Supplementary material

Additional Computational Methods

A Principal component analysis (PCA) xxx QUELLE is used to alter the coordinates of a given dataset to its eigenvectors. This matrix rotation results in a new set of basis vectors called principal components (PCs) - the eigenvectors - that are orthogonal and show little correlation. Sorting the PCs by their associated eigenvalue, the PCs explaining the most variance can easily be identified, as they have the highest eigenvalue. By displaying the data set in a coordinate system span by the n most variant PCs, the dimensionalty of the data set is reduced to $\mathbb{R}^n$ with the lowest loss in variance.

Linear regression is a statistical model that uses measurable values to predict an outcome. For this purpose, a linear function serves as basis to build the linear regression equation [@lm]. The coefficients for each variable are estimated by their correlation and slope with the predicted parameter. Lastly, all coefficients as well as the intersect are optimized for the data set with a least sum of squares method.

Additional Figures

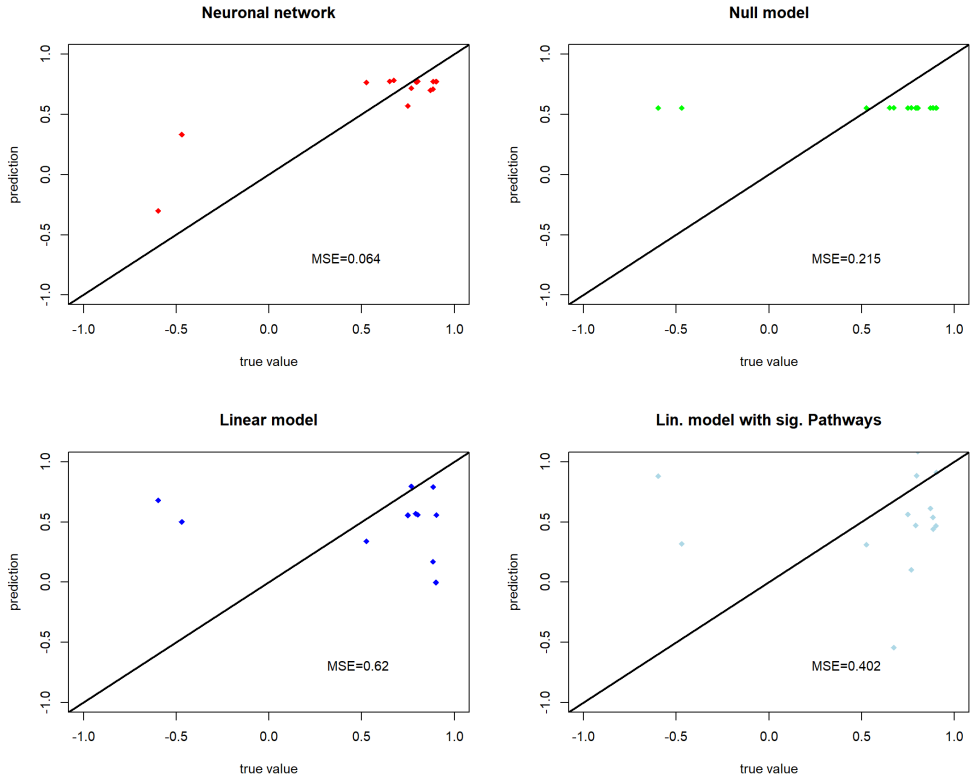

Figure 1: Regression results for various models on THCA GSEA test data. True values are plotted against predicted values, black slope indicate a perfect prediction.

Packages

```
## Warning: Paket 'readxl' wurde unter R Version 4.1.3 erstellt
```

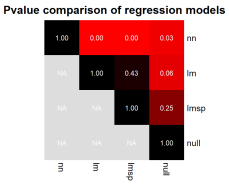

Figure 2: F-test comparison of various regression models. p-values are obtained from a two-sided variance test and displayed as heatmap. nn = neuronal network, lm = linear regression, lmsp = linear regression with only significant pathways, null = null model.

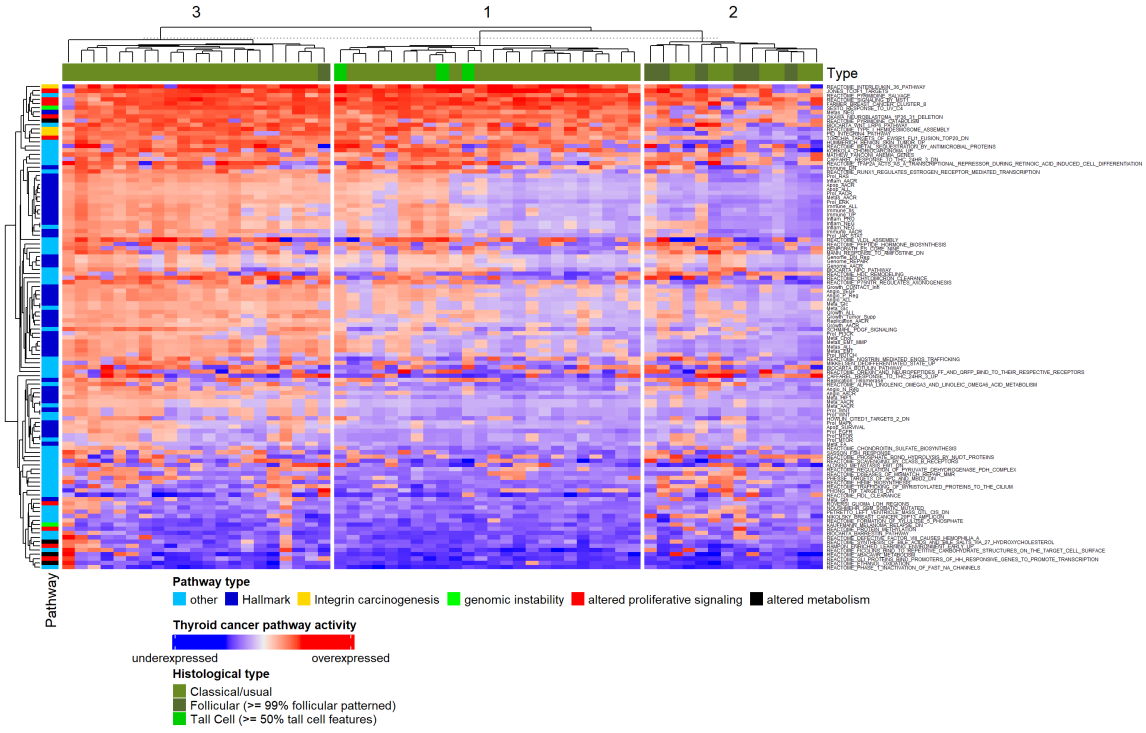

Figure 3: Pathway activity of the 50 most variant, hallmark, and 20 most significantly altered pathways for each patient. Column clusters were obtained by k-means clustering with k=3. Pathway activities were computed via GSEA of THCA expression data. For all pathway activities see figure (XXX in the apendix).

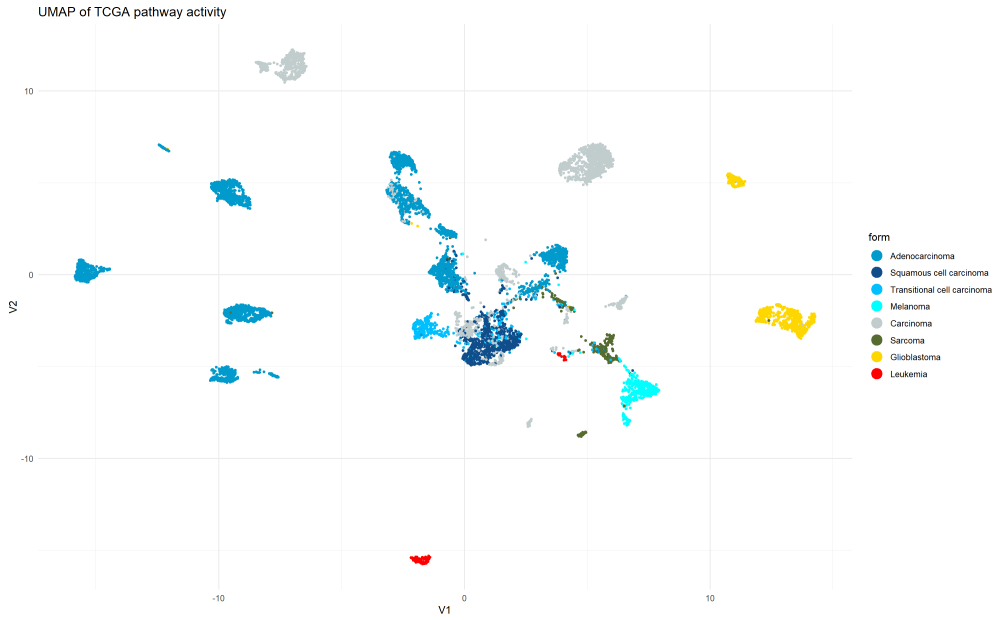

Figure 4: UMAP of TCGA pathway activity, colored by histological type

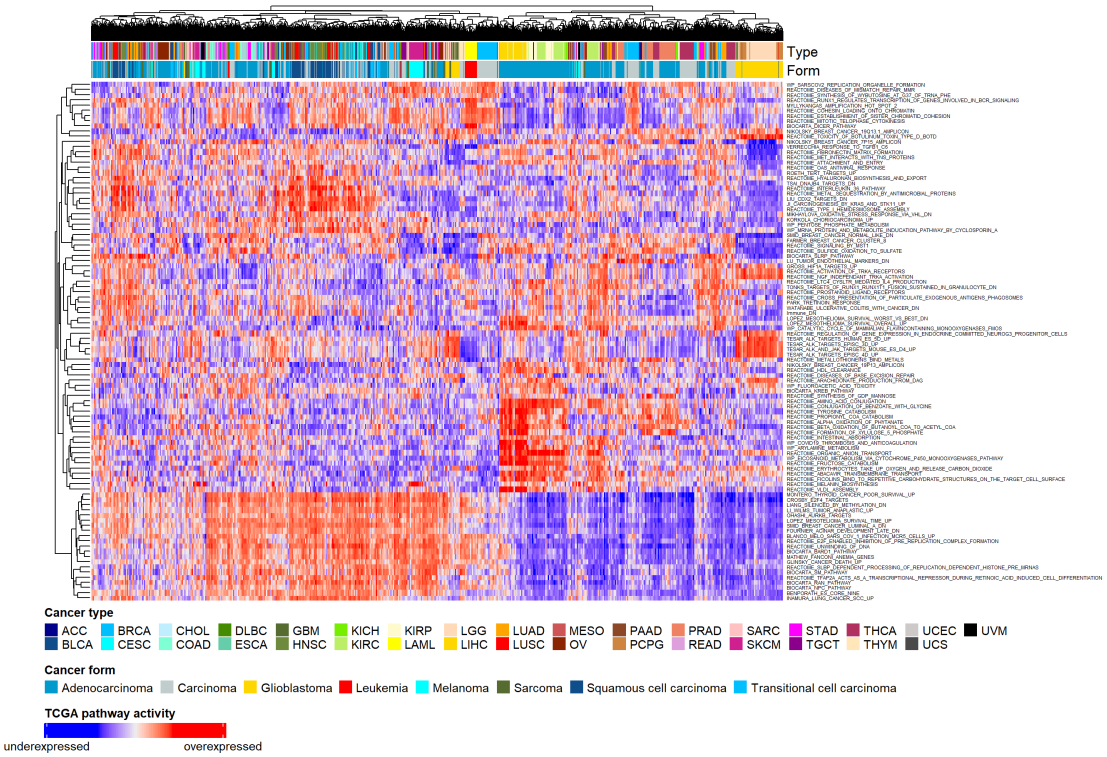

Figure 5: Pathway activity of the 100 most variant pathways for each patient. Column and row clusters were obtained by complete hierachical clustering. Pathway activities were computed via GSVA of pan-cancer expression data. For all pathway activities see figure (XXX in the apendix).

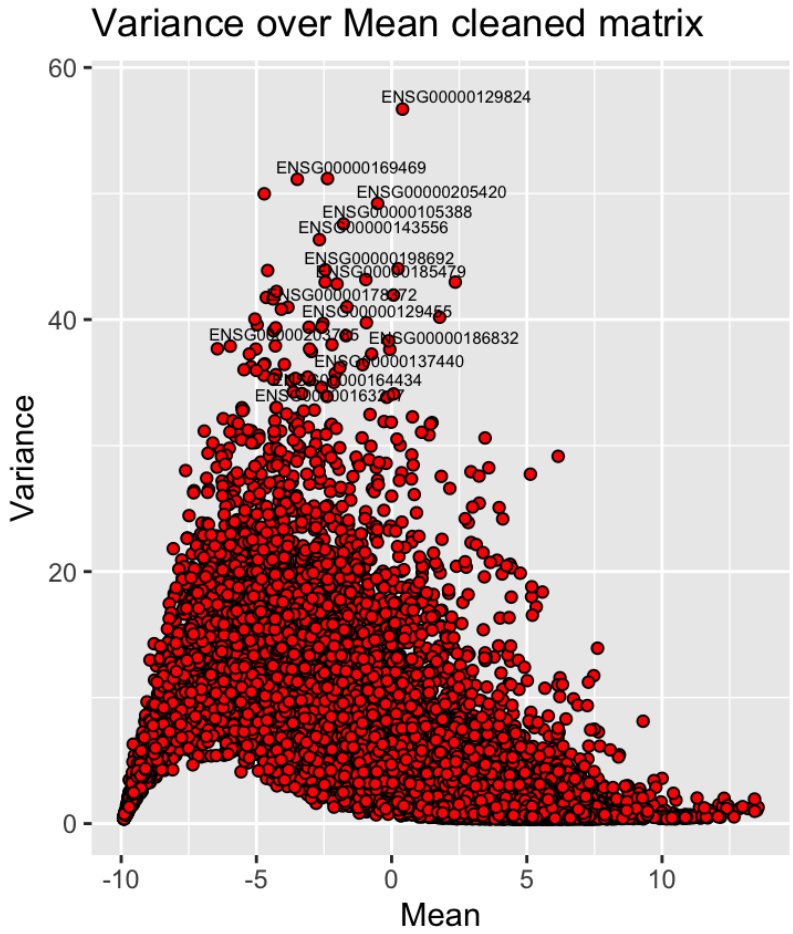

Figure 6: Mean-variance plot of cleaned TCGA expression data. Y-axis shows variance of a gene expression, x-axis shows the log2 mean of a gene expression. Genes with variance greater 33 are labelled with their ENSEMBL-ID

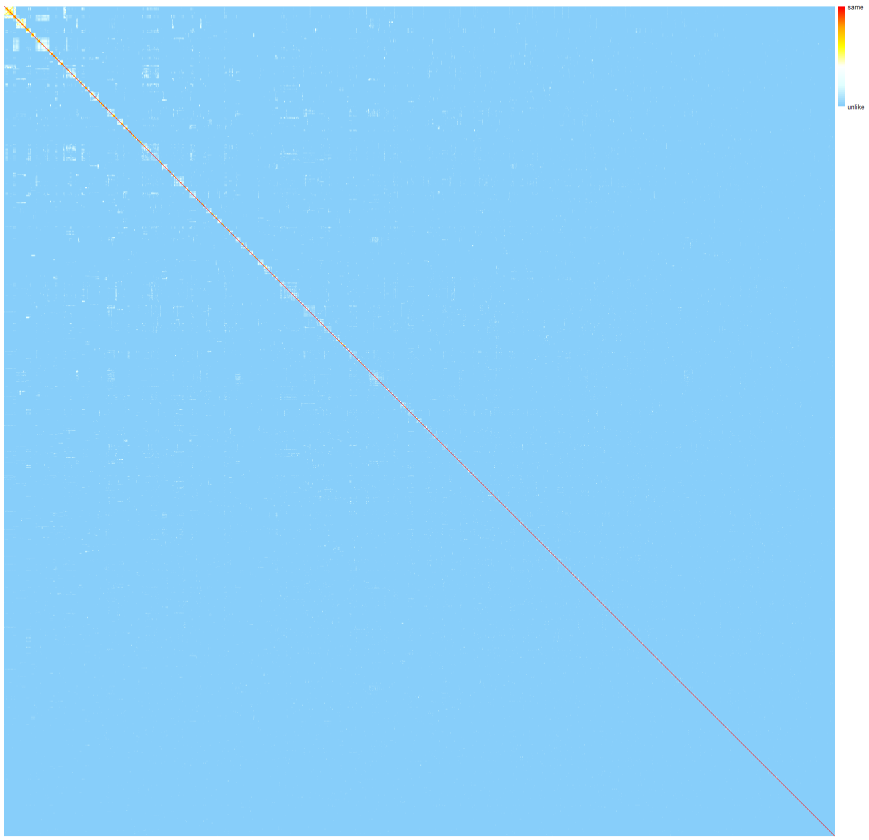

Figure 7: Heatmap, displaying jaccard index obtained by caomparing uncleaned metabolic pathways. Pathways with a high Jaccard index are colored red to white. Pathways without similarity are colored blue.

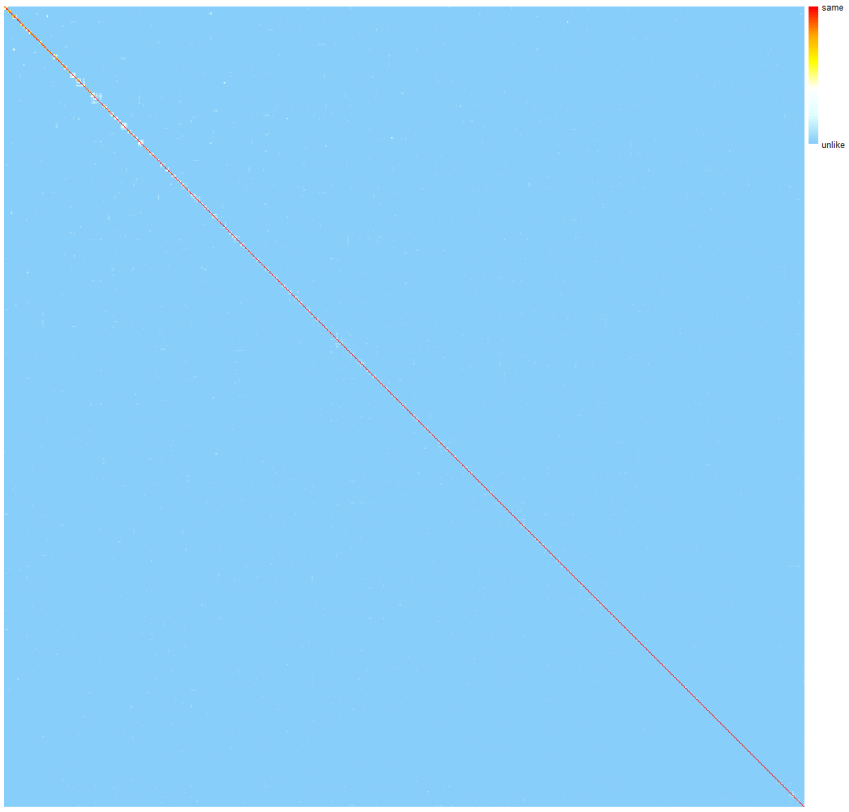

Figure 8: Heatmap, displaying jaccard index obtained by comparing cleaned metabolic pathways. Pathways with a high Jaccard index are colored red to white. Pathways with low similarity are colored blue.

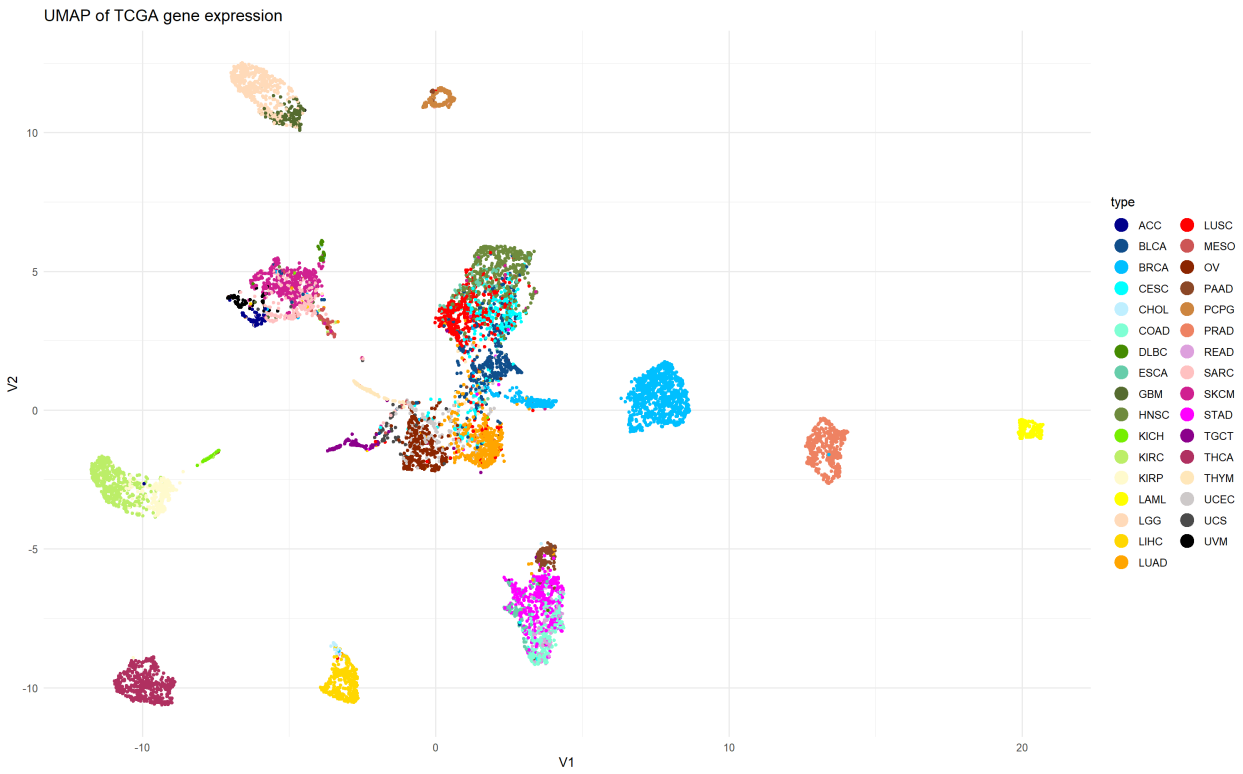

Figure 9: UMAP performed for gene expression data, colored by cancer type

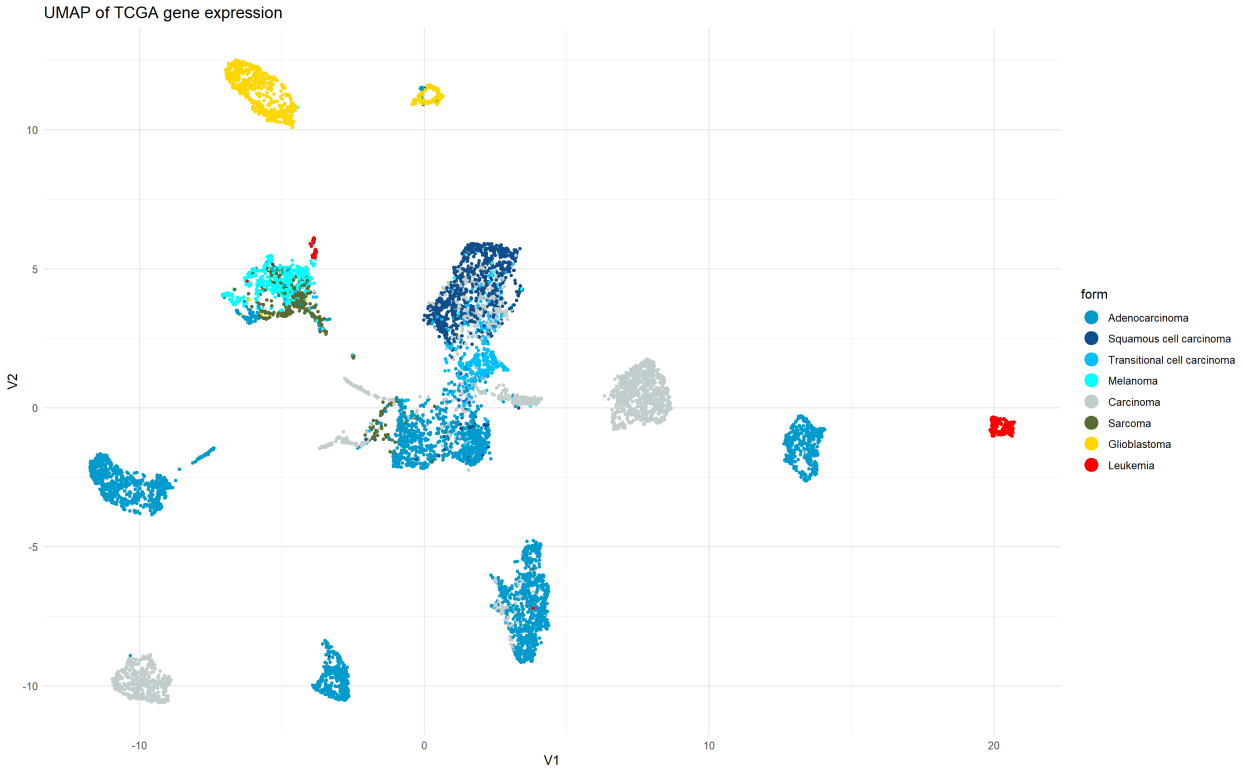

Figure 10: UMAP performed for gene expression data, colored by hiytological type

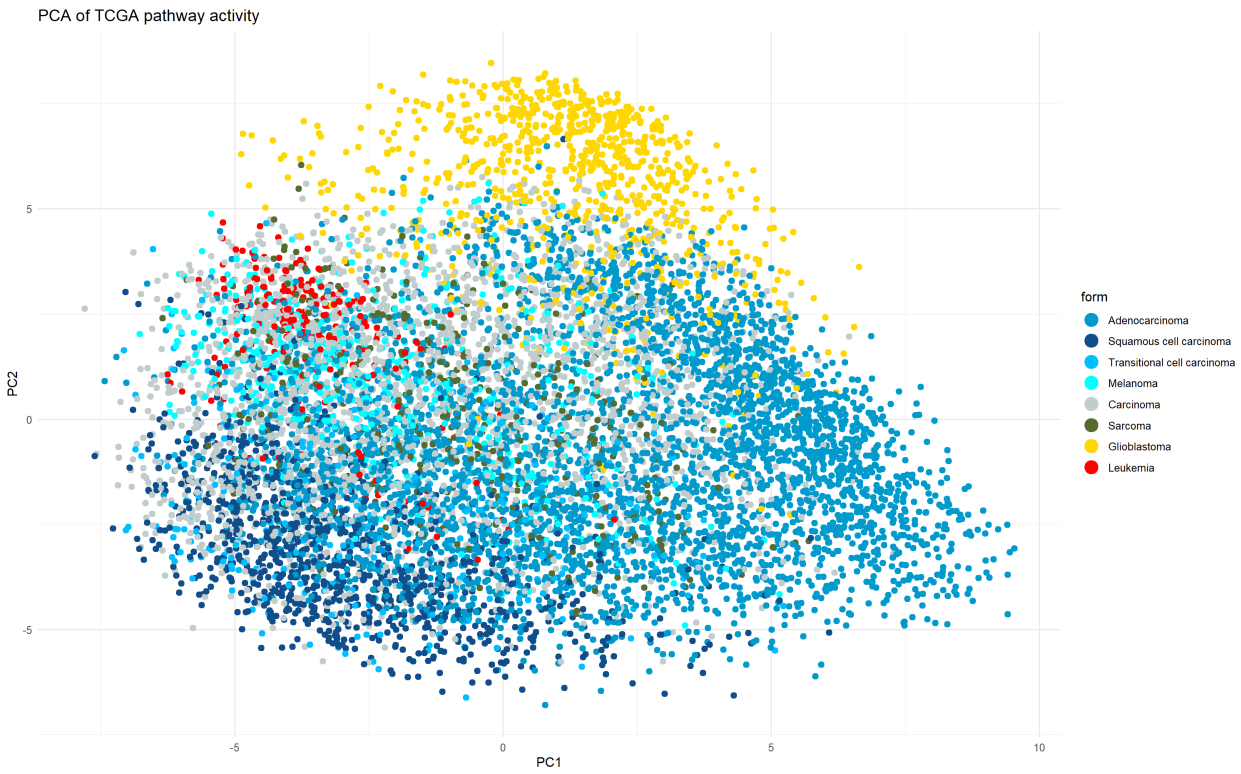

Figure 11: Results of PCA, PC 1 and 2 are shown, samples are colored by histological type.

Figure 12: Results of PCA, PC 1 and 2 are shown, samples are colored by cancer type.

Table 1: Packages used in the analysis.

Package	Usage	Authors
biomart	renaming genenames to ensembleIDs	Durinck *et al.,* 2009
msigdbr	downloading canonical pathways and included genes from MSigDB	Bhuva *et al.,* 2022
dplyr	editing of dataframes	Wickham *et al.,* 2022
ggplot2	creation of plots with detailed options	Wickham *et al.,* 2022
pheatmap	creation of heatmaps with detailed options	Kolde, 2019
vioplot	creation of violinplots	Kolde, 2019
VennDiagram	creation of VENN-diagrams	Chen, 2022
fgsea	performing a GSEA	Korotkevich *et al.,* 2019
GSVA	performing a GSVA	Hänzelmann *et al.,* 2013
ComplexHeatmap	creation of heatmaps with detailed options	Gu *et al.,* 2016
metaplot	creating of data-driven plots	Bergsma, 2019
gridExtra	implementing of "grid" graphics	Auguie and Antonov, 2017
umap	creating UMAPs	Konopka, 2022
gage	application of GSEA	Luo *et al.,* 2009
psych	performing an iterative factor analysis	Revelle, 2022
cluster	performing a cluster analysis	Maechler *et al.,* 2022
MASS	implementing of neural network	Ripley *et al.,* 2022
neuralnet	training of neural networks	Fritsch *et al.,* 2019
AnnotationDbi	translating ensemble ids into gennames	Pagès *et al.,* 2022
org.Hs.eg.db	translating ensemble ids into gennames	Carlson, 2019