Ruprecht-Karls-Universität Heidelberg
Fakultät für Biowissenschaften
Bachelorstudiengang Molekulare Biotechnologie

Data Analysis Report
Topic 2, Group 1

Our way to a potential Biomarker

Data Science Project SoSe 2022

Supervisor: Carl Hermann
Tutor: Wangjun Hu

Autores: Ekin Ören, Yoana Onishtenko, Linh Trinh, Junona Sachov
Deadline: 18.07.2022

Contents

Abbreviations

Table 1: Abbreviations

Abbreviations	
BRCA	Breast Cancer Gene
DSS1	Split hand / Split foot protein 1
EGF-R	Epidermal growth factor receptor
GSE	Gene Set Enrichment Analysis
GSVA	Gene Set Variation Analysis
HR	Homologous Recombination
PCA	Principal Component Analysis
UMAP	Uniform Manifold Approximation and Projection for Dimensional Reduction
WU-CM	WU_CELL_MIGRATION
YTT-UP	YAMAZAKI_TCEB3_TARGETS_UP.

1 Introduction

1.1 Cancer and BRCA

Up until this day cancer remains a crucial research topic in the scientific fields. The focus of this project lies specifically on breast cancer. It is the most common malignant tumor and the second capital reason for cancer death among women worldwide (L. Wang, S. Zhang and X. Wang, 2021). Breast cancer can be identified during screening with mammography and clinically classified into four stages depending on the tumor size, if it has formed metastasis or spread to other types of tissues (Giuliano *et al.*, 2017). Whereas conventional therapies, such as surgery, radiotherapy, and chemotherapy, are the most common options they usually have limited success. More promising therapy approaches are personalized therapies like immunotherapy, i.e., HER2-targeted therapies (Clifton, 2021) or epigenetic therapies. These therapies are a promising tool for overcoming clinical resistance to traditional treatments for breast cancer but are still in the pre-clinical setting and show most promise in use in combinations with other treatments (Brown *et al.*, 2022). Especially in the era of personalized medicine and improved genomic analysis using methods of data analysis, i.e., to identify biomarkers for epigenetic therapies, underline the importance of the combined force of computational mathematical analyses via the creation of computer-generated models or simulations and the foundational experimental laboratory.

1.2 Hallmarks

> "The hallmarks constitute an organizing principle for rationalizing the complexities of neoplastic disease. They include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis."*Hallmarks of cancer: the next generation (D. Hanahan, R. A. Weinberg, 2011)*

The Hallmarks describe the distinction between tumor and regular cells, comprising six organic abilities received throughout the multistep development of human tumors. Underlying those hallmarks are the “emerging hallmarks” and the “enabling hallmarks,” like genome instability, which generates the genetic variety that expedites their acquisition, and inflammation, which fosters multiple hallmark functions.

1.3 Methabolic pathways - ascendany of tumerigenesis

An additional core point of our project is the focus on metabolic pathways since metabolic activities are altered in cancer cells relative to normal cells.These reprogrammed activities improve cellular fitness to provide a selective advantage during tumorigenesis. (@ Hanahan and and, 2011). The significant differences in the relative use of different types of energy production in normal cells and tumors, observed by O. Warburg manifests itself in tumor cells during the glycolysis, where pyruvate is mainly converted to lactic acid and energy is produced anaerobically. This phenomena

takes place even if there is sufficient oxygen for supporting the mitochondrial function. Hence, the conversion of pyruvate to lactic acid by fermentation, even in the presence of oxygen, is called aerobic glycolysis or the Warburg effect (Coller, 2014).

1.4 Mathematical Tools

In order to perform a significant examination of our data we implemented different mathematical tools such as dimensional reduction, Wilcoxon test, Jaccard index, Bonferroni correction and fold change. First, we started with computing the variance, defining how far our samples are from the average value and continued with dimensionality reduction via Principal Component Analysis (PCA) and Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP). The goal of PCA is to extract important information from a data frame and represent it as a new set of orthogonal variables called principal components (Abdi and Williams, 2010). In comparison to PCA, the UMAP method reduces dimensions of the data nonlinearly, indicating the highest reproducibility and the most meaningful organization of cell clusters (Becht *et al.*, 2019). For the comparison of BRCA and other tumor types, we computed p-values via Wilcoxon paired test. Wilcoxon paired test is a nonparametric statistical hypothesis used to compare the position of two populations using two matched samples (Conover, 1999). The p-value helps us to estimate the probability of getting a test statistic as large or larger assuming both distributions are the same [Hung:1997wx]. Since in some cases p-value correction was necessary, we used the Bonferroni method to reduce the chances of obtaining false-positive results (type I errors) when multiple pairwise tests are performed on a single set of data (Ranstam, 2016). Furthermore, we also used Jaccard index to compare members for two sets to see which members are shared and which are distinct (Fletcher and Islam, 2018). For measuring how much quantity changes between the original and subsequent measurements we computed fold change, which is defined as the ratio of the two quantities (Feng *et al.*, 2012).

1.5 Project Outline

Aiming to investigate the possibility of identifying breast cancer-relevant biomarkers, during the course of this project, we explored the gene expression data of a broad selection of patients in a step-by-step manner summarized in our project outline (supplements). We aimed to gradually concentrate the descriptive capacity of our data by data cleaning and dimensional reduction methods, including PCA and UMAP, eventually grouping and collectively scoring genes in the context of pathways by the GSVA method. Afterwards, we compared the respective differential activation scores of several tumor types. Furthermore, we plotted the pathways in the context of p-value and fold change, thus identifying the pathways most relevant to BRCA and tumor progression and possible trends connecting these pathways. With this information, we selected one pathway we hypothesize could be scientifically and clinically relevant as a biomarker. Hence, we built a regression model predicting the behaviour of this pathway by selecting other highly relevant pathways. In conclusion, we propose that further use and optimization of this regression model could lead to the identification of previously unknown interconnections between these pathways, as well as the identification of biomarkers for targeted therapy and diagnostic tools employed in the treatment of breast cancer.

2 Material and Methods

2.1 Material

2.1.1 Data Sets

For this project we were provided with four data sets. The first one being the "tcga_tumor_log2TPM" a data frame containing about 60000 rows, representing the gene expression data from RNA-seq for almost 10,000 TCGA (The Cancer Genome Atlas) cancer patients, representing 33 different tumor types in the columns in the data frame. The second data set "tcga_tumor_annotation" is also a data frame which contains all of the patients from the previous datasets with 37 clinical annotations (for example gender, race, tissue type etc.) The third provided data set is a R object "tcga_tumor_normal_datascience_proj_2022" containing expression data of matched tumor and normal tissue for five tumor types (BRCA, KIRC, LUAD, PRAD, THCA), each of them are specific for one tumor type, each of them have three smaller dataframe showing gene expression of tumor samples and normal samples from the same patients, and clinical annotations for these patients. Last given data set is a list "hallmarks_genesets" containing 46 Hallmark genesets, each of them containing various amounts of genes associated with this set.

2.1.2 Used Packages

{We used the R version 4.2.0 and the table of used packages are attached as supplementary (see Sup.8).}

2.2 Methods

2.2.1 Data Exploration

2.2.1.1 Cleaning and Filtering

Data cleaning is the process of deleting inaccurate, corrupted, malformed, duplicated, or incomplete information in a dataset. First, we checked for missing values in our datasets. Secondly, we removed the low variance genes in the "tcga_tumor_log2TPM" dataset and then performed biotype filtering.

2.2.1.2 Descriptive analysis

Descriptive analysis is a type of data analysis that helps you describe, visualize, or summarize data points in a constructive way to explore datasets and observe possible patterns. Firstly, we computed the histogram that consists of numbers of missing values for each variable in the "tcga_tumor_annotation." Furthermore, we visualized the variety of multiple annotations in the same dataset via histograms. For our "tcga_tumor_log2TPM" dataset we did PCA and UMAP to reduce the dimension and explore the structure of the data.

2.2.2 Selecting and Retrieving Pathways from Database

To create our pathway matrix we first collected all the hallmark pathways in our "hallmarks_genesets" list. Additionally, we identified in the literature 46 metabolic pathways that behave differently in the presence of cancer and retrieved them from the MSigDB (Molecular Signature Database). In the last stage, we inserted all the pathways from the "C2: curated gene sets" (6366 gene sets) database. The gene sets in the pathway matrix are carefully selected from a variety of sources, consisting of online pathway databases and biomedical literature. In order to implement this matrix for further analysis, we reduced the number of pathways by filtering out the pathways that have less than 15 overlapping genes in them.

2.2.3 Gene Set Scoring (GSVA)

Gene Set Variation Analysis (GSVA) is a GSA method considered to contribute to the current need for GSE methods for RNA-seq data. It is applied for the estimation of variation of pathway activity over a sample population in an unsupervised manner and provides higher power compared to other methods for detecting subtle changes in pathway activity over samples (Hänzelmann *et al.*, 2013). This method was applied on genesets constructed by matching the cleaned "tcga_tumor_log2tpm" data frame genes to their respective pathways by matching Ensembl-IDs through the MSigDB database. Thus, an activity score was assigned and each pathway per patient. A second pathway matrix was constructed from only the BRCA patients where normal tissue and tumor tissue were both included. Lastly, we checked via a Venn diagram the percentage of overlap of the genes from the original data set ("tcga_tumor_log2tpm") and different stages of our pathway matrix.

2.2.4 Pan Cancer Analysis

Pan cancer analysis includes the assessment of frequently mutated genes and other genomic observations that are common to many different types of cancer, regardless of the origin of the tumor. First of all, we computed UMAP for the cleaned "tcga_tumor_log2TPM' ' dataset and for the pathway activity score data frame. By using the annotation information for cancer type and pathological stage we coloured both plots to assess the distribution of annotations within clusters. Moreover, we created a heatmap for the activity scores in our pathway matrix and categorized the pathways into Metabolic, Hallmark or C2: curated genes.In order to identify pathways with large fold changes that are also statistically and biologically significant, we plotted the BRCA pathway matrix in comparison to other tumor types and the final pathway matrix, on volcano plots where each dot was representative of a pathway. Pathways were separated by vertical and horizontal thresholds and annotated accordingly. Lastly, correlation analysis was carried out in order to identify overlapping genes between pathways using the Jaccard index.

2.2.5 Focused Analysis

In Focused Analysis we investigated BRCA tumor and normal tissue via UMAP plots annotated by menopause, pathological stage and sample type in order to test for significant clustering of these parameters. A Volcano plot between the tumor and non-tumor samples of the BRCA pathway matrix was carried out and subsequently split by vertical and horizontal thresholds. Additionally, we computed a heatmap out of activity score in BRCA for tumor and normal samples for all pathways and the top 200 differentially active pathways.

2.2.6 Regression Analysis

Focused analysis provided us with a list of potentially significant pathways for regression analysis. A comparison of these pathways with previously identified significant pathways in pan-cancer analysis further narrowed the selection of our potential regression analysis pathways. Literature was considered to identify the potential clinical significance of the overlapping pathways, which lead to the selection of one pathway

as a prediction target. The rest of the significant pathways were used as variables to predict the chosen pathway. The patients were split into a training and a testing group to train the algorithm and compute the accuracy of the prediction consecutively. Afterwards, parameter reduction by the exclusion of low significance parameters and k-fold cross validation was carried out to solve overfitting issues and increase precision.

3 Results

3.1 Data cleaning and exploration

As previously mentioned our first task was to reduce our "tcga_tumor_log2tpm" to a more manageable one, in order to perform significant analysis. Starting with a data frame consisting of 60498 genes and 9741 patients we reduced the number of the genes to 14674. The starting point was to explore the distribution of the data by calculating the variance and the mean from all available patients and plotting them against each other as in Sup.1.A-C. After checking our data for missing values, we didn't encounter any, hence, filtering for missing values was not necessary. Aiming to concentrate our analysis only on relevant genes, we performed biotype filtering removing all the non-protein coding genes from our data. In order to determine the biotypes of all the genes, we used the biomaRt package (Sup.8) and the ensembles IDs we were given in our "tcga_tumor_log2tpm" data. As a second step, we also cleared all the low variance genes since they are considered not crucial for the distribution (Sup. 8C). Considering that we removed more than 70% of the genes in the original data it was necessary to guarantee that the behaviour of our data remained the same. This we confirmed by plotting the variance against the mean after each step of cleaning (Sup.1) Additionally we investigated the significant information in the "tcga_tumor_annotation" data frame via bar plots. In contrast to the "tcga_tumor_log2tpm" data frame here we encountered a lot of missing information as seen in Sup.2.A. Consequently we excluded the annotations with a significant amount of missing values from our plan for further analysis. As shown in Sup.2 we have potential clustering and regression analysis information in the form of histological grades, race, new tumor event, margin status, clinical stages, pathological stages and cancer type.

3.2 Dimensional reduction

After successfully cleaning our data and analysis of our annotations we continued exploring our data by dimensional reduction. Our first attempt consisted of linear dimensional reduction of the genes in our filtered "tcga_tumor_log2tpm" data via PCA. First, we computed 14 principal components and tested which of them carry the most significant information using the Elbow method (Liu and Deng, 2021) which brought us to the conclusion that we can plot our data using the first two principal components. For more detailed analysis we also applied cancer type annotations to colour respectively each patient in our plot. As visualized in Fig 1.A. We encountered two clusters, however, the information in the annotation is not distributed in a polarized manner. Aiming to observe more significant clustering we also performed non-linear reduction via UMAP (Fig. 1.B). Subsequently, we were able to see very distinct clusters and polarized distribution in terms of tumor type. Furthermore, we encountered several clusters consisting of only one tumor type (BRCA, PCPG, TGCT, THCA). UMAP proved to be more successful in terms of keeping the information in the reduced dimensions, so we continued using it for further analysis.

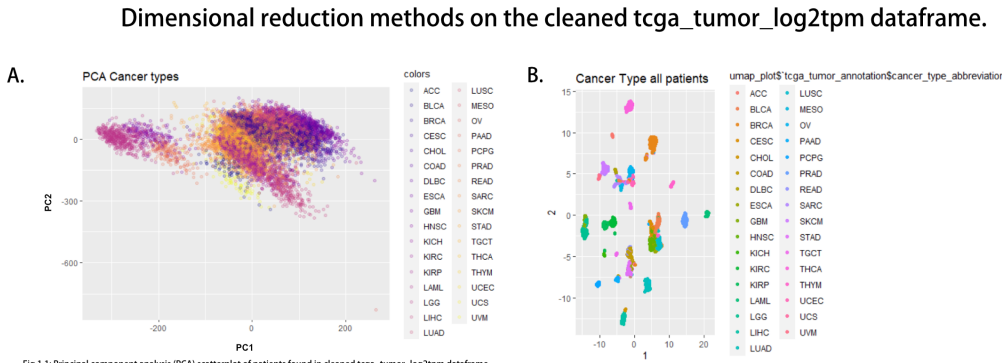


Fig 1.1: Principal component analysis (PCA) scatterplot of patients found in cleaned tcga_tumor_log2tpm dataframe. Dimensional reduction through principal component analysis was carried out on the cleaned tcga_tumor_log2tpm dataset, after which the most descriptive components (PC1 and 2) were plotted against each-other and colored with their respective annotations showing their tumour types.

Fig.1.2: Uniform Manifold Approximation and Projection (UMAP) scatterplot of patients found in cleaned tcga_tumor_log2tpm dataframe. Dimensional reduction through Uniform Manifold Approximation and Projection was carried out on the cleaned tcga_tumor_log2tpm dataset, after which the most descriptive components (PC1 and PC2) were plotted against each-other and colored with their respective annotations showing their tumour types.

3.3 Gene set scoring

After creating our pathway list and selecting the gene sets accordingly we matched our filtered "tcga_tumor_log2tpm" data with the corresponding pathways. An activity score was assigned and the final pathway matrix computed, consisting of patients as columns and pathway names as rows. The same was performed with only BRCA patients. In order to visualize and quantify how many of the genes are kept within the pathway matrix in each stage of the pathway matrix creation we created venn diagrams for each stage of our pathway matrix creation (Fig 2). This allows us to see the percentage comparisons of overlapping genes between our cleaned tcga_tumor_log2tpm dataframe. In Figure 2.A and 2.B we see the overlap of all available pathways we had collected, both separately and grouped. The number of pathways was further reduced to ease computational load, address concerns about the inclusion of geneset scores of very few (lower than 15 genes) pathways leading to an overrepresentation of certain genes, as well as concerns regarding misrepresentation of pathways that had very little matching genes with our cleaned tcga_tumor_log2tpm data frame. Pathways which had less than 15 matching genes with our cleaned tcga_tumor_log2tpm dataframe were excluded, and the remaining 1010 pathways were subjected to GSVA scoring to generate a final pathway matrix of 1010 pathways and 9741 patients. The overlap of all genes in the final pathway, the previous selection of all available pathways, and our cleaned tcga_tumor_log2tpm data frame can be seen in fig 2.C, and a more direct comparison of the final pathways and our cleaned tcga_tumor_log2tpm dataframe in fig 2.D.The initial 54 % overlap was reduced to 9 % after the exclusion of pathways carrying less than 15 overlapping genes. In order to test the significance of our pathway matrix, we plotted UMAP annotated by cancer type and compared it to our cleaned"tcga_tumor_log2tpm" data (Sup.3). We could observe similar clustering of both which verified the biological significance of the pathway matrix. Additional confirmation we encountered in the comparison between UMAP plots for our data cleaned "tcga_tumor_normal_datascience_proj_2022" for BRCA patients and on the corresponding pathway activity matrix (Fig 3). The plot represents every dot as a patient and gives us output for two very distinguished clusters. By applying annotation information for tumor and normal tissue we proved that the clustering in both gene expression data and pathway matrix activity data corresponds to similarities within tumor and normal tissue annotations.

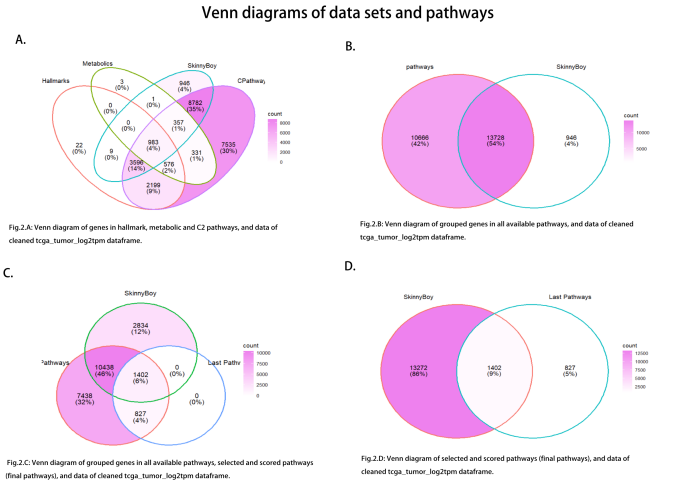


Fig.2.A: Venn diagram of genes in hallmark, metabolic and C2 pathways, and data of cleaned tcga_tumor_log2tpm dataframe.

Fig.2.B: Venn diagram of grouped genes in all available pathways, and data of cleaned tcga_tumor_log2tpm dataframe.

Fig.2.C: Venn diagram of grouped genes in all available pathways, selected and scored pathways (final pathways), and data of cleaned tcga_tumor_log2tpm dataframe.

Fig.2.D: Venn diagram of selected and scored pathways (final pathways), and data of cleaned tcga_tumor_log2tpm dataframe.

Fig.2.A-D. : Seen are the percentages of overlapping genes between our cleaned tcga_tumor_log2tpm dataframe and different stages of our pathway matrix, allowing us to visualize and quantify how many of the genes are kept within the pathway matrix in each stage of the pathway matrix creation.

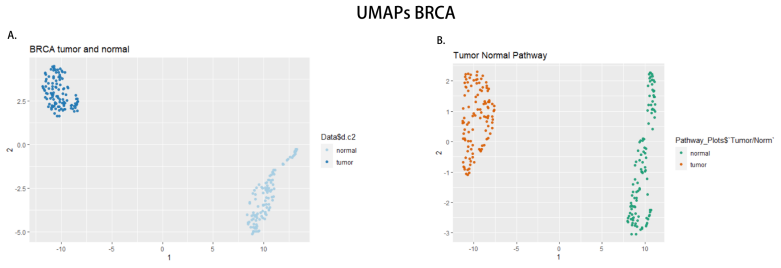


Fig.3.A: UMAP of BRCA tumor and normal tissue for gene expression values.
Dimensional reduction method was applied on our BRCA gene expression data and colored respectively tumor and normal tissue

Fig.3.B: UMAP of BRCA tumor and normal tissue for pathway activity values
Dimensional reduction method was applied on our BRCA pathway activity matrix and colored respectively tumor and normal tissue

3.4 Pan cancer analysis

In the pan-cancer analysis we investigated common genetic abnormalities in many different cancers. Initially, we coloured the UMAP of the final pathway matrix with a focus on interesting annotations in order to identify a potential annotative target for regression analysis. In Sup.4.A clear clusters of different cancer types including BRCA can be seen. The UMAP of the pathway matrix was then coloured by pathological stages for all tumor types in Sup.4.B. Lastly, only BRCA patients are coloured by gender and menopause status in Sup.4.C and Sup.4.D. Unfortunately, there was no clear cluster to be seen, thus no conclusive information was delivered.

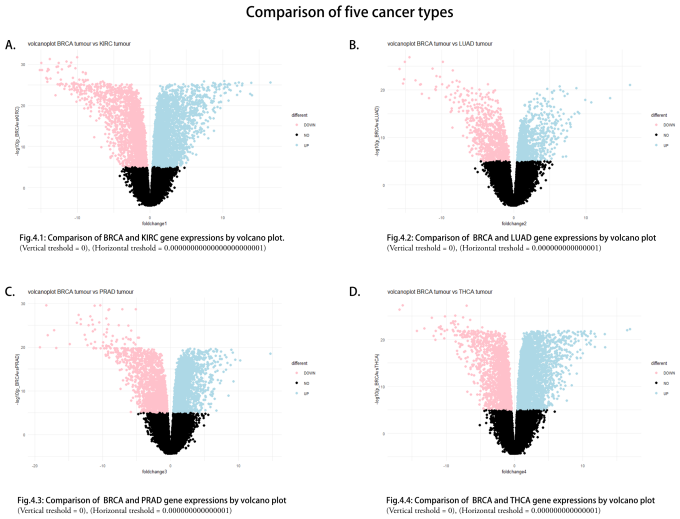


Fig.4.1: Comparison of BRCA and KIRC gene expressions by volcano plot. (Vertical treshold = 0), (Horizontal treshold = 0.000000000000000000001)

Fig.4.2: Comparison of BRCA and LUAD gene expressions by volcano plot (Vertical treshold = 0), (Horizontal treshold = 0.000000000000001)

Fig.4.3: Comparison of BRCA and PRAD gene expressions by volcano plot (Vertical treshold = 0), (Horizontal treshold = 0.0000000000000001)

Fig.4.4: Comparison of BRCA and THCA gene expressions by volcano plot (Vertical treshold = 0), (Horizontal treshold = 0.0000000000000001)

Afterwards, the four other major cancer types present in our database were compared with BRCA in terms of gene expression by plotting each gene in a scatter plot of respective p values scaled by logarithm, and fold changes. The result is shown in Fig.4. A horizontal threshold that separates the visibly differing genes from non-differing genes was selected individually for each plot, as well as a vertical threshold separating the upregulated (right) genes from the downregulated (left) was generated to separate and annotate the data, the respective threshold values are given below the individual volcano plots. The comparatively upregulated (in pink), downregulated (in blue) and non-changing genes (in black) were labelled as such,and coloured accordingly. This separation provided us with lists of genes with large fold changes, representing the degree of quantity change between the expression of the gene in BRCA and in other cancer types, as well as being statistically significant. Different cancer types were compared not only by gene expression but also by pathway activity. To do so, fold changes and p-value are calculated and each pathway is plotted in a scatter plot, as shown in Sup.5.A. This provided us with lists of gene sets that are commonly or differentially active between BRCA and PRAD as well as between BRCA and KIRC. The target for regression analysis is coloured in red in the plot, which we will go into detail in the next part of the report. In the next step, the mean GSVA score for all pathways was calculated, the highest of them being KEGG_REGULATION_OF_AUTOPHAGY, indicating that a lot of the genes belonging to this gene set have a higher expression level than the genes outside this geneset in cancer development. Then twenty highest-GSVA scoring pathways were plotted in Fig.5, indicating potential biomarkers for cancer. In the last step of the pan-cancer analysis, we plotted a heatmap for the activity score of all pathways and all patients and categorized them into 33 tumor types (Fig.6.A). Then the means of the activity score were computed for all of the pathways and were plotted in a heatmap, as shown in Fig.6.B. This enables us to identify pathways that are highly active or less active in which cancer type. Finally, only the highest variant pathways are plotted in Fig.6.C, indicating the most crucial pathways in cancer development.

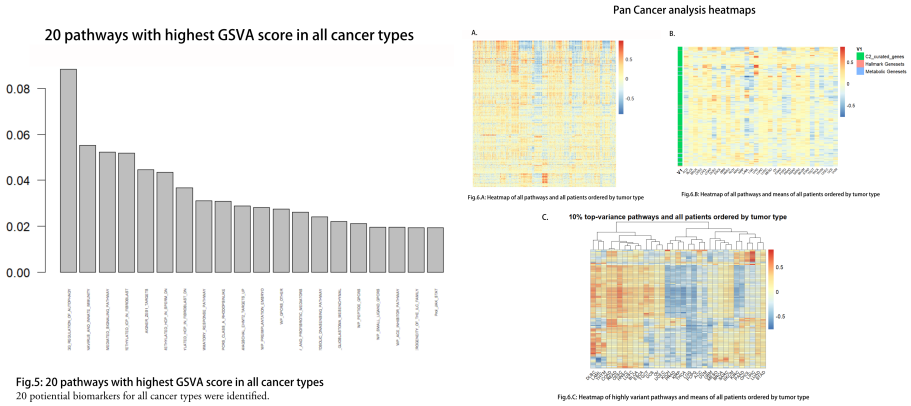


Fig.5: 20 pathways with highest GSVA score in all cancer types
20 potiential biomarkers for all cancer types were identified.

Fig.6.A: Heatmap of all pathways and all patients ordered by tumor type

Fig.6.B: Heatmap of all pathways and means of all patients ordered by tumor type

Fig.6.C: Heatmap of highly variant pathways and means of all patients ordered by tumor type

3.5 Focused analysis

In the focused analysis we examined breast cancer in greater detail. The correlation values of all of the pathways for BRCA patients were computed and plotted in a heatmap in Sup.6.A. Then the overlapping number of genes in each geneset was computed. These scores were normalized by the Jaccard index and plotted in a heatmap in Sup.6.B, which displays that the genes in these genesets do not overlap greatly. However, Sup.6.A shows that these pathways are strongly correlated with each other. In the next step, pathway activity scores for all pathways in normal tissue and tumor tissue from BRCA patients were plotted in a heatmap in Fig.7.B. These pathways are categorized into C2_curated_genes, Hallmark Genesets as well as Metabolic Genesets. We identified five distinct clusters of genesets and retrieved the names of the pathways in these clusters in order to further investigate them. Finally, we compared BRCA pathway activity scores between tumor and normal tissue and plotted them in a volcano plot, as shown in Fig.8.A. By choosing the threshold so high, we were able to visually see the most upregulated and most downregulated pathways in BRCA patients. Afterwards, we compared the pathway activity score between BRCA and all tumor types. The result can be seen in Fig.8.B. After the selection of YAMAZAKI_TCEB3_TARGETS_UP (YTT-UP) as our regression analysis target, we coloured the YTT-UP showing that it is one of the most upregulated gene sets in BRCA tumor and normal tissue comparison. By colouring the same pathway in the volcano plot of BRCA pathway activity in tumor tissue and the final pathway matrix for tcga_tumor_log2tpm we again identified it as upregulated.

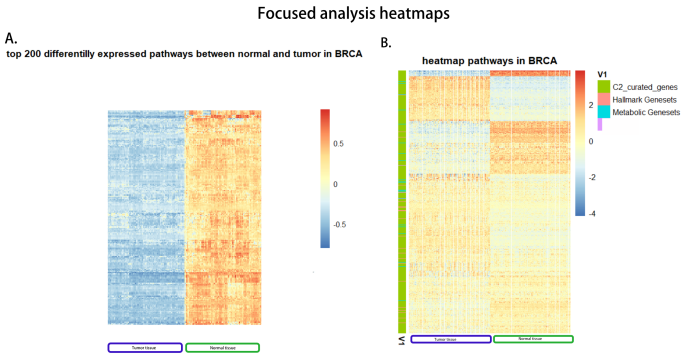


Fig.7.A: Normal tissue and tumor tissue from BRCA patients plotted against top 200 differentially expressed pathways.
Tumor samples are shown on the left and normal samples on the right. The clear distinction in genset scores between tumor and normal, and a clear clustering of pathways can be seen. Five pathway groups appear distinct. Group 1 appears intensely active in normal tissue, group 3 is also significantly more active in normal tissue than in tumor tissue. Group 2 appears more active in tumor tissue, while group 4 and 5 do not show a significant distinction in activity. Pathways are annotated and split by expression patterns in order to separate group 1-4.The distinction in pathway activity between tumor (left) and normal (right) tissue can be seen clearly.

Fig.7.B: Normal tissue and tumor tissue from BRCA patients plotted against all pathways, selected by computing fold changes.
We determined the 200 pathways with the highest fold changes between tumor and normal tissue activation scores, and plotted them in a heatmap. The distinction in activity between tumour (left) and normal (right) tissue can be seen clearly. These five clusters of pathways are retrieve in order to further research the reation for different patterns in the future.

3.6 Regression Analysis

The look for a target for regression analysis by colouring the UMAP by interesting annotation was not successful, as shown in Sup.4. After comparing the BRCA pathway matrix for tumor and normal tissue in volcanos (Fig.8), a list of 85 pathways that are up-regulated and down-regulated in breast cancer were provided. Then the list was narrowed down by choosing only upregulated and down-regulated pathways in comparison between BRCA tumor patients and PRAD and KIRC tumor patients (Sup.5). This process narrowed the list down to two pathways, WU_CELL_MIGRATION (WU-CM) and YTT-UP. After literature research, WU-CM is excluded for being strongly tied to bladder cancer (Wu *et al.*, 2008). Therefore, YTT-UP is chosen as the target for regression analysis, and 85 significant pathways in BRCA that were identified before are used to predict YTT-UP. The correlation among these pathways was investigated by plotting the correlation heatmap (Fig.9).

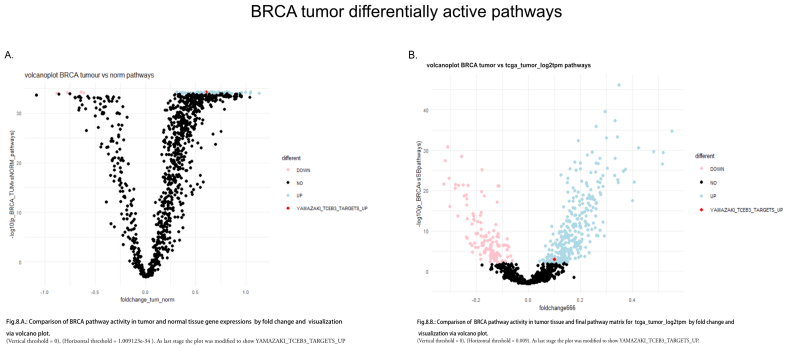


Fig.8.A.: Comparison of BRCA pathway activity in tumor and normal tissue gene expressions by fold change and visualization via volcano plot.
(Vertical threshold = 0), (Horizontal threshold = 1.009123e-34). At last stage the plot was modified to show YAMAZAKI_TCEB3_TARGETS_UP

Fig.8.B.: Comparison of BRCA pathway activity in tumor tissue and final pathway matrix for tcga_tumor_log2tpm by fold change and visualization via volcano plot.
(Vertical threshold = 0), (Horizontal threshold = 0.009). At last stage the plot was modified to show YAMAZAKI_TCEB3_TARGETS_UP.

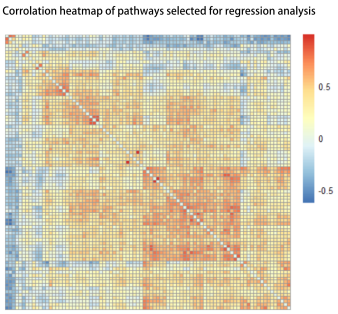


Fig.9: The Correlation of GSVA means of the 85 selected pathways are plotted against eachother here to visualize their relation to eacother.

The data was split into two groups, one is the training set and the other is the testing set. Originally, the model performed successfully on the training set but poorly on the testing set, indicating an overfitting problem. In order to solve that, variables that are used to predict the target pathway that has a p-value larger than 0.5 were excluded, and the process was performed five times with random patients for the training set and testing set each time. We ended up with 25 pathways with the highest significance. The final result is shown in Tab.2 and Sup.7.

Table 3.1: Results for regression.

	Adjusted R-squared values	**RMSE training model**	**RMSE test data sets**
#1	0.8727	0.0410006	0.07613583
#2	0.8696	0.043647	0.07622888
#3	0.8417	0.04716991	0.07467378
#4	0.8458	0.4701736	0.06281068
#5	0.8275	0.04413988	0.07238331

4 Discussion

In this project we aimed to explore the genetic expression patterns tied to tumorigenesis, by identifying differentially expressed genes and pathways in cancer patients. First, the data was cleaned of low-variance (below 75 percent) genes and non-protein coding biotypes, then dimensionally reduced until a matrix of pathway scores per patient was generated. In our analysis, we used the Molecular Signature Database (MSigDB) as the source of the gene set to group gene expression data under the pathway. This is to provide a very comprehensive and annotated collection of gene sets (Liberzon *et al.*, 2011). As a means of quality control, the efficiency of our Pathway matrix was tested with Venn diagrams of overlapping genes as well as UMAP clustering plots and compared to the plots of all genes. The similarity of the clusters suggested that the information was preserved in the final pathways.

Afterwards, the analysis was carried out in three stages, namely, the pan-cancer analysis focused analysis and regression analysis. In the pan-cancer analysis, we aimed to investigate the differential activation patterns of all cancer cells and compare them in the context of their tumor type. In the focused analysis, we determined the differentially active pathways in a comparison of BRCA tumor and normal tissues. The comparison of differentially active pathways identified in pan-cancer analysis and those identified in the focused analysis, as well as literature research on these groups allowed us to pick a target pathway YTT-UP for regression analysis. Finally, in the regression analysis, we used a multinomial regression model to predict the activation of this pathway. In the pan-cancer analysis stage of our project, we investigated the distribution of several annotations within these clusters to see if clustering was correlated with tumor type, menopause state etc. While tumor types were separated distinctively, we saw that annotations other than tumor type were not separated in a conclusive manner, suggesting that our data was most relevant for the tumor type. The lack of conclusive annotative distribution suggests that annotations were not strongly tied to our pathway information and therefore would be insufficient targets for regression analysis. The pathways with higher variance than %75 of other pathways were also identified and clustered with a heatmap into smaller gene groups in the context of tumor type. These 101 pathways were used to separate our cancer types into five distinct families. In Fig. 6C we can see that tumor types can be separated into five major groups in terms of behaviour also shown in Tab.3. In further research these families and the top pathway can be used to understand similarities and differences between tumor types and predict the success of treatment in previously untested tumor types.

Table 4.1: Cancer type clustering.

Group	**Cancer type**
#1	DLBC, LAML, THYM
#2	COAD, READ, CESC, HNSC, LUSC, BLCA, ESCA, TGCT, UCS, OV, UCEC
#3	KIHC, PRAD, KIRP, THCA, LGG, PCPG, ACC
#4	UVM, GBM, MESD, BRCA, SARC, SKCM
#5	KIRC, PAAD, CHOL, LIHC, LUAD, STAD

We also identified the top 20 most active pathways as potential biomarkers in all tumour types. As expected many known oncogenes like GATA transcription factors (Zheng and Blobel, 2010), and cancer-related elements like genes in cancer gene neighbourhoods (CRISP2 in WEBER_METHYLATED_HCP_IN_FIBROBLAST_DN (Weber *et al.*, 2007) pathways were found in these genesets. Furthermore, drivers of innate immunology were also significantly activated. In one gene set we tied 14 interferons tied to covid, to cancer (https://www.wikipathways.org/instance/WP4912_r115852, 16.09.2022, 12:11). We have confirmed the driving forces of tumorigenesis in terms of pathways, and thus became even more confident that our pathway matrix was accurate.

In focused analysis first, we computed the correlation values between the mean pathway scores for BRCA patients. The reason for a strong correlation between pathway groups was further investigated by computing the overlapping number of Ensemble IDs in each pathway, normalized by the Jaccard index. the comparison of these data (Sup.6B) suggests that the reason for the high correlation is because they are upregulated or downregulated in a similar manner.

Furthermore, normal tissue and tumor tissue from BRCA patients were plotted against all pathways and the top 200 pathways (which were selected by taking those with the highest fold changes.) The clear distinction is seen in Fig.7.B allows us to divide the pathways into five clusters. We annotated the pathways and split them by the clustering patterns in order to investigate more about them in the literature. Among the group of pathways that are strongly downregulated in breast cancer tissue, a lot of them are related to methylation in CpG island promoters in germ-specific genes in primary fibroblast or sperm. DNA methylation is linked to long-term gene silence and is crucial for growth, chromosomal inactivation, and genomic imprinting (Andersen *et al.*, 2012). The study has shown loss of X chromosome inactivation (XCI) is frequently seen in basal-like subtype and BRCA1 mutation-associated breast cancers, as well as ovarian cancers (Kang *et al.*, 2015). Furthermore, methylation of CpG island is known to regulate gene expression. As a result of abnormal promoter methylation, tumor suppressor genes are silenced and oncogenes are activated, which is regarded as a hallmark of cancer (Moarii *et al.*, 2015). DNA methylation also contributes to gene expression in the transforming growth factor-beta (TGF-beta) pathway, which is well established in cancer-associated fibroblasts (Zhang *et al.*, 2017). Another pathway we encountered is the tamoxifen metabolism pathway. Tamoxifen is a selective estrogen receptor modulator that blocks the cytoplasmatic oestrogen receptor via a competitive inhibition, which leads to a reduction of cell division activity in oestrogen-dependent tissues (Shagufta and Ahmad, 2018). However, many patients relapse due to resistance to tamoxifen, a cornerstone of adjuvant breast cancer therapy (Cronin-Fenton *et al.*, 2014), the resistance process is still underexplored and is the subject of further research (Barazetti *et al.*, 2021).

As for regression analysis, we decided to investigate the YTT-UP pathway for reasons that were explained in the result. To our knowledge, no one had researched the involvement of this pathway in cancer before. We found out that the YTT-UP pathway was a group of genes that are overexpressed under the lack of Elongin-A. Further investigation in the pathway revealed an interesting 20 gene transcription factors group which had no overlap with known hallmark genes. The 20 gene transcription factor group is largely involved in cellular differentiation, and pre-mRNA splicing (NCBI Resource, 2016; Safran *et al.*, 2021–). We believe their joined regulatory function compensates for the lack of Elongin-A in RNA metabolism, which may lead to the heightened growth rate. This means the upregulation of YTT-UP transcription factors without the lack of Elongin-A in cancer cells might contribute to intensified growth and proliferation rate of aggressive cancers. Additionally, we theorize that the differing transcriptive signature tied to the activation of YTT-UP-transcription factors plays a role in cell differentiation during epithelial-mesenchymal transition (EMT).

In regression analysis, we generated a successful regression model that predicts the geneset activation score of YTT-UP from 25 pathways with the highest significance. We investigated these pathways further and found parallels in the thyroid cancer and melanoma-related genesets. The highest correlation value between YTT-UP and other pathways (0.7368), was with the ALONSO_METASTASIS_EMT_UP geneset which is an EMT-related gene set. The second highest correlation (0.5424cor) is with AMIT_EGF_RESPONSE_480_HELA which is the genes upregulated in response to epithelial growth factor. The ties between growth signalling and the growth-promoting transcription signature of YTT-UP transcription factors are as expected. The strong ties between this transcription signature and cell differentiation in EMT remain to be investigated, and potentially a target for anti-metastatic therapy.

5 References

Abdi, H, and Williams, LJ and (2010). Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics 2, 433–459.

Andersen, IS, Reiner, AH, Aanes, H, Aleström, P, and Collas, P and (2012). Developmental features of DNA methylation during activation of the embryonic zebrafish genome. Genome Biology 13, R65–.

Barazetti, JF, Jucoski, TS, Carvalho, TM, Veiga, RN, Kohler, AF, Baig, J, Al Bizri, H, Gradia, DF, Mader, S, and Carvalho De Oliveira, J and (2021). From micro to long: Non-coding RNAs in tamoxifen resistance of breast cancer cells. Cancers 13, 3688–3688.

Becht, E, Mcinnes, L, Healy, J, Dutertre, C-A, Kwok, IWH, Ng, LG, Ginhoux, F, and Newell, EW and (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology 37, 38–44.

Brown, LJ, Achinger-Kawecka, J, Portman, N, Clark, S, Stirzaker, C, and Lim, E and (2022). Epigenetic therapies and biomarkers in breast cancer. Cancers 14, 474–474.

Clifton, G, GT; Peoples (2021). Immunotherapy as a partner for HER2-directed therapies. EXPERT REVIEW OF ANTICANCER THERAPY 21, 739–746.

Coller, HA (2014). Is cancer a metabolic disease? Am J Pathol 184, 4–17.

Conover, WJ (1999). Practical nonparametric statistics, 3rd edition.

Cronin-Fenton, DP, Damkier, P, and Lash, TL and (2014). Metabolism and transport of tamoxifen in relation to its effectiveness: New perspectives on an ongoing controversy. Future Oncology 10, 107–122.

Feng, J, Meyer, CA, Wang, Q, Liu, JS, Shirley Liu, X, and Zhang, Y and (2012). GFOLD: A generalized fold change for ranking differentially expressed genes from RNA-seq data. Bioinformatics 28, 2782–2788.

Fletcher, S, and Islam, MZ and (2018). Comparing sets of patterns with the jaccard index. Australasian Journal of Information Systems 22, –.

Giuliano, AE, Connolly, JL, Edge, SB, Mittendorf, EA, Rugo, HS, Solin, LJ, Weaver, DL, Winchester, DJ, and Hortobagyi, GN and (2017). Breast cancer-major changes in the american joint committee on cancer eighth edition cancer staging manual. CA: A Cancer Journal for Clinicians 67, 290–303.

Hanahan, D, and and, R (2011). Hallmarks of cancer: The next generation. Cell 144, 646–674.

Hänzelmann, S, Castelo, R, and Guinney, J and (2013). GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics 14, 7–7.

Kang, J, Lee, HJ, Kim, J, Lee, JJ, and Maeng, L-S and (2015). Dysregulation of x chromosome inactivation in high grade ovarian serous adenocarcinoma. PLOS ONE 10, e0118927–.

Liberzon, A, Subramanian, A, Pinchback, R, Thorvaldsdóttir, H, Tamayo, P, and Mesirov, JP (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics 27, 1739–1740.

Liu, F, and Deng, Y (2021). Determine the number of unknown targets in open world based on elbow method. IEEE Transactions on Fuzzy Systems 29, 986–995.

Moarii, M, Boeva, V, Vert, J-P, and Reyal, F and (2015). Changes in correlation between promoter methylation and gene expression in cancer. BMC Genomics 16, –.

NCBI Resource, C (2016). Database resources of the national center for biotechnology information. Nucleic Acids Res 44, D7–19.

Ranstam, J and (2016). Multiple p -values and bonferroni correction. Osteoarthritis and Cartilage 24, 763–764.

Safran, M, Rosen, N, Twik, M, BarShir, R, Stein, TI, Dahary, D, Fishilevich, S, and Lancet, D (2021–). The GeneCards suite. In: Practical Guide to Life Science Databases, ed. I Abugessaisa, and T Kasukawa, Singapore: Springer Nature Singapore, 27–56.

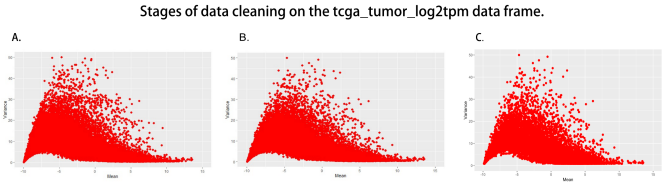
Shagufta, and Ahmad, I (2018). Tamoxifen a pioneering drug: An update on the therapeutic potential of tamoxifen derivatives. European Journal of Medicinal Chemistry 143, 515–531.

Weber, M, Hellmann, I, Stadler, MB, Ramos, L, Pääbo, S, Rebhan, M, and Schübeler, D and (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nature Genetics 39, 457–466.

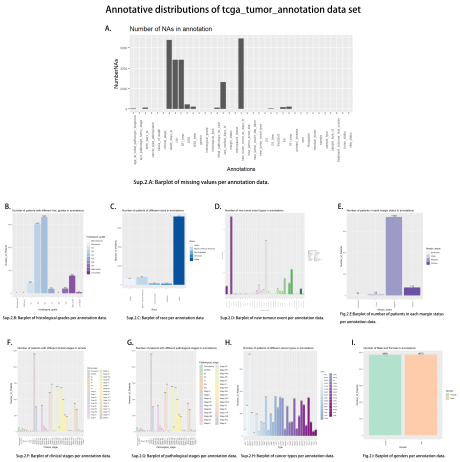
References

Wu, Y, Siadaty, MS, Berens, ME, Hampton, GM, and Theodorescu, D (2008). Overlapping gene expression profiles of cell migration and tumor invasion in human bladder cancer identify metallothionein 1E and nicotinamide n-methyltransferase as novel regulators of cell migration. Oncogene 27, 6679–6689.
Zhang, M-W, Fujiwara, K, Che, X, Zheng, S, and Zheng, L and (2017). DNA methylation in the tumor microenvironment. Journal of Zhejiang University-SCIENCE B 18, 365–372.
Zheng, R, and Blobel, GA and (2010). GATA transcription factors and cancer. Genes & Cancer 1, 1178–1188.

6 Supplements



Sup.1. A-C: Scatterplots of variance and mean of genes in tcga_tumor_log2tpm data set pre data cleaning (A.), after biotype filtering (B.) and post data cleaning (C.). For B. non-protein coding genes were removed. For C. compleate data cleaning included biotype and variance filtering was performed. During varience filtering genes with a variance lower than that of %75 of other genes were excluded.

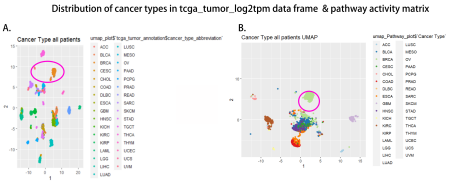


Sup.2.A: Barplot of missing values per annotation data.

Sup.2.B: Barplot of histological grades per annotation data. Sup.2.C: Barplot of race per annotation data Sup.2.D: Barplot of non-tumour event per annotation data Fig.2.E Barplot of number of patients in each margin status per annotation data.

Sup.2.F: Barplot of clinical stages per annotation data. Sup.2.G: Barplot of pathological stages per annotation data. Sup.2.H: Barplot of cancer types per annotation data. Fig.2.I: Barplot of genders per annotation data.

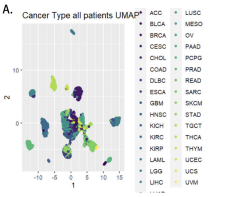
Sup .2. A-I.: Distributions of tcga_tumor_annotation data set. A selection of exploratory barplots visualizing the disparity of annotation data tied to the tcga_tumor_log2tpm dataframe. The high prevalence of NAs in certain annotations, as well as the low disparity and high segmentation seen in most annotative data can be observed.



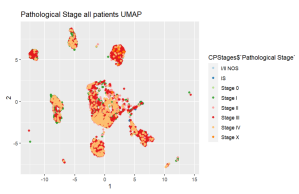
Sup.3.A: UMAP on tcga_tumor_log2tpm colored by cancer type annotation Sup.3.B: UMAP on final pathway matrix colored by cancer type annotation

Sup.3.A-B.: First UMAP on tcga_tumor_log2tpm and on final pathway matrix was computed. By plotting similar clustering can be observed, verifying the biological significance of our final pathway matrix. For color information we used cancer types annotation and subsequently similar BRCA clusters were noticed in both data frames.

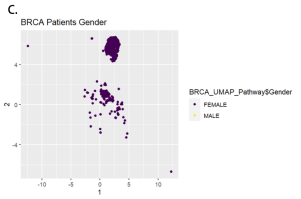
Distributions of interesting annotative data inside the UMAP of final pathway matrix.



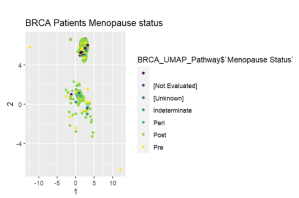
Sup.4.A: Distribution of cancer types inside the UMAP of final pathway matrix.
Clusters of non homogenous distribution and uniform tumor type clusters.



Sup.4.B: Distribution of Pathological stage inside the UMAP of final pathway matrix.
Non conclusive distribuon of pathological stages can be observed. This annotation is not strongly tied to our pathway information.

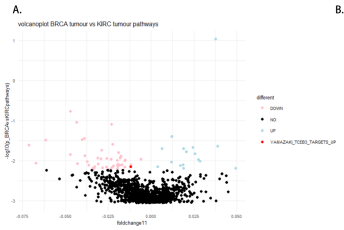


Sup. 4.C: Distribution of Gender stage of only BRCA patients inside the UMAP of final pathway matrix.
This plot reveals the lack of male patients in this dataset.

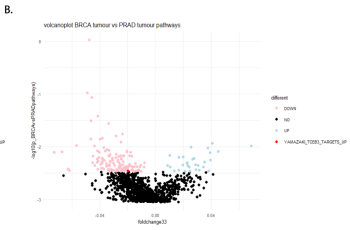


Sup.4.D: Distribution of menapause status of only BRCA patients inside the UMAP of final pathway matrix.
Non conclusive distribuon of menopause status can be observed. This annotation is not strongly tied to our pathway information.

Pan Cancer volcano plots

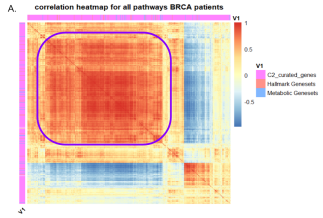


Sup.5.A.: Comparison of BRCA and KIRC pathway activity by volcano plot
(Vertical treshold = 0), (Horizontal treshold = 0.1526032)

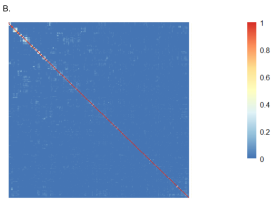


Sup.5.B.: Comparison of BRCA and PRAD pathway activity by volcano plot
(Vertical treshold = 0), (Horizontal treshold = 0.2692998)

Correlation analysis of the pathways

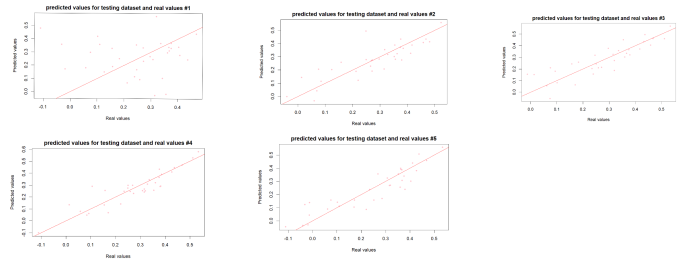


Sup.6.A.: Correlation heatmap of final pathway matrix.
correlation values between the mean pathway scores in final pathway matrix were computed and plotted. An agreggation of highly correlated pathways can be seen. The group was identified and pathway data was extracted for further analysis and literature investigation.



Sup.6.B.: Overlap heatmap of genes per pathway, normalized by jaccard index.
No significant overlap can be seen, suggesting that pathways with a high number of overlapping genes were not included in the pathway matrix.

Regression improvement (plots for testing model)



Sup.7.A: Regression improvement (plots for testing models).

Table 6.1: Sup.8: Used packages

Package name	**Usage**	**Author**
BiocManager	install and update Bioconductor packages	Martin Morgan
BiomaRt	retrieve Ensembl	Steffen Durinck, Wolfgang Huber
Dplyr	used for PCA	Hadley Wickham, Romain François, Lionel Henry, Kirill Müller
ggVennDiagram	used to generate all of our Venn plot (for pathway matrix comparison)	Chun-Hui Gao
ggplot2	creating graphics, like bar plots, volcano plots, box plots, histograms	Hadley Wickham, *et all.*
GSVA	perform non-parametric, unsupervised method for estimating variation of gene set enrichment through the samples of a expression data set (for giving gene set scores)	Justin Guinney, Robert Castelo
gtools	Functions to assist in R programming	Ben Bolker, Gregory R. Warnes, Thomas Lumley
magrittr	Used for grouping age pf patients in annotations	Stefan Milton Bache, Hadley Wickham
msigdbr	Provides the 'Molecular Signatures Database (MSigDB) gene sets (to draw Ensembl ID's for our pathways) Retrieve C2_curated_genes to create a pathway matrix	Igor Dolgalev
pheatmap	Implementation of heatmaps that offers more control over dimensions and appearance. Used for the generation of all our heatmaps	Raivo Kolde
RColorBrewer	Used to extend color pallets for plotting, i.e, for the barplots of the annotations	Erich Neuwirth
tibble	Used for grouping age pf patients in annotations	Kirill Müller, Hadley Wickham
tidyr	Create tidy data. Used for grouping age pf patients in annotations	Hadley Wickham, Maximilian Girlich
umap	Uniform manifold approximation and projection, used for dimension reduction	Tomasz Konopka
uwot	An implementation of Uniform manifold approximation and projection	James Melville