Silesian University of Technology

Faculty of Automatic Control, Electronics and Computer Science

Politechnika
Śląska

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

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

Faculty of Automatic Control, Electronics and Computer Science
Silesian University of Technology

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

In recent years, researchers have become more and more aware of the reproducibility crisis that the whole scientific community is facing. Despite comprehensive benchmarks and guidelines provided, the reproducibility problem is also an important issue for RNA-seq data analysis. 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 The problem is compounded by outdated but simply well-known tools remaining in frequent use long after they have been superseded by better methods, and by this affects the quality of the results. 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. 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 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. Although the human genome is the most studied and complete, there are still new annotation updates released, differing in the number of genes and transcripts. 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. 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.

The aim of this work is to investigate best practices in RNA-seq data analysis and to develop an approach to improve the reproducibility and robustness of the 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. What is unique about this solution is that it integrates multiple stages of RNA-seq data analysis, in contrast to existing workflows that focus on particular modules (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. That part is missing from standard analysis pipelines; however it is crucial for better understanding and annotation of gene models. The solution presented here bridges that gap, providing end-to-end workflow- from raw reads, through alignment, differential gene expression and 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.

Finally, 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, 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 work presents results for three different data sets- ‘real-world’ example of next-generation sequencing (NGS) data, NGS data used for benchmarking and microarray data. The main data set was composed of 88 samples, 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. There were four biological replicates for each condition. 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 reference data set was composed of RNA samples A and B from the SEQC2 consortium, 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. 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.
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).
Each data set provided new insights and improvements into build the pipeline and at the same time confirmed the choice of methods. All of them confirmed that there is no all- purpose solution both in terms of data analysis method and choice of the laboratory techniques. That is why it is very beneficial to provide options in high- throughput analysis pipelines, so that one can choose a path best suited for the purpose of the experiment. This is especially important when dealing with real- life scenarios when the signal might be weak, and the true distortions in the data might be unknown.

The pipeline uses the Snakemake workflow management system 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. Almost all pipeline's dependencies are installed via the Conda package and environment management system. With all requirements defined in the YAML file, Conda automatically builds a new environment, resolves conflicts, and downloads all dependencies. Additionally, pipeline consists of an R Markdown script for microarray data analysis and a set of R Markdown scripts for RNA-seq data analysis. General pipeline overview is presented in Figure 1.

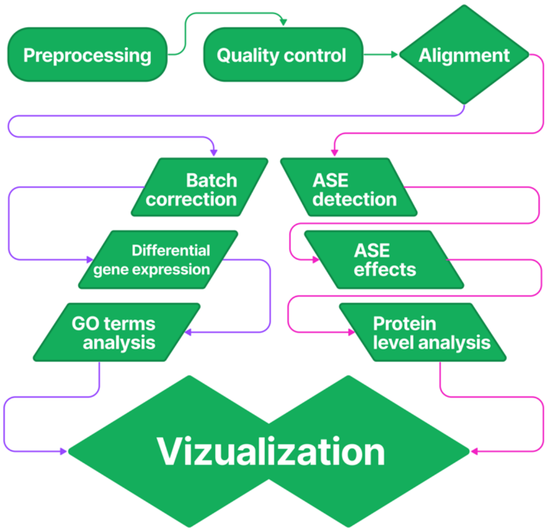


Figure 1: General pipeline overview.

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. The second and third option utilizes illuminaHumanv4.db package as annotation and VSN or quantile normalization provided by beadarray package.
- **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 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. The pipeline consists of following steps:

- **Quality control**

  Initial quality control on FASTQ files is performed with FastQC, then also alignment quality report is produced by MultiQC.
- **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 contains also quantification with StringTie2 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 usedfor 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.

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. 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 from different mice, it was obvious that it was affected by both known and hidden confounding factors. 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. Should we indicate that SHAM and PNSL samples come from different batches (Figure 2a) or should we treat them together (Figure 2b)? 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. In Figure 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. 3a) but even lower- 6% when applied together(Fig. 3b) 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 previous studies. 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.

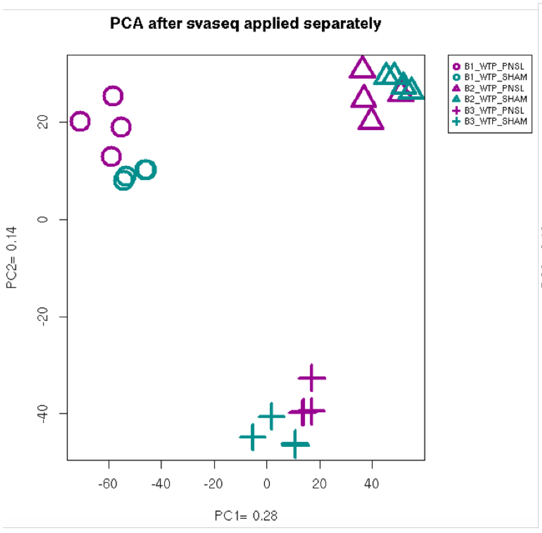


Figure 2a: PCA with 7 factors removed, applied separately for each batch

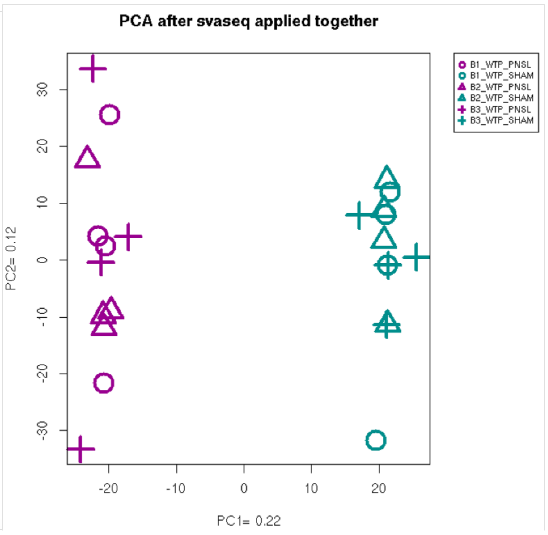


Figure 2b: PCA with 7 factors removed, applied together for each batch

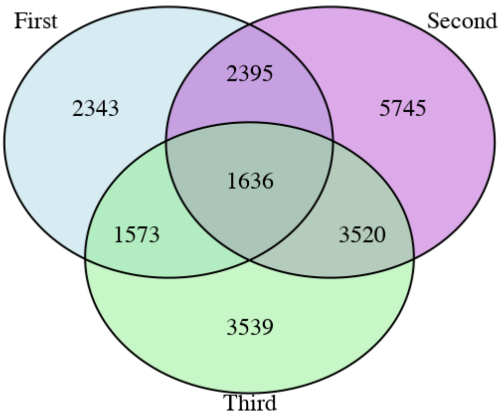


Figure 3a: Venn diagram for reaproducubility between 3 batches- SVAseq applied separately

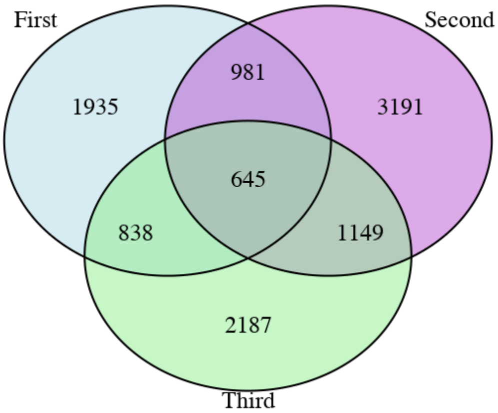


Figure 3b: Venn diagram for reaproducubility between 3 batches- SVAseq applied together

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

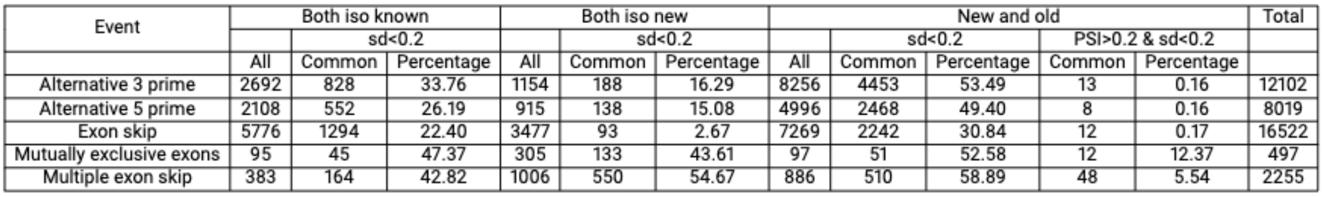
It can be very beneficial to think about the possibility of batch effects occurring in our data and to adequately account for them. The first issue to solve would be to carefully think about the questions we need to answer, and which comparisons should be made. In addition, choosing the right number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

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).

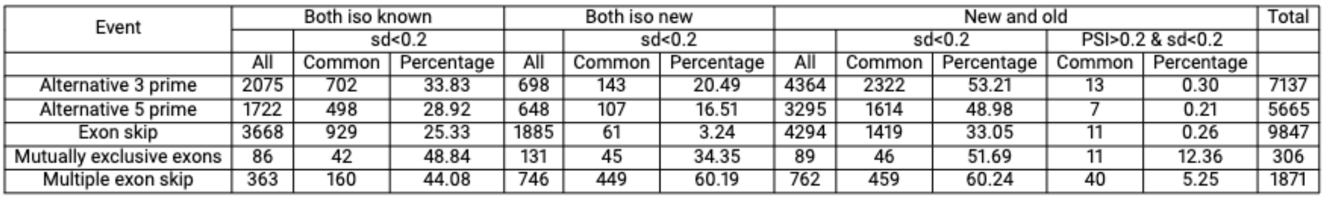
Spladder, among other metrics, reports the PSI (percent spliced in) value for each event and sample. This is the ratio of the signal supporting given event and sum of the signals for both events. Using PSI and appropriate threshold, False Positives number can be reduced. A set of criteria to choose valid events was chosen. The first one, applied for all three groups, was setting a threshold on PSI standard deviation. Those events that had it below 0.2 in all 4 replicates were treated as valid. 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 new ASE constitutes at least 1/4 of already known ASE.

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 (Table 1). They also affect a large number of genes (Table 2), but further analysis showed that although they appear to affect different processes and functions, they are all mainly connected to the nervous system as the overlap for Cellular Component GO Terms was higher than for others and also the terms were related to nervous system.



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 1: 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 2: Table showing number of genes containing detected ASEs depending on a type and group and also common number of genes.

Furthermore, the functional analysis with the combination of events from different groups showed that new events might provide new information for annotation. Figure 4 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. Protein phosphorylation is a major factor in signal transduction pathways. Transmembrane transporter activity could be related to signaling via neurotransmitters in the nervous system.

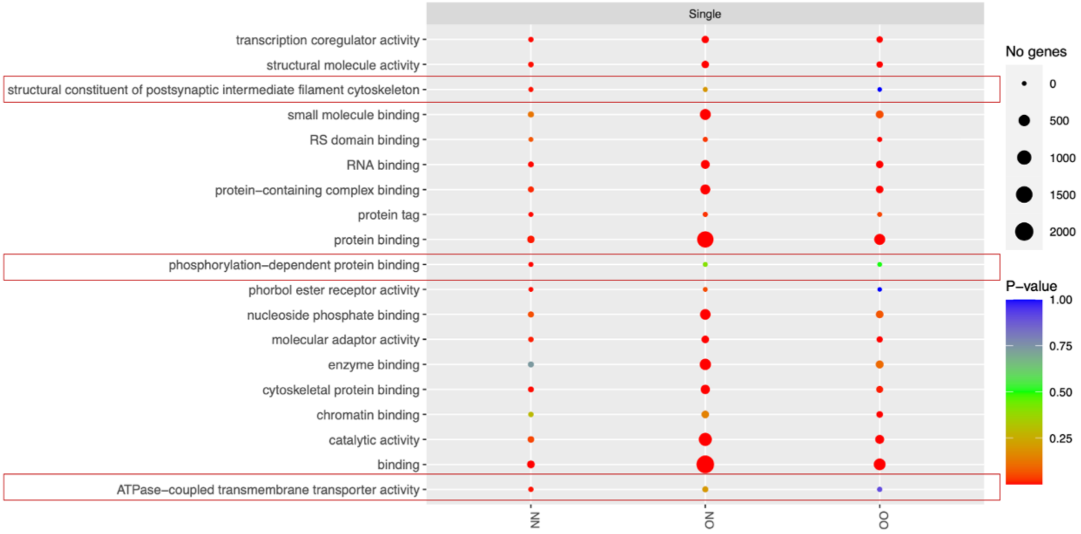


Figure 4: Top 10 MF GO terms for alternative 3' event shown across three groups.

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 (one of which is presented further) 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.

For 93 previously selected nASE of different types Bisbee and InterProscan analysis for new+old group was performed. Table 3 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 3: 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. 5 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. 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.

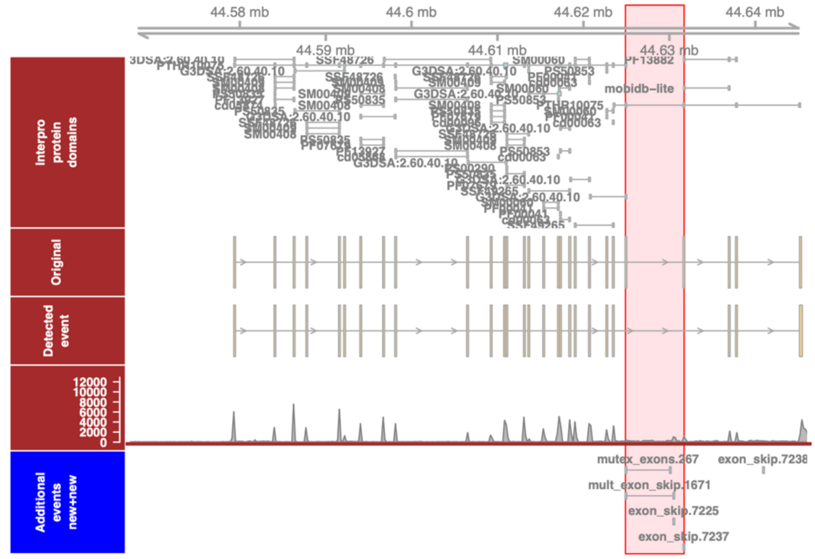


Figure 6: Visualization of multiple exon skip in the Nrcam gene.

The part for benchmarking RNA-seq was used to investigate how Spladder will work on the data where the signal is supposed to be strong. 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). 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.

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.

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.
2. Analysis recommendations based on studies with artificially created data sets should be applied to real-life problems with caution.
3. RNA- seq and microarray approaches both have strengths and weaknesses and should be used interchangeably, depending on the scientific problem.

This work showed also that the complexity of the mouse, but also the human genome, is not yet fully understood. Despite the presence of many confounding factors it was possible to detect common patterns in the data. It was obvious that the common Gene Ontology terms between different types of alternatively spliced events (ASEs) were related to the nervous system. It appears that the functions of the detected ASEs and the processes in which they are involved are characteristic of a given event type. Results indicate that there are many events not present in reference yet, and that great improvements can still be made in the field of reference transcript annotation. 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. Despite that a comprehensive tool for studying this aspect is still missing. As a result of this studies, a pipeline covering different aspects of high-throughput data analysis and bridging that gap was developed. It will help to improve reproducibility by providing more control over methods and parameters used at different stages of analysis but also by addressing the need of automatic detection and analysis of AS and its consequences. As AS plays also huge role in many diseases this tool can provide new insights into alterations occurring in conditions like Parkinson's disease, SMA or different cancer types.