Schildkrötenkrebs

# Introduction

# Introduction

2019 starben 230,242 Menschen in Deutschland an Krebs

. Um Tumore zu erkennen und eine besserer Behandlung zu finden, ist die Entwicklung neuer Behandlungsmethoden von hoher Wichtigkeit. Dazu ist es essenziell die mutationstechnischen uRsachen für Krebsentwicklung zu identifizieren. Dafür können transcriptomic profiling methods wie RNA-seq verendet werden.

The provided data in the following analysis originates from a transcriptomic profiling methods like RNA-seq. Transcriptomic profiling sequences all the RNA that has been generated by transcription of the cells DNA. The difference to sequencing of DNA is, that it only sequences those genes, that are going to be expressed in that cell.

#### RNA-seq

RNA-seq is performed by cleaning of RNA, fragmentation, translation of RNA to cDNA, sequencing of cNDA and comparing with the reference genome. The advantage of RNA-seq is that it includes information about gene expression, that is especially important in the analysis of tumors such as epigenetic changes (e.g. epigenetic gene silencing) or fusion proteins.

The results from RNA-seq used for the analysis originate from the cancer genome atlas (TCGA)

## Thyroid carcinoma

Thyroid carcinoma (THCA) incidence increased dramatically over the past few years (<https://jamanetwork.com/data/journals/intemed/936342/m_ied170005f1.png>) REFERENCE, deswegen schauen wir uns die Gene, die THCA verursachen im unserer Analyse näher an. Allerdings sind nur 1% der deutschen Tumoren THCA.

Thyroid cancers werden in verschiedene Typen aufgeteilt, der häufigste Typ ist papillary thyroid cancer (PTC) with 80% of total Thyroid cancers. The most common mutation in PTC is the V600E-Mutation of the RAF-Kinase, what causes a constant activation and by intracellular signaling it promotes tumor cell growth and thereby growth of the tumor.

Die Aufgabe der Schilddrüse ist es, Hormone zu sythetisieren und somit Köpertemperatur und Metabolismus zu kontrollieren. Thyroxine spielt dabei eine wichtige Rolle im Metabolismus, indem es die metabolic rate stimuliert. [Thyroxine - Higher - Coordination and control - The human endocrine system - Edexcel - GCSE Biology (Single Science) Revision - Edexcel - BBC Bitesize](https://www.bbc.co.uk/bitesize/guides/z3gxb82/revision/3#:~:text=Thyroxine%20is%20produced%20from%20the%20thyroid%20gland%2C%20which,development.%20Its%20levels%20are%20controlled%20by%20negative%20feedback.) xxx. Was auch als eine underactive thyroid (= hypothyroidsm) bezwichnet wird und zu folgen wie Headaches, Nausea, depression führen kann. In der kommenden Analyse wird die Aktivität des thyroxine biosythese pathways in thyroid cancers untersucht. [Thyroxine Deficiency? 17 Signs - ProgressiveHealth.com](https://www.progressivehealth.com/thyroxine-deficiency.htm) Dadurch dass man weiß dass Thyroxine nicht vorhanden, kann man das ersetzen

the 3 histological types, was bringt uns das, die 3 Typen zu kennen??

2 Subtypen (differentiated (papilarry fulicular thca, weil die wie ursprünglihce thyroid cells ausshen) undifferentited, sehen einfach anders aus als andere thyroid cells, nicht klar abzugrenzen,

das andere bedeuted aus den

neue klassifikationen nur von PTC (BRAF like und RAS like) 80%s

## Hallmarks of cancer

Hallarks: was sind hallmarks? - Hallmarks sind Eigenschaften von TUmoren, die in jedem Tumor nachgewiesen werden können. Dazu gehören unter anderen die folgenden EIgenschaften: Resisting sell death, inducung angiogenesis, enabling replicative immortality, activating invasion and metastasis evading growth surpressors. Diese wurden zuerst von Hanahan und Weinberg 2011 veröffentlicht. Im laufe der Zeit kommen immer mehr hallmakrs dazu/es werden neue entdeckt.

Epigeentic profiles = auch epigenetische veränderungen werden in die Expressionsdate mit inebezogen, das wäre sehr sinnvol für die ANalyse, wird hier aber nicht beachetet

Was ist unsere Fragestellung?? und warum ist genau die so interessant?? und wie haben wir das erreicht=

## Computational tools

Um die Aktivität von Pathways zu vergleichen wurde eine

GSE methods um gene zu analysieren die bei RNA- seq analyse´siert wurden (fangen an mit einer ranked gene list =>x single enrichment score für jedes Geneset (2 genesets werden verglichen also zb normales und tumorgewebe)

#### Gene Set Variation Analysis

The Gene Set Variation Analysis (GSVA) is performed with the same intention as the GSEA, with the difference, that there is no reference expression data needed like in the GSEA. Because there was no expression data provided for comparison in the TCGA analysis, GSVA was used. There a various solutions to perform GSVA, one of them is to z-transform the provided expression data of the analysed tissue, so the mean over all patients of each gene is zero and the standard deviation is 1. The distribution of these z-transformed genes is like a t-distribution and can be used to calculate the log2FC and continue like in the GSEA.

#### Gene Set Enrichment Analysis

To analyse and compare the activity of pathways a Gene Set Enrichment Analysis (GSEA) is performed. Therefore a reference is needed to compare the activity of certain pathways in tumor tissue with the activity in normal tissue.

First, the log2FC is calculated for each gene and ranked in a vector for each patient, beginning with the highest log2FC. This resulted in a MATRIX/DATAFRAME/LIST. A high log2FC implies, that the the gene is higher expressed in tumor tissue than in normal tissue. In the next step the activity of each pathway in each patient is calculated. Therefore the ranked vectors of each patient containing the log2FCs are used. The function checks for each gen if it is included in the analysed pathway or not. If it is included, the log2FC of that gene is added to the cumulative function, if it is not included, the log2FC is subtracted. This results in a cumulative function, that has a peak at a certain place. In this place of the ranked vector is also a gene located, the expression value of that gene is saved as the enrichment score of the analysed pathway and the patient belonging to the used vector.

GSEA was performed with the package xxx.

#### UMAP (**Uniform Manifold Approximation and Projection for Dimension Reduction)**

The UMAP is a method to reduce the dimension of a multidimensional data set. Compared to the PCA, the structure of the data in higher dimensions is maintained. Thereby the UMAP keeps the overall structure of the data set, therefore clusters are easier to identify.

The problem of the UMAP is, that although the overall structure is conserved, the distance between the individual points is not proportional to the real distance in the data set.

kleine KLumpen sind wahrscheinlcich auch im richtigen RAum zusammen

#### PCA (Principle component analysis)

Reduce the dimension of a given data set. The dimensions are summarized in principal components (PCs) which do not correlate. Because the PCs summarize the dimensions, the first PCs explain most of the variance of the data set and thereby can be selected to explain the data. Still, one has to keep in mind, that by reducing the dimensions, not all of the variance is explained and some of the information is lost in the process. The ideal number of PCs can be determined with an elbow-plot. In our analysis we use a PCA as a foundation for the UMAP, because the UMAP can not work with correlated dimensions. Furthermore it is used to detect the most important pathways, which explain most of the first PCs.

In the analysis, a PCA is performed for the pan cancer analysis on the TCGA gene expression data, to find similarities and differences in pathway activity for each tumor type. Furthermore a PCA is performed for the focused analysis of THCA and normal tissue.

### Jaccard index

The Jaccard index is the Intersection, divided by the union of two sets.

### Our Analysis

In the following 2 analyses are performed, a pan cancer analysis and a focused analysis about THCA.

#### Pan Cancer Analysis

For the pan cancer analysis 3 data sets are provided. One containing expression data of 60,000 genes in 10,000 tumor patients, another one with clinical annotations concerning those patients and one with hallmark pathways and their included genes. In the following analysis this data is cleaned by removing NAs, biotype filtering and low-variance filtering. After that a descriptive analysis is performed. Those two steps lead to the actual analysis, a gene set variation analysis to detect significantly altered pathways compared to the other pathways in tumor tissue. In the end a linear regression analysis is performed to predict pathway activity based on other pathways??? xxx Furthermore a neuronal network is built to improve prediction.

#### Focused analysis on THCA patients

Furthermore a analysis of THCA patients is performed. For this analysis a data set containing the gene expression of 60 patients in tumor an normal tissue and their clinical annotations. First the data is cleaned and described like the pan cancer data, to prepare the data for the gene set variation analysis, which is also performed for the THCA data in the bigger pan cancer data set, to confirm results from the smaller data set. In this analysis a linear regression analysis is performed to predict the activity of thyroxine biosynthesis. The results are also improved with a neuronal network.

## Cancer

You can cite one or multiple authors. One author Kumar *et al.* (2017). Write in **bold** or in *italic* or in both ***bolditalic***. You can also write inline code, e.g. Seurat::RunUMAP.

## LUAD

Some information Kumar *et al.* (2017)

## Computational Tools

### Gene Set Variation Analysis

### Gene Set Enrichment Analysis

### UMAP

### PCA

## Our Analysis

### Pan Cancer

### Focused Analysis

### Related Work

# Material and Methods

## Our data sets

For the analysis four data sets were provided.

The first data set is a Gene expression data frame. The Gene expression data frame contains 60,000 genes and and their expression in 10,000 patients. It is derived from The Cancer Genome Atlas (TCGA). The expression of the genes was obtained by RNA-seq.

The second data frame contains 37 clinical annotations like Tumor type, age, gender, etc. for each of the 10,000 patients from the Gene expression data frame.

The third object is a list that contains five lists for the focused analysis, one list for each tumor type (BRCA, KIRC, LUAD, PRAD, THCA). For the focused analysis, the THCA data was used. The THCA data contains 3 data frames, each one with information about the same 60 patients. The first data is a gene expression matrix from THCA tissue, the second data contains the gene expression from normal tissue and the third data frame contains the clinical annotations like age and gender. Gene expression data was obtained by RNA-seq.

The fourth object contains 46 pathways involved in phenotypes partly included in the hallmarks of cancer and the genes involved in those pathways.

SIND DIE DATEN NORMALISIERT –> Normalisiert glaub ich (Anna) ODER ALS COUNTS? . –> test-Zitate

## Metabolic pathway selection

From the Molecular Signature Database (MSigDB) metabolic pathways were selected. First, they were compared to the given Hallmark-Pathways in order to select pathways that differ to the Hallmark-Pathways. The goal was to identify more pathways, that are important for the development of cancer. Therefore it was important that as many genes from the selected pathways as possible are also included in the provided Hallmark pathways. To identify the relevant pathways, the intersection of genes was calculated and the genes with an intersection of at least 99% were maintained for further analysis.

xxx??????????????????????

To avoid duplicates in between the metabolic pathways and between the Hallmark pathways and the metabolic pathways, the pathways were checked for duplicates with the Jaccard index. Pathways with a sum of Jaccard indices beyond the 1-sigma range were discarded.

## Preprocessing

### Deleting Not Available Values (NA’s)

Deleting of NA’s was done with the R-function na.omit(x).

### Low-Variance Filtering

Low variance filtering is performed to delete genes with a low variance in gene expression from the data set. It is performed to delete genes that are expressed the same in all cancer types (pancancer analysis) or the same in normal cells. To calculate the variance of the gene expression of a gene, the r-function var(x) is used and genes with a lower variance than a certain threshold value are removed.\ For the focused analysis the variance of the gene expression for each gene in tumor tissue was calculated. Genes with a variance beneath a certain threshold were deleted in the data sets of tumor and normal tissue.

### Biotype Filtering

The biotype filtering was conducted for the pancancer data and the focussed analysis data. The biotype of each gene was determined (protein coding, RNA,…) and compared with the biotypes of pathways. To allow an appropriate comparison of the expression data and further reduce the data, only biotypes were kept that are available in the pathways. The biotype can be determined with the R-function checkbiotypes(x) from the package biomaRt .

#### Selection of metabolic pathways

(da eine hohe jaccard summe eine hohe überschneidung mit anderen pathways bedeutet. In einer heatmap sind hohe Jacccard indices weiß bis rot gefährbt. Ein niedriger Jaccard index ist blau gefärbt.)

To test for duplicate pathways in the selected metabolic pathways compared to the hallmark pathways and the compared tp the metabolic pathways themselves, the Jaccard index between to pathways were calculated.\ There were a few duplicates between the metabolic and Hallmark pathways. Those metabolic pathways with a high Jaccard index were discarded. The success of the cleaning was checked by again calculating the Jaccard index between the metabolic and the hallmark pathways. The values of the Jaccard index were then illustrated in a heatmap . It can be assumed, that the selection of relevant pathways was successful because the pathways differ between each other. The number of metabolic pathways could be reduced from xxx to 600.

## Descriptive analysis

### Mean-variance plot

In a mean-variance plot the variance is plotted against the mean of expression values of the single genes. Therefore, the variance and mean were calculated by the R-function var(x) and mean(x).

### Violin plot

To check the distribution of a data set and compare it with other data sets violin plots are used. Violin plots are tilted and mirrored density plots of gene expression values. The y-axis shows the gene expression value and the x-axis shows the amount of genes with a certain gene expression value.

#### Volcano plot

A volcano plot is used to identify significantly differentially expressed genes. In the following analysis a volcano plot is used to identify which genes are significantly diferentially expressed in the analysed THCA tissue, compared to the analysed normal tissue. Therefore the mean of each gene is calculated for normal and THCA tissue and used for the calculation of the Log2-Foldchange (Log2FC) in the following way:

In the next step, a two-sided t test was performed to determine the significance of a difference in expression.

To avoid the accumulation of type 1 errors, a bonferroni correction was performed. n is the number of genes in the cleaned data set for focused analysis:

In the volcano plot the -log10 of the calculated p values is plotted against die Log2FC. Genes with a a lower p-value than the corrected alpha-value are significantly differentially expressed. If the Log2FC is additionally higher than 0.1, the genes are significantly over expressed in tumor tissue, if the Log2FC is higher lower than -0.1, the genes are significantly under expressed in tumor tissue.

## Comparing of Pathways

After reducing the number of Genes of the Gene Expression data frame and the THCA data by data cleaning, a descriptive analysis was performed with a Mean-variance plot and five violin plots of the TGCA data frame. The descriptive analysis of the THCA data was performed with a volcano plot and the distribution of the Tumor-specific data was displayed with violin plots.

### PCA

The package xxx is used to perform the PCA. Therefore the data obtained from the GSEA was used. After performing the PCA, the results were plotted to visualize the different clusters.

The PCA was performed for pathway and gene activity. For analysis of the gen activity the package xxx was used.

Dazu wurde noch analysiert, wie die Pathways auf die PCs verteilt sind.

## TCGA data

What kind of data do we have? ## Used Packages

show a table!

# Results

## Preprocessing

#### Deleting NA’s

Deletion of NA’s was applied to the three gene expression data frames (pan cancer and tumor and normal tissue data). Because the dimension of our data frames did not change during this process, it was assumed that there were no NA’s in the data sets.

#### Low-variance filtering

The goal of the analysis was to identify the genes that show a significantly different expression in certain tumor types (Pan cancer analysis) or in comparison from normal and tumor tissue (THCA Analysis). Therefore genes with a similar expression in all patients are not relevant. Probably, these would be mostly housekeeping genes.

The histogram of the logarithmised variances of the pancancer data is displayed in . The threshold of -1 was fixed and all genes with a lower variance were omitted. Doing so, the number of genes reduced from 60,000 to 19,000 genes.

The low-variance filtering of the THCA dataset was done in a similar way. The gene expression data of the cancer tissue was used to obtain the logarithmised variances of each gene. Genes with a lower variance than -1.25 were deleted in the tumor tissue and the normal tissue data. This resulted in a reduction from about 20,000 genes to 15,000 genes in both data frames.

#### Biotype filtering

The biotype of the genes from the selected metabolic pathways, of the genes of the hallmark pathways and of the genes of the gene expression matrix was determined. Only genes with the same biotypes were kept. The hallmark pathways and the selected metabolic pathways were almost all protein-coding, so only the protein-coding genes of the expression matrices were maintained.

**Biotypes of Hallamrk genes**

**Biotypes of Hallamrk genes**

**Biotypes of Gene expression Data frame**

**Biotypes of Gene expression Data frame**

**Biotypes of metabolic pathways**

**Biotypes of metabolic pathways**

### Descriptive analysis

#### Mean-variance Plot

In the mean-variance plot, displayed in Figure @ref(fig:showmeanvariance), genes with a very high variance and non-zero mean were annotated their ensemble ids.

Mean-variance plot of cleaned TCGA expression data. Y-axis shows variance of a genes expression, x-axis shows mean of a genes expression

Mean-variance plot of cleaned TCGA expression data. Y-axis shows variance of a genes expression, x-axis shows mean of a genes expression

#### Violin Plots

For descriptive analysis 5 violin plots were created for 5 different tumor types with the gene expression data from the TCGA Matrix, to compare the distribution of the data of 5 different tumor types. The violinplot can be seen in Figure @ref(fig:showviolinplots). The white point in the middle of each plot shows the 50% quantile. For all tumor types it is located in the middle of the gene expression value 0 and 5. In this area is also the highest amount of genes for every tumor type. Going to the top or bottom the curve flattens because only a few genes are expression very high or very low. One can see, that the the distributions of all 5 tumor types are very similar. It can be concluded, that the other 28 tumor types of this data set are distributed in a similar way.

Mean-variance plot of cleaned TCGA expression data

Mean-variance plot of cleaned TCGA expression data

#### Volcano plot

For the construction of the volcano plot the data for THCA from the data set for the focused analysis was used. The volcano plot is displayed in Figure @ref(fig:showvolcanoplot). Not significantly differentially expressed genes were marked green, significantly over expressed genes are marked blue and significantly under expressed genes are marked red. The gene with a very low p-value differ the most from tumor to normal tissue and are annotated with their name.

Volcano plot of THCA expression data

Volcano plot of THCA expression data

#### GSEA

The results of the GSEA for THCA tissue can be seen in Figure @ref(fig:GSEAHeat)

GSEA performed on the THCA expression data, annotated with Pathway type, histoligical type and cluster.

GSEA performed on the THCA expression data, annotated with Pathway type, histoligical type and cluster.

Horizontal sind die einzelnen Patienten angeordnet, vertikal die pathways, die Heatmap stellt die stärke der Expression der einzelnen pathways in den einzelnen Patienten dar. Rot markierte pathways sind überexprimiert, blaue sind unterexprimiert. Die Achsen wurden mit dem pathway typ, dem hiytological typ und dem cluster annotiert. Es bilden sich innnerhalb der Patienten 3 Cluster, die auf Gemiensamkeiten in der PAthwayaktivität zurückgeführt werden können. Vielleicht hängt die Clusterbildung mit dem histological type zusammen, da eindeutig erkennbar ist, dass in Cluster 3 nur der classical type vorhanden ist, während in Cluster 2 vermehrt Tall cell type auftritt und Folicular type vermehr in Cluster 1 auftritt.

### pan cancer für THCA

Um die Clusterbildung zu bestätigen wurde die gleiche Analyse für die THCA daten aus dem großen gene expression dataframe durchgeführt. Auch in dieser Analyse formten sich 3 Cluster, die in zum Teil auf die hiytological types zurückzuführen sind. xxx

#### GSVA

xxx Was wir aus der Heatmap sehen???

PCA of TCGA expression data, colored by tumor type

PCA of TCGA expression data, colored by tumor type

#### PCA

Die PCA wurde anhand der Pathway Aktivitäten durchgeführt und reduziert auf die ersten 2 PCs in Figure @ref(fig:PCAPanType) dargestellt. Es sind ansatzweise cluster, aber keine eindeutigen Strukturen zu erkennen. Die zusammensetzung der PCs und die Erklärung der AVrianz durch die einzelnen VAriablen kann Figure @ref(fig:) entnommen werden. BESCHREIBUNG DER PCA UND DER VARIANZVERTEILUNG. Der Plot wurde auch nach Form des Tumors angefärbt und in Figure @ref(fig:PCAPanForm) dargestellt. Hier ist eine klarere Struktur zu erkennen als bei Anfärbung nach Tumortyp. Daraus lässt sich schließen, dass zumindest die PC1 und PC2 inkludierten Pathways mit der Form des Tumors zusammenhängen.

Um die Ergebnisse zu überprüfen wurde die PCA auch auf die Genaktivität angewendet, es war kein Unterschied zu erkennen.

PCA of TCGA expression data, colored by tumor type

PCA of TCGA expression data, colored by tumor type

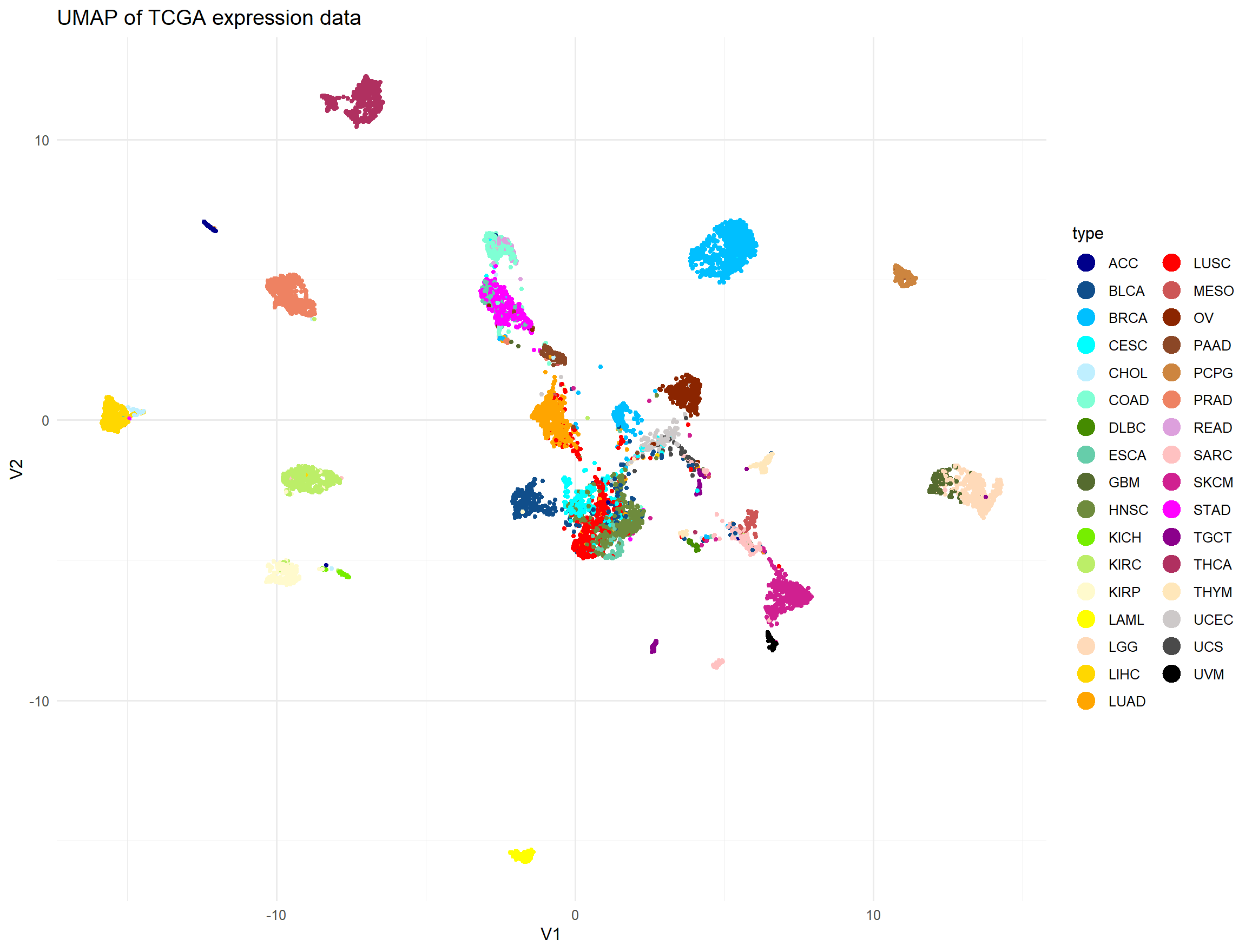
PCA of TCGA expression data, colored by form of tumor

PCA of TCGA expression data, colored by form of tumor

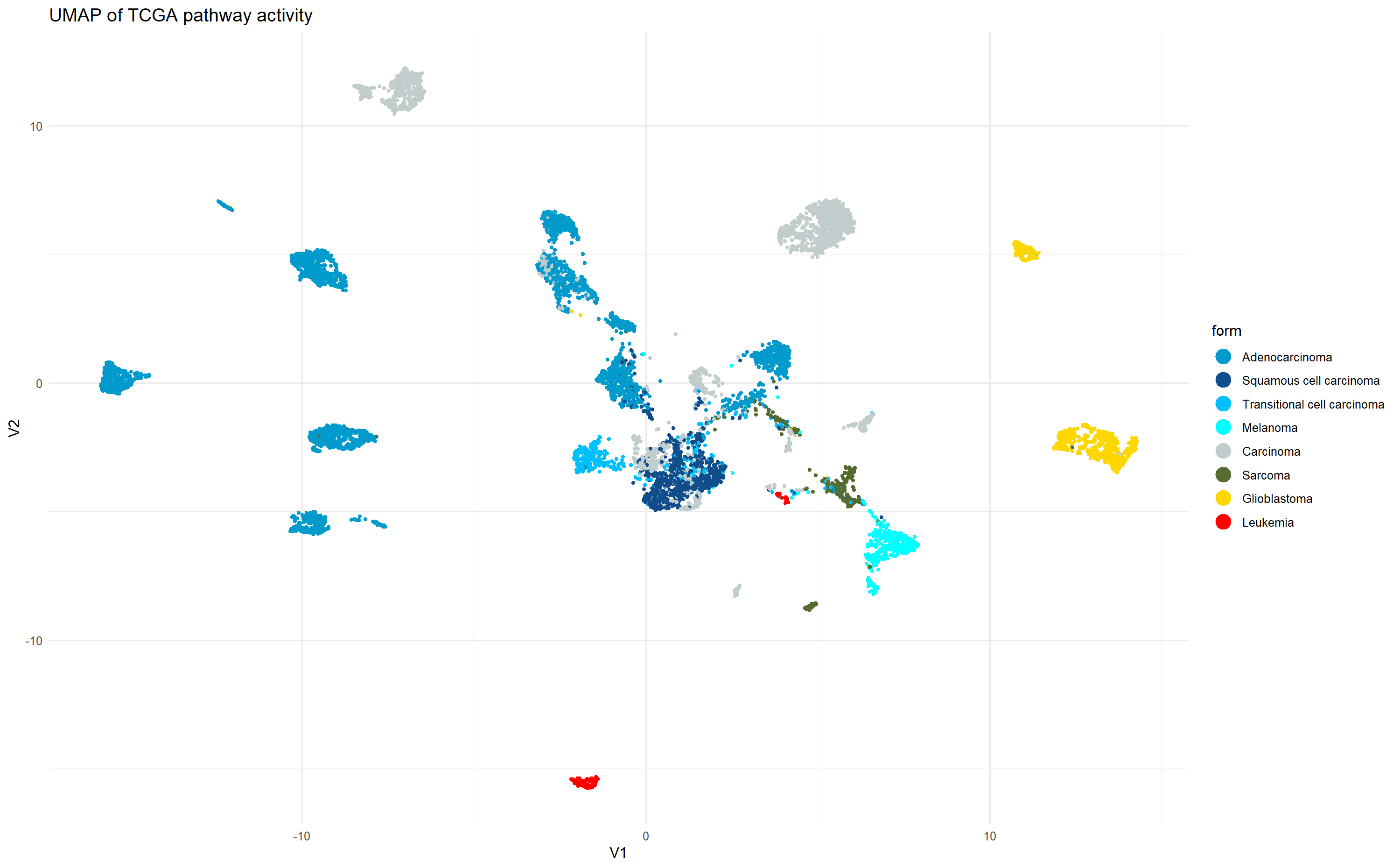
#### UMAP

DA durch die PCA keine eindeutigen Cluster erkennbar wurden, wurden die Ergebnisse der PCA genutzt, um eine UMAP zu erstellen. Die Ergebnisse sind in Figure @ref(fig:UMAPPanType) zu sehen. Hier sind eindeutige Cluser zu erkennen, auch wenn die Abstände zwischen den Clustern nicht proportional zu den wahren Unterschieden sind, kann daraus geschlossen werden, dass die einzelnen Tumortypen charakteristische Pathwayaktivitäten haben, wie zum Beispiel KIRC. Viele Tumortypen haben aber auch ähnliche Pathwayaktivitäten wie zum Beispiel GBM und SARC.

Um die Ursache dafür zu finden wurde auch hier die UMAP nach Form des Tumors markiert/gefärbt. Die Ergebnisse sind in Figure @ref(fig:UMAPPanForm) zu sehen. Es ist eindeutig zu erkennen, dass die Pathwayaktivität auch mit der Form des Tumors zusammenhängt. Zum Beispiel sind Glioblastome eindeutig von Adenocarcinomen getrennt. Allerdings ist auch hier das große Cluster in der Mitte nicht eindeutig einer Tumorform zuzuordnen.



UMAP of TCGA expression data, colored by tumor type



UMAP of TCGA expression data, colored by form of the tumor

#### Regression analysis

### Focused analsis

#### PCA

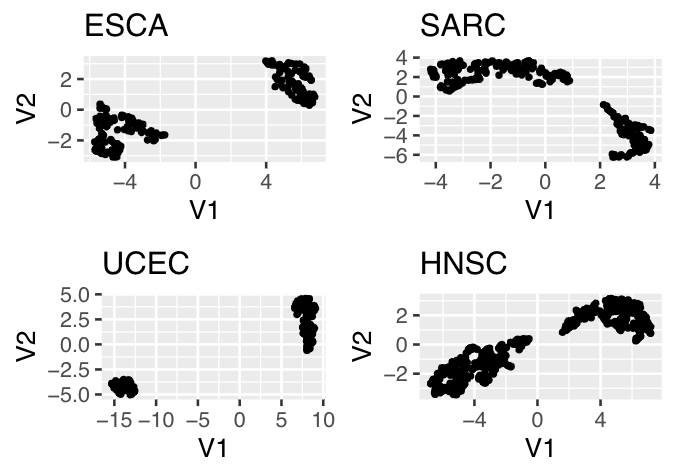
PCA was also performed for focused analysis, um Untergruppen, die mit der Pathwayaktivität oder Genaktivität zusammenhängen, innerhalb der Thyroid tumors zu finden. Dafür wurden zuerst die Ergebnisse der GSEA verwendet. Der Versuch Cluster zu finden gestaltete sich hier leider schwerer, da weder durch Betrachtung des Thyroid Tumortyps, noch durch Betrachtung der Stage ein eindeutiger Zusammenhang erkannt werden konnte.

## 33 tumor types are showing disting clusters in UMAP

hello world!

## blb alsdjflaskdf umap of some tumortypes

Figure generation. You can do it with knitr or with latex formatting. This is knitr:



Title. Description

easier alternative: this is latex formatting. In Figure you can see an UMAP. (+ label your figures, equations etc and then reference with /)

# Discussion

## Immune pathways are significantly upregulated in X

/ref{RNAseq} /ref{histological\_types} (Stark *et al.*, 2019) # References

Kumar, SK, Rajkumar, V, Kyle, RA, Duin, M van, Sonneveld, P, Mateos, M-V, Gay, F, and Anderson, KC (2017). Multiple myeloma. Nat Rev Dis Primers 3, 17046.

Stark, R, Grzelak, M, and Hadfield, J (2019). RNA sequencing: The teenage years. Nature Reviews Genetics 20, 631–656.

Zavidij, O et al. (2020). Single-cell RNA sequencing reveals compromised immune microenvironment in precursor stages of multiple myeloma. Nat Cancer 1, 493–506.

# Appendix

## Plots

hello ## Code world

#createn einer liste mit allen patienten in dfs sortiert nach krebs  
cancers = list();cancers = vector('list',length(table(tcga\_anno$cancer\_type\_abbreviation)))  
names(cancers) = names(table(tcga\_anno$cancer\_type\_abbreviation))  
i=1  
for (i in 1:length(cancers)){  
 cancers[[i]] = tcga\_exp\_cleaned[,tcga\_anno$cancer\_type\_abbreviation == names(cancers)[i]]  
}  
#function die einen krebstypen df und genesets als input nimmt und ein df mit pvalues ausgibt  
enrichment = function(expressiondata, genesets = genesets\_ids){  
 ESmatrix = sapply(genesets, FUN = function(x){  
 ins = na.omit(match(x,rownames(expressiondata)))#indices der gene im aktuellen set  
 outs = -ins#indices der gene nicht im aktuellen set  
 #gibt einen vektor der für jeden patienten den pval für das aktuelle gene enthält  
 res = NULL  
 for (i in 1:ncol(expressiondata)){#testet für jeden patienten  
 res[i] = wilcox.test(expressiondata[ins,i],expressiondata[outs,i],'two.sided')$p.value  
 }  
 return(res)  
 })  
 row.names(ESmatrix) = colnames(expressiondata); return(ESmatrix)  
}  
pvalueslist = lapply(cancers, enrichment)#für die tests für jeden krebstypen durch

get\_top10pathways\_from\_pvalues = function(df\_p\_values, length\_genesets) {  
   
 require(ggplot2)  
   
 results <- list()  
   
 df\_p\_values\_log10 <- -log10(as.data.frame(df\_p\_values))  
   
 mean\_pathway <- as.data.frame(apply(df\_p\_values\_log10, 1, mean))  
 rownames(mean\_pathway) <- rownames(df\_p\_values\_log10)  
   
 ordered\_score <- mean\_pathway[order(-mean\_pathway[ ,1]), 1]  
 top\_10 <- data.frame(ordered\_score[1:10])  
 colnames(top\_10) <- "mean\_pathway"  
   
 ordered\_names <- order(-mean\_pathway[ ,1])  
 top\_10\_names <- ordered\_names[1:10]  
 top\_10$pathway\_names <- row.names(mean\_pathway)[top\_10\_names]  
   
 results[[1]] <- top\_10  
   
 results[[2]] <- ggplot(data = top\_10, aes(x = mean\_pathway, y = reorder(pathway\_names, mean\_pathway)))+  
 geom\_bar(stat = "identity")+  
 coord\_cartesian(xlim =c(3, 3.75))+  
 labs(title = names(df\_p\_values),  
 x = "mean p-value pathway",  
 y = "pathway name")  
   
 pathway\_size <- order(-mean\_pathway[ ,1])  
 top\_10\_size <- pathway\_size[1:10]  
 top\_10$pathway\_size <- length\_genesets[top\_10\_size]  
   
 results[[3]] <- ggplot(data = top\_10, aes(x = mean\_pathway, y = reorder(pathway\_names,  
 mean\_pathway)))+  
 geom\_point(aes(size = pathway\_size))+  
 labs(title = names(df\_p\_values),  
 x = "mean p-value pathway",  
 y = "pathway name")  
   
 return(results)  
}