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

13 Juli 2022

Fabian Strobel, Lottida Phondeth, Laura Lange, Carla Welz

Contents

1 Introduction

1.1 Hallmarks of cancer

1.2 Prostate adenocarcinoma

... LOTTI

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R. (Alberts 2015)

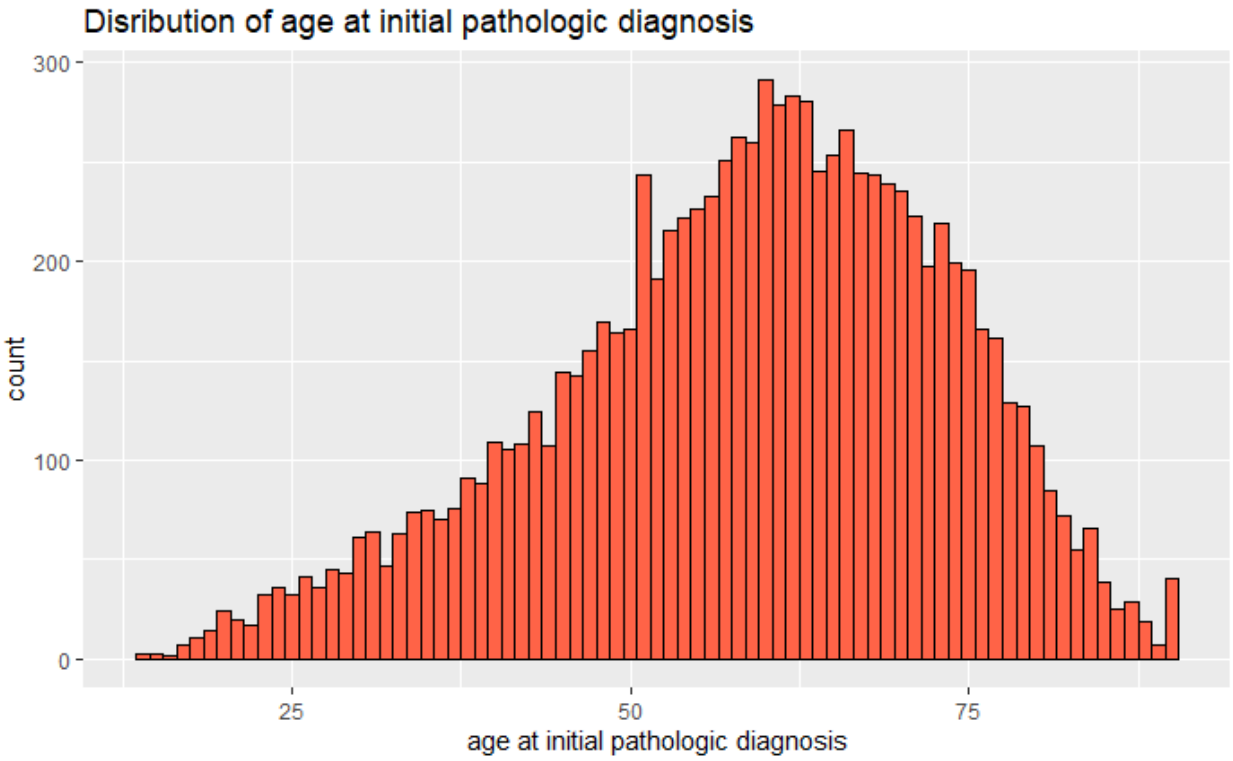

Figure 1: Caption for the picture.

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 about 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 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*. 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.

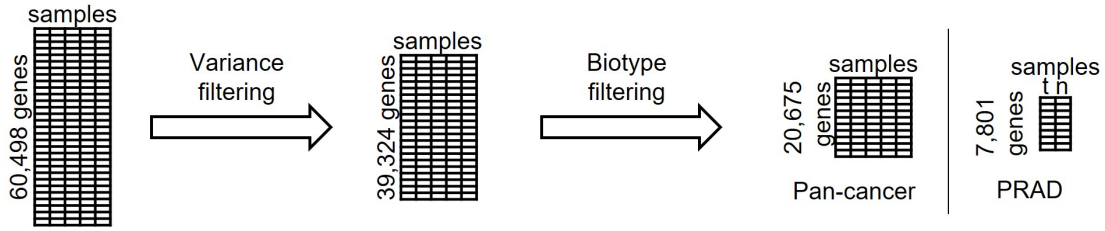

Figure 2: Preprocessing.

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 in 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 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

2.4.1 descriptive analsis –> CARLA

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 (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.

2.4.2 linear regression

2.5 Focused analysis: Prostate adenocarcinome

2.5.1 descriptive analysis CARLA

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. ### Volcano Plots LAURA

2.5.2 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 input. 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).

2.5.3 GSEA/GSVA

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.

2.5.4 PRAD pathway activity matrix –> HEATMAP

After GSVA PCA and UMAP were applied on the pathway activity matrix. Furthermore, the pathway activity of the differentially expressed genes was visualized in the figure.

---

3 Results

3.1 Preprocessing

By calculating the mean and variance for each gene of the original RNAseq data frame (Figure XA, XB), an overview about the data was possible. To reduce the dimensionality variance and biotype filtering were performed, resulting in a data frame with only a third of the starting number of genes. The remaining genes were visualized in a mean variance plot (Figure XC) showing a smaller number of genes with very large variance compared to the rest.

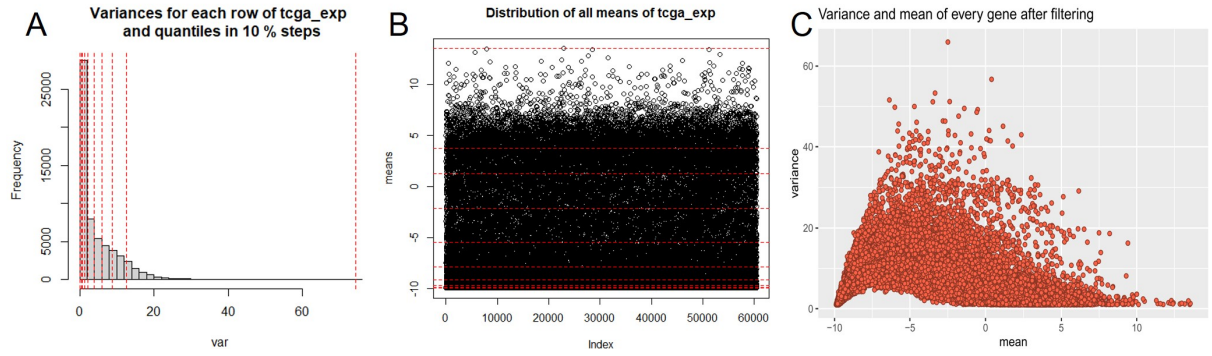

Figure 3: Preprocessing results.

3.2 Gene set analysis

To combine the information of RNAseq into pathways the overlap between the experimental data and the available gene sets must be considered. In the pan-cancer analysis 12,248 RNAseq genes (59 %) were included in the gene sets (Figure XA). From only the ten percent of genes with the highest variance in expression 65 % were present within the pathways (Figure XB). A greater overlap was achieved at the focused PRAD analysis where 6,591 RNAseq genes (84 %) occurred within the gene sets (Figure XC). Furthermore, the Jaccard metric combined with a heatmap shows the similarity between the gene sets used during the study (Figure XD). Particularly the ontology gene sets display the greatest likeness, as well as a small part of the KEGG pathways and given pathway.

3.3 Pan-cancer analysis

For pan-cancer analysis PCA and UMAP were applied on the cleaned RNAseq data frame. The two-dimensional UMAP components were then plotted and colored according to the 33 different cancer types (Figure XA). Around two thirds of the cancer types seemed to build individual clusters. But not every cluster was pure of one cancer type. Furthermore, the were one large and two smaller clusters consisting of multiple cancer types where no differentiation was possible. Trying to characterize the cancer types with pathway activity, GSVA was applied with the former mentioned chosen pathways. Next, PCA and UMAP were also applied and plotted (Figure XB) which allowed a comparison between the plots. After GSVA only ten sperate clusters were visible. Whereas some of the clusters before GSVA reappeared as individual small clusters, many cancer types formed one enormous cluster. To find out more about the differences between the enormous cluster and the smaller ones the pathway activity for the nine pathways with the largest variance across all cancer types was dyed (Figure XC). These pathways include cell cycle, DNA replication and repair, the E2F transcription factor family, proteasome, and the oncogenes myc and kras. In general, the enormous cluster and one other small cluster showed predominantly increased pathway activity (BLCA, BRCA, CESC, COAD, DLBC, ESCA, HNSC, LAML, LUAD, LUSC, MESO, OV, PAAD, READ, SARC, SKCM, STAD, TGCT, THYM, UCEC, UCS, UVM). In contrast, the remaining clus-

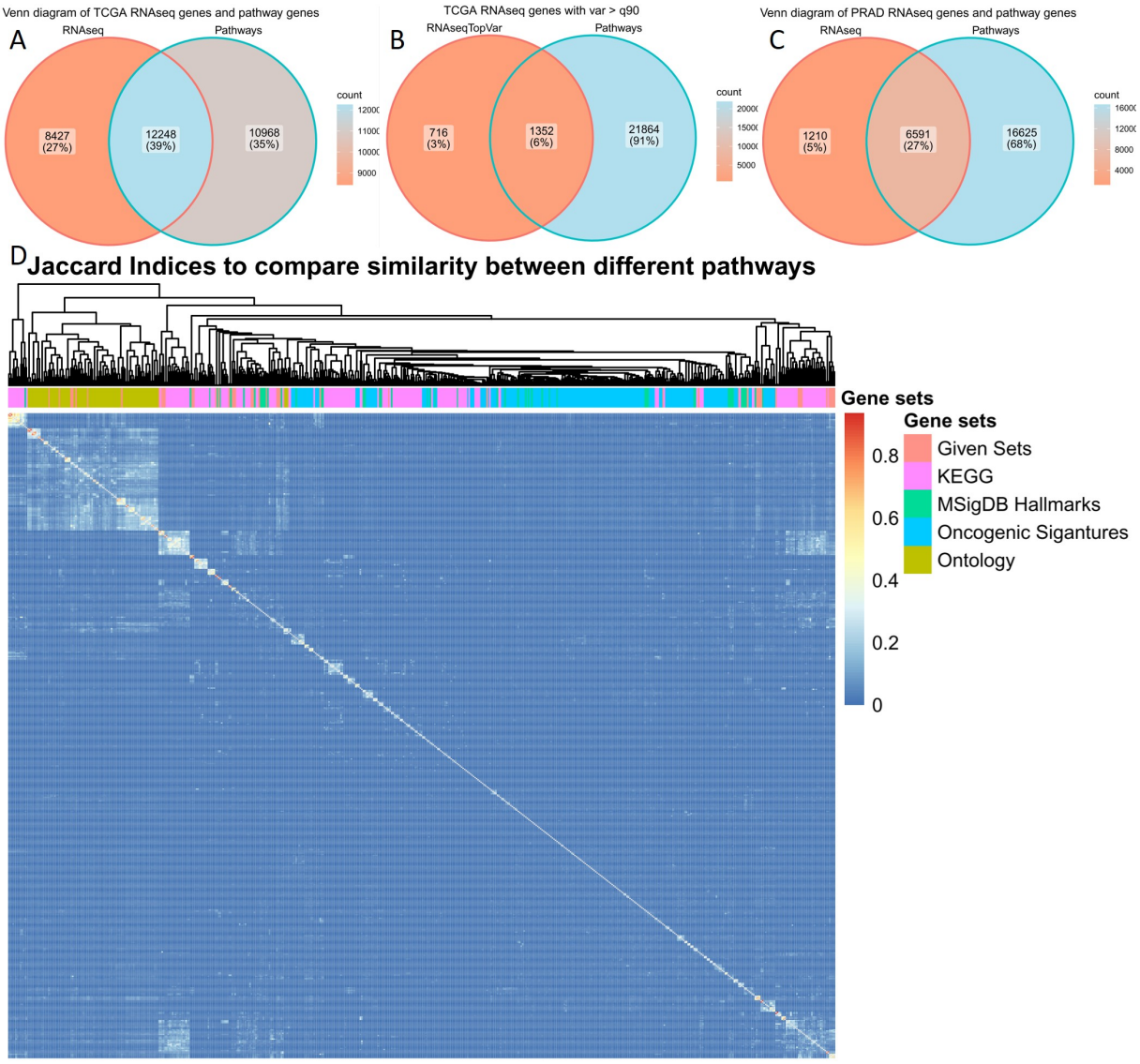

Figure 4: Gene set analysis.

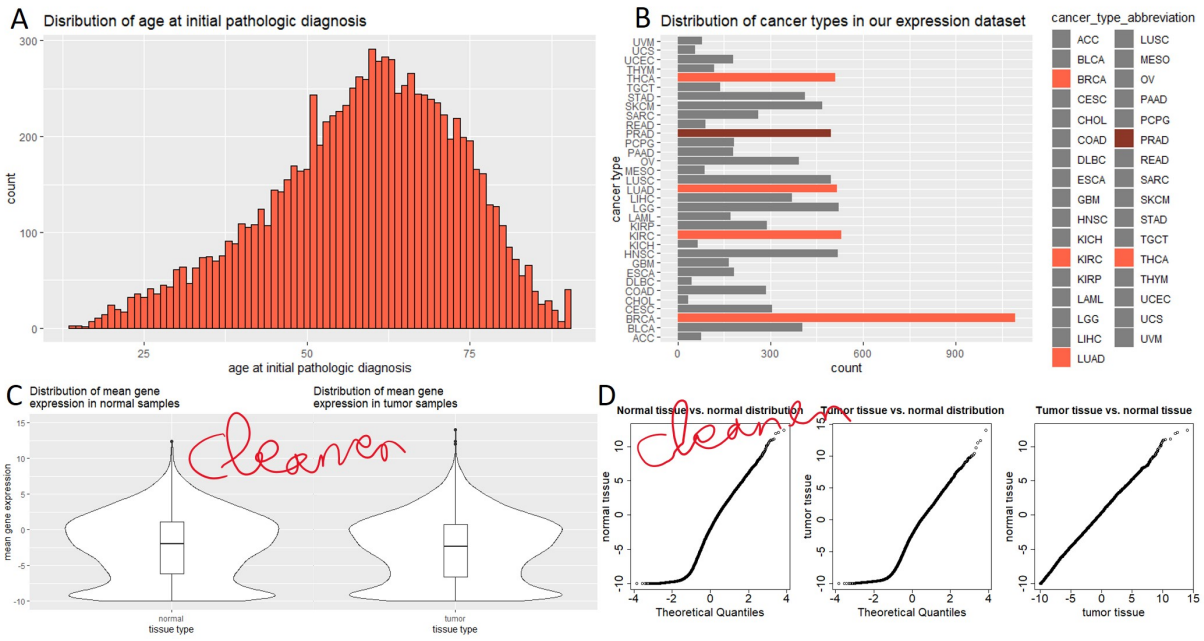

Figure 5: Descriptive analysis. CARLA NEUE PLOTS

ters showed rather decreased pathway activity (ACC, CHOL, GBM, KICH, KIRC, KIRP, LGG, LIHC, PCPG, PRAD, THCA).

---

4 Discussion

4.1 kjhtgrfedws

4.2 uztrwwret

Refrences

Alberts, Bruce. 2015. *Molecular Biology of the Cell*. Book. 6. ed. New York, NY [u.a.]: Garland Science.

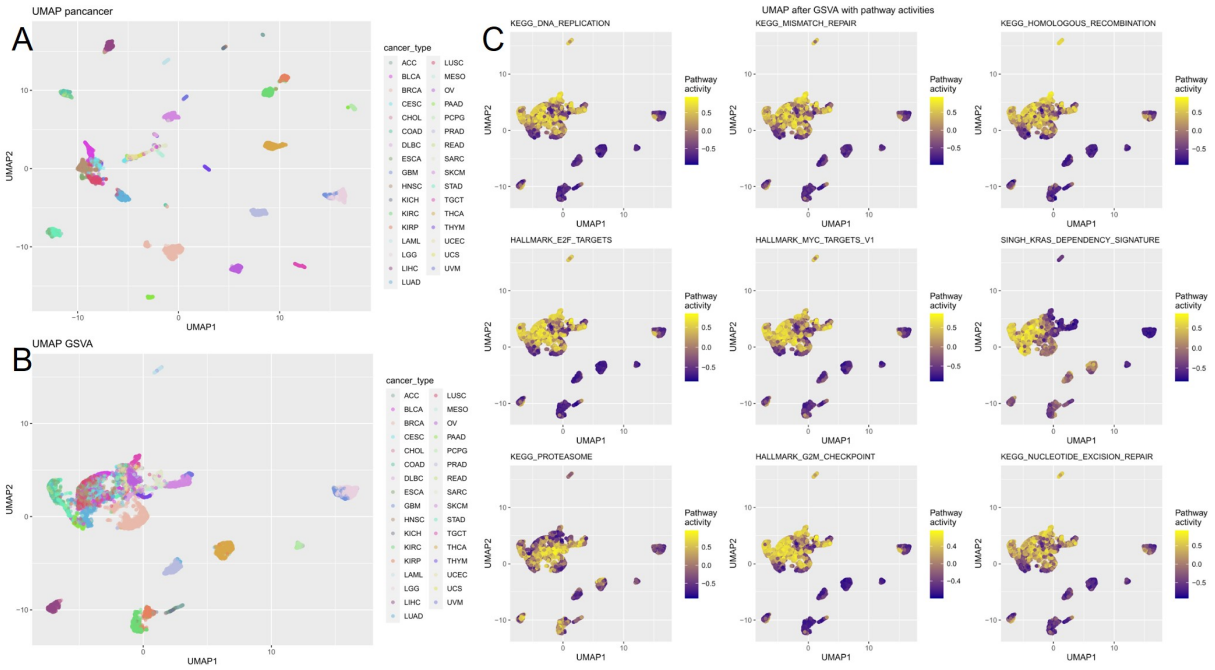

Figure 6: Pan-cancer GSVA and UMAP.

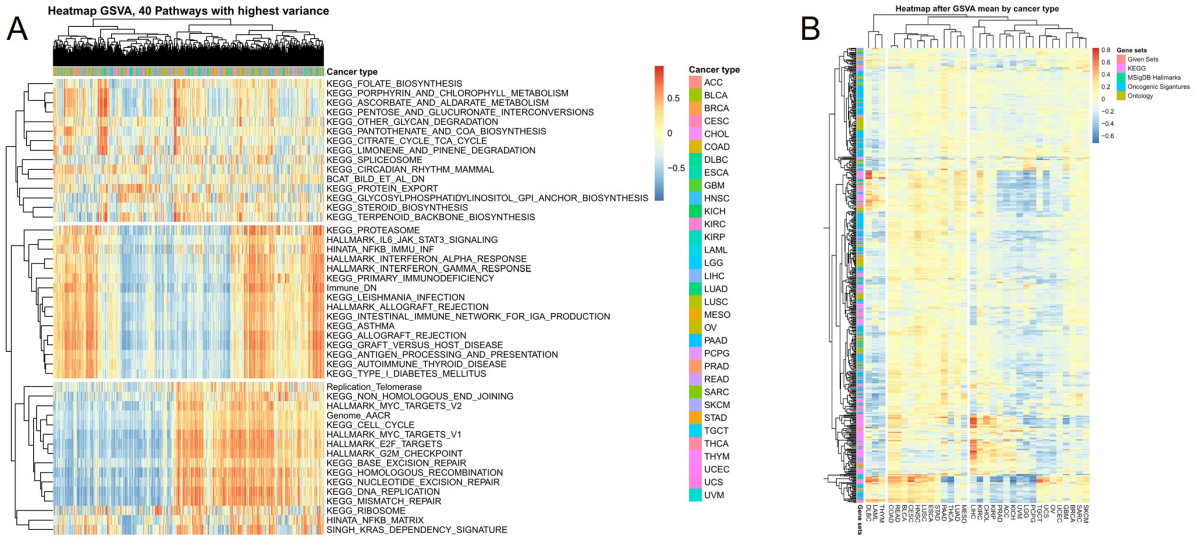

Figure 7: GSVA Top Var heatmap and mean over cancer types.

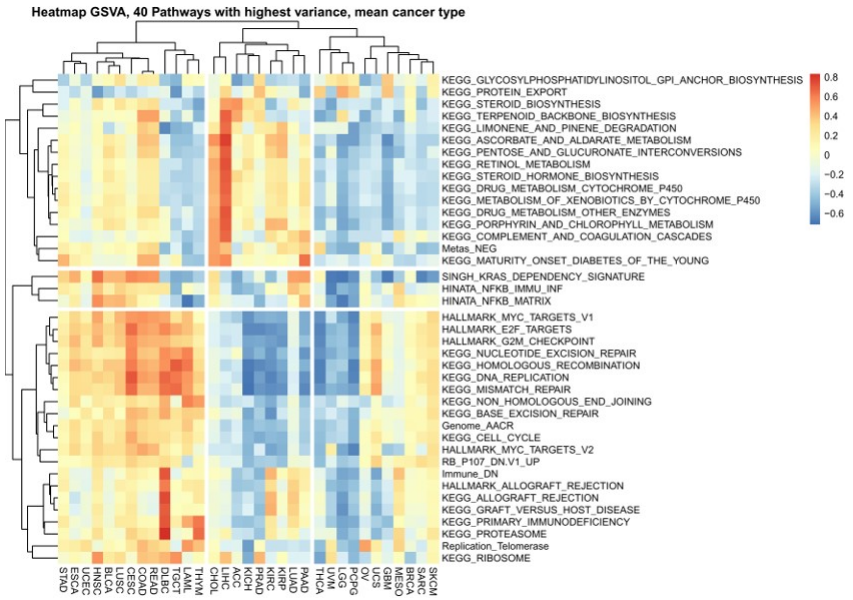

Figure 8: GSVA topVar combined with mean over cancer types.

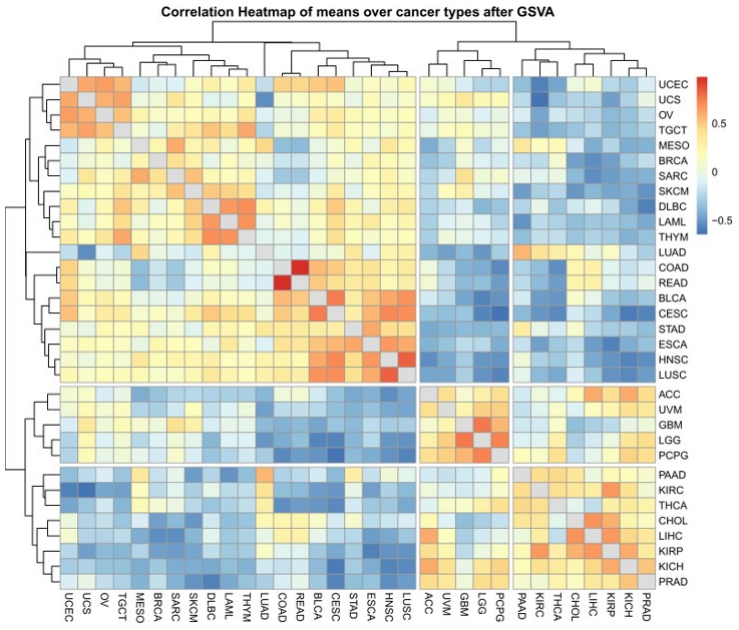

Figure 9: Correlation between cancer types after GSVA.

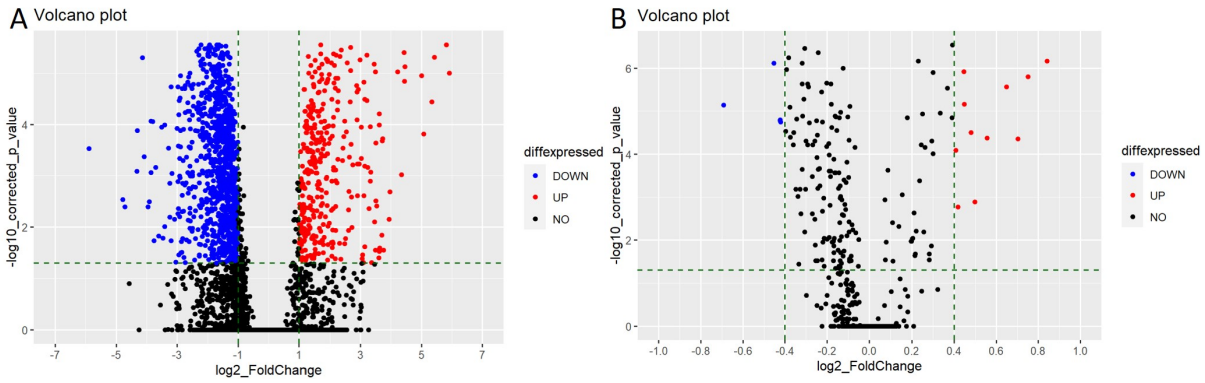

Figure 10: Volcano plots.

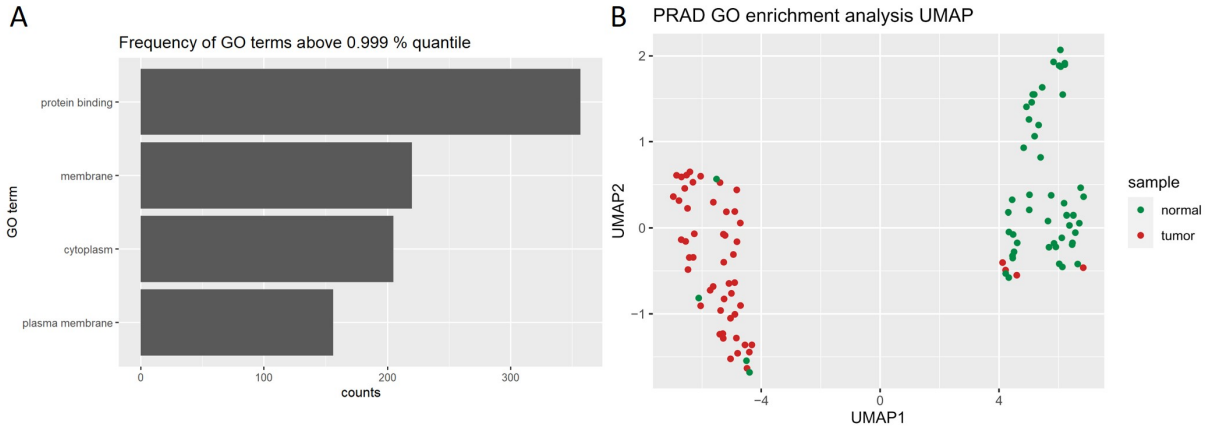

Figure 11: GOEA.

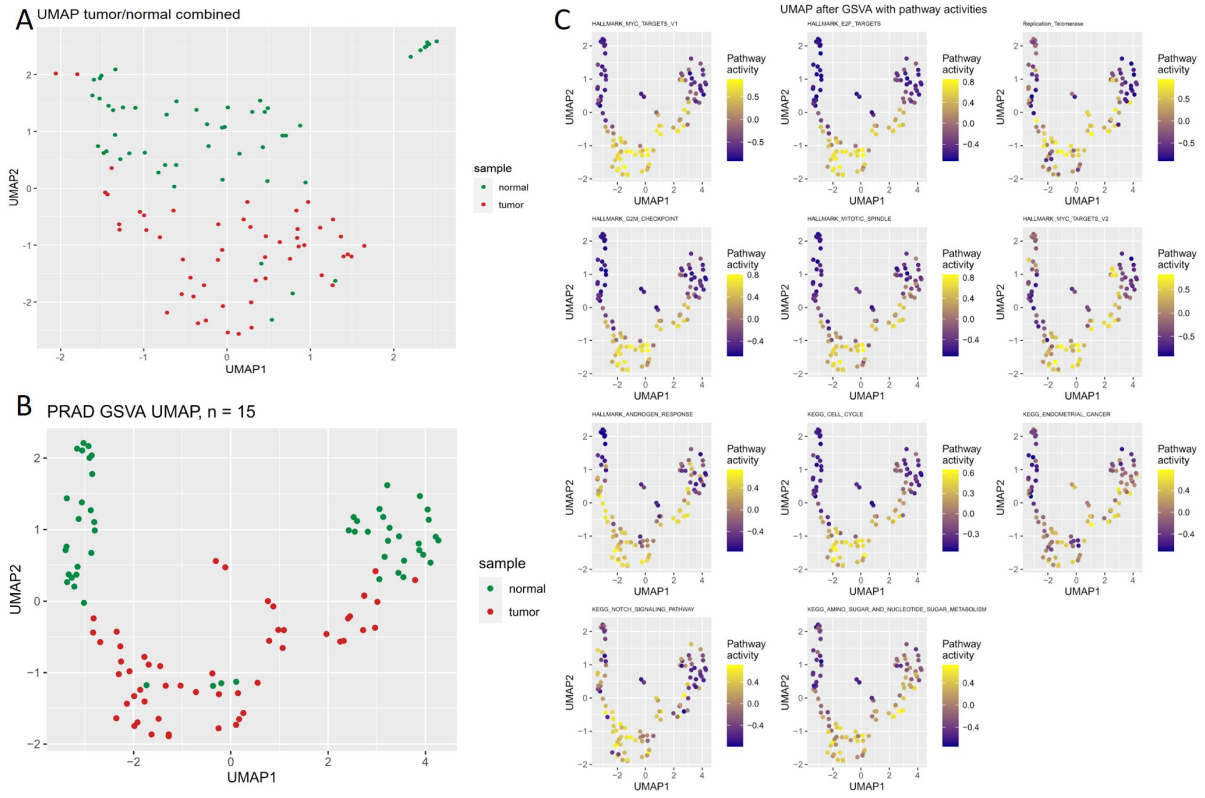

Figure 12: PRAD GSVA UMAP.