Differential Gene Expression for Ovarian Serous Adenocarcinoma for Women Between the Ages of 25 and 50

By: Vaishnavi Sathiyamoorthy

# Introduction

Ovarian serous adenocarcinoma (OV) is a form of epithelial ovarian cancer. 96% of individuals with ovarian cancer have a TP53 mutation. This gene codes tumor suppressor proteins that prevent cancer development. Studies have shown that gene expression patterns are associated with longer or shorter survival (The Cancer Genome Atlas Research Network, 2023). The purpose of this project is to determine which genes are differentially expressed in samples between the ages of 25-50 that survived for less than or equal to 5 years and greater than 5 years?

# Methods

Sample IDs were taken from the GDC database (Grossman et al., 2016). The database was filtered down to TCGA-OV, age of diagnosis between 25 and 50, gender is female, cancer genus is true, vital status is dead, and data source was open. Group 1 was filtered down to days until death from 0 to 1825 (≤ 5 years). Group 2 was filtered down to days until death from 1826 (> 5 years) (The Cancer Genome Atlas Research Network, 2023).

The analysis was done in the programming language R (R Core Team, 2021) in R Studio (Posit team, 2023). The pipeline used was based on Bentley’s code and modified to fit the parameters of this project (Bentley, 2023). This script can be found in S1. Based on the sample IDs, the gene expression data was downloaded using the package TCGAbiolinks (Colaprico et al., 2016). Group 1 was determined to be the first level and group 2 was determined to be the second level. The samples from both groups were checked to ensure that they were mutually exclusive. Next, it was checked if there were any loci that had more than 90% of raw counts of 0. It was also checked whether any of the samples had read raw counts that varied drastically from the average raw counts. Then, the read depths were normalized using the estimated size factor function from the DESeq2 package and checked if any samples had read depths that were significantly different from the mean read depth (Love et al., 2014).

Principal component analysis of the variance between group 1 and group 2 samples were done using prcomp where the values are scaled. This was done using the stats package (R Core Team, 2023). Then, DESeq2 was used to calculate differential expression across the samples and loci using default parameters (Love et al., 2014). The samples were grouped so that group 1 consisted of individuals who died within 5 years of diagnosis and group 2 consisted of individuals who died after 5 years of diagnosis. A volcano plot of the differential expression analysis was created using ggplot2. The genes that had adjusted p-values of < 0.01 were considered differentially expressed genes between the two groups. Finally, violin plots that compared the expression of each of the significant genes were created using ggplot2 (Wickham, 2016). The expression data was taken from the normalized read counts.

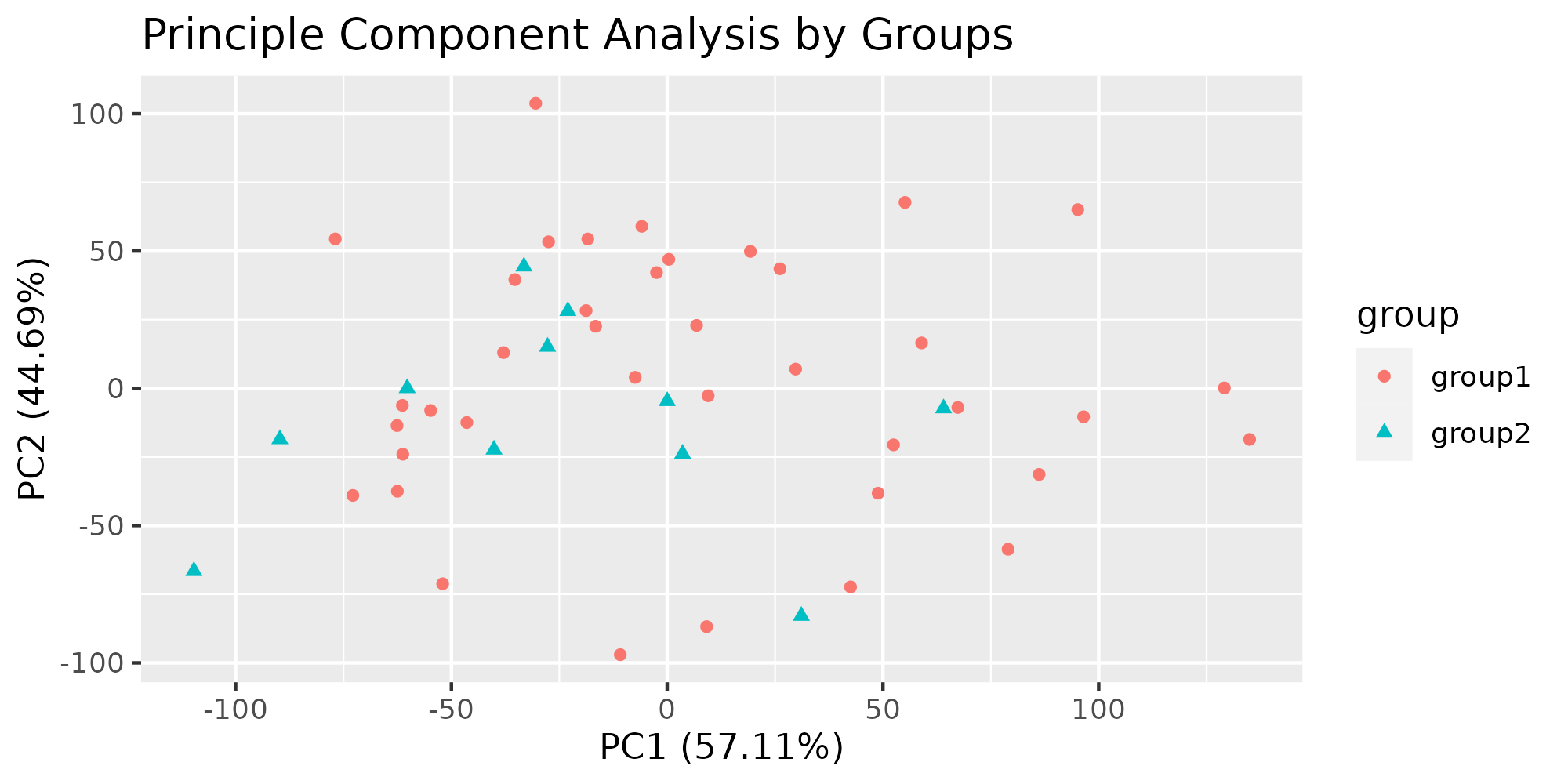
# Results

After downloading the dataset and checking for further filtering of the samples, there were 39 samples for those that died within 5 years of diagnosis (group 1) and 11 samples for those that died after 5 years of diagnosis (group 2). The grouping information can be found in Table1, S2a, and S2b. There were 60,660 genes in total. After filtering the loci that had more than 90% of raw counts as 0, there were 43,953 genes, which can be found in S3.

**Table 1.** Barcodes for each group.

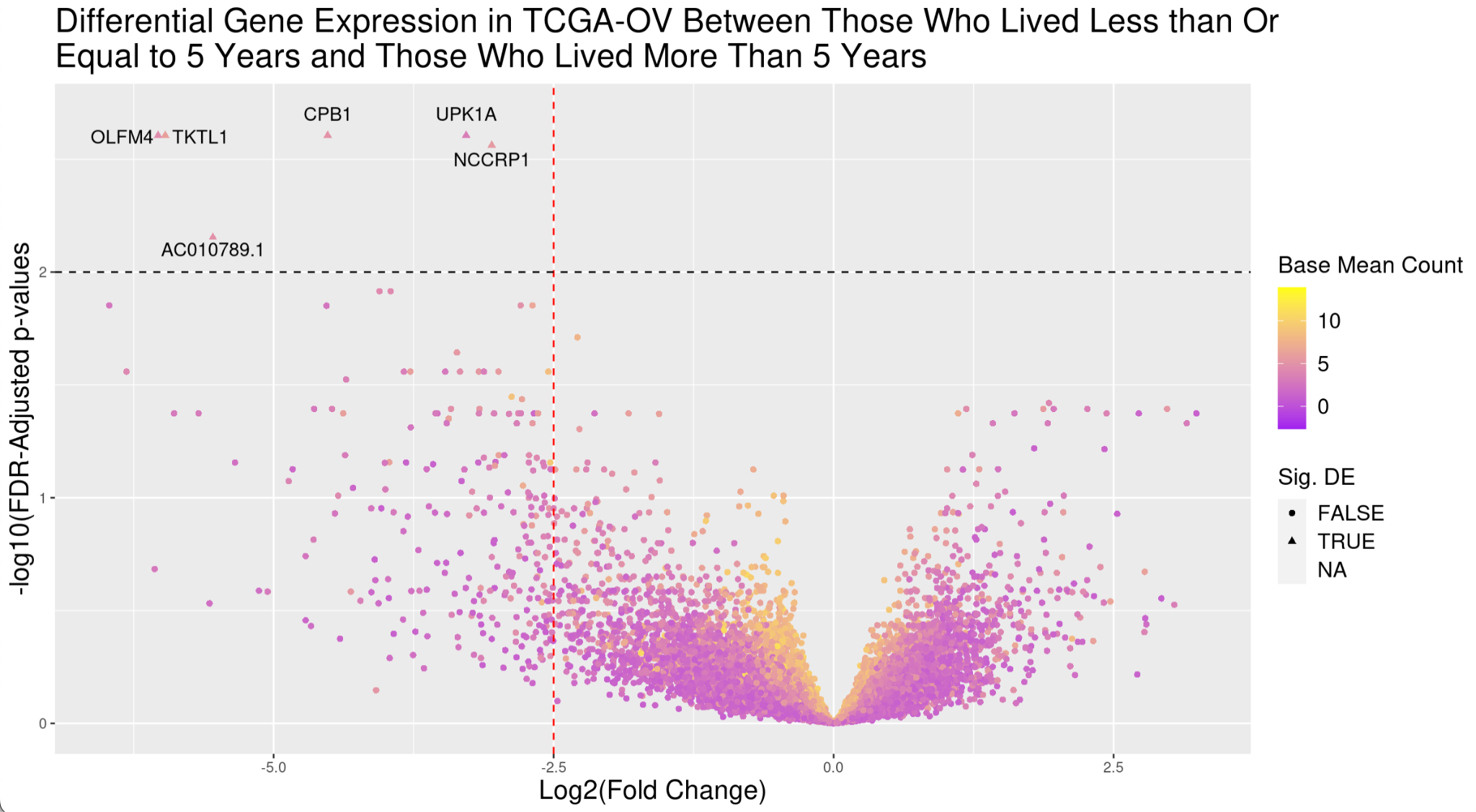
| **Group 1** |  | **Group 2** |
| --- | --- | --- |
| TCGA-61-1736-01B-01R-1568-13 | TCGA-29-1705-01A-01R-1567-13 | TCGA-24-2036-01A-01R-1568-13 |
| TCGA-13-0720-01A-01R-1564-13 | TCGA-23-1027-01A-02R-1564-13 | TCGA-13-0887-01A-01R-1564-13 |
| TCGA-04-1655-01A-01R-1566-13 | TCGA-30-1718-01A-01R-1567-13 | TCGA-13-0884-01B-01R-1565-13 |
| TCGA-13-0801-01A-01R-1564-13 | TCGA-23-2084-01A-02R-1568-13 | TCGA-24-1556-01A-01R-1566-13 |
| TCGA-04-1514-01A-01R-1566-13 | TCGA-10-0931-01A-01R-1564-13 | TCGA-25-1328-01A-01R-1565-13 |
| TCGA-61-2109-01A-01R-1568-13 | TCGA-13-0893-01B-01R-1565-13 | TCGA-04-1350-01A-01R-1565-13 |
| TCGA-13-1477-01A-01R-1565-13 | TCGA-09-2045-01A-01R-1568-13 | TCGA-13-1403-01A-01R-1565-13 |
| TCGA-29-1768-01A-01R-1567-13 | TCGA-13-0765-01A-01R-1564-13 | TCGA-24-1546-01A-01R-1566-13 |
| TCGA-29-1696-01A-01R-1567-13 | TCGA-13-1485-01A-02R-1565-13 | TCGA-29-1688-01A-01R-1566-13 |
| TCGA-23-1113-01A-01R-1564-13 | TCGA-25-1315-01A-01R-1565-13 | TCGA-23-2081-01A-01R-1568-13 |
| TCGA-10-0934-01A-02R-1564-13 | TCGA-24-1557-01A-01R-1566-13 | TCGA-24-1555-01A-01R-1566-13 |
| TCGA-24-1105-01A-01R-1565-13 | TCGA-61-1918-01A-01R-1568-13 |  |
| TCGA-25-1314-01A-01R-1565-13 | TCGA-13-1506-01A-01R-1565-13 |  |
| TCGA-25-2404-01A-01R-1569-13 | TCGA-24-1550-01A-01R-1566-13 |  |
| TCGA-29-1694-01A-01R-1567-13 | TCGA-61-1724-01A-01R-1568-13 |  |
| TCGA-13-1405-01A-01R-1565-13 | TCGA-04-1348-01A-01R-1565-13 |  |
| TCGA-10-0937-01A-02R-1564-13 | TCGA-24-2293-01A-01R-1568-13 |  |
| TCGA-23-1021-01B-01R-1564-13 | TCGA-13-0766-01A-02R-1564-13 |  |
| TCGA-24-1103-01A-01R-1565-13 | TCGA-13-1501-01A-01R-1565-13 |  |
| TCGA-13-0725-01A-01R-1564-13 |  |  |

Principal component analysis was done to see if the two sample groups were similar with the first two principal components. 57.11% of the variation is accounted for by principal component 1 and 44.69% of the variation in the data is accounted for by principal component 2. Based on the scatterplot in figure 1, it is evident that the first two principal components do not identify the two sample groups clearly. Thus, the differential gene expression results that are obtained from these samples are more accurate.



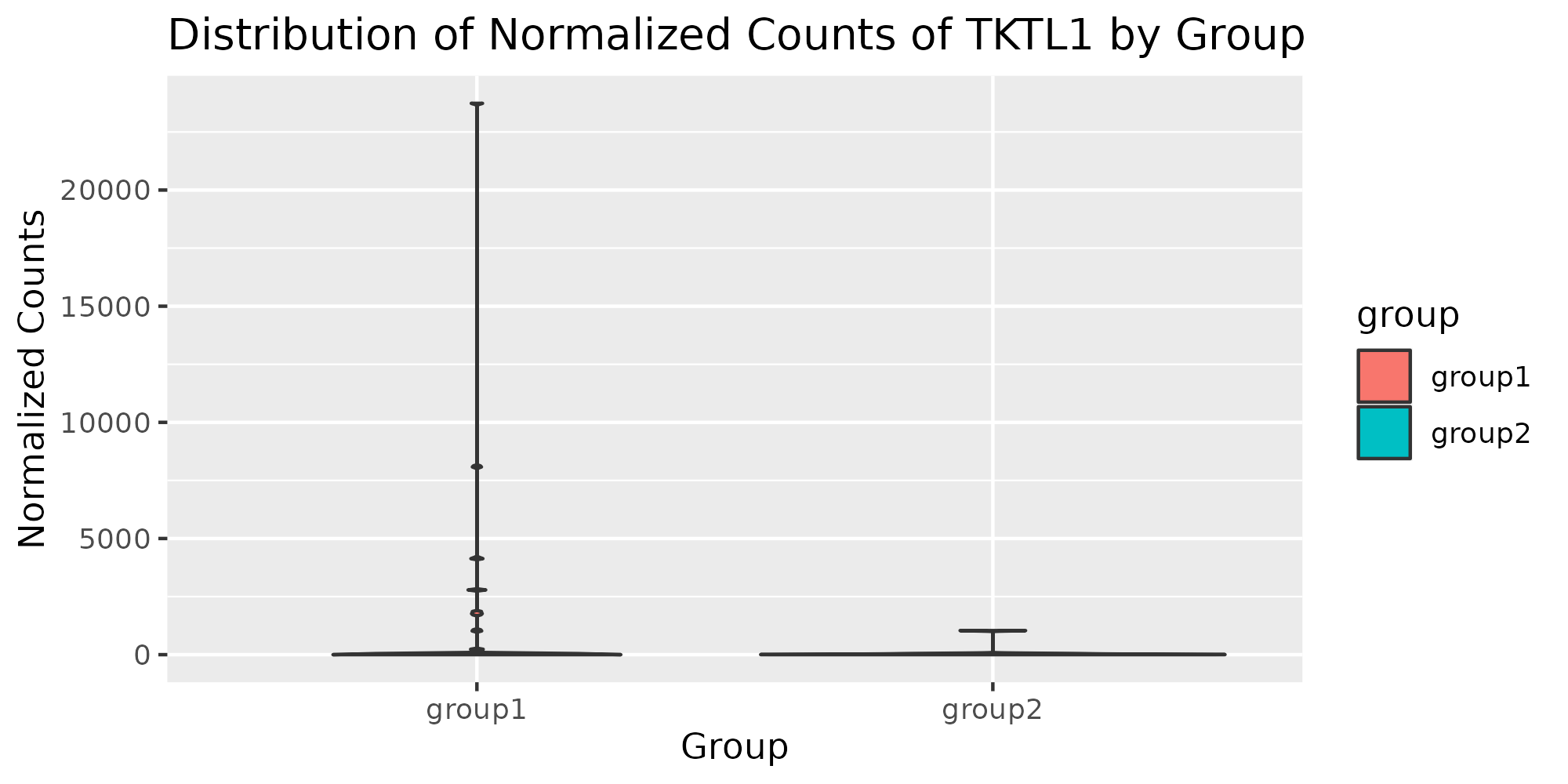
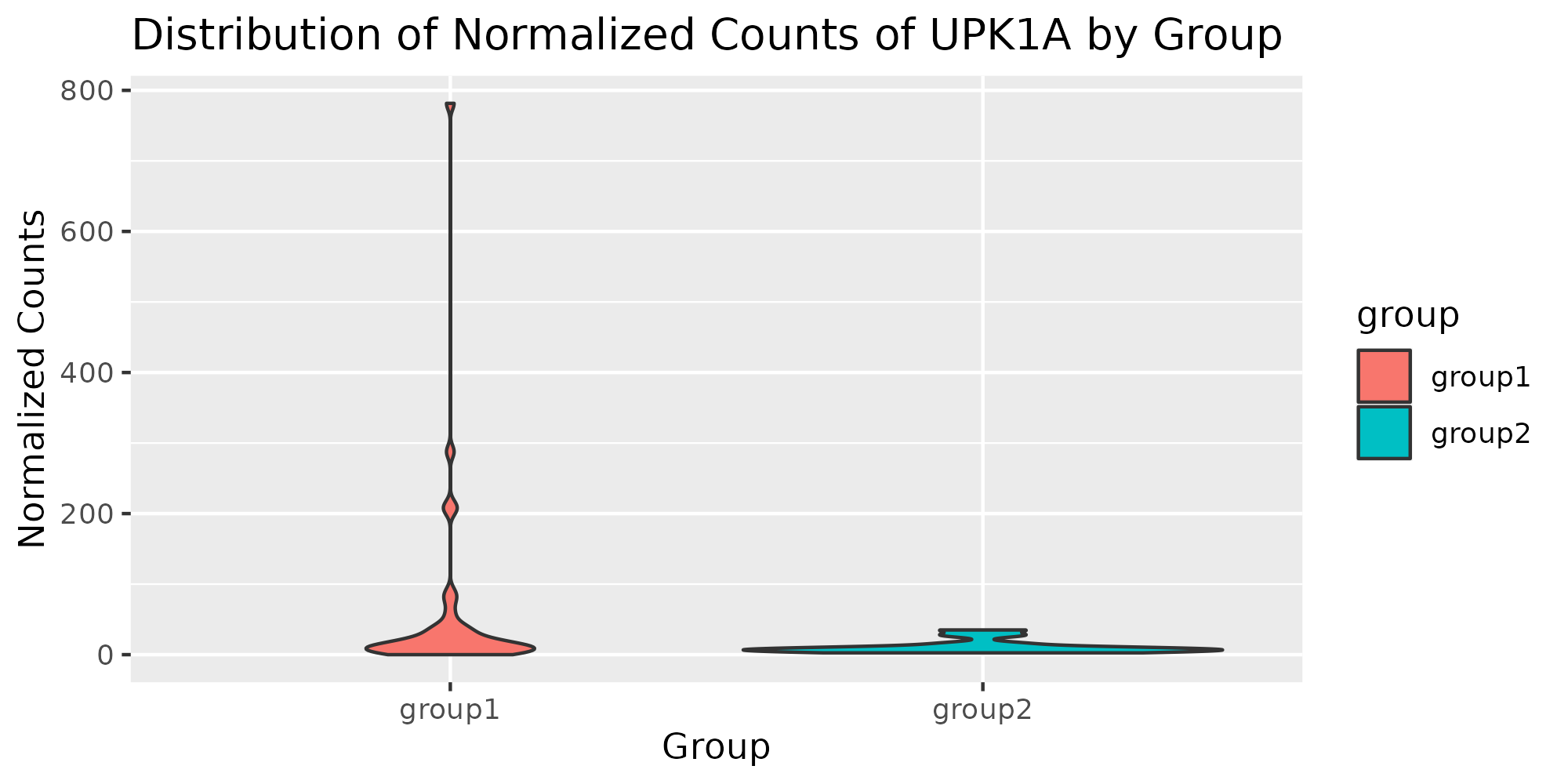
**Figure 1.** This scatterplot shows the datapoint with principal components 1 and 2 on the x and y axes. The data points are colored based on the sample groups. Group 1 is in red and the data points are shaped as circles. Group 2 is in blue and shaped as triangles. 57.11% of the variation is accounted for by principal component 1 and 44.69% of the variation in the data is accounted for by principal component 2.

Differential gene expression analysis was done across 43,452 genes. The adjusted p-value was less than 0.01 for 6 genes: TKTL1, OLFM4, UPK1A, CPB1, NCCRP1, and AC010789.1. These genes can be found in S4. These genes had log 2 fold change values less than -2.5. These results have been plotted in figure 2. These genes were further analyzed in downstream analysis.

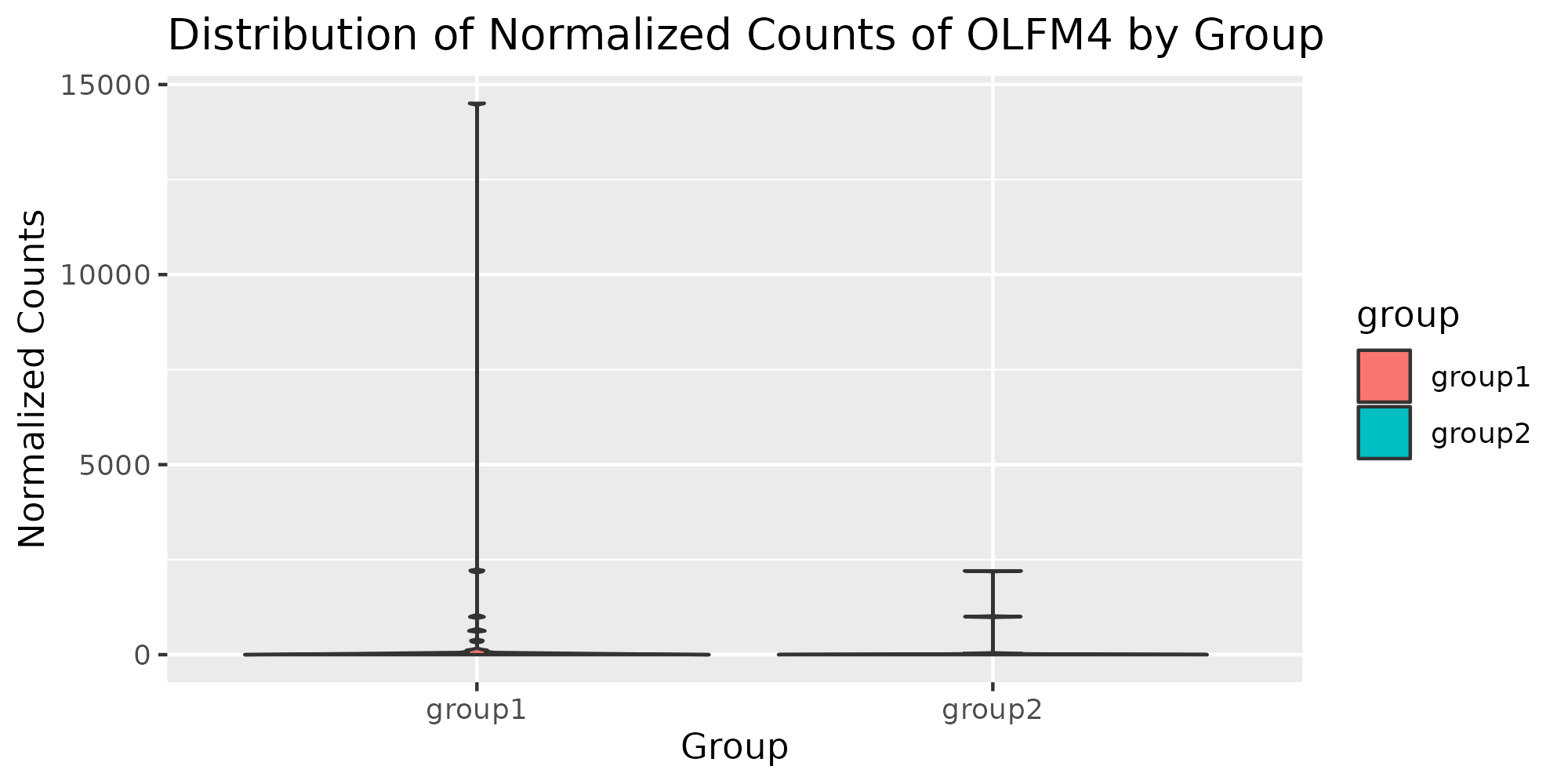


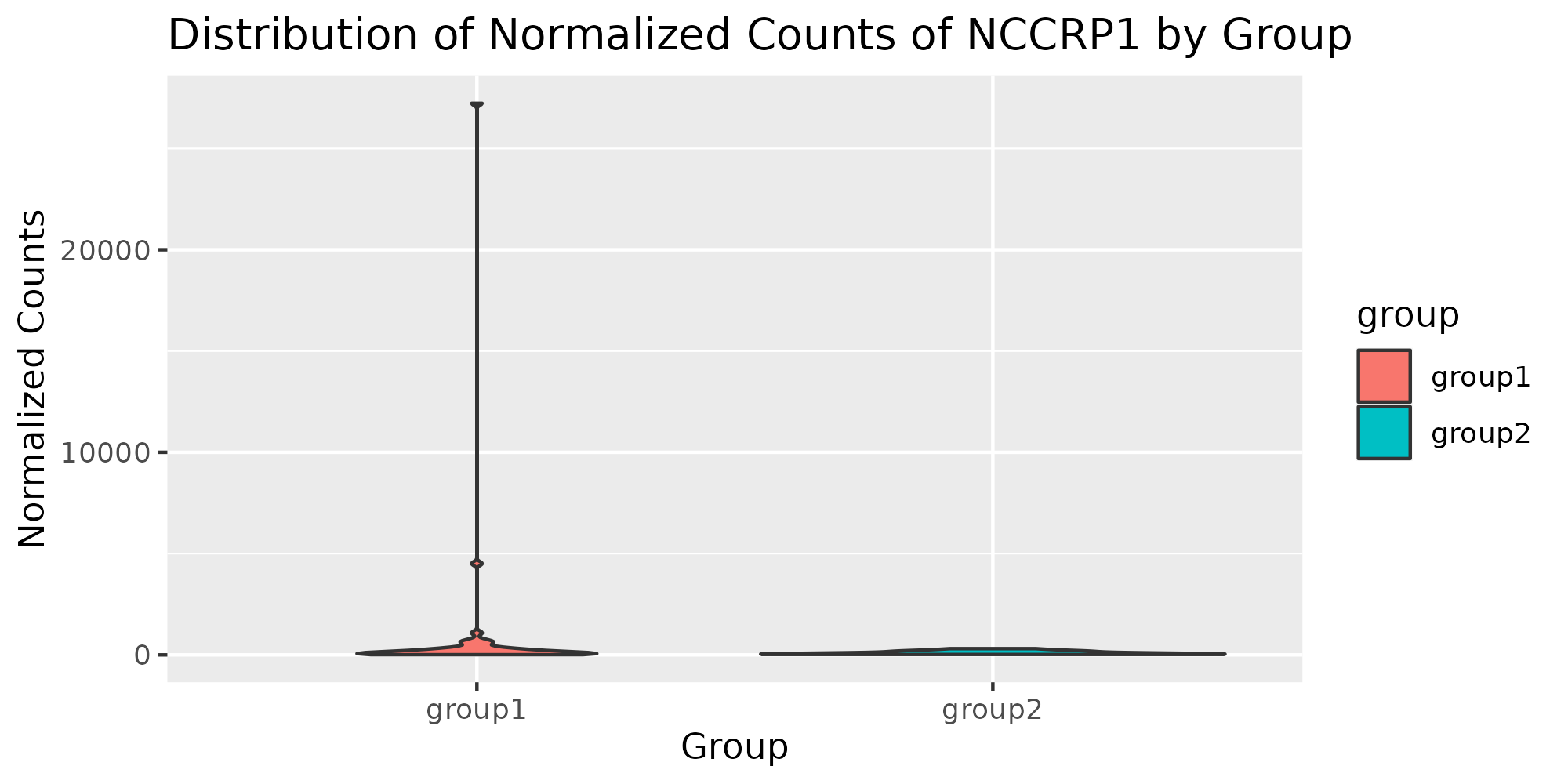
**Figure 2.** This volcano plot showcases the expression of genes in TCGA-OV samples between those who lived for 5 or less years and more than 5 years. The 6 genes that had an adjusted p-value of less than 0.01 are shown as triangles. The mean base count is shown based on the color of the data points. A vertical line in red indicates points that have a log 2 fold change less than -2.5. A horizontal line shows where the adjusted p-value is less than 0.01.

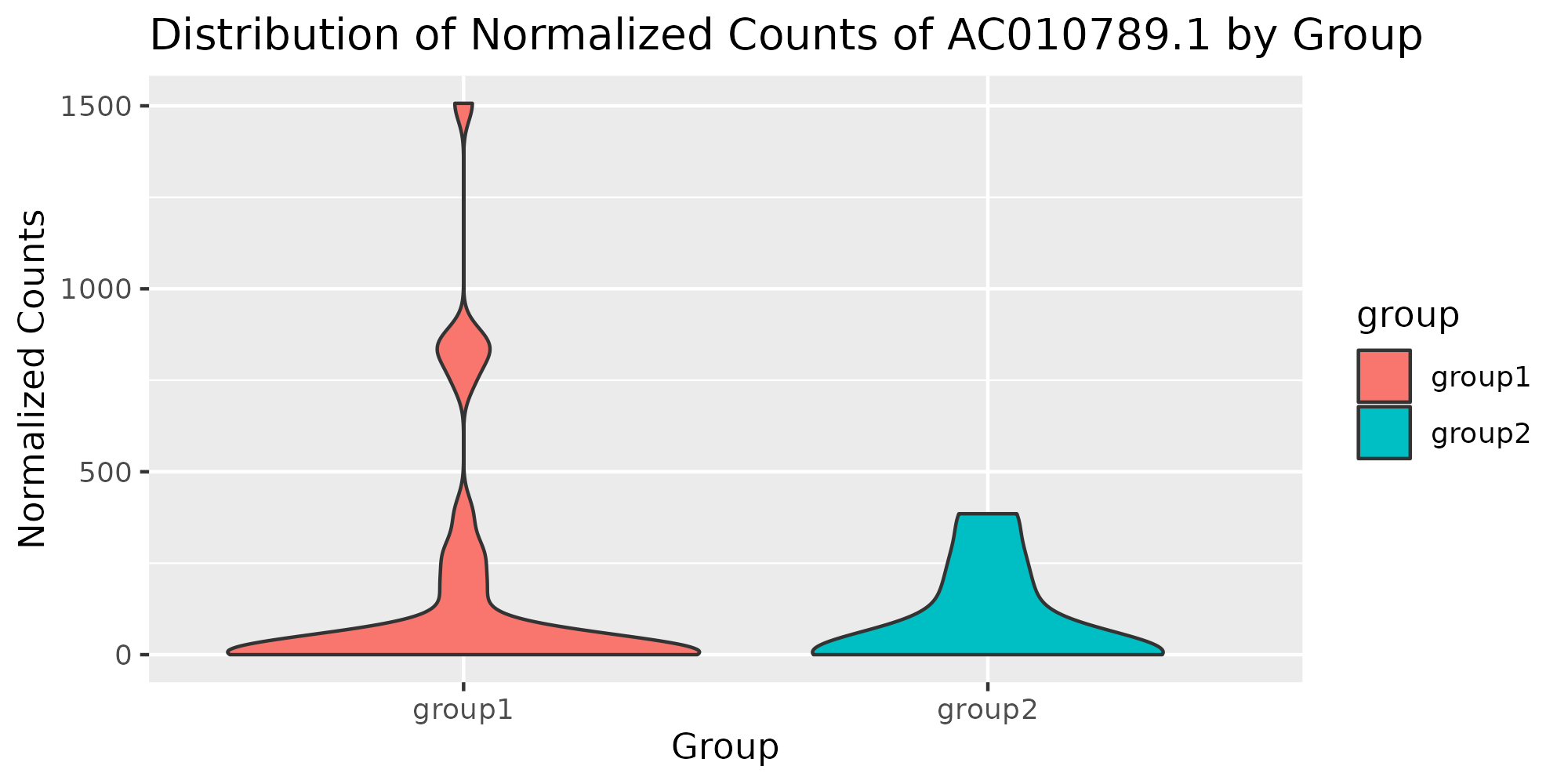
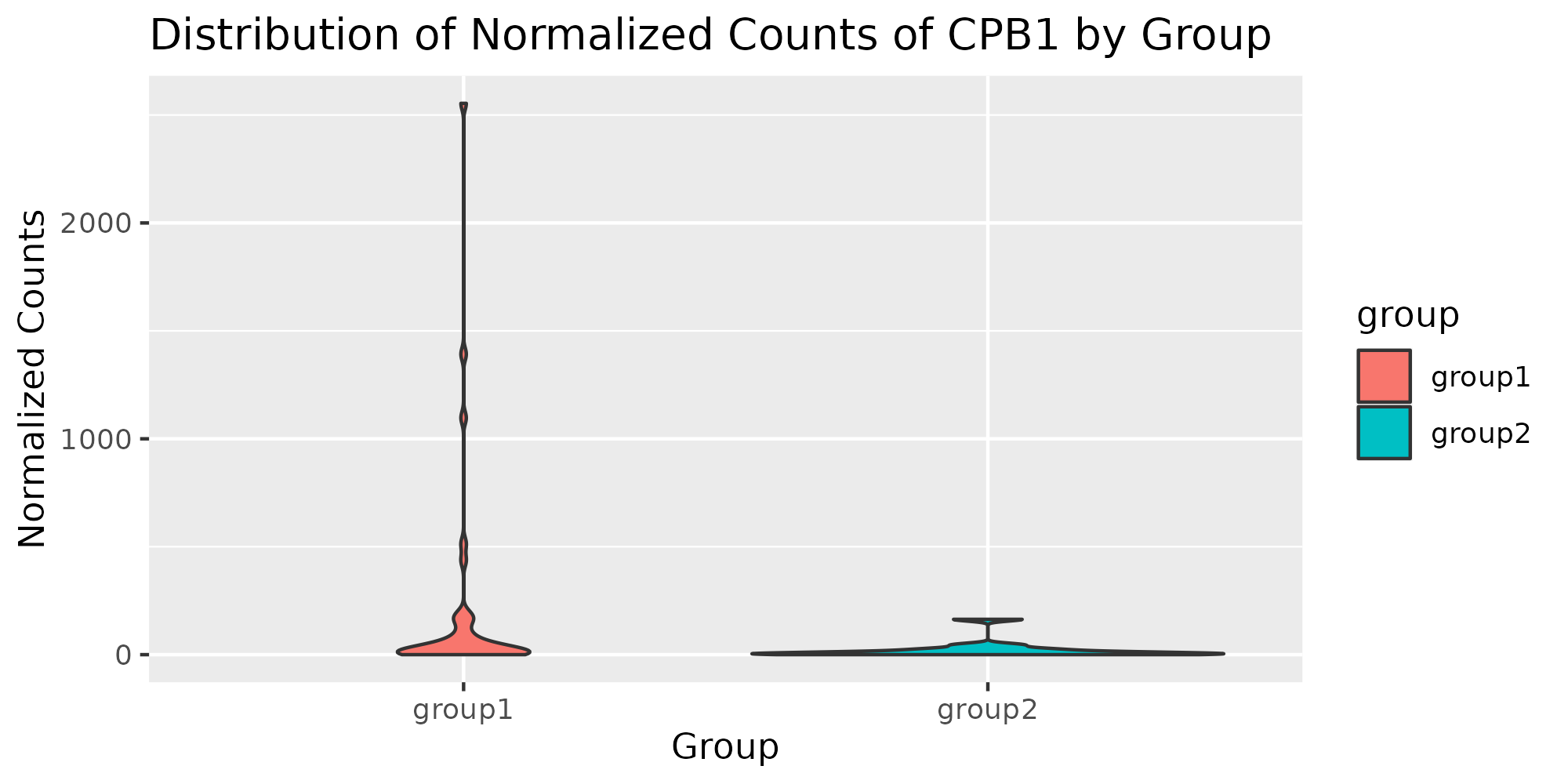
To determine the normalized read counts for each of the significant genes for the groups, violin plots were created. For all the significant genes, some samples in group 1 had greater expression of TKTL1, OLFM4, UPK1A, CPB1, NCCRP1, and AC010789.1 compared to group 2. However, not all samples followed this trend. These trends can be seen in figure 3.



1. **b.**



**c. d.** 



**e. f.**

**Figure 3.** These violin plots show the normalized count of reads for each of the significantly expressed genes. Group 1 is shown on the left of each plot in red and group 2 is shown on the right of each plot in green. These plots are shown for gene **a.** UPK1A **b.** TKTL1 **c.** OLFM4 **d.** NCCRP1 **e.** CPB1 **f.** AC010789.1.

# Discussion

Differential gene expression in women with OV of the ages of 25 to 50, whose vital status was dead, had cancer genus, and had an open data source was done between those who survived for 5 years or less and those who survived for more than 5 years. Women who survived 5 years or less had greater expression in TKTL1, OLFM4, UPK1A, CPB1, NCCRP1, and AC010789.1.

TKTL1, also known as transketolase like 1, is a gene that codes for the transketolase enzyme. Transketolase connects the pentose phosphate pathway to glycolysis. This generates sugar phosphates for nucleotide metabolism and helps produce NADPH. A change in activity of this enzyme has been associated with various types of cancer (Deshpande, 2019). Thus, it is feasible that increased expression of TKTL1 leads to more aggressive OV.

In addition, OLFM4 is a gene that codes for Olfactomedin-4. This is a glycoprotein “that facilitates cell adhesions.” This protein is also an antiapoptotic factor, therefore promoting tumor growth (“Olfm4 olfactomedin 4 [homo sapiens (human)]”, 2011). Unsurprisingly, this gene had increased expression within individuals that survived for 5 years or less.

Furthermore, UPK1A, formally known as Uroplakin 1A, codes for uroplakin proteins, which are involved in mediating signal transduction. This protein also regulates “cell development, activation, growth, and motility.” Interestingly, uroplakin proteins potentially play a role in suppressing tumors (“UPK1A Uroplakin 1a [homo sapiens (human)]”, 2013). Although there is evidence to suggest that the protein coded by UPK1A influences cancer growth, the tumor suppressing nature of the protein makes it unclear why higher expression of this gene in group 1 can be associated with a shorter survival time.

Moreover, CPB1 is a gene that codes for carboxypeptidase B1. This protein is only found in the pancreas. There seems to be no link between acute pancreatitis and OV (“CPB1 carboxypeptidase B1 [homo sapiens (human)]”, 2008).

In addition, NCCRP1 codes for Non-specific cytotoxic cell receptor protein 1 and regulated cell proliferation (“NCCRP1 NCCRP1, F-box associated domain containing [homo sapiens (human)]”, 2022). This gene is considered a biomarker for esophageal cancer, but is indicated as those with lower expression of this gene tend to have a higher occurrence of cancer (Miwa et al., 2017). Thus, it is unclear why those who had a shorter survival time with OV have a higher expression of this gene.

Finally, AC010789.1 is a long non-coding RNA (lncRNA). While the function of lncRNAs is uncertain, they have been associated with tumor progression. This gene has been indicative of other cancers, such as colorectal cancer (Duan et al., 2020). Therefore, higher expression of this gene for those with OV and shorter survival time is logical.

# Conclusion

Differential gene expression in women with OV of the ages of 25 to 50, whose vital status was dead, had cancer genus, and had an open data source was done between those who survived for 5 years or less and those who survived for more than 5 years. Women who survived 5 years or less had greater expression in TKTL1, OLFM4, UPK1A, CPB1, NCCRP1, and AC010789.1. Through analyses, it is clear that higher expression of TKTL1, OLFM4, UPK1A, and AC010789.1 can be linked to ovarian cancer. However, further research needs to be done to determine whether higher expression of CPB1 and NCCRP1 was coincidental with shorter survival time or if these genes play a role in ovarian cancer progression.

# Supplementary Section

1. **R Script:** <https://drive.google.com/file/d/1Bv38LUcxkLOJgR7rXmiPEFEmhr_BNlvY/view?usp=sharing>
2. **Case IDs/Barcodes:**
3. Group 1: <https://drive.google.com/file/d/1G8CZAkP3HtL5QKqgS-y2CdOM_PMmjytv/view?usp=share_link>
4. Group 2: <https://drive.google.com/file/d/12PxMaiaoUneCxURNj9PR5Q_jG93WJRql/view?usp=share_link>
5. **All Genes Analyzed:** <https://drive.google.com/file/d/1WGdq24xUetD1t6x_dHU829CZoR-gjTh7/view?usp=share_link>
6. **Significant Genes Analyzed:** <https://drive.google.com/file/d/1PncFxOzm0sBb2sGiOjgGA0iWhuVgjGW6/view?usp=share_link>

# References

1. The Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. *Nature*, *474*(7353), 609–615. https://doi.org/10.1038/nature10166.
2. Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, 375(12), 1109-1112.
3. R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
4. Posit Team. (2023). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. http://www.posit.co/
5. Bentley, N. (2023a). *FinalProject2023c\_OverviewDocument. Bio321G*. Retrieved November 13, 2023, from https://docs.google.com/document/d/1lp9zjXXNFVJniMlJkTKI06xskzrwEihgRkhHirDCGao/edit?usp=sharingFinalProject2023a\_Overview
6. Colaprico, A., Silva, T. C., Olsen, C., Garofano, L., Cava, C., Garolini, D., ... Noushmehr, H. (2016). TCGAbiolinks: An R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research, 44(8), e71*. https://doi.org/10.1093/nar/gkv1507
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.
8. R Core Team (2023). *\_R: A Language and Environment for Statistical Computing\_*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
9. Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.
10. Deshpande, G. P., Patterton, H.-G., & Faadiel Essop, M. (2019). The human transketolase-like proteins TKTL1 and TKTL2 are bona fide transketolases. *BMC Structural Biology*, *19*(1). https://doi.org/10.1186/s12900-018-0099-y
11. U.S. National Library of Medicine. (2011, March). *Olfm4 olfactomedin 4 [homo sapiens (human)]*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/10562
12. U.S. National Library of Medicine. (2013, July). *UPK1A Uroplakin 1a [homo sapiens (human)]*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/11045
13. U.S. National Library of Medicine. (2008, July). *CPB1 carboxypeptidase B1 [homo sapiens (human)]*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/1360
14. U.S. National Library of Medicine. (2022, April). *NCCRP1 NCCRP1, F-box associated domain containing [homo sapiens (human)]*. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/342897
15. Miwa, T., Kanda, M., Koike, M., Iwata, N., Tanaka, H., Umeda, S., Tanaka, C., Kobayashi, D., Hayashi, M., Yamada, S., Fujii, T., Fujiwara, M., & Kodera, Y. (2017). Identification of NCCRP1 as an epigenetically regulated tumor suppressor and biomarker for malignant phenotypes of squamous cell carcinoma of the esophagus. *Oncology Letters*, *14*(4), 4822–4828. https://doi.org/10.3892/ol.2017.6753
16. Duan, W., Kong, X., Li, J., Li, P., Zhao, Y., Liu, T., Binang, H. B., Wang, Y., Du, L., & Wang, C. (2020). LncRNA AC010789.1 promotes colorectal cancer progression by targeting microrna-432-3p/zeb1 axis and the WNT/β-catenin signaling pathway. *Frontiers in Cell and Developmental Biology*, *8*. https://doi.org/10.3389/fcell.2020.565355