Gene expression profile analysis in gastric cancer: the role of collagen and the extracellular matrix

Linda Cova

Github: https://github.com/lindacova/NBDA

Network-Based Data Analysis
Prof. Lauria
University of Trento
A.Y. 2021-2022

Abstract

One of the main advantages of high-throughput sequencing is the possibility of exploiting the same results for a number of different purposes. In this project, a gene expression dataset obtained from a public repository was analysed in order to potentially discover something different from the aim it was originally generated for. Several classification models were evaluated on this data and the best-performing one was used to select the most relevant genes to perform network and functional enrichment analysis. Thanks to this, the role of collagen in relation to gastric cancer was highlighted.

Introduction

Gastric cancer is one of the leading causes of cancer-related deaths worldwide. The 5-year survivals in gastric cancer patients can be less than 20%, making this type of cancer as one of the most lethal cancers[1]. The early stages of gastric cancer are usually asymptomatic or associated with nonspecific symptoms, for this reason early detection is problematic and innovative methods for an early diagnosis are possibly be the key to decrease the number of fatalities caused by this disease.

The main diagnostic techniques rely on gastrointestinal endoscopy, imaging techniques such as FDG-PET and staging laparoscopy. In addition to that, the identification of a set of biomarkers related to this pathology could improve diagnosis, prognosis, prediction of recurrence and treatment response[2]. Currently applied biomarkers for gastric cancer include, for example, cell adesion molecules (CEA), glycolipid antigen CA19-9 and proto-oncogene HER2. Other markers were explored as methastasis biomarkers and non-invasive biomarkers were identified in circulating cell-free DNA, microRNA, long-noncoding RNA and exosomes[3].

This project aims at applying classification models on a publicly-available gene expression dataset in order to select the most relevant genes for the discrimination between healthy and tumoral tissue samples. Functional enrichment analysis and network-based analysis are then exploited to detect the functions and pathway that these genes influence the most and that could thus be useful for the identification of biomarkers for the developement novel therapeutic or diagnostic techniques.

Methods

All data processing was performed with the R programming language (version 4.1.3).

0.1 Dataset selection

The data on which this project is carried out is retrieved from the *Recount3* database[4], which provides gene expression datasets obtained from uniformly processed RNA-Seq experiments. It was fetched with the *recount3* R package. The chosen dataset is SRP133891, that includes 68 tissue samples, half from patients diagnosed with gastric cancer and half from cancer-free patients. All individuals are korean and aged between 45 and 88, 28 are female and 40 are male. The samples were collected at the department of internal medicine of the Hallym University Medical Center and sequenced with pair-ended sequencing by the Illumina HiSeq 2500 machine. For each sample, the raw counts of 63865 genes are reported.

0.2 Data pre-processing

First, genes showing very low or none expression (less than 10 total counts over all the samples) were filtered out. The normalization was performed with *geTMM*[5], which allows for both inter- and intrasample analyses with the same normalized data set. This method performs correction for sequencing depth, RNA composition and gene length. It was carried out with the *edgeR* R package [6]. Finally, the obtained values were visualized with a boxplot to determine that a scaling was still required. A logarithmic scaling was applied in order to align all the medians of the samples and their quartiles.

PCA is a method for dimensionality reduction useful for a first visualization of the dataset. The first principal components and their percentages of variance explained were plotted thanks to the R packages *ggplot2* [7] and *factoextra*[8]

0.3 Data clustering

The first classification models tried out are two unsupervised clustering methods: the K-means and the hierarchical clustering. For the purpose of this project it was unlikely for this sort of methods to provide outstanding results, however they were nevertheless evaluated and proved useful for means of data visualization. Given the knowledge that the dataset includes two groups of patients, the K-means algorithm was implemented with the tuning parameter `k=2`. For hierarchical clustering the distance matrix was computed with euclidean distances and the distance between clusters with average linkage.

0.4 Supervised methods for classification

For the supervised methods, the data was split into training and test set and the cross-validation were performed thanks to the *caret* R package[9].

Random Forest Random forest requires the tuning of 2 parameters: *ntree* and the *mtry*. The first is the number of trees to grow and the second is the number of features considered when building each tree. First, the ideal number of trees was determined as 200 (Supplementary FigureS1) with the *randomForest* R packege[10]. Next, cross-validation was employed to determine `mtry=251` as the ideas value and to fit the optimal model in the test set for evaluation.

Linear Discriminant Analysis In order to perform a LDA, a feature selection was necessary to reduce the number of genes considered. A T-test was performed thanks to the *genefilter* R package[11]. LDA was fitted on the features with a P-value lower than 0.1 with a 10-fold cross-validation.

Lasso regression Lasso was tuned and fitted with a 10-fold cross-validation. Multiple values were tested for the tuning parameter *lambda*, while *alpha* was fixed at 1. The optimal value resulted to be `lambda=0.3`.

Scudo The Scudo (R package *rScudo*[12]) method compares gene signatures from different individuals and identifies the most important genes for each sample as the most and less expressed. Due to time constraints, no cross-validation was performed to tune the parameters for this model, instead 25 features from the top and the bottom of each signature were considered.

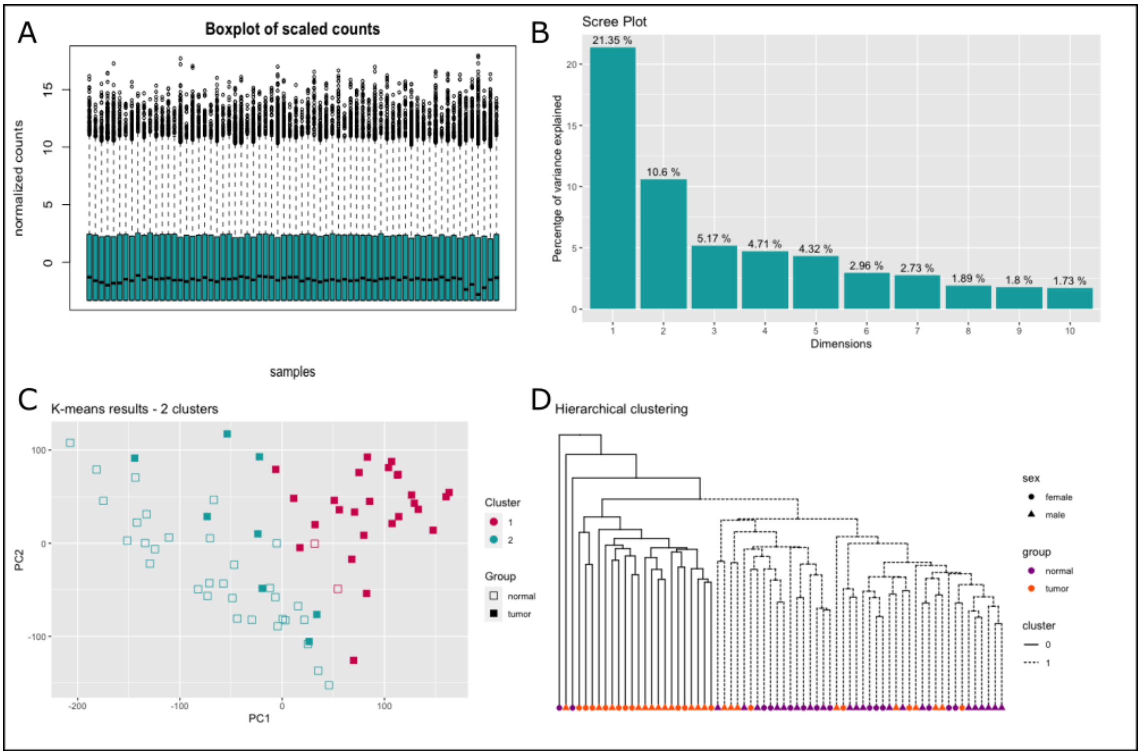


Figure 1: Pre-processing and data clustering

0.5 Models comparison and feature selection

Each fitted model was evaluated considering two metrics: the accuracy and the area under the ROC curve (AUC). For the supervised models, these values were computed after predicting the classification of the samples in the test set using the model fitted on the training set with the optimal parameters. The R package *pROC*[13] was used for the ROC curves and AUC. A feature selection was performed considering the most relevant genes for the best performing model.

0.6 Functional enrichment analysis

The functional enrichment analysis is performed with *gprofiler2* [14], with the objective of finding out the molecular functions more represented among the most important genes. For this purpose, the 200 genes with highest importance score according to the random forest model fitted before are considered. This tool performs statistical enrichment analysis to find over-representation of functions and pathways from different databases. This is done with the hypergeometric test followed by P-value correction for multiple testing.

0.7 Biological Network Analysis

This analysis was carried out with the R package *pathfindR*[15], which allows active subnetwork-oriented pathways analysis: groups of active interconnected genes that are mostly significant. In order to extract the most relevant pathways, *pathfindR* uses both the P-value of the gene list provided as input (computed with *genefilter* as before) and information from the protein-protein interaction network.

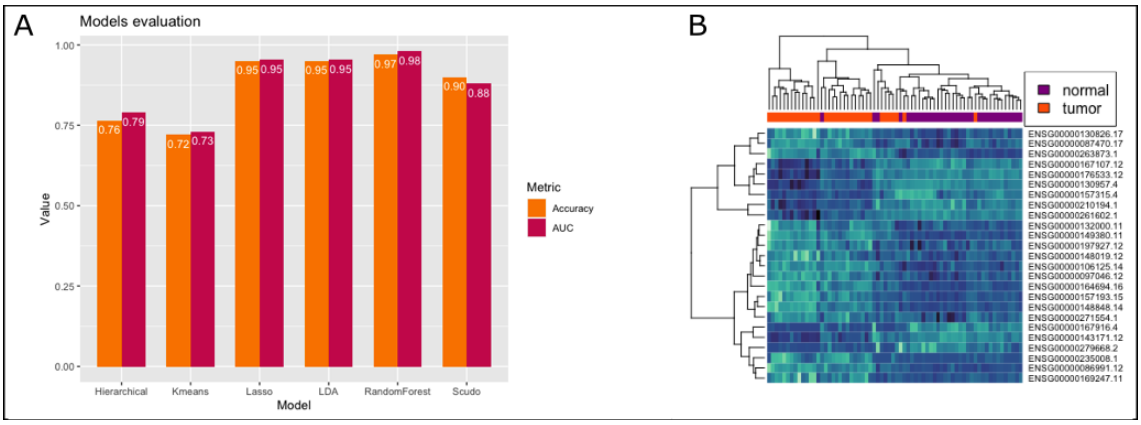


Figure 2: Model evaluation results and feature selection

Results

Pre-Processing After the pre-processing steps, the data distribution for each individual was correctly scaled, as it shows in Figure 1(A). The PCA results were satisfying, with the first principal component explaining 21.35% of variance (Figure 1(B). Different combinations of PCs were plotted (Supplementary Figure S3) but no further interesting results were obtained.

Classification Methods Both the **K-means** and the **hierarchical clustering** show a partially correct classification of the samples, with respectively: `accuracy=0.85`, `AUC=0.86` and `acc=0.76`, `AUC=0.79`. The results are depicted inn Figure 1 (C-D).

The **random forest** was the better-performing model, with `accuracy=0.97`, `AUC=0.98`.

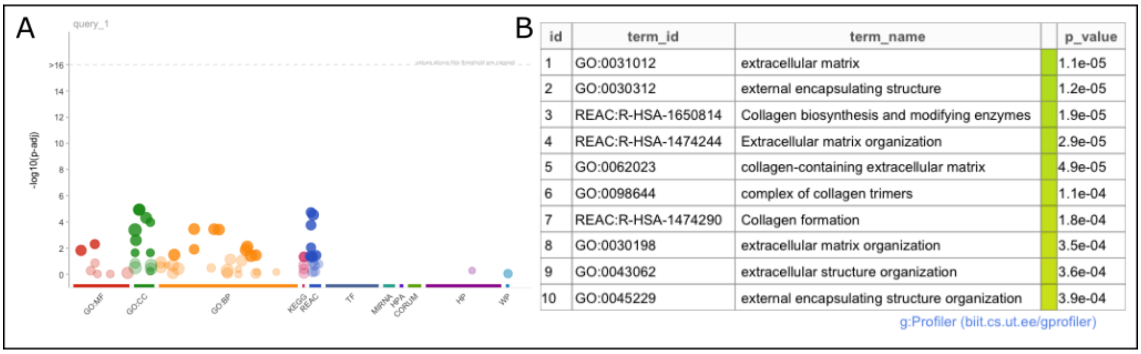
All the other methods had a slightly lower performance: for **LDA** and **lasso** `accuracy=0.95`, `AUC=0.95` and for **scudo** `accuracy=0.90`, `AUC=0.88`. The network resulting from the scudo model is depicted in Supplementary Figure S2 and shows a good classification, although it was only tested on a small test set.

These results are summarized in figure 2 (A) and in the Supplementary Table 1.

Feature selection Since random forest was the best performing model, a feature selection was performed considering the importance score assigned to each gene when fitting the random forest. The most important genes are represented in Figure 2 (B), and the heatmap shows how these genes are indeed differentially expressed between the two groups of samples.

Functional Enrichment Analysis The results of the functional enrichment analysis are reported in Figure 3. Panel (A) shows how there are no terms with an extremely low P-value, but some significant results were found nevertheless in all the databases considered by *gprofiler*. In particular, panel (B) shows the 10 entries with the lowest P-value: all of them are terms from the *Gene Ontology* and from *Reactome* and they are related to the biosynthesis and processing of collagen and to the extracellular matrix structure.

Biological Network Analysis The *pathfindR* analysis was carried out on the *KEGG*, *Gene Ontology* and *Reactome* databases. Among those three, *Reactome* produced the most interesting results, confirming the findings of the previous paragraph. In particular,



id	term_id	term_name		p_value
1	GO:0031012	extracellular matrix		1.1e-05
2	GO:0030312	external encapsulating structure		1.2e-05
3	REAC:R-HSA-1650814	Collagen biosynthesis and modifying enzymes		1.9e-05
4	REAC:R-HSA-1474244	Extracellular matrix organization		2.9e-05
5	GO:0062023	collagen-containing extracellular matrix		4.9e-05
6	GO:0098644	complex of collagen trimers		1.1e-04
7	REAC:R-HSA-1474290	Collagen formation		1.8e-04
8	GO:0030198	extracellular matrix organization		3.5e-04
9	GO:0043062	extracellular structure organization		3.6e-04
10	GO:0045229	external encapsulating structure organization		3.9e-04

Figure 3: Functional enrichment analysis

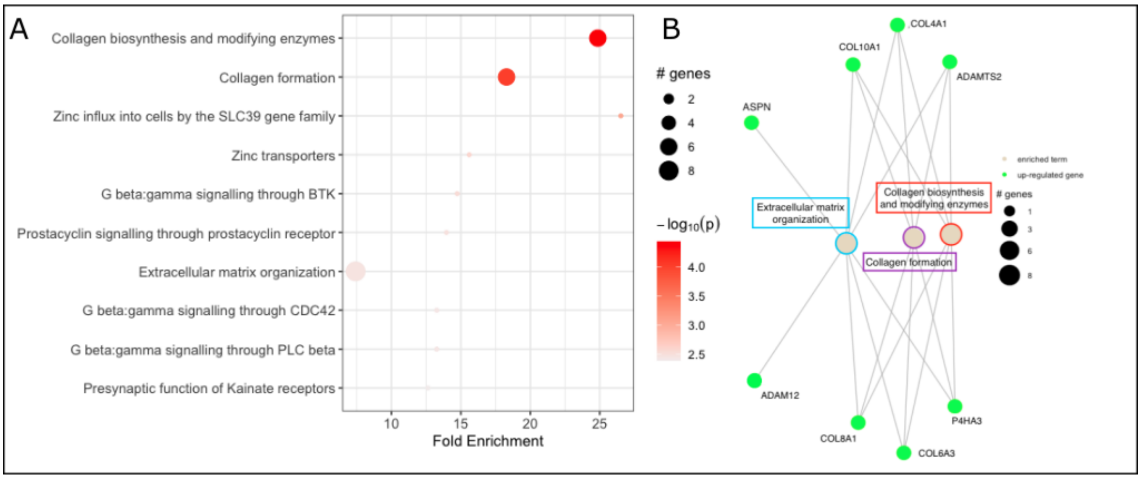


Figure 4: Biological network analysis

there are three terms with significant P-value and 4 genes linked to them: ”Collagen biosynthesis and modifying enzymes”, ”Collagen formation” and ”Extracellular matrix organization”. These results are represented in Figure 4. The network in panel (B) shows the genes related to these terms. Four of them are collagen genes: COL10A1,COL4A1, COL8A1 and COL6A3, ASPN encodes for a cartilage extracellular protein, ADAMTS2 encodes for a protein that is proteolytically processed to generate the mature procollagen N-proteinase, ADAM12 is involved in cell-cell and cell-matrix interactions and P4HA3 encodes a component of prolyl 4-hydroxylase, a key enzyme in collagen synthesis.

Discussion and conclusions

The functional enrichment analysis conducted on the most important genes according to the random forest, which was the best performing classification model among those tested during this project, produced results consistent with the network-based analysis performed on a list of genes selected by T-test. Both these analyses highlighted the role of the extracellular matrix (ECM) and, more in particular, of collagen, in relation to gastric cancer: genes connected to pathways and functions relative to the ECM appear to be differentially expressed in tumor tissue samples. This outcome is not surprising since the components, their proportions and their assembly determine the rigidity, porosity, and other properties of each tissue. It is thus expected that alterations in ECM composition, and subsequently on its mechanical and biochemical properties, will strongly affect cellular communication, function, and behavior and lead to cancer onset[16]. Among the genes that were found to be related to the ECM composition by *pathfindR*, COL4A1 was already studied as potential biomarker for gastric cancer due to its overexpression in tumor samples and its association to poor prognosis and methasiasis[17]. Since collagen genes appear to play crucial roles in the progression of gastric cancer, they may be utilized for diagnosis and therapy.

In conclusion, starting from a gene expression database, it is possible to evaluate and select a classification model with an adequate level of accuracy and to gain biological insight regarding pathways and cellular functions related to a disease.

Supplementary Materials

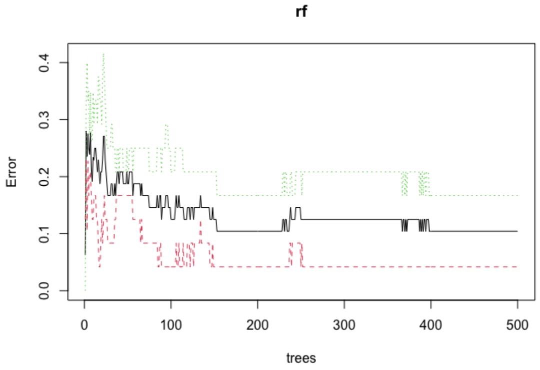


Figure S1: OOB error of the random forest model reaches a plateau with a number of trees aound 200

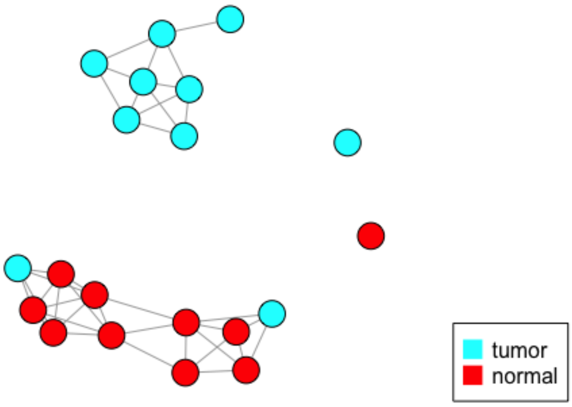
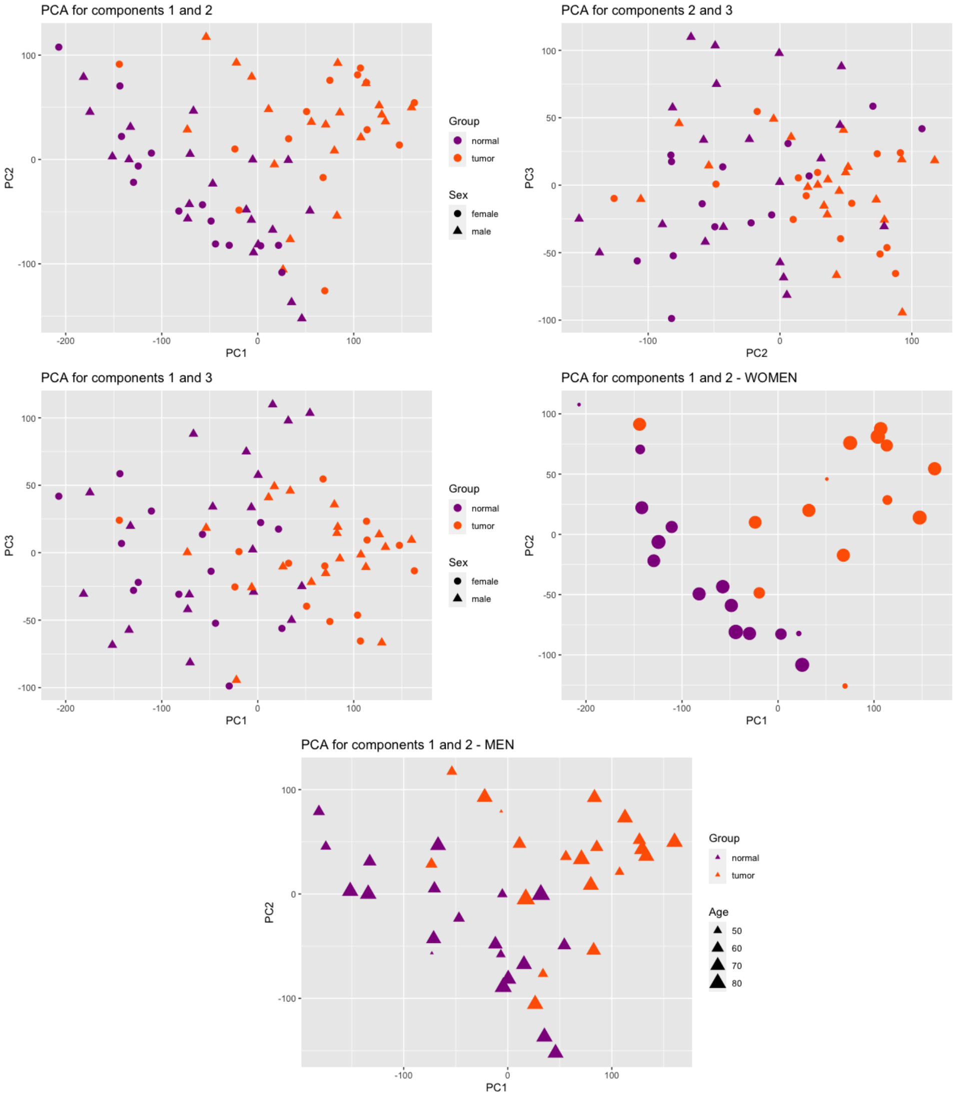


Figure S2: Scudo results

	accuracy	**AUC**
Kmeans	0.8529412	0.8642857
Hierarchical	0.7647059	0.7897727
LDA	0.9500000	0.9545455
Lasso	0.9500000	0.9545455
SCUDO	0.9000000	0.8800000
Random Forest	0.9705882	0.9800000

Table 1



PCA for components 1 and 2

PCA for components 2 and 3

PCA for components 1 and 3

PCA for components 1 and 2 - WOMEN

PCA for components 1 and 2 - MEN

Figure S3: PCA results

References

[1] Pelayo Correa. “Gastric Cancer: Overview”. In: *Gastroenterol Clin North Am.* (2013).

[2] Kitagawa Y. Takahashi T Saikawa Y. “Gastric cancer: current status of diagnosis and treatment”. In: *Cancers (Basel)* (2013).

[3] Yashiro M. Matsuoka T. “Biomarkers of gastric cancer: Current topics and future perspective”. In: *World J Gastroenterol.* (2018).

[4] Chen Wilks Zheng. “recount3: summaries and queries for large-scale RNA-seq expression and splicing”. In: *Genome Biol* (2021).

[5] van de Werken Smid Coebergh van den Braak. “Gene length corrected trimmed mean of M-values (GeTMM) processing of RNA-seq data”. In: *BMC Bioinformatics* (2018).

[6] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data”. In: *Bioinformatics* (2010).

[7] Hadley Wickham. “ggplot2: Elegant Graphics for Data Analysis”. In: (2016).

[8] Alboukadel Kassambara and Fabian Mundt. “factoextra: Extract and Visualize the Results of Multivariate Data Analyses”. In: (2020). R package version 1.0.7.

[9] Max Kuhn. “caret: Classification and Regression Training”. In: (2022). R package version 6.0-92.

[10] Andy Liaw and Matthew Wiener. “Classification and Regression by randomForest”. In: *R News* (2002).

[11] Robert Gentleman et al. “genefilter: genefilter: methods for filtering genes from high-throughput experiments”. In: (2021). R package version 1.76.0.

[12] Matteo Ciciani, Thomas Cantore, and Mario Lauria. “rScudo: an R package for classification of molecular profiles using rank-based signatures”. In: *Bioinformatics* (2019).

[13] Xavier Robin et al. “pROC: an open-source package for R and S+ to analyze and compare ROC curves”. In: *BMC Bioinformatics* (2011).

[14] Liis Kolberg et al. “gprofiler2– an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler”. In: *F1000Research* (2020). R package version 0.2.1.

[15] Ege Ulgen, Ozan Ozisik, and Osman U Sezerman. “pathfindR: An R Package for Comprehensive Identification of Enriched Pathways in Omics Data Through Active Subnetworks”. In: *Frontiers in Genetics* (2019).

[16] Ana Margarida et al Moreira. “The Extracellular Matrix: An Accomplice in Gastric Cancer Development and Progression”. In: *Cells* (2020).

[17] Qiang-Nu et al. Zhang. “A panel of collagen genes are associated with prognosis of patients with gastric cancer: an integrated bioinformatics analysis and experimental validation.” In: *Cancer management and research* (2019).