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

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. The reason behind that is very often lack of applying those guidelines and best practices and also using tools, which are outdated but simply well-known. Another, more complex reason behind the lack of reproducibility in RNA-seq could be 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.

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 based on existing research and available tools, but modifies and combines them into a custom pipeline. 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). As there are no gold standards in high-throughput data analysis methods, it also provides different options for some stages. In addition, it also incorporates tools for alternative splicing (AS) analysis (AS detection and analysis of its consequences at the transcript and protein level) and visualization. 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.

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. 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.

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.