BIO-463: Mini project

Reproducing analysis from the paper: Gene Expression Profile on Human Mesenchymal Stromal Cells Exposed to Hypoxic and Pseudohypoxic Preconditioning - An Analysis by RNA Sequencing

Marija Zelic
Section of Life Sciences Engineering

May 2024

1 Introduction

Mesenchymal stromal cells (MSCs) are a subgroup of stem cells with multipotent capabilities, most frequently found in the bone marrow and other musculoskeletal system's connective tissues, which they can repair. Due to their properties, MSCs have become a promising tool in developing modern and efficient future treatment strategies [1]. However, the benefits of MSC administration to patients in clinical trials are less than expected. In the last decade, priming or preconditioning MSCs has gained credibility to enhance MSC therapeutic potential. Priming with hypoxia represents a pivotal approach for enhancing MSC functionality, as it stimulates the secretion of essential growth factors, such as vascular endothelial growth factor (VEGF) and HGF, which are crucial for angiogenesis and tissue regeneration [2]. In this paper [3], authors wanted to compare gene regulatory effects induced by physical hypoxia (culturing cells for 6h in a gas mixture composed of 2% O_2) and by pharmacological inhibition of only hypoxia-inducible factor prolyl hydroxylase 2 (PHD2) by Vadadustat (AKB-6548) in a concentration of 40μM.

For RNA isolation, six populations of bone marrow MSCs were cultured (three populations per each condition - Control, Hypoxia and Vadadustat), therefore 18 samples in total. Data used to reproduce the differential gene expression (DEG) analysis figures were openly available in the NCBI Gene expression omnibus (GEO) repository with accession number GSE180371 as relative expression of transcripts, quantified for each donor using the Salmon v0.8.1 method. On the other hand, gene set enrichment analysis (GSEA) was performed on normalized counts generated in DESeq2, which can be found in the paper's supplementary materials.

2 Methods

2.1 Differential gene expression analysis

DEG analysis is one of the most important steps in RNA-seq workflow. Its goal is to determine which genes are expressed at different levels between conditions by performing statistical analysis on read count data. For instance, we utilize statistical testing to assess whether the observed variation in read counts for a particular gene is significant, indicating it surpasses what could be attributed to the natural random fluctuation. Specifically in this analysis, to import transcript abundance datasets for DEG from the relative expression of transcripts data, the tximport pipeline was used [4] and DESeq2 software version 1.44.0 [5].

Regarding the *tximport* function parameters: files were specified as character vectors of file names, the type of software used to generate transcript-level abundances is "salmon" and the tx2gene parameter specifying data frame associating transcripts with gene IDs for gene-level summarization was obtained from Homo_sapiens.GRCh37 reference genome (Basic Gene Annotation GTF files [6]). For other parameters, default values were preserved.

For differential gene expression analysis, *DESeqDataSetFromTximport* function was used. As parameters, it takes the result of *tximport* function, metadata data frame that includes information about sources of variation per each sample - in this case, condition (Control, Hypoxia and Vadadustat) and design formula that should capture all the major sources of variation (here, condition).

2.2 Gene Set Enrichment Analysis

GSEA represents a powerful analytical method for interpreting gene expression data, first described in the paper [7]. Given a priori defined set of genes (e.g., genes linked to the cell cycle control or map kinase signalling, etc.), the ultimate goal of GSEA is to determine whether the members of this dataset are randomly distributed over the inputted list of genes or primarily found at the top or bottom, i.e., whether genes from the dataset are enriched in the experimental dataset or control. It differs from other enrichment analysis methods as it does not necessitate

selecting DEGs by applying various thresholds, which can significantly influence the final outcomes. To obtain plots, GSEA was performed using GSEA software version 4.3.3.

The analysis used h.all.v.7.4.symbols.gmt (Hallmark) and c5.all.v.7.4.symbols.gmt (Gene Ontology) gene set databases with the following settings: 1000 permutations, collapse dataset to gene symbols - true (Human_ENSEMBL _Gene_ID_MSigDB.v7.4.chip), permutation type - gene_set, enrichment statistics - weighted, metric for ranking genes - Signal2Noise, gene list sorting mode - real, gene list ordering mode - descending, max size of gene sets - 500, and min size of gene sets - 15.

3 Results/Discussion

We first display the results of differentially expressed genes obtained in DESeq2. A general way to provide a global view of gene expression levels is using volcano plots, in which the log-transformed adjusted p-values are plotted on the y-axis and $\log_2$ Fold Change values on the x-axis. In Figures 1 to 3 three groups were analyzed: Vadadustat vs. Hypoxia, Hypoxia Vs. Control, and Vadadustat vs. Control. Dots in red with depicted HGNC symbols next to them are genes considered significantly differentially expressed - with an adjusted p-value smaller than 0.05. A key purpose of volcano plots is to identify genes significantly upregulated or downregulated between different conditions specified by the design formula in the DESeq2 software. In terms of the biological question at hand, volcano plots provide useful insights into the numbers and profile of genes altered by preconditioning with two hypoxia methods.

The plot in Figure 4 represents the numbers of upregulated and downregulated genes in each of the three comparisons. Upregulated and downregulated genes are identified by first selecting differentially expressed genes with adjusted p-values less than 0.05. Genes with $\log_2$ Fold Change above 0 are classified as upregulated, while those with $\log_2$ Fold Change below 0 are classified as downregulated. These numbers are important because they provide insights into the molecular response of the system under study. For instance, we can observe that preconditioning with physical hypoxia resulted in much fewer DEGs compared to the case of pharmacological hypoxia.

The plots in Figures 5a, 8a, 11a represent results of GSEA Hallmark analysis showing enriched gene sets for each of three conditions, Hypoxia vs. Control, Vadadustat vs. Control and Vadadustat vs. Hypoxia, respectively. Bars in red indicate significant enrichment at FDR $< 25\%$, while bars in grey represent gene sets with FDR $> 25\%$. Essentially, FDR provides the probability of significantly enriched gene sets (i.e. ones with significant nominal p-value) being a false positive. A positive normalized enrichment score (NES) value indicates enrichment in the Hypoxia, Vadadustat and again Vadadustat phenotype, respectively.
The plots in the Figures 5b, 5c, 8b, 8c, 11b, 11c depict top 50 marker genes for each phenotype in the comparison of Hypoxia vs. Control, Vadadustat vs. Control and Vadadustat vs. Hypoxia, respectively. Expression values are represented as colors and range from red (high expression), pink (moderate), light blue (low) to dark blue (lowest).
Figures 6, 9, and 12 display enrichment plots for the top five datasets enriched in GSEA Hallmark Hypoxia vs. Control, Vadadustat vs. Control and Vadadustat vs. Hypoxia analysis, respectively. They show the profile of the running ES score and positions of gene set members on the rank-ordered list. The enrichment score reflects the degree to which a gene dataset is overrepresented at the extremes (top or bottom) of the entire ranked gene list.
Finally, the plots in Figures 7, 10, and 13 reflect gene sets significantly enriched (FDR q-value < 0.25) in the comparison Hypoxia vs. Control, Vadadustat vs. Control and Vadadustat vs. Hypoxia, respectively, using gene ontology biological processes database. Gene sets were ordered by decreasing NES, where positive NES indicates enrichment in Hypoxia, Vadadustat and again Vadadustat, respectively. Also, number of genes assigned to each data set is given next to the bar.
From the comparative analysis of the results obtained by GSEA, and related to the current biological inquiry we can identify changes in the expression of genes belonging to different categories in pharmacological and physical hypoxia. However, GSEA does not indicate the direction of changes in gene expression, in terms of upregulated and downregulated genes, which differential gene expression analysis certainly provides.

4 Conclusion

Reproducibility analysis of the study by Zielniok et al. [3] using the available data partially matched reported results. We were not able to recreate identical plots related to differential gene expression analysis. Even though patterns in the volcano plot and identified significant genes are similar to those reported in the paper, corresponding adjusted p-values are not identical. A major reason behind this is an insufficiently detailed pipeline of differential gene expression analysis. We are only informed about the usage of *tximport* pipeline and DESeq2 software. Therefore, we are unaware of the parameters they specified, potential filtering or scaling, i.e. data preprocessing of any kind, variations in the design formula, etc. Similarly, some important thresholds, such as the one determining significant genes and specifying differences between upregulated and downregulated genes are also not indicated. Consequently, there is a mismatch in the number of upregulated and downregulated genes. The plot shown in Figure 4 was generated using a $\log_2$ Fold Change threshold of 0 to closely resemble, in terms of the relative counts, results reported in the paper. However, thresholds of greater than 1 for upregulated genes and less than -1 for downregulated genes are typically considered more appropriate. Applying these more stringent thresholds yielded counts of upregulated and

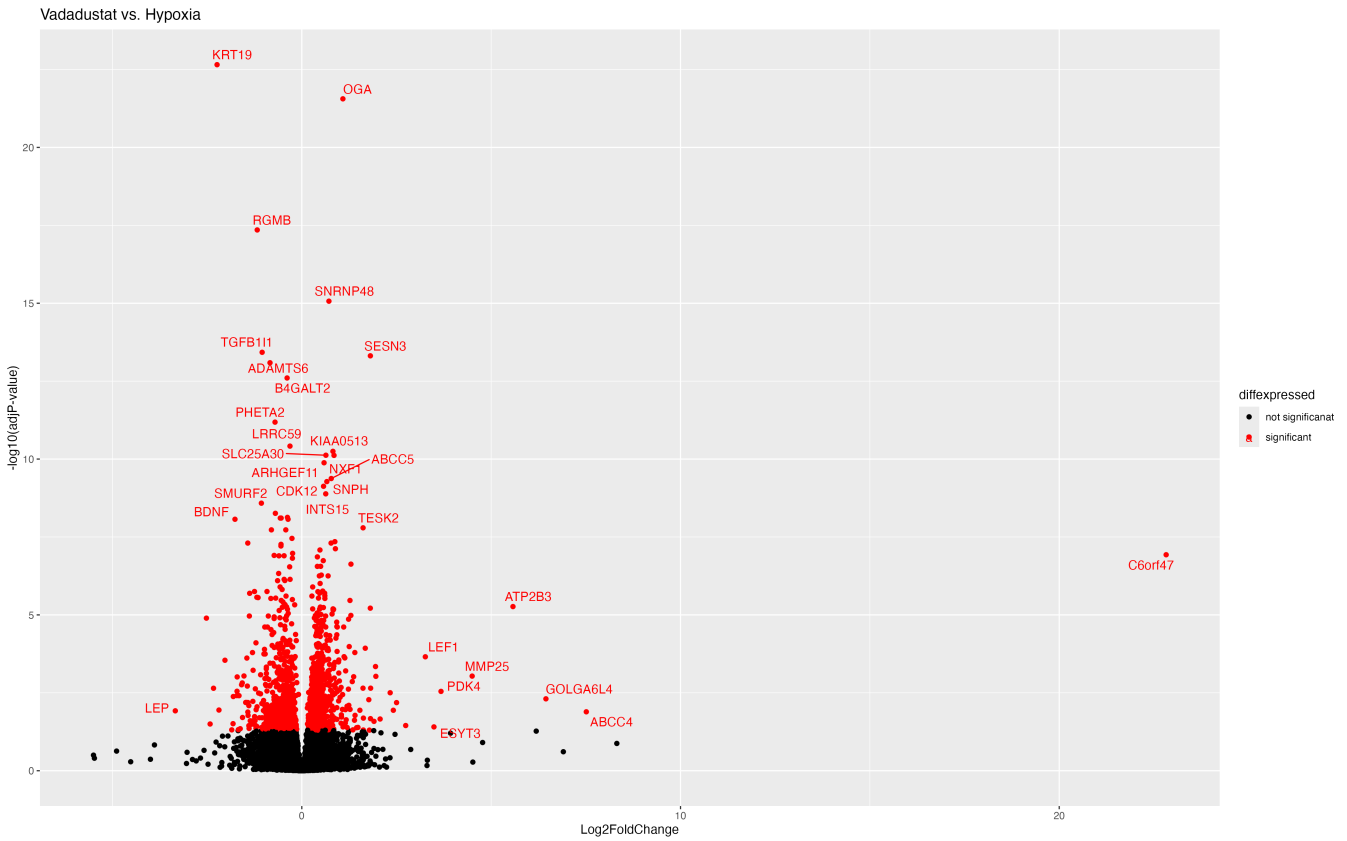


Figure 1: Volcano plot showing the significance of differential gene expression data (adjusted p-value) versus magnitude of expression change ($\log_2$ Fold Change) from comparison Vadadustat vs. Hypoxia.

downregulated genes that are not comparable to those reported in the original study.

On the other hand, the results obtained for GSEA Hallmark analysis and heat map of the top 50 marker genes match the plots presented in the paper. The justification for this lies in the use of normalized counts obtained from DESeq2, as provided by the authors in the supplementary material, along with a detailed explanation of their pipeline and parameter specifications. However, setting the parameter "collapse data set to gene symbol" to false caused an error in the GSEA software, likely because we used DESeq normalized counts with ENSEMBL gene names instead of gene symbols, which the authors might have used. Additionally, there is a potentially contentious statement in the GSEA plot descriptions: "Bars in red indicate significant enrichment at FDR $< 25\%$, bars in grey represent gene sets with FDR $> 25\%$ and a nominal p-value $< 5\%$." The relevance of the nominal p-value $< 5\%$ is unclear, as our analysis produced the same results without filtering based on the nominal p-value. It is worth noting that gene set enrichment using gene ontology biological processes does not give the same results, specifically not all gene sets with FDR q-value < 0.25 reported in the original study are present in our analysis.

Based on the reported findings, we can conclude that the reproducibility of the results was not a straightforward procedure due to the vague description of the analysis pipeline. Overall, results were not robust on unclear parameters and function selection.

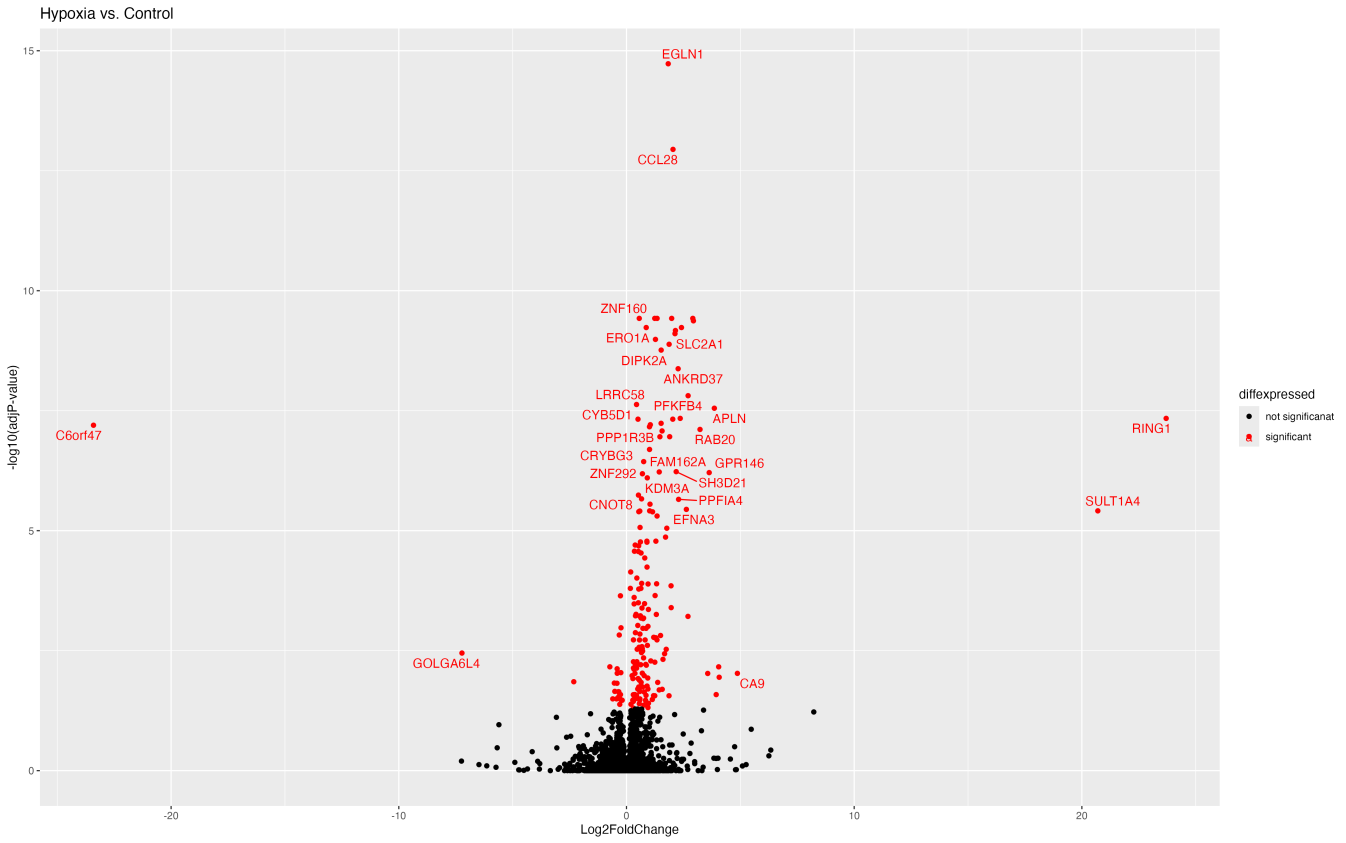


Figure 2: Volcano plot showing the significance of differential gene expression data (adjusted p-value) versus magnitude of expression change ($\log_2$ Fold Change) from comparison Hypoxia vs. Control.

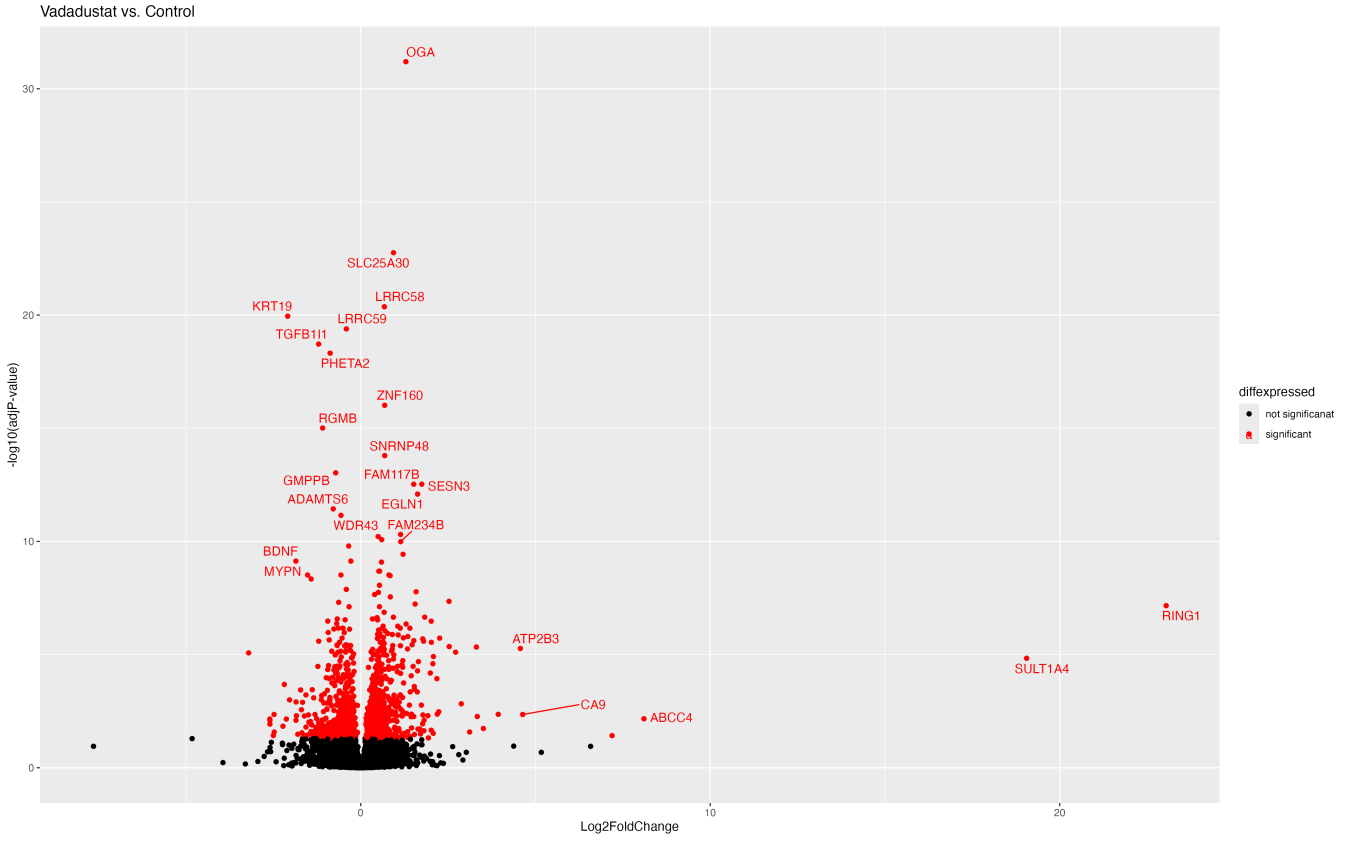


Figure 3: Volcano plot showing the significance of differential gene expression data (adjusted p-value) versus magnitude of expression change ($\log_2$ Fold Change) from comparison Vadadustat vs. Control.

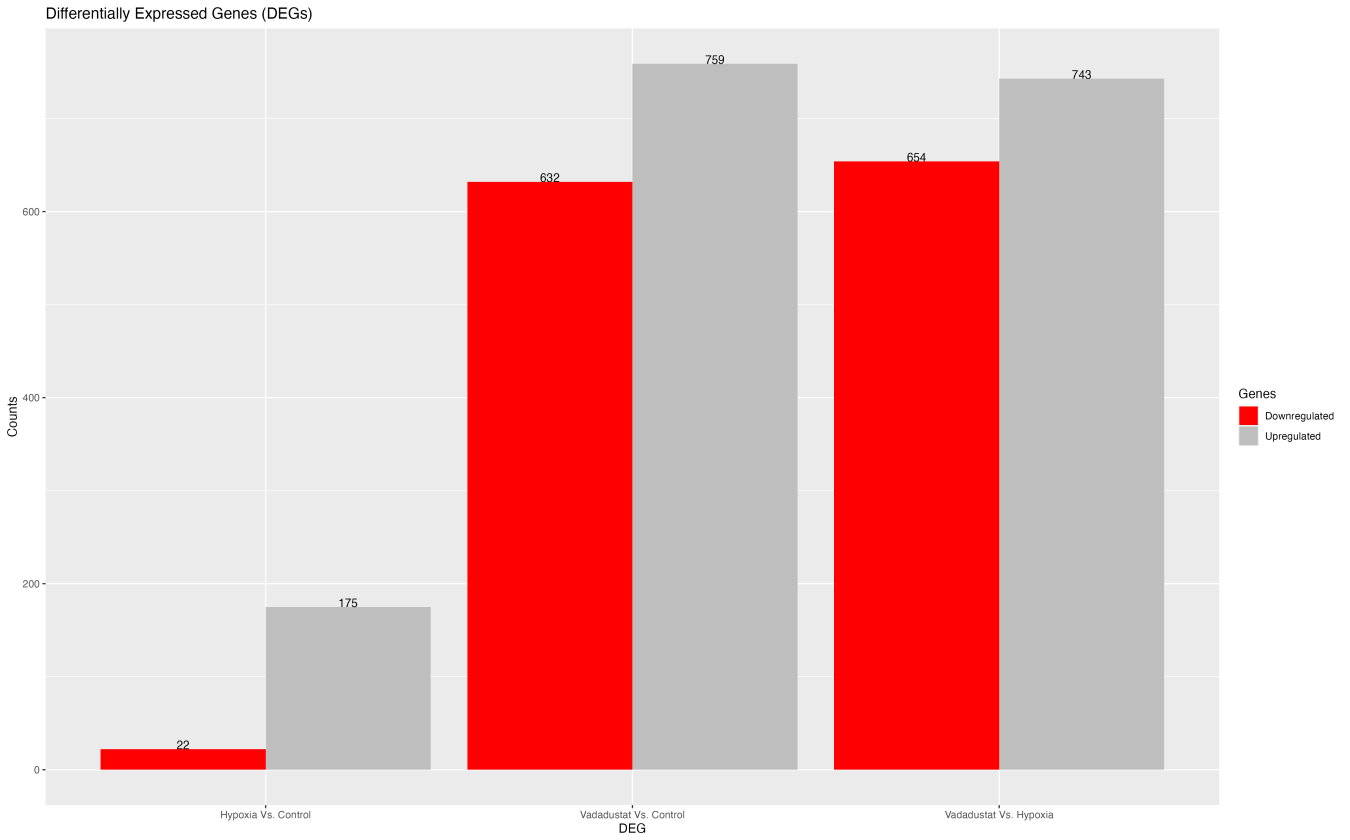
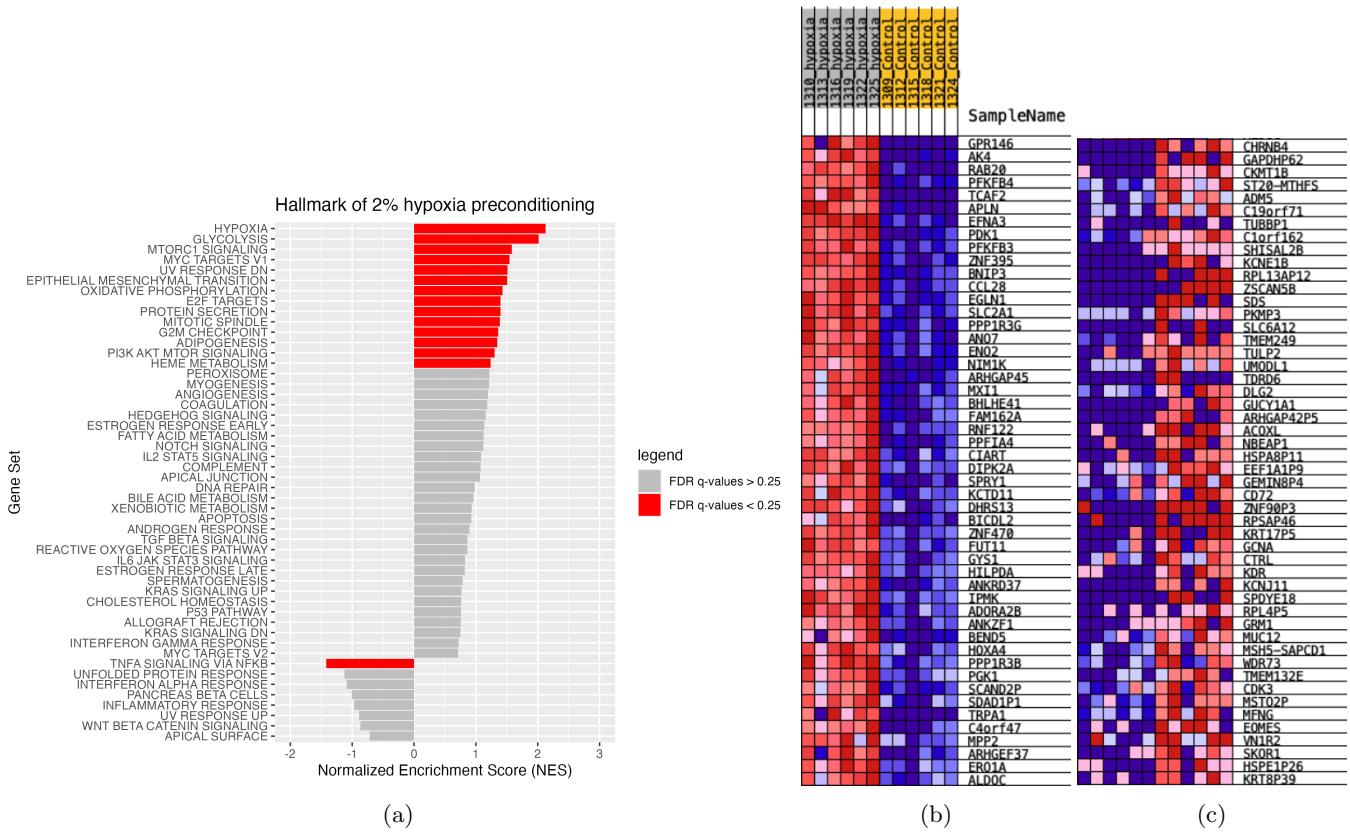


Figure 4: The number of upregulated and downregulated DEGs obtained from all three comparisons.



(a) (b) (c)

Figure 5: Gene set enrichment analysis for Hypoxia vs. Control. (a) Results of GSEA Hallmark analysis showing enriched gene sets. (b) and (c) Heat map of the top 50 marker genes for each phenotype in the comparison of Hypoxia vs. Control.

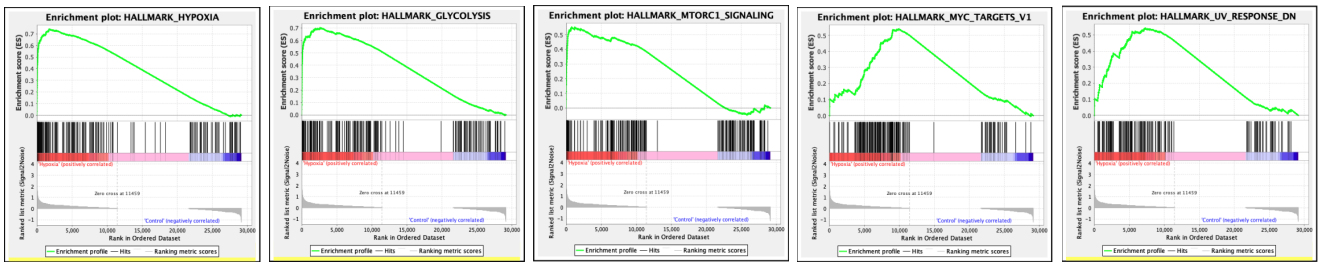


Figure 6: Enrichment plots for top five data sets enriched in GSEA Hallmark Hypoxia vs. Control analysis.

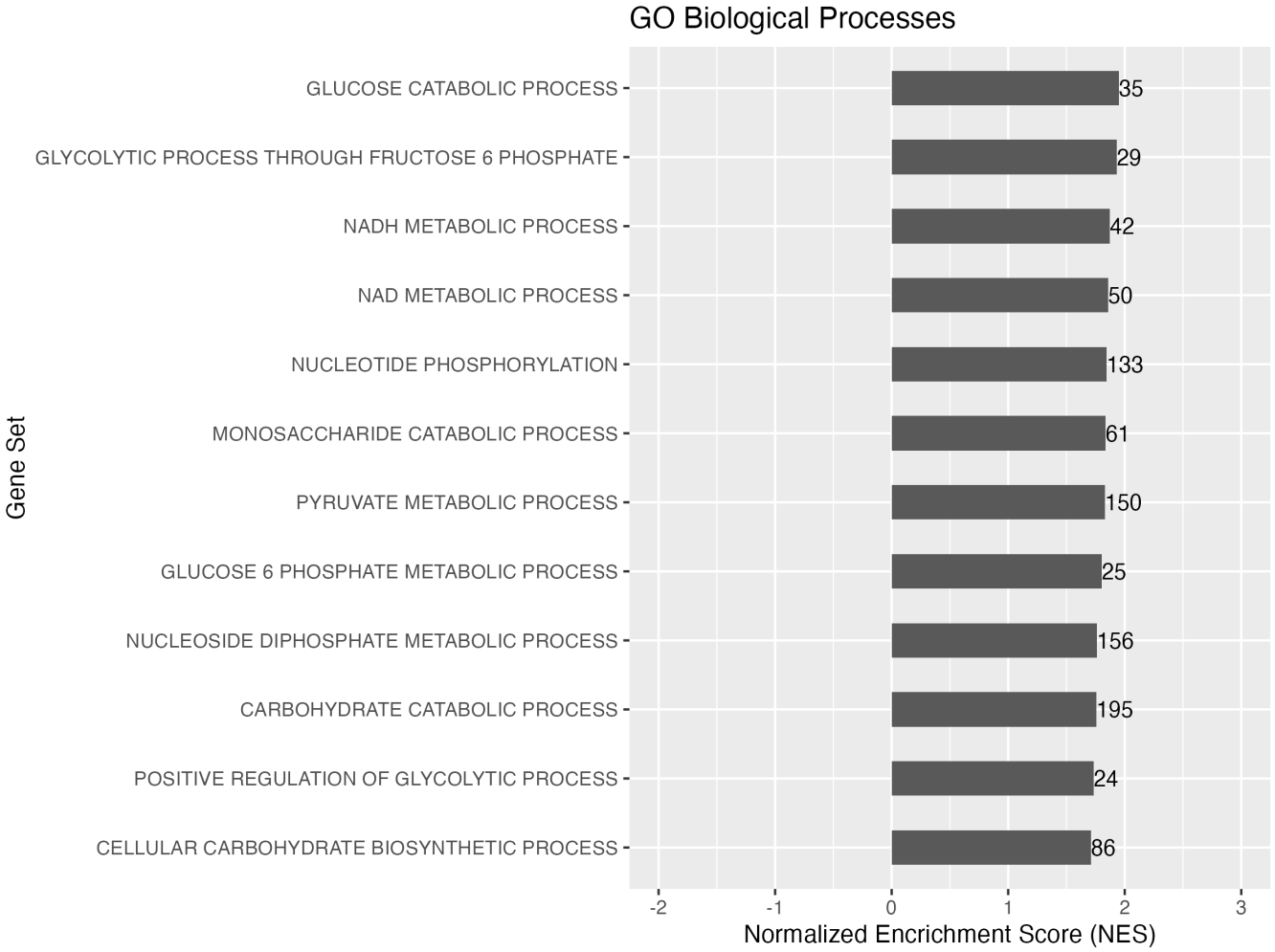
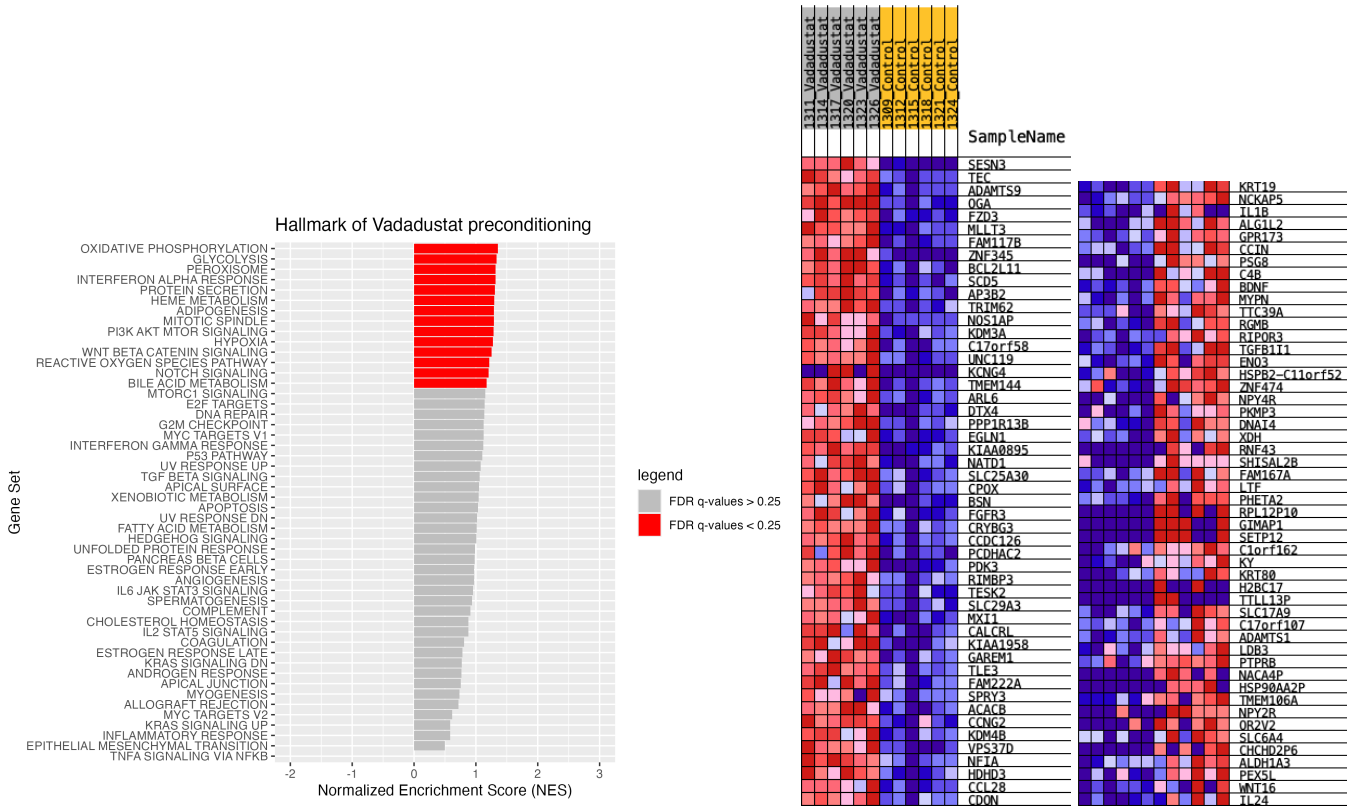


Figure 7: Gene sets significantly enriched in the comparison of Hypoxia vs. Control using the gene ontology biological processes.



(a) (b) (c)

Figure 8: Gene set enrichment analysis for Vadadustat vs. Control. (a) Results of GSEA Hallmark analysis showing enriched gene sets. (b) and (c) Heat map of the top 50 marker genes for each phenotype in the comparison of Vadadustat vs. Control.

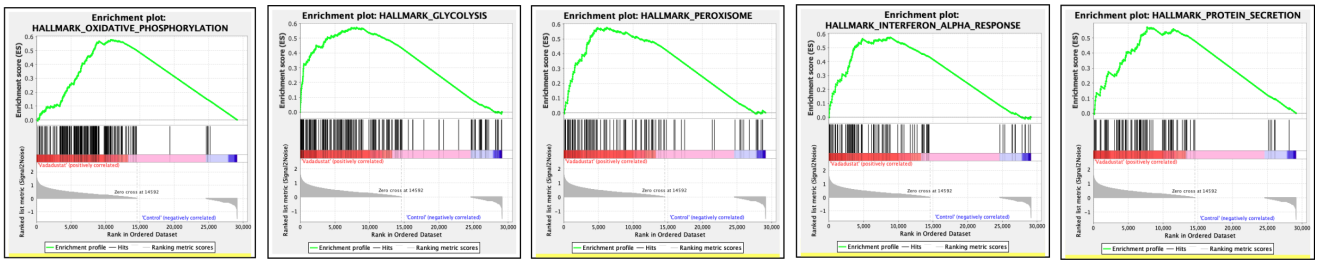


Figure 9: Enrichment plots for top five data sets enriched in GSEA Hallmark Vadadustat vs. Control analysis.

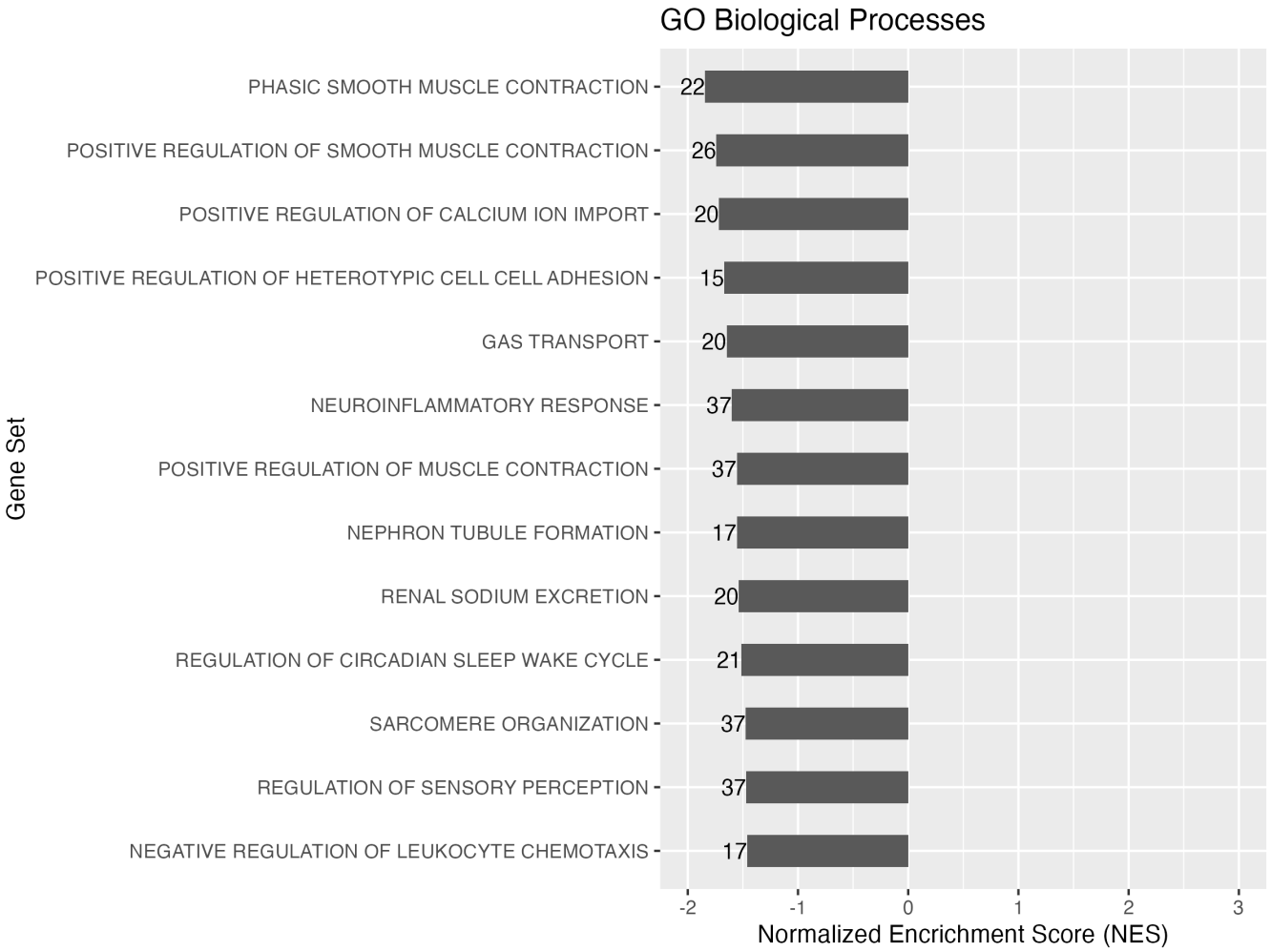
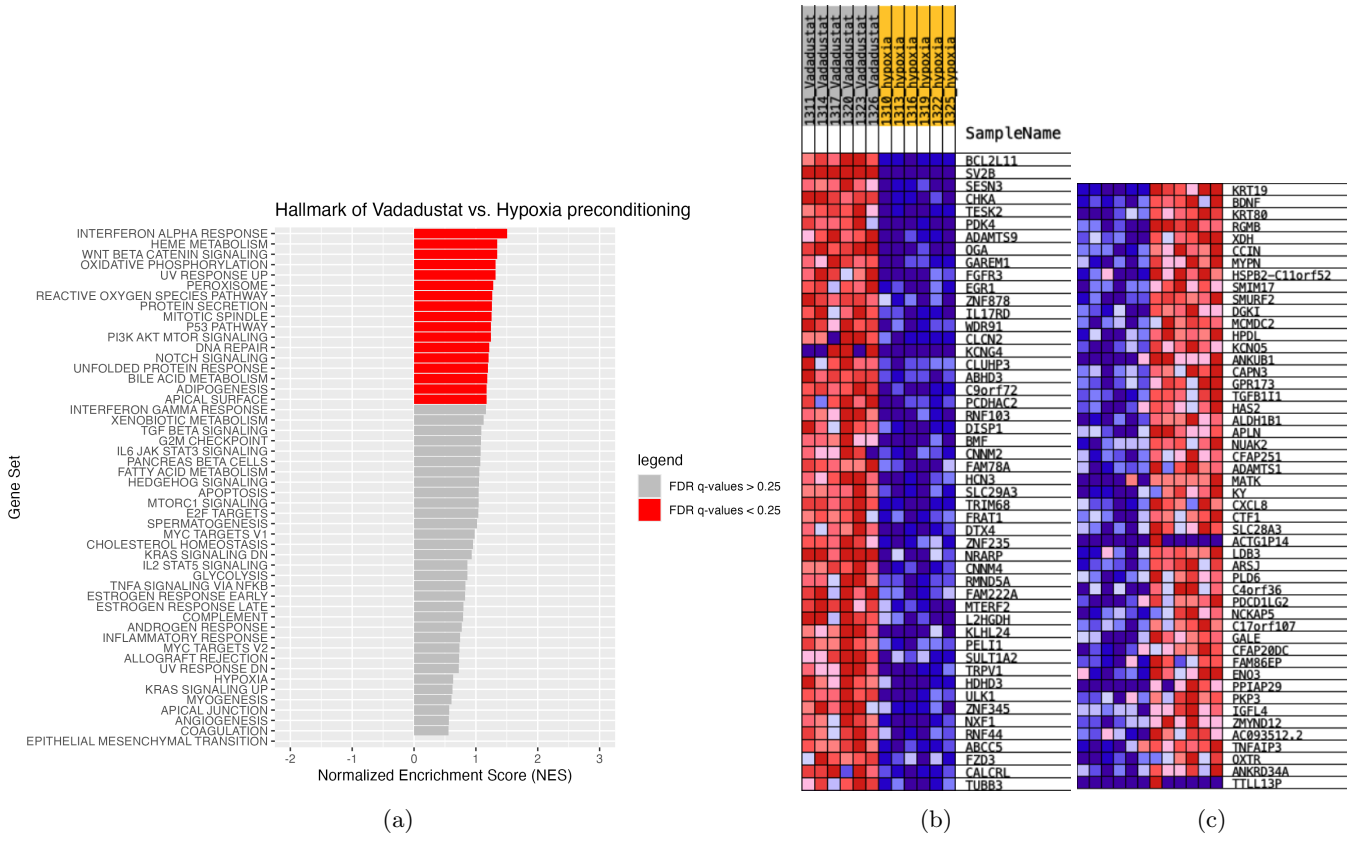


Figure 10: Gene sets significantly enriched in the comparison of Vadadustat vs. Control using the gene ontology biological processes.



(a) (b) (c)

Figure 11: Gene set enrichment analysis for Vadadustat vs. Hypoxia. (a) Results of GSEA Hallmark analysis showing enriched gene sets. (b) and (c) Heat map of the top 50 marker genes for each phenotype in the comparison of Vadadustat vs. Hypoxia.

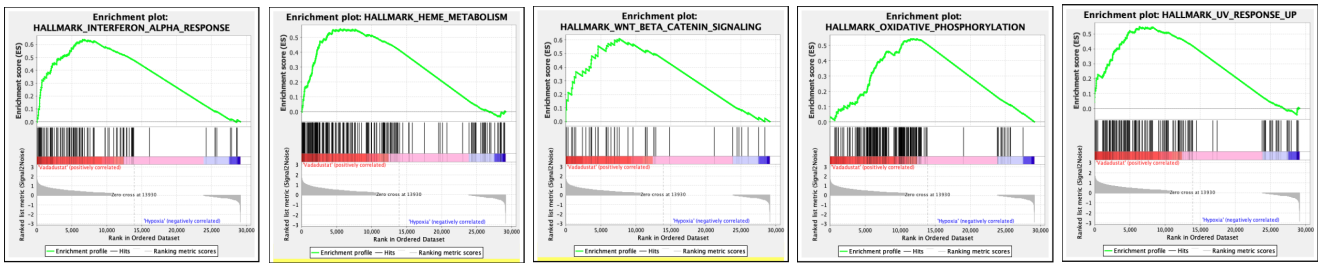


Figure 12: Enrichment plots for top five data sets enriched in GSEA Hallmark Vadadustat vs. Hypoxia analysis.

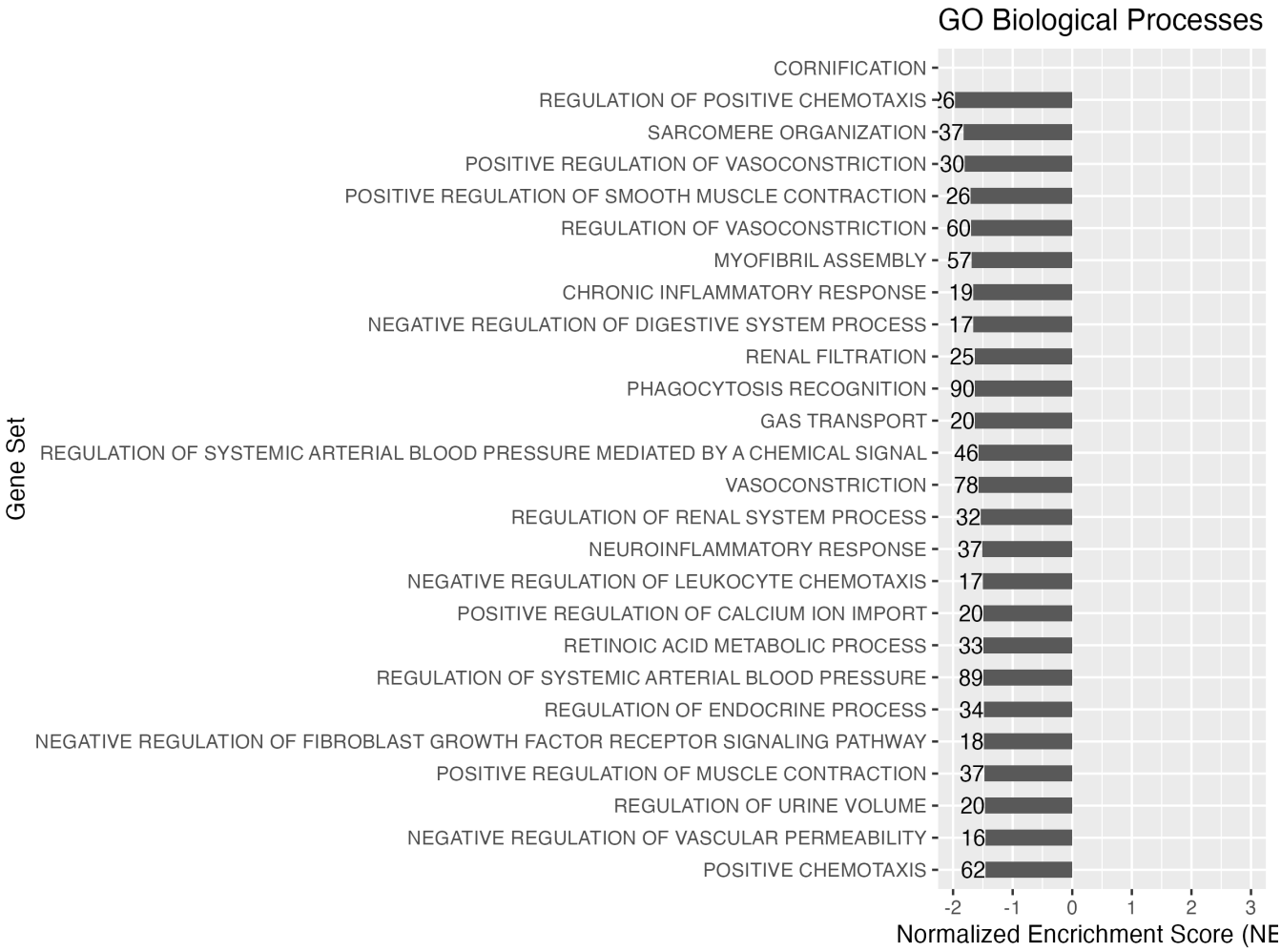


Figure 13: Gene sets significantly enriched in the comparison of Vadadustat vs. Hypoxia using the gene ontology biological processes.

References

[1] A. Musiał-Wysocka, M. Kot, and M. Majka, "The pros and cons of mesenchymal stem cell-based therapies," *Cell Transplant.*, vol. 28, no. 7, pp. 801–812, Jul. 2019.

[2] Z. Zhilai, M. Biling, Q. Sujun, D. Chao, S. Benchao, H. Shuai, Y. Shun, and Z. Hui, "Preconditioning in lowered oxygen enhances the therapeutic potential of human umbilical mesenchymal stem cells in a rat model of spinal cord injury," *Brain Res.*, vol. 1642, pp. 426–435, Jul. 2016.

[3] K. Zielniok, A. Burdzinska, V. Murcia Pienkowski, A. Koppolu, M. Rydzanicz, R. Zagozdzon, and L. Paczek, "Gene expression profile of human mesenchymal stromal cells exposed to hypoxic and pseudohypoxic preconditioning-an analysis by RNA sequencing," *Int. J. Mol. Sci.*, vol. 22, no. 15, p. 8160, Jul. 2021.

[4] Michael I. Love, Charlotte Soneson, Mark D. Robinson, "Importing transcript abundance with tximport," 2024, [Online; accessed 22-May-2024]. [Online]. Available: https://bioconductor.org/packages/devel/bioc/vignettes/tximport/inst/doc/tximport.html

[5] M. I. Love, W. Huber, and S. Anders, "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2," *Genome Biol.*, vol. 15, no. 12, p. 550, 2014.

[6] "Human, release 46 (grch37)," [Online; accessed 22-May-2024]. [Online]. Available: https://www.gencodegenes.org/human/release_46lift37.html

[7] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov, "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 102, no. 43, pp. 15 545–15 550, Oct. 2005.