Ruprecht-Karls-University Heidelberg
Faculty for Life Sciences
Molecular Biotechnology

Thyorid cancer: Comparison of linar model and neuronal network (xxx)

3-Sätze-Zusammenfassung

sfsf

Data Science Project SoSe 2022

Autoren Anna Lange, David Matuschek, Jakob Then, Maren Schneider
Abgabetermin 20.07.2022

Abstract

In the recent years bioinformatic methods became a tool of utmost importance in medical research. To define specific genes and pathways in different cancer types or histological types pan-cancer analysis are done. A focused analysis is done to specify different subcategories within a certain cancer type and to identify targets for targeted therapy. The main methods in identifying up- or down-regulated pathways are GSEA and GSVA. GSVA of TCGA expression data reveals four clusters of cancer types, which are defined by different histological types like glioblastoma and adenocarcinoma. The histological types therefore seems to correlate with a specific set of pathways being especially enriched in certain cancer types. Furthermore, a GSVA of Thyroid cancer expression data shows that thyroid carcinogenisis is associated with the up-regulation of proliferative signalling pathways like the hedgehog pathway and alpha6beta4 integrin signaling pathway and associated pathways such as IL-36 signaling. It also showed the down-regulation of a pathway that is associated with an increased MAP-kinase acitivity. It is based on those proliferative signalling pathways that three subclusters form inside of the THCA patients from the pan-cancer data. One THCA subtype that could be linked to the follicular histological subtype is defined by increased mTOR and MAPK acitivity, while having low alpha6beta4 activity. In contrast another THCA subtype is defined by a low mTOR and MAPK activity, but a high alpha6beta4 activity. The third THCA subtype is linked to enhanced activity of both of these proliferative signalling pathways. These results promise better results is treatment, as a more percise diagnosis of the distinct THCA subtypse is possible To improve the understanding of THCA and thereby hopefully improve patients prognosis, this project focuses on finding genes that have a significantly different expression in THCA compared to other cancers and especially to normal tissue.

Abbreviations

xxxxxx

Contents

Abstract 2

Abbreviations 3

1 Introduction 5
- 1.1 Biological background 5
- 1.2 Computational tools 6
- 1.3 The analysis xxx nochaml lesen !!! 8

2 Materials and Methods 10
- 2.1 Preprocessing of expression data 10
- 2.2 Methods for descriptive analysis 11
- 2.3 Dimension reduction and pathway enrichment analysis 12
- 2.4 Regression analysis 12
- 2.5 Environment 13

3 References 14

4 Appendix 16

1 Introduction

1.1 Biological background

In 2019 230,000 cancer deaths were documented in Germany[1]. To detect and fight tumors, the development of new treatment and detection methods is essential. For that it is beneficial to find a similarities in mutational causes across different tumors by using transcriptomic profiling methods like RNA-seq. In transcriptomic profiling all the RNA that has been generated by transcription of a cells DNA is sequenced (Alberts and Walter, 2015).

The Hallmarks of Cancer are properties of tumors, that can be detected in each tumor. Among others those are: resisting cell death, inducing angiogenesis, enabling replicative immortality, activating invasion and metastasis evading growth suppressors were the first detected hallmarks (Hanahan and Weinberg, 2011).

The observed tumors can be classified into different histological types. Carcinoma, which can be further subcategorized into adenocarcinomas, squamous cell carcinoma, transitional cell carcinoma. Carcioma derive from epithelial cells. Melanoma are skin tumors, sarcoma derive from connective or supportive tissue cells, glioblastoma are brain tumors and leukemia affect bloodcells (Alberts and Walter, 2015).

RNA-sequencing (RNA-seq) is performed by cleaning of RNA, fragmentation, translation of RNA to cDNA, sequencing of cDNA and comparison with a 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 (Alberts and Walter, 2015). The results from RNA-seq used for the analysis stem from data from the cancer genome atlas (TCGA).

Thyroid carcinoma (THCA) incidence increased dramatically over the past few years (Cabanillas *et al.*, 2016). The main tasks of the thyroid gland are synthesizing hormones and regulating body temperate and metabolism (Tsibulnikov *et al.*, 2020). Most THCA derive from thyroid cells and result in the thyroid gland losing its function. Thyroid cancer can occur in two

[1] https://www.krebsinformationsdienst.de/tumorarten/grundlagen/krebsstatistiken.php

different types, differentiated and undifferentiated thyroid cancer. Those two types again have histological subtypes. Papillary thyroid cancer (PTC), the most common THCA, follicular thyroid cancer (FTC) and a tall cell variant (TCV) are subtypes of differentiated thyroid cancer (DTC). Medullary and anaplastic thyroid cancer are subtypes of undifferentiated thyroid cancer (UTC). Prevalence of DTCs is clearly higher than of UTCs (Prete *et al.*, 2020). Regarding the presented DTCs, PTCs have the best clinical prognosis (Lin, 2007), while TCV cancers have the worst clinical outcome (Coca-Pelaz *et al.*, 2020). Therefore, the detection of the tumor type would be important and for more specific therapy options. Even though, all thyroid cancers are treated with thyroidectomy and radioactive iodine, the additional therapy differs for each histological type (Kant *et al.*, 2020).

Integrin is a cellular adhesion molecule, that binds to laminin in the extracellular matrix (Liberzon *et al.*, 2015a). Together with other proteins they form hemidesmosomes. Thereby, integrin is essential for the integrity between cells. An important step in the development of malignant cancer is the invasion into healthy tissue. Thus, the detachment of the extracellular matrix from of the surrounding cells is essential and alterations of integrin are very common in cancer cells (Rabinovitz and Mercurio, 1996).

1.2 Computational tools

To analyse how the activity of a gene set differs between two sets of gene expression data, a Gene Set Enrichment Analysis (GSEA) is performed. For this, the genes in the expression data have to be ranked decreasingly by a certain metric. Such metrics can include the log2 fold change between the sample expression data and a reference set or the associated p-values for each gene. After ranking, a cumulative sum of all expression values in the ranked sample is computed. If a gene is present in the gene set to be analysed the expression value of that gene is added to the running sum. However, if the current gene does not lie in the gene set the value is subtracted. The extremum of this running sum is termed the enrichment score of the gene set. It is positive if the gene set is overexpressed in the sample compared to the reference data and negative vice versa. (Reimand *et al.*, 2019)

The Gene Set Variation Analysis (GSVA) is performed with the same intention as the GSEA - to analyse the gene set activities in gene expression data. However, no reference data is required to successfully perform GSVA. There are various approaches to GSVA, one of them is performed by (Hänzelmann *et al.*, 2013a) by following five steps. First, the cumulative density distribution of a gene over all samples is estimated. Then the expression statistic of a gene in a sample based on the cumulative density distribution is calculated to bring all of the expression values

to the same level. The third step is to rank the genes based on the expression statistic and to normalize the ranks with z-transformation. Lastly, the enrichment score is computed based on the obtained ranked list by calculating the Kolmogorov-Smirnov-like rank statistic for each gene set. (Hänzelmann *et al.*, 2013a)

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, sthe 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 dataset is reduced to $\mathbb{R}^n$ with the lowest loss in variance.

The Uniform manifold approximation and projection for dimension reduction (UMAP) is a method to reduce the dimension of a multidimensional data set. Compared to PCA, UMAP preserves the global structure of the data better and is much faster than other comparable techniques like t-SNE (Maaten and Hinton, 2008). The algorithm starts by setting up a high-dimensional graph representation of the data. From each data point, a radius is extended and when two radii come into contact the points are connected in the graph. The radius is chosen individually for each point based on the distance to the nearest neighbor. The algorithm goes on until k points are connected or n iterations are reached. The resulting clustered high-dimensional graph is then optimized for a visualization in low-dimensions. A disadvantage of UMAP is that although the overall structure is conserved, the distances between the individual points are not proportional to the real distance in the data set. This arises from the non-linear dimensional reduction. (Sharma *et al.*, 2021)

The Jaccard index is the intersection, divided by the union of two sets. Therefore, it can be used to identify the similarity of the sets.

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.

As an alternative to the linear regression a neuronal network can be used. In general, a deep learning network consists of an input layer, multiple hidden layers, and an output layer consisting

of various neurons (Riedmiller). The input layer contains as much neurons as input numbers are given for each sample. The output is for a regression analysis a singular neuron. The number of neurons in each hidden layer and the number of hidden layers vary and must be tested to give best results. The activation of each neuron can be described as a linear composition of all the inputs x_i from the previous layer associated with a weight w_i and a bias:

$$Activation = \sum_{i=1}^{n} w_i x_i + bias$$

To obtain neuron activations in the range of 0 and 1, a min/max-scaling is performed on the input data. The "learning effect" of the network is achieved by optimizing the randomly chosen weights and biases via gradient decent. To do so, for each training iteration the error of the network is computed by a cost function:

$$Costfunction = \frac{1}{2m} \sum_{i=1}^{m} (x - y)^2$$

m is the number of samples, y the output and x the expected output.

Next, the cost function value must be reduced. Therefore, its gradient is computed, and all weights and biases are adjusted accordingly in a process called backpropagation. In resilient backpropagation, only the sign of the gradient is used, to avoid harmful effects of its magnitude. For the next samples those steps are repeated to reach the minimum of the cost function. A drawback of this method is that gradient decent only identifies local minima of the cost function. To find a global minimum the training has to be repeated with various initial weights and biases. After such a minimum is identified to network performs optimally for the data set.

1.3 The analysis xxx nochaml lesen !!!

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 contains clinical annotations concerning those patients and the last one contains hallmark pathways and their included genes. In the following analysis the data is cleaned by removing NAs, biotype filtering and low-variance filtering. After that a descriptive analysis is performed. After that a gene set variation analysis to detect significantly altered pathways compared to the other pathways in tumor tissue and a linear regression analysis is performed to predict pathway activity based on other pathways??? xxx Furthermore a neuronal network is built to improve prediction.

An analysis of THCA patients is performed. This analysis is done on a data set containing the gene expression data 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. GSVA is performed on 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 other pathways based on thyroxine biosynthesis (nicht mehr!!!!) xxx. A better prediction can be achieved with a neuronal network.

2 Materials and Methods

In the course of this project two separate analysis are performed: a pan-cancer analysis focusing on differences between cancer types and a focused analysis investigating THCA.

For the analysis four data sets were provided. For pan-cancer analysis a gene expression data frame with normalized and log2 transformed bulk RNA-seq expression data for 60,489 genes in 9741 patients with 33 different forms of cancer was used. The data was derived from The Cancer Genome Atlas (TCGA). Complementing the TCGA expression data is an annotation data frame with 37 clinical annotations regarding tumor type, tumor stage, gender, age, etc. for all patients.

The third piece of data is a list containing five lists for the focused analysis, one list for each tumor type (BRCA, KIRC, LUAD, PRAD, THCA). For our focused analysis, only the THCA data were used. The THCA list consists of three data frames: The first two contain normalized and log2 transformed bulk RNA-seq expression data for 19,624 genes in 59 THCA patients for carcinogenic and homeostatic tissue. The third data frame complements the data with the respective clinical annotations.

The last object contains 46 pathways associated with the hallmarks of cancer in form of a list of string vectors.

To perform enrichment analysis later on, 6366 canonical pathways were selected from the Molecular Signatures Database (MSigDB)(Liberzon *et al.*, 2015b) with the `msigdbr::msigdbr()` function. As not to introduce a bias during enrichment analysis, the similarity of MSigDB pathways among themselves as well as with the hallmark pathways was computed with the Jaccard index. Pathways with a Jaccard index greater than the 1σ range were discarded.

2.1 Preprocessing of expression data

All expression data were checked for missing values with the `na.omit()` function. Subsequently, low variance filtering was performed for TCGA and THCA tumor expression data. The variances

of expression were computed for every gene across all samples and then, genes with variances below a threshold were discarded to reduce dimensionality.

Next, biotype filtering was performed for pan-cancer and THCA expression data to reduce dimensionality further. Only genes sharing biotypes with the hallmark pathways were kept for the the following analysis. The biotypes of the genes were retrieved using the `biomart::getBM()` function from the biomaRt package (Durinck *et al.*, 2009). To allow for an appropriate comparison within all pathways, only MSigDB pathways in which over 99% of their respective genes were present in the filtered expression data were selected as final pathways.

2.2 Methods for descriptive analysis

In a mean-variance plot the variance is plotted over the mean of expression values of single genes across all patients. Thus, the variance and mean were calculated for each gene in the THCA expression data. The final plot was created with the package ggplot2 **??**.

Jaccard index is a method to describe the similarity between two quantities. To compute it, the intersection of all gene EnsembleIDs from two compared pathways was divided by their union. We used this method to determine the degree in which pathways are similar to each other.

A volcano plot is used to identify genes displaying significantly different expression in cancerous versus homeostatic tissues. First, the log2 fold change (Log2FC) is calculated for each gene across all samples in the THCA expression data in the following way:

$$log2FC = mean(normaltissue) - mean(tumortissue)$$

Next, a two-sided t-test was performed with the `t.test()` function to determine the significance of a difference in expression. To avoid the accumulation of type one errors, a Bonferroni correction was performed. n is the number of genes in the cleaned data set for focused analysis:

$$\alpha = \frac{0.025}{n}$$

In the volcano plot the -log10 of the calculated p-values is plotted against the Log2FC. Genes with a lower p-value than the corrected significance level α are significantly differently expressed. If the Log2FC is additionally positive, the genes are significantly overexpressed in tumor tissue, if the Log2FC is negative, the genes are significantly underexpressed in tumor tissue.

2.3 Dimension reduction and pathway enrichment analysis

The GSEA was used to identify enriched pathways in THCA tumor tissue. Here, GSEA was performed with the package "fgsea" (Korotkevich *et al.*, 2019). First, the expression values were ranked in decreasing order by Log2FC for every patient. Log2FC was chosen as the ranking metric as it is easy to compute and shows a high sensitivity. **xxx Quelle: Ranking metrics in gene set enrichment analysis: do they matter?** Secondly, using the ranked Log2FC vectors, the enrichment score of each pathway was calculated for each patient with the `fgseamultilevel()` function.

As no normal tissue reference data was provided for the TCGA expression data, pathway activities were computed via GSVA. The analysis was performed with the `gsva()` function from the "GSVA" package (Hänzelmann *et al.*, 2013b). To give a general overview over the differences in expression of THCA and homeostatic thyroid tissue GSVA, the THCA expression data were also analysed by GSVA. To do so, tumor and normal expression data were combined into a singular data frame of which enrichment scores were computed with `gsva()`. Then, the GSVA data was split again and the log2FC between the two matrices was computed and taken as pathway activity.

PCA was performed to provide an uncorrelated data set for the subsequent UMAP. For the TCGA GSVA pathway activity data the `prcomp()` function was used. To verify the results, PCA was performed on TCGA expression data, as well. In this case `Seurat::RunPCA()` from the Seurat package was used to minimize computation time (Hao *et al.*, 2021).

UMAP analysis was done on the principle components from previous PCA to identify and visualize clusters in TCGA GSVA and expression data. This was achieved with the `umap()` function from the package "umap" (Konopka, 2022) running on all PCs from TCGA GSVA ans expression data. The computational effort is lower in UMAP than in PCA. UMAP works on uncorrelated features provided by the PCA.

2.4 Regression analysis

For THCA pathway activity regression analysis a highly variant and significantly altered pathway was selected. To prepare the data appropriately the THCA GSEA data set was divided into a training and test data set containing 44 and 15 samples respectively. A linear regression analysis was performed on the training data with the `glm()` function. To do so, the correlation of all pathways was computed and pathways with high correlations are omitted. Subsequently, the 10% of most variant pathways are selected as variables for the regression model. sA second

model was introduced by computing the p-values of all coefficients and selecting only those pathways contributing significantly to the model were kept.

A neural network was implemented to predict the pathway activity using the `neuralnet()` function from the "neuralnet" package (Fritsch *et al.*, 2019). For identification of the best initial conditions, 25 different networks are generated, each with 2 hidden layers and different combinations of neurons per layer. For each combination the networked was trained on the min-max-scaled training data and the best network was determined by the lowest mean squared error (MSE) in the test data.

2.5 Environment

The R version 4.0.1 was used, the table of used packages is attached in the appendix (see table @ref(tab:packagesused)).

3 References

Alberts, J, B., and Walter, P (2015). Molecular biology of the cell, New York: Garland science.

Cabanillas, ME, McFadden, DG, and Durante, C (2016). Thyroid cancer. Lancet 388, 2783–2795.

Coca-Pelaz, A et al. (2020). Papillary thyroid cancer-aggressive variants and impact on management: A narrative review. Adv Ther 37, 3112–3128.

Durinck, S, Spellman, PT, Birney, E, and Huber, W (2009). Mapping identifiers for the integration of genomic datasets with the r/bioconductor package biomaRt. Nature Protocols 4, 1184–1191.

Fritsch, S, Guenther, F, and Wright, MN (2019). Neuralnet: Training of neural networks.

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

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

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

Hao, Y et al. (2021). Integrated analysis of multimodal single-cell data. Cell 184, 3573–3587.e29.

Kant, R, Davis, A, and Verma, V (2020). Thyroid nodules: Advances in evaluation and management. Am Fam Physician 102, 298–304.

Konopka, T (2022). Umap: Uniform manifold approximation and projection.

Korotkevich, G, Sukhov, V, and Sergushichev, A (2019). Fast gene set enrichment analysis. bioRxiv.

Liberzon, A, Birger, C, Thorvaldsdóttir, H, Ghandi, M, Mesirov, JP, and Tamayo, P (2015b). The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425.

Liberzon, A, Birger, C, Thorvaldsdóttir, H, Ghandi, M, Mesirov, JP, and Tamayo, P (2015a). The molecular signatures database (MSigDB) hallmark gene set collection. Cell Syst 1, 417–425.

Lin, JD (2007). Papillary thyroid carcinoma with lymph node metastases. Growth Factors 25, 41–49.

Maaten, L van der, and Hinton, G (2008). Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605.

Prete, A, Borges de Souza, P, Censi, S, Muzza, M, Nucci, N, and Sponziello, M (2020). Update on fundamental mechanisms of thyroid cancer. Front Endocrinol (Lausanne) 11, 102.

Rabinovitz, I, and Mercurio, AM (1996). The integrin alpha 6 beta 4 and the biology of carcinoma. Biochem Cell Biol 74, 811–821.

Reimand, J et al. (2019). Pathway enrichment analysis and visualization of omics data using g:profiler, GSEA, cytoscape and EnrichmentMap. Nature Protocols 14, 482–517.

Riedmiller, MA Rprop - description and implementation details.

Sharma, S, Quinn, D, Melenhorst, JJ, and Pruteanu-Malinici, I (2021). High-dimensional immune monitoring for chimeric antigen receptor t cell therapies. Current Hematologic Malignancy Reports 16, 112–116.

Tsibulnikov, S, Maslov, L, Voronkov, N, and Oeltgen, P (2020). Thyroid hormones and the mechanisms of adaptation to cold. Hormones (Athens) 19, 329–339.

4 Appendix