Gene Set Enrichment Analysis of Pancreatic Cancer-Related Datasets

Gene Set Enrichment Analysis (GSEA) is an analytical approach used to identify significantly associated gene sets enriched with phenotypic changes. Given a reference gene set *S*, for instance genes sharing a common biological function, GSEA determines whether genes in *S* are populated at the top or bottom of a ranked expression dataset of interest *L* and calculates p-values by permutations of phenotypes (1). Ultimately, if genes in *S* are not randomly distributed across *L*, then they are likely to be enriched with phenotypic changes. This paper discusses methodologies, significance, and limitations of GSEA. Additionally, it presents three GSEA experiments performed for various pancreatic cancer related datasets and discusses their findings and results.

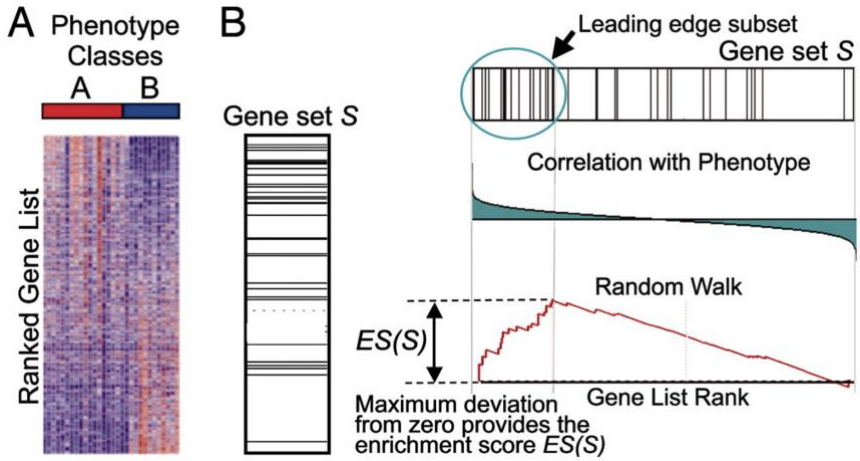

Figure 1. GSEA workflow **(1)**

Methodologies (1)

1. *Obtain a ranked list L (**A** in **Figure 1**)*: Given a raw expression dataset and phenotype labels, such as disease and control, differential expression analysis is performed. This can be done using any available R functions and packages, for instance the DESeq function from the DESeq package. Then a suitable ranking metric is used to rank the genes. A common ranking metric is $-\log_{10}(\text{p-value})*\text{sign}(\log_2\text{FoldChange})$, which ranks significant and upregulated genes at the top, significant and downregulated genes at the bottom, and non-significant genes in the middle. In other words, genes are ranked based on their ability to distinguish between the two classes.

2. *Select a reference gene set S (**B** in **Figure 1**)*: Some caveats include ensuring that *S* is highly relevant and encompasses genes in *L*. Any genes with minimal chances of selection should be excluded from *S*. Moreover, members of the set with known technical biases should also be removed (2).

3. *Calculate an Enrichment Score (ES)*: During a random walk down *L*, a running-sum statistic is increased if a gene in *L* also exists in *S*, and decreased otherwise. The magnitude of the increase depends on the correlation of the gene with the phenotype.

4. *Estimate the significance level of ES*: Phenotype labels are permuted, and the ES is recomputed for the permuted data to generate a null distribution of ES. The nominal p-value of the observed ES is then calculated relative to this null distribution.
5. *Adjust ES to account for multiple hypothesis testing*: To identify gene sets that exhibit significant enrichment related to the phenotype, steps 3. and 4. are repeated for each gene set *S* in the database. Once this is done, ES for each gene set is then normalized to a normalized enrichment score (NES). False

Discovery Rate (FDR) for each NES is finally calculated to identify gene sets that are statistically significant and relevant to the phenotype under investigation.

Significance in biomedical research

Traditional methods of analyzing differential essential genes rely too much on specific thresholds, either fold-change (FR) or p-value, to select and compare genes, leading to loss of subtle but significant information changes in biological processes like metabolic pathways, transcriptional programs, and stress responses. For example, using a strict p-value threshold of 0.05 might exclude genes with p-values slightly above this cutoff (e.g., 0.051), which could still be biologically relevant and contribute to the overall understanding of the disease process. On the contrary, GSEA utilizes a ranked list of all genes *L* as input and analyzes large-scale genes through identifying pathways and processes. This provides a more comprehensive gene pattern instead of focusing on high-score but unreproducible single genes. When computing the significance of ES, GSEA permutes class labels to preserve gene-gene correlation, which results in a more accurate null model. Therefore, GSEA retains statistical significance while increasing the biological relevance and practical value of results by emphasizing the biological significance of the gene set (1).

Aims

1. To write codes to perform GSEA on three different pancreatic cancer expression datasets curated from the Gene Expression Omnibus (GEO) database (3) to obtain enriched gene sets. The reference database used is KEGG_LEGACY from MSigDB (4). To perform GSEA, the GSEABase package (5) will be utilized and the steps outlined in **Methodologies** will be followed.

2. To compare the similarities and differences in the obtained enriched gene sets across different expression datasets. This analysis may provide insights into the broad spectrum of genes associated with pancreatic cancer.

3. To rationalize findings in 2. by doing literature review on pancreatic cancer related gene expression profiles.

Data Preparation

Dataset Selection

A systematic search was conducted using the key phrase “Pancreatic Cancer, Homo sapiens, Arrays” in the Gene Expression Omnibus (GEO), which identified 246 datasets. The selection was refined based on the following detailed criteria: First, the datasets had to focus specifically on mRNA expression. Second, they needed to provide a comparison between patients diagnosed with pancreatic cancer (PC) and control subjects. Third, the datasets were required to contain more than three samples to ensure statistical robustness. Lastly, the datasets needed to compare pancreatic cancer samples to normal controls derived from the same patients, ensuring consistency and relevance in the control data. Based on these criteria, three gene expression datasets, as listed in Table 1, were identified and selected for further analysis in this study.

Table 1. Selected Datasets for Pancreatic Cancer Study

GEO Accession	Study Design	Region	Number of Tumor Samples	Number of Normal Samples
GSE11838	RNA profiling from pancreatic cancer tumor and normal samples	USA	4	4
GSE147717	Transcriptomic profiling of Trip12 mRNA-inhibited PANC-1 cells and control cells	France	4	4
GSE22780	RNA profiling from pancreatic cancer tumor and normal samples	USA	8	8

Database Selection

The KEGG_LEGACY database has extensive coverage, allowing for a thorough exploration of potential mechanisms and pathways involved in pancreatic cancer. This subset of Canonical Pathways gene sets is derived from the KEGG pathway database and is considered a legacy gene set since the introduction of more recent KEGG MEDICUS data. GMT files for the KEGG_LEGACY database can be downloaded in various formats, including Gene Symbols, NCBI (Entrez) Gene IDs, and JSON bundles.

Experiments & Results

Experiment 1 : GSE11838

Dataset 1 (GSE11838), titled "*Pancreatic Tumor vs Various Tissue Normals*," includes a total of 107 samples from various tissues. For this study, a subset of 8 samples was selected, comprising 4 normal pancreatic tissue samples and 4 pancreatic tumor samples. The samples were collected from biopsies at different stages and normal tissues from various organ sites, including commercial normal samples. These samples were arrayed on 2-channel Agilent 60 mer Oligo arrays (human v1 & v2) and were median normalized for cross comparison.

Pathway	Size	ES	NES	p-value	p.adjust	q-value
KEGG_RIBOSOME	78	0.7992775	2.4184555	1.000000e-10	1.120000e-08	8.631579e-09
KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION	27	0.8265611	2.0680914	2.910637e-07	1.629957e-05	1.256170e-05
KEGG_VIBRIO_CHOLERAE_INFECTION	17	0.8500960	1.9442623	6.418804e-06	2.396354e-04	1.846814e-04
KEGG_TYPE_I_DIABETES_MELLITUS	19	0.8357834	1.9500022	9.407026e-06	2.633967e-04	2.029937e-04
KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION	18	0.8145871	1.8847670	5.686645e-05	1.273808e-03	9.816944e-04

Table 2. Top 5 Most Significant Genes Ranked by Adjusted P-values for the GSE11838 Experiment

In the GSEA analysis and adjusted p-value evaluation of 112 gene sets, **Table 2** lists the top 5 most enriched gene sets. The KEGG_RIBOSOME gene set was the most significant, with a maximum enrichment score (ES) of 0.7992775, indicating that this gene set was highly enriched at the top of the ranked list. In other words, this gene set was upregulated in the experimental samples compared to the control samples. Other notable pathways include KEGG_ANTIGEN_PROCESSING_AND_PRESENTATION, KEGG_VIBRIO_CHOLERAE_INFECTION, KEGG_TYPE_I_DIABETES_MELLITUS, and KEGG_PATHOGENIC_ESCHERICHIA_COLI_INFECTION, all of which show strong statistical significance with low adjusted p-values and q-values.

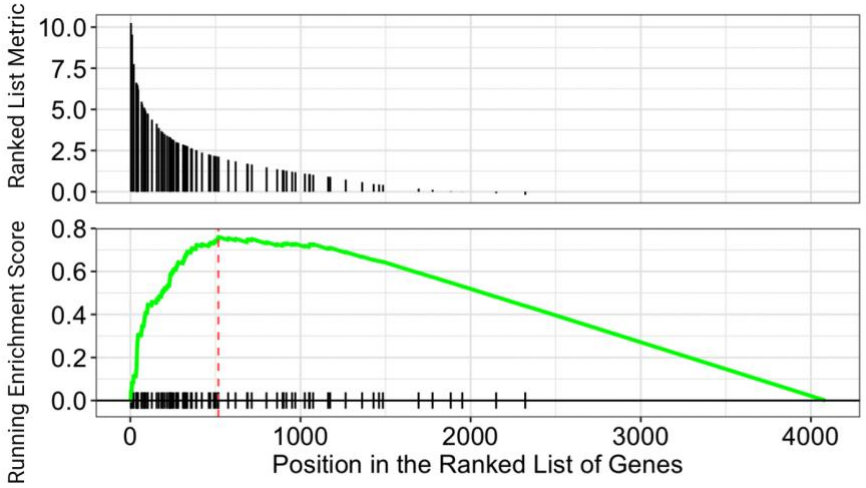

Figure 2. ES plot for the most significant gene set, KEGG_RIBOSOME, for the GSE11838 experiment

Figure 2 shows the visualization results of the KEGG_RIBOSOME gene set from the GSEA analysis. The top panel displays the distribution of the ranked list metric for all genes. Each vertical bar represents a gene, and the height of the bar indicates the value of the metric. Higher bars represent genes with stronger differential expression between experimental and control samples. The peak of the green line represents the maximum enrichment point of the KEGG_RIBOSOME gene set. This peak indicates that a large number of genes in the KEGG_RIBOSOME set cluster near the top of the ranked list, corresponding to a high level of upregulation in the experimental samples compared to the controls. This upregulation supports the high protein synthesis demands of cancer cells, as cancer cells need more ribosomes to grow and survive.

Experiment 2 : GSE147717

The title of this dataset is "*Transcriptomic profiling of human pancreatic cancer-derived cell lines PANC-1 depleted of TRIP12 mRNA compared to control cells.*" Its overall goal is to investigate the role of TRIP12 (Thyroid hormone Receptor Interacting Protein), an E3 ubiquitin ligase involved in numerous cellular processes, on the expression of pancreatic cancer-derived cell lines. TRIP12 controls the stability of a transcription factor that is important in pancreatic cancer initiation, and thus inhibition of TRIP12 may elucidate its role on the tumorigenicity and progression of pancreatic cancer. The dataset contains four experimental samples, which are PANC-1 cell + TRIP12 inhibited, and four control samples.(6)

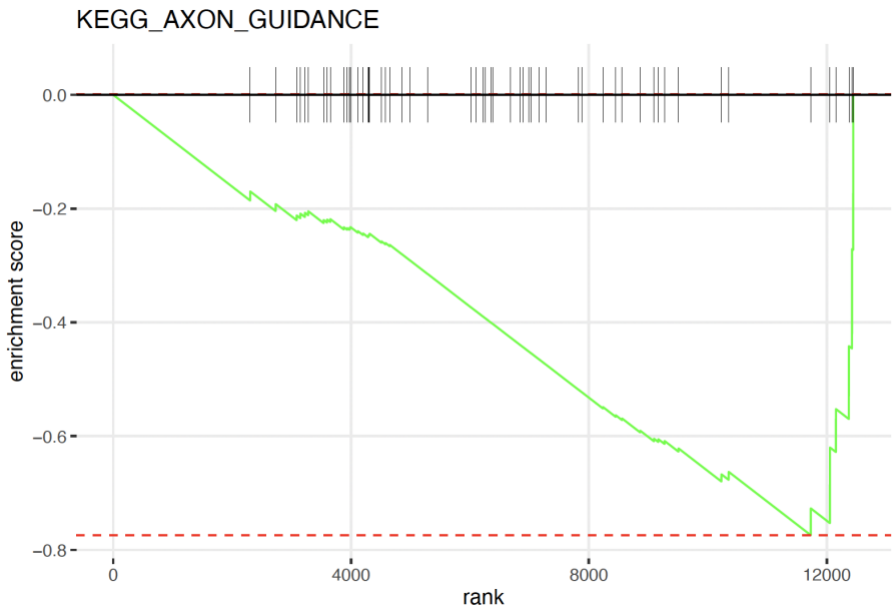

Figure 3. ES plot for the most significant gene set, KEGG_AXON_GUIDANCE, for the GSE147717 experiment

	pathway	pval	padj	log2err	ES	NES	size	leadingEdge
18	KEGG_AXON_GUIDANCE	2.765210e-07	0.0000434138	0.67496286	-0.7741687	-1.9917756	58	EPHB2, S....
108	KEGG_PATHWAYS_IN_CANCER	5.332086e-06	0.0004185687	0.61052688	-0.5771043	-1.7496235	154	IKBKB, J....
92	KEGG_NEUROTROPHIN_SIGNALING_PATHWAY	2.270161e-05	0.0011594611	0.57561026	-0.7106704	-1.8507652	61	IKBKB, M....
95	KEGG_NOD_LIKE_RECEPTOR_SIGNALING_PATHWAY	2.954041e-05	0.0011594611	0.57561026	-0.8477028	-1.9266959	25	IKBKB, M....
17	KEGG_AUTOIMMUNE_THYROID_DISEASE	9.552410e-05	0.0027796002	0.53843410	-0.8996613	-1.8648786	16	HLA-C, H....

Figure 4. Top 5 most significant genes ranked by p-adjusted values for the for the GSE147717 experiment

Experiment 3 : GSE22780

The title for dataset 3 is "*Affymetrix Arrays Interrogated with Tumor/Normal Pancreatic Samples*". The topic for this dataset is to study which gene causes pancreatic cancer when it goes through the 3p12 chromosome. The researchers took both normal samples and tumor samples from the same untreated, retrospective pancreatic adenocarcinoma cases. Finally, 8 pairs of experimental samples with matched tumor tissue and its adjacent normal tissue were obtained.

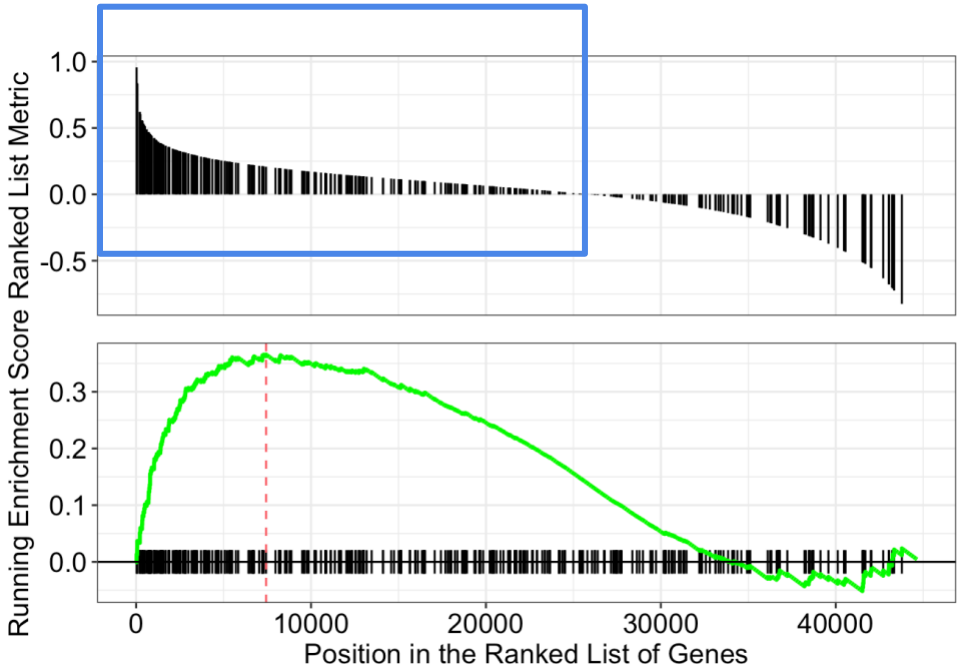

Figure 5. ES plot for the most significant gene set, KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY, for the GSE22780 experiment

Table 3. Top 5 Most Significant Genes Ranked by Normalized ES for the GSE22780 Experiment

Pathway	Size	ES	NES	p-val	Adj. p-val	q-val
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	133	0.6254735	2.952990	1.000000e-10	1.027778e-09	4.678363e-10
KEGG_HEMATOPOIETIC_CELL_LINEAGE	86	0.6264805	2.806109	1.000000e-10	1.027778e-09	4.678363e-10
KEGG_VIRAL_MYOCARDITIS	68	0.6365436	2.718834	1.000000e-10	1.027778e-09	4.678363e-10
KEGG_LEISHMANIA_INFECTION	69	0.6197319	2.677187	1.000000e-10	1.027778e-09	4.678363e-10
KEGG_LEUKOCYTE_TRANSENDOTHELIAL_MIGRATION	115	0.5671623	2.572168	1.000000e-10	1.027778e-09	4.678363e-10

Discussion

Experiment 1 : GSE11838

This study focuses on the gene enrichment analysis of pancreatic cancer. In this dataset's experiment, while the KEGG_RIBOSOME gene set was identified as the most significantly enriched, showing a high level of upregulation in the experimental samples compared to the controls, the KEGG_PANCREATIC_CANCER gene set did not show significant enrichment. The p-value for KEGG_PANCREATIC_CANCER was 0.7, and the adjusted p-value was 0.8. This lack of significance may be attributed to several factors. Different subtypes of pancreatic cancer exhibit distinct gene expression patterns, and the samples used in this analysis may not have included key subtypes. Additionally, pancreatic cancer involves multiple complex biological pathways, making it challenging to capture all relevant gene sets in a single analysis.

Experiment 2 : GSE147717

Out of the 29 significant gene sets identified by GSEA and p-adjusted values, KEGG_AXON_GUIDANCE was the most significant. The ES plot of this gene set (**Figure 4**) showed a negative decreasing ES score with a maximum enrichment of -0.774, indicating that gene set enrichment was at the bottom of the ranked list. In other words, this gene set was downregulated in the experimental samples, where TRIP12 was inhibited in the PANC-1 cell lines, compared to the control samples. The gene lines also showed that most genes appeared past the 4000^{th} rank in the ranked list. The leading edge of this gene set was the EPHB2 gene. The second most significant gene set was KEGG_PATHWAYS_IN_CANCER. KEGG_PANCREATIC_CANCER was the 23^{rd} most significant gene set.

Axon guidance pathways are crucial for neuronal development but can also play roles in cancer cell migration, invasion, and metastasis.(7) This experiment demonstrated that TRIP12 has a regulatory role in pathways associated with axon guidance, which can influence key aspects of pancreatic cancer cell behavior, including migration and invasion. It also suggested that bioinformaticians may want to research more about the role of EPHB2 gene in axon guidance in pancreatic cancer cell lines specifically. The high rank of KEGG_PATHWAYS_IN_CANCER indicated that the experiment indeed focused on pancreatic cancer. The presence of the KEGG_PANCREATIC_CANCER in one of the significant gene sets also corroborated this observation.

Experiment 3 : GSE22780

Based on the study, the KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY gene set is identified as the most significant across all test gene sets. Natural killer (NK) cells are a crucial component of the immune system, capable of inhibiting the growth and spread of various cancers, including pancreatic cancer. In the upper part of **Figure 5**, the bars above 0, highlighted in blue, represent the normal phenotype, while the bars below 0, highlighted in red, represent the tumor phenotype. In the lower part of **Figure 5**, the green line represents the running-sum statistics of the significant gene set, indicating changes in the natural killer cell gene set. The peak of the ES score is located at the top of the ranking list, indicating higher enrichment of NK cells in the normal phenotype. The trend gradually decreases to a negative value which demonstrates a downregulation or inactivation is happening in the tumor phenotype. This condition is consistent with the role of NK cells in preventing the spread of cancer. When NK cells are deactivated, there is a higher likelihood of cancer development. Meanwhile, the presence of the KEGG_PANCREATIC_CANCER with adjusted p-value equals to 2.228148e-05 shows this study on dataset 3 is highly significant in the progression of pancreatic cancer and highlights its potential as a target for further research and therapeutic intervention.

Taken in the context of the data, this result suggests that chromosome 3p12 may play a significant role in tumor formation. Given the importance of NK cells, genes on chromosome 3p12 might influence NK cell function, potentially leading to inactivation and aiding the development of pancreatic cancer. Understanding this relationship could lead to the development of new immunotherapy strategies to improve outcomes in pancreatic cancer treatment, such as enhancing the function of NK cells or blocking the immune evasion mechanisms of pancreatic cancer cells. Currently, there are already products like immune checkpoint inhibitors available.

Overall Discussion

Pancreatic cancer is a challenging, complex, multifactorial disease. Based on the study, each experiment obtained different significant gene sets enriched in the dataset are related to pancreatic cancer, including Protein Synthesis, Invasion and Metastasis, and Immune Evasion. This indicates that the appearance of pancreatic cancer is not caused by a single factor but by combined factors involved in multiple biological processes.

Furthermore, the analysis results heavily depend on the design and selection of datasets. As mentioned before, dataset 1 collected samples from normal tissues of various organs, potentially diluting or hiding specific pancreatic cancer gene expressions, making them less significant. In contrast, datasets 2 and 3 specifically collected samples from the pancreas, resulting in more significant and interpretable pancreatic cancer gene expressions.

Limitations

In our study, several limitations were identified. Firstly, differences between datasets and insufficient sample sizes can affect the consistency and comparability of the analysis results. This makes it more challenging to compare results across different datasets. Additionally, due to the necessity of performing numerous statistical tests and making complex biological interpretations, validating the results becomes more difficult and less straightforward.

From a technical perspective, GSEA has its own set of limitations. It relies on predefined gene sets, which means it may overlook recent discoveries or novel gene interactions. The method is sensitive to background noise, which can affect accuracy if the data quality is poor. Furthermore, GSEA assumes that genes act independently of each other, but in biological systems, genes often interact in complex ways. Lastly, the reference implementation of GSEA often struggles with accurately estimating very small p-values. This limitation is significant because GSEA involves multiple hypothesis testing, where the correction for multiple comparisons relies on precise p-value calculations. Inaccurate p-value estimation can impair the method's sensitivity, making it less reliable in identifying truly significant gene sets, especially in studies involving a large number of gene sets. This can result in a higher likelihood of overlooking important findings.(8)

Conclusion

This project has demonstrated the key uses and effectiveness of GSEA as a powerful tool for bioinformaticians to analyze gene expression datasets. With GSEA, bioinformaticians can identify which gene sets are enriched with phenotypic changes and determine which genes contribute most to the observed enrichment score. Additionally, they can ascertain whether the enriched gene sets are upregulated or downregulated in experimental samples compared to controls, providing deeper insights into the data. They can also compare the significance of each gene set by ranking their Normalized Enrichment Score (NES), an aspect this project can further improve on. Overall, this project showcased three examples of GSEA applied to three different pancreatic cancer-related data sets, highlighting its effectiveness as a tool for bioinformaticians.

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci. 2005 Oct 25;102(43):15545–50.

2. Tilford CA, Siemers NO. Gene Set Enrichment Analysis. In: Nikolsky Y, Bryant J, editors. Protein Networks and Pathway Analysis [Internet]. Totowa, NJ: Humana Press; 2009 [cited 2024 May 2]. p. 99–121. Available from: https://doi.org/10.1007/978-1-60761-175-2_6

3. Home - GEO - NCBI [Internet]. [cited 2024 May 2]. Available from: https://www.ncbi.nlm.nih.gov/geo/

4. KEGG_LEGACY [Internet]. [cited 2024 Jun 2]. Available from: https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=CP:KEGG_LEGACY

5. Bioconductor [Internet]. [cited 2024 May 2]. GSEABase. Available from: http://bioconductor.org/packages/GSEABase/

6. GEO Accession viewer [Internet]. [cited 2024 Jun 2]. Available from: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147717

7. Stoeckli ET. Understanding axon guidance: are we nearly there yet? Development. 2018 May 14;145(10):dev151415.

8. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinformatics. 2017 Mar 4;18(1):151.