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

Cancer Hallmark and Metabolic Pathways in Cancer
Topic 02 Team 03
Exploration of Lung Adenocarcinoma (LUAD)

Data Science Project SoSe 2022

Autoren Paul Brunner, Marie Kleinert, Felipe Stünkel, Chloé Weiler

Abgabetermin 18.07.2022

1 Introduction

1.1 Biological Background

1.2 The Pan Cancer Project

The Cancer Genome Atlas (TCGA) is a publicly available collection of datasets that store the most important cancer-causing genomic alterations in order to create an 'atlas' of cancer genomic profiles. (Tomczak et al., 2015). In 2012 TCGA Research Network launched the Pan-Cancer analysis project as a globally coordinated initiative whose main objective is to assemble coherent, consistent TCGA data sets across twelve different tumor types, one of which being lung adenocarcinoma (LUAD). Each tumor type is characterized using six different genomic, proteomic, epigenomic, and transcriptional platforms. Data collected from thousands of patients is analysed and interpreted in an attempt to gain a deeper understanding of the genomic changes that drive a normal cell to become cancerous. In the future, the aim is to analyse additional tumor types beyond the twelve original ones in the hopes that the pan cancer project will one day inform clinical decision-making and aid in the development of novel therapeutic options

$$@weinstein2013$$

.

1.3 Jaccard index

The Jaccard index is a widely known measurement for the similarity between finite sample sets. The restricted domain is defined between zero and one. A Jaccard index close to one indicates a high similarity of the sample sets

$$@jaccard1901$$

.

$$J(A, B) = \frac{A \cap B}{A \cup B}$$

1.4 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a procedure used to perform linear dimension reduction. The goal is to reduce the dimension of a given data set whilst losing as little information as possible by retaining a maximum of the standardized data set's variation (Ringnér, 2008).

Principal components (PC) are a set of new orthogonal variables that are made up of a linear combination of the original variables. Principal components display the pattern of similarity of the observations and of the variables as points in maps (Abdi and Williams, 2010). By convention, the PCs are ordered in decreasing order according to the amount of variation they explain of the original data (Ringnér, 2008). It is important to note that all PCs are uncorrelated.

PCA is a useful tool for genome-wide expression studies and often serves as a first step before clustering or classification of the data. Dimension reduction is a necessary step for easy data exploration and visualization (Ringnér, 2008).

1.5 Uniform Manifold Approximation and Projection (UMAP)

Uniform manifold approximation and projection (UMAP) is a k-neighbour graph based algorithm that is used for nonlinear dimension reduction (Smets et al., 2019) (McInnes et al., 2018).

After data normalization, the Euclidean distances between points in a two-dimensional space of the graph are calculated and a local radius is determined (Vermeulen et al., 2021). In general the closer two points are to each other, the more similar they are. UMAP makes a density estimation to find the right local radius. This variable radius is smaller in high density regions of data points and larger in low density regions. In general, the density is higher when the k-nearest neighbour is close and vice versa. The number of k-nearest neighbours controls the number of neighbours whose local topology is preserved. Precisely, a large number of neighbours will ensure that more global structure is preserved whereas a smaller number of neighbours will ensure the preservation of more local structure (McInnes et al., 2018).

UMAP is a newer method tan PCA and it is generally believed to be easier to interpret and group data than by using PCA. Furthmore, UMAP has the advantage of not requiring linear data (Milošević et al., 2022).

1.6 Gene Set Enrichment Analysis (GSEA)

Gene set enrichment analysis (GSEA) is a computational method that is used to determine whether two gene expression states are significantly different from each other or not. In this project we compared gene expression profiles between healthy and tumorous tissue of LUAD(Subramanian et al., 2007).

Two data sets are compared and the genes are sorted from the most to the least differential expression between the data sets according to their p-values. This creates a ranked list L.

Referring to an *a priori* defined set of gene sets S, the goal is to locate for each pathway of S where its corresponding genes fall in L and find a discerning trend. If the genes of a given pathway are randomly distributed in L then the pathway is assumed to not significantly contribute to the particular tumor's phenotype. However if the genes are primarily clustered at the top or the bottom of L then a phenotypic significance of the given pathway can be assumed.

To determine the location of the genes, an enrichment score is calculated for each pathway. For this, a running-sum statistic is calculated as the list L is ran through. The running-sum is increased every time a gene belonging to the pathway in question is encountered and decreased otherwise. An enrichment score is thus calculated for each pathway. The enrichment score is defined as the maximum deviation from zero of the running-sum.

Lastly, adjustment for multiple hypothesis testing is performed by normalizing the enrichment score for each pathway to account for its size and a normalized enrichment score is obtained.

GSEA is a useful tool for interpretation of gene expression data.

(Subramanian et al., 2005)

1.7 Gene Set Variation Analysis (GSVA)

Gene set variation analysis (GSVA) is an unsupervised method to estimate pathway activities based on gene expression data. Contrarily to the aforementioned GSEA, GSVA does not rely on phenotypic characterisation of the data sets into two categories but rather quantifies enrichment in a sample-wise manner which makes GSVA the better choice to perform on the tcga_exp data set.

GSVA estimates a cumulative distribution for each gene over all samples. The gene expression values are then converted according to these estimated cumulative distributions into scaled values. Based on these new values, the genes are ranked in each sample. Next, the genes are classified into two distributions and a Komogorow-Smirnow statistic is calculated to judge how similar the distributions are to each other and to obtain an ES.

The GSVA corresponds to either the maximum deviation between both running sums or the GSVA score can be defined as the difference of the maximum deviations in the positive and in the negative direction. A highly positive or negative GSVA score indicates that the studied gene set is positively or negatively enriched compared to the genes not in the gene set, respectively. If the GSVA score for a given gene set is close to zero, then the gene set is probably not differentially expressed compared to the genes not in the gene set.

(Hänzelmann et al., 2013a)

2 Abbreviations

GSEA	gene set enrichment analysis
GSVA	gene set variation analysis
LUAD	lung adenocarcinoma
PC	principal component
PCA	principal component analysis
RNA-seq	RNA sequencing
TCGA	The cancer genome atlas
UMAP	uniform manifold approximation and projection

3 Methods

3.1 Our Data

At the beginning of this project we were given four datasets of which two contained RNA-seq data, one contained clinical annotations pertaining to one of the RNA-seq data frames and one contained a list of gene sets for cancer hallmark analysis.

The first RNA-seq dataset is a data frame containing RNA-seq data from almost 10,000 TCGA cancer patients for 33 different tumor types. The data stored within that data frame was used to perform pan cancer analysis and to create a logistic regression model. The second RNA-seq dataset is a smaller data frame containing the TCGA expression data of tumor tissue and the corresponding healthy tissue for five different cancer types. A focused analysis was performed on this second dataset.

RNA sequencing (RNA-seq) data makes it possible to go beyond static genome analysis and to gain an insight into the transcriptional landscape of a cell. Studying gene expression profiles of a given cell, regulated by RNA synthesis, enables researchers to gain a deeper understanding of how gene expression is regulated in cells and its impact on the cell's phenotype (Marguerat and Bähler, 2010).

All expression data was log_2(TPM) transformed. Log_2 transformation is a commonly used tool to reduce skewness in data and to make it more conform to a normal distribution.Here, TPM stands for 'Transcripts per million' and refers to a method of RNA-seq normalization in which one first accounts for gene length before adjusting for sequencing depth, or count bias, in the data. A possible perk of TPM is the reduction of type I and type II errors which would otherwise falsify downstream analysis results by accounting for gene length first (Yuen In and Pincket, 2022).

3.2 Overview of used packages

Table 3.1: Tab. 1: Used packages in alphabetical order.

Package Name	Application	**Reference**
babyplots	create interactive 3D visualizations	Trost (2022)
base	basic R functions	R Core Team (2022a)
bayesbio	calculate Jaccard coefficients	McKenzie (2016)
BiocParallel	novel implementations of functions for parallel evaluation	(Morgan et al., 2021)
biomaRt	access to genome databases	Durinck et al. (2009)
blorr	building and validating binary logistic regression models	Hebbali (2020)
cinaR	combination of different packages	Karakaslar and Ucar (2022)
cluster	cluster analysis of data	Maechler et al. (2021)
ComplexHeatmap	arrange multiple heatmaps	(Gu et al., 2016)
edgeR	assess differential expression in gene expression profiles	Chen et al. (2016)
EnhancedVolcano	produce improved volcano plots	(Blighe et al., 2021)
enrichplot	visualization of gene set enrichment results (GSEA)	Yu (2022)
FactoMineR	perform principal component analysis (PCA)	Lê et al. (2008)
fgsea	Run GSEA on a pre-ranked list	Korotkevich et al. (2019)
ggplot2	visualization of results in dot plots, bar plots and box plots	Wickham (2016)
ggpubr	formatting of ggplot2-based graphs	Kassambara (2020)
grid	implements the primitive graphical functions that underlie the ggplot2 plotting system	R Core Team (2022b)

Package Name	Application	**Reference**
gridExtra	arrange multiple plots on a page	Auguie (2017)
GSVA	Run GSVA on a data set	(Hänzelmann et al., 2013b)
gplots	plotting data	(Warnes et al., 2022a)
gtools	calculate foldchange, find NAs, logratio2foldchange	Warnes et al. (2022b)
knitr	creation of citations using write_bib	Xie (2014)
limma	“linear models for microarray data”	Ritchie et al. (2015)
msigdbr	provides the ‘Molecular Signatures Database’ (MSigDB) gene sets	Dolgalev (2022)
parallel	allows for parallel computation through multi core processing	R Core Team (2022c)
pheatmap	draw clustered heatmaps	Kolde (2019)
RColorBrewer	provides color schemes for maps	Neuwirth (2022)
ROCR	visualizing classifier performance	Sing et al. (2005)
scales	helps in visualization: r automatically determines breaks and labels for axes and legends	Wickham and Seidel (2022)
Seurat	visualize gene set enrichment results in dot plots	Satija et al. (2022)
tidyverse	collection of R packages, including ggplot2	Wickham et al. (2019)
uwot	performs dimensionality reduction and Uniform Manifold Approxiamtion and Projection (UMAP)	Melville (2021)

3.3 Gene Set Extraction

The Molecular Signature Database (MSigDB) is a database offering a variety of annotated gene sets publicly available for analysis. The import of gene sets from MSigDB into

RStudio can easily be performed using the R package “msigdbr” (Dolgalev, 2022) which allows the extraction of species-specific gene sets of the category of interest. In a final step, the prefix corresponding to the source of the gene set was removed. The resulting output was a list containing all selected gene sets with the comprising genes saved in a vector and each element named after the pathway stored in it. This workflow was conducted to extract curated (C2) and ontology (C5 BP) gene sets which were used for focused analysis as well as pan cancer analysis. The curated gene sets that regulate the metabolism of cells were also used for comparison with known pathways deregulated in cancer cells.

Contents

4 Results

4.1 Cancer hallmark pathways

By calculating the Jaccard index for each metabolism gene set to each hallmark gene set, the similarity between these pathways was measured and visualized in a heatmap (**Fig. XXX**). This highlights that there is a general low similarity between the selected pathways and only a few gene sets show a slightly higher similarity which don't exceed an index of 0.2. The most shared genes are found in the alanine aspartate and glutamate metabolism with glutamine metabolism. Additionally, large overlaps are found in purine and pyrimidine metabolism with genome repair and down regulation as well as in lipid and fatty acid metabolism with VEGF-induced angiogenesis.

5 Discussion

6 Outlook

6.1 Analysis of metastasis formation in LUAD

About a third of LUAD patients already present with brain metastases at the time of diagnosis and about half of all patients will eventually develop brain metastases (Shih et al., 2020). As a cancer's ability to metastasize dramatically impacts a patient's chances of survival, a comprehensive analysis of the genetic alterations that most often lead to metastasis formation would allow to single out patients at risk for developing metastases and to treat them accordingly.

The comparison of genomic expression profiles of cancer cells taken from the metastases with those taken from the primary tumor might reveal which mutations enable a lung cancer cell to metastasize.

6.2 Smoker vs non-smoker

It is a well-known fact that lung cancer is a smoker's disease. However, in recent years studies have found that lung cancer incidence is decreasing in smokers and increasing in non-smokers. Furthermore, the same study states that the genomic profile of lung cancer in non-smokers differs from that in smokers. (Qiu et al., 2015). Inspired by this study, a possible next step would be to subgroup the data into smokers and non-smokers and to compare the two groups in order to determine which pathways are differentially expressed and if the results of Qui *et al* can be replicated with our data.

6.3 Identification of Immune Subtypes

Many different research groups have already set out to subtype LUAD according to immune signature defined for instance by PD-L1 expression or immune cell infiltration. Generally, the more pronounced a cancer's immune signature is, the better the cancer will respond to immunotherapy and the better a patient's chances of survival (Xu et al., 2020). Based on the findings of the likes of Xu *et al.*, our own data could be divided according to immune signature and determining trends within each subgroup such as survival rate could be identified.

6.4 Find defining trends between LUAD clusters

Performance of UMAP and PCA on our data (**Figure XXX**) showed that LUAD forms two distinct clusters. Further analysis may reveal the genetic differences within LUAD that lead to clustering as well as the genetiv similarities shared by samples belonging to the same cluster. Additionally, response to therapy might differ between clusters and thus patients belonging to one cluster or another might face vastly different chances of survival.

6.5 Prediction of cancer stage

Over the course of this project we trained a logistic regression model to predict whether an individual is at risk of eventually developing LUAD. This model could be further sophisticated to additionally predict the cancer stage on the grounds of a patient's genomic profile as well as other factors like age or smoking habits. With enough training, this model could ideally be used as a less invasive alternative to the current diagnostic methods and therefore help determine an adequate treatment plan with reduced patient trauma.

6.6 Epigenetics

The term epigenetics describes hereditary changes in gene expression that are not due to changes in the DNA sequence. The most common epigenetic alterations in cancer cells include global hypomethylation of repetitive DNA sequence regions and hypermethylation of tumor suppressor genes which are consequently inactivated (Esteller, 2008). Analysis of the differences in methylation patterns between tumorous and healthy tissue would allow to determine which tumor suppressor genes are most commonly inactivated in LUAD. Furthermore differences in methylation patterns between LUAD and other cancer types could further help to distinguish the process of LUAD development from that of other cancers.

6.7 Experimental validation

The ultimate step of any analysis would be to seek empirical confirmation of novel findings. Since TCGA holds immunohistochemistry stained samples of tumor tissue it would be possible to directly study in what way different genomic profiles impact the development of tumors *in vivo*.

7 References

Abdi, H., and Williams, L.J. (2010). Principal component analysis. WIREs Computational Statistics *2*, 433–459.

Auguie, B. (2017). gridExtra: Miscellaneous functions for "grid" graphics.

Blighe, K., Rana, S., and Lewis, M. (2021). EnhancedVolcano: Publication-ready volcano plots with enhanced colouring and labeling.

Chen, Y., Lun, A.A.T., and Smyth, G.K. (2016). From reads to genes to pathways: Differential expression analysis of RNA-seq experiments using rsubread and the edgeR quasi-likelihood pipeline. F1000Research *5*, 1438.

Dolgalev, I. (2022). Msigdbr: MSigDB gene sets for multiple organisms in a tidy data format.

Durinck, S., Spellman, P.T., 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.

Esteller, M. (2008). Epigenetics in cancer. New England Journal of Medicine *358*, 1148–1159.

Gu, Z., Eils, R., and Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics.

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

Hänzelmann, S., Castelo, R., and Guinney, J. (2013a). GSVA: Gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics *14*, 1–15.

Hebbali, A. (2020). Blorr: Tools for developing binary logistic regression models.

Karakaslar, O., and Ucar, D. (2022). cinaR: A computational pipeline for bulk 'ATAC-seq' profiles.

Kassambara, A. (2020). Ggpubr: 'ggplot2' based publication ready plots.

Kolde, R. (2019). Pheatmap: Pretty heatmaps.

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

Lê, S., Josse, J., and Husson, F. (2008). FactoMineR: A package for multivariate analysis. Journal of Statistical Software *25*, 1–18.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K. (2021). Cluster: Cluster analysis basics and extensions.

Marguerat, S., and Bähler, J. (2010). RNA-seq: From technology to biology. Cellular and Molecular Life Sciences *67*, 569–579.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction (arXiv).

McKenzie, A. (2016). Bayesbio: Miscellaneous functions for bioinformatics and bayesian statistics.

Melville, J. (2021). Uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction.

Milošević, D., Medeiros, A.S., Stojković Piperac, M., Cvijanović, D., Soininen, J., Milosavljević, A., and Predić, B. (2022). The application of uniform manifold approximation and projection (UMAP) for unconstrained ordination and classification of biological indicators in aquatic ecology. Science of The Total Environment *815*, 152365.

Morgan, M., Wang, J., Obenchain, V., Lang, M., Thompson, R., and Turaga, N. (2021). BiocParallel: Bioconductor facilities for parallel evaluation.

Neuwirth, E. (2022). RColorBrewer: ColorBrewer palettes.

Qiu, M., Xu, Y., Wang, J., Zhang, E., Sun, M., Zheng, Y., Li, M., Xia, W., Feng, D., Yin, R., et al. (2015). A novel lncRNA, LUADT1, promotes lung adenocarcinoma proliferation via the epigenetic suppression of p27. Cell Death & Disease *6*, e1858–e1858.

R Core Team (2022a). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

R Core Team (2022b). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

R Core Team (2022c). R: A language and environment for statistical computing (Vienna, Austria: R Foundation for Statistical Computing).

Ringnér, M. (2008). What is principal component analysis? Nature Biotechnology *26*, 303–304.

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research *43*, e47.

Satija, R., Butler, A., Hoffman, P., and Stuart, T. (2022). SeuratObject: Data structures for single cell data.

Shih, D.J., Nayyar, N., Bihun, I., Dagogo-Jack, I., Gill, C.M., Aquilanti, E., Bertalan, M., Kaplan, A., D'Andrea, M.R., Chukwueke, U., et al. (2020). Genomic characterization of human brain metastases identifies drivers of metastatic lung adenocarcinoma. Nature Genetics *52*, 371–377.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: Visualizing classifier performance in r. Bioinformatics *21*, 7881.

Smets, T., Verbeeck, N., Claesen, M., Asperger, A., Griffioen, G., Tousseyn, T., Waelput, W., Waelkens, E., and De Moor, B. (2019). Evaluation of distance metrics and spatial autocorrelation in uniform manifold approximation and projection applied to mass spectrometry imaging data. Analytical Chemistry *91*, 5706–5714.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences *102*, 15545–15550.

Subramanian, A., Kuehn, H., Gould, J., Tamayo, P., and Mesirov, J.P. (2007). GSEA-P: a desktop application for Gene Set Enrichment Analysis. Bioinformatics *23*, 3251–3253.

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review the cancer genome atlas (TCGA): An immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia *2015*, 68–77.

Trost, N. (2022). Babyplots: Easy, fast, interactive 3D visualizations for data exploration and presentation.

Vermeulen, M., Smith, K., Eremin, K., Rayner, G., and Walton, M. (2021). Application of uniform manifold approximation and projection (UMAP) in spectral imaging of artworks. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy *252*, 119547.

Warnes, G.R., Bolker, B., and Lumley, T. (2022b). Gtools: Various r programming tools.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., et al. (2022a). Gplots: Various r programming tools for plotting data.

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis (Springer-Verlag New York).

Wickham, H., and Seidel, D. (2022). Scales: Scale functions for visualization.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L.D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. Journal of Open Source Software *4*, 1686.

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In Implementing Reproducible Computational Research, V. Stodden, F. Leisch, and R.D. Peng, eds. (Chapman; Hall/CRC),.

Xu, F., Chen, J., Yang, X., Hong, X., Li, Z., Lin, L., and Chen, Y. (2020). Analysis of lung adenocarcinoma subtypes based on immune signatures identifies clinical implications for cancer therapy. Molecular Therapy-Oncolytics *17*, 241–249.

Yu, G. (2022). Enrichplot: Visualization of functional enrichment result.

Yuen In, H.L., and Pincket, R. (2022). Transcripts per million ratio: A novel batch and sample control method over an established paradigm. arXiv e-Prints arXiv–2205.

8 Appendix