ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

REPORT

GENOMICS AND BIOINFORMATICS

Viola Renne

27th May 2024

1 Introduction

In many human cancers, inhibiting oncogenic kinases is a crucial component of therapy. Tang et al. 2021 presents a genome-wide knockout screen in EGFR-mutant glioblastoma (GBM) cells treated with neratinib, a type II EGFR kinase inhibitor. Since responses to neratinib vary among individuals, the aim is identify signaling or metabolic pathways that modulate the response to EGFR kinase inhibitors in EGFR-mutant GBM. The research identifies the amino acid starvation sensor GCN2 and its downstream effector ATF4 as key factors, showing that neratinib activates the GCN2/ATF4 signaling pathway.

2 Method

To generate the main figure (Extended Data Fig. 4a,b), I conducted a differential gene expression analysis, using the provided RNA-seq count matrix and DESeq2 package. Initially, I compared gene expression at 0h (control - untreated) with both 6h and 72h (treated). The volcano plot was then generated using DGE results, with adjusted p-values recalculated using the p.adjust function with Benjamini-Hochberg set as the adjustment method. I maintained the option to specify a genesetFile and introduced a parameter, modifications, which, when set to false, reproduces the exact figure from the paper 1. To achieve this, I used the krige amino acid deprivation gene list from the paper. When the modifications parameter is set to true, the top 10 upregulated and downregulated genes are labelled 2. This process involves first identifying significant genes, which are those with adjusted p-values below the specified alpha threshold of 0.05. Then, the significant genes are sorted by their log2 fold change, allowing for the selection of the top (upregulated) and bottom (downregulated) 10 genes.

To visualize differentially expressed genes in a different way, I generated a heatmap using the pheatmap package (pheatmap package). The heatmap displays normalized gene expression values of the top 10 upregulated and top 10 downregulated genes (from the comparison between 72H and 0H) across all six samples. Clustering is performed on both the rows (genes) and columns (samples) to identify groups of genes with similar expression patterns among samples and groups of samples with similar expressions patterns among genes. The pheatmap function allows the selection of the distance method for clustering. For the first figure (6), I used the default method, which is Euclidean distance, while for the second figure (7), I used Pearson correlation as similarity metric.

Additionally, I have chosen to reproduce Extended Figures 4c-d, which were generated using Gene Set Enrichment Analysis (GSEA Software). This analysis uses a list of ranked genes (from highest to lowest log fold change in our case) and a specific gene set. In this study, the gene set used is the krige amino acid deprivation, which contains genes up-regulated in HL-60 cells (acute promyelocytic leukemia, APL) after amino acid deprivation MSigDB krige amino acid deprivation. Unlike pathway enrichment analysis, which focuses solely on differentially expressed genes, GSEA considers all genes. It evaluates the distribution of genes from the provided gene set within the ranked list to determine if they are predominantly upregulated (towards the top) or downregulated (towards the bottom). The output includes a p-value and an enrichment score (ES), which is the maximum deviation from zero. The ES indicates how much the gene set is over-represented at either the top or bottom of the ranked list. A positive ES suggests upregulation, while a negative ES suggests downregulation. GSEA software also provides an enrichment plot where the x-axis shows the position in the ranked gene list and the y-axis displays the running enrichment score 3. This plot highlights the hits, which are the positions of genes from the provided gene set in the ranked list. It also identifies the leading edge subset, which includes the genes that contribute most to the ES. When comparing different gene sets, the normalized ES (NES) is crucial as it accounts for the number of genes in each set.

3 Discussion

The paper demonstrates that GCN2 is involved in the GCN2/ATF4 signaling pathway, which is recognized for promoting apoptosis, typically in response to prolonged hyperactivation of the ISR. To investigate the kinetics of ATF4 gene target induction by neratinib, a differential gene expression analysis was conducted. The analysis revealed that the genes induced by neratinib included numerous known ATF4-regulated genes, among them several metabolite transporters (Extended Data Fig. 4a,b 1).

The heatmap provides additional information compared to the volcano plot, as it shows the expression levels across all six samples rather than the log2 fold change and adjusted p-values. The first 10 rows of the heatmap represent upregulated genes, whose expression increases over time (from 0H to 72H), while the last 10 rows represent downregulated genes. Additionally, the heatmap helps identify groups of genes with similar expression patterns across samples by examining the clustering performed on the rows. For both Euclidean distance and Pearson correlation, upregulated genes are first grouped with other upregulated genes, and similarly, downregulated genes

are grouped together. The order in which the different genes are added to a cluster (and so, their similarity) appears to be consistent between the two methods. The clustering applied to the columns (i.e., across different samples) initially groups replicates from the same hour together, as expected due to their higher similarity, despite considering only a subset (20 genes) of the entire transcriptome.

Gene set enrichment analysis showed significant enrichment of the amino acid deprivation response gene set at both 6 and 72 hours post-treatment, indicating sustained ISR activation by neratinib.

4 CONCLUSION

For the volcano plot, due to the lack of corresponding numerical values, the comparison was only visual. It appears that Figure 1 is identical to the Figure provided by the paper.

I have compared the results of GSEA stated by the study and obtained with my analysis in 1. Performing the GSEA made me notice that the krige amino acid deprivation gene has been updated and now includes 29 genes [MSigDB krige amino acid deprivation], an increase from the original 27 genes list used in the paper (with the additional genes being RETN and CXCL8). I have adjusted the list accordingly to achieve the exact results of the paper shown in Figures 1 and 5. Using the original list from the paper, I was able to reproduce all the results for 6H except for the Normalized ES. It is important to note that the results from the paper are sourced from the Source Data Extended Data Figures, not the values displayed in the figure. For 72H, the results from the paper are not available, and this comparison is not possible 2.

The table does not include the following values: NOM p-val, FDR q-val, and FWER p-val. This is because the NOM p-value was consistently 0.0, likely due to its extremely small size, resulting in an approximation of zero. FDR q-val and FWER p-val are generated when the gene sets database includes multiple lists, however, these results are only available for the updated list and not for the original list. Additionally, the contents of the gene sets database may have changed, preventing me from reproducing the same results as in the paper.

Even if the updated list contains 29 genes, only 24 are present in our ranked list (first row of Tables 1 and 2). Conversely, using the original list from the paper, all 27 genes are found. To investigate why only 24 of the 29 genes in the updated list were found, I noticed that some genes had numbers added to their names (usually only number 1) in the updated list that were not present in the original list from the paper. After removing these numbers, the count increased to 28. The missing gene is RETN, which, although present in the provided count matrix, is excluded because all read counts are zero.

Overall, the results appear robust.

REFERENCES

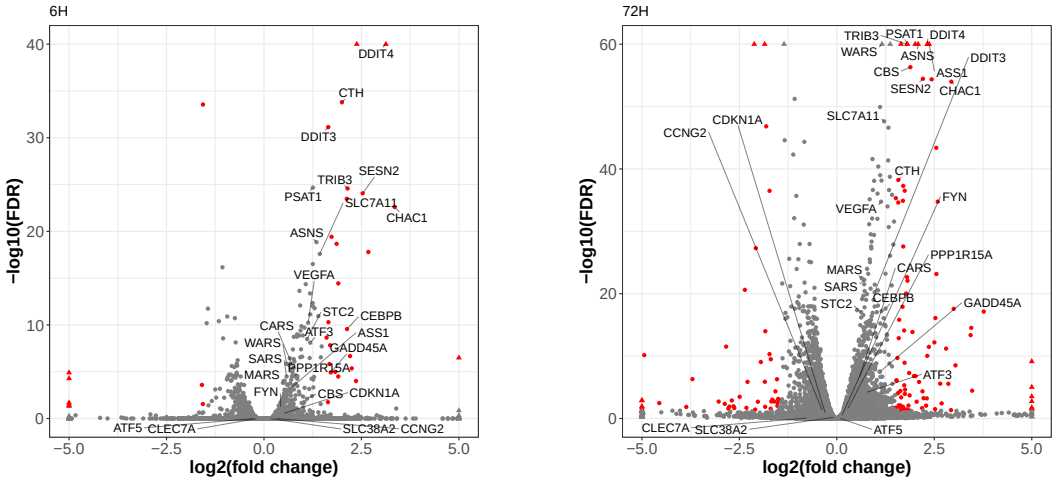
Tang, Colin P. et al. (Dec. 2021). 'GCN2 kinase activation by ATP-competitive kinase inhibitors'. In: *Nature Chemical Biology* 18.2, pp. 207–215. ISSN: 1552-4469. DOI: 10.1038/s41589-021-00947-8. URL: http://dx.doi.org/10.1038/s41589-021-00947-8.

pheatmap (package). *pheatmap: Pretty Heatmaps*. Accessed: 2024-05-23. URL: https://cran.r-project.org/web/packages/pheatmap/index.html.

GSEA (Software). *Gene Set Enrichment Analysis*. Accessed: 2024-05-23. URL: https://www.gsea-msigdb.org/gsea/index.jsp.

MSigDB (krige amino acid deprivation). *KRIGE_AMINO_ACID_DEPRIVATION*. Accessed: 2024-05-23. URL: https://www.gsea-msigdb.org/gsea/msigdb/human/geneset/KRIGE_AMINO_ACID_DEPRIVATION.html.

5 Figures and Tables

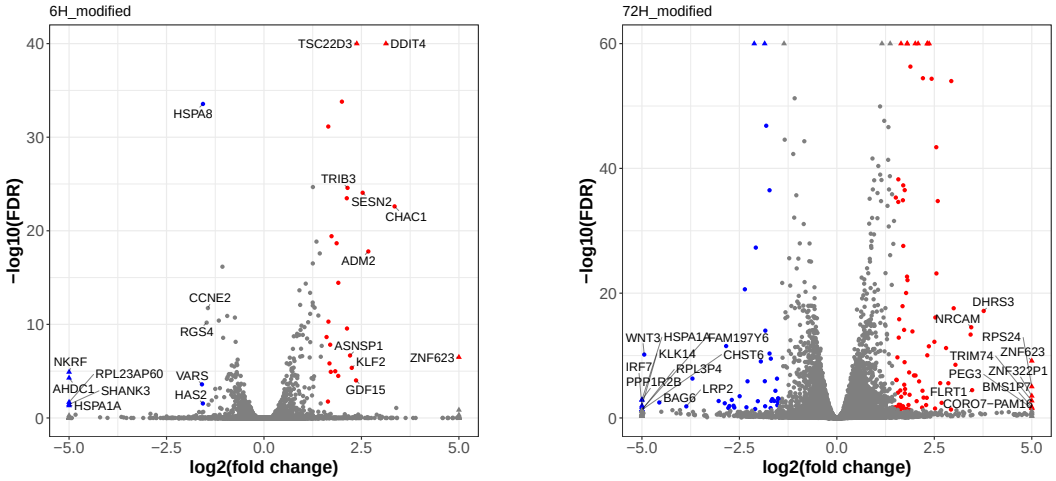


(A) Volcano plot of 6H vs 0H

(B) Volcano plot of 72H vs 0H

Figure 1

Volcano plots showing differential expression of genes in SF268 cells treated with 675 nM neratinib or vehicle for 6 hours (A) and 72 hours (B) with gene labels corresponding to the enriched krige amino acid gene set. Significant upregulated and downregulated genes are highlighted in red.



(A) Volcano plot of 6H vs 0H

(B) Volcano plot of 72H vs 0H

Figure 2

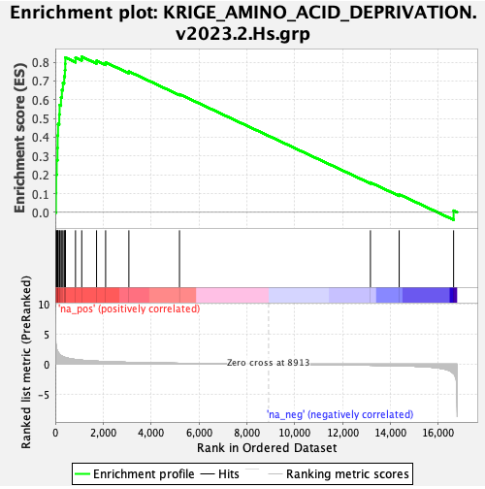
Volcano plots showing differential expression of genes in SF268 cells treated with 675 nM neratinib or vehicle for 6 hours (A) and 72 hours (B) with gene labels corresponding to the top 10 upregulated and downregulated genes. Significant upregulated genes are highlighted in red, while downregulated genes are highlighted in blue.

TABLE 1
GSEA results 6H

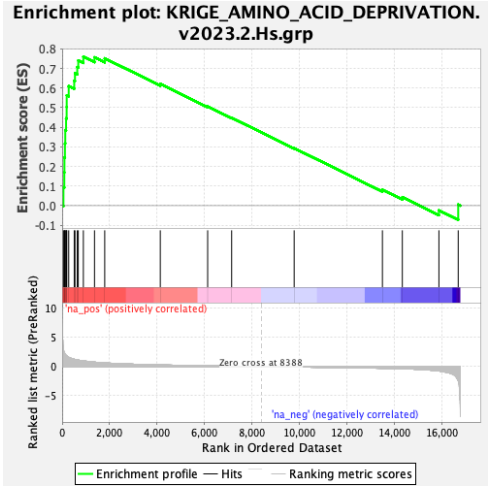
NAME	SIZE	ES	NES	RANK AT MAX	LEADING EDGE
Updated list 6H	24	0.8293936	2.0709746	1105	tags=71%, list=7%, signal=76%
Updated list with modifications of names 6H	28	0.81065124	2.0652525	1790	tags=79%, list=11%, signal=88%
List from paper 6H	27	0.8543334	2.1825404	1790	tags=81%, list=11%, signal=91%
Results from paper	27	0.8543335	2.2001698	1790	tags=81%, list=11%, signal=91%

TABLE 2
GSEA results 72H

NAME	SIZE	ES	NES	RANK AT MAX	LEADING EDGE
Updated list 72H	24	0.76025623	1.996554	893	tags=58%, list=5%, signal=62%
Updated list with modifications of names 72H	28	0.75051296	2.0244715	2040	tags=71%, list=12%, signal=81%
List from paper 72H	27	0.8167544	2.2218797	2040	tags=74%, list=12%, signal=84%

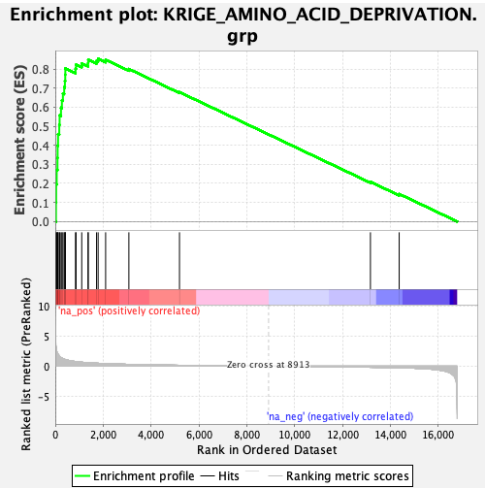


(A) GSEA plot for 6 vs 0h

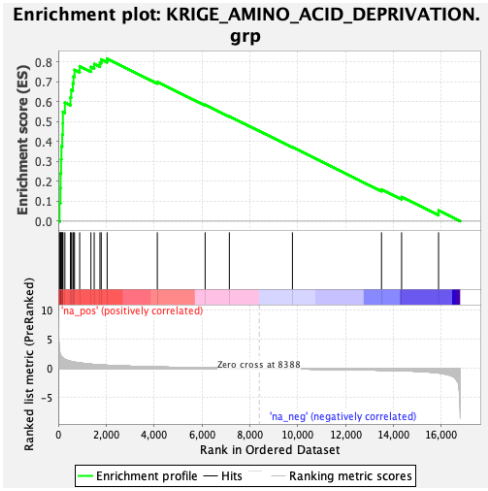


(B) GSEA plot for 72 vs 0h

FIGURE 3
Gene set enrichment analysis results showing differential expression of genes in SF268 cells treated with 675 nM neratinib or vehicle for 6 hours and 72 hours with updated gene set list

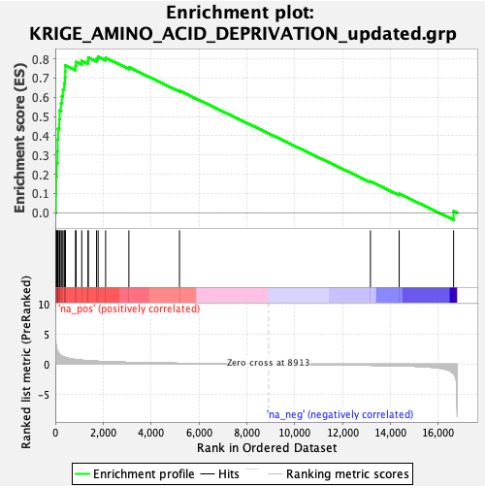


(A) GSEA plot for 6 vs 0h

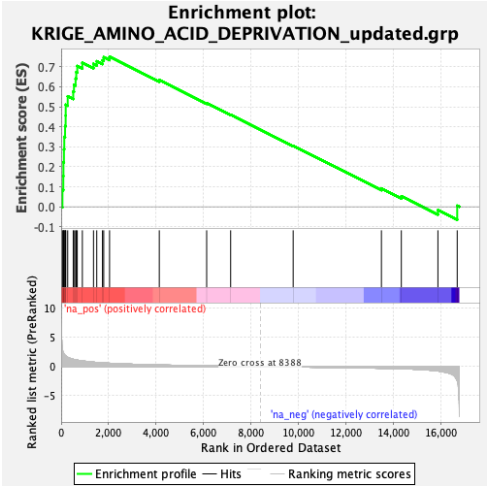


(B) GSEA plot for 72 vs 0h

FIGURE 4
Gene set enrichment analysis results showing differential expression of genes in SF268 cells treated with 675 nM neratinib or vehicle for 6 hours and 72 hours with same gene set list from the paper



(A) GSEA plot for 6 vs 0h



(B) GSEA plot for 72 vs 0h

FIGURE 5

Gene set enrichment analysis results showing differential expression of genes in SF268 cells treated with 675 nM neratinib or vehicle for 6 hours and 72 hours with updated gene set list, but with names modified

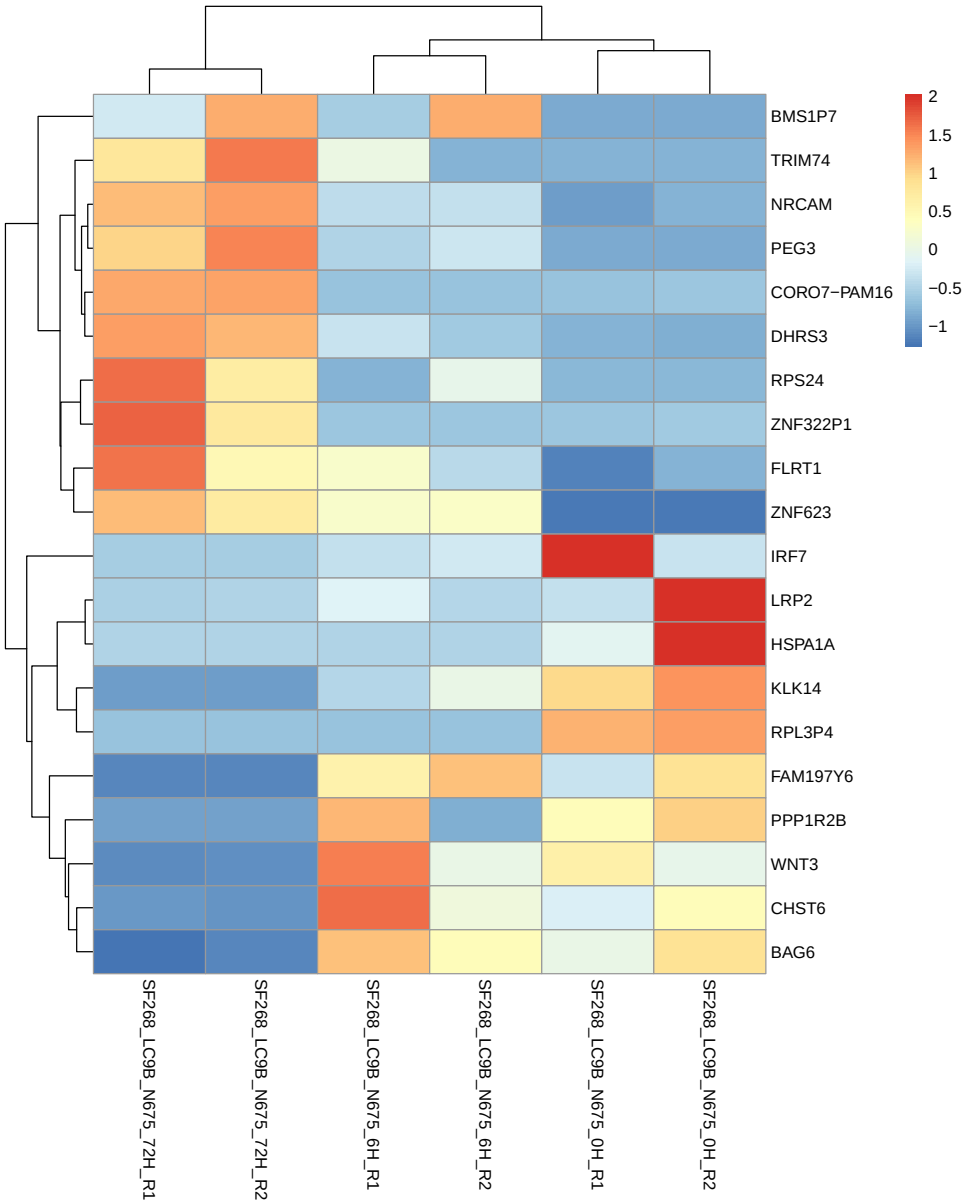


FIGURE 6
Heatmap of the top 10 upregulated and top 10 downregulated genes across all six samples using euclidean distance for rows clustering

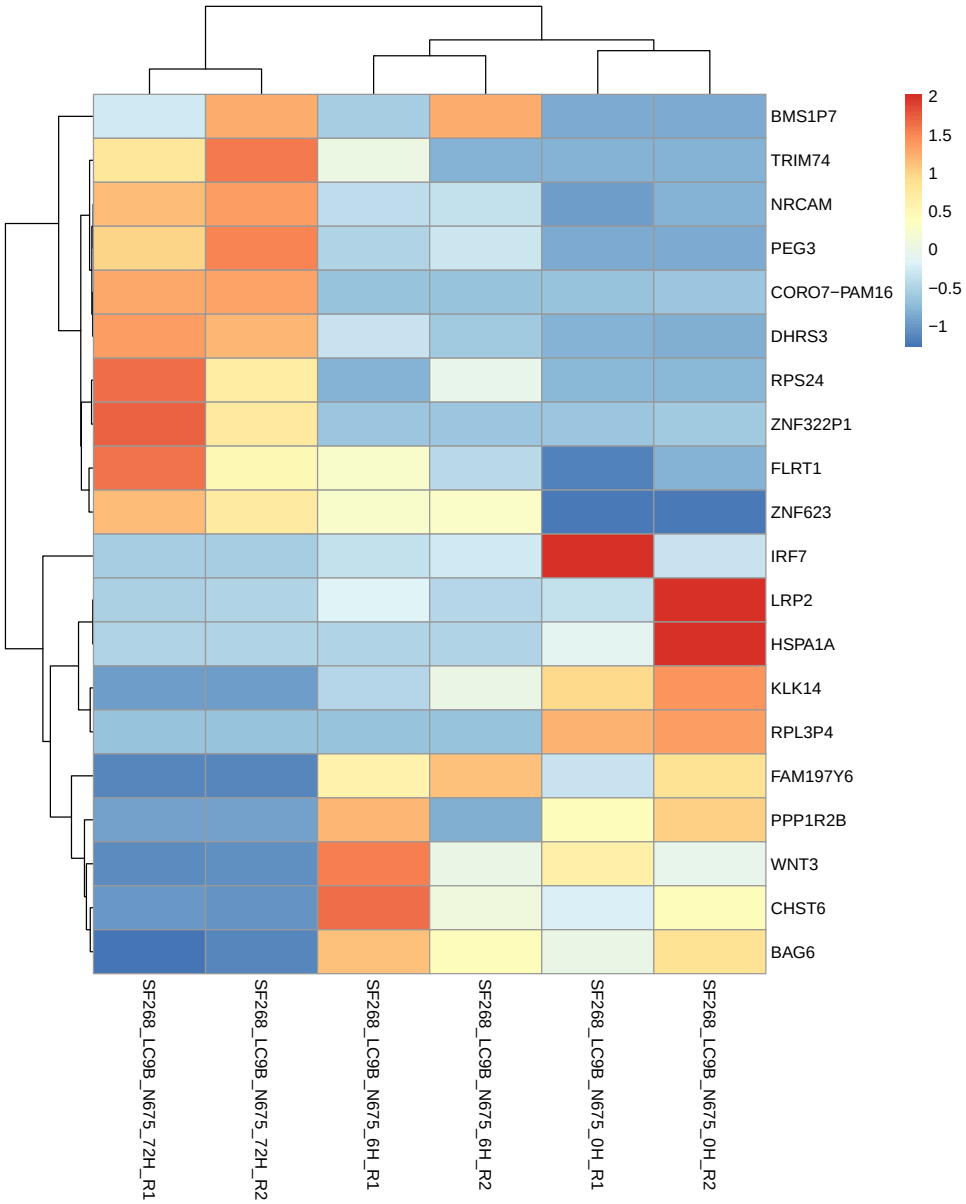


FIGURE 7
Heatmap of the top 10 upregulated and top 10 downregulated genes across all six samples using Pearson correlation for rows clustering