Generating reproducible analysis pipelines in biomedical informatics

The Ohio State University
College of Medicine

Caroline He, Qin Ma PhD, Megan McNutt BS
Bioinformatics and Mathematical Biosciences Lab. https://u.osu.edu/bmbl/
Department of Biomedical Informatics

Background

As a high school senior with a little coding experience, understanding basic java and html, R language is a brand new concept to me, I initially perceived this endeavor as a significant challenge, however, my passion for science compelled me to enroll in a wide range of science courses during my time in school, ultimately steering me toward this internship opportunity. At school, other than regular biology and chemistry which is mandatory, I went beyond than the science credit requirement, challenged myself with AP Biology and AP Environmental Science. These advanced courses significantly enriched my understanding of concepts related to single cells, proving invaluable during my internship.

And a short background about the project I'm currently working on, the GitHub Notebook serves as a collection of bioinformatics data analysis examples. This notebook encompass both data and code for reference. Which is focusing on three major aspects

- Apply current best practices in single-cell analysis using real data.
- Generate comprehensive reports to share with biological collaborators.
- Find code snippets to quickly produce results and figures.

Objectives

- Gain a deep understanding of the inner workings of the tasks I'm currently performing.
- Enhanced job adaptability resulting in increased efficiency and quicker task completion

Contributions

- Updating code
  - replacing Reactome with KEGG in "downstream_analysis_template.Rmd"
  - replacing Deseq2 with Limma in "downstream_analysis_template_using_limma.Rmd"
  - updating visualization codes (enrichment bar plot, volcano plot, heatmap)
- Figuring out how to access working features
  - including gaining access on Github Desktop, to update information

References

https://www.e3s-conferences.org/articles/e3sconf/pdf/2021/47/e3sconf_icepe2021_03058.pdf
https://www.gsea-msigdb.org/gsea/msigdb/human/collections.jsp#C2
https://github.com/OSU-BMBL/BMBL-analysis-notebooks/tree/master/RNAseq_workflow/Example_Data

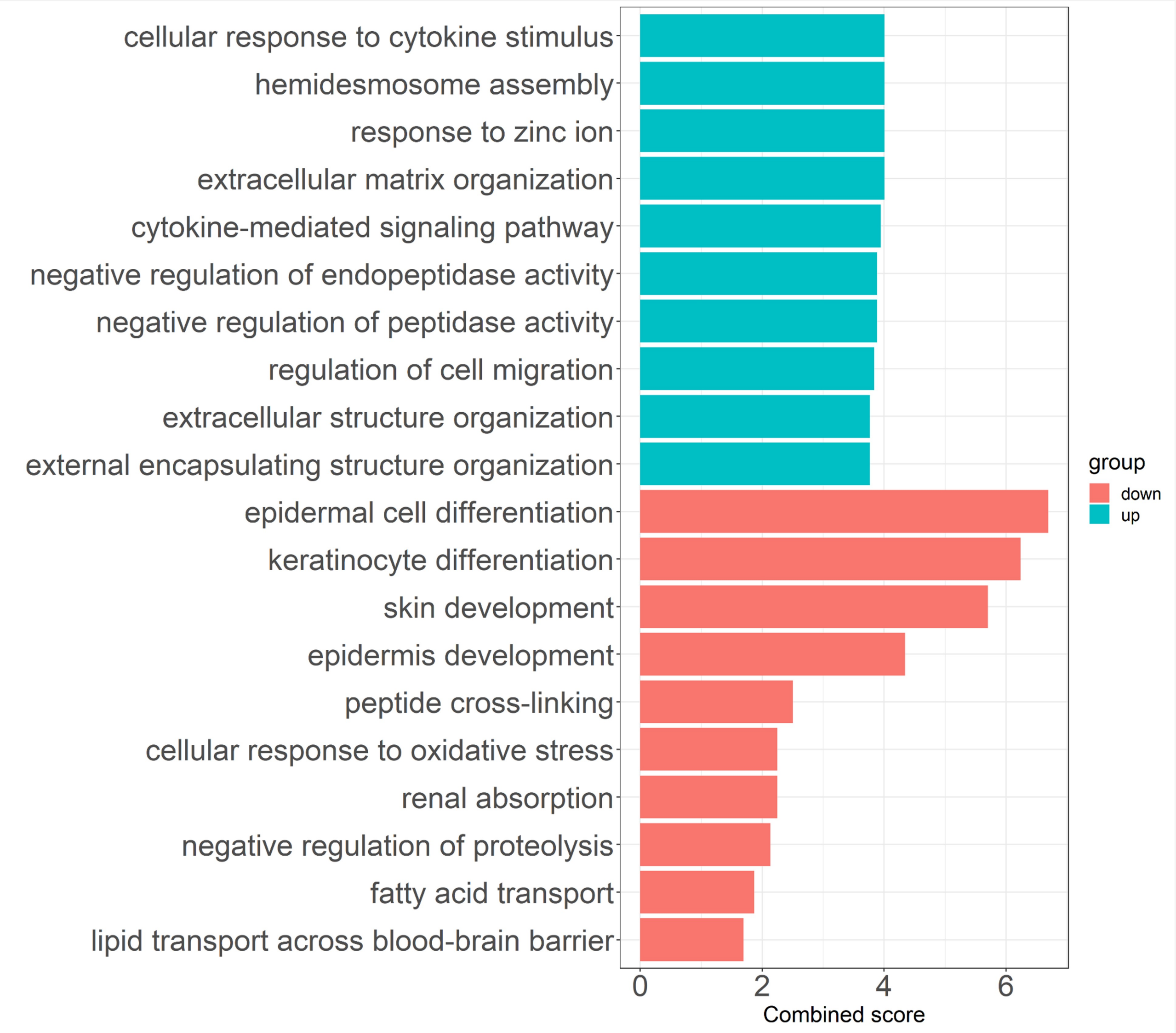

Figure 1.bar plot showing the enrichment score of top 10 up-regulated and down-regulated pathways

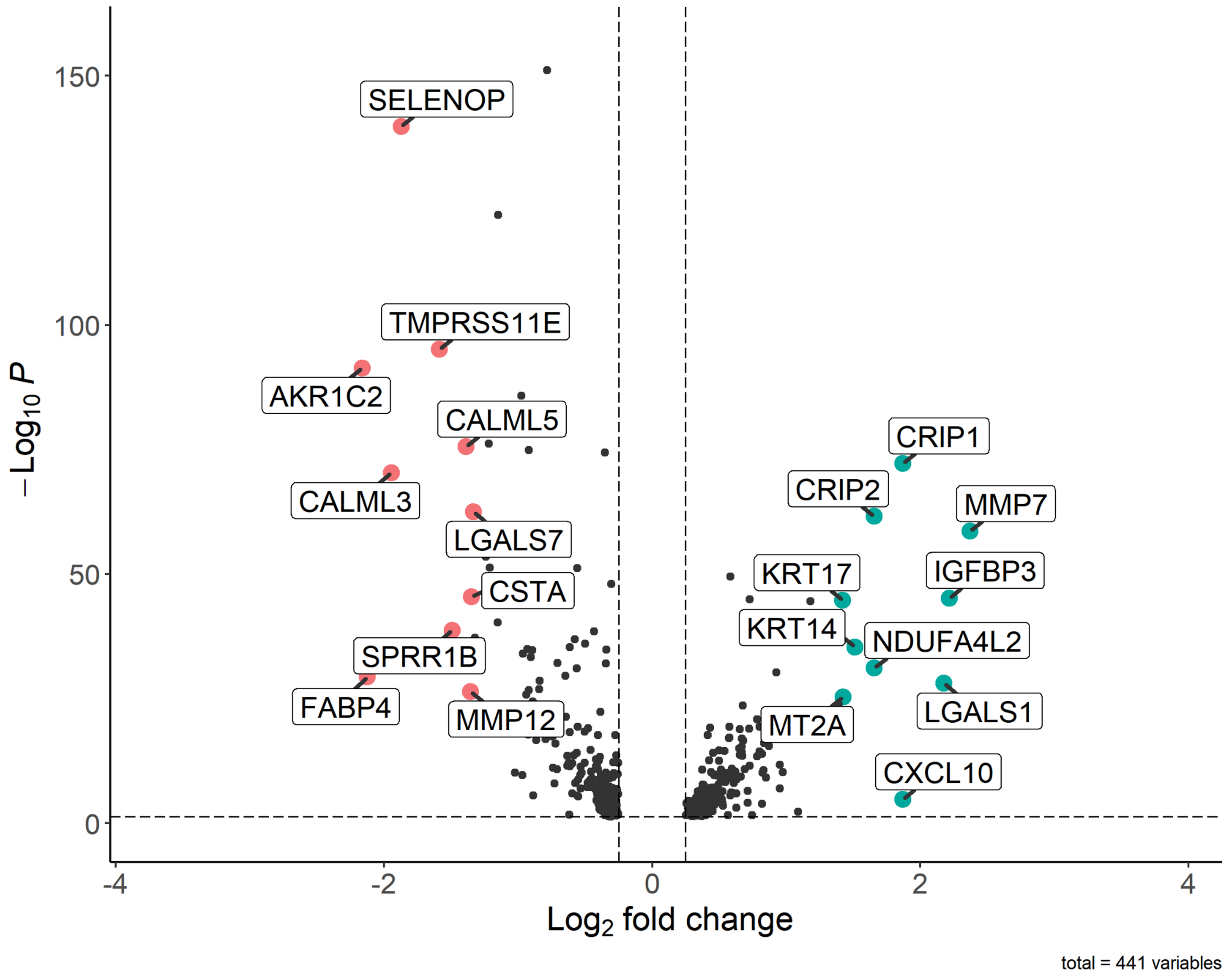

Figure 2. Volcano plot showing differentially expressed genes (DEGs). Both top 10 up-regulated and down-regulated DEGs were highlighted.

Methods

- Gained understanding of data analysis functions in R to better understand tasks
  - limma and DESeq2 are quite frequently used in gene expression analysis, they still have many differences.
  - DESeq2 uses the non-linear model, while limma uses the linear model.
- Understanding the difference between the KEGG and REACTOME databases
  - KEGG-> metabolic pathways
  - REACTOME-> pathways and processes
  - Genes can be seen as overexpressed in KEGG

Results

What I learned and what I hoped to improve

- I learned the basic function of R, including running it, checking for errors, etc.
- I also learned when replacing the code, how to google for the replacement, and utilize outside resources to solve my problems
- I hope to enhance my comprehension of the code, pinpointing specific errors and comprehending the function of entire code segments. During my efforts to replace code, I often encounter difficulties because I lack a full understanding of what went wrong or how the replacement should be implemented. Most likely I'm just trying different things out until it works.

Outcomes

- Updated 5 separate workflows in GitHub
- Gained an understanding of R, the research process, and GitHub

Acknowledgements

Qin Ma, PhD

Megan McNutt

Caroline He

I would like to thank Dr. Ma for the opportunity, and Ms. Megan McNutt for the instruction and expert insight