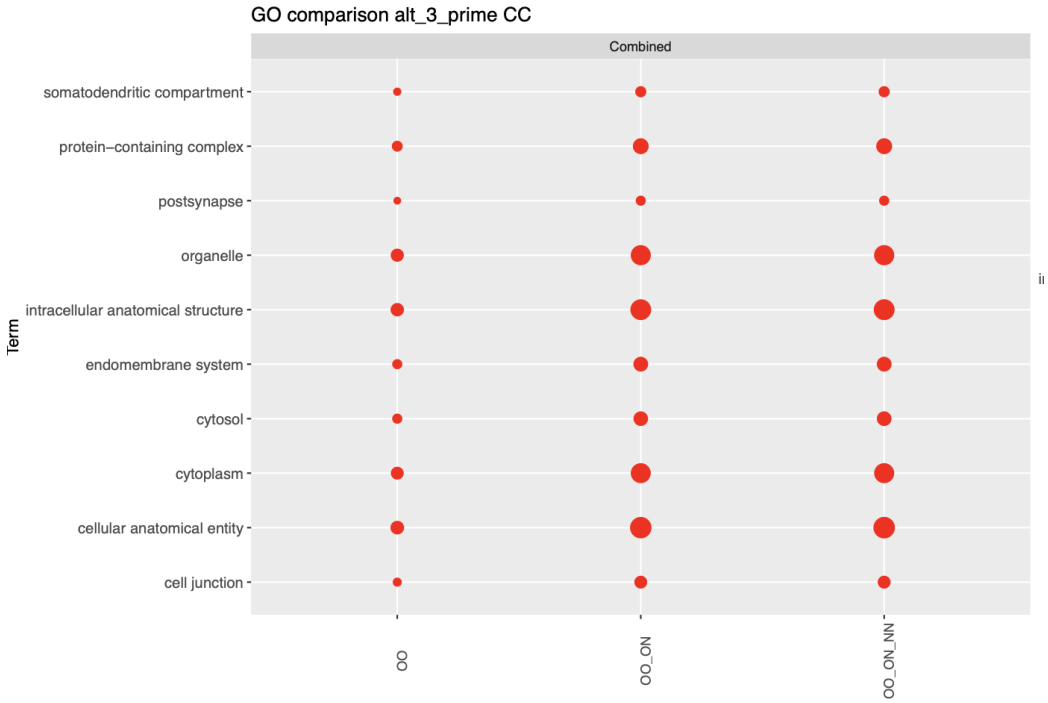


Figure 4.14: Top 10 MF GO terms for alternative 3' event shown across three groups.

4.1.7 Protein level implications of nASE

For 93 previously selected nASE of different types `Bisbee` and `InterProscan` analysis for new+old group was performed. Table 4.14 summarizes ORF and amino acid effects on proteins introduced with the new transcript version. Premature stop was mainly caused by substitution. Four events caused protein loss, and seven were silent. For most of events, `InterProscan` was able to find and assign reference protein domains. In the next step, several chosen events were visualized.

Event	Premature Stop		In frame				Protein loss		Total	Assigned by InterProScan
	Insertion	Substitution	Deletion	Insertion	Substitution	Silent	Start loss	Stop loss		
Alternative 3 prime	0	3	3	4	2	1	0	0	13	12
Alternative 5 prime	0	2	2	2	1	0	1	0	8	7
Exon skip	0	3	4	4	1	0	0	0	12	12
Mutually exclusive exons	0	4	0	0	8	0	0	0	12	12
Multiple exon skip	1	21	3	12	2	6	0	3	48	41

Table 4.14: Table showing the type of changes introduced by nASE from new+old group, and also the number of modified transcripts, which were assigned domains by InterProScan.

An example in Fig. 4.15 was made for the multiple exon skip event in Nrcam gene, which, among others, is involved in neuron- neuron adhesion and promotes directional signaling during axonal cone growth. It may play a general role in cell-cell communication. The plot is divided into 5 parts:

- additional events- other than the main investigated event events, also selected as valid ones,
- alignment track- read support provided by all 88 samples,
- detected event- transcript with novel event incorporated,

- original- original transcript,
- Interpro protein domains- domains assigned for the original transcript.

The part with event of interest is marked in red rectangular box, and the close-up is also available in Fig. 4.16. We can see that the peaks for exons reported as missing in multiple exon skip event are indeed much smaller than the peaks for rest of the exons. The protein track shows that two domains are affected by this event. They are described by `InterProScan` as neuronal cell adhesion molecule.

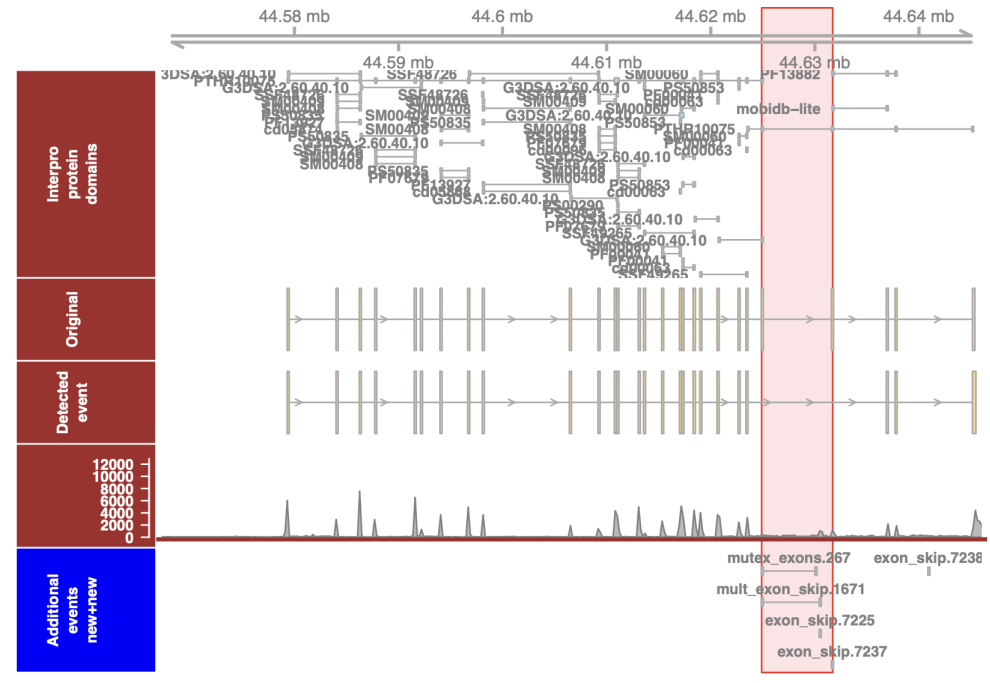


Figure 4.15: Visualization of multiple exon skip in the Nrcam gene.

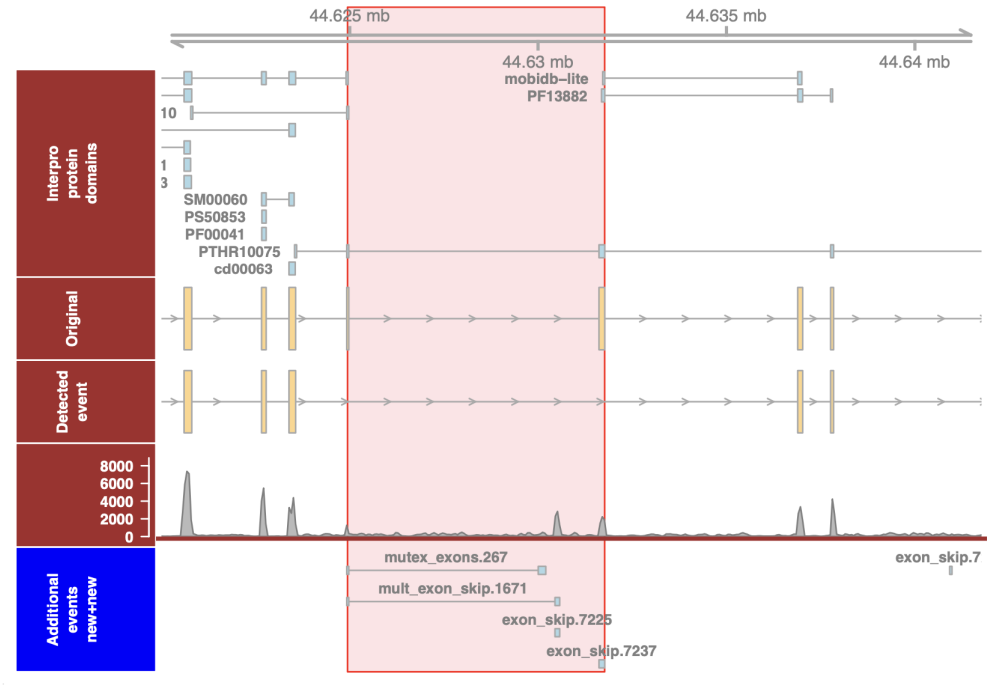


Figure 4.16: Visualization of multiple exon skip in the Nrcam gene- close up.

Another interesting event is presented in Figure 4.17. It is a mutually exclusive exon event detected in the Gria1 gene. We can see that two peaks related with two mutually exclusive exons are approximately 1/2 height of the other two exons visible on the plot. The change caused by this event results in premature stop. These exons are related to the protein domains assigned by `Interproscan` to the NMDA receptor signature, which is a glutamate receptor and ion channel found in neurons [53].

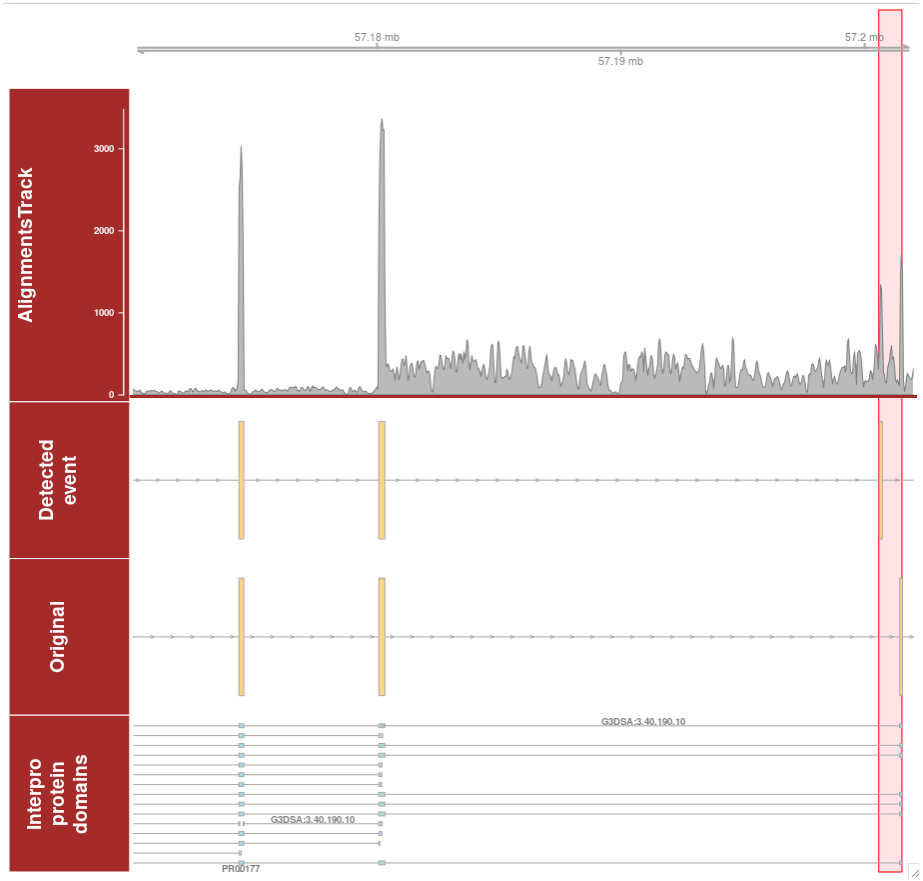


Figure 4.17: Visualization of mutually exclusive exons in the Gria1 gene-close up.

4.1.8 Discussion

Results for differential expression showed that analysis recommendations based on studies with artificially created data sets, where signal changes are high, should be applied to "real-life" problems with caution. In the case of data with low signal values, the reproducibility can vary significantly, de-

pending on a chosen method, and obtaining results as high as the reported over 90% concordance in the DEG lists [23] may not always be possible. Also, filtering methods should be adjusted to signal values, not applied arbitrarily, based on benchmarking papers.

It can be very beneficial to think about the possibility of batch effects occurring in our data and to adequately account for them. The first issue to solve would be to carefully think about the questions we need to answer and which comparisons should be made. In addition, choosing the right number of factors to adjust for is crucial.

Even though the reproducibility for differential expression analysis is very low, alternative splicing analysis revealed a lot of AS events consistent for all 88 samples. They also affect a large number of genes, but further analysis in sections 4.1.3 - 4.1.5 showed that although they appear to affect different processes and functions, they are all mainly connected to the nervous system. Furthermore, the functional analysis with the combination of events from different groups in Section 4.1.6 showed that new events might provide new information for annotation.

Still observation that different ASE types are potentially associated with exclusive sets of functions seems to be understudied. In discussion with a few experts it was pointed out as plausible, but no publication clearly talking about this phenomenon was found.

Further validation with long-read sequencing, for example, is needed for confirmation, but also for visualization of how exactly whole transcripts

look like, as we can see that many of them can occur in the same gene and overlap with each other, and here we can only study short fragments. However, visualizations in Section 4.1.7 seem to confirm, at least for the new + old group, that the reads support detected events and, in addition, they tend to occur in genes related to the nervous system and even affect important protein domains.

The results from this chapter show that even applying best practices might not be enough to receive stable and reproducible quantitative results. This is due to the complexity of RNA-seq data generation and analysis but as shown in section 4.1.2- 4.1.7 might also be caused by the incomplete mouse transcriptome annotation. The performed analysis provided, however, a number of interesting qualitative observations and can be a good starting point for further studies of alternatively spliced events, which ultimately will lead to better differential expression estimation.

4.2 Reference NGS data

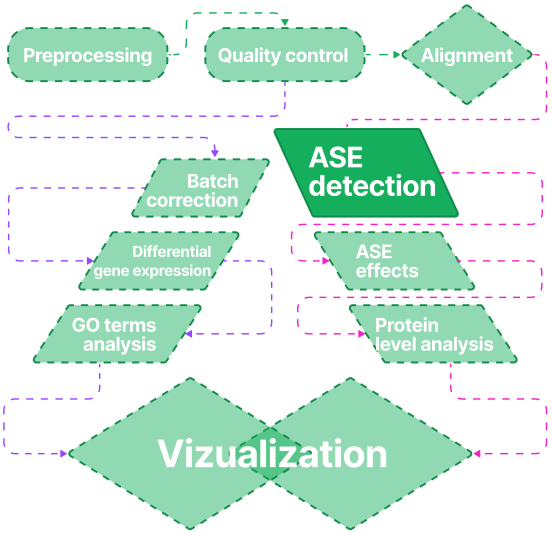


Figure 4.18: For SEQC data set only Spladder analysis was run.

Here we wanted to investigate how `Spladder` will work on the benchmarking data where the signal is suppose to be strong. As stated in Section 3.2.2 we were using data generated by targeted short reads sequencing where panels A1, A2, and R1 were used. `Spladder` was run on those using as reference either the AceView annotation or AceView extended by the SEQC2 consortium (with `IsoQuant` run on long reads, see Section 3.7.2). Figure 4.19 shows how detected by `Spladder` ASE in SEQC2 short reads data overlap with original and extended annotation, depending on the reference used. It is important to note that for reference (AceView or AceViewExtended) we report all exon-exon junctions (introns) present in the

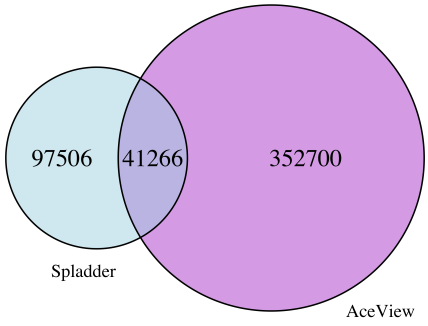
reference, while `Spladder` reports only those directly involved in the alternative splicing events.

- `Spladder` run with AceView annotation reported 138,772 junctions in total. Approximately 1/3 of those junctions are consistent with AceView and constitute 10.5% of junctions in reference (Figure 4.19a). To have a better understanding, we need, however, to focus only on genes present on the targeting panels, as those are enriched in the samples. Analysis of the remaining ones will be highly affected by the lower effective sequencing depth of those.

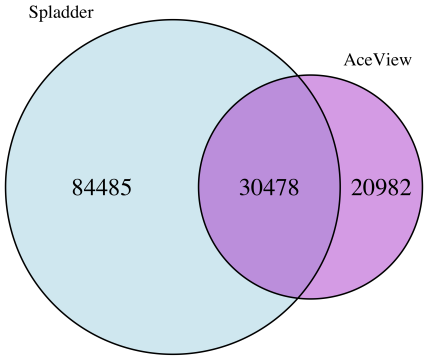
- In Figure 4.19b we focus on 2343 genes that were effectively targeted by panel A1, A2 or R1. It can be seen that about 60% of junctions reported for those genes from the AceView annotation are among those seen by `Spladder` in the short read data. This level should be considered reasonable as: i) not all junctions in a gene are involved in ASE, ii) it was estimated that about 80% of the transcriptome is active / detectable at a specific time point [107]. Interestingly, `Spladder` is seeing about 85k of new junctions which is 87% of all new junctions reported by `Spladder` (see Figure 4.19a). This from one site confirms the targeting efficiency while from the other site shows that even such comprehensive annotation as AceView is still not covering the full transcriptome landscape.

- In Figure 4.19c we show results when `Spladder` was run and com-

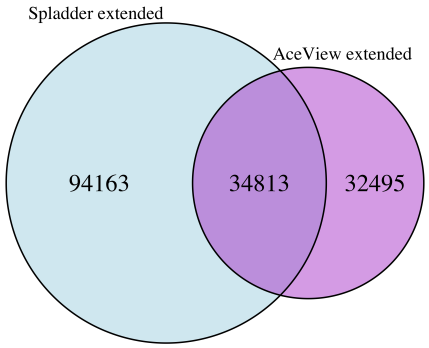
pared with the extended AceView annotation. It is interesting to note that the number of known junctions seen by `Spladder` in short reads data increased by about 4.5k, while number of not seen by `Spladder` junctions increased by 11.5k and now the fraction of junctions from annotation seen by `Spladder` dropped to about 50%. As we are talking about the same set of genes, it is unlikely that in the extended set we were adding new transcripts with dominating fraction of junctions not taking part in alternative splicing. Also, this extended set of alternative transcripts has been obtained from the same samples, so those are definitely expressed. Thus the only explanation is that short reads sequencing technology is unable to detect some junctions and the reason could be that effective depth of targeted long read sequencing is higher than one for short reads. Also, the number of new junctions involved in ASE detected by `Spladder` increased. This might be related with two Spladder features. First of all, it filters out all introns which originate from more than one gene. Second, it comes with a certain redundancy- `Spladder` reports all possible variants of a given event. Introducing a new transcript with extended annotation might then cause several new events reported for previously filtered out regions.



(a) AceView and Spladder.



(b) Spladder and Aceview filtered for panel genes.



(c) Spladder extended and AceView extended filtered for panel genes.

Figure 4.19: Venn diagrams comparing alternative splicing events found by Spladder in SEQC2 short reads data with those existing in: i) AceView annotation, ii) additional set of transcripts identified by IsoQuant in SEQC2 long reads data and filtered out based on study design ground truth, iii) AceView extended (i+ii). Spladder was run either with original AceView as reference or with extended AceView as reference.

4.2.1 Discussion

The results for the benchmarking data sets provide solid proof that `Spladder` results are reliable and it is a good choice for the developed pipeline. The percentage of known introns detected by `Spladder` for targeted genes is in line with the fraction of transcriptome expected to be expressed at any given time point. It also confirmed the efficiency of targeted sequencing as 87% of the reported junctions originated from panel genes.

Despite extending the reference annotation, `Spladder` still reports almost 95k new junctions. One must bare in mind that transcripts expanding annotation were chosen in a very rigorous approach, where about 90% of the initially reported transcripts were rejected. `Spladder` results might be an indicator that those junctions are correct, but need to be further validated. Once again we see evidence that even comprehensive annotations are still incomplete.

4.3 Microarray data

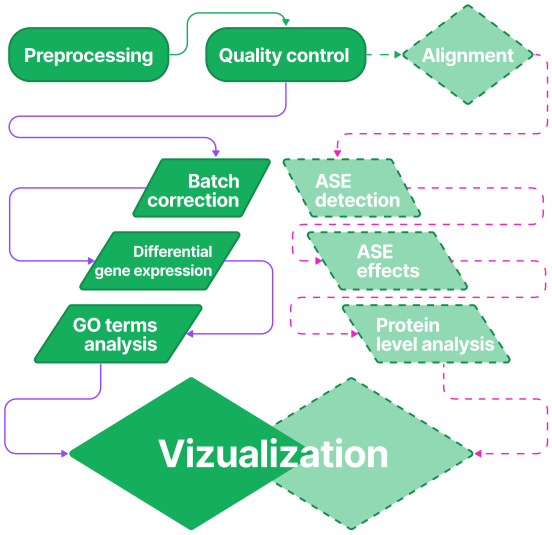
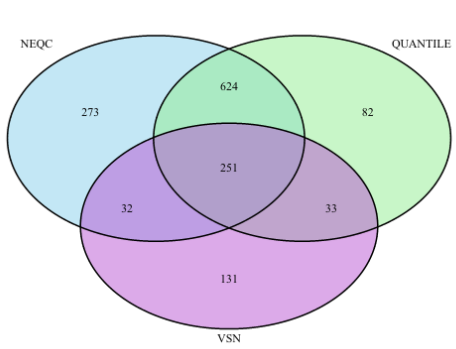


Figure 4.20: Part of the general pipeline used for microarray data preprocessing (only dark green boxes).

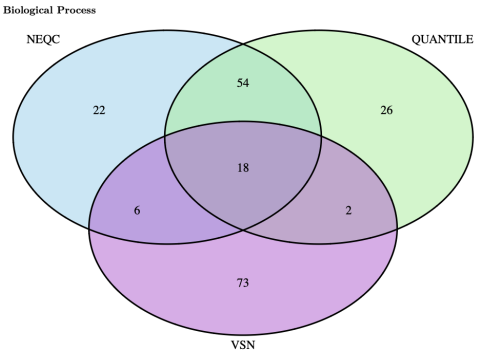
This part of the project can perfectly demonstrate both the similarities in NGS and microarray data analysis and the power of the `SVA` algorithm and the challenges it might entail. Besides normalization, quality control, and `SVA` instead of `SVAseq`, the rest of the pipeline uses the same approaches as for RNA-seq.

The venn diagram in Figure 4.21a demonstrates how different normalization methods can influence the obtained lists of differentially expressed genes. We can see that `NEQC` and `quantile` method share a huge portion of genes, that is because the difference is only in the background correc-

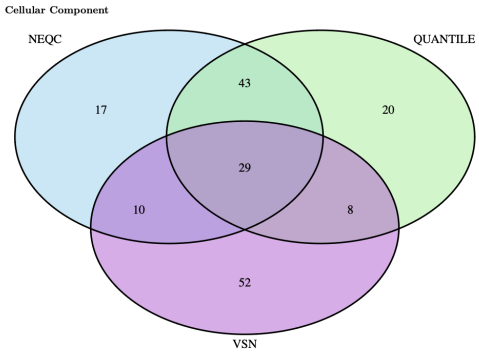
tion step. The common part- 251 genes constitute only about 18% of all genes detected with different methods. These observations are also true for further Gene Ontology analysis as shown on Figures 4.21b, 4.21c and 4.21d.



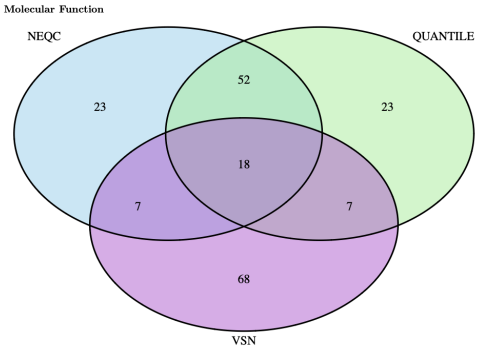
(a) Venn for DEGs found with different methods.



(b) Venn diagram for BP GO terms.



(c) Venn diagram for CC GO terms.

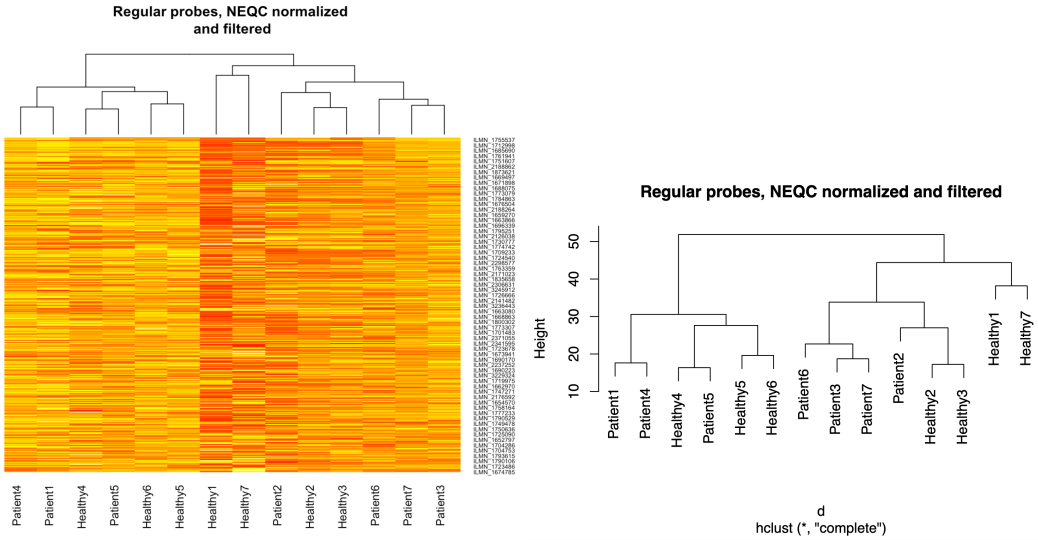


(d) Venn diagram for MF GO terms.

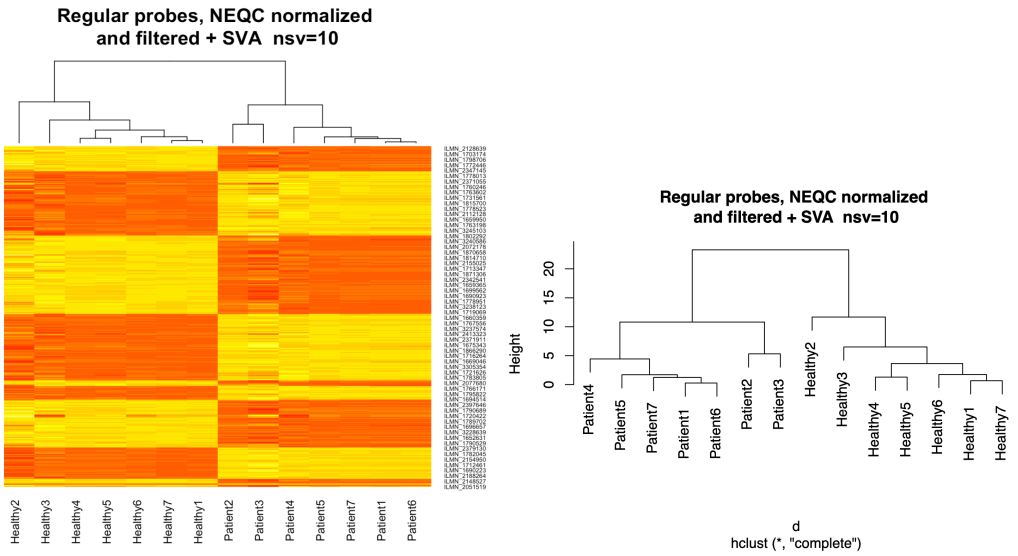
Figure 4.21: Differences in results for DEGs and GO terms caused by different normalization methods.

As explained in 4.1.1 `SVA` is a useful tool but should be applied with caution. Figure 4.22 once again presents how misleading `SVA` results can be. Although removing the maximum number of hidden factors (10 in this

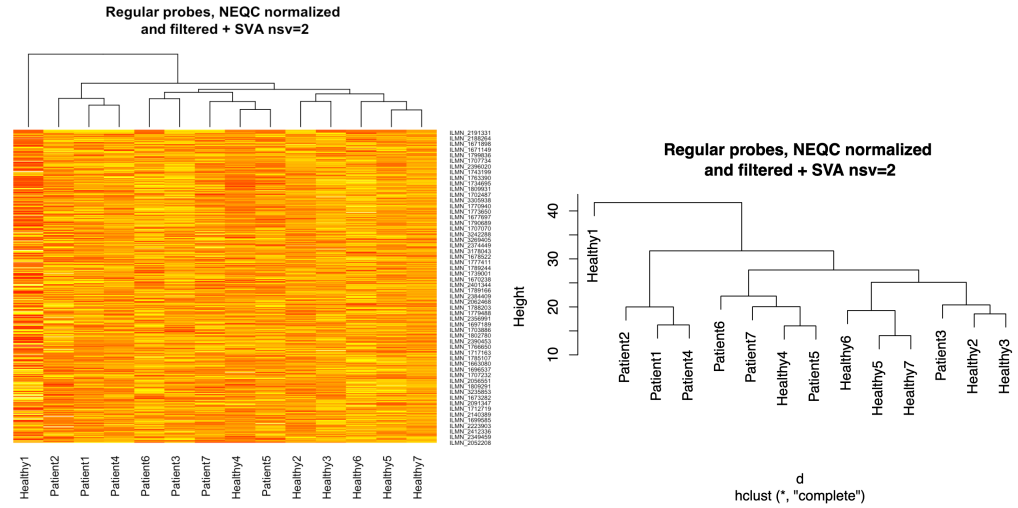
case in Figure 4.22b) produces a perfect separation between Patient and Healthy samples, it also reduces the variability within the group so much that any change between groups is treated as significant, causing the error of the limma algorithm and making differential expression analysis impossible. Adjusting for a smaller number of hidden factors allows us to remove unwanted variation without 'overfitting' the data. In this case, two factors are enough to roughly separate two sample groups.



(a) Heatmap and dendogram before SVA



(b) Heatmap and dendogram with 10 factors removed.



(c) Heatmap and dendogram with 2 factors removed.

Figure 4.22: Plots showing changes in the results of the gene clustering, depending on whether or not SVA was used and how many surrogate variables were removed.

4.3.1 Discussion

Microarray data analysis confirmed findings from RNA-seq results. Batch correction methods can be very useful if applied correctly, but the data produced for medical experiments, rather than benchmarking, can still be very problematic and despite applying best practices, we may also obtain low reproducibility.

Based on previous experiences with RNA-seq data, a complete microarray analysis pipeline was built in a relatively short amount of time. The fact that those technologies can borrow methods from each other is a huge advantage, as microarrays are an older technology with a variety of well-established and robust preprocessing algorithms. Knowledge of both approaches is crucial as they should not be treated as concurrent methods but rather be used interchangeably, depending on a scientific problem. That is why developing and incorporating steps for microarray data analysis into the pipeline can be very beneficial.

Chapter 5

Summary

Building and developing solutions that combine best practices can make the analyses more reliable and reproducible. For these tools to be useful, it is crucial to apply the methods in the correct way. Sections 4.1.1 and 4.3 demonstrated how many pitfalls can arise when only one step in the data analysis pipeline is used incorrectly.

As a result of this studies, a comprehensive pipeline covering different aspects of high- throughput data analysis and providing a variety of different approaches was developed. It includes all the major steps required for proper preprocessing and further analysis of both microarrays and sequencing data. Initial focus was placed on improving the reproducibility of differential gene expression detection; however, the research shifted towards alternative splicing analysis, a topic which was found to be very broad and also with no established solutions. A fairly exhaustive ap-

proach, covering different stages of the analysis was developed. Solution was based on already existing tools; however, not all of them were compatible with others. `Bisbee` software was a good starting point, but since it is not further supported, it had to be modified to suit our needs. With a set of additional scripts, a complete approach was developed-from alignment to possible effects on the protein level providing quantitative and qualitative analysis of RNA-Seq data. This solution is much needed, not only to improve the reproducibility of the results but also to fill the gap of automatic detection and analysis of AS and its consequences. This stage is missing from available workflows, but can provide a lot of new information and help to expand and better quantify the transcriptional landscape. This work is mainly focused on AS in the nervous system and shows that many isoforms are not yet known. AS is also known to play an important role in many diseases. Thus, an automatic approach for alternative splicing analysis can facilitate new insights into alterations that occur in conditions such as Parkinson's disease, SMA, or different types of cancer.

As mentioned before, there is no single all-purpose solution when it comes to high-throughput data analysis. However, the pipeline presented in this study benefits from the fact that it was created in the course of analysing multiple data sets and each provided new insights and caused new improvements. The fact that different challenges were solved with the same approaches is a validation of pipeline's comprehensiveness. As it combines in a unique way the quantitative and qualitative approaches for

RNA-Seq data analysis it can be used for many use cases or at least be a good starting point for others in follow-up studies.

5.1 Thesis conclusions

In the course of the projects, a deeper understanding of the challenges and limitations encountered in the analyses of real-world biomedical data sets was gained. These reflect the complexities of experiments and unwanted variations. The critical assessment of the analysis process and obtained results has allowed to outline four major conclusions. Those complement the developed pipeline in a form of best practices/guidelines and are constituting the most important outcome of this thesis.

1. **Currently there are no gold standards in the analysis of data from high-throughput technologies.**

   This was confirmed by the results for the 3 data sets in Sections 4.1 - 4.3. Depending on the algorithms used, the results obtained for differential gene expression can vary considerably for microarrays (Section 4.3) and RNA-seq (Section 4.1). This challenge is true for both the choice of data analysis methods and also the laboratory techniques (different types of sequencing or microarrays). Section 4.2 showed that the results obtained with long reads differ from those obtained with short reads. As here we were working on benchmarking data sets, the source of some discrepancy can be inferred and

understood. Still, it is very important to choose the technology best tailored to our needs and validate the results with other approaches, as they can often complement each other.

2. **Analysis recommendations based on studies with artificially created data sets should be applied to real-life problems with caution.**
   Data sets used for benchmarking tend to have strong signal differentiating between compared conditions, as well as comprehensive metadata documentation, those two very important aspects are often not fulfilled when it comes to real-life scenarios. There, we need to face challenges resulting from poor data quality or low signal levels. Data set presented in Section 4.1 is an example of such case. Results are not stable and it is not possible to filter out lowly expressed transcripts, which are known to contain a lot of noise, because of generally low signal values, without losing True Positives. Such cases might require more complex approaches (as careful accounting for hidden confounding factors), and results can still be uncertain. Still, the developed pipeline allows for all this steps and also for qualitative analysis to get as much as possible even from quantitatively poor data.

3. **RNA- seq and microarray approaches both have strengths and weaknesses and should be used interchangeably, depending on the scientific problem.**

It is worth noting that solutions already developed for microarrays can often be a good starting point for RNA-seq data analysis, as they have already been tested and proven to provide meaningful results; however, one has to bare in mind the differences in nature of both technologies and incorporate appropriate alterations. `Limma` is a great tool for differential gene expression analysis, developed and well established for microarrays, and can be successfully applied to RNA-seq data, after accounting for their discrete nature. `Limma` can also easily incorporate the results obtained from `SVA` which is yet another staple tool in microarray analysis and has also been adjusted to fit the sequencing data.

4. **Great improvements can still be made in the field of reference transcript annotation.**

   Last but in the context of the thesis topic, the most important conclusion is that the complexity of the mouse, but also the human genome, is not yet fully understood. Human gene model is more comprehensive and better studied, thus it might be harder to find new and trustworthy isoforms. Despite that, results for benchmarking data set in Section 4.2 report thousands of previously unannotated junctions for human reference. Mouse reference, on the other hand, remains not fully annotated, and new findings are more confident. Despite the presence of many confounding factors, such as different sample batches and knockouts induced in different structures, it was possi-

ble to detect common patterns in the data. Interestingly, an overlap between different types of ASEs at the level of gene and functional analysis was low except Cellular Component tree where terms were related to the nervous system. It implies that the functions of the detected ASEs and the processes in which they are involved are characteristic for a given event type. This is an observation that does not yet have any evidence in the literature, however, was pointed out as possible in discussion with a few experts. As mentioned before, results are based on short read sequencing and we cannot infer whole transcripts. Nevertheless, they indicate that there are many events not present in reference, and they originate from genes related to the nervous system. That is a strong indication that long-read sequencing experiment is much needed to validate those events and possibly expand existing annotation, providing a better understanding of the mouse gene model. Such conclusions have also been confirmed in the literature, and work has already begun for some brain structures [55]. Also, further protein domain analysis connected with dedicated visualization has shown that these events have support in the reads and could have an impact on important protein domains. This observation is in line with several recent articles reporting many novel alternatively spliced events occurring in different regions of the brain and also other tissues in different species [101, 19, 36]. Great improvements can still be made in the field of reference transcript annotation,

as even for model organisms, the reference gene models are not mature yet. And here, the developed pipeline can be the best choice to start with.

5.2 Scientific manuscripts arising from this Thesis Research and related work

During the course of this PhD work, two articles have been published in international SCI-listed scientific journals:

- Foox, Nordlund, Lalancette, et al. [38] in *Genome Biology*, where I was responsible for genome-scale sequence analysis,
- Chlebanowska et al. [17] in *Journal of Molecular Sciences*, where I was responsible for bioinformatics analysis, including microarray differential gene expression analysis.

A third publication Deshpande et al. [26] is in the final stages of the review process at *Frontiers in Genetics*. It provides an overview of the state-of-the-art methods in RNA-seq data analysis. I was responsible for several sections with a focus on characterizing differential gene expression analysis methods and best practices there.

5.3 Availability of data and code

The data sets presented in Section 4 are from novel primary studies that have not been published before. For each of them, manuscripts are in the final stages of preparation, and the data will be published together with the main study manuscripts. The pipeline developed for this thesis is available on GitHub (`https://github.com/aagatam/Pipeline`).

Chapter 6

Acknowledgments

I would like to thank my supervisor, Dr. Paweł Łabaj, for his support and inspiration, not only during my PhD studies, but essentially for almost 10 years during different stages of my scientific career. I would have never thought about pursuing my doctorate if Paweł had not introduced me to the world of transcriptomics and planted the seed in my head to develop in this direction.

I am also very thankful to Prof. Ryszard Przewłocki from the Institute of Pharmacology of the Polish Academy of Sciences for his advice, patience in explaining biological foundations, and countless ideas on how to approach neuropathic pain dataset.

This work has been co-funded by the European Union through the European Social Fund grant (POWR.03.02.00-00-I029)

References

[1] Adrian Alexa and Jorg Rahnenfuhrer.
topGO: Enrichment Analysis for Gene Ontology.
R package version 2.50.0. 2022.

[2] J Allaire et al. *rmarkdown: Dynamic Documents for R*.
R package version 2.19. 2022.

[3] *Anaconda Software Distribution*. Version Vers. 2-2.4.0. 2020.
URL: `https://docs.anaconda.com/`.

[4] Simon Andrews et al. *FastQC*. Babraham Institute.
Babraham, UK, 2012.

[5] M. Ashburner et al. “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.”
In: *Nat Genet.* 25(1) (2000), pp. 25–9.

[6] M. Baker. “ 1,500 scientists lift the lid on reproducibility.”
In: *Nature* 533 (2016), pp. 452–454.

[7] Y Benjamini and Y Hochberg. “Controlling the false discovery rate: a practical and powerful approach to multiple testing.”
In: *Journal of the Royal Statistical Society.* (1995), pp. 289–300.

[8] Y Benjamini and D Yekutieli. “The control of the false discovery rate in multiple testing under dependency.”
In: *Annals of Statistics.* (2001), 1165–1188.

[9] Koen Van den Berge et al. “A hitchhiker’s guide to single-cell transcriptomics and data analysis pipelines.” In: *Annual Review of Biomedical Data Science* 2 (2021), pp. 139–173.

[10] BJ. Blencowe.
“Alternative splicing: new insights from global analyses”.
In: *Cell* 126.5 (2006), pp. 37–47.

[11] B M Bolstad et al.
“A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.”
In: *Bioinformatics* 19 (2003), pp. 185–93.

[12] BM Bolstad. “Low Level Analysis of High-density Oligonucleotide Array Data: Background, Normalization and Summarization.”
In: *Dissertation. University of California, Berkeley* (2004).

[13] G.M. Boratyn, J. Thierry-Mieg, D. Thierry-Mieg, et al. “ Magic-BLAST, an accurate RNA-seq aligner for long and short reads”.
In: *BMC Bioinformatics* 20 (2019), p. 45.

[14] N. Bray, H. Pimentel, P. Melsted, et al.
“Near optimal probabilistic RNA-seq quantification”.
In: *Nature Biotechnology* 34 (2016), pp. 525–527.

[15] JH Bullard et al.
“ Evaluation of statistical methods for normalization and

differential expression in mRNA-Seq experiments."
In: *BMC Bioinformatics* 11:94 (2010).

[16] Michael Burrows and David J. Wheeler.
"A block sorting lossless data compression algorithm."
In: *Digital Equipment Corporation* (1994).

[17] Paula Chlebanowska et al. "Origin of the Induced Pluripotent Stem Cells Affects Their Differentiation into Dopaminergic Neurons".
In: *International Journal of Molecular Sciences* 21.16 (2020), pp. 1422–0067.

[18] Su Chun-Hao, D Dhananjaya, and Tarn Woan-Yuh.
"Alternative Splicing in Neurogenesis and Brain Development".
In: *Frontiers in Molecular Biosciences* 5.3 (2018), p. 12.

[19] Michael B Clark et al.
"Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain."
In: *Molecular psychiatry* 25(1) (2020), pp. 37–37.

[20] Matthew Cobb.
"60 years ago, Francis Crick changed the logic of biology."
In: *PLoS Biol.* 15(9) (Sept. 2017), e2003243.

[21] A. Conesa, P. Madrigal, S. Tarazona, et al. "A survey of best practices for RNA-seq data analysis., year = 2016".
In: *Genom Biol* 17 (), p. 13.

[22] MAQC Consortium.
“The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. ”
In: *Nat Biotechnol.* 24(9) (2006), pp. 1151–61.

[23] SEQC/MAQC-III Consortium. “ A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium”.
In: *Nat Biotechnol* 32.13 (2014), 903–914.

[24] Fiona Cunningham et al. “Ensembl 2022”.
In: *Nucleic Acids Research* 50.D1 (2021), pp. D988–D995.

[25] Sebastian Deorowicz and Szymon Grabowski.
“Compression of DNA sequence reads in FASTQ format”.
In: *Bioinformatics* 27.6 (2011), pp. 860–862.

[26] Dhrithi Deshpande et al.
RNA-seq data science: From raw data to effective interpretation.
2020. URL: `https://arxiv.org/abs/2010.02391`.

[27] S. Djebali, C. Davis, A. Merkel, et al.
“Landscape of transcription in human cells.”
In: *Nature* 489 (2012), 101–108.

[28] Li Dongmei. *Statistical Methods for RNA Sequencing Data Analysis*.
Rochester, NY, USA: Clinical, Translational Science Institute, University of Rochester School of Medicine, and Dentistry, 2019.

[29] M Dunning, A Lynch, and M Eldridge. “illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4)”. In: *R package version 1.26.0.* (2015).

[30] M Dunning et al. “beadarray: R classes and methods for Illumina bead-based data.” In: *Bioinformatics* 23(16) (2007), 2183–4.

[31] B.P. Durbin et al. “A variance-stabilizing transformation for gene-expression microarray data”. In: *Bioinformatics* 18 (2002), S105–S110.

[32] C Everaert et al. “Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data”. In: *Sci Rep.* 7(1) (2017), p. 1559.

[33] Philip Ewels et al. “MultiQC: summarize analysis results for multiple tools and samples in a single report”. In: *Bioinformatics* 32.19 (2016), pp. 3047–3048.

[34] Mölder F, Jablonski KP, Letcher B, et al. “Sustainable data analysis with Snakemake.” In: *F1000Research* 10 (2021), p. 33.

[35] Rohmatul Fajriyah. “Paper review: An overview on microarray technologies.” In: *Bulletin of Applied Mathematics and Mathematics Education* 1(1) (2021), pp. 21–30.

[36] Wen Feng et al.
“Profiling Novel Alternative Splicing within Multiple Tissues Provides Useful Insights into Porcine Genome Annotation”.
In: *Genes* 11 (2020).

[37] Paolo Ferragina and Giovanni Manzini.
“Opportunistic data structures with applications.”
In: *Foundations of Computer Science,Proceedings. 41st Annual Symposium on. IEEE* (2000), pp. 390–398.

[38] J. Foox, J. Nordlund, C. Lalancette, et al.
“The SEQC2 epigenomics quality control (EpiQC) study.”
In: *Genome Biol* 22 (2021), p. 332.

[39] A Frankish et al. “GENCODE 2021”.
In: *Nucleic Acids Res.* 49(D1 (2021), pp. D916–D923.

[40] A Frankish et al. “GENCODE reference annotation for the human and mouse genomes”.
In: *Nucleic Acids Res.* 47(D1 (2019), pp. D766–D773.

[41] Leonard P. Freedman, Iain M. Cockburn, and Timothy S. Simcoe.
“The Economics of Reproducibility in Preclinical Research.”
In: *PLOS Biology* 13(6) (2015), e1002165.

[42] Glenn K. Fu et al.
“Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations.”

In: *Proceedings of the National Academy of Sciences* 111.5 (2014), pp. 1891–1896.

[43] L Gautier et al.
“Affy–analysis of Affymetrix GeneChip data at the probe level.”
In: *Bioinformatics* 20 (2004), pp. 307–315.

[44] R.C. Gentleman, V.J. Carey, and D.M.and others Bates.
“Bioconductor: open software development for computational biology and bioinformatics.” In: *Genome Biol* 5 (2004), R89.

[45] Jelle Goeman and Aldo Solari.
“Multiple hypothesis testing in genomics.”
In: *Statistics in medicine* 33(11) (2014), pp. 1946–78.

[46] S. Grossmann, S. Bauer, and M. Robinson PN.and Vingron.
“Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.”
In: *Bioinformatics* 23(22) (2007), pp. 3024–31.

[47] K. L. Gunderson et al. “Decoding randomly ordered DNA arrays.”
In: *Genome Res.* 14(5) (May 2004), pp. 870–7.

[48] R F Halperin, A Hegde, J D Lang, et al. “Improved methods for RNAseq-based alternative splicing analysis.”
In: *Sci Rep* 11 (2021), p. 10740.

[49] Stephanie C. Hicks and Rafael A. Irizarry.
“When to use Quantile Normalization?” In: *bioRxiv* (2014).

[50] P. G. Higgs and Teresa K. Attwood.
Bioinformatyka i Ewolucja Molekularna.
Warszawa: Wydawnictwo Naukowe PWN, 2012.

[51] Sepp Hochreiter et al.
“A new summarization method for Affymetrix probe level data.”
In: *Bioinformatics* 22,8 (2006), pp. 943–9.

[52] S. Holm. “A simple sequentially rejective multiple test procedure.”
In: *Scandinavian Journal of Statistics* (1979), 65–70.

[53] Holger Husi. “NMDA Receptors, Neural Pathways, and Protein Interaction Databases”. In: *Human Brain Proteome*. Vol. 61. International Review of Neurobiology. Academic Press, 2004, pp. 49–77.

[54] H. Jin, YW Wan, and Z. Liu. “Comprehensive evaluation of RNA-seq quantification methods for linearity”.
In: *BMC Bioinformatics* 18(Suppl 4) (2017), p. 117.

[55] A. Joglekar, A. Prjibelski, A. Mahfouz, et al.
“A spatially resolved brain region- and cell type- specific isoform atlas of the postnatal mouse brain ”.
In: *Nat Commun* 12.8 (2021), p. 463.

[56] Philip Jones et al.
“InterProScan 5: genome-scale protein function classification”.
In: *Bioinformatics* 30.9 (2014), pp. 1236–1240.

[57] W. Jones, B. Gong, N. Novoradovskaya, et al.
“A verified genomic reference sample for assessing performance of cancer panels detecting small variants of low allele frequency.”
In: *Genom Biol.* 22 (2021), p. 111.

[58] A Kahles et al. “SplAdder: identification, quantification and testing of alternative splicing events from RNA-Seq data”.
In: *Bioinformatics* 15.11 (2016), pp. 1840–1847.

[59] H Keren, G Lev-Maor, and G. Ast. “Alternative splicing and evolution: diversification, exon definition and function”.
In: *Nat Rev Genet.* 11.7 (2010), pp. 345–55.

[60] D. Kim, B. Langmead, and S. Salzberg.
“ HISAT: a fast spliced aligner with low memory requirements.”
In: *Nature Methods* 12 (2015), pp. 357–360.

[61] D. Kim, J.M. Paggi, C. Park, et al. “ Graph-Based Genome Alignment and Genotyping with HISAT2 and HISAT-genotype”.
In: *Nat Biotechnol* 37(8) (2015), pp. 907–915.

[62] E Korpelainen et al. *RNA-seq data analysis a practical approach.*
London: CRC Press Taylor Francis Group, 2014.

[63] ES Lander et al.
“Initial sequencing and analysis of the human genome.”
In: *Nature* 409 (2001), 860–921.

[64] C.W. Law, Y. Chen, W. Shi, et al. “voom: precision weights unlock linear model analysis tools for RNA-seq read counts.” In: *Genome Biol* 15 (2014), R29.

[65] Cosmin Lazar et al. “Batch effect removal methods for microarray gene expression data integration: a survey”. In: *Briefings in Bioinformatics* 14.4 (2012), pp. 469–490.

[66] Jeffrey T Leek. “svaseq: removing batch effects and other unwanted noise from sequencing data.” In: *Nucleic Acids Research* 42(21) (2014), e161.

[67] Jeffrey T Leek and John D Storey. “Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis.” In: *PLoS genetics* 3(9) (2007), pp. 1724–35.

[68] J Li and R. Tibshirani. “Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. ” In: *Stat Methods Med Res.* 22(5) (2013), pp. 519–36.

[69] S Li et al. “Detecting and correcting systematic variation in large-scale RNA sequencing data. ” In: *Nat Biotechnol.* 32 (2014), 888–95.

[70] Diane Lipscombe. “Neuronal proteins custom designed by alternative splicing”. In: *Current Opinion in Neurobiology* 15.10 (2005), pp. 358–363.

[71] MI Love, W Huber, and S Anders. “Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.” In: *Genome Biology* 15 (2014), p. 550.

[72] Shir Mandelboum et al. “Recurrent functional misinterpretation of RNA-seq data caused by sample-specific gene length bias.” In: *PLOS Biology* 17(11) (2019), e3000481.

[73] Sergey Nurk et al. “The complete sequence of a human genome”. In: *Science* 376.6588 (2022), pp. 44–53.

[74] Nuala A. O’Leary et al. “Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation”. In: *Nucleic Acids Research* 44.D1 (2015), pp. D733–D745.

[75] K. Owzar, W. T. Barry, and S. H. Jung. “Statistical considerations for analysis of microarray experiments.” In: *Clinical and translational science* 4(6) (2011), 466–477.

[76] Q. Pan et al. “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing”. In: *Nat Genet* 40.6 (2008), 1413–1415.

[77] R. Patro, S. Mount, and C Kingsford. “Sailfish enables alignment-free isoform quantification from RNA-seq reads using lightweight algorithms.” In: *Nat Biotechnol* (2014), pp. 462–464.

[78] R. Patro et al. “Salmon provides fast and bias-aware quantification of transcript expression.” In: *Nature methods* (2017), 417–419.

[79] M. Pertea, D. Kim, G. Pertea, et al. “Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown.” In: *Nature Protocols* 11 (2016), pp. 1650–1667.

[80] M. Pertea et al. “StringTie enables improved reconstruction of a transcriptome from RNA-seq reads”. In: *Nature Biotechnology* 3(3) (2015), pp. 290–295.

[81] Belinda Phipson et al. “ROBUST HYPERPARAMETER ESTIMATION PROTECTS AGAINST HYPERVARIABLE GENES AND IMPROVES POWER TO DETECT DIFFERENTIAL EXPRESSION.” In: *The annals of applied statistics* 10,2 (2016), pp. 946–963.

[82] Andrey Prjibelski et al. “Accurate isoform discovery with IsoQuant using long reads.” In: *Nat Biotechnol* (2023).

[83] Bushra Raj and Benjamin J. Blencowe. “Alternative splicing in the mammalian nervous system: recent insights into mechanisms and functional roles”. In: *Neuron* 87.1 (July 2015), pp. 14–27.

[84] Matthew E Ritchie et al. “limma powers differential expression analyses for RNA-sequencing and microarray studies.” In: *Nucleic acids research* 43,7 (2015), e47.

[85] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.” In: *Bioinformatics* 26 (2010), 139–140.

[86] M.D. Robinson and A. Oshlack. “A scaling normalization method for differential expression analysis of RNA-seq data.” In: *Genom Biol* 11 (2010), R25.

[87] F. Sanger, S. Nicklen, and A. R. Coulson. “DNA sequencing with chain-terminating inhibitors.” In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467.

[88] R. Schmid, P. Baum, C. Ittrich, et al. “Comparison of normalization methods for Illumina BeadChip HumanHT-12 v3.” In: *BMC Genomics* 11 (2010), p. 349.

[89] L. Shi et al. “Cross-platform comparability of microarray technology: intra-platform consistency and appropriate data analysis procedures are essential”. In: *BMC bioinformatics* 6 (2005), S12.

[90] Wei Shi, Alicia Oshlack, and Gordon K. Smyth.
"Optimizing the noise versus bias trade-off for Illumina whole genome expression BeadChips".
In: *Nucleic Acids Research* 38 (2010), e204–e204.

[91] *Statistics about the GENCODE Release 42, url = https://www.gencodegenes.org/human/stats_42.html, Accessed = 2022-11-14*.

[92] *Statistics about the GENCODE Release M31, url = https://www.gencodegenes.org/mouse/stats_M31.html, Accessed = 2022-11-14*.

[93] O Stegle et al. "Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses." In: *Nat Protoc.* 7(3) (2012), pp. 500–7.

[94] Yunxia Sui et al. "Background Adjustment for DNA Microarrays Using a Database of Microarray Experiments".
In: *Journal of Computational Biology* 16(11) (2009), pp. 1501–1515.

[95] T Svitkina et al. "Regulation of the postsynaptic cytoskeleton: roles in development, plasticity, and disorders. J Neurosci."
In: 30(45) (2010), pp. 14937–42.

[96] S Tarazona et al.
"Differential expression in RNA-seq: A matter of depth."
In: *Genome Res.* 21(12) (2011), 2213–23.

[97] Kevin Ushey, JJ Allaire, and Yuan Tang.
reticulate: Interface to 'Python'. https://rstudio.github.io/reticulate/, https://github.com/rstudio/reticulate. 2023.

[98] CK Vuong, DL Black, and Zheng S.
"The neurogenetics of alternative splicing".
In: *Nat Rev Neurosci*. 17.9 (2016), pp. 265–281.

[99] Eric T. Wang et al.
"Alternative isoform regulation in human tissue transcriptomes".
In: *Nature* 456.2 (2008), pp. 470–476.

[100] Nobumoto Watanabe and Osada Hiroyuki.
"Phosphorylation-dependent protein-protein interaction modules as potential molecular targets for cancer therapy."
In: *Current drug targets* 13 (2012), pp. 1654–8.

[101] D.J. Wright, N.A.L. Hall, N. Irish, et al.
"Long read sequencing reveals novel isoforms and insights into splicing regulation during cell state changes."
In: *BMC Genomics* 23 (2022), p. 42.

[102] Y Xie, X Wang, and M. Story. "Statistical methods of background correction for Illumina BeadArray data."
In: *Bioinformatics* 25(6) (2009), pp. 751–7.

[103] Y. H. Yang et al.
"Normalization for cDNA microarray data: a robust composite

method addressing single and multiple slide systematic variation."
In: *Nucleic acids research* 30(4) (2002), e15.

[104] G. Yeo, D. Holste, G. Kreiman, et al.
"Variation in alternative splicing across human tissues".
In: *Genom Biol* 5.4 (2004), R74.

[105] DR Zerbino, A Frankish, and P. Flicek. " Progress, Challenges, and Surprises in Annotating the Human Genome. "
In: *Annu Rev Genomics Hum Genet.* 21 (2020), pp. 55–79.

[106] J. Zyla, M. Marczyk, J. Weiner, et al.
"Ranking metrics in gene set enrichment analysis: do they matter?"
In: *BMC Bioinformatics* 18 (2017), p. 256.

[107] Paweł P. Łabaj et al. "Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling".
In: *Bioinformatics* 27.13 (2011), pp. i383–i391.

[108] P.P. Łabaj and D.P. Kreil. " Sensitivity, specificity, and reproducibility of RNA-Seq differential expression calls."
In: *Biol Direct* 11 (2016), p. 66.

Appendices

Appendix A

Attached USB drive content

Attached USB drive contains:

- a pdf file containing this thesis,
- source code for analysis pipeline,
- a pdf file containing additional plots for Section 4.1.6.

Appendix B

List of Figures

List of Figures

Appendix C

List of Tables

List of Tables

Appendix D

Supplementary Material

Figure S1: Barplots showing summary statistics for different events for group with both known events.

Figure S2: Barplots showing summary statistics for different events for group with both new events.

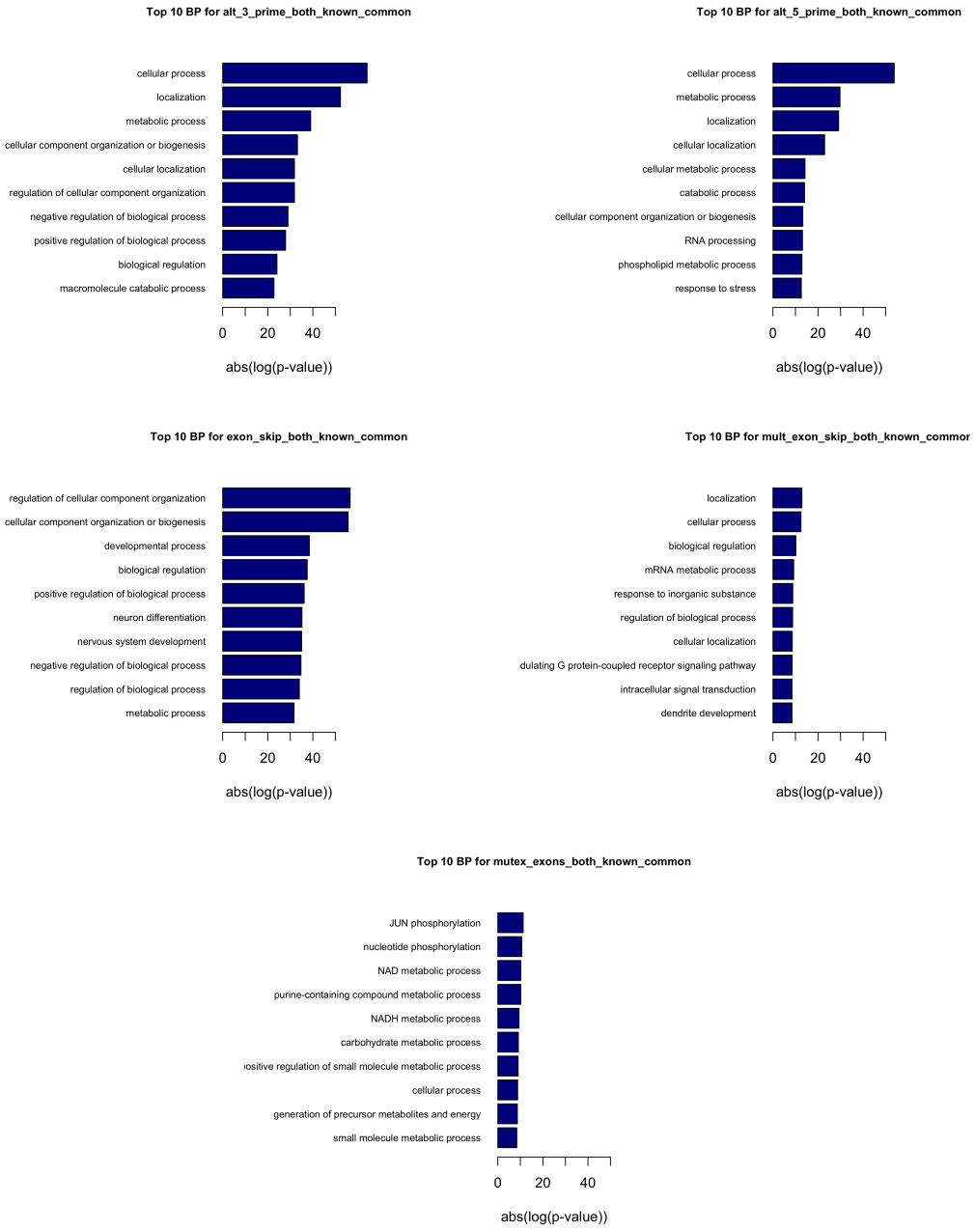


Figure S3: Top 10 BP terms for old+old group.

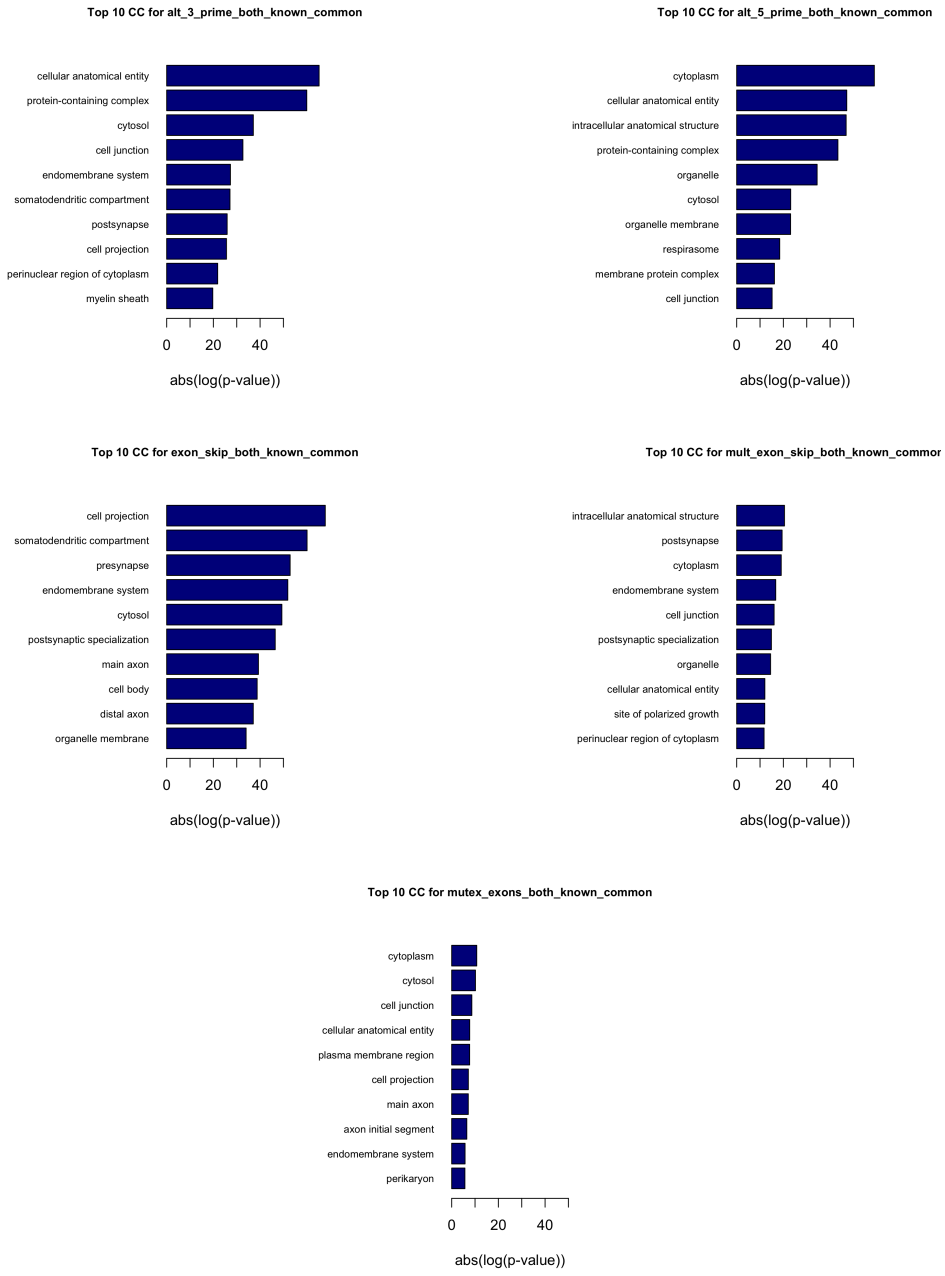


Figure S4: Top 10 CC terms for old+old group.

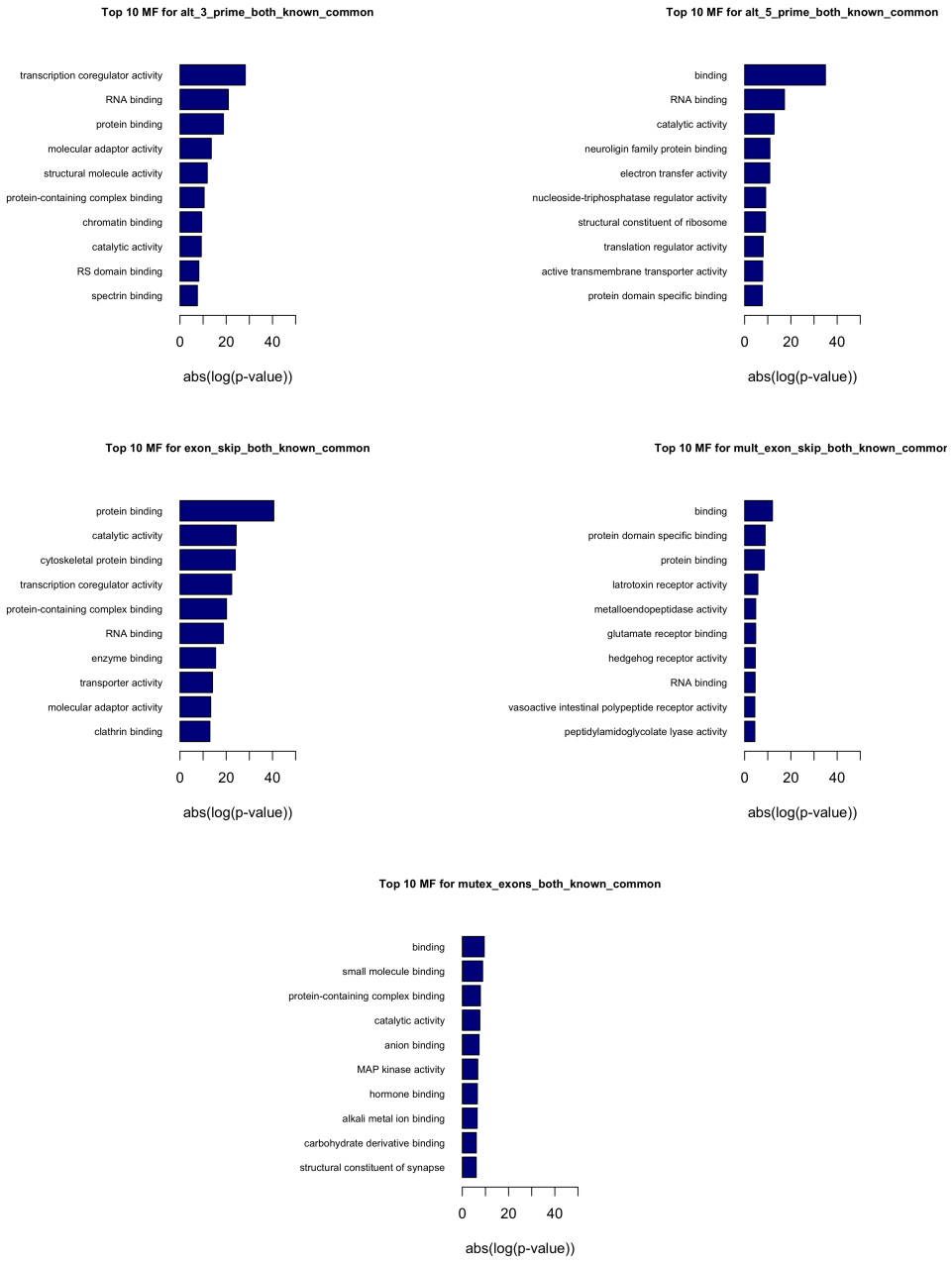


Figure S5: Top 10 MF terms for old+old group.

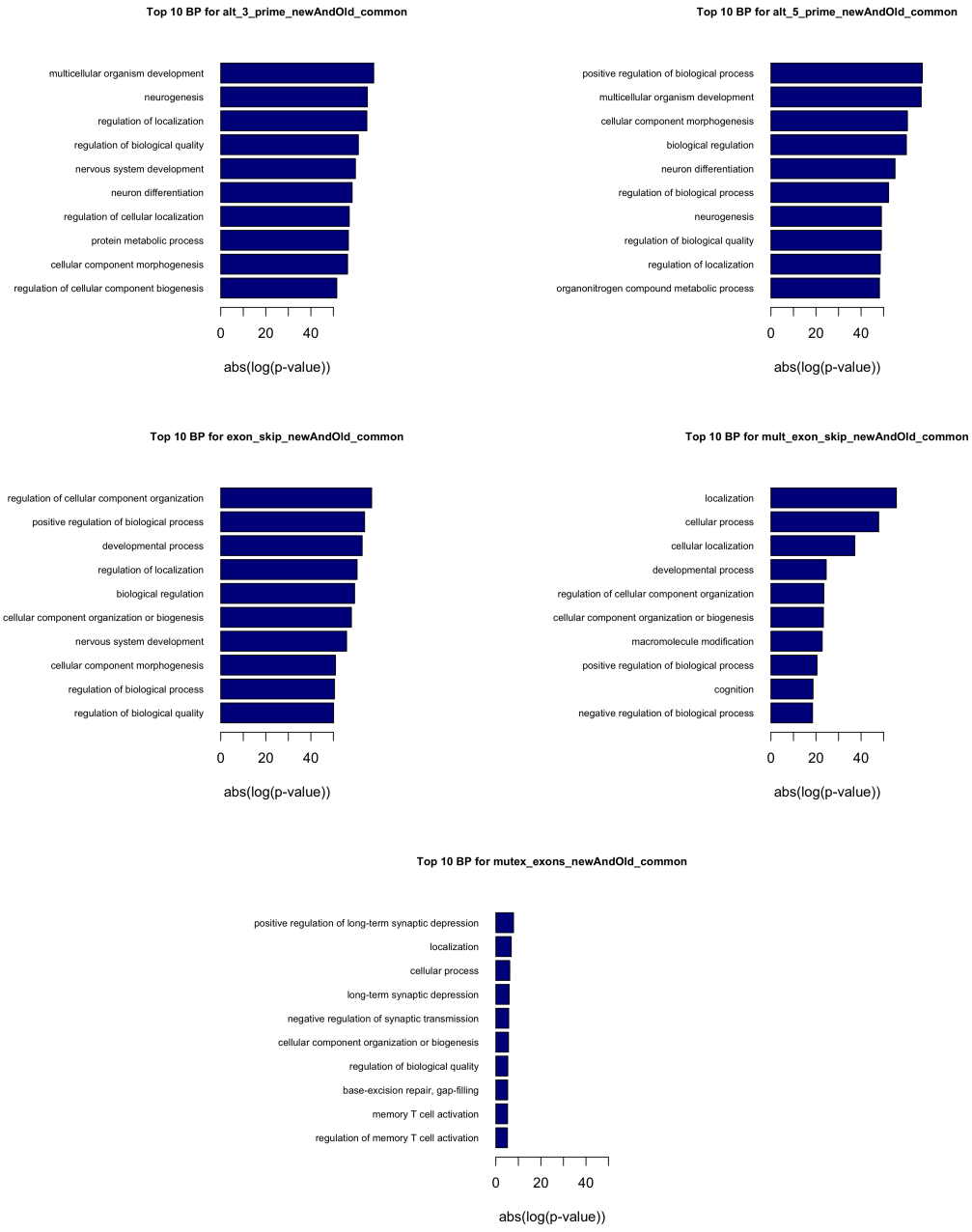


Figure S6: Top 10 BP terms for new+old group.

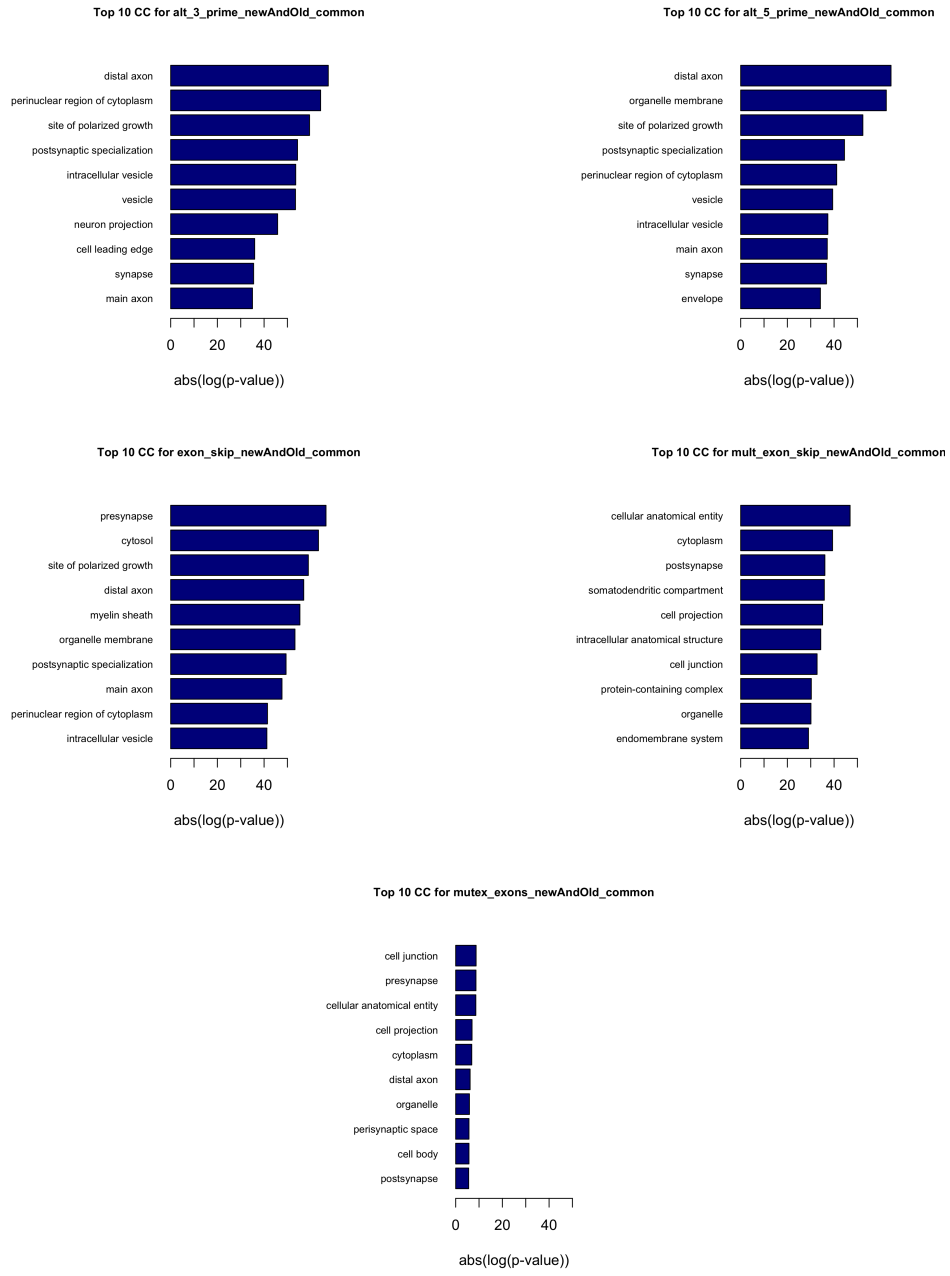


Figure S7: Top 10 CC terms for new+old group.

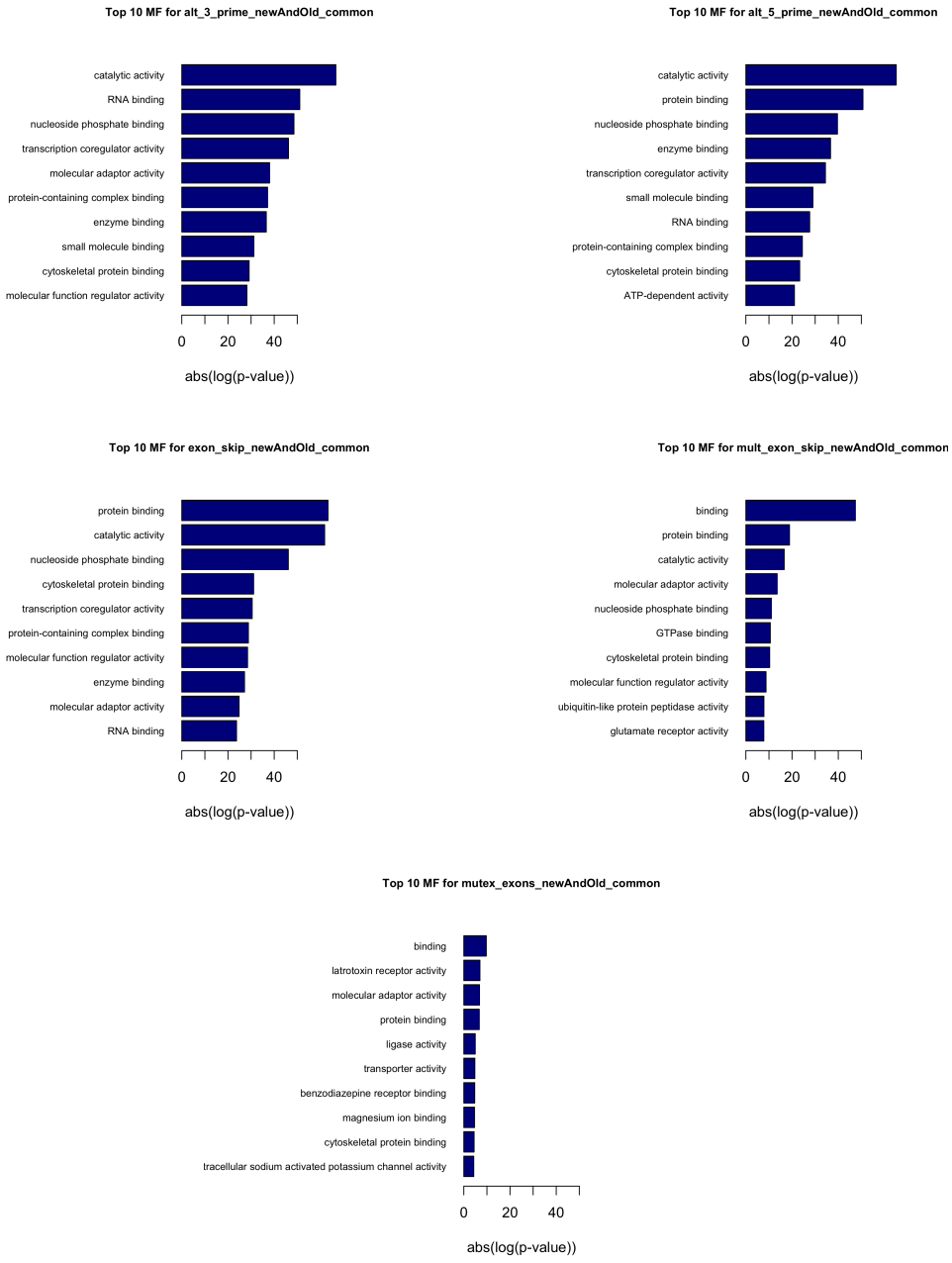


Figure S8: Top 10 MF terms for new+old group.

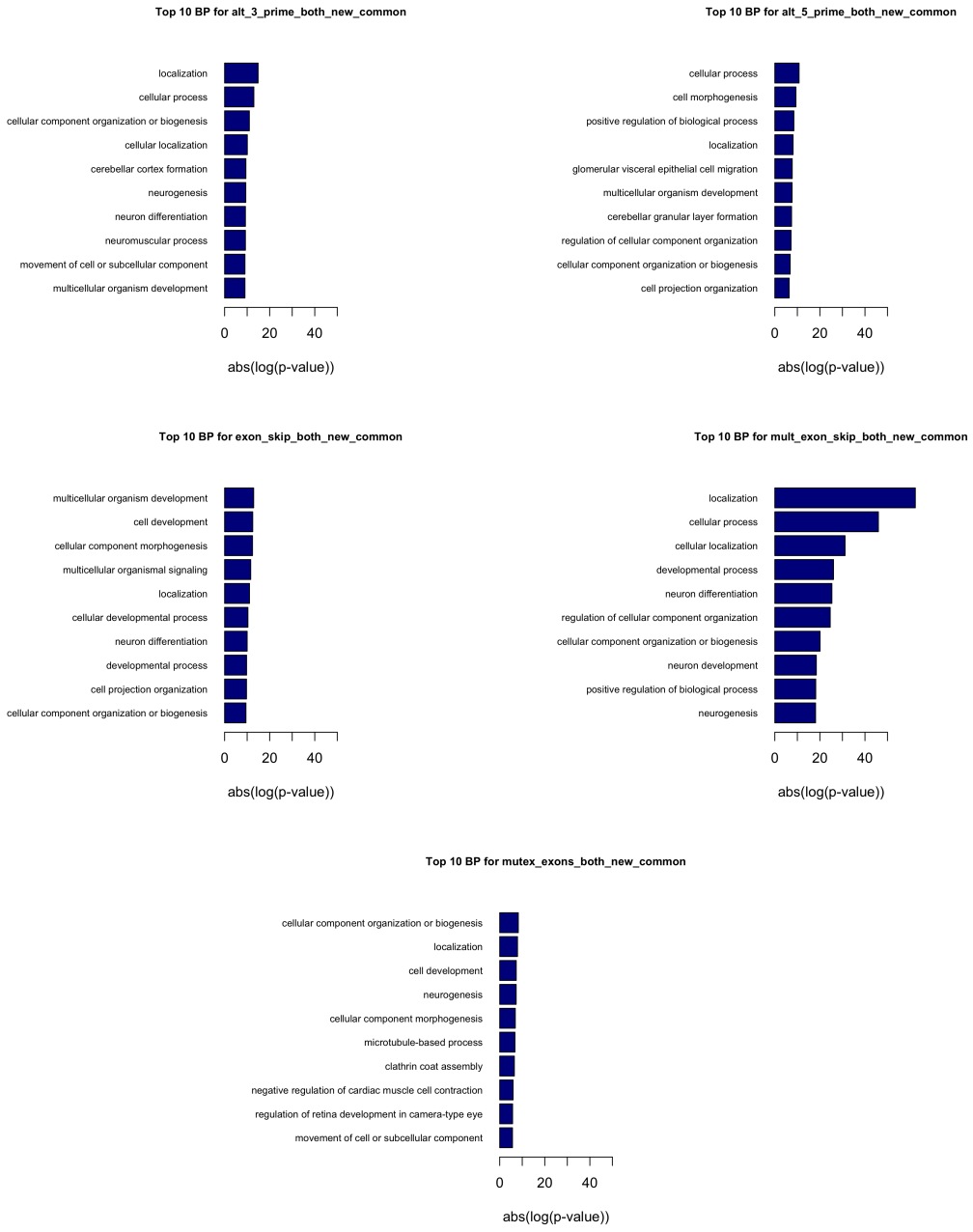


Figure S9: Top 10 BP terms for new+new group.

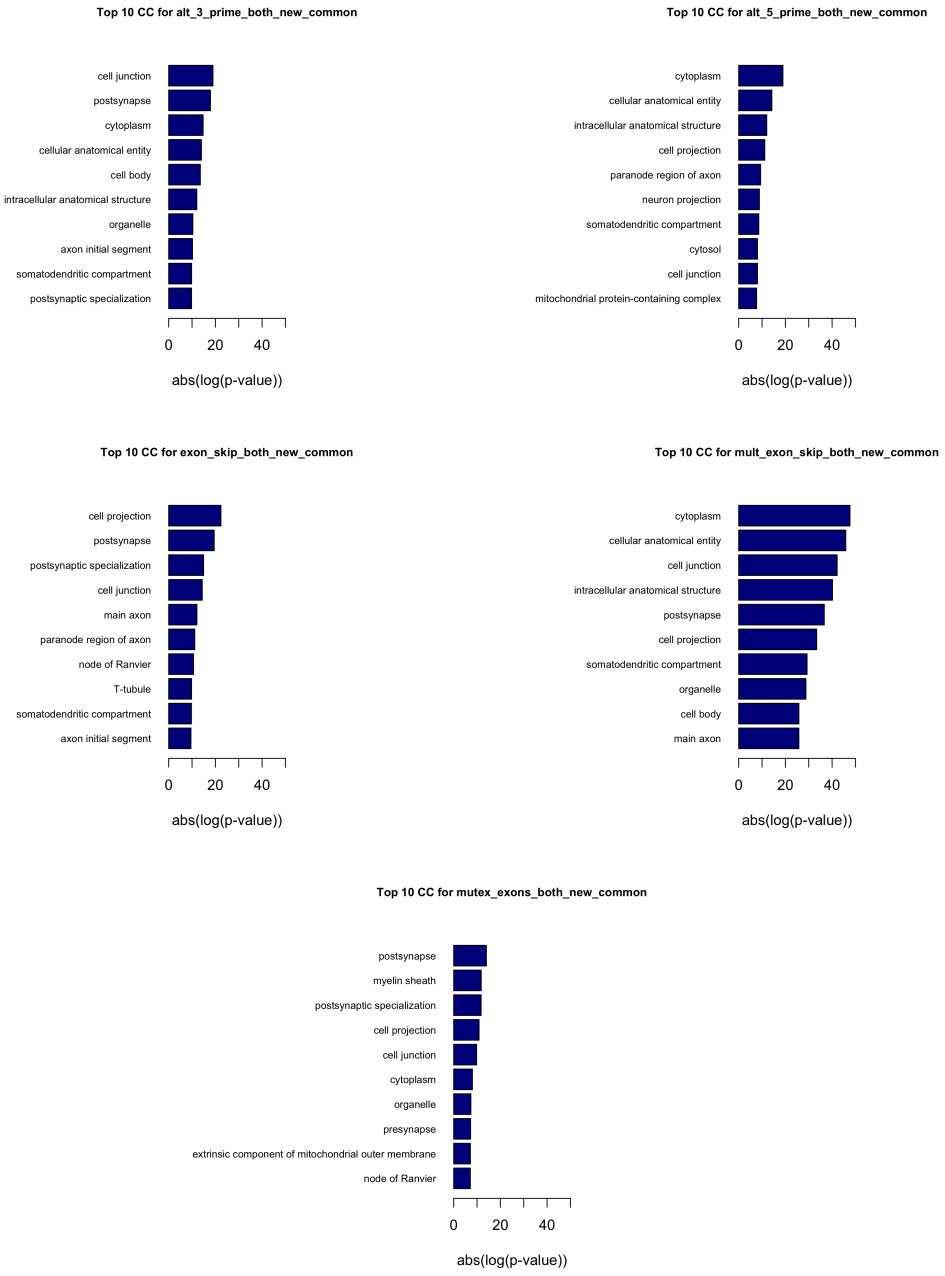


Figure S10: Top 10 CC terms for new+new group.

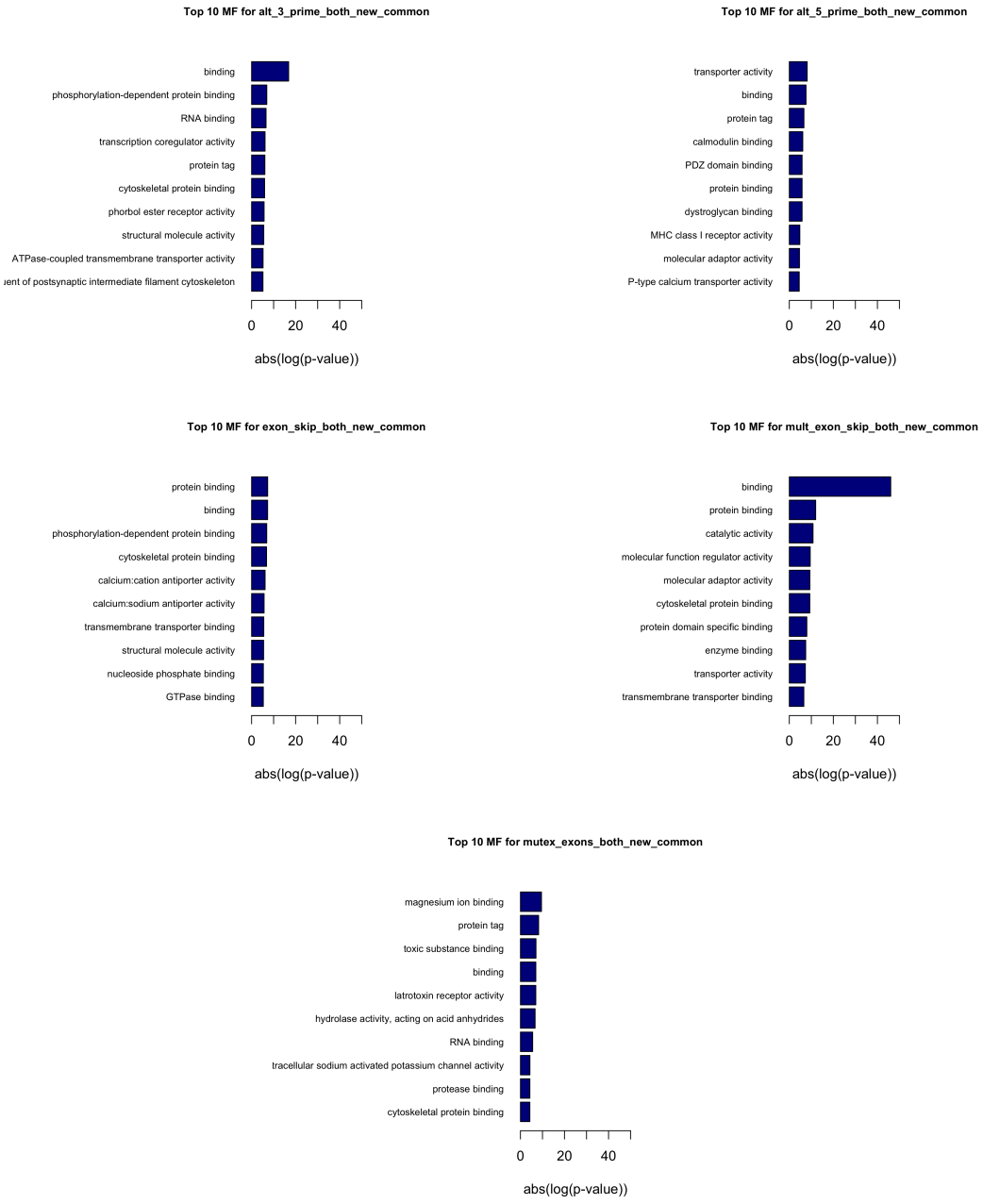


Figure S11: Top 10 MF terms for new+new group.