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

07/16/22

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 (Ahmad, Cherukuri, and Choyke 2021).

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, Carroll, and Cooperberg 2011). The most recurrent diagnosed prostate cancer type is the so-called prostate adenocarcinoma (PRAD) (Li et al. 2016). Common and effective ways to treat prostate cancers pose 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, Cherukuri, and Choyke 2021).

1.3 Analysis - an overview

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. 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, Castelo, and Guinney 2013; Subramanian et al. 2005). 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 (PCA) 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. 2015; 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. 2015; 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 a 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, where 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 to 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.

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

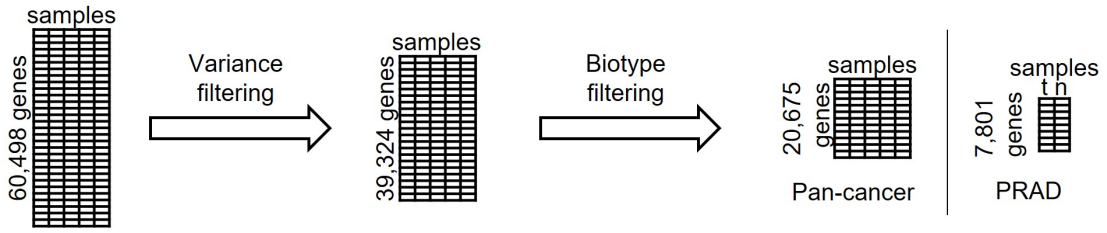

Figure 1: Preprocessing of RNAseq data.

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 of the specific 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. 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 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. (Hänzelmann, Castelo, and Guinney 2013). For this 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 function's 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 pathway's 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, nonparametric Wilcoxon Signed Rank Test was performed. This statistical test compares two paired groups that do

not need to be normally distributed. The computed p-value was 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 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).

2.5.2 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 analysis of gene sets

By calculating the mean and variance for each gene of the original RNAseq data frame, an overview of 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 2A) showing a smaller number of genes with a very large variance compared to the rest. 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 2B). From only the ten percent of genes with the highest variance in expression, 65 % were present within the pathways. A greater overlap was achieved at the focused PRAD analysis where 6,591 RNAseq genes (84 %) occurred within the gene sets (Figure 2C). Furthermore, the Jaccard metric combined with a heatmap shows the similarity between the gene sets used during the study (Figure 2D). Particularly the ontology gene sets display the greatest likeness, as well as a small part of the KEGG pathways and given pathway.

3.2 Pan-cancer analysis

The distribution of age at the initial diagnosis is left-skewed around the mean of 59 years (Figure 3A). This can also be seen in the quartiles. The first quartile lies at 50 years, the third at 69 years. Taking a look at the distribution of cancer types within the pan-cancer RNAseq data, BRCA brings the most (1094) patients (Figure 3B). The median number of samples per cancer type is 262 (mean = 295). The cancer types ACC, CHOL, DLBC, KICH, MESO, READ, UCS, and UVM have less than 100 samples.

For closer pan-cancer analysis PCA and UMAP were applied to the cleaned RNAseq data frame. The two-dimensional UMAP components were then plotted and colored according to the 33 different cancer types (Figure 4A). 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,

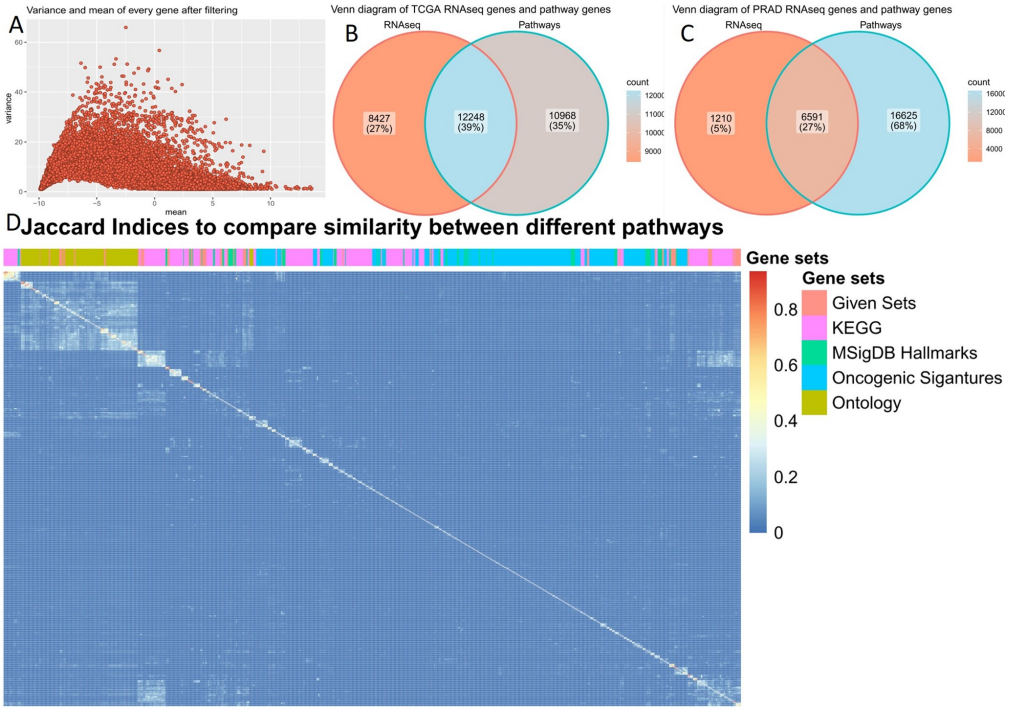

Figure 2: Preprocessing and gene sets. *A* Mean-Variance-Plot after preprocessing. *B* Venn diagram for pan-cancer RNAseq genes and gene sets. *C* Venn diagram for focused analysis RNAseq genes and gene sets. *D* Heatmap for pathway similarity based on Jaccard metric.

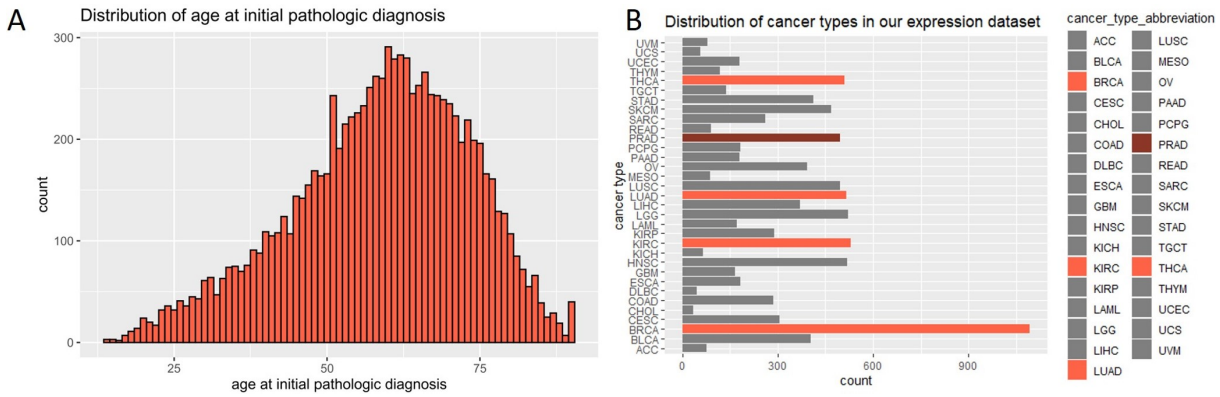

Figure 3: Descriptive analysis. *A* Age at cancer diagnosis. *B* Frequency of cancer types in pan-cancer analysis.

GSVA was applied with the former mentioned chosen pathways. Next, PCA and UMAP were also applied and plotted (Figure 4B) which allowed a comparison between the plots. After GSVA only ten separate 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 4C). 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 clusters showed rather decreased pathway activity (ACC, CHOL, GBM, KICH, KIRC, KIRP, LGG, LIHC, PCPG, PRAD, THCA). When UMAP was performed for each separate cancer type, some showed subclusters within a single cancer type. This was the case for BRCA, ESCA, KIRC, LAML, LUAD, TGCT, THYM. After GSVA these clusters were no longer visible except for BLCA, TGCT.

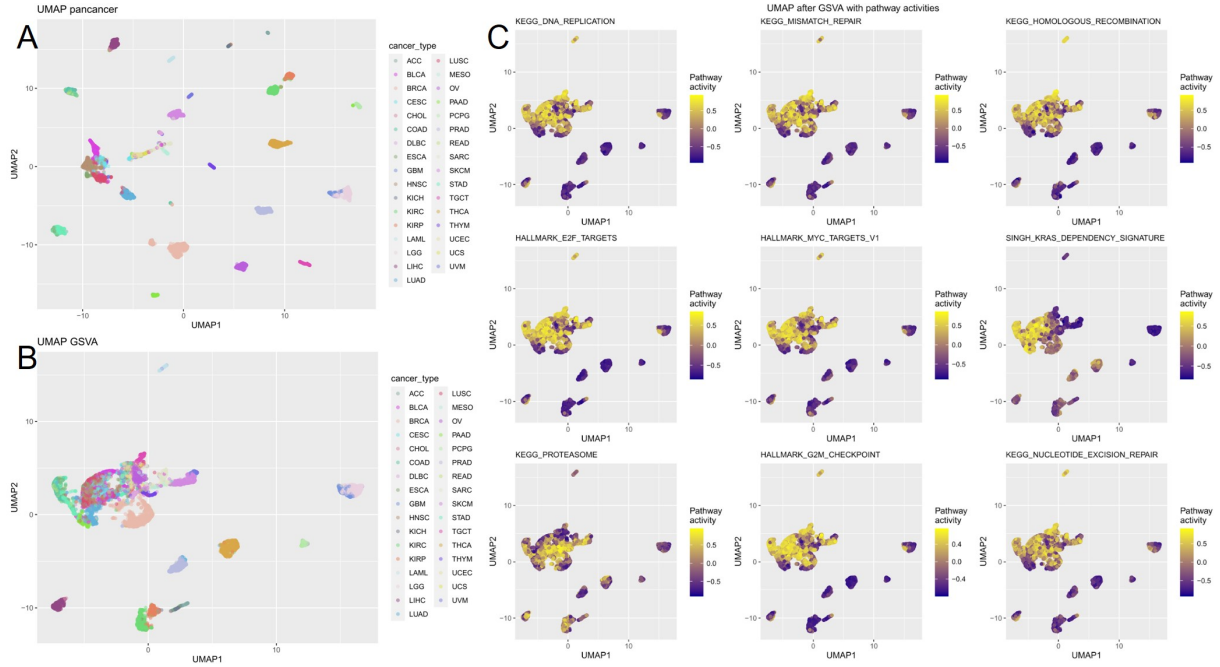

Figure 4: Pan-cancer analysis. *A* UMAP for RNAseq data with cancer types. *B* UMAP for pathway activity matrix (GSVA) with cancer types. *C* Pathway activity of 9 pathways with highest variance in pathway activity among all cancer types.

The pan-cancer clusters were mostly verified by a correlation heatmap, where the mean pathway activity for every cancer type was computed (Figure 5). The two main clusters match the clusters from the UMAP with two exceptions: PAAD and UVM were clustered

with the down regulated cancer types from the UMAP. A closer look at the correlation heatmap also shows very strong correlation between COAD and READ as well as between HNSC and LUSC.

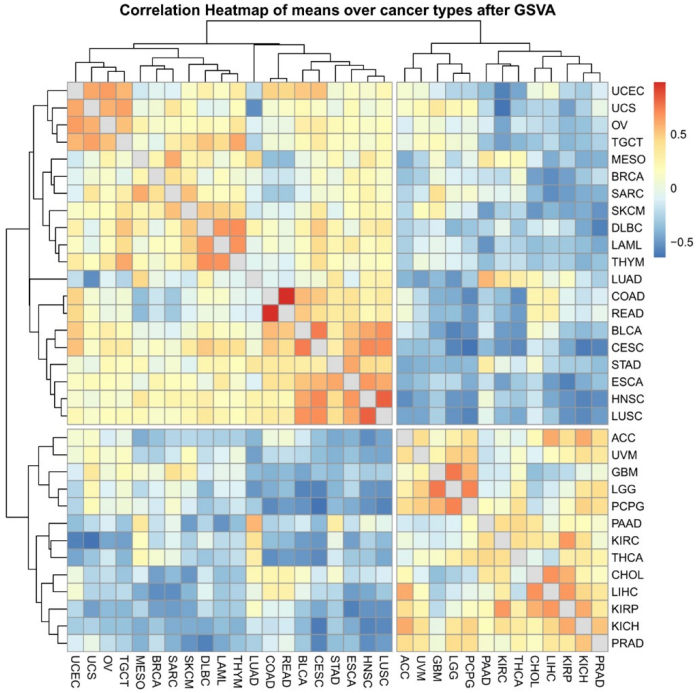

Figure 5: Pathway activity correlation between cancer types after GSVA.

Looking at the 30 pathways with the highest variance among all patients, three major cluster were identified (Figure 6). First, a rather metabolism-oriented group, second, a DNA and cell cycle-oriented group and third, an immune system-oriented group. For these 30 pathways there were no clusters for specific cancer types recognizable.

Based on the pathway activity matrix, another goal was to build a linear regression model. Here, it was tried to predict a pathway's activity based on other pathways' activities.

3.3 Focused analysis

To show differences in gene expression between tumor and normal PRAD samples, a volcano plot was created (Figure 8A). Each gene in this volcano plot is represented by one point. The 998 genes that are significantly downregulated in the tumor tissue are colored in blue. On the other side 347 genes, which are significantly upregulated in the tumor tissue are shown in red. So, there are almost three times as many downregulated than upregulated genes. For GOEA the GO terms of the 500 differentially expressed genes from the volcano plot with the highest significance, the genes above upper horizontal line in Figure XA, were determined. 484 genes were matched with at least one GO term. Most GO terms had a very low frequency.

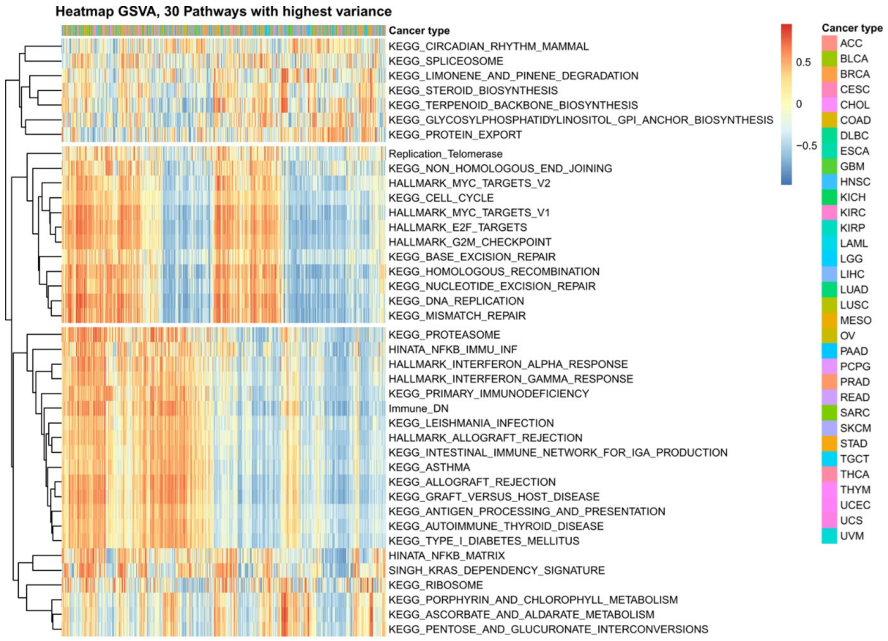

Figure 6: Pathway activty (after GSVA) per sample for the 30 gene sets with the highest variance among all cancer types.

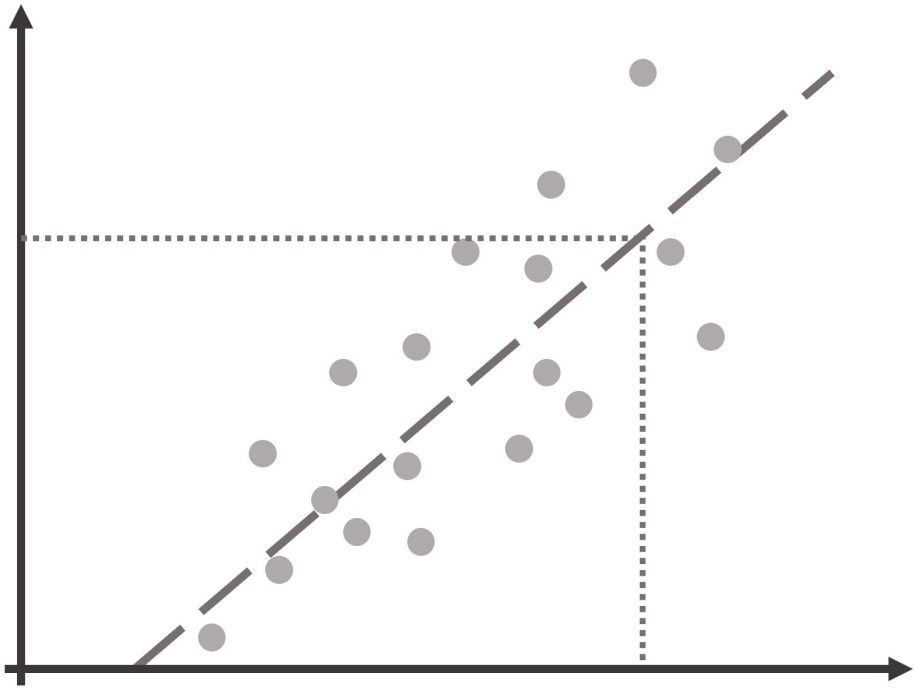

Figure 7: Linear regression.

Some GO terms with the highest frequency were for example “protein binding” (357 counts), “membrane” (221), “cytoplasm” (205), or “nucleus” (132) (Figure 8B). With the created gene sets per GO term GSVA was applied followed by PCA and UMAP. Plotting two UMAP components for each sample showed two separated cluster (Figure 8C left). Annotating the sample type revealed that the two clusters are formed by the tumor and normal samples. Both clusters show four samples from the opposite tissue source, respectively. Looking at the “normal” samples, there is a horizontal gap visible withing the “normal” cluster. This gap is far greater when performing the GSVA with the gene sets from pan-cancer analysis (Figure 8C right). Here, there are two separated clusters of normal samples and one widely scattered tumor cluster containing again four normal samples.

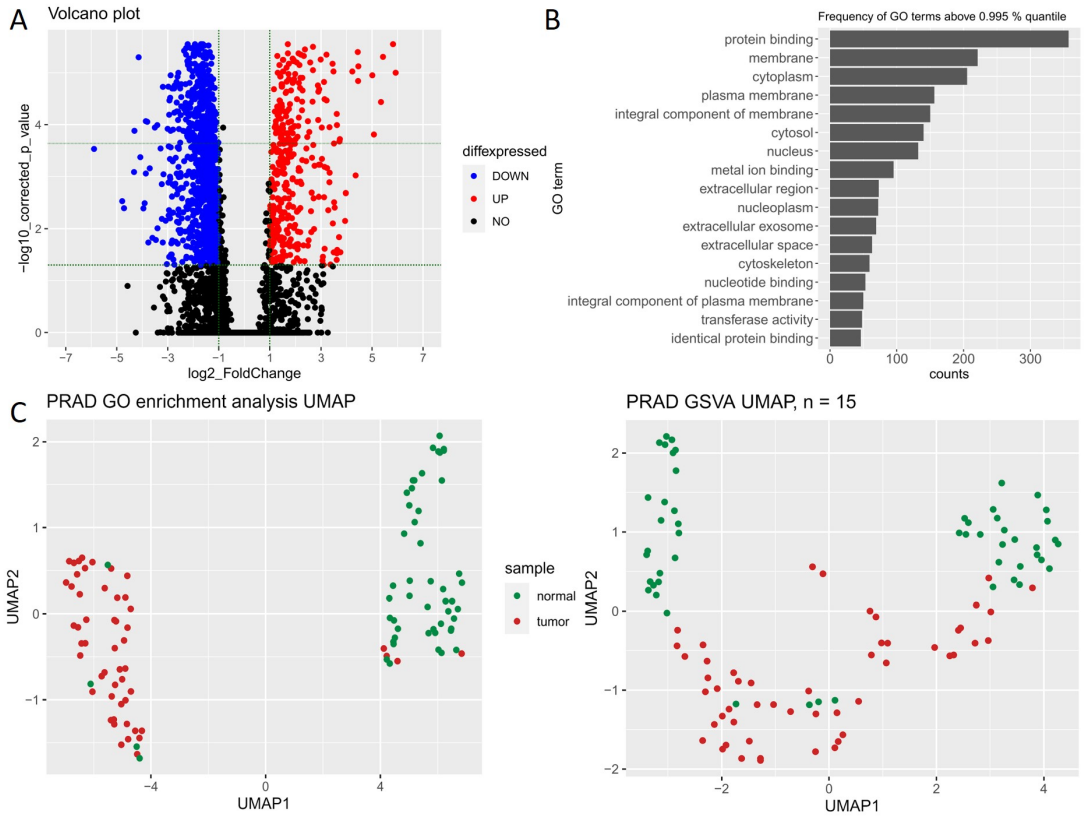

Figure 8: Focused analysis. *A* Volcano plot for differences in gene expression between tumor and normal PRAD samples. *B* GO term frequency of 500 differentially expressed genes with highest significance. *C* UMAP after GSVA with GO term gene lists (left) and with gene sets (right).

After GSVA, there was also an updated volcano plot created. Here each dot in the volcano plot represents one pathway (Figure 9A). Among our selected gene sets there are five downregulated and eleven upregulated gene sets in the tumor tissue. This means there are twice as much upregulated than downregulated gene sets. The eleven upregulated gene sets differ in the gene set size as well as the Fold Change (Figure 9B). The gene set

HALLMARK_MYC_TARGETS_V2 has the highest fold change of 0.848. However, this gene set contains 236 genes and is the largest of the upregulated gene sets and in total only 3.3% of the genes in the gene set are differentially expressed. The gene set HALLMARK_E2F_TARGETS has the second largest fold change with 0.751. Of the 218 genes 11.0% are upregulated. The gene sets HALLMARK_G2M_CHECKPOINT (13,7%), HALLMARK_ANDROGEN_RESPONSE (11.8 %) and HALLMARK_E2F_TARGETS (11.0%) contain the most differentially expressed genes. Dying the samples according to the pathway activity of six of the upregulated pathways, the differences between tumor and normal samples can also be visualized (Figure 9C). The two normal clusters are rather downregulated, but do not look identical. The tumor samples are mostly upregulated.

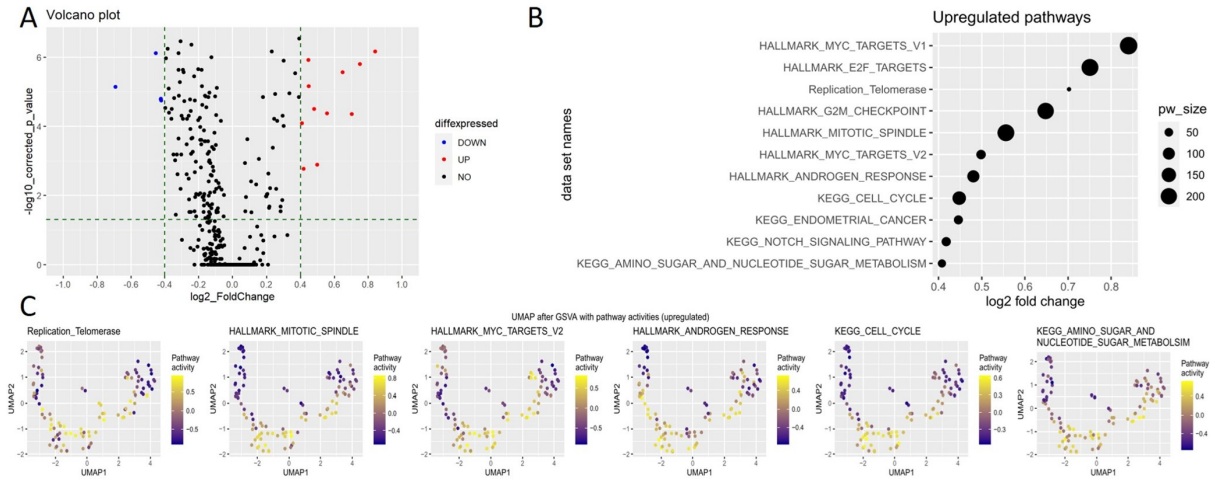

Figure 9: Pathway activity in PRAD. *A* Volcano plot for differences in pathway activity between tumor and normal PRAD samples. *B* Pathway size and fold change for upregulated pathways. *C* Pathway activity of 6 upregulated pathways in tumor and normal samples.

4 Discussion

4.1 Preprocessing and analysis of gene sets

The analysis with the Venn diagrams showed that not every RNAseq gene is represented at least once in any gene set. This has the consequence that information will be lost whenever a method only uses the genes included within the gene sets. This effect is stronger in the pan-cancer analysis than in the focused analysis. Not only shows this a downside of reducing the data into pathways but also stresses how much information about transcriptomics are yet to be discovered. The dissimilarity between the gene sets which was illustrated with

the Jaccard heatmap was the try to increase the overlap with the RNAseq genes. A careful analysis with the chosen gene sets was possible and performed.

4.2 Preprocessing and analysis of gene sets

The analysis with the Venn diagrams showed that not every RNAseq gene is represented at least once in any gene set. Consequentially information will be lost whenever a method only uses the genes included within the gene sets. This effect is stronger in the pan-cancer analysis than in the focused analysis. Not only does this show a downside of reducing the data into pathways, but also stresses how much information about transcriptomics is yet to be discovered. The dissimilarity between the gene sets, which was illustrated with the Jaccard heatmap, was the try to increase the overlap with the RNAseq genes. It shows that most ontology gene sets are more similar than others as they were clustered together. The reason might be that these gene sets are quite large since gene ontology terms like "chromosome" or "cell cycle" are very broad and include cellular structures.

4.3 Pan-cancer analysis

The comparison of the UMAP plots before and after GSVA showed that information seemed to be lost within the pathway activity matrix. On the one hand some cancer types of clusters were preserved, but on the other hand many clusters were lost or condensed into one enormous cluster. Dying the clusters according to highly variant pathway activity indicates the possibility to split the cancer types into two main groups. This is a very unsatisfying result, hinting that either the pathways used in the study did not cover enough of the RNAseq information or some cancer types are too alike to be separated based on the used pathways. Eventually, the hallmarks of cancer (Hanahan and Weinberg, 2011) are characteristics shared among most cancers. The next step would be to refine this analysis. The same trend was also visible when checking the cancer types by themself. Based on the clinical annotations, the cancer ESCA could be separated by the subtypes Esophagus Adenocarcinoma and Esophagus Squamous Cell Carcinoma and TGCT by Seminoma and Non-seminoma. But only TGCT and BLCA showing clusters after GSVA again emphasizes the loss of information. Another thing to consider with GSVA is other diseases. The expression of pathways substantially differs between different tissues or points of time. But not only cancer can change the expression patters within cells. Many other concomitant diseases might influence a tissue's transcriptome. The cancer type correlation heatmap supported mostly the findings from UMAP after GSVA. The two strongly correlating pairs were rectum adenocarcinoma (READ) and colon adenocarcinoma (COAD), and head and neck squamous cell carcinoma (HNSC) and lung squamous cell carcinoma (LUSC). READ and COAD not-only share a close localization,

but also share histological, genetic and methylation patterns (Tamas et al., 2015). HNSC and LUSC have multiple molecular characteristics in common, although they are genetically and heterogeneous diverse (Polo et al., 2016). The heatmap for 30 pathways with the highest variance in their activity allowed clustering into three main groups: metabolism, DNA/cell cycle and immune response. These three are part of the hallmarks of cancer (Hanahan and Weinberg, 2011). For a more precise analysis it would be necessary to group the used gene sets to the 10 hallmarks. Nevertheless, it would be different to makes these assignments since they also overlap with each other. Interestingly, in this heatmap there are no clear clusters for the 33 cancer types. This suggests that these pathways activity varies even among same cancers so much that they could not be clustered together.

Linear regression

4.4 Focused analysis: Prostate adenocarcinoma

4.4.1 LAURA volcano plot genes

Knowledge about GO terms of the differentially expressed genes between normal and tumor samples of PRAD allowed an insight into the cellular functions. Computing the GO term frequency showed that, changes caused by cancer cannot be pinned down to a restricted part of a cell, but rather causes changes within the whole cell. The GOEA resulted in two almost pure sample clusters. The four respective exceptions could not be explained with known clinical annotations. It is also to be considered that only the 484 most significantly differentially expressed genes were part of the GO gene sets. The role of the remaining RNAseq data remained unanswered. One thing that caught the eye was that the normal samples were showing a horizontal gap within the UMAP cluster. The gap between the two clusters with normal samples is by far greater after the GSVA with the gene sets. Again, no clinical annotations explained this effect. The GSVA used many more pathways and genes than the GOEA. The four normal samples within the tumor samples are also visible, here. In general, even though the Venn diagram showed a better overlap with the RNAseq data in the focused analysis than in the pan-cancer analysis, the clustering results are more scattered. The reasons are unclear. One explanation might be that the GSVA uses one combined data frame as input. On major difference was that the pan-cancer analysis only had cancer samples, the focused analysis used normal samples. Another interesting fact is that in the pan-cancer analysis PRAD was part of the scattered clusters with downregulated pathway activity. But some of the pathways reappeared in the focused analysis as upregulated compared to normal cells. Hence, the pathways seem to be upregulated to a certain level in PRAD, but not as much as compared to cancers of the enormous pan-cancer cluster. Perhaps the pathway activity is still too similar between tumor and normal cells in PRAD. The volcano plot for the pathways also shows that only a small number of pathways is

differentially expressed between the sample types. The GSVA considered by which degree the genes are up or downregulated. However, it does not correct against the pathway size which also effects the results.

4.4.2 LAURA volcano plot pathways

Downregulation of genes is more common -> gene knock out can be done by unspecific deactivating mutations. Gene upregulation is only possible by specific activating mutations Cancer gene sets contain more genes that are potentially upregulated -> specific activating mutations can cause cancer faster that deactivating mutations (2 alleles)

By taking the pathway size into account, one gene set containing plenty of differentially expressed genes that specifically stands out is the "HALLMARK_ANDROGEN_RESPONSE". Healthy prostate epithelium cells that potentially could give rise to PRAD have a unique metabolic background themselves, for they contain large amounts of zinc that prevent the oxidation of citrate, which thus can inhibit the citric acid cycle (TCA). Thus, making normal prostate epithelial cells dependent on different energy sources such as glycolysis. The control of the previous steps mentioned has been described to be controlled by the androgen receptor (AR) response. Once bound to an androgen hormone, such as testosterone, the AR acts as a transcription factor by traveling to the nucleus and thus activating the expression of metabolism-promoting genes in prostatic epithelial cells.
Mutations in the androgen response pathway could potentially wreak havoc in favor of cancer progression. These mutations could lead to the promotion of the TCA and therefore could lead to dependence on oxidative phosphorylation. This metabolic shift has been noticed in early prostate adenocarcinomas. Since this project revolved around RNA-seq data it can't definitively be stated that the cause for the upregulation of the "HALLMARK_ANDROGEN_RESPONSE" is the result of mutations, but it is apparent that crucial similarities between literature (Ahmad et al., 2021) and the RNA-seq analysis could be found. On top of that, the plausible differential expression within the genes of the "HALLMARK_G2M_CHECKPOINT" and the "HALLMARK_E2F_TARGETS" can also be substantiated by literature. The G2M checkpoint itself is important for halting the cell cycle once DNA replication has been faulty whereas E2F itself is a DNA promotor binding protein that promotes the transcription of protein-coding genes. Therefore, irregularities in this checkpoint could potentially lead to tumor growth progression (Alberts, 2015).
Generally, the upregulation of the tumor samples in the six UMAP graphs is concordant with the prior discussed revelations regarding the androgen response or even the G2M checkpoint-associated genes being upregulated. On the contrary, it is not completely clear why the normal samples in all these graphs can be separated into two clusters. It must be questioned though whether the gap between the two clusters increased because of their dissimilarities or because these samples simply didn't belong to the same cluster regardless of the severity of

dissimilarities.

---

5 Refrences

6 Appendix

6.1 used packages

6.2 Cancer type abbreviations

Ahmad, F., M. K. Cherukuri, and P. L. Choyke. 2021. "Metabolic Reprogramming in Prostate Cancer." Journal Article. *Br J Cancer* 125 (9): 1185–96. https://doi.org/10.1038/s41416-021-01435-5.

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

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." Journal Article. *Nature Genetics* 25 (1): 25–29. https://doi.org/10.1038/75556.

Bechis, S. K., P. R. Carroll, and M. R. Cooperberg. 2011. "Impact of Age at Diagnosis on Prostate Cancer Treatment and Survival." Journal Article. *J Clin Oncol* 29 (2): 235–41. https://doi.org/10.1200/JCO.2010.30.2075.

Cunningham, Fiona, James E Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Olanrewaju Austine-Orimoloye, et al. 2021. "Ensembl 2022." Journal Article. *Nucleic Acids Research* 50 (D1): D988–95. https://doi.org/10.1093/nar/gkab1049.

GOC, Gene Ontology Consortium. 2021. "The Gene Ontology Resource: Enriching a GOld Mine." Journal Article. *Nucleic Acids Res* 49 (D1): D325–d334. https://doi.org/10.1093/nar/gkaa1113.

Hanahan, D., and R. A. Weinberg. 2000. "The Hallmarks of Cancer." Journal Article. *Cell* 100 (1): 57–70. https://doi.org/10.1016/s0092-8674(00)81683-9.

———. 2011. "Hallmarks of Cancer: The Next Generation." Journal Article. *Cell* 144 (5): 646–74. https://doi.org/10.1016/j.cell.2011.02.013.

Hänzelmann, Sonja, Robert Castelo, and Justin Guinney. 2013. "GSVA: Gene Set Variation Analysis for Microarray and RNA-Seq Data." Journal Article. *BMC Bioinformatics* 14 (1): 7. https://doi.org/10.1186/1471-2105-14-7.

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
kableExtra	Zhu H (2021). _kableExtra: Construct Complex Table with 'kable' and Pipe Syntax_. R package version 1.3.4, <https://CRAN.R-project.org/package=kableExtra>
car	John Fox and Sanford Weisberg (2019). An {R} Companion to Applied Regression, Third Edition. Thousand Oaks CA: Sage. URL: https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Figure 10: R packages.

Abbreviation	Name
GBM	Glioblastoma multiforme
LGG	Brain Lower Grade Glioma
CESC	Cervical squamous cell carcinoma and endocervical adenocarcinoma
LUAD	Lung adenocarcinoma
COAD	Colon adenocarcinoma
LAML	Acute Myeloid Leukemia
BRCA	Breast invasive carcinoma
ESCA	Esophageal carcinoma
SARC	Sarcoma
KIRP	Kidney renal papillary cell carcinoma
STAD	Stomach adenocarcinoma
PRAD	Prostate adenocarcinoma
SKCM	Skin Cutaneous Melanoma
UCEC	Uterine Corpus Endometrial Carcinoma
HNSC	Head and Neck squamous cell carcinoma
KIRC	Kidney renal clear cell carcinoma
LUSC	Lung squamous cell carcinoma
THYM	Thymoma
LIHC	Liver hepatocellular carcinoma
THCA	Thyroid carcinoma
MESO	Mesothelioma
READ	Rectum adenocarcinoma
PAAD	Pancreatic adenocarcinoma
OV	Ovarian serous cystadenocarcinoma
TGCT	Testicular Germ Cell Tumors
PCPG	Pheochromocytoma and Paraganglioma
UVM	Uveal Melanoma
UCS	Uterine Carcinosarcoma
BLCA	Bladder Urothelial Carcinoma
ACC	Adrenocortical carcinoma
KICH	Kidney Chromophobe
CHOL	Cholangiocarcinoma
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma

Figure 11: Cancer type abbreviations and their names.

Jolliffe, Ian. 2011. "Principal Component Analysis." Book Section. In *International Encyclopedia of Statistical Science*, edited by Miodrag Lovric, 1094–96. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-04898-2_455.

Kanehisa, M. 2019. "Toward Understanding the Origin and Evolution of Cellular Organisms." Journal Article. *Protein Sci* 28 (11): 1947–51. https://doi.org/10.1002/pro.3715.

Kanehisa, M., M. Furumichi, Y. Sato, M. Ishiguro-Watanabe, and M. Tanabe. 2021. "KEGG: Integrating Viruses and Cellular Organisms." Journal Article. *Nucleic Acids Res* 49 (D1): D545–d551. https://doi.org/10.1093/nar/gkaa970.

Kanehisa, M., and S. Goto. 2000. "KEGG: Kyoto Encyclopedia of Genes and Genomes." Journal Article. *Nucleic Acids Res* 28 (1): 27–30. https://doi.org/10.1093/nar/28.1.27.

Köhler, Sebastian, Michael Gargano, Nicolas Matentzoglu, Leigh C Carmody, David Lewis-Smith, Nicole A Vasilevsky, Daniel Danis, et al. 2020. "The Human Phenotype Ontology in 2021." Journal Article. *Nucleic Acids Research* 49 (D1): D1207–17. https://doi.org/10.1093/nar/gkaa1043.

Korotkevich, Gennady, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. 2021. "Fast Gene Set Enrichment Analysis." Journal Article. *bioRxiv*, 060012. https://doi.org/10.1101/060012.

Levandowsky, Michael, and David Winter. 1971. "Distance Between Sets." Journal Article. *Nature* 234 (5323): 34–35. https://doi.org/10.1038/234034a0.

Li, Y., J. Mongan, S. C. Behr, S. Sud, F. V. Coakley, J. Simko, and A. C. Westphalen. 2016. "Beyond Prostate Adenocarcinoma: Expanding the Differential Diagnosis in Prostate Pathologic Conditions." Journal Article. *Radiographics* 36 (4): 1055–75. https://doi.org/10.1148/rg.2016150226.

Liberzon, Arthur, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. 2015. "The Molecular Signatures Database (MSigDB) Hallmark Gene Set Collection." Journal Article. *Cell Systems* 1 (6): 417–25. https://doi.org/10.1016/j.cels.2015.12.004.

Litwin, M. S., and H. J. Tan. 2017. "The Diagnosis and Treatment of Prostate Cancer: A Review." Journal Article. *JAMA* 317 (24): 2532–42. https://doi.org/10.1001/jama.2017.7248.

McInnes, Leland, and John Healy. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." Journal Article. *ArXiv e-Prints* 1802.03426.

Skloot, Rebecca, Cassandra Campbell, Bahni Turpin, and Random House Audio Publishing. 2010. "The Immortal Life of Henrietta Lacks." Audiovisual Material. Random House Audio,.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." Journal Article. *Proc Natl Acad Sci U S A* 102 (43): 15545–50. https://doi.org/10.1073/pnas.0506580102.