National and Kapodistrian University of Athens

Data Science & Information Technologies

Bioinformatics - Biomedical Data Science

Introduction to Bioinformatics
Final Project

Recreation of RNA-seq Analysis: "TAF10 Interacts with the GATA1 Transcription Factor and Controls Mouse Erythropoiesis" *by Papadopoulos P. et al.*

Student
Glykeria Spyrou

Professors
Dr. Martin Reczko
Dr. Alexandros Dimopoulos

Contents

List of Figures

List of Tables

Introduction

The present report is a recreation of the RNA-seq analysis that was part of the scientific paper with title: "TAF10 Interacts with the GATA1 Transcription Factor and Controls Mouse Erythropoiesis" by Petros Papadopoulos and his colleagues [19].

1.1 Background

In the given publication the researchers investigated how certain proteins complexes can contribute in the process of erythropoiesis. Overall, the TAF10 protein was suggested to interact with GATA1 towards and also play a vital role towards the control of red blood cell formation. It was also proposed that there is evident cross-talk between the transcription factor II D (TFIID) and the SAGA complex.

TFIID is a complex that is composed of the TATA box binding protein (TBP) and 13 other conserved TBP-associated factors, that serve as coactivators in the initiation of transcription by polymerase II [10]. TAF10 is an essential subunit for the activation of gene expression by the estrogen receptor, thus making it a critical component in the progression of the cell cycle and additional differentiation processes of the cell [20].

1.2 RNA Sequencing

RNA sequencing is a widely-used and powerful tool that exploits Next-Generation Sequencing and deep-sequencing technologies, aiming to reveal the transcriptome profile of a given sample of an organism [24]. In biomedical research RNA-seq is vastly used to quantify genes and also identify those transcribed in different quantities between two or more conditions, which can indicate regulatory malfunctions in the sampled tissues. Moreover, this technique aims to discover not only novel transcripts and their isoforms, but also detect allele-specific expression patterns [13].

RNA-sequencing requires multiple steps that combine both wet lab and bioinformatics techniques and methods. Briefly, those steps involve the extraction of RNA from a cell or tissue, quality control of the extracted RNA, library preparation, sequencing, and finally raw data analysis.

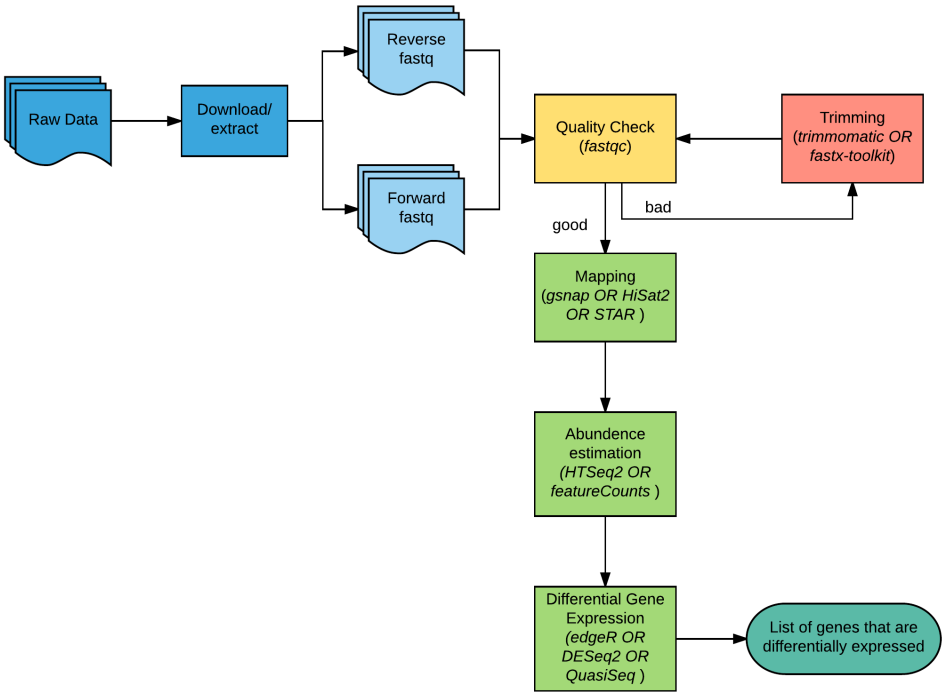

Figure 1.1: Overview of the bioinformatic analysis of RNA-seq data. Photo credit: Chudalayandi [6].

Methods

2.1 Data Description

In this experiment, tissue samples from the livers of two *TAF10KO*^{cEry} mice lacking the erythroid cell-specific *EpoR-Cre* allele, three heterozygous *TAF10KO*^{cEry} and two wild-type (WT) were sequenced 2.1.

FASTQ file	Sample Type
SRR1982462	TAF10KO
SRR1982463	TAF10HET
SRR1982464	WT
SRR1982465	TAF10HET
SRR1982466	TAF10HET
SRR1982467	WT
SRR1982468	TAF10KO

Table 2.1: Table of the FASTQ files and the corresponding sample type.

2.2 Software & Packages

In the present report, the most part of the analyses described were conducted at the cloud infrastructure of Hypatia [9] through command line interface (CLI) tools. For a seamless utilization of bioinformatics tools throughout this project, a Conda environment was configured, where tools were installed through the *bioconda* channel under *python3*. A small part was conducted locally through RStudio (2023.12.0+369 "Ocean Storm" Release) using R under the version 4.3.2. ("Eye Holes") [22]. Additional tools and packages used are listed in the table below 2.2.

Packages/Tools	Creators	Version
AnnotationDbi	Pagès et al. [18]	1.64.1
bioconda	Grüning et al. [8]	2.11.1
Bowtie	Langmead and Salzberg [14]	2.4.4.1
clusterProfiler	Wu et al. [27]	4.10.0
DESeq2	Love, Huber, and Anders [16]	1.42.0
dplyr	Wickham et al. [26]	1.1.4
FastQC	Andrews [3]	0.12.1
ggplot2	Wickham [25]	3.5.0
ggpubr	Kassambara [12]	0.6.0
HTSeq	Anders, Pyl, and Huber [2]	0.11.3
HTSlib	Bonfield et al. [4]	1.3.1
Miniconda	*Anaconda Software Distribution* [1]	23.11.0
minion	Dongen [7]	-
org.Mm.eg.db	Carlson [5]	3.18.0
samtools	Li et al. [15]	1.3.1
SRA toolkit	NCBI [17]	3.0.10
TopHat	Trapnell, Pachter, and Salzberg [23]	2.2.1

Table 2.2: List of packages used in the analysis, their current versions and the creators.

2.3 Data Acquisition

Raw data (fastq) from the experiment are openly distributed through the SRA portal of NCBI. Data can directly be downloaded in the VM through the `sra-toolkit` CLI tool by using the functions `prefetch` to download the SRA files and then `fastq-dump` to convert the SRA files into FASTQ, along with the flags `--gzip` for compression and `--split-3` for the creation of separate files of paired-end data.

2.4 Quality Control

The FASTQ file format is the one used for sequence reads coming from NGS technologies, where we can also find quality score for the nucleotide of each read. Quality control is a highly important task before any analysis, which can be conducted through *FastQC*. It performs several analyses on the given raw sequence data with some of them being basic statistics, per base quality score, per sequence GC content, adapter content, and many more. For the present analysis *FastQC* was used as a command line tool using the following command for all FASTQ files.

2.5 Adapter Scan

Adapter scanning in FASTQ files is a crucial step in the process of quality control. As it is known, adapters are short DNA sequences being used in the sequencing process to facilitate cDNA synthesis and amplification. From the previous step of quality control it was shown that there was no adapter presence in the FASTQ files. However, an additional tool that performs adapter scanning was used to

ensure their absence. For this specific task, we used *minion*. This tool uses De Bruijn graphs to assemble sequences from sequencing data, which returns the sequence density of the inferred adapter, a measure (fanout) of the number of distinct prefixes (length 3) observed in the candidate sequence, and some additional measures associated with density and fanout.

Regarding the mice raw RNA-seq data, several candidate sequences were returned of significantly indifferent lengths, with very low scores in terms of sequence density and fan-out score, in both criteria outpouts. The lengths of the sequences ranged from 13 to 511 bases, where sequence density didn't exceed the score of 1. Both suggested sequences per strand were inserted into BLAST to further investigate their origin, and evidently both were associated with hemoglobin proteins. Therefore, the FASTQ files were not imposed into adapter trimming.

2.6 Spliced Alignment

The next crucial part of this procedure is alignment of the paired-end data to the reference genome. In this step, *TopHat2* and *Bowtie2* were used to perform the alignment to the UCSC mm10 mouse reference genome (ftp link). After the completion of alignment for all samples, the output consists of several BAM files, and the one utilized is `accepted_hits.bam`. Finally, all BAM files were indexed.

2.7 Gene Quantification

In order to better understand the gene expression activity of each sequenced sample, it is required to perform gene quantification and estimate the abundance of each gene. For this step, *htseq-count* was used to perform the quantification by using the flags of `-f bam -r pos -m union -s no -a 20 -i gene_id`, in order to specify the format of the input data (`-f`), the criterion under which the BAM files were sorted (`-r`), the mode for overlapping in more than one features (`-m`), the minimum quality score for reads to be kept (`-a`), and the attribute from the GTF file to be used as feature id (`-i`). The annotation file exploited was the mm10, which included ENSEMBL gene and transcript identifiers (ftp link). After the end of the quantification procedure, .txt files were acquired for each sample containing the counts for each ENSEMBL gene identifier.

2.8 Differential Gene Expression Analysis

Differential gene expression analysis, i a crucial method that manages to identify significant changes in the expression of genes between two or more experimental conditions. However, before applying this method, we conducted a principal component analysis (PCA). PCA is a technique that simplifies high-dimensional data by transforming them into a lower-dimensional space in order to explore the variation of the expression data [11]. Visualizing transformed data through PCA is also useful for the exploration of similarities in data points (in our case, samples) and the identification of clusters.

For this purpose, we used the *DESeq2* package in R, through which both PCA and DEA were conducted [16]. First of all, all expression data were store properly into a `DESeqDataSet` data structure. They underwent factor-size normalization, low-count gene filtering, differential expression analysis and variance stabilizing transformation.

Regarding differential expression analysis, *DESeq2*, uses a linear negative binomial model for the calculation of the differences in the expression data per condition. Prior to normalization, the three conditions that describe the samples were factorized, in order to perform pairwise comparisons. For the determination of the top differentially expressed genes a p-value threshold equal to 0.05 was used, while logFC wasn't taken into account for this task in order to avoid excluding genes of high biological importance, which can have a great impact on the phenotype of a condition even with minimal changes.

2.9 Integrative Genomics Viewer (IGV)

The Integrative Genomics Viewer (IGV) software [21] was used to visualize the top differentially expressed genes among the given conditions of the biological samples. For the comprehensive visualization of the significant genes, all sample BAM files were loaded, the mm10 genome was selected withing the software and the annotation file (.gtf) file that was used in gene quantification step was also added. Prior to loading the annotation file it was necessary to sort and index it.

2.10 GOterm Enrichment Analysis

Gene Ontology (GO) term enrichment analysis describes a statistical method used in order to identify overrepresented biological mechanisms by GO terms, within a set of genes that interest the researchers. Through this technique we have the ability to identify specific gene lists that by highlighting their contribution in several functional categories. Those categories are biological processes (BP), molecular functions (MF), and cellular components (CC). Additionally, we can track underlying molecular pathways critical to the conditions we are studying.

For this method, the R package *clusterProfiler* was exploited [27], for which all statistically significant genes (p-value<0.05) were used for enrichment through the `enrichGO()` function. It was applied for each set of significant genes coming from the pairwise comparisons that were conducted during DEA, in terms of all three functional categories.

Results

3.1 Quality Control

The output reports of the FastQC are depicted in the table 3.1, where the summary, basic statistics and the per base sequence quality are shown. In both strands of each sample, overall quality doesn't drop below the "green"/acceptable quality scores region, while there was no adapter content found.

3.2 PCA

The initial dataset of 48440 transcripts underwent filtering and normalization by size factors. The amount of transcripts in the dataset were 21646 after filtering, for which differential expression analysis was performed. Then, the dataset underwent blind variance stabilization transformation, in order to perform principal component analysis. In the following figure, the PCA plot of the dataset is depicted (Figure 3.1).

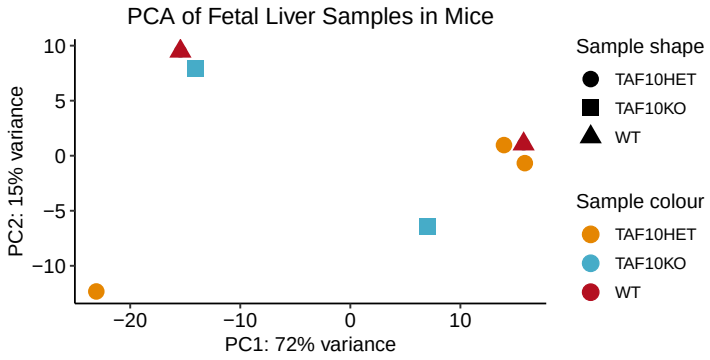

Figure 3.1: Principal Component Analysis plot. The horizontal and the vertical axis correspond to the first and the second principal components accordingly, and data points represent the samples.

The first principal component accounts for the 72% of the variance, while the second principal

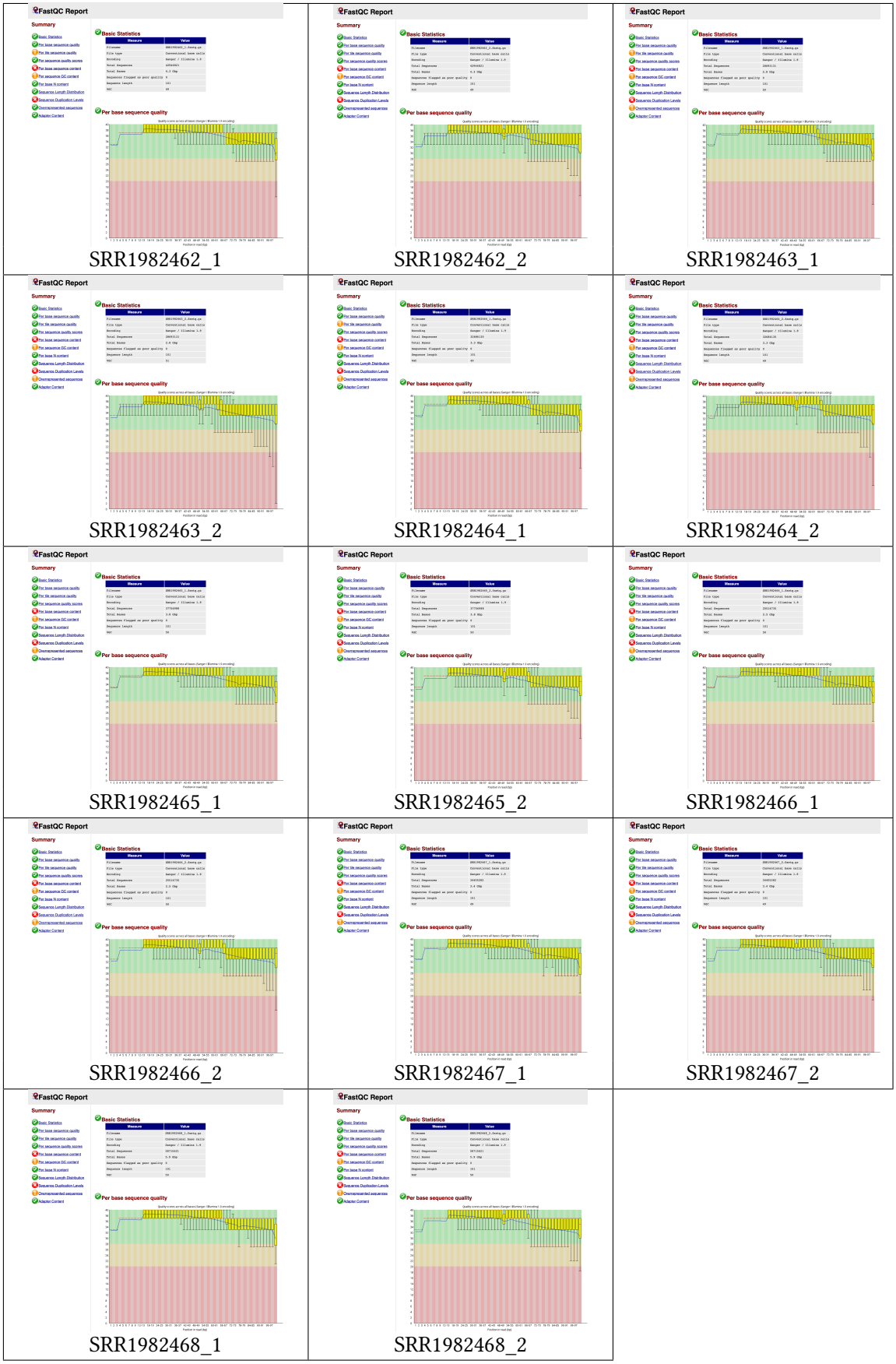

Table 3.1: FastQC reports of all samples, where "_1" and "_2" indicate forward and reverse strands accordingly. The show plots refer to the per base sequence quality.

component accounts for the 15% of the variance in the expression data. As shown on the plot, and in contrast to the original paper, no definitive clusters of the sample groups are depicted. More specifically in the left part of the plot, it appears as if TAF10KO and TAF10HET have switched places, thus making impossible to cluster TAF10HET with WT in order to perform the differential expression analysis between TAF10HET/WT and TAF10KO. As a result, in the present analysis, by using the `contrast` argument in the `results()` offered by the *DESeq2* package, we managed to perform pairwise comparisons of the three different conditions based on their sample type.

3.3 Differential Gene Expression Analysis

3.3.1 TAF10KO vs TAF10HET

Between the conditions TAF10KO and TAF10HET, out of the 21646 transcripts with nonzero total count and an adjusted p-value less than 0.1, the total number or upregulated genes was 87 (0.4%), the number of downregulated genes was 77 (0.36%), the number of outliers was 13 (0.06%), and the total number of low counts was 5876 (27%).

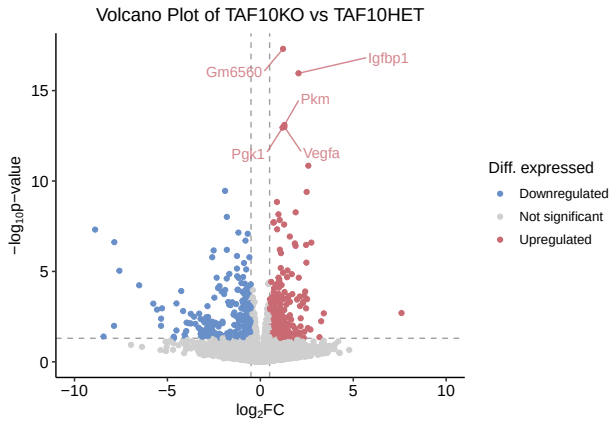

Figure 3.2: Volcano plot of the DEA results between TAF10KO and TAF10HET. Dashed vertical lines at logFC $= -1$ and logFC $= 1$ delineate the threshold for gene regulation. A horizontal dashed line at $-\log_2(0.05)$ highlights the significance threshold. Non-significant transcripts are shown in gray, while those with a p-value < 0.05 and logFC > 0.05 are colored in pink, indicating upregulation, and genes with p-value < 0.05 and logFC < -0.05 are colored in sky blue, indicating downregulation. Labels show the top 5 differentially expressed genes.

3.3.2 WT vs TAF10KO

Between the conditions WT and TAF10KO, out of the 21646 transcripts with nonzero total count and an adjusted p-value less than 0.1, the total number or upregulated genes was 36 (0.17%), the number of downregulated genes was 61 (0.28%), the number of outliers was 13 (0.06%), and the total number of low counts was 5876 (27%).

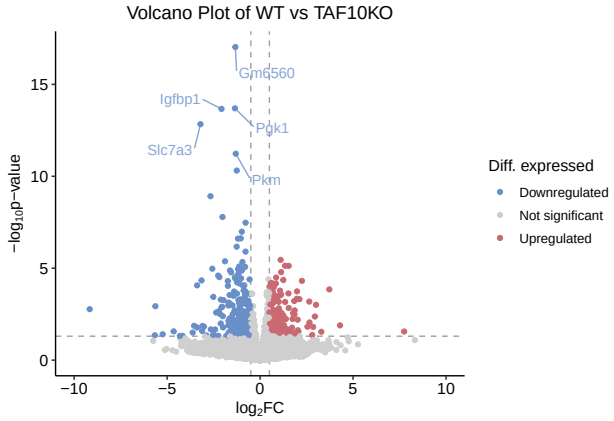

Figure 3.3: Volcano plot of the DEA results between WT and TAF10KO. Dashed vertical lines at logFC $= -1$ and logFC $= 1$ delineate the threshold for gene regulation. A horizontal dashed line at $-\log_2(0.05)$ highlights the significance threshold. Non-significant transcripts are shown in gray, while those with a p-value < 0.05 and logFC > 0.05 are colored in pink, indicating upregulation, and genes with p-value < 0.05 and logFC < -0.05 are colored in sky blue, indicating downregulation. Labels show the top 5 differentially expressed genes.

3.3.3 WT vs TAF10HET

Between the conditions WT and TAF10KO, out of the 21646 transcripts with nonzero total count and an adjusted p-value less than 0.1, no genes were found unregulated, 5 (0.023%) were considered downregulated, the number of outliers was 13 (0.06%), and the total number of low counts was 0.

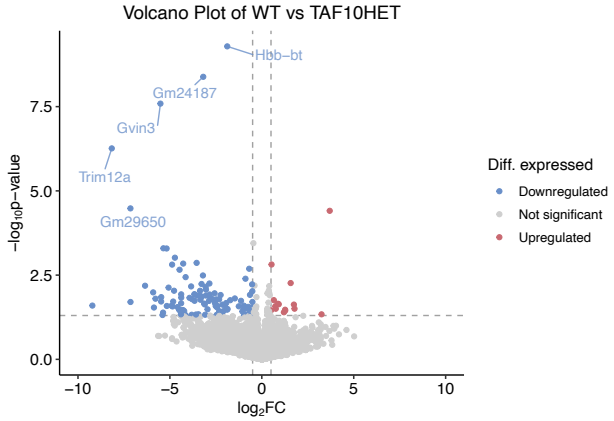

Figure 3.4: Volcano plot of the DEA results between WT and TAF10HET. Dashed vertical lines at logFC $= -1$ and logFC $= 1$ delineate the threshold for gene regulation. A horizontal dashed line at $-\log_2(0.05)$ highlights the significance threshold. Non-significant transcripts are shown in gray, while those with a p-value < 0.05 and logFC > 0.05 are colored in pink, indicating upregulation, and genes with p-value < 0.05 and logFC < -0.05 are colored in sky blue, indicating downregulation. Labels show the top 5 differentially expressed genes.

Therefore, since the comparison between WT and TAF10HET brought about so low metrics, with the majority of the transcripts determined as non-significant, an additional DEA was conducted by considering the control group to be the union of TAF10HET and WT.

3.3.4 WT/TOAF10HET vs TAF10KO

Between the conditions CONTROL and TAF10KO, out of the 21646 transcripts with nonzero total count and an adjusted p-value less than 0.1, the total number or upregulated genes was 216 (1%), the number of downregulated genes was 179 (0.83%), the number of outliers was 61 (0.28%), and the total number of low counts was 5869 (27%).

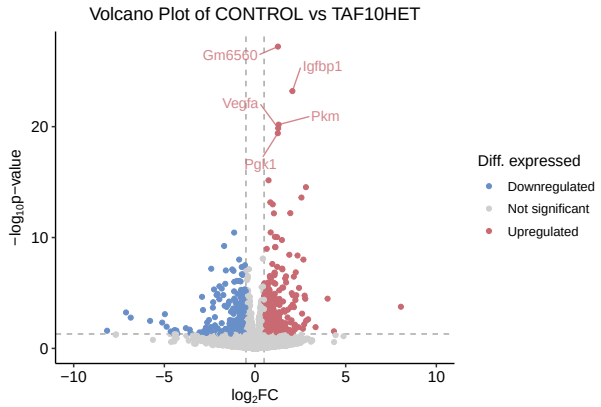

Figure 3.5: Volcano plot of the DEA results between WT/TAF10HET and TAF10KO. Dashed vertical lines at logFC $= -1$ and logFC $= 1$ delineate the threshold for gene regulation. A horizontal dashed line at $-\log_2(0.05)$ highlights the significance threshold. Non-significant transcripts are shown in gray, while those with a p-value < 0.05 and logFC > 0.05 are colored in pink, indicating upregulation, and genes with p-value < 0.05 and logFC < -0.05 are colored in sky blue, indicating downregulation. Labels show the top 5 differentially expressed genes.

3.4 IGV: Top 5 DEGs

From the final DEA analysis that was performed for WT/TAF10HET (control) against TAF10KO, the top 5 differentially expressed genes were *Gm6560*, *Igfbp1*, *Pkm*, *Vegfa*, and *Pgk1*. For those genes, IGV was used to search the genomic regions of those genes and further observe them. In the figures that follow (3.6 3.7, 3.8, 3.9, 3.10) below, the top 5 differentially expressed genes between WT/TAF10HET and TAF10KO are shown through the IGV interface.

3.5 GOterm Enrichment Analysis

In the table of figures (Table: 3.2), are three barplots containing the top 20 categories for each of the different components that GOterm enrichment analysis was conducted. The first is Biological Processes (BP), the second is Cellular Components (CC), and the third is Molecular Functions (MF). In each barplot we can also observe based on the colour gradient the p-value of each category, for which the amount of genes were found to be overrepresented in.

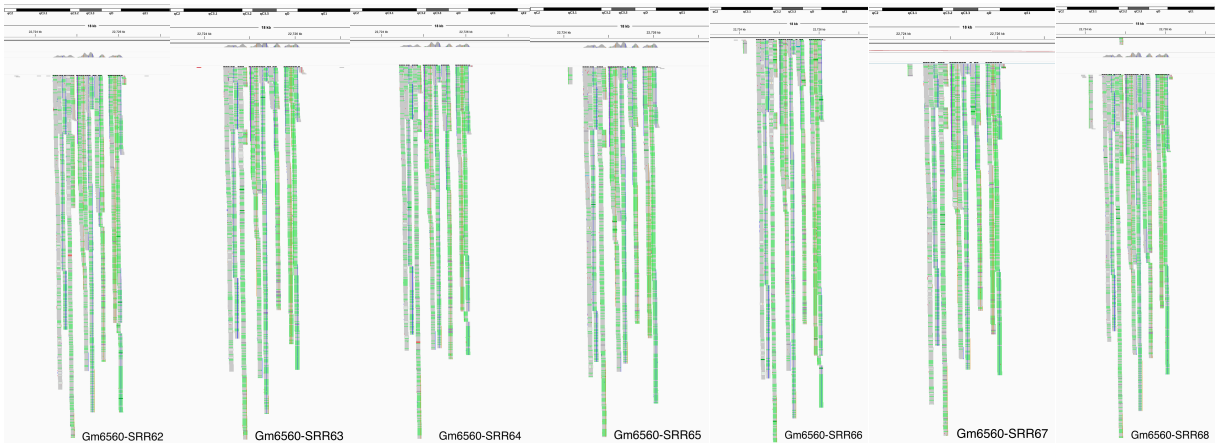

Figure 3.6: IGV: *Gm6560*

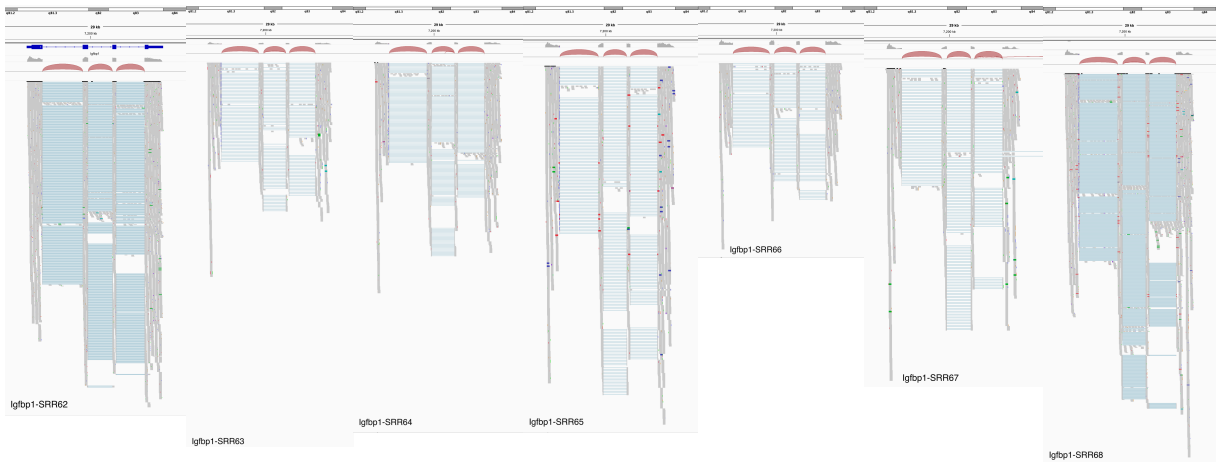

Figure 3.7: IGV: *Igfbp1*

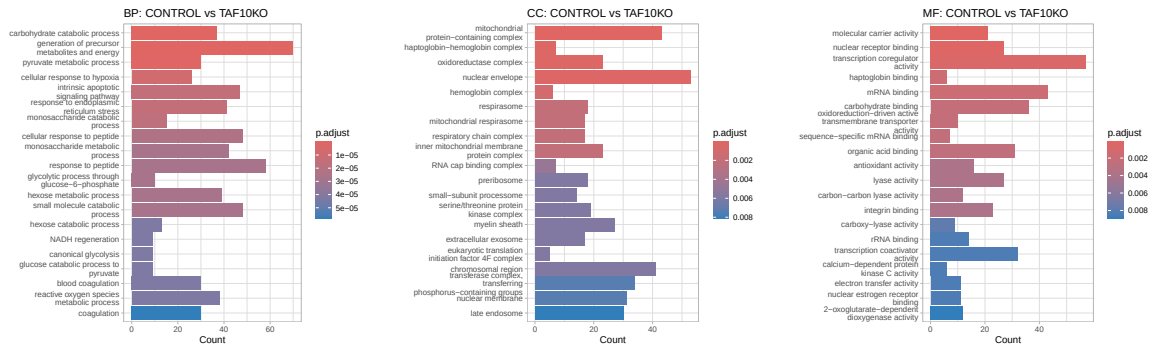

Table 3.2: Top 20 categories for Biological Processes (BP), Cellular Components (CC), and Molecular Functions (MF) gene set enrichment analysis.

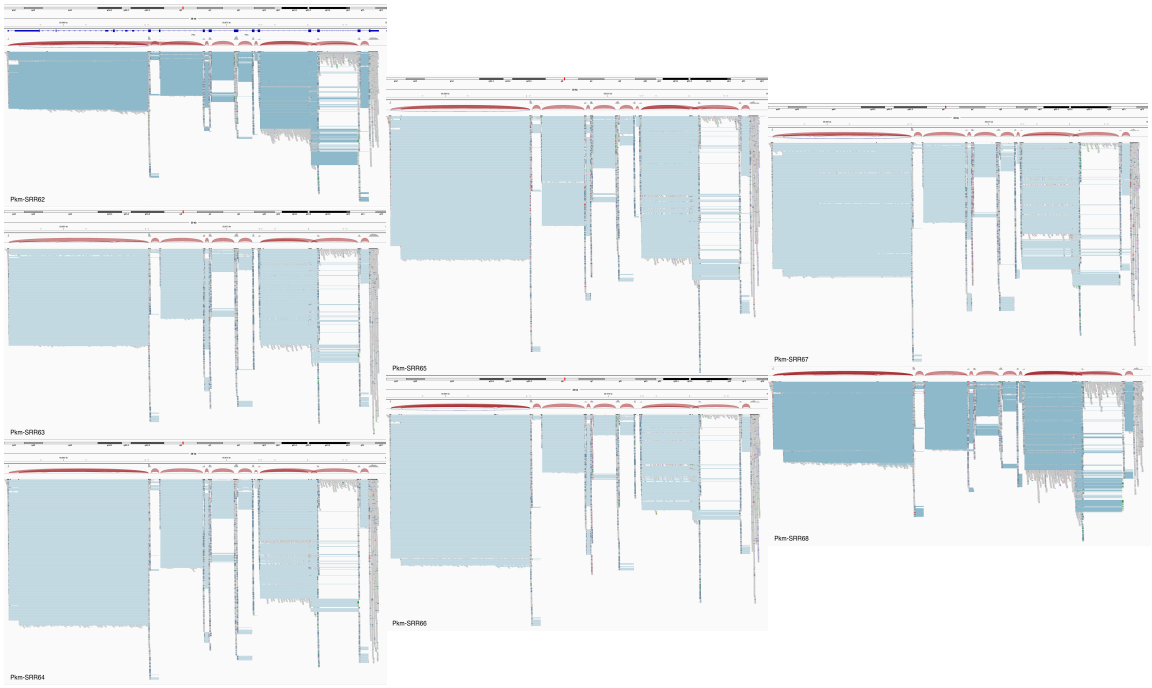

Figure 3.8: IGV: *Pkm*

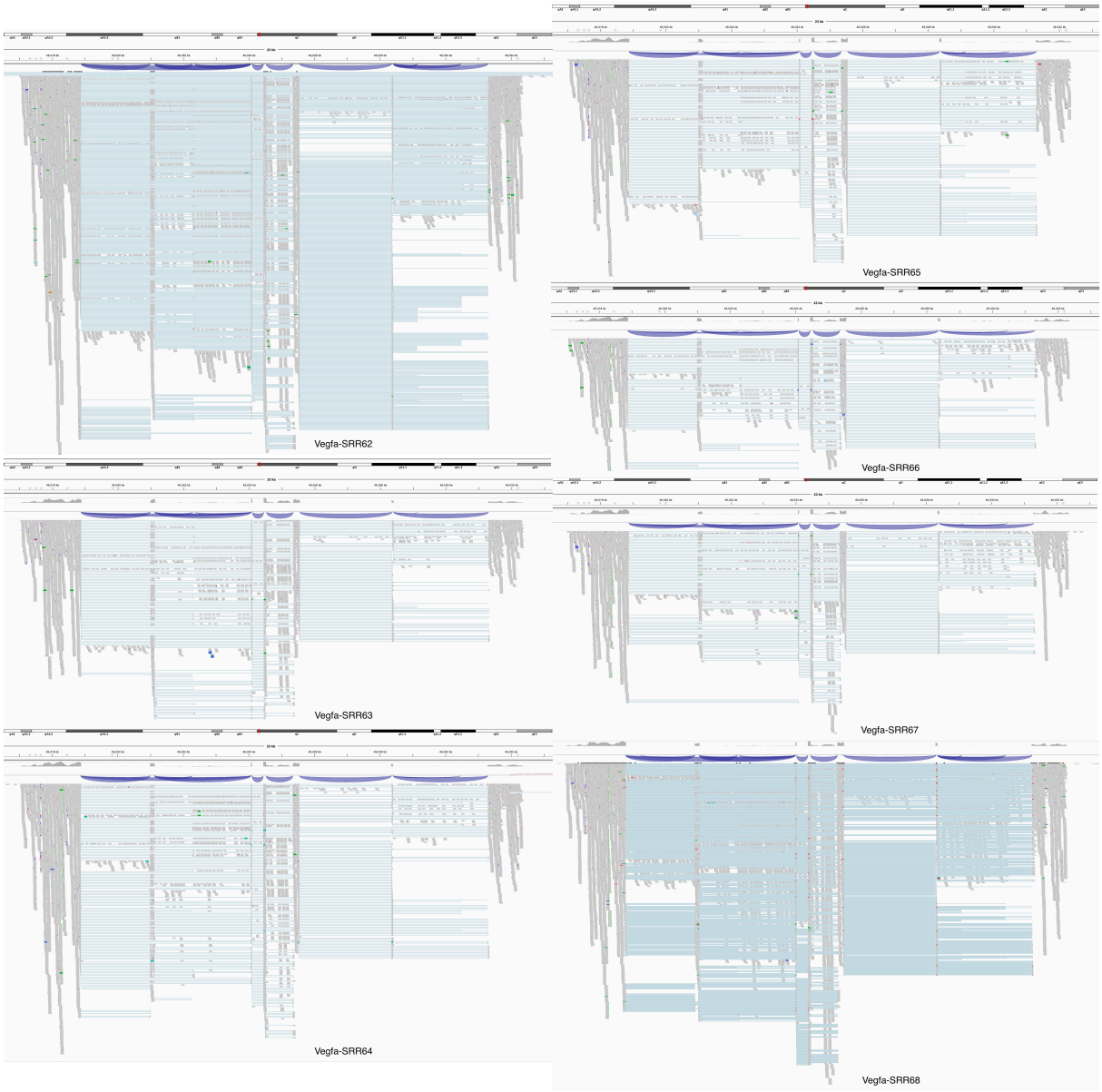

Figure 3.9: IGV: *Vegfa*

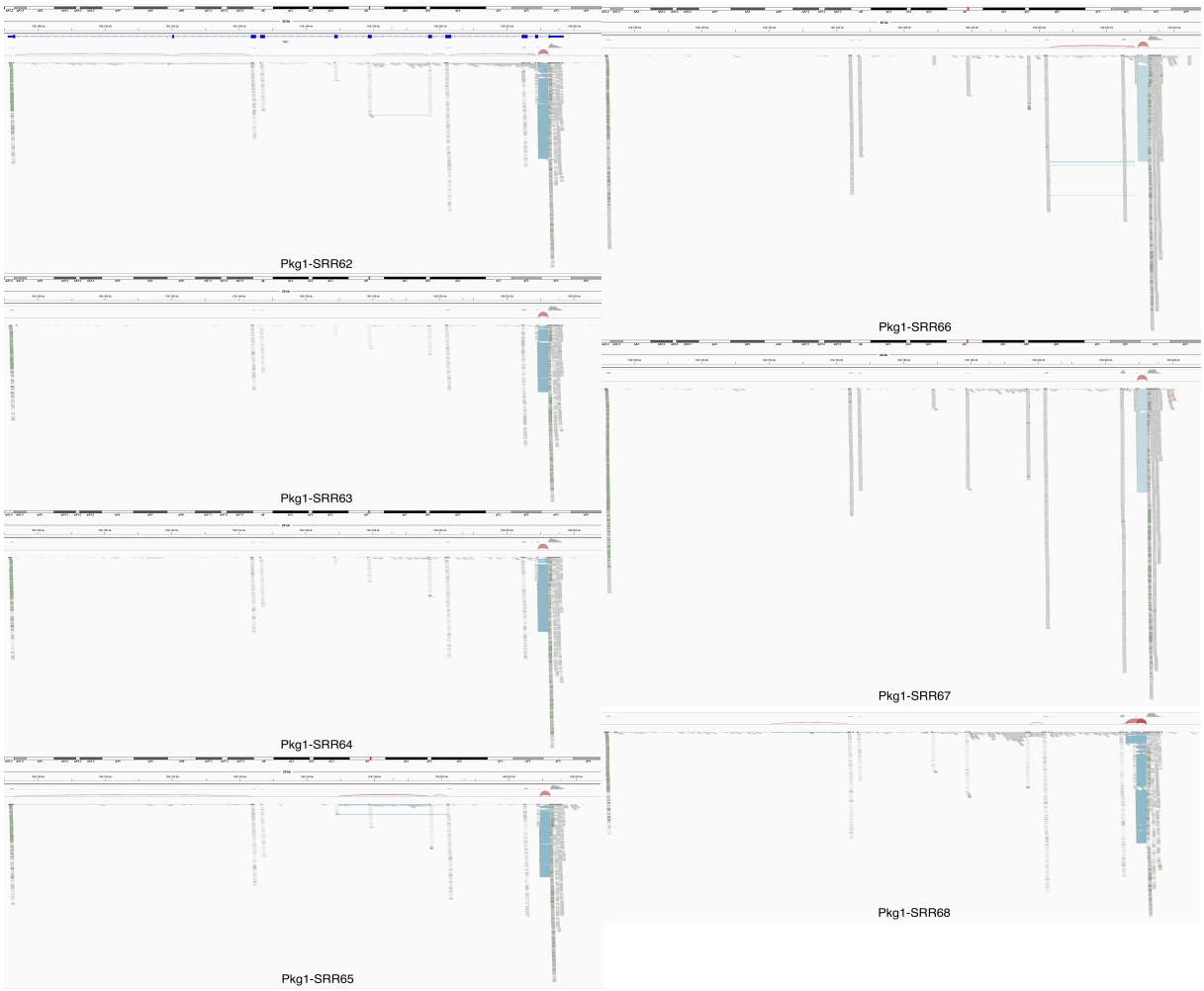

Figure 3.10: IGV: *Pkg1*

Bibliography

[1] *Anaconda Software Distribution*. Version 3-23.11. Anaconda Inc., 2024. URL: https://docs.anaconda.com/.

[2] Simon Anders, Paul Theodor Pyl, and Wolfgang Huber. "HTSeq—a Python Framework to Work with High-Throughput Sequencing Data". In: *Bioinformatics* 31.2 (2015), pp. 166–169. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btu638.

[3] Simons Andrews. *FastQC: A Quality Control Tool for High Throughput Sequence Data*. 2010. URL: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

[4] James K. Bonfield et al. "HTSlib: C Library for Reading/Writing High-Throughput Sequencing Data". In: *Gigascience* 10.2 (2021), giab007. ISSN: 2047-217X. DOI: 10.1093/gigascience/giab007. pmid: 33594436.

[5] Marc Carlson. *Org. Mm. Eg. Db: Genome Wide Annotation for Mouse. R Package Version 3.8.2*. 2019.

[6] Siva Chudalayandi. *RNA Sequence Analysis*. Bioinformatics Workbook. 2023. URL: https://bioinformaticsworkbook.org/dataAnalysis/RNA-Seq/RNA-SeqIntro/RNAseq-using-a-genome.html.

[7] Stijn van Dongen. *Minion*. 2015. URL: https://gensoft.pasteur.fr/docs/reaper/15-065/minion.html.

[8] Björn Grüning et al. "Bioconda: Sustainable and Comprehensive Software Distribution for the Life Sciences". In: *Nat Methods* 15.7 (2018), pp. 475–476. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0046-7.

[9] *HYPATIA*. URL: https://hypatia.athenarc.gr/.

[10] Arup Kumar Indra et al. "TAF10 Is Required for the Establishment of Skin Barrier Function in Foetal, but Not in Adult Mouse Epidermis". In: *Developmental Biology* 285.1 (2005), pp. 28–37. ISSN: 0012-1606. DOI: 10.1016/j.ydbio.2005.05.043.

[11] "Principal Component Analysis for Special Types of Data". In: *Principal Component Analysis*. Ed. by I. T. Jolliffe. Springer Series in Statistics. New York, NY: Springer, 2002, pp. 338–372. ISBN: 978-0-387-22440-4. DOI: 10.1007/0-387-22440-8_13.

[12] Alboukadel Kassambara. *Ggpubr: 'ggplot2' Based Publication Ready Plots*. Version 0.6.0. 2023. URL: https://cran.r-project.org/web/packages/ggpubr/index.html.

[13] Kimberly R. Kukurba and Stephen B. Montgomery. "RNA Sequencing and Analysis". In: *Cold Spring Harb Protoc* 2015.11 (2015), pdb.top084970. ISSN: 1940-3402, 1559-6095. DOI: 10.1101/pdb.top084970. pmid: 25870306.

[14] Ben Langmead and Steven L. Salzberg. "Fast Gapped-Read Alignment with Bowtie 2". In: *Nat Methods* 9.4 (2012), pp. 357–359. ISSN: 1548-7105. DOI: 10.1038/nmeth.1923.

[15] Heng Li et al. "The Sequence Alignment/Map Format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp352. pmid: 19505943.

[16] Michael I. Love, Wolfgang Huber, and Simon Anders. "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2". In: *Genome Biology* 15.12 (2014), p. 550. ISSN: 1474-760X. DOI: 10.1186/s13059-014-0550-8.

[17] NCBI. *Ncbi/Sra-Tools: SRA Tools*. 2023. URL: https://github.com/ncbi/sra-tools.

[18] Hervé Pagès et al. *AnnotationDbi: Manipulation of SQLite-based Annotations in Bioconductor*. Version 1.64.1. Bioconductor version: Release (3.18), 2023. DOI: 10.18129/B9.bioc.AnnotationDbi.

[19] Petros Papadopoulos et al. "TAF10 Interacts with the GATA1 Transcription Factor and Controls Mouse Erythropoiesis". In: *Mol Cell Biol* 35.12 (2015), pp. 2103–2118. ISSN: 1098-5549. DOI: 10.1128/MCB.01370-14. pmid: 25870109.

[20] PubChem. *TAF10 - TATA-box Binding Protein Associated Factor 10 (Human)*. URL: https://pubchem.ncbi.nlm.nih.gov/gene/TAF10/human.

[21] James T. Robinson et al. "Integrative Genomics Viewer". In: *Nat Biotechnol* 29.1 (2011), pp. 24–26. ISSN: 1546-1696. DOI: 10.1038/nbt.1754.

[22] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2023.

[23] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. "TopHat: Discovering Splice Junctions with RNA-Seq". In: *Bioinformatics* 25.9 (2009), pp. 1105–1111. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btp120.

[24] Zhong Wang, Mark Gerstein, and Michael Snyder. "RNA-Seq: A Revolutionary Tool for Transcriptomics". In: *Nat Rev Genet* 10.1 (2009), pp. 57–63. ISSN: 1471-0064. DOI: 10.1038/nrg2484.

[25] Hadley Wickham. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org/.

[26] Hadley Wickham et al. *Dplyr: A Grammar of Data Manipulation*. Version 1.1.2. 2023. URL: https://cran.r-project.org/web/packages/dplyr/index.html.

[27] Tianzhi Wu et al. "clusterProfiler 4.0: A Universal Enrichment Tool for Interpreting Omics Data". In: *Innovation* 2.3 (2021). ISSN: 2666-6758. DOI: 10.1016/j.xinn.2021.100141. pmid: 34557778.