Exploring the Immune Microenvironment and Prognostic Role of BTK in Lung Adenocarcinoma

Kindle Zhang

January 2025

A reproducibility study based on TCGA data mining Practicum Supervisors: Dr.Zhonghua Liu

1 Overview and Student Role

This is a reproducible study. By replicating a published bioinformatics paper [1], I aim to gain a deeper understanding of methods such as DEG analysis and enrichment analysis, particularly in their application to gene-level and cell-level cancer treatment. Through this practicum, I hope to build foundational knowledge in this field and prepare for future research in related areas.

During this practicum, my primary responsibilities involve conducting literature reviews, collecting and organizing relevant data, reproducing the code based on the methodology outlined in the original paper under the guidance of Professor Zhonghua Liu, and ultimately preparing materials for a final presentation and report.

2 Background

Lung adenocarcinoma (LUAD) is the most common subtype of non-small cell lung cancer (NSCLC), particularly prevalent among non-smokers and females. Its development is closely associated with somatic genetic alterations, many of which are actionable targets for therapy. Key driver mutations include EGFR, KRAS, ALK, ROS1, and BRAF, among others [2]. These mutations not only contribute to tumor initiation and progression but also guide the selection of targeted treatments [3]. With the widespread use of next-generation sequencing (NGS), molecular profiling of LUAD has become essential in clinical decision-making, enabling personalized therapy and significantly improving patient outcomes.

3 Methods

3.1 Raw Data

In the original study, the authors used data from 2022, including 551 LUAD cases. In contrast, this study utilized the latest LUAD dataset from TCGA (2024 release). After data cleaning, a total of 19,934 genes were consistently retained across both the counts and TPMs tables, with 589 matched samples.

Transcriptome RNA-seq data of 589 LUAD cases (normal samples, 58 cases; tumor samples, 513 cases) and the corresponding clinical data were downloaded from TCGA database (https://portal.gdc.cancer.go with level 3. The survival information of samples comes from **xena** web.

In addition, the RNA-seq counts data will be downloaded concurrently to facilitate our differential expression (DEG) analysis.

3.2 Generation of Score

Using the ESTIMATE algorithm from the corresponding R package, I employed TPMs data to estimate the proportion of immune and stromal components in the tumor microenvironment (TME) for each sample. As a result, ImmuneScore, StromalScore, and ESTIMATEScore (the sum of immune and stromal scores) were calculated for each sample. A higher ImmuneScore or StromalScore reflects a higher abundance of immune or Stromal components in the tumor microenvironment.

I also estimated tumor purity to support further downstream analyses.

3.3 Survival Analysis KM curve and COX Regression Analysis

In this study, samples were stratified based on their respective scores into high and low groups for ImmuneScore, StromalScore, and ESTIMATEScore. Kaplan–Meier survival curves were then employed to assess whether these scores could serve as reliable indicators for predicting patient survival outcomes. The log rank test was the statistical significance test; $p < 0.05$ was considered significant.

3.4 Generation of DEGs

Two rounds of differential expression gene (DEG) analysis were conducted in this study. In the first round, all samples were divided into high and low ImmuneScore groups to identify DEGs. In the second round, after identifying key genes, the gene BTK was selected for further analysis.

Samples were then grouped based on high and low BTK expression levels, and a second DEG analysis was performed.

DEGs with fold change larger than 1 after transformation of log2 (high-score group/low-score group) and false discovery rate $(FDR) < 0.05$ were considered significant.

3.5 GO and KEGG Enrichment Analysis

GO and KEGG enrichment analyses were performed with the help of the R packages clusterProfiler, enrichplot, and ggplot2.

3.6 Heatmaps

Heatmaps of DEGs were produced by R language with package pheatmap.

3.7 Difference Analysis of Scores With Clinical Stages

Clinicopathological data corresponding to the LUAD samples were obtained from TCGA. Statistical analyses were conducted using R, applying either the Wilcoxon rank-sum test or the Kruskal–Wallis test, depending on the number of clinical stages being compared.

3.8 PPI Network Construction

PPI network was constructed by STRING database. The gene used in this part is the intersection genes got from DEGs in Immune Score group and Stromal score group. Nodes with confidence of interactive relationship larger than 0.95 were used for building network.

3.9 Gene Set Enrichment Analysis

GSEA analysis was performed using R packages released by the Broad Institute. The reference gene sets used for enrichment analysis, including the Hallmark and C7 collections, were downloaded from the Molecular Signatures Database (MSigDB).

In contrast to the GO and KEGG enrichment analyses, which were based on selected DEGs, GSEA utilized the entire transcriptome of all tumor samples. And only gene sets with NOM $p < 0.05$ and FDR $q < 0.06$ were considered as significant.

3.10 Cibersort

I applied CIBERSORT to estimate the relative proportions of 22 immune cell types in tumor samples and explored their association with BTK gene expression. A rainbow plot was used to visualize the immune cell composition across samples, while correlation plots illustrated the relationships between BTK expression and specific immune cell types. To further assess these associations, we performed Spearman correlation tests and conducted group-wise difference analyses to identify statistically significant differences in immune cell proportions between high and low BTK expression groups.

4 Results

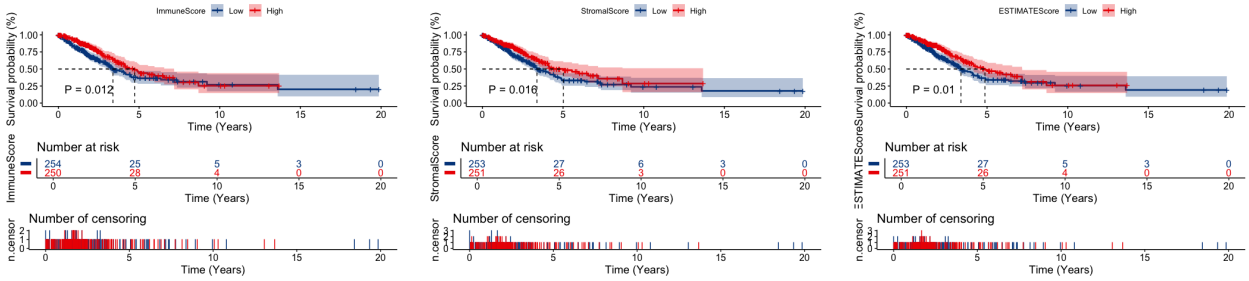
4.1 general analysis process

To investigate the composition of the tumor microenvironment (TME) in LUAD, transcriptome RNA-seq data from 589 cases were downloaded from the TCGA database. The proportions of tumor-infiltrating immune cells (TICs) and the levels of immune and stromal components were estimated using the CIBERSORT and ESTIMATE algorithms. Differentially expressed genes (DEGs) associated with ImmuneScore and StromalScore were identified and subjected to protein–protein interaction (PPI) network construction and univariate COX regression analysis. Key genes were selected based on the intersection of core nodes from the PPI network and the top significant variables from the COX analysis, resulting in the identification of BTK, CD19, F2 and CD79A. BTK was prioritized for subsequent analyses, including survival analysis, correlation with clinicopathological features, multivariate COX modeling, Gene Set Enrichment Analysis (GSEA), and evaluation of its association with TIC profiles.

Finally, I packaged the code into an R package, allowing the cancer type and the selected hub gene (which was BTK in the original study) to be specified as input parameters, in order to improve efficiency and scalability.

4.2 Scores Were Correlated With the Survival Rate

To assess the relationship between immune/stromal components and survival outcomes, Kaplan–Meier survival analyses were performed for ImmuneScore, StromalScore, and ESTIMATEScore. Higher ImmuneScore or StromalScore indicated greater infiltration of immune or stromal cells within the TME, while ESTIMATEScore reflected their combined abundance. A score and tumor purity table for all 589 LUAD samples was generated after this procedure. However, only 513 samples with a tumor were used to make a survival analysis.



(a) Immune Score (b) Stromal Score (c) Estimate Score

Figure 1: Kaplan-Meier survival curve

As shown in Figure 1, a higher ImmuneScore was associated with better overall survival. Although StromalScore also showed a significant correlation (Figure 1b), its log rank test p-value is larger than ImmunerScore's, suggesting that immune components may be more predictive of LUAD prognosis.

4.3 Scores Were Correlated With the Clinic–Pathological Staging of Patients

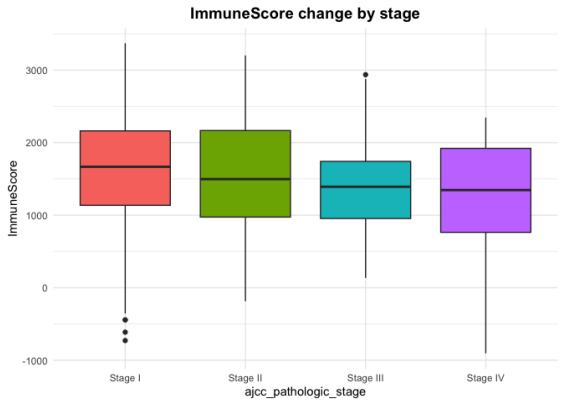
As the previous section demonstrated that the three scores could serve as reliable indicators for predicting patient survival, this part of the study aimed to explore how the ESTIMATEScore varies across different clinical stages and pathological subtypes by integrating it with clinicopathological data.

The T denotes the size and extent of invasion of the primary tumor, the N denotes the regional lymph nodes and the M denotes the distant metastasis. These results suggested that the ratio of immune and stromal components was associated with the progress of LUAD, such as invasion and metastasis. According to the Figure 2, as tumor progression advanced or tumor size increased, ImmuneScore showed a decreasing trend across samples. A Kruskal–Wallis test was conducted to assess differences in ImmuneScore among tumor stages, revealing a statistically significant difference ($p - value < 0.05$), suggesting that immune infiltration may diminish as the tumor develops.

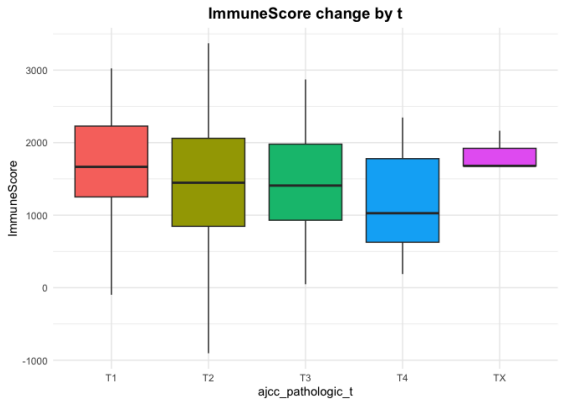
Similar plots with StromalScore and EstimateScore are showed in Appendix.

4.4 Identification of DEGs Common to ImmuneScore and StromalScore

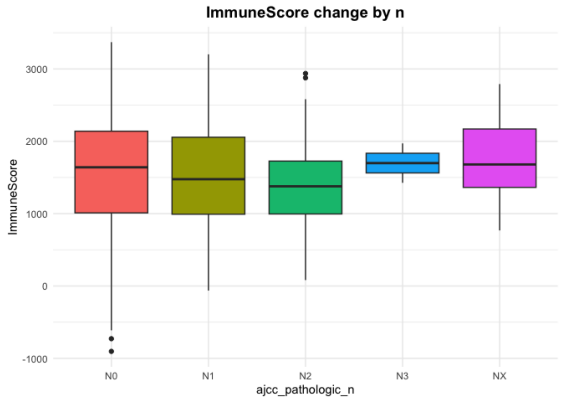
To ascertain the exact alterations of gene profile in TME regarding immune and stromal components, the comparison analysis between high-score and low-score samples were carried out. Genes were considered differentially expressed if they met the criteria of adjusted p-value $(padj) < 0.05$ and false discovery rate $(FDR) < 0.06$. Only genes satisfying both thresholds were retained for



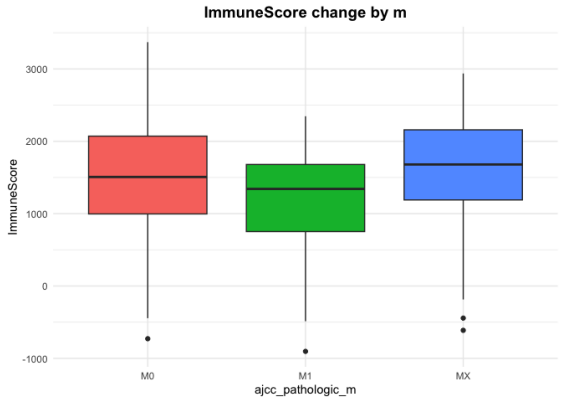
(a) Cancer Stage



(b) T



(c) N

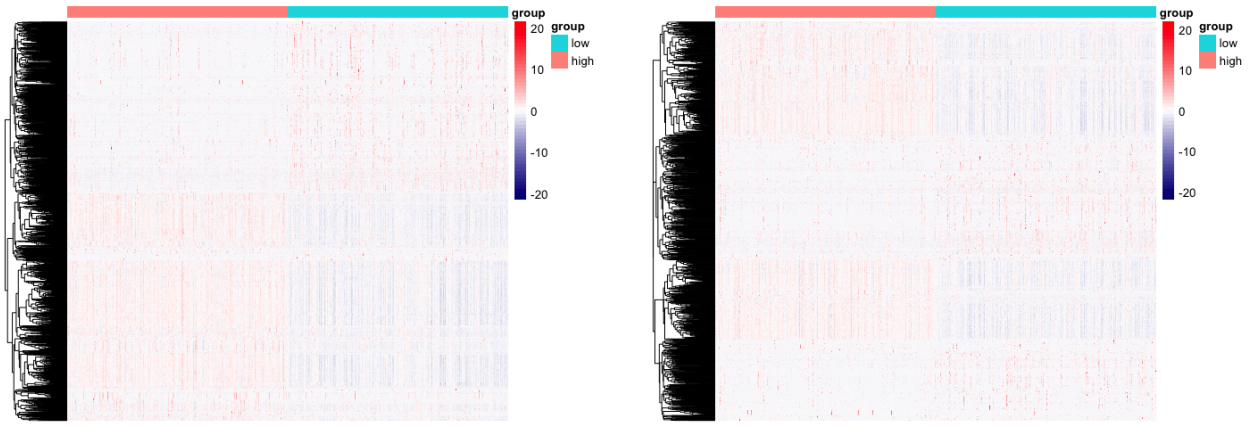


(d) M

Figure 2: ImmuneScore status among different groups

downstream DEG analysis.

Two heatmaps Figure 3 were generated to visualize differential gene expression between groups. The left half represents the high-score group, while the right half represents the low-score group. The color of each dot reflects the corresponding log2 fold change (log2FC) value: red indicates up-regulation (higher log2FC), whereas blue indicates downregulation (lower log2FC). The intensity of the color corresponds to the magnitude of expression change.



(a) Immune Score (b) Estimate Score

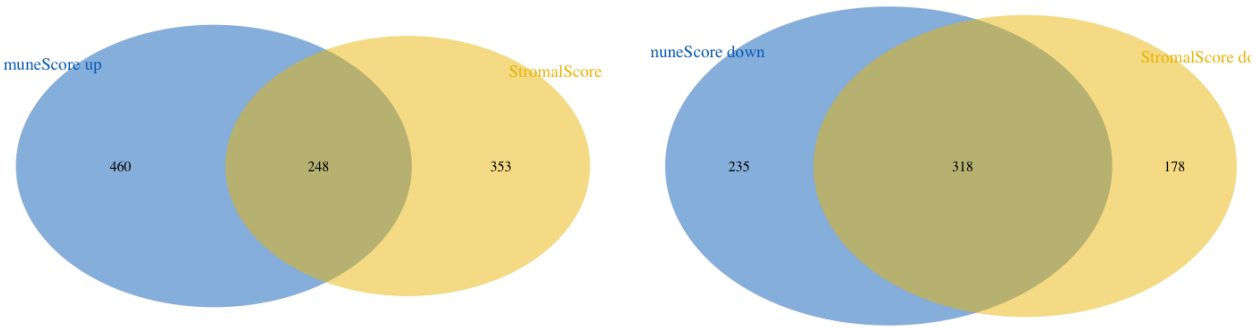
Figure 3: DEGs' heatmap for Score

Following DEG analysis of 19,934 genes, only a small subset was retained for further investigation, facilitating more focused downstream analysis. Differentially expressed genes (DEGs) were categorized based on their log2 fold change: genes with $log2FC > 1$ were assigned to the upregulated group, while those with $log2FC < -1$ were assigned to the downregulated group. For the ImmuneScore-based DEGs, a total of (460 + 248 = 708) genes were upregulated, and (235 + 318 = 553) were downregulated. For the StromalScore-based DEGs, (253 + 248 = 501) genes were upregulated, and (318 + 178 = 496) were downregulated. 248 + 318 These DEGs (total 566 genes) were possibly determinate factors for the status of TME.

Figure 4 shows the venn diagram. These 566 genes were ultimately identified as the key differentially expressed genes of interest and served as the primary focus of this study.

4.5 Enrichment Analysis by GO and KEGG

Based on the 566 DEGs, functional enrichment analysis was performed. GO functional enrichment and KEGG pathway analysis each highlighted distinct biological insights: GO focused on gene functions, while KEGG emphasized pathway-level organization. As shown in Figure 5, the GO analysis revealed that these DEGs were predominantly enriched in immune-related biological



(a) up group (b) down group

Figure 4: DEG both in Immune and Stromal Score

processes, such as immune response–regulating cell surface receptor signaling pathway, indicating a strong involvement of the immune system in the underlying biological mechanisms.

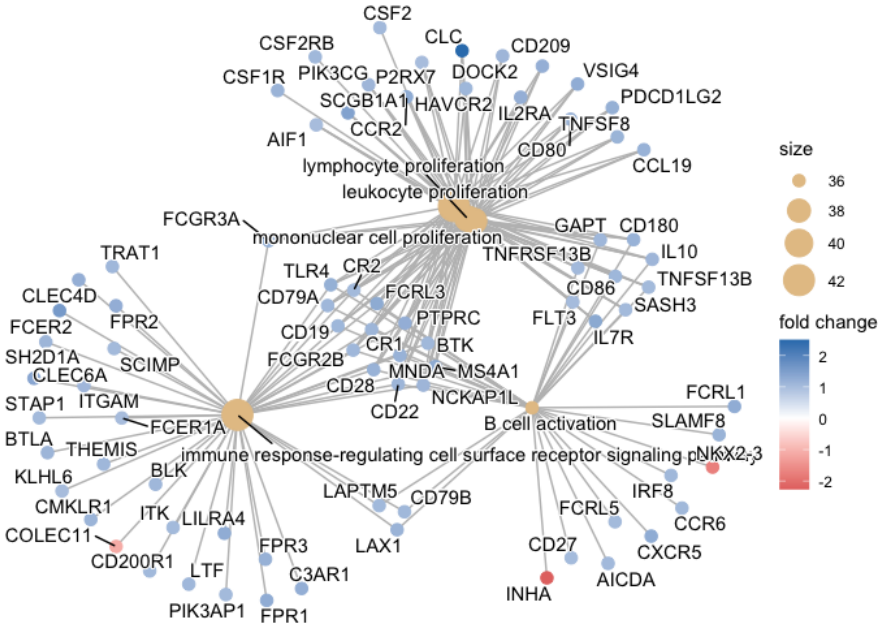


Figure 5: Go function result

These findings suggest that the DEGs identified based on ImmuneScore and StromalScore are indeed involved in modulating the immune system to a certain extent.

4.6 Intersection Analysis of PPI Network and Univariate COX Regression

To further refine the analysis, the initial set of 566 DEGs was filtered down, as investigating all 566 genes simultaneously would be time-consuming and resource-intensive.

Two complementary approaches were applied: protein–protein interaction (PPI) network analysis and a univariate Cox regression model. The intersection genes identified by both methods were selected as the final genes of interest for downstream analysis.

Figure 6 shows the top 30 gene ranked by their number of nodes. Figure 7 shows the result of cox univariate model.

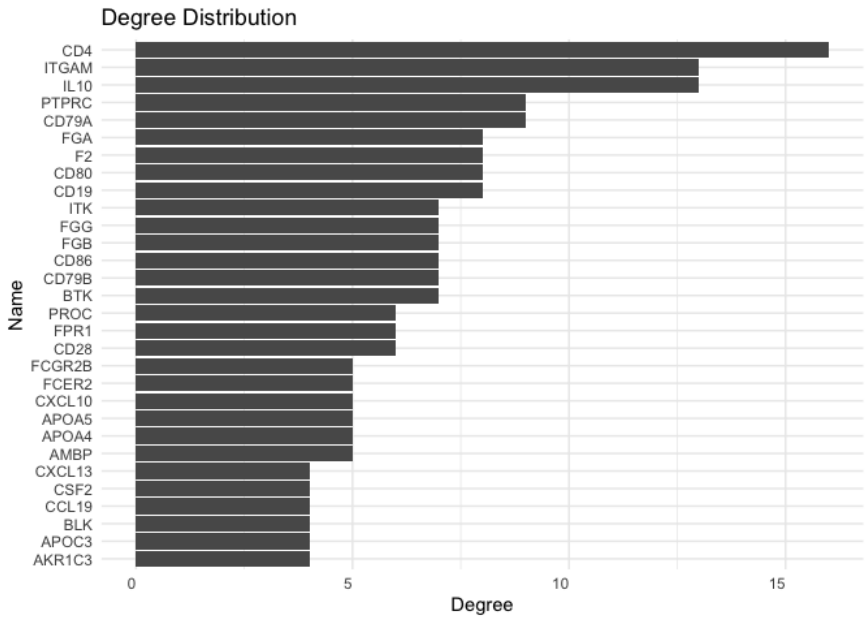


Figure 6: PPI result

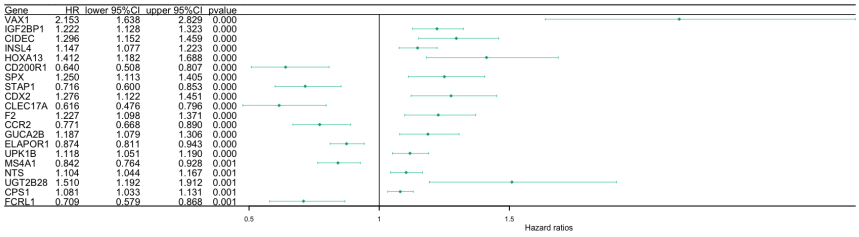


Figure 7: Cox univariate model

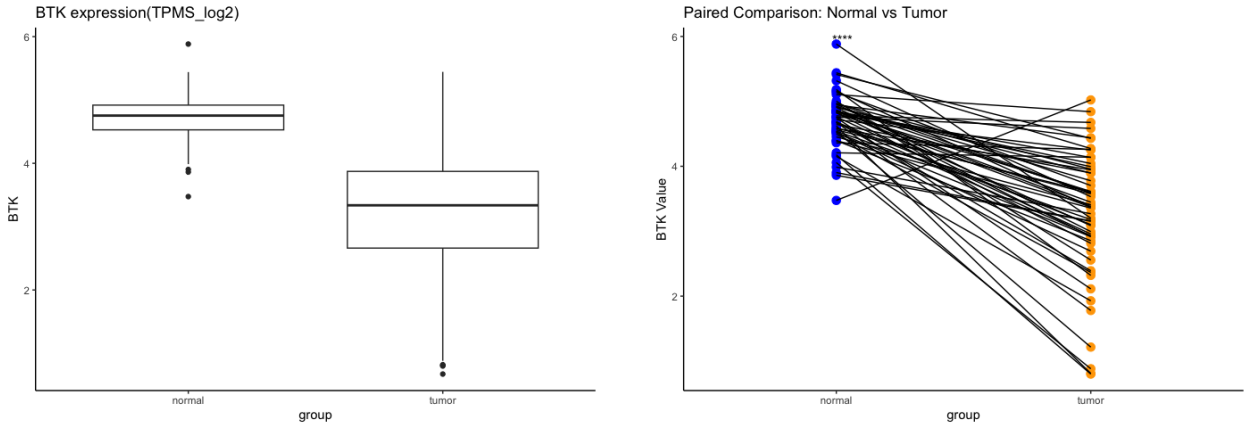
According to the results shown in the figure, only four genes—F2, CD19, BTK, and CD79A—were identified as playing central roles in the PPI network and exhibiting significant associations with

survival in the univariate Cox regression analysis. These genes were thus selected as key candidates for further investigation.

This study primarily focused on exploring the role of the BTK gene. However, the author also developed a corresponding R package, enabling users to investigate similar patterns across other cancer types or genes of interest.

4.7 The Correlation of BTK Expression With the Survival and Classification of TNM Stages in LUAD Patients

After identifying the genes of interest, a series of downstream analyses were conducted. These included a box plot (Figure 8a) to visualize expression differences, a paired box plot (Figure 8b) for matched comparisons, and a Kaplan–Meier survival curve (Figure 9) to evaluate the prognostic significance of the selected genes.



(a) box plot

(b) paired box plot

Figure 8: BTK expression in tumor and normal group

These results suggest that patients with high BTK expression tend to have a better overall survival, highlighting its potential as a favorable prognostic biomarker.

Consistent findings were also observed when integrating clinical and pathological data, further supporting the association between BTK expression and improved survival. The corresponding results are presented in the appendix.

4.8 BTK Had Potential to Be an Indicator of TME Modulation

A second round of DEG analysis was conducted, this time grouping all tumor samples based on BTK expression levels, using the median BTK expression as the cutoff to define high and low expression groups. This grouping yielded a new set of DEGs. Subsequently, Gene Set Enrichment

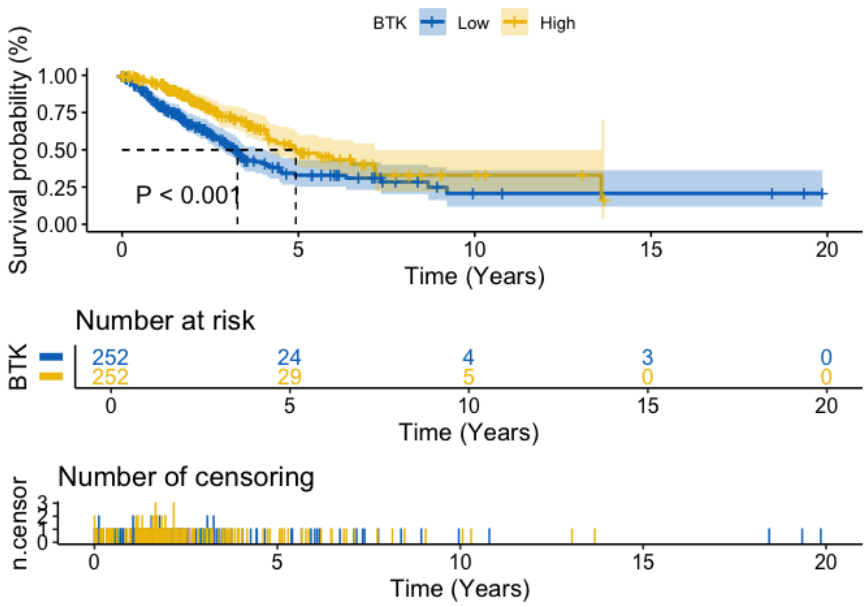
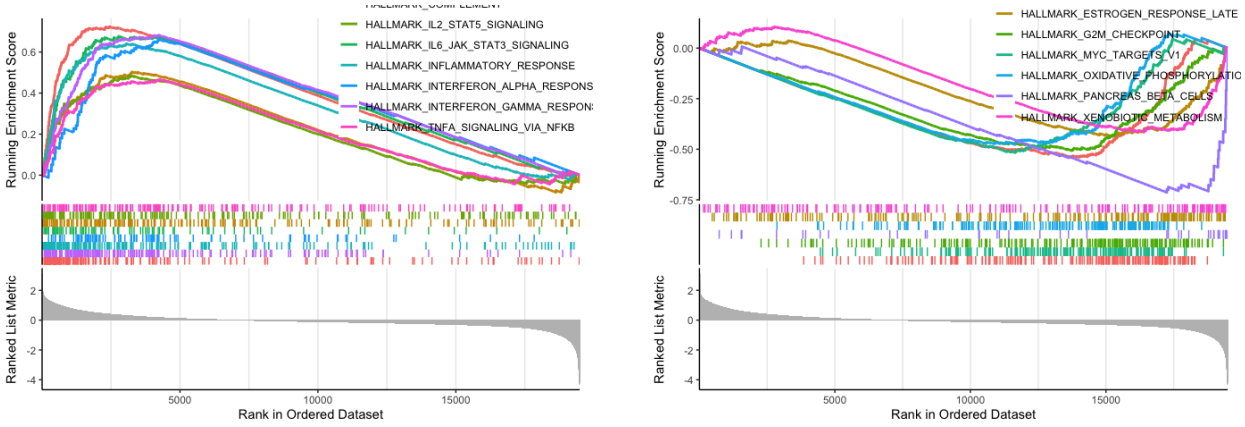


Figure 9: BTK expression Cox model

Analysis (GSEA) was performed on these DEGs. Figure 10 illustrates the gene enrichment results under the HALLMARK gene set, showing the distinct pathways enriched in the BTK high and BTK low expression groups.



(a) hall set BTK high

(b) hall set BTK high

Figure 10: GSEA result in high and low BTK group

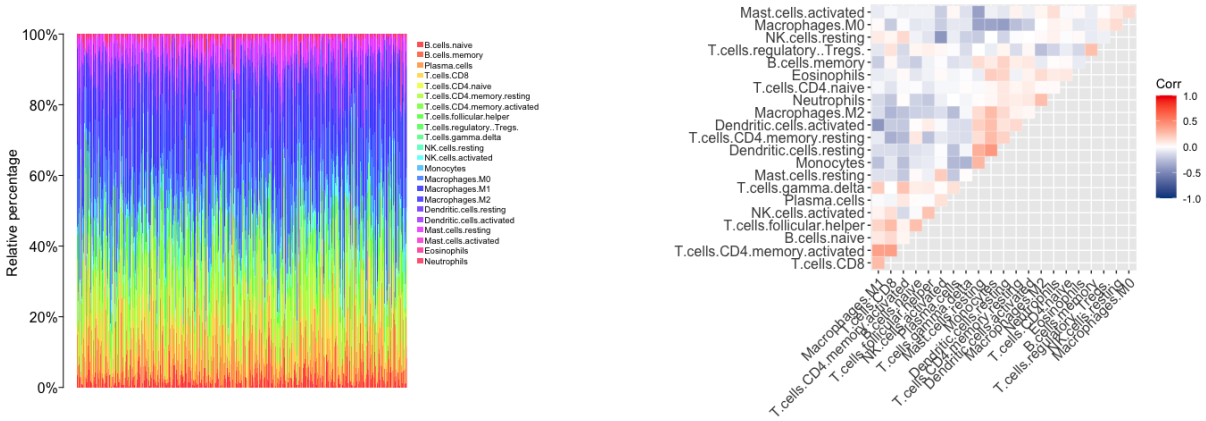
It was observed that in samples with high BTK expression, the enriched DEGs were predominantly involved in immune-related functions, such as the interferon response and inflammatory response. In contrast, samples in the low BTK expression group showed enrichment in metabolism-related pathways, including processes such as pancreatic beta cell function.

The GSEA results based on the C7 (immunologic signatures) gene set are presented in the

appendix.

4.9 Cibersort on TICs

There are 22 distinct types of tumor-infiltrating immune cells (TICs) in the human body. To explore the relationship between BTK expression and the immune microenvironment from a different perspective, the study investigated the association between BTK and the composition of TICs. Figure 11a presents a rainbow plot illustrating the proportion of each immune cell type across all tumor samples, while Figure 11b shows a correlation plot depicting the interrelationships among the proportions of different TICs.



(a) TICs proportion status

(b) TICs proportion correlation

Figure 11: Rainbow plot and correlation plot

4.10 Get Intersection Gene from Difference Test and Correlation Test

To identify which TICs are significantly associated with BTK expression, both a difference test and Spearman correlation analysis were performed. Figure 12 illustrates the results of the difference test, highlighting immune cell types with significantly different proportions between high and low BTK expression groups. For example, the proportion of naive B cells was significantly lower in the high BTK expression group. This relationship appears to be statistically significant, as indicated by the presence of two asterisks ($p < 0.01$) in the plot.

Figure 13 specifically depicts the positive correlation between the proportion of B.cells.naive and BTK expression, suggesting a potential immunological link between BTK activity and B cell infiltration. The coefficient of correlation is -0.14 which means there is a significant negative correlation between BTK expression and naive B cells' proportion. ($p < 0.01$)

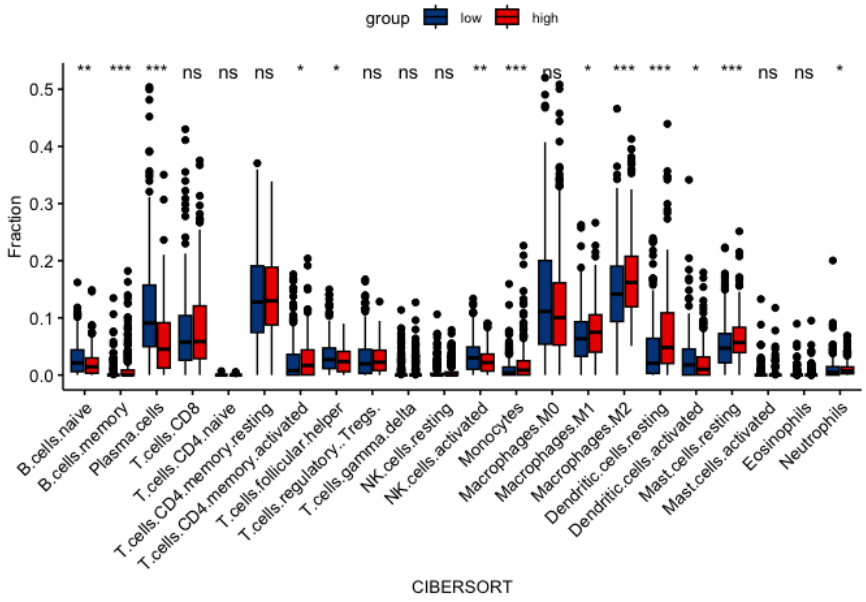


Figure 12: difference test on BTK and TICs

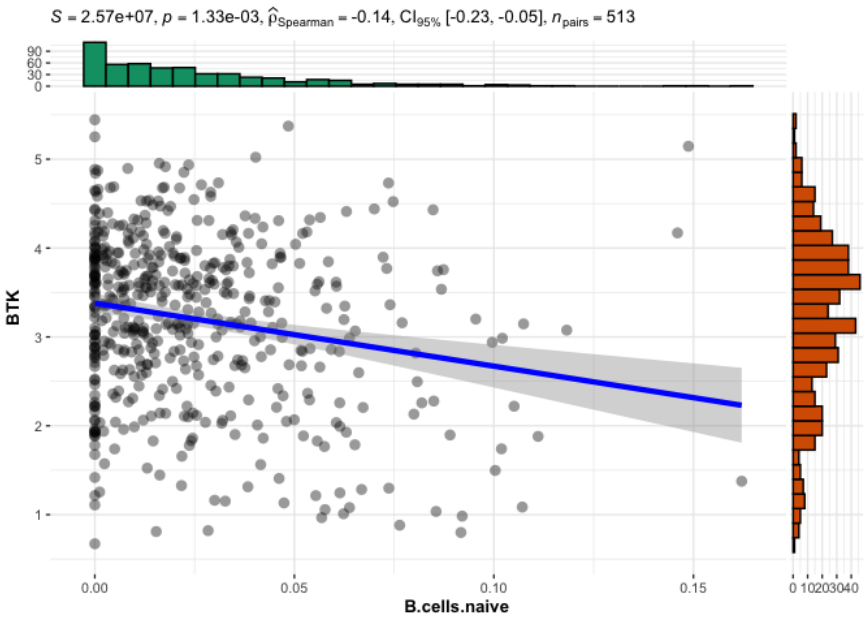


Figure 13: spearman correlation test for B.cells.naive

Based on these analyses, we identified the intersection TICs that showed significant associations with BTK expression in both the difference test and the Spearman correlation analysis.

5 Conclusions/Discussion

In summary, BTK expression appears to be a promising prognostic biomarker for LUAD. Its expression level correlates with immune-related features such as the ImmuneScore and the proportion of tumor-infiltrating immune cells (TICs), both of which are linked to patient survival.

These findings suggest that BTK may reflect the immune landscape of the tumor microenvironment and could help predict clinical outcomes. While other genes, such as F2, also show potential prognostic relevance, BTK stands out and warrants further investigation.

Importantly, this raises a broader question worth exploring in future studies: Can BTK serve not only as a prognostic marker, but also as an indicator of tumor microenvironment (TME) status and potential responsiveness to immunotherapy?

6 Reference

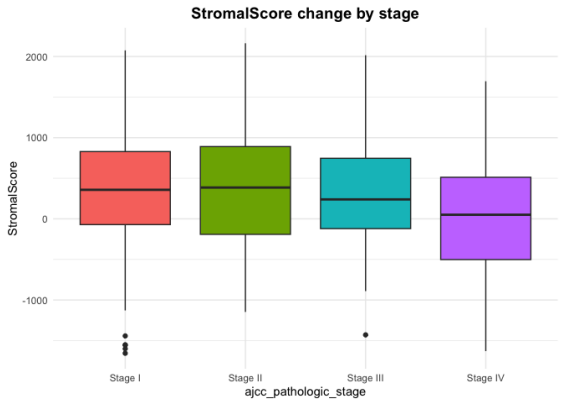
References

[1] Chen, L., Liu, D., & Lv, Z. (2021). BTK has potential to be a prognostic factor for lung adenocarcinoma and an indicator for tumor microenvironment remodeling: A study based on TCGA data mining. *Frontiers in Cell and Developmental Biology*, 9, 668592. https://doi.org/10.3389/fcell.2021.668592

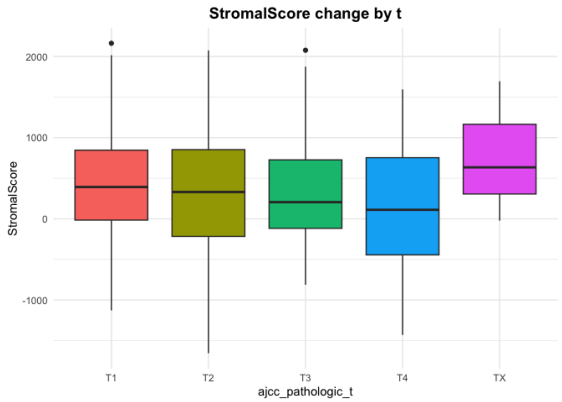
[2] Herbst, R. S., Morgensztern, D., & Boshoff, C. (2018). The biology and management of non-small cell lung cancer. *Nature*, 553(7689), 446–454. https://doi.org/10.1038/nature25183

[3] Hanna, N. H., Robinson, A. G., Temin, S., et al. (2021). Therapy for Stage IV Non–Small-Cell Lung Cancer Without Driver Alterations: ASCO and OH (CCO) Joint Guideline Update. *Journal of Clinical Oncology*, 39(9), 1040–1091. https://doi.org/10.1200/JCO.20.03158

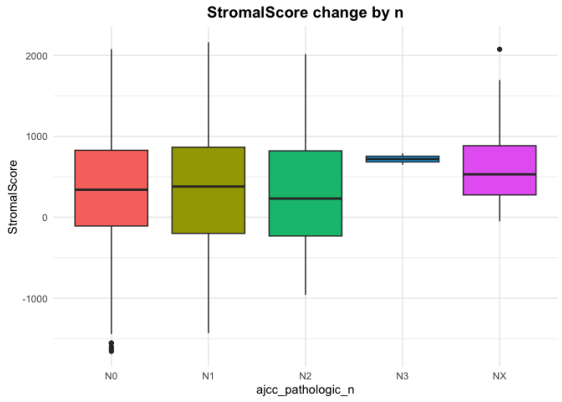
7 Appendix



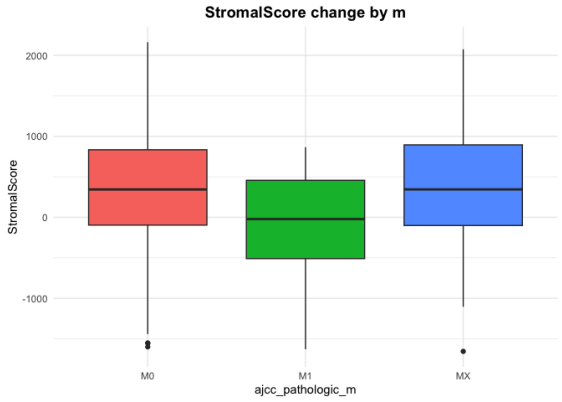
(a) Cancer Stage



(b) T

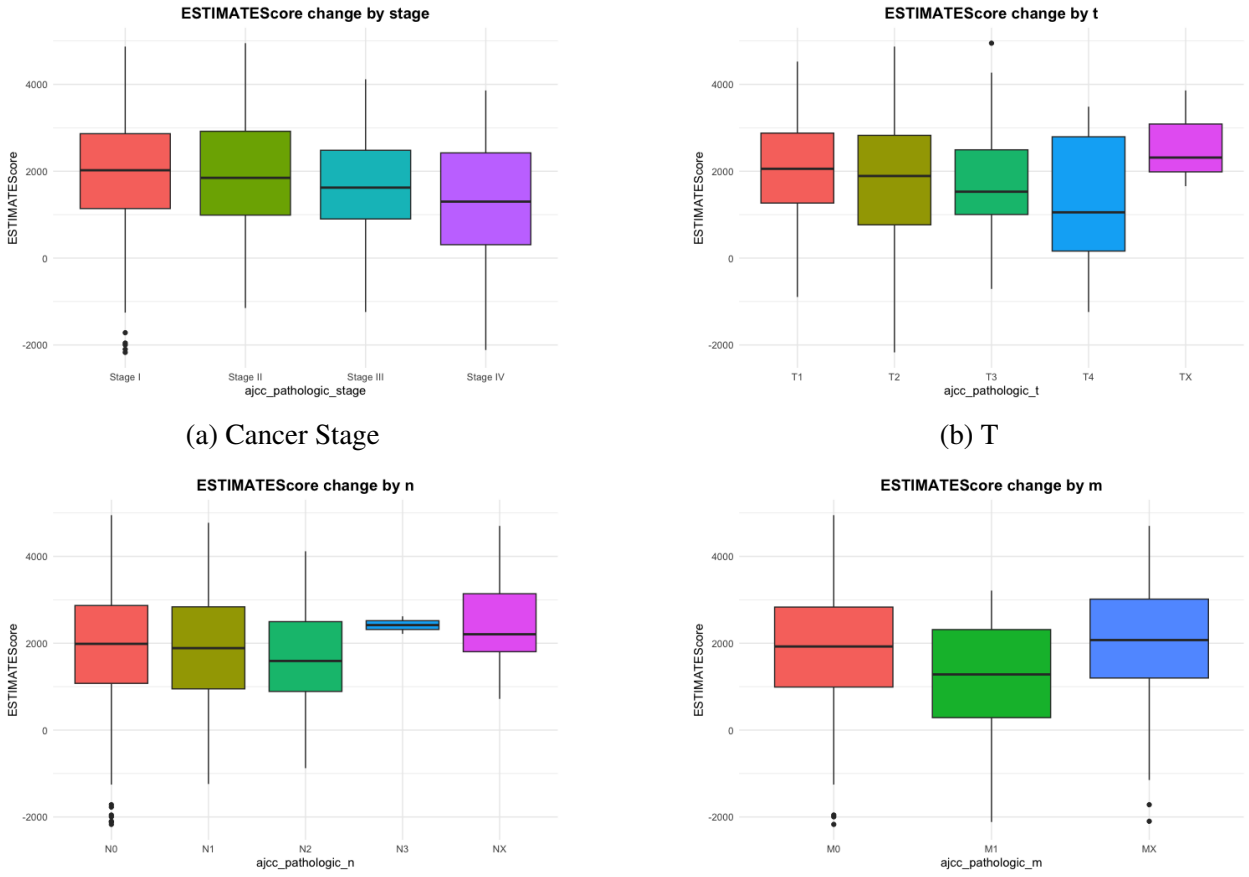


(c) N



(d) M

Figure 14: StromalScore status among different groups



(a) Cancer Stage (b) T

(c) N (d) M

Figure 15: EstimateScore status among different groups

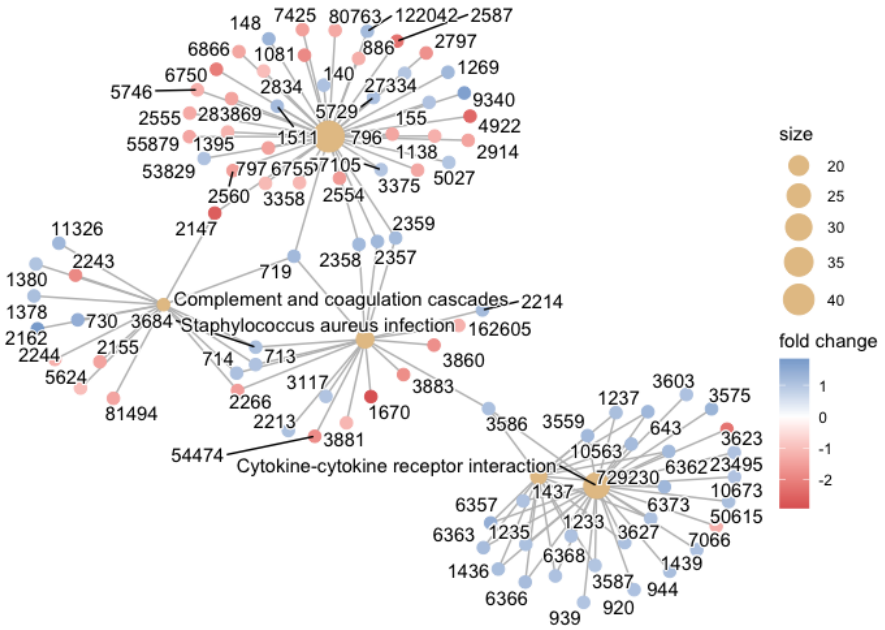
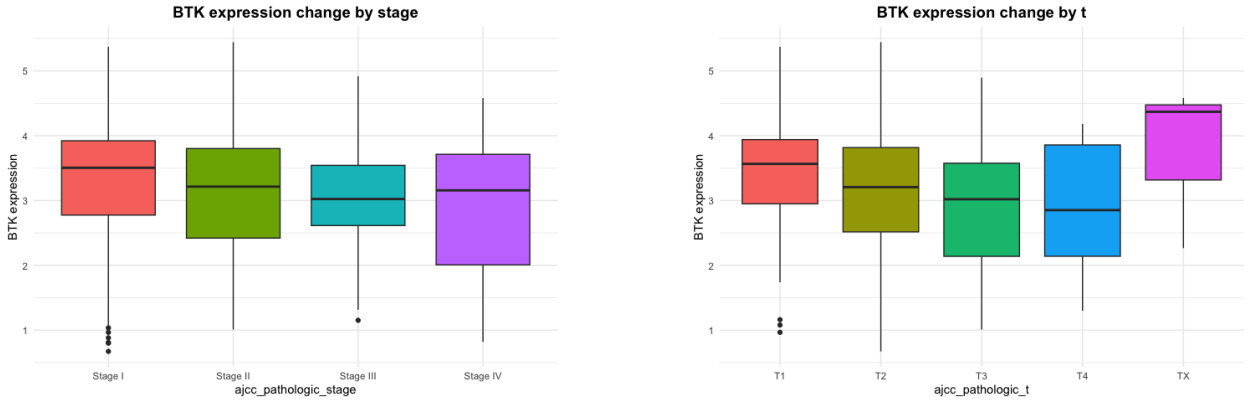
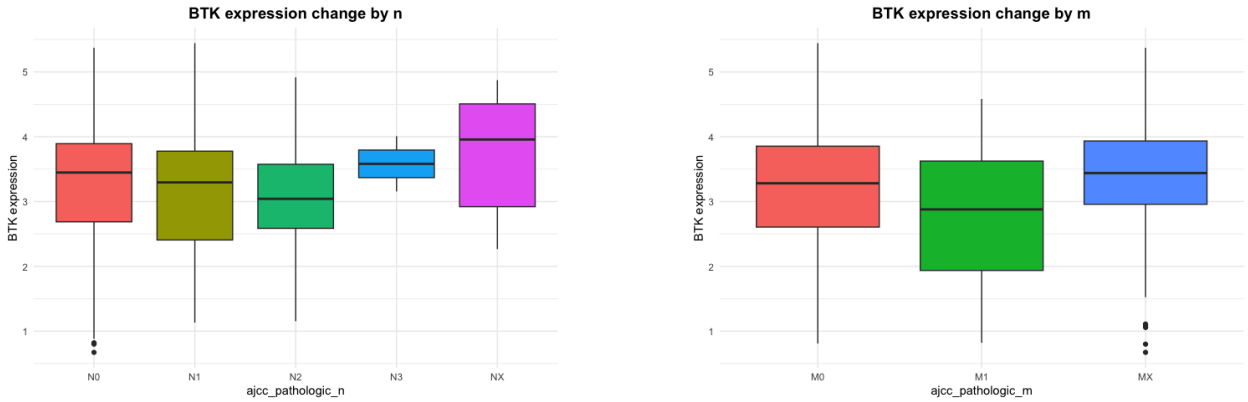


Figure 16: KEGG function result



(a) Cancer Stage

(b) T



(c) N

(d) M

Figure 17: BTK expression among different groups

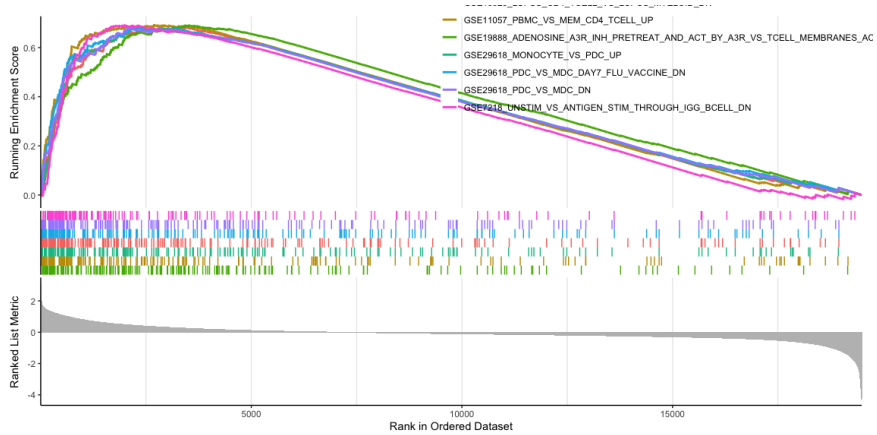


Figure 18: GSEA result with C7 gene set