Optimization of an RNA-seq Differential Expression Workflow

Master's Thesis in Bioinformatics - June 15 (2025)

Carolina F. Jørgensen
Bioinformatics Research Center (BiRC)
Department of Molecular Biology and Genetics
Aarhus University
Aarhus, Denmark
201707243@post.au.dk

Supervisors:
Thomas Bataillon - BiRC
Stine Rye Østergaard - Former Lead Data Scientist at BioXpedia

Keywords: Differential Gene Expression Analysis · DESeq2 · Empirical Bayes methods · Genomic Control · Multiple testing correction

ABSTRACT

This thesis has two primary objectives: (1) to optimize an RNA-seq differential expression analysis workflow originally developed by the company BioXpedia and (2) to apply the optimized workflow to experimental data provided by a research team at *Rigshospitalet*, with the aim of generating biologically relevant results. The results obtained from the analysis of the experimental data were used iteratively to assess and guide improvements during the workflow optimization process.

The finalized optimized workflow consists of three scripts: (1) the first script is responsible for data formatting, (2) the second performs differential gene expression analysis using the *DESeq2* package, incorporating LFC shrinkage, genomic control and improved multiple testing correction, and (3) the third script conducts gene enrichment analysis and compares the results to those obtained in the original analysis. Key enhancements to the original BioXpedia pipeline include the integration of LFC shrinkage, genomic control and more robust methods for multiple testing adjustment.

The research group was primarily interested in eight specific comparisons derived from the experimental data. The key findings related to these comparisons are as follows: (1) Comparison 1 exhibited a particularly strong signal, indicating a substantial number of differentially expressed genes, (2) Comparison 7 showed a weaker, yet similar signal and (3) Comparison 4 demonstrated a modest but non-negligible signal, suggesting potential relevance. The results indicate that the treatment of interest to the research group targets lean individuals and has the greatest effect when there has been no meal intake.

Applying both the initial and optimized workflows to the experimental data demonstrated that the optimized version substantially reduces noise and, hence, false discoveries. The incorporation of LFC shrinkage, genomic control and a more refined approach to multiple testing correction appears to enhance both the sensitivity and specificity in identifying differentially expressed genes. As a result, the optimized workflow enables more reliable detection of truly differentially expressed genes, leading to a more accurate interpretation of the underlying biological signals. The workflow is adaptable to other transcriptomic studies, particularly in contexts where an exploratory yet statistically robust method is required.

1 Acknowledgements

I would like to express my gratitude to my supervisors, Prof. Thomas Bataillon and Stine Rye Østergaard, for their invaluable guidance and encouragement throughout the course of this thesis. Their expertise and consistent engagement have been instrumental in shaping both the direction and quality of this thesis. I am especially grateful for their willingness to discuss ideas and for creating a supportive research environment that allowed me to grow academically.

Furthermore, I acknowledge the assistance of OpenAI's ChatGPT. It was used to improve the clarity and academic tone in certain parts of the thesis, as well as to refine code and implement minor modifications. The support provided by ChatGPT was limited to mainly aesthetic enhancements and did not influence the scientific content or the conclusions drawn in this thesis.

Contents