An analysis of knockout method’s impact on resulting differentially expressed genes (DEGs) from bulk gene knockout RNA-seq data

Conner Webber, James Simpson, Austin Maggert, Lindsay Ding

**Abstract:**

This report examines the impact of individually knocking out four genes—MXD1, RUNX1, NCOA3, and BHLHE40—on the expression of other genes by identifying differentially expressed genes (DEGs) through DESeq2 analysis. The study compares DEG profiles across three experimental conditions: KO vs. WT, CE vs. WT, and PTC vs. WT. DESeq2 is used to identify DEGs, and visualizations are generated to illustrate gene expression changes and highlight differences among the groups.

**Introduction:**

This study investigates differentially expressed genes (DEGs) by analyzing the effects of gene modification techniques on four specific genes: **MXD1**, **RUNX1**, **NCOA3**, and **BHLHE40**. The analysis is conducted using the Bulk RNA-seq Study 2 dataset from the MorPhiC project [9]. Our goal is to compare how samples that have undergone three different gene modification strategies—KO (gene deletion), CE (critical exon deletion), and PTC+1 (premature termination codon with frameshift)—affect gene expression levels in relation to the wild-type (WT) control sample.

Each knockout strategy disrupts gene function in a distinct way:

* **KO (Knockout):** Removes most or all protein-coding exons, completely preventing protein production.
* **CE (Critical Exon Deletion):** Deletes key exons, which may partially disrupt gene function or lead to a truncated, nonfunctional protein.
* **PTC+1 (Premature Termination Codon +1):** Uses CRISPR-Cas9 to introduce a premature stop codon and a degenerate base in an early exon, potentially producing a truncated or nonfunctional protein [2].

By comparing gene expression levels across these experimental conditions, we aim to determine the impact of each knockout method on gene regulation and identify differentially expressed genes.

**Methods:**

1. Data preprocessing

Since DESeq2 requires raw count data and performs its own normalization, we first prepared the metadata. This involved converting the MorPhiC dataset into a structured CSV file containing essential details such as sample ID, knockout strategy (KO, CE, PTC), gene identifier, and cell scheme.

After loading both the raw count data and metadata into our analysis environment, we ensured proper alignment between the datasets. Maintaining consistent sample order and matching raw counts to their respective metadata was crucial for accurate analysis.

To enhance reliability, we applied the following filtering criteria:

* Retained only samples that had at least three replicates across all comparison groups (KO, CE, PTC, and WT).
* Removed low-expression genes, keeping only those with ≥10 counts in at least three samples across all conditions.

These filtering steps helped reduce noise and improve analytical accuracy by minimizing the impact of low-expression genes. Requiring at least three replicates is crucial in ensuring the most accurate results while using DESeq2 [9].

1. DESeq2 DEG Analysis

With preprocessing complete, we conducted differentially expressed gene (DEG) analysis using DESeq2, focusing on four genes: MXD1, RUNX1, NCOA3, and BHLHE40. These genes were chosen within the same cell differentiation scheme to reduce variance.

Experimental conditions were categorized into four groups: KO, CE, PTC, and WT. To facilitate comparison, WT was set as the reference group in all pairwise analyses (KO vs. WT, CE vs. WT, PTC vs. WT).

For each comparison, DESeq2 calculated:

* Log2 fold changes (to measure the magnitude of expression differences).
* Adjusted p-values (to assess statistical significance).

To focus on biologically significant changes, we applied the following thresholds:

* Log2 fold change ≥ |0.5| (genes with absolute log2 fold changes below 0.5 were excluded).
* Adjusted p-value < 0.05 (to control for false positives).

Only genes that met both criteria were classified as significantly differentially expressed and considered for further investigation.

1. Visualization

To interpret DEG results, we generated volcano plots and UpSet plots:

* Volcano plots display genes based on their log2 fold change (x-axis) and statistical significance (adjusted p-value, y-axis), highlighting upregulated and downregulated genes.
* UpSet plots illustrate the overlap between differentially expressed genes in each experimental group, showing the number of unique and shared DEGs across conditions.

These visualizations help identify patterns of gene regulation across different knockout strategies and provide insights into how gene function is disrupted.

1. Batch Effects

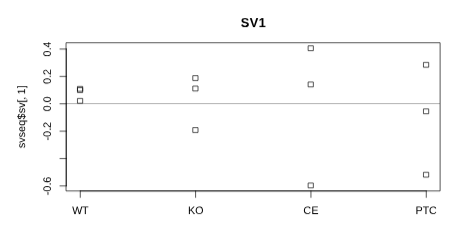
To strengthen the confidence in our analysis, the package svaseq [5] was used alongside DESeq2 as per an RNA-seq workflow vignette provided by Michael Love to adjust for batch effects. The svaseq package was able to find a single significant surrogate variable for the model handling MXD1.

Figure 1: The hidden variation plotted

against knockout method for each sample

Since each sample used the same parent cell line, KOLF2.2J, differentiated into similar cell schemes, we expected batch effects to have a minimal effect on the resulting DEGs. However, there were almost no overlapping DEGs when comparing the inclusion and exclusion of batch effects. Furthermore, the data seemed to skew further away from our reference, the MorPhiC data portal. Therefore, we decided not to include batch effects in our analysis, but they may be important to consider in downstream analysis.

**Results:**

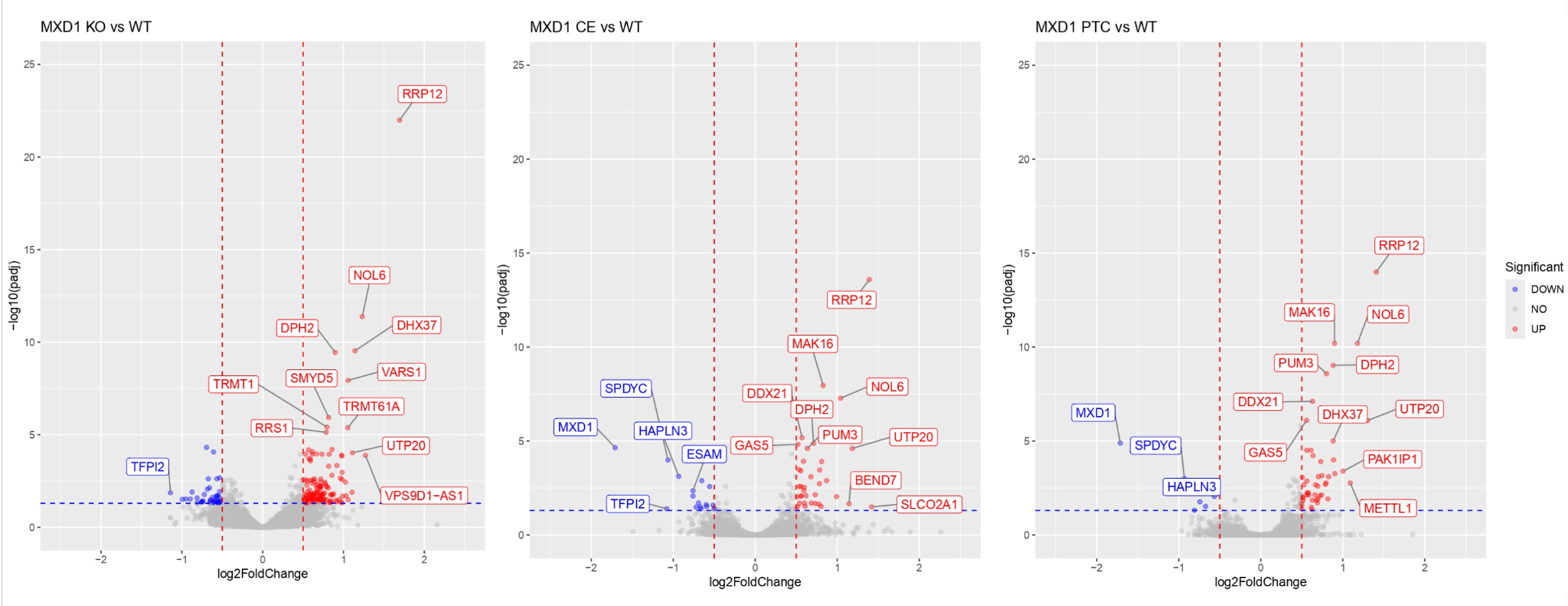
I. MXD1

Figure 2: Volcano plots for MXD1 comparisons plotting adjusted

p-value on the y-axis and log2FoldChange on the x-axis.

Out of all knockout methods used in samples targeting MXD1, full protein coding deletion had significantly more DEGs, while all of them had more up-regulated than down-regulated genes, as seen in Figure 2. Even though there is visible variation in the number of significant DEGs, all methods lead to the identification of the most significantly differentially expressed genes, such as RRP12, NOL6, and DPH2.

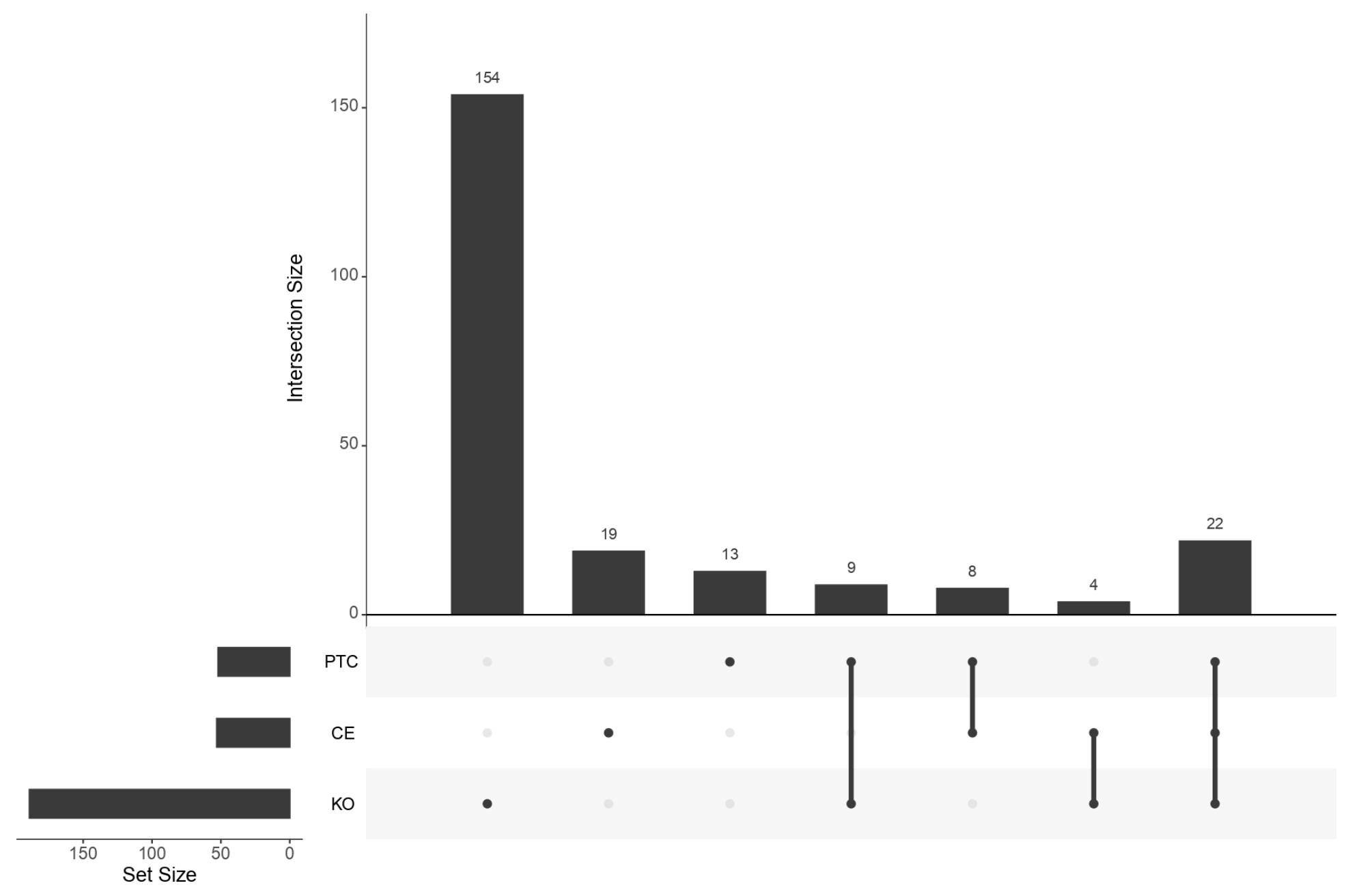
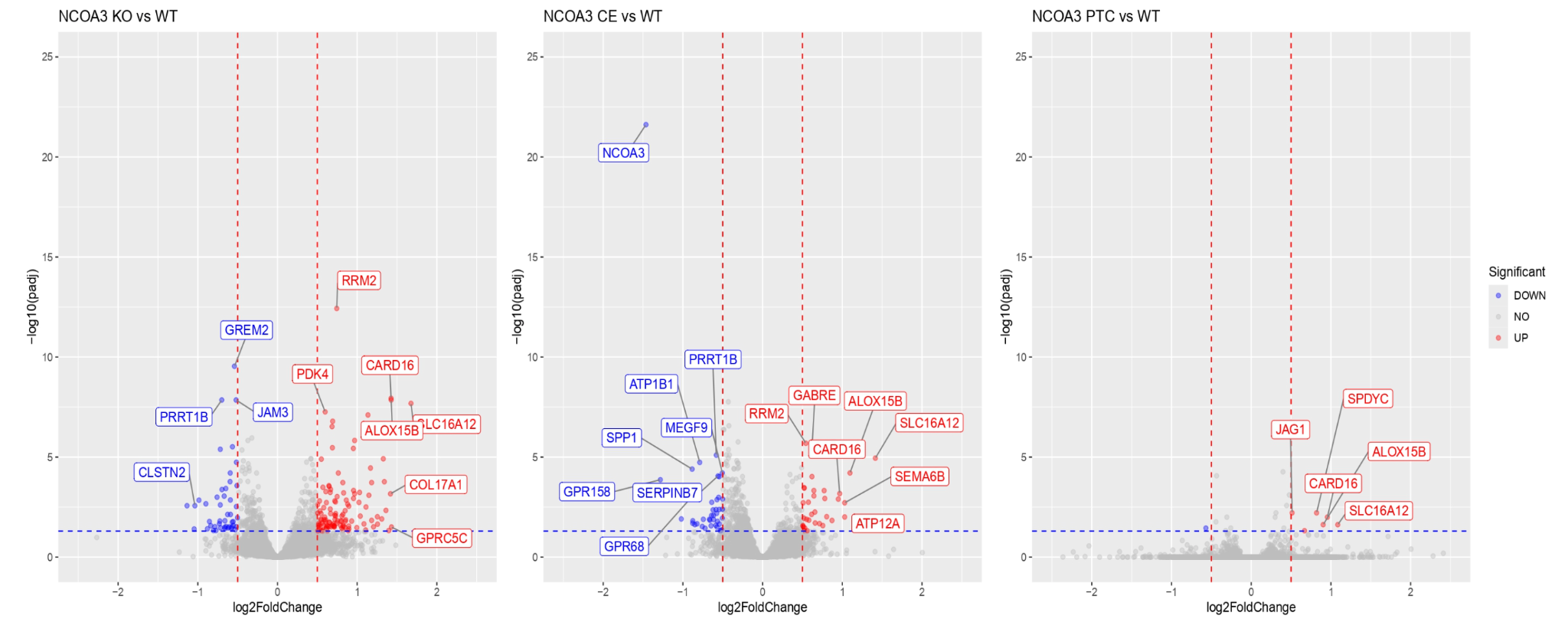
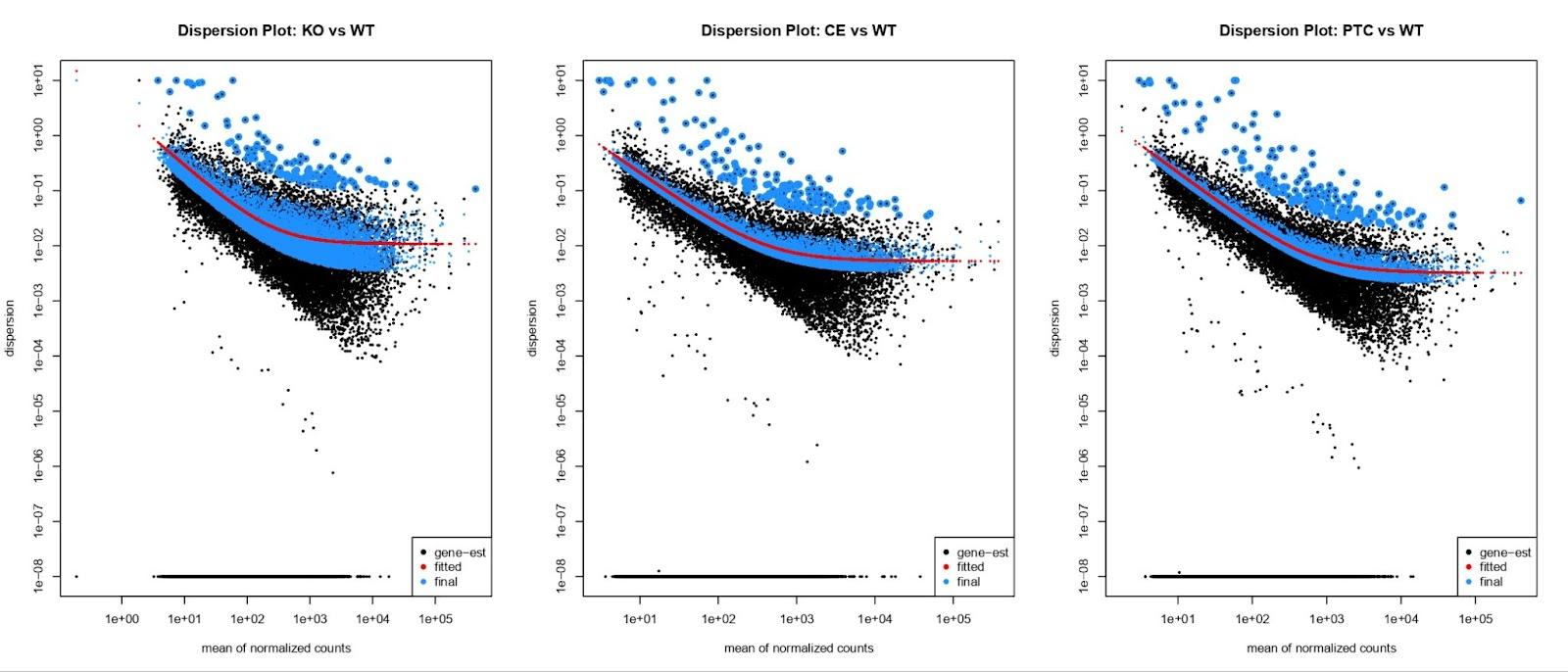


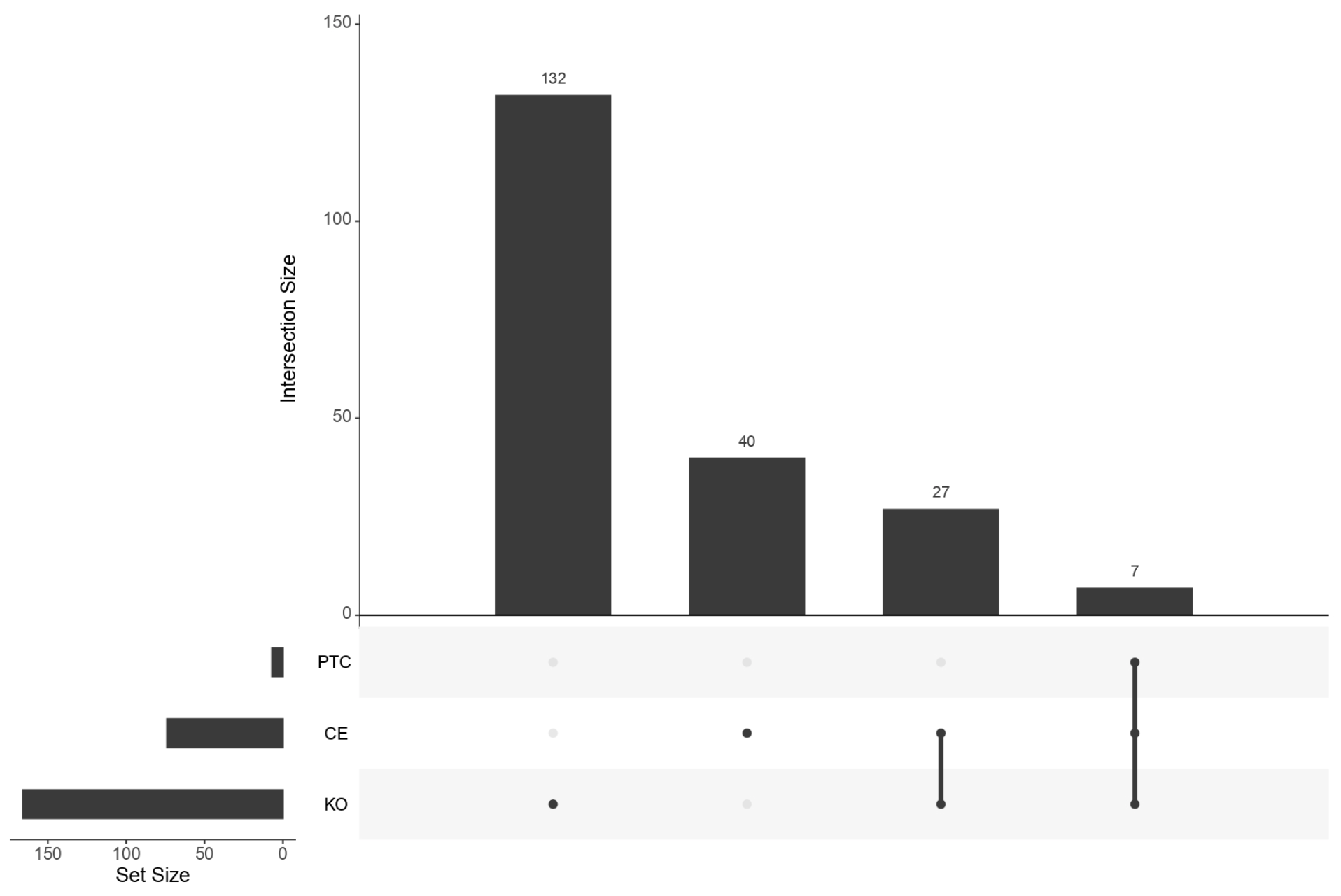
Figure 3: An UpSet plot showing the difference in frequency of

significant DEGs for each knockout method targeting MXD1.

.

The UpSet plot in Figure 3 better displays the sizable increase in significant DEGs for the KO vs WT comparison. There is about a tenfold increase in unique DEGs comparing KO to both the CE and PTC methods. While the genes BHLHE40 and RUNX1 did not share a similar pattern in gene expression between knockout methods, there is a considerable similarity with NCOA3, as shown in the following figures.

II. NCOA3



1. KO vs WT

For NCOA3, KO produced the greatest amount of DEGs, with a total of 166. 118 of these are upregulated, while only 48 are downregulated.

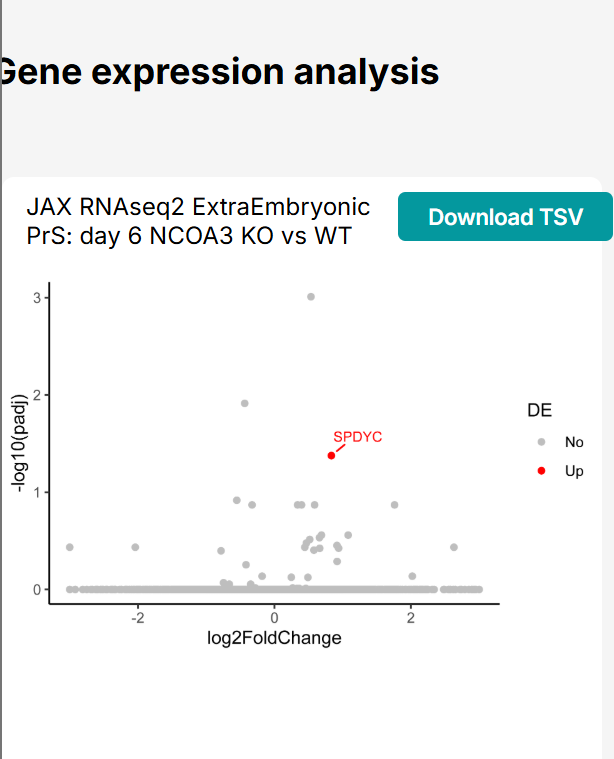
1. CE vs WT

Produced 74 DEGs and has the most even split with 34 being upregulated and 40 downregulated.

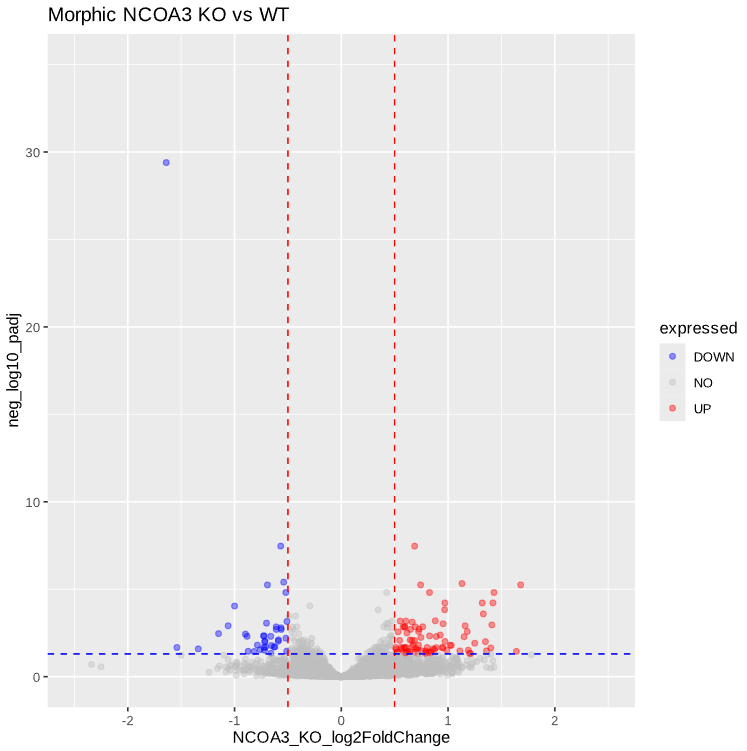
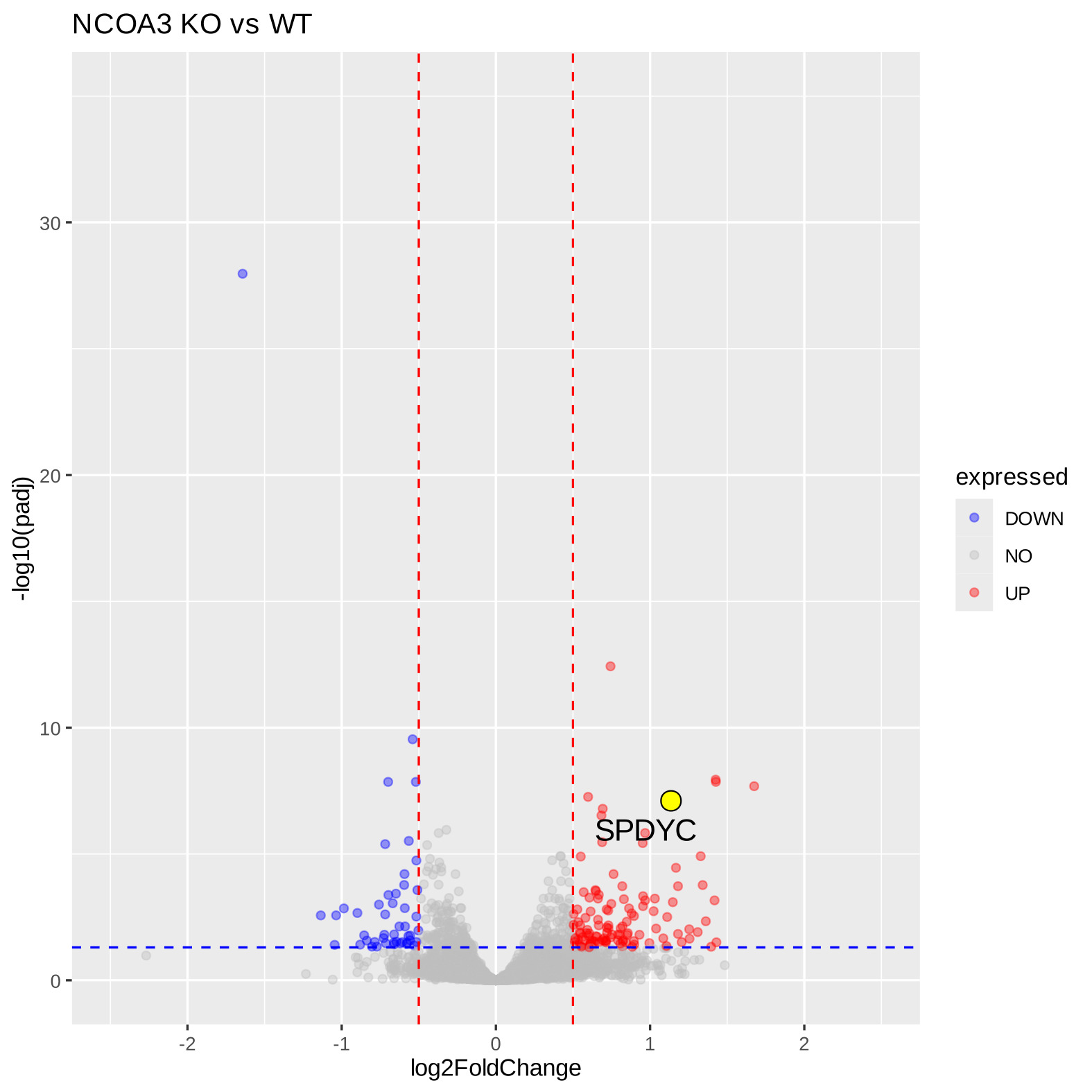
1. PTC vs WT

Only 7 DEGs are produced using PTC+1, 6 of them being upregulated and a single downregulated gene.

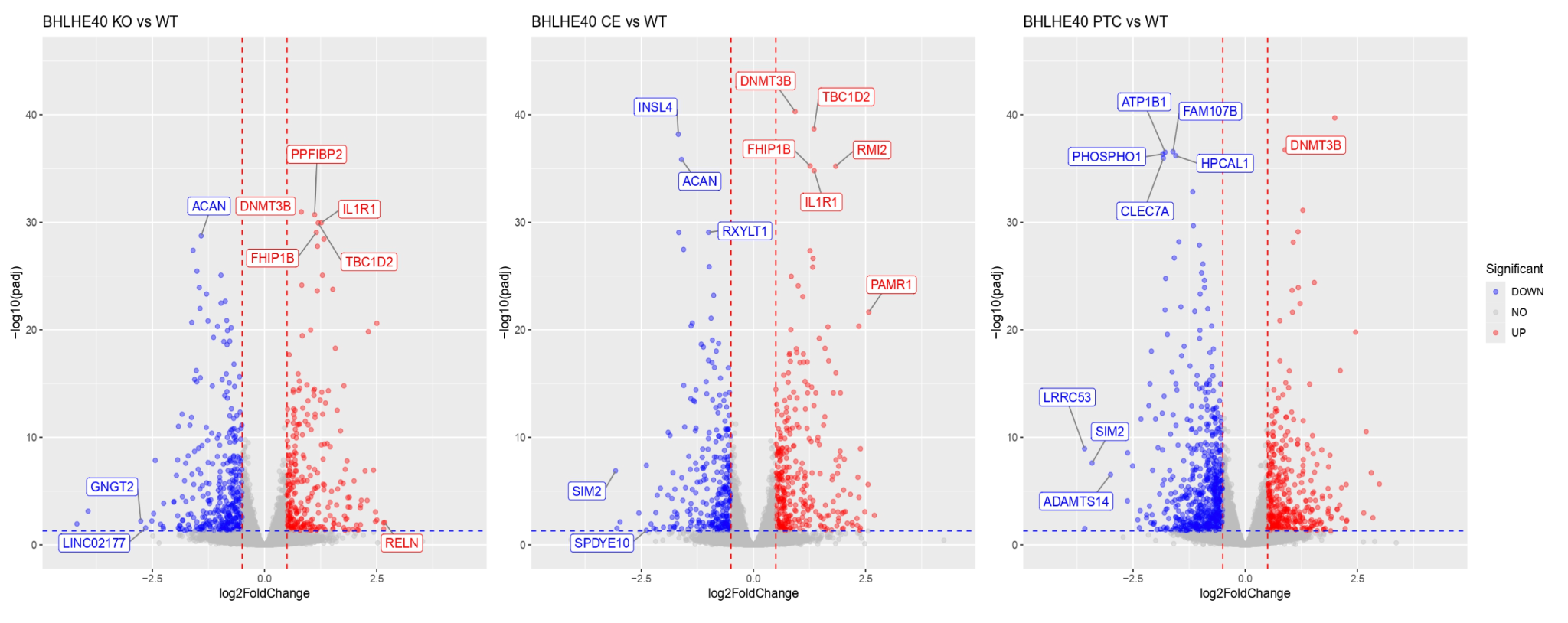
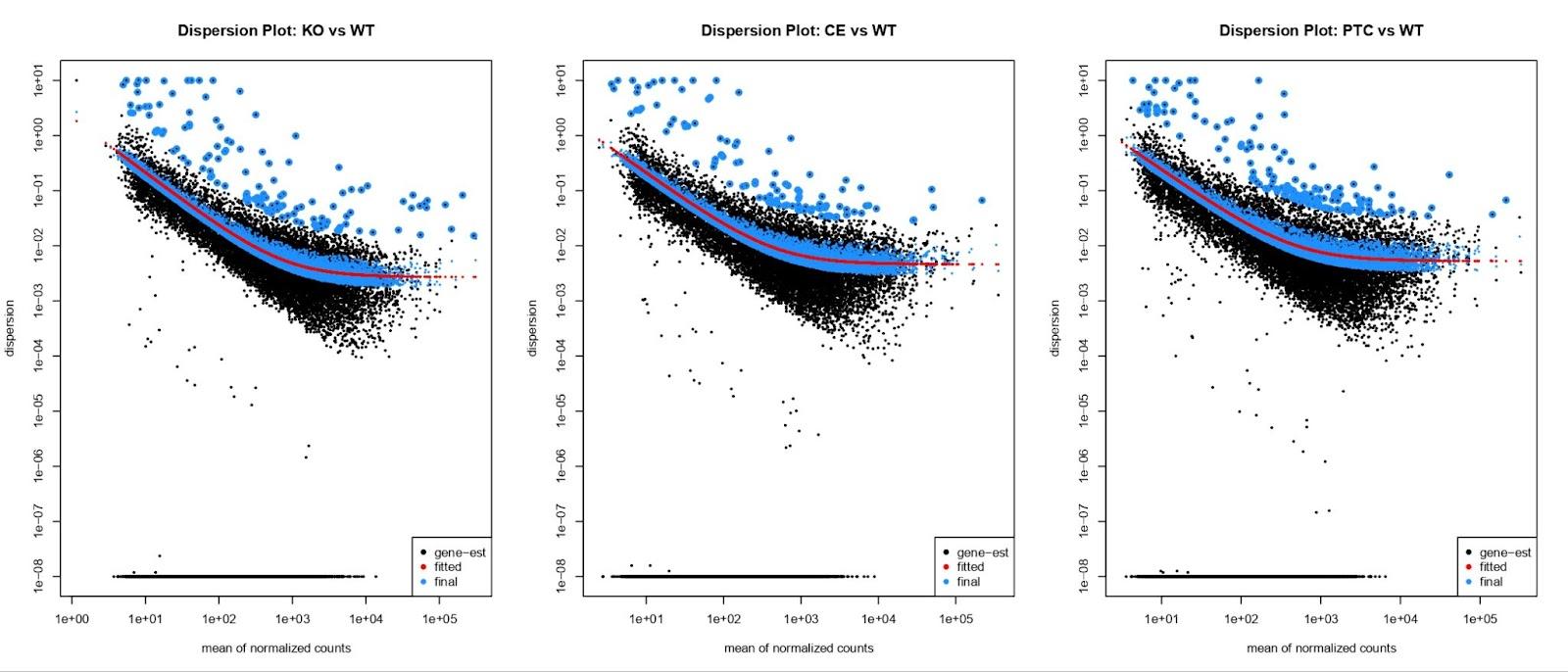
The Morphic project generated a volcano plot for NCOA3 that stands out from the others they have analyzed, both in appearance and content. Notably, the plot lacks an enrichment analysis and displays only a single significantly upregulated differentially expressed gene (SPDYC).



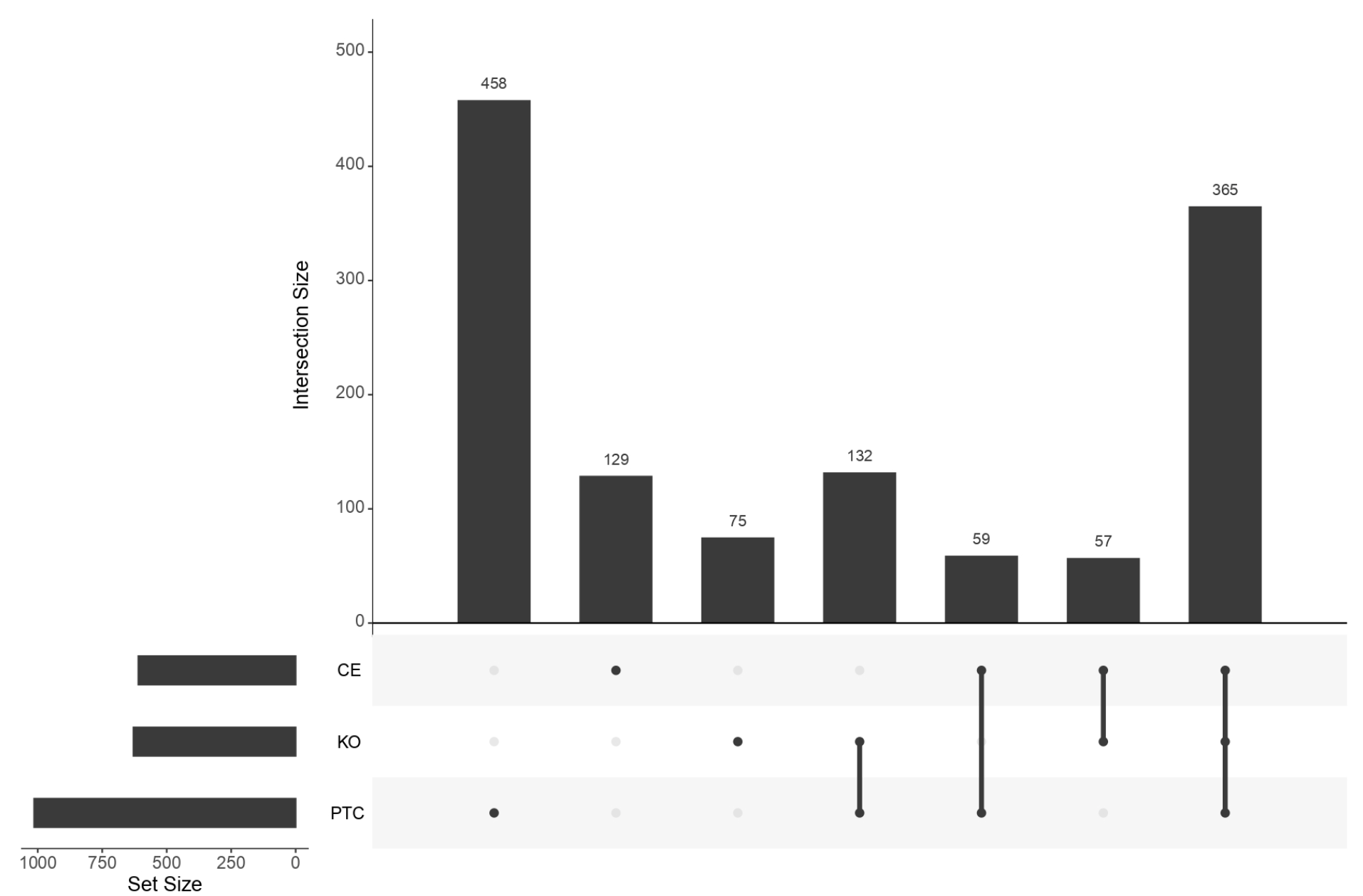
When plotting Morphic’s normalized values using the same method shown in their graph (KO), the resulting volcano plot revealed many more DEGs and took on a more typical pattern expected from such a plot. This graph aligned more closely with the results obtained from analyzing the raw counts using DESeq2, and the graph scaled significantly larger than the one displayed.



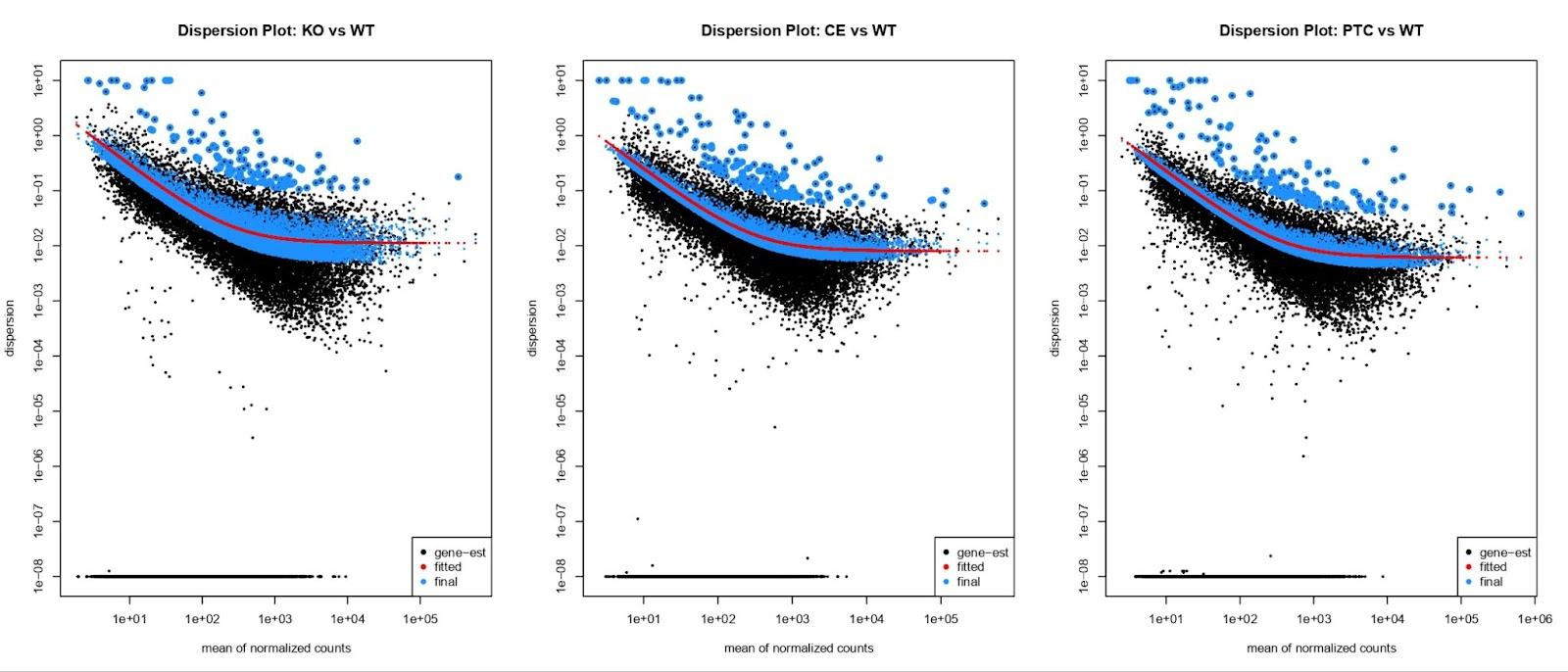
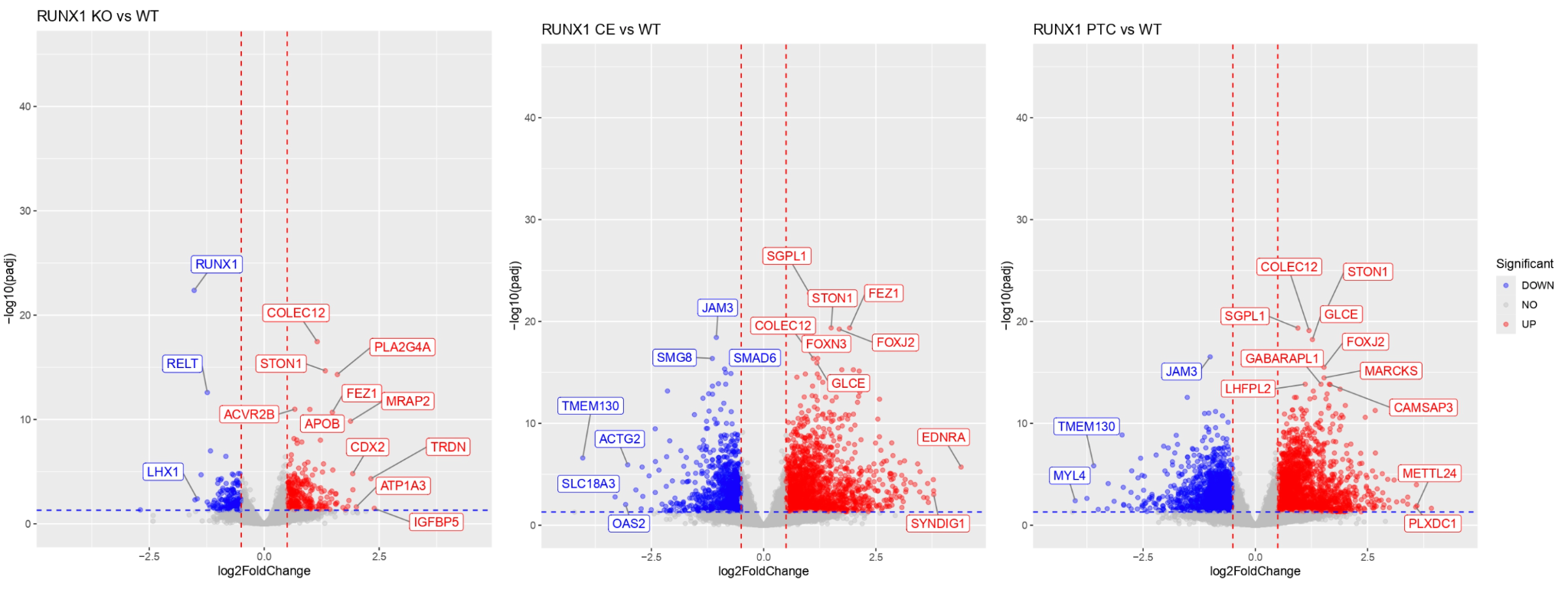
III. BHLHE40

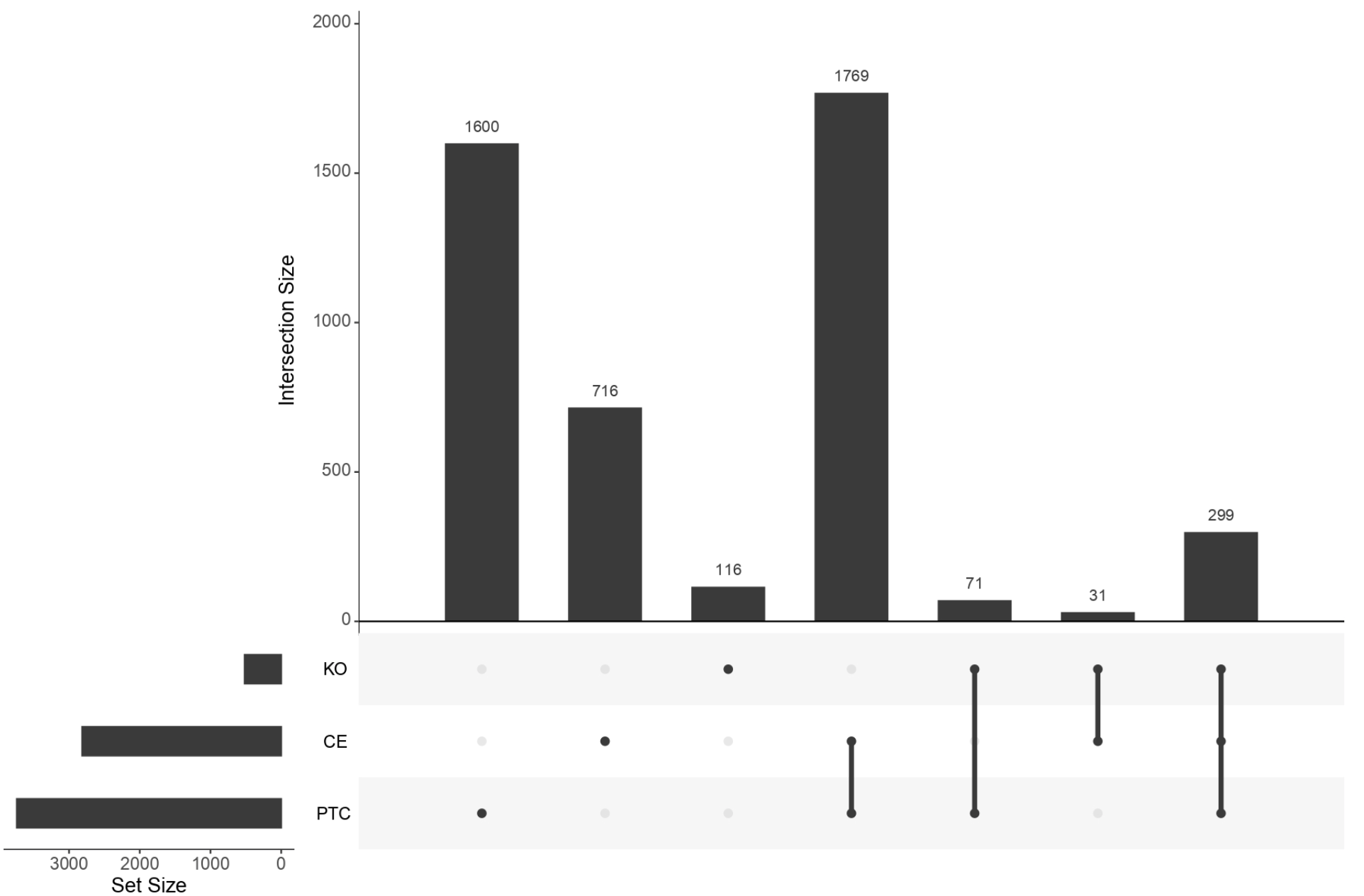




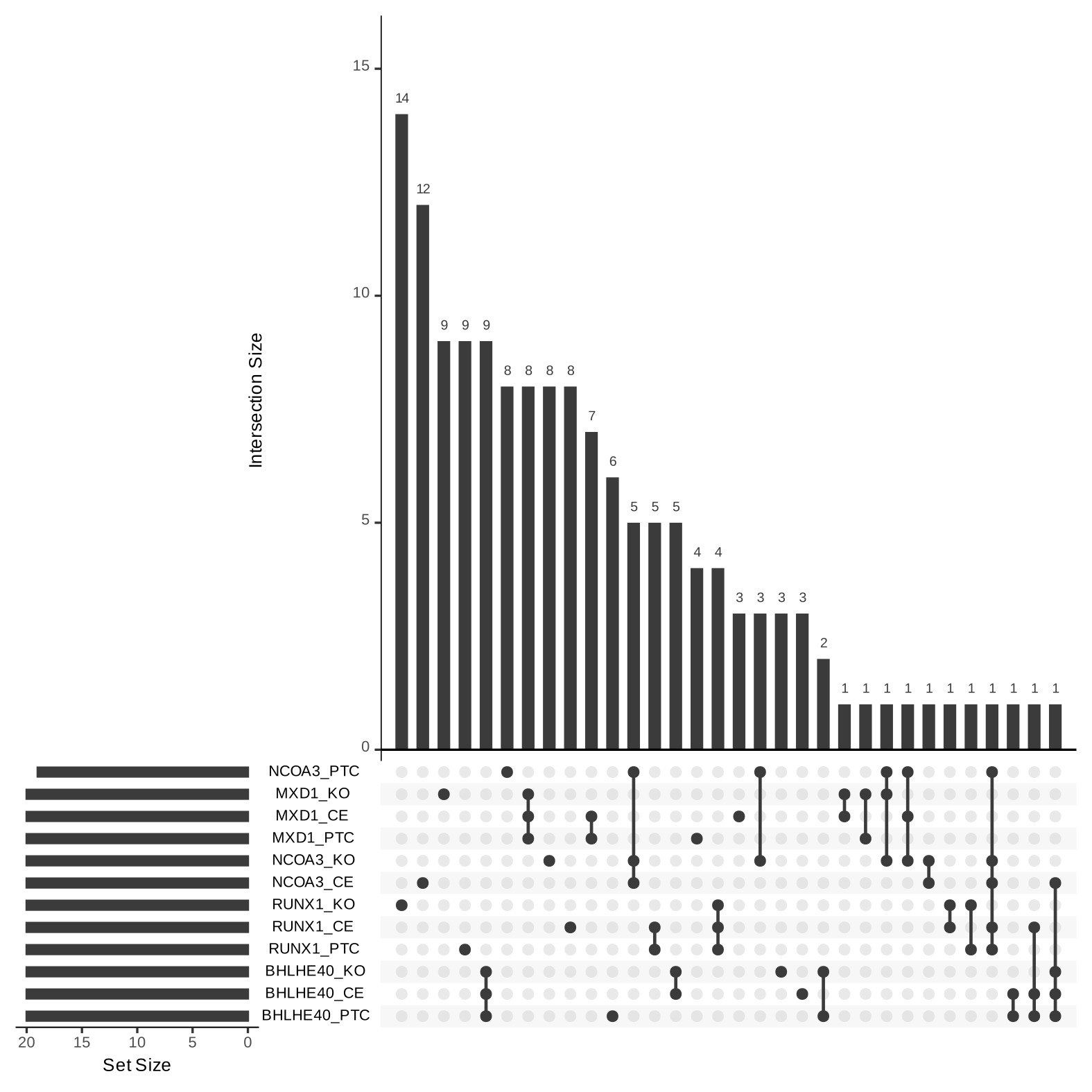
For BHLHE40, the dispersion plot, as per the foundational DESeq2 study, represents a downward trend following the Love et al. paper’s discussion on genes having similar average expression strength having similar dispersion [6]. The dispersion plots from all three experimental groups are similar. Supporting the similarity in dispersion plots, the upset plot reveals that the PTC overlaps with both KO and CE, having the second highest number of DEGs. The highest number of DEGs in the upset plot is PTC alone, which can be supported by the results shown in volcano plots. PTC has the most DEGs compared to KO and CE, especially downregulated genes. Overall, the results illustrate that PTC has the most impact on DEGs compared to the other groups. From the plots, PTC may have some similar characteristics to KO and CE.

IV. RUNX1





1. From the dispersion plots, we notice the expected downward curve and can see that genes with low expression show higher dispersion and vice versa. The KO plot has fewer highly dispersed genes than the CE or PTC plots, suggesting less variability. PTC has the most highly dispersed genes.
2. From the volcano plots, we similarly see PTC resulted in the most DEGs, followed closely by CE, while KO yielded fewer DEGs.
3. The upset plot provides a clearer representation of the DEGs without delineation between up or down regulation. From this plot, we see the highest category was the DEGs shared by PTC and CE, followed by the DEGs exclusively from PTC.
4. It is notable that the knockout of RUNX1, especially via CE and PTC+1, yielded significantly more DEGs than the other three genes from our analysis.

JAM3 appeared downregulated in the top 20 DEGs for all NCOA3 KO methods and RUNX1’s PTC+1 and CE. The downregulation for JAM3 suggests dependence on these genes, though it is stronger in RUNX1 than in NCOA3 knockouts. 

**Discussion:**

Our results highlight that different CRISPR-based knockout methods (KO, CE, and PTC) distinctly influence gene expression profiles. As previously discussed, the different genes we analyzed did not all trend the same way across different knockout methods, but all the genes showed different DEG patterns across the knockout methods. The PTC+1 method yielded the most DEGs for both RUNX1 and BHLHE40. This could be due to work involving nonsense-mediated decay (NMD) eliminating transcripts or another mRNA quality control mechanism having possible interactions with these genes in particular [2]. The lack of significant DEGs attributed to critical exon deletion could be due to the possibility of exon skipping creating semi-functional or fully functional proteins [12]. The KO method yielded the most DEGs for both MXD1 and NCOA3. This may indicate that these genes don’t have any redundant backup pathways. While there is not a distinct pattern to the frequency of significant DEGs related to knockout methods, the notion that there are possible differences is important to note when working with these studies and warrants more research on the topic.

**Future Work:**

The impact of knockout method on DEGs within the extraembryonic lineage could be further explored by the addition of more genes within the used set of studies, alongside further downstream analysis. Our workflow included the addition of annotations for each gene and the inclusion of batch effects if warranted. The next step in this study would likely be pathway enrichment analysis to discern the function of up-regulated and down-regulated genes and ascertain the possible relationship between not just the knocked-out gene itself but also between knockout methods.

**Conclusion:**

This report analyzed the impact of different CRISPR knockout methods (KO, CE, and PTC+1) on gene expression in four target genes. The results showed that each method produced distinct differentially expressed gene (DEG) patterns, with PTC+1 leading to more DEGs in some cases, likely due to nonsense-mediated decay, while KO resulted in more DEGs for others, likely due to the absence of compensatory pathways. These findings highlight the importance of selecting the appropriate knockout strategy based on experimental goals. Future studies should expand the analysis to additional genes and validate these findings through functional experiments. Additionally, this report may be of use in tandem studies to better understand connections between genes and pathways.

**References:**

1. Gehlenborg N (2019). \_UpSetR: A More Scalable Alternative to Venn and

Euler Diagrams for Visualizing Intersecting Sets\_. R package version

1.4.0, <[https://CRAN.R-project.org/package=UpSetR](https://cran.r-project.org/package=UpSetR)>.

1. Hwang J, Kim YK. When a ribosome encounters a premature termination codon. BMB Rep. 2013 Jan;46(1):9-16. doi: 10.5483/bmbrep.2013.46.1.002. PMID: 23351378; PMCID: PMC4133823.
2. H. Wickham. ggplot2: Elegant Graphics for Data Analysis.

Springer-Verlag New York, 2016.

1. Lawrence M, Huber W, Pages H, Aboyoun P, Carlson M, et al. (2013)

Software for Computing and Annotating Genomic Ranges. PLoS Comput

Biol 9(8): e1003118. doi:10.1371/journal.pcbi.1003118

1. Leek JT, Johnson WE, Parker HS, Fertig EJ, Jaffe AE, Zhang Y, Storey

JD, Torres LC (2023). \_sva: Surrogate Variable Analysis\_.

doi:10.18129/B9.bioc.sva <https://doi.org/10.18129/B9.bioc.sva>, R

package version 3.50.0, <https://bioconductor.org/packages/sva>.

1. Love, M.I., Huber, W., Anders, S. Moderated estimation of fold change

and dispersion for RNA-seq data with DESeq2 Genome Biology 15(12):550

(2014)

1. M. Lawrence, R. Gentleman, V. Carey: "rtracklayer: an {R} package for

interfacing with genome browsers". Bioinformatics 25:1841-1842

1. Morgan M, Ramos M (2024). \_BiocManager: Access the Bioconductor

Project Package Repository\_. R package version 1.30.25,

<[https://CRAN.R-project.org/package=BiocManager](https://cran.r-project.org/package=BiocManager)>.

1. Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA. 2016 Jun;22(6):839-51. doi: 10.1261/rna.053959.115. Epub 2016 Mar 28. Erratum in: RNA. 2016 Oct;22(10):1641. doi: 10.1261/rna.058339.116. PMID: 27022035; PMCID: PMC4878611<https://rnajournal.cshlp.org/content/22/6/839>
2. Slowikowski K (2024). \_ggrepel: Automatically Position

Non-Overlapping Text Labels with 'ggplot2'\_. R package version 0.9.6,

<[https://CRAN.R-project.org/package=ggrepel](https://cran.r-project.org/package=ggrepel)>.

1. Study 2- RNA-seq of male KOLF2.2J hiPSC-derived trophoblast cell lines homozygous null for seven different transcription factors [**GSE288289**](https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE288289) Gene expression and metadata.
2. Mou, H., Smith, J.L., Peng, L. *et al.* CRISPR/Cas9-mediated genome editing induces exon skipping by alternative splicing or exon deletion. *Genome Biol* **18**, 108 (2017). <https://doi.org/10.1186/s13059-017-1237-8>