Project 2 RNAseq Report: The Role of TYK2 in Pancreatic Beta-Cell Development

Addison Yam

October 24, 2024

1 Introduction

Type 1 Diabetes (T1D) is an autoimmune disorder characterized by the destruction of insulin-producing beta-cells in the pancreas. Genetic studies have frequently implicated variations in the Tyrosine Kinase 2 (*TYK2*) gene, which plays a critical role in mediating inflammatory signals, such as interferon-alpha (IFN-α). However, the precise mechanism by which TYK2 influences beta-cell development and response to stress remained unclear. To address this, Chandra et al. (2022) conducted a study to delineate the specific functions of TYK2 in two key areas: its role in guiding beta-cells through proper developmental stages and its modulation of their response to the damaging IFN-α signal [1]. The authors employed comprehensive RNA sequencing (RNA-seq) to capture genome-wide expression changes. This was followed by bioinformatic analyses, including differential expression to identify genes that were upregulated or downregulated, and pathway analysis (e.g., Reactome) to group these genes into broader biological processes. These techniques revealed that TYK2 regulates a network of processes vital for beta-cell integrity, including extracellular matrix organization and signaling by receptor tyrosine kinases.

2 Methods

RNA-seq data was analyzed from six samples, comprising three biological replicates each of wild-type (WT) control and TYK2 knockout (KO) experimental conditions at the S5 (endocrine progenitor) stage. The analysis pipeline was executed using R (v4.4.3) and Bioconductor packages.

Sequencing read quality was assessed using FastQC (v0.12.1) [2] with default parameters, and results were aggregated using MultiQC (v1.25) [3] with default parameters. Paired-end reads were aligned to the human reference genome (GRCh38) using STAR (v2.7.11b) [4] with default parameters, producing BAM alignment files for each sample. Gene-level counts were quantified from these aligned reads using VERSE (v0.1.5) [5] with default parameters, counting reads mapping to exons. Ensembl gene IDs were converted to gene symbols using a custom Python script with the pandas library (v2.3.3) [6], and individual count files were merged into a single gene-by-sample matrix.

The raw count matrix was filtered to remove lowly expressed genes using the Counts-Per-Million (CPM) method, retaining genes with a CPM ≥ 1 in at least three samples. This reduced the dataset from 63,241 to 13,197 genes, focusing subsequent analysis on actively transcribed genes. Normalization and differential expression analysis were performed using the `DESeq2` package (v1.50.0) [7] with default parameters. The experimental design was specified as $\sim$ condition. Significant differentially expressed genes (DEGs) were defined using an adjusted p-value (padj) cutoff of < 0.05. For functional characterization, these significant genes were further subset into upregulated and downregulated lists using an absolute log2 fold-change threshold of ≥ 1.

Pathway enrichment analysis was conducted using two complementary approaches. First, the list of significant DEGs was submitted to the Reactome database via Enrichr [8]. Second, Gene Set Enrichment Analysis (GSEA) was performed using the `fgsea` package (v1.32.4) [9] with default parameters. For GSEA, all genes were ranked by their log2 fold change, and the analysis was run against the C2 canonical pathways gene set from MSigDB (v2025.1) [10]. A padj cutoff of < 0.05 was used to identify significantly enriched pathways. All visualizations, including principal component analysis (PCA) plots, heatmaps, and volcano plots, were generated using `ggplot2` (v3.5.1) [11] and `pheatmap` [12].

3 Deliverables

3.1 Quality Control Evaluation

The initial quality assessment, aggregated by MultiQC, indicated that the sequencing data was of high quality and suitable for downstream analysis. The total number of reads per sample ranged from 84.5 million to 118.9 million. While the "Per Base Sequence Content" metric flagged warnings or failures for all samples—potentially indicating a random hexamer bias during library preparation that could be remedied by trimming—all other metrics were satisfactory. There was no significant adapter contamination, and overrepresented sequences were not alarming. Most importantly, the alignment rates were consistently high, with all samples achieving near or above 92% uniquely mapped reads, indicating a strong match to the reference genome.

3.2 Filtering the Counts Matrix

To focus the analysis on biologically relevant genes, the counts matrix was filtered using a Counts-Per-Million (CPM) threshold. Genes were retained if they had a CPM ≥ 1 in at least three samples. This filtering step removed 50,044 genes, resulting in a final set of 13,197 genes, which represents 20.87% of the original total. The distribution of log2(CPM+1) values before and after filtering demonstrates that the majority of genes had very low or zero expression across samples. The filtering process effectively removed this background noise, enriching for genes that are statistically more likely to be actively transcribed.

3.3 Differential Expression Analysis

Differential expression analysis with DESeq2 identified 1,116 genes as significantly differentially expressed (padj < 0.05) between TYK2 KO and WT samples. When an additional fold-change threshold of $|\log2FC| \geq 1$ was applied to determine biological relevance, 48 genes were classified as upregulated and 100 as downregulated in the KO condition. The top 10 most significantly differentially expressed genes, ranked by adjusted p-value, are presented in Table 1.

Table 1: Top 10 Differentially Expressed Genes (TYK2 KO vs. WT)

Gene ID	Gene Name	baseMean	log2FC	stat	pvalue	padj
ENSG00000129824	RPS4Y1	10735.37	-8.75	-65.44	0.000	0.000
ENSG00000250616	YPEL3-DT	549.24	-2.40	-22.47	7.47e-112	4.93e-108
ENSG00000251129	LINC02506	1717.14	-1.64	-20.51	1.60e-93	7.02e-90
ENSG00000111860	CEP85L	2092.55	-1.30	-18.73	2.80e-78	9.25e-75
ENSG00000287596	ENSG00000287596	6998.90	-0.71	-15.33	4.88e-53	1.29e-49
ENSG00000173262	SLC2A14	640.61	4.61	15.03	4.99e-51	1.10e-47
ENSG00000108439	PNPO	128.88	-4.25	-14.79	1.66e-49	3.13e-46
ENSG00000291093	SVIL-AS1	178.43	6.14	14.66	1.08e-48	1.79e-45
ENSG00000073803	MAP3K13	1474.66	1.00	14.16	1.71e-45	2.50e-42
ENSG00000273748	ENSG00000273748	187.59	-6.97	-13.79	2.79e-43	3.68e-40

The results were visualized using a volcano plot (Figure 4), which highlights the distribution of gene expression changes. While the majority of significant genes exhibited modest fold changes, a subset, including notable genes like *KRAS*, *LAMB2*, and *SPP1*, showed strong regulation, suggesting they are key targets for further functional investigation of TYK2's role.

3.3.1 Pathway Enrichment Analysis via Enrichr (Reactome)

To understand the biological processes disrupted by TYK2 knockout, the list of significant DEGs was analyzed using the Reactome database through Enrichr. This analysis revealed a coherent picture of TYK2's influence. The most significantly enriched pathways for the upregulated genes were heavily associated with structural organization and cellular signaling, including **Extracellular Matrix Organization**, **Collagen Degradation**, and **Signaling by Receptor Tyrosine Kinases**. This suggests that TYK2 deletion may lead to a remodeling of the cellular environment and dysregulated growth factor signaling. Conversely, downregulated genes were significantly enriched in pathways related to **Neuronal System** function and, crucially, **Regulation of Beta-Cell Development**. The enrichment of beta-cell development pathways, involving key transcription factors like *NEUROG3*, *NKX2-2*, and *INSM1*, directly aligns with the study's focus and indicates that TYK2 is essential for the transcriptional program driving beta-cell maturation.

3.3.2 Pathway Enrichment Analysis via GSEA

To complement the threshold-based Enrichr analysis, a pre-ranked GSEA was performed using the fgsea package. This method is sensitive to coordinated, subtle changes in gene expression across entire pathways. The top enriched pathways from the GSEA (Figure 1) strongly corroborated and expanded upon the Enrichr results. Pathways like **PID P53 DOWNSTREAM PATHWAY**, **REACTOME ASSEMBLY OF COLLAGEN FIBRILS**, and **PID AVB3 INTEGRIN PATHWAY** were positively enriched (NES > 0), indicating these processes are upregulated in the TYK2 KO. This reinforces the finding that TYK2 loss promotes ECM-related and pro-motility pathways. On the other hand, pathways such as **REACTOME NEURONAL SYSTEM** and **WP CELL LINEAGE MAP FOR NEURONAL DIFFERENTIATION** were negatively enriched (NES < 0), consistent with the downregulation of neuro-developmental processes observed in the Enrichr analysis, which overlaps significantly with beta-cell developmental pathways.

3.3.3 Comparison of Enrichment Methods

The results from the Enrichr (Reactome) and GSEA analyses show remarkable concordance, despite their different methodological approaches. Both methods identified **Extracellular Matrix Organization** and **Neuronal/Beta-Cell Development** as key biological themes affected by TYK2 deletion. The primary difference lies in GSEA's ability to identify additional, more specific signaling pathways (e.g., **P53 Downstream Pathway**, **Integrin pathways**) that were not the top hits in the Enrichr analysis. This synergy confirms the robustness of the conclusion that TYK2 is a critical regulator of both the structural microenvironment and the core transcriptional network essential for endocrine progenitor cell development.

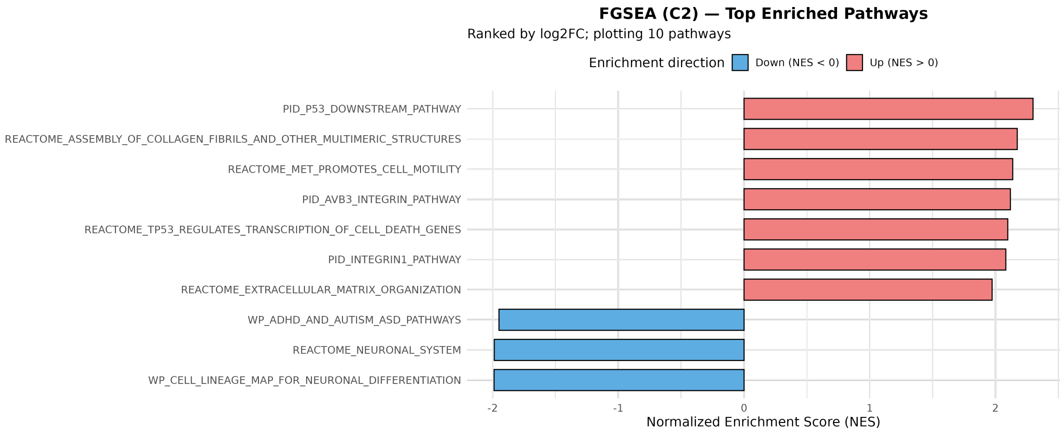

Figure 1: Bar plot of the top enriched pathways from the FGSEA analysis, ranked by Normalized Enrichment Score (NES). Pathways with NES > 0 (red) are upregulated in TYK2 KO, while those with NES < 0 (blue) are downregulated.

3.4 RNAseq Quality Control

To assess the overall quality of the experiment and the normalization procedure, variance-stabilizing transformation (VST) was applied to the filtered counts, followed by Principal Component Analysis (PCA) and sample-to-sample distance calculation. The PCA plot (Figure 2) showed a clear separation between WT and TYK2 KO samples along the first principal component (PC1), which accounted for 84% of the total variance. The tight clustering of biological replicates within each condition indicates high reproducibility and low technical noise. This demonstrates that the dominant source of variation in the dataset is the experimental condition (TYK2 KO), which is ideal for differential expression analysis. The sample-to-sample distance heatmap (Figure 3) further confirms this, showing that replicates cluster together and are distinctly separate from the other condition.

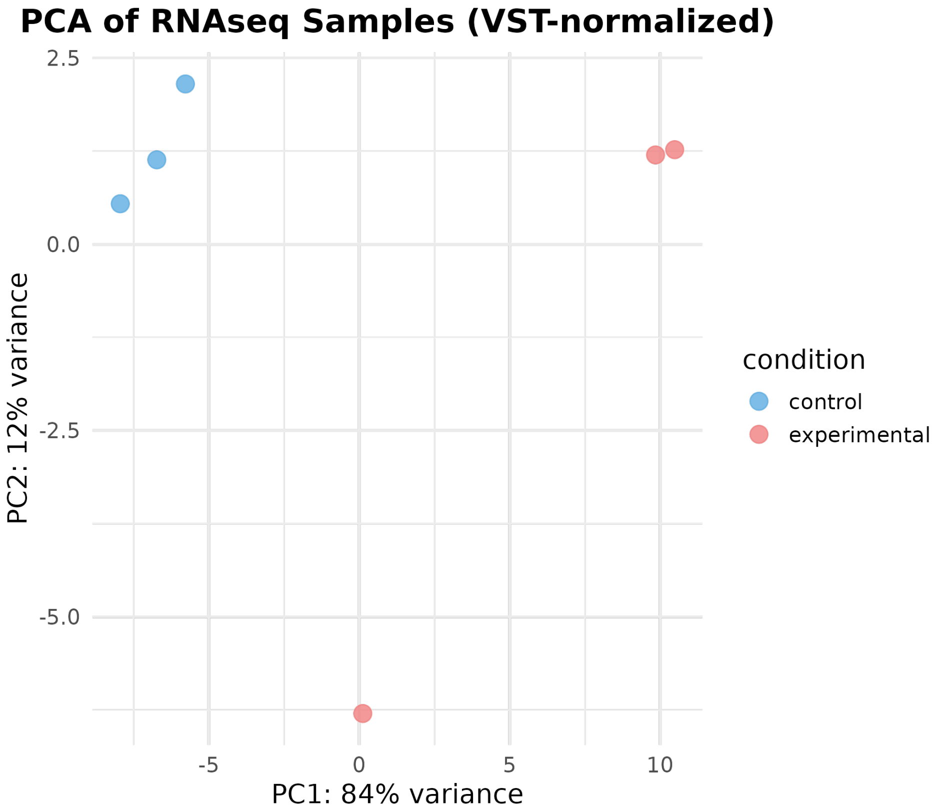

Figure 2: PCA of RNA-seq samples using VST-normalized counts. PC1 (84% variance) clearly separates the TYK2 KO (experimental) and WT (control) samples.

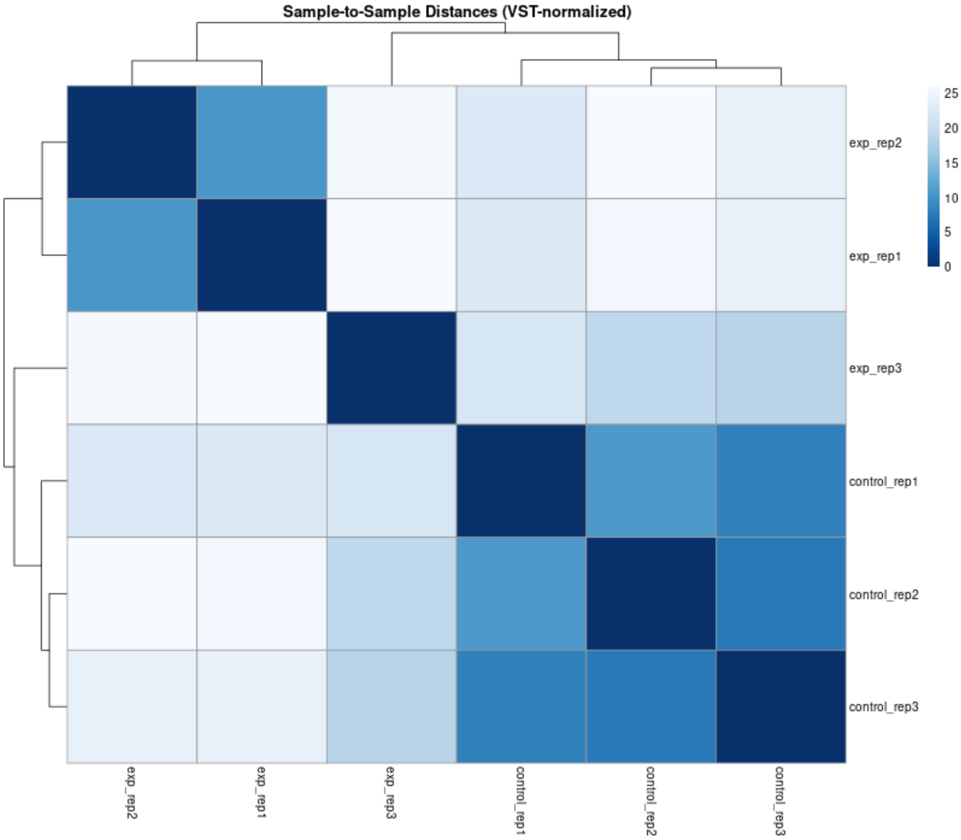

Figure 3: Heatmap of sample-to-sample Euclidean distances using VST-normalized counts. The block structure confirms that biological replicates are most similar to each other and that the two conditions are well separated.

3.5 Replication of Publication Figures

3.5.1 Replication of Figure 3C: Volcano Plot

A volcano plot was generated to replicate the style of Figure 3C from Chandra et al. (Figure 4). Our plot visualizes the 1,116 significant DEGs, with the 48 upregulated and 100 downregulated genes (using $|\log2FC| \geq 1$ and padj < 0.05) highlighted. Key genes such as *KRAS*, *LAMB2*, and *SPP1* are labeled. A direct comparison with the original publication reveals a notable discrepancy in the number of significant genes: we identified a total of 1,116 DEGs, while Chandra et al. reported 731 (319 upregulated, 412 downregulated). This difference is likely attributable to variations in the bioinformatic pipelines, such as the choice of read quantification tool (VERSE vs. the author's method), normalization techniques, or the specific statistical thresholds applied, underscoring how methodological choices can influence analytical outcomes.

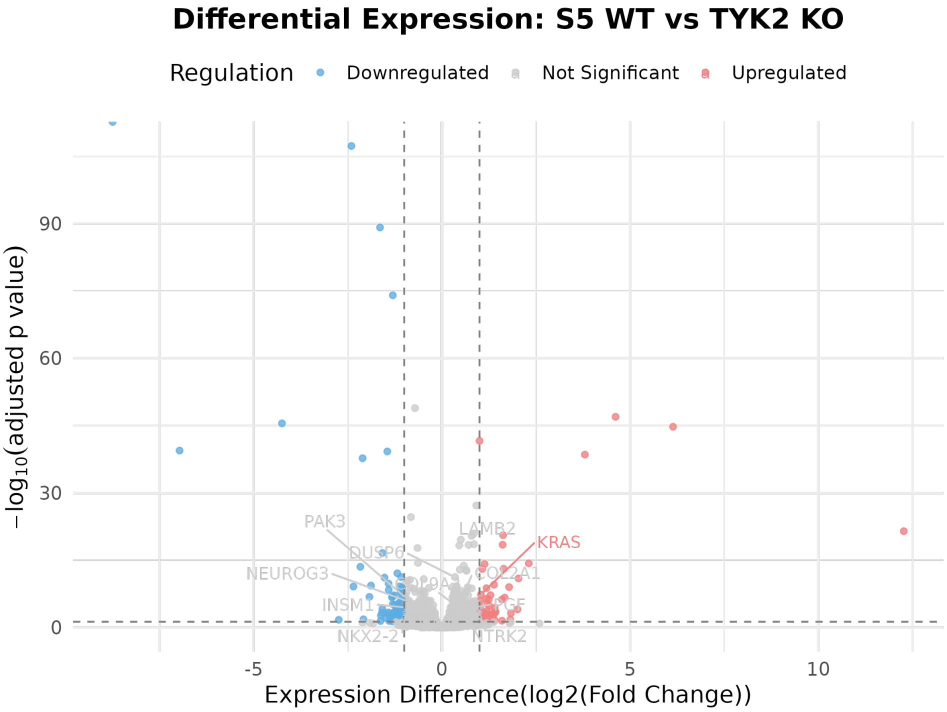

Figure 4: Volcano plot of differential expression in S5 TYK2 KO versus WT samples. Genes are colored by regulation status (red: upregulated, blue: downregulated, gray: not significant). Dashed lines indicate the significance (padj = 0.05) and fold-change (|log2FC| = 1) thresholds.

3.5.2 Replication of Figure 3F: Pathway Enrichment

A bar plot was generated to mimic the style of Figure 3F, displaying the top enriched Reactome pathways from our Enrichr analysis (Figure 5). The plot shows pathways ranked by their combined score and colored by their adjusted p-value, with a clear distinction between the biological themes of upregulated (e.g., ECM) and downregulated (e.g., Neuronal System, Beta-Cell Development) processes. While the exact pathways and their order differ from the original publication, the overarching biological conclusions are consistent. Both analyses implicate TYK2 in regulating processes critical for cell-matrix interactions and developmental fate decisions in pancreatic progenitors. The differences likely stem from the use of different pathway databases and enrichment tools, but they converge on a similar high-level biological narrative.

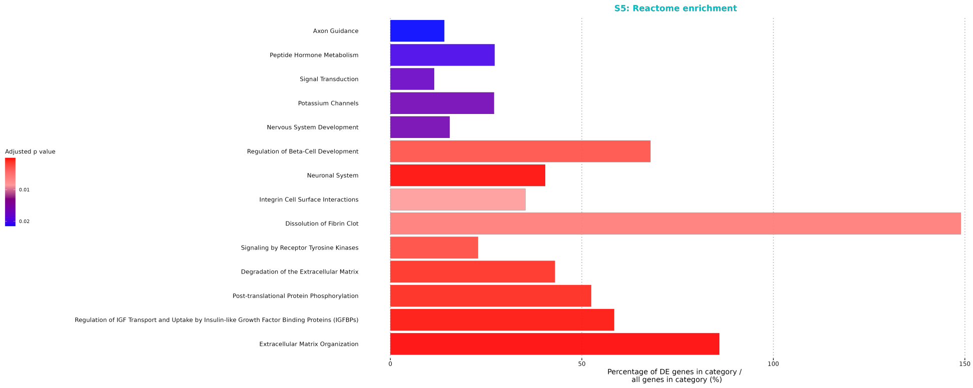

Figure 5: Bar plot of top enriched Reactome pathways from Enrichr analysis. The bar length represents the combined score, and the color indicates the adjusted p-value. The plot highlights the key biological processes affected by TYK2 deletion.

4 Future Directions

This analysis confirms the role of TYK2 at the S5 stage, but several questions remain that would be compelling to pursue:

1. **Time-Course Analysis:** How does the transcriptional disruption caused by TYK2 KO evolve across all developmental stages (hiPSC to immature islets)? A longitudinal analysis could pinpoint the exact stage when beta-cell development first goes awry.

2. **IFN- Challenge:** The original study also subjected cells to IFN-α. Analyzing the KO vs. WT response to this cytokine at the S5 stage would directly reveal how TYK2 deficiency impairs the beta-cell stress response, a key factor in T1D etiology.

3. **Validation and Mechanism:** The top differentially expressed genes, such as *KRAS*, should be validated using an independent method like qPCR. Furthermore, functional experiments (e.g., CRISPRa/i) to modulate these candidate genes could establish a direct causal link in the TYK2 regulatory network.

5 Conclusion

This project successfully replicated key aspects of the bioinformatic analysis from Chandra et al., investigating the role of TYK2 in pancreatic beta-cell development. The RNA-seq data was of high quality, and the differential expression analysis revealed a significant transcriptional impact upon TYK2 knockout. While the total number of significant genes differed from the original publication, the subsequent pathway enrichment analyses using both Enrichr and GSEA converged on a highly consistent biological story. The results robustly demonstrate that TYK2 is a critical regulator of pathways involved in extracellular matrix organization and, most importantly, the core transcriptional program governing beta-cell development. This provides a mechanistic foundation for understanding how genetic variation in *TYK2* can contribute to Type 1 Diabetes risk.

6 References

References

[1] Chandra, V., Ibrahim, H., Halliez, C. *et al.* The type 1 diabetes gene TYK2 regulates β-cell development and its responses to interferon-α. *Nat Commun* **13**, 6363 (2022). https://doi.org/10.1038/s41467-022-34069-z

[2] Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics, 32(19), 3047–3048. https://multiqc.info/

[4] Dobin, A., Davis, C. A., Schlesinger, F., *et al.* (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics, 29(1), 15–21. https://github.com/alexdobin/STAR

[5] https://github.com/verse/VERSE

[6] McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the 9th Python in Science Conference. https://pandas.pydata.org/

[7] Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://bioconductor.org/packages/DESeq2

[8] Kuleshov, M. V., Jones, M. R., *et al.* (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Research, 44(W1), W90–W97. https://maayanlab.cloud/Enrichr/

[9] Korotkevich, G., Sukhov, V., *et al.* (2021). Fast gene set enrichment analysis. bioRxiv. https://bioconductor.org/packages/fgsea

[10] Liberzon, A., Birger, C., *et al.* (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Systems, 1(6), 417–425. https://www.gsea-msigdb.org/gsea/msigdb

[11] Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. https://ggplot2.tidyverse.org

[12] Kolde, R. (2019). pheatmap: Pretty Heatmaps. https://cran.r-project.org/package=pheatmap