Ruprecht-Karls-Universität Heidelberg

Fakultät für Biowissenschaften

Bachelorstudiengang Molekulare Biotechnologie

Cancer Hallmark and Metabolic Pathways differ over Cancer types and in Prostate Adenocarcinoma patients

Data Science Project SoSe 2022

14 Juli 2022

Fabian Strobel, Lottida Phondeth, Laura Lange, Carla Welz

Contents

1 Introduction

1.1 Hallmarks of cancer

Immortality has long been associated with cancer cells and has been a key component of research, tracing back to the discovery and distribution of HeLa cells (Skloot et al., 2010). It was in the year 2000 that the researchers Hanahan and Weinberg defined the characteristics of cancer cells in their publication "The hallmarks of cancer", therefore shaping our understanding of cancer (Hanahan and Weinberg, 2000). They described qualities such as immortalization, immune evasion, or angiogenesis need to apply to cells to be considered a cancer cell. Additional characteristics were published in their 2011 postulation "Hallmarks of cancer: the next generation" (Hanahan and Weinberg, 2011). One major hallmark that offers a myriad of pharmacological interventions is the "deregulation of cellular energetics". Altering metabolic pathways, therefore providing an energy supply, thus supporting cell proliferation, can be seen in various cancer types such as prostate adenocarcinoma.

1.2 Prostate adenocarcinoma

According to the GLOBOCAN 2020 estimates stemming from the International Agency for Research on cancer, prostate cancer is the second most common cancer found in men worldwide - making up about 1.4 million cases of the 10.1 million new cases of all combined cancers diagnosed in males. While the prevalence of prostate cancer is clear, the cause of it is not. However, it has been noticed that prostate cancer can be found more commonly in older males and therefore age could pose a potential risk factor (Bechis et al., 2011). The most recurrent diagnosed prostate cancer type is the so-called prostate adenocarcinoma (PRAD) (Li et al., 2016). The most common and effective ways to treat prostate cancers, in general, are surgery and radiation therapy. These treatments only apply to a non-metastatic disease progression. Metastatic prostate cancer calls for androgen deprivation therapy (Litwin and Tan, 2017). The absence of these androgen hormones leads to a significant decrease in the progression of prostate adenocarcinoma. However, once these cancer cells find a way to regain the activity of components that are part of the androgen receptor triggered signaling pathway, regardless

of the androgen hormone absence, it is even harder to treat it. Therefore, it is ever so more important to understand the metabolic changes in prostate adenocarcinomas (Ahmad et al., 2021).

1.3 Transcriptomic profiling

The data analysis project revolves around RNA-seq data derived from the cancer genome atlas. It gives an overview of the transcriptome, meaning the RNAs present in a sample such as a tissue. Additionally, the RNA-seq data quantifies the transcripts of a specific gene and therefore describes a good way to determine the relevance of certain genes in specific samples, especially since the amount of reads of a gene determines the generation of certain proteins.

1.4 The best way to analyze RNA-seq data

Usually, RNA-seq data comprises copious amounts of genes, which is why analyzing single genes has been a common practice. One major challenge that could result from solely focusing on a few genes is to understand the impact certain expressional changes of a specific gene can have on a pathway that is made up of a cascade of genes. To avoid this altogether analyzing RNA-seq data as gene sets and not as single genes is more practical. Methods that are based on analyzing gene sets such as the Gene Set Enrichment Analysis (GSEA) or the Gene Set Variation Analysis GSVA are therefore favorable (Hänzelmann et al., 2013; Subramanian et al., 2005).

1.5 General questions and the path to their answers

By analyzing the given RNA-seq data the goal was to identify patterns between cancer types of the data set, as well as potentially find significant differences in pathways between normal and cancer cells. Additionally, quantifying and clustering pathway activity was to be achieved. Using common methods such as a principal component analysis and heatmaps are the building blocks of cluster finding. Newer methods that were applied pose the Uniform

Manifold Approximation and Projection (UMAP) as a means for dimensional reduction and formation of clusters and the GSVA as a means for characterizing and quantifying the pathway performance in specific samples (McInnes and Healy, 2018). Visualizing and selecting differentially expressed genes between two different phenotypes such as cancer or normal cells can be done with volcano plots.

---

2 Methods

2.1 Initial raw data

During our project, we used four given data sets. The first was an R-object consisting of a list of gene sets for cancer hallmarks. Second, a pan-cancer RNA-seq gene expression data frame for 9,741 patients of 33 various cancer types based on data generated by the "The Cancer Genome Atlas" Research Network: https://www.cancer.gov/tcga. In addition, there was an R-object containing 37 clinical annotations regarding the RNA-seq patients. And fourth, for a focused analysis of PRAD, an R-object with RNA-seq gene expression data of matched tumor tissue and normal tissue of 52 PRAD patients was used. To get a broader view of the cancer hallmarks and metabolic activities, additional gene sets were chosen from the Molecular Signatures Database (MSigDB) (Liberzon et al., 2011; Subramanian et al., 2005) after literature review. Trying to get a large overlap with the genes from the RNA-seq, 509 additional gene sets were used, resulting in a total number of 555 gene sets used during the study. These include 50 hallmark gene sets (Liberzon et al., 2015), 186 curated gene sets from the KEGG pathway database (Kanehisa, 2019; Kanehisa et al., 2021; Kanehisa and Goto, 2000), 189 oncogenic signature gene sets (Liberzon et al., 2011; Subramanian et al., 2005) and the 84 largest ontology gene sets (Ashburner et al., 2000; GOC, 2021; Köhler et al., 2020) as of June 2022.

2.2 Preprocessing

The RNA-seq data came in a log2(TPM) format which served as normalization technique. The original pan-cancer data frame, which contained 60,498 genes, was preprocessed as follows (Figure 1): After confirming the absence of missing values, the means and variances for all genes were calculated. To remove rather constant genes across all cancer types, variance filtering was performed, were all genes with a variance below the 35 % quantile were discarded. Following this, the biotypes of the remaining 39,324 genes were identified by using the EnsDb.Hsapiens.v79 package (Citation). 98 genes which could not be attributed with a biotype were also removed. For the rest, the frequency of each occurring biotype was counted. For further analysis, the most interesting biotypes within the RNA-seq data frame were kept. These include short non-coding RNAs like small nuclear RNAs, micro RNAs, ribosomal RNAs and small nucleolar RNAs, which are known to have important functions in molecular biology (Alberts, 2015). Furthermore, long non-coding RNAs, which are longer than 200 bp in length and might possess regulatory functions (Alberts, 2015; Cunningham et al., 2021) and protein coding genes were retained. The latter also included T cell receptor genes and immunoglobulin genes that both undergo somatic recombination and were listed with separate biotypes (Cunningham et al., 2021). All chosen biotypes also appeared within the pathways. After removing the other biotypes from the pan-cancer data frame, it only contained 20,675 genes. Similarly, the combined PRAD data frame was variance filtered using a 60 % quantile threshold. In contrast to the pan-cancer analysis biotypes, only protein coding genes and long non-coding genes were present and therefore kept resulting in a final data frame with 7,801 genes.

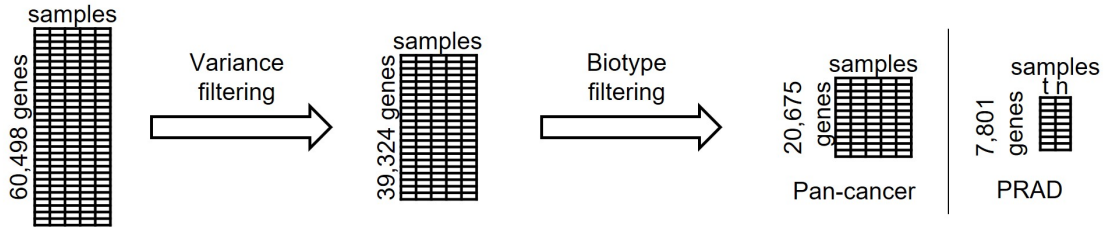

Figure 1: Preprocessing.

2.3 Analysis of gene sets

To get a sense of the meaningfulness of the used gene sets in the study, their overlap with the genes of the RNAseq was investigated using Venn diagrams. Before analyzing the RNA-seq data, a comparison of the gene sets was performed. Therefore, it was necessary to convert all gene names into the same format. The gene symbols in the pathways of the given gene sets were converted into Ensemble gene IDs using the *EnsDb.Hsapiens.v79 package*. Regarding the additional gene sets, the genes were also imported as Ensembl gene IDs. Next, all gene sets were combined into one list. To take a closer look at the gene sets, the similarity between them was investigated. As a metric which compares how many genes are shared between the different gene sets, the Jaccard Indices

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \, with 0 \leq J(A,B) \leq 1$$

were computed for each combination. These result in a value between zero and one, which is the ratio between the intersection and union of the two respective pathways (Levandowsky and Winter, 1971). After computing the Jaccard indices for every combination, a heatmap was created, which also clustered the gene sets.

2.4 Pan-cancer analysis

To start off, different descriptive plots were produced based on the given data. For example, the data was checked for normality. Next, a histogram for age at initial cancer diagnosis or the frequency of cancer types throughout the pan-cancer analysis. For further dimension reduction a Principal Component Analysis (PCA) (Jolliffe, 2011) was applied over the full pan-cancer data using the Seurat package. With the resulting first 50 principal components, Uniform Manifold Approximation and Projection (UMAP) (McInnes and Healy, 2018) with a set seed (123) was performed with the uwot package. By using “cosine” (why) as metric two UMAP components were calculated and later plotted in a two-dimensional plot. The same analysis was used after dividing the whole data into the 33 different cancer types with 35 principal components. Next, Gene Set Variation Analysis (GSVA) was carried out.

GSVA allows summarizing single genes into defined pathways or gene sets. This reduces the dimensionality of the RNAseq data and eases biological interpretability. The input for the GSVA algorithm is a data frame of log2 RNA-seq counts and a list of gene sets. To calculate a ranking-score a cumulative density function is estimated for each gene over all samples. The ranking-score is the probability for the gene expression in the corresponding sample. These scores are then used to create a ranked list of genes for each sample. The ranked list is the basis to access the GSVA enrichment score using the Kolmogorov-Smirnov (KS) like random walk statistic. In total 2 random walks are done. The first one regards the genes in the gene set. The algorithm iterates over each gene in the ranked list and checks if it is in the gene set. The ranks of the genes present in the gene set are added to a running sum. For the genes that are not present the running sum is kept as it is. In the second random walk the value 1 is added to the running sum if the genes that are not present in the gene set. If the gene is in the gene set, the running sum remains unchanged. The enrichment score is the difference between the largest positive and the largest negative deviations. *Weighted Kolmogorov Smirnov testing: an alternative for Gene Set Enrichment Analysis???* (Hänzelmann et al., 2013). For our study, this method resulted in a pathway activity matrix with the samples as columns and 552 pathways as rows. Three pathways were discarded by the GSVA package because of a too small intersection between pathway genes and RNAseq genes, since the minimum size of the resulting gene sets was set to 3. Otherwise, the functions default settings were used. Trying to determine the win or loss of information caused by the GSVA a PCA followed by UMAP was done for the pathway activity matrix with the same settings as mentioned before. To explain the resulting clusters, the pathway activity for the pathways with the highest variance was visualized in a heatmap. Another heatmap was created in which the mean pathways activity for each cancer type was determined and plotted against the pathways. Finally, a linear regression model was built to predict...

2.5 Focused analysis: Prostate adenocarcinoma

For the focused analysis, the normality was also checked using violin plots and qq-plots. As with the pan-cancer RNAseq data frame, PCA and UMAP were applied on a combined data frame of tumor and normal samples from RNAseq. To get a first impression of the differences between the samples, volcano plots were created. Volcano plots are one way to visualize the difference between two conditions and identify genes or pathways that are differentially expressed and whose change in expression is statistically significant. To get a symmetrical distribution around zero the log2 fold change between tumor and normal tissues is calculated for each gene or pathway.

$$Foldchange = log_2 \frac{mean\,tumor}{mean\,normal}$$

The log2 fold change is plotted on the x-axis. Genes with a log2 Fold Change higher than +/- 1 were defined to be differentially expressed and pathways with a log2 Fold Change of +/- 0.4 were defined to be differentially expressed. Additionally, a paired Wilcoxon test was used to compute a p-value, which was then adjusted due to multiple testing using the Bonferroni correction. Before the test an alpha of 0.05 was defined. To get a higher score for the significant tests the -log10 p-value was plotted on the y-axis.

2.5.1 PRAD gene ontology enrichment analysis

To get a better understanding of the RNAseq gene's functions, a gene ontology enrichment analysis (GOEA) was performed. Based on the volcano plot with the genes the 500 differentially expressed genes (UP/DOWN) with the smallest p values (highest significance) were extracted. For these 500 genes the gene ontology (GO) terms were identified using the packages biomaRt and GO.db. (Figure XA, 0.999 quantile). It is important to know that one gene can have multiple GO terms. For every GO term a list with the corresponding genes was created but GO terms with less than 10 corresponding genes were discarded. Finally, the enrichment analysis was performed using the GSVA package with the combined tumor and normal RNAseq data of focused analysis and the GO term gene lists as gene set list in-

put. All other parameters were equal to the pan-cancer GSVA. On the resulting GO activity matrix were then again PCA and UMAP applied with the same settings as before (Figure XB, UMAP).

2.5.2 PRAD Gene Set Variation Analysis (and GSEA)

Since GO terms alone do not give information about further interactions or pathways, in the next step the use of Gene Set Enrichment Analysis (GSEA) and Gene Set Variation Analysis (GSVA) were investigated. Unlike the GSVA the Gene Set Enrichment Analysis (GSEA) uses the distinction between two different phenotypes to determine a ranking score for each gene of each patient (Korotkevich et al., 2021; Subramanian et al., 2005). The Log2 Foldchange can be used to determine this score to enable the ranking of the genes of each patient. Since approximately 28% of the Log2 Foldchange showed to have ties the fGSEA couldn't be applied here. When applying the GSVA, 20 gene sets were discarded by the function. After GSVA, a new volcano plot was created and PCA and UMAP were applied on the pathway activity matrix. Furthermore, the pathway activity of six selected upregulated genes from a pathway volcano plot was visualized. The shown pathways are pathways that did not already appear in the pan-cancer plots. When calculating the pathway activity with the GSVA an error occurs because larger gene sets automatically get a higher enrichment score. One opportunity to correct this error would be to count how many genes of the gene set are included in the TCGA data frame and determine how many differentially expressed genes are in the gene set. The formula

$$Pathway{-}size{-}ratio = number of differentially expressed genes / number of genes in pathway$$

provides the percentage of differentially expressed genes in the pathway.

---

3 Results

3.1 Preprocessing and gene set analysis

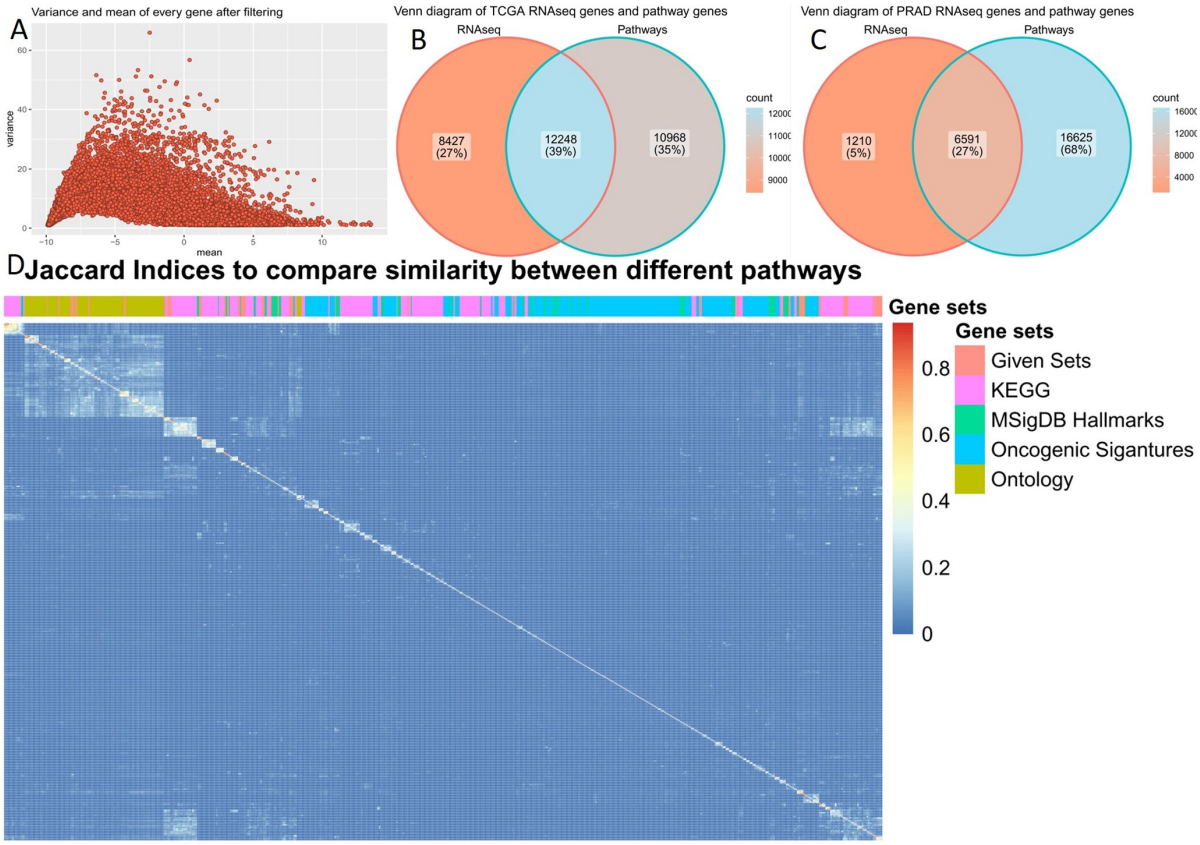

Figure 2: Preprocessing results.

3.2 Pan-cancer analysis

3.3 Focused analysis

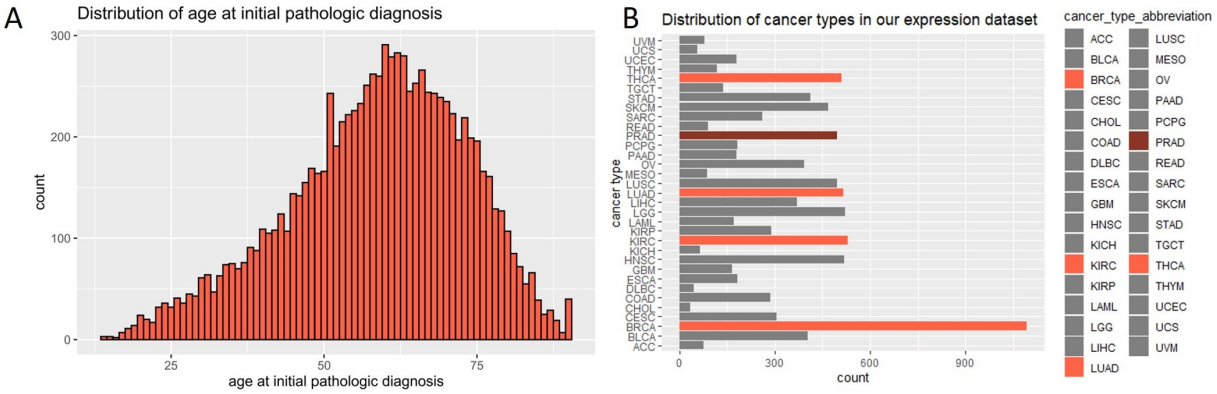

Figure 3: Descriptive analysis. CARLA NEUE PLOTS

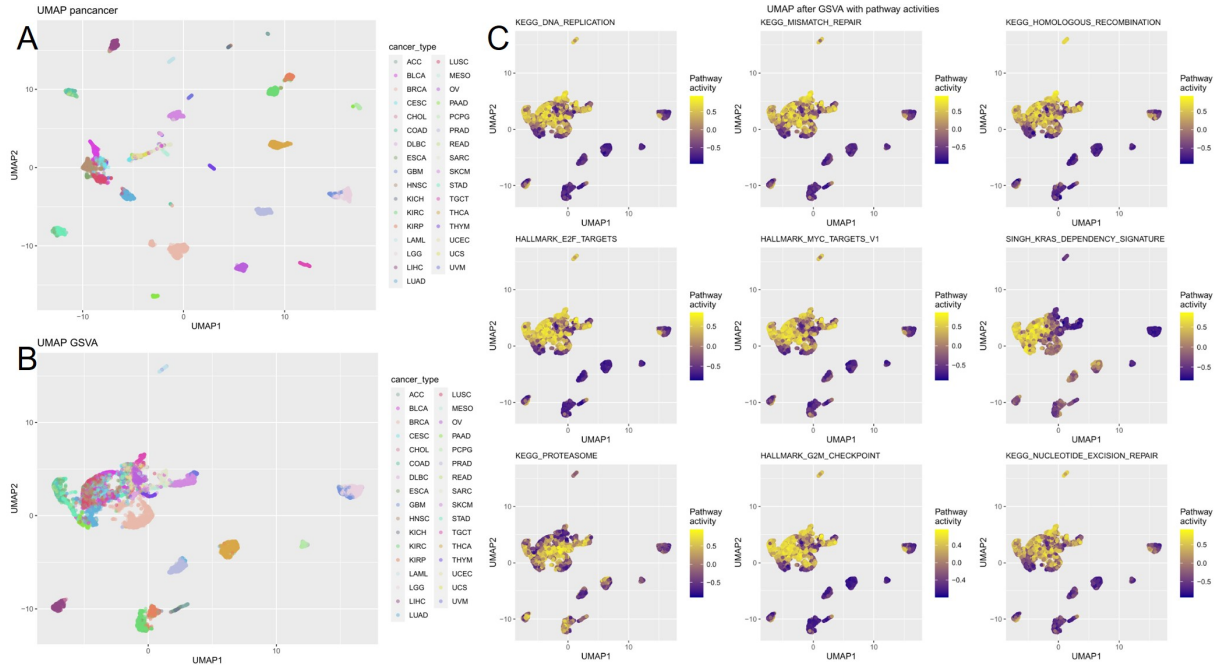

Figure 4: Pan-cancer GSVA and UMAP.

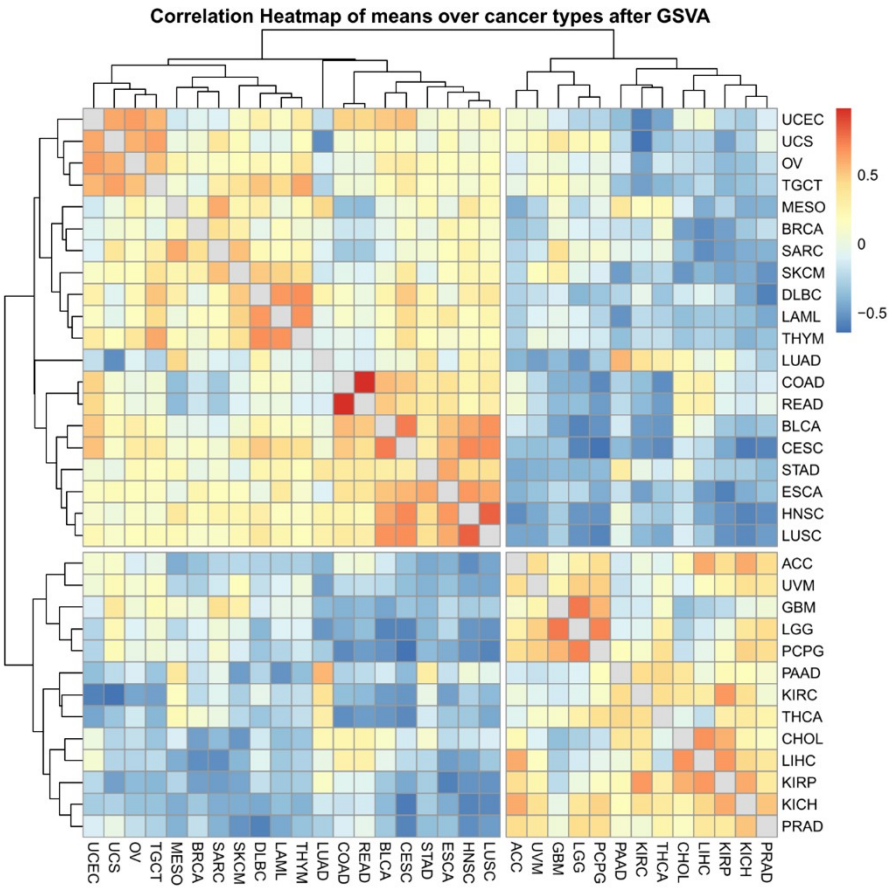

Figure 5: Correlation between cancer types after GSVA.

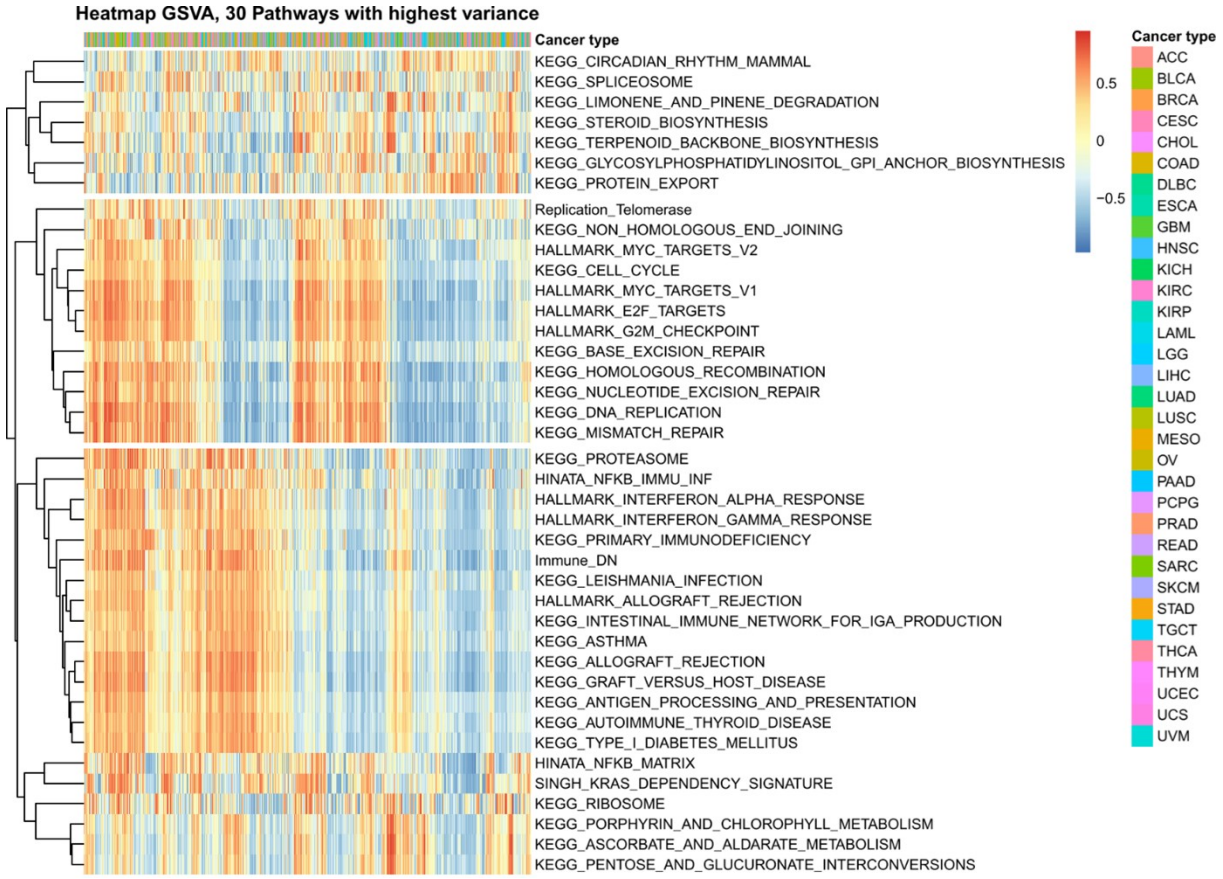

Figure 6: GSVA 30 topVar pws.

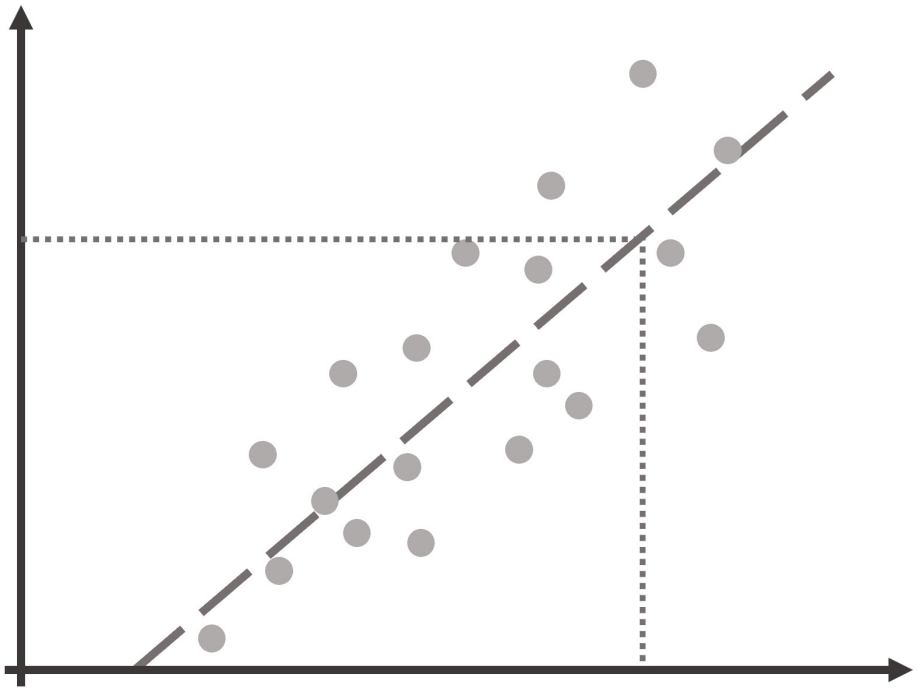

Figure 7: Linear regression.

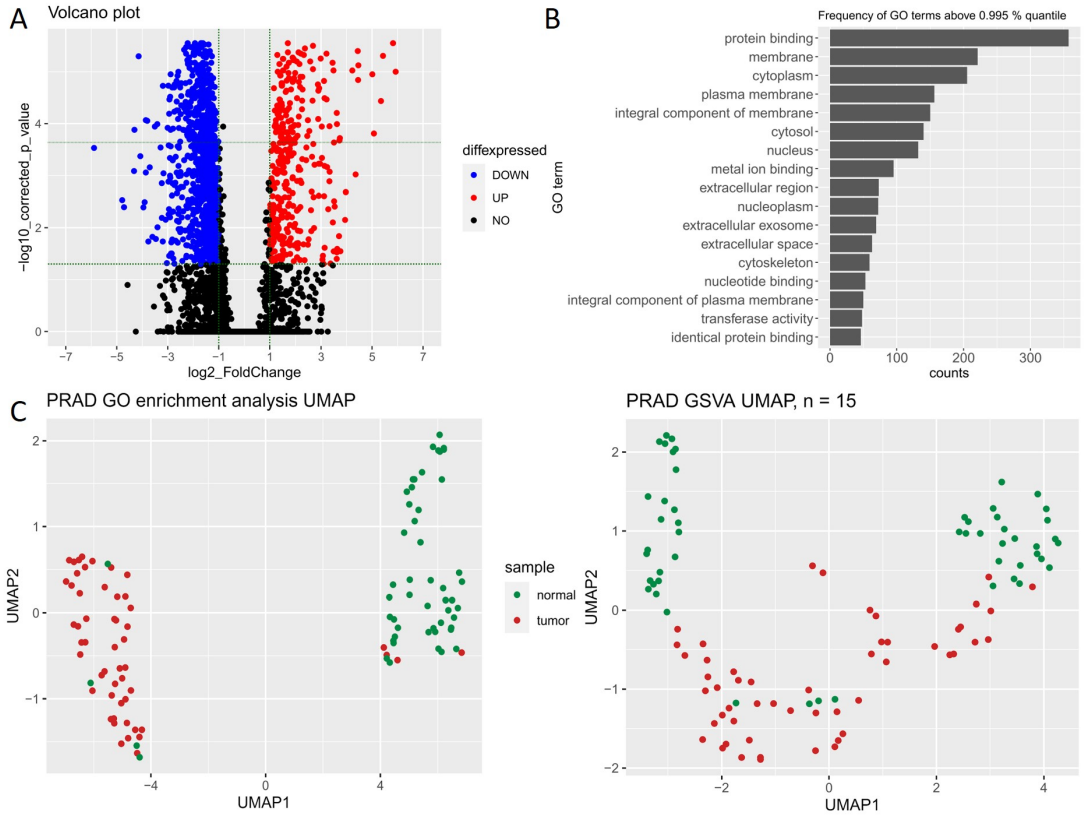

Figure 8: Panel name

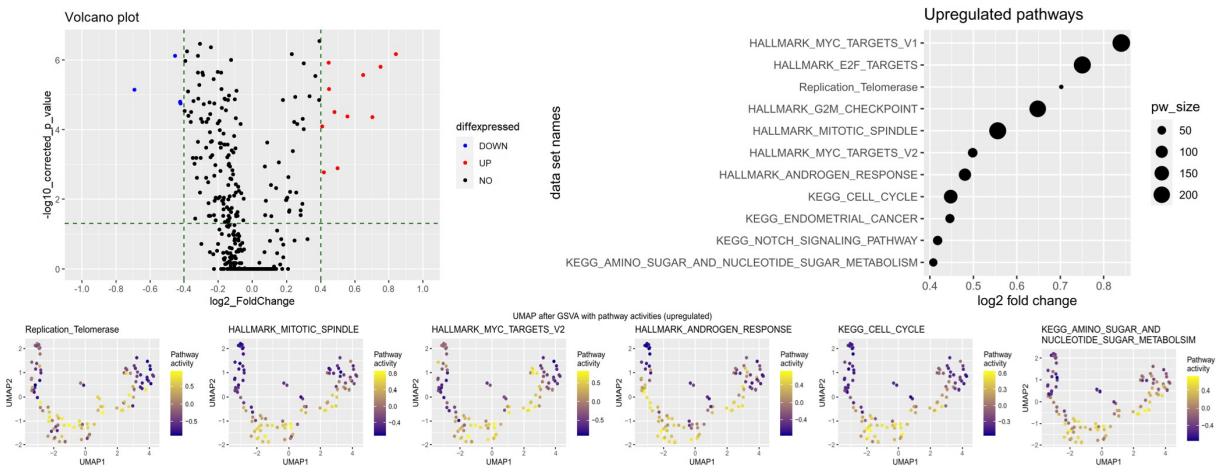

Figure 9: Panel name

4 Discussion

4.1 kjhtgrfedws

4.2 uztrwwret

---

5 Refrences

6 Appendix

6.1 used packages

packages	
ggplot2	H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016
ggVennDiagram	Gao C (2021). _ggVennDiagram: A 'ggplot2' Implement of Venn Diagram_. R package version 1.2.0,<https://CRAN.R-project.org/package=ggVennDiagram>
gridExtra	Auguie B (2017). _gridExtra: Miscellaneous Functions for Grid Graphics_. R package version 2.3, <https://CRAN.R-project.org/package=gridExtra>
randomcoloR	Ammar R (2019). _randomcoloR: Generate Attractive Random Colors_. R package version 1.1.0.1, <https://CRAN.R-project.org/package=randomcoloR>
pheatmap	Kolde R (2019). _pheatmap: Pretty Heatmaps_. R package version 1.0.12, <https://CRAN.R-project.org/package=pheatmap>
msigdbr	Dolgalev I (2022). _msigdbr: MSigDB Gene Sets for Multiple Organisms in a Tidy Data Format_. R package version 7.5.1, <https://CRAN.R-project.org/package=msigdbr>
EnsDb.Hsapiens.v79	Rainer J (2017). _EnsDb.Hsapiens.v79: Ensembl based annotation package_. R package version 2.99.0.
biomaRt	Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009)
GO.db	Carlson M (2022). _GO.db: A set of annotation maps describing the entire Gene Ontology_. R package version 3.15.0.
Seurat	Hao and Hao et al. Integrated analysis of multimodal single-cell data. Cell (2021) [Seurat V4] Stuart and Butler et al. Comprehensive Integration of Single-Cell Data. Cell (2019) [Seurat V3] Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol (2018) [Seurat V2] Satija and Farrell et al. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol (2015) [Seurat V1]
uwot	Melville J (2021). _uwot: The Uniform Manifold Approximation and Projection (UMAP) Method for Dimensionality Reduction_. R package version 0.1.11, <https://CRAN.R-project.org/package=uwot>
dplyr	Wickham H, François R, Henry L, Müller K (2022). _dplyr: A Grammar of Data Manipulation_. R package version 1.0.9, <https://CRAN.R-project.org/package=dplyr>
fgsea	G. Korotkevich, V. Sukhov, A. Sergushichev. Fast gene set enrichment analysis. bioRxiv (2019), doi:10.1101/060012
BiocManager	Morgan M (2022). _BiocManager: Access the Bioconductor Project Package Repository_. R package version 1.30.18, <https://CRAN.R-project.org/package=BiocManager>
data.table	Dowle M, Srinivasan A (2021). _data.table: Extension of 'data.frame'_. R package version 1.14.2, <https://CRAN.R-project.org/package=data.table>
GSVA	Hänzelmann, S., Castelo, R. and Guinney, A. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics, 14:7, 2013

Figure 10: R packages.