Gene expression analysis of lung adenocarcinoma and matched adjacent non-tumor lung tissue

Daniele Capelli - Student Number: 256212

Data Analysis and Exploration - Academic Year 2024-25

Abstract

Lung cancer is the main cause of cancer-related deaths in the world, and one of its most common subtypes is given by adenocarcinoma. In this report we analyze gene expression profiling array data taken from 59 matched lung adenocarcinoma and non-tumor lung samples. We look at supervised and unsupervised classification techniques to find the differences between these two groups, and we try to interpret the results from a biological point of view through functional analysis. Finally, we integrate methylation array data from the same samples and we perform a multi-omics analysis to compare the two layers.

Our results are compared to those reported in [3]. All code used in the analysis is publicly available at [1].

Contents

1 Introduction

Lung cancer is the leading cause of cancer-related death worldwide, and one of its most common subtypes is given by adenocarcinoma. This is increasingly recognized as a clinically and molecularly heterogeneous disease, and it is thought to be highly common both in smokers and never-smokers patients. Being able to recognize this tumor is becoming a challenging topic in medicine in order to develop and find targeted interventions that could benefit a patient in the best possible way.
In this document we will re-analyze the results got by Selamat et al., 2012 [3]. We will focus on 30 lung adenocarcinoma tumors and 29 non-tumor lung (NTL) tissues taken from 59 (ill) patients, with the aim of identifying the main differences between these two groups at the molecular level. To reach our goal, we will analyze two different types of data [2]:

- Expression profiling by array data, which were collected using the Illumina HumanWG-6 v3.0 expression beadchip (GPL6884).
- DNA Methylation data (relative to the same patients and the same tissues), which were collected using the Illumina HumanMethylation27 BeadChip (HumanMethylation27_270596_v.1.2) (GPL8490).

In order to take into consideration the highly heterogeneity linked to lung adenocarcinoma, we considered an highly heterogeneous population sample, made of smokers and non-smokers patients with different age, gender and ethnicity and that presented the tumors in different stages.

2 Expression Profiling by Array Data

We start our analysis by studying the expression profiling array data from the Gene Expression Omnibus (GEO), accession number GSE32867 [2].
The original dataset includes data on both adenocarcinoma tumors and matched adjacent non-tumor lung (NTL) tissues taken from 59 patients. In order to improve the computational complexity of our analysis and to reduce the correlation that would be naturally present when considering tissues from the same patients, we decide to partition the patients into two groups. The first one, made of 30 patients chosen at random, will contain the adenocarcinoma tissue data, while the second one, made of the 29 remaining patients, will include the NTL tissue data.

2.1 Exploratory Analysis

First of all we notice that the data are correctly stored by observing that we have a negligible number of missing data. Then we check the boxplots to see if the data are normally distributed: after removing the outliers, just to get a better visual representation, we obtain the results in Figure 1. We see that the results are pretty good, and so we do not need to apply a normalization procedure.
The next step in our analysis is to perform a Principal Component Analysis with the aim of reducing the dimensionality of the data: after looking at the variance explained by this technique (see Figure 2a), we decide to focus on the first two components. If we plot the data in correspondence of these two components we get the results in Figure 2b: the result is really good, as the two groups of tissues (the adenocarcinoma ones, A, and the healthy ones, C) are well separated.

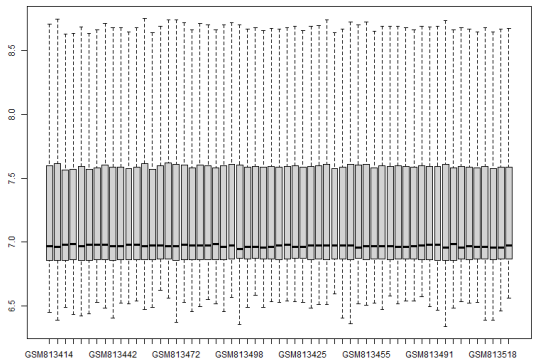
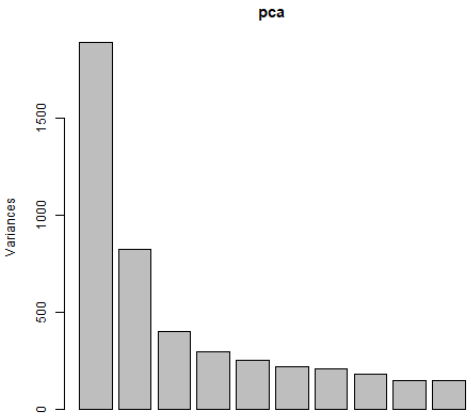
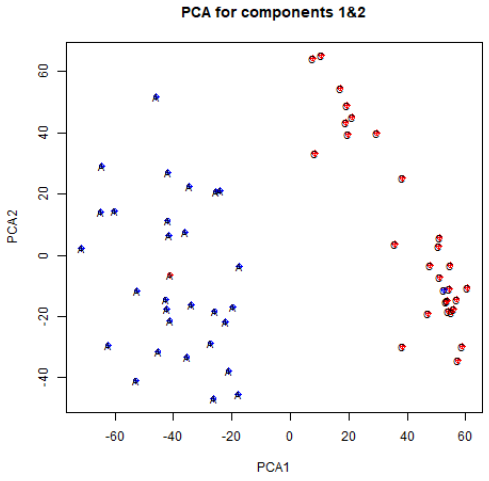


Figure 1: Boxplot of the expression profling by array data without outliers.



(a) Variance explained by PCA



(b) Data along the first two components

Figure 2: Principal Component Analysis results.

2.2 Clustering Techniques

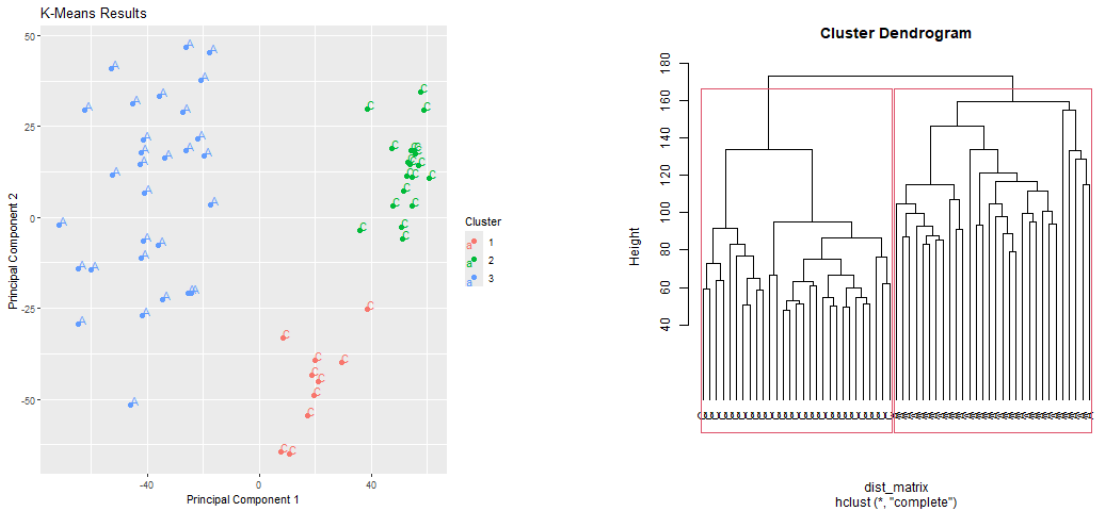
After this preliminary analysis, it is interesting to apply clustering techniques to our data to discover if there could be unknown groups behind them.

2.2.1 K-means clustering

We start by considering a k-means approach: as we are studying a division in two groups (tumor and non-tumor), it makes sense to test what happens when $k = 2$: with this choice we get a perfect division (see the Appendix).
However, by looking at Figure 2b it seems that we could divide our samples in three groups: with $k = 3$, we get the results in Figure 3a. We notice that one of the groups includes only adenocarcinoma tissue data and the other two contains the NTL tissues: this could be an interesting starting point for a new clinical trial. As a matter of fact, we remember that these healthy tissues are collected from patients that are affected by adenocarcinoma. So we could try to delve into this topic by studying if something effectively causes this division and if the outcome of the disease is somehow influenced by it.

2.2.2 Hierarchical Clustering

The second clustering technique that we want to test is given by the hierarchical clustering approach. We can try different combinations for the distance functions, but we decide to keep the complete linkage one. We get the results in Figure 3b: here we get a nice division into the two classes (adenocarcinoma and NTL tissue), but if we perform a three cluster division we do not get a nice result as in the k-means approach.



(a) 3-Means Clustering

(b) Hierarchical Clustering

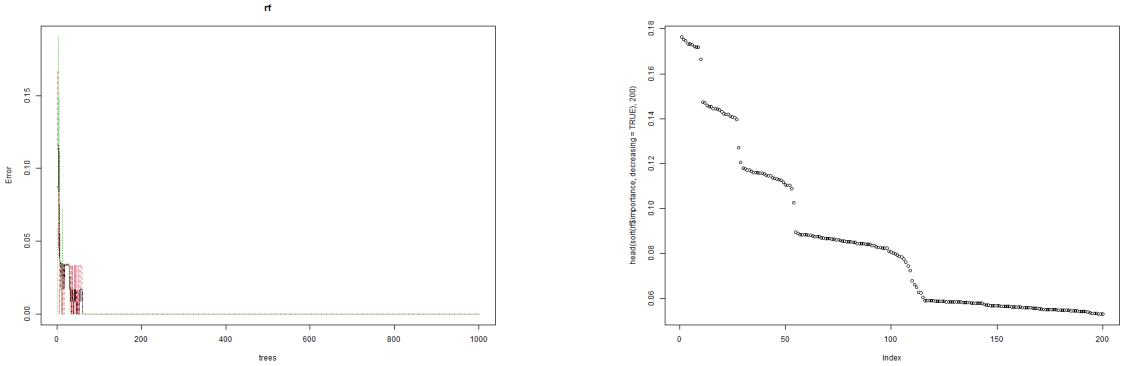
Figure 3: Clustering techniques applied to the data.

2.3 Supervised Learning Techniques

The next step of our analysis is to look at supervised learning classification methods to see if we can find a way to divide the healthy and non-healthy tissues: we will have a look at the Random Forests, LDA, LASSO and Scudo techniques.

2.3.1 Random Forest

The first supervised technique that we apply to our dataset is a Random Forest approach: we decide to construct 1000 trees that will be "merged" together into our forest. In Figure 4a we see how the error decreases as we add trees: we notice that the error drops really quickly and, in particular, after about 100 trees this quantity becomes negligible. This is a sign that the Random Forest fits really well our data, and it is confirmed by the fact that we get a perfect accuracy value of 1 (see Section 2.3.5).
In addition to its capability of capturing the differences between the two groups, the random forest approach is also useful because it outputs a measure of variable importance: we see the result for the 200 most important variables in Figure 4b. From this plot we can conclude that we have more or less 200 important variables in our data: these will be useful later on when we will do the functional analysis.



(a) Random Forest error

(b) Importance Plot of the first 200 variables

Figure 4: Random Forest results.

2.3.2 Linear Discriminant Analysis

The second approach we apply to our dataset is the Linear Discriminant Analysis (LDA): we fit the model with a 10-fold cross validation using the `caret` library in `R`. As in the random forest approach we get a perfect classification: as a matter of fact, both the AUC and the accuracy values are 1. This perfect classification is confirmed by the LDA projection plot (see the Appendix).

2.3.3 LASSO

The third approach we apply to our dataset is the LASSO regression: after performing a 10-fold cross-validation to tune the parameters, we select 0.2 as the regularization one (see Figure 5). Studying the results we notice that we get a perfect accuracy value of 1, and so again we got a model that is perfectly able to capture the differences in the two groups.

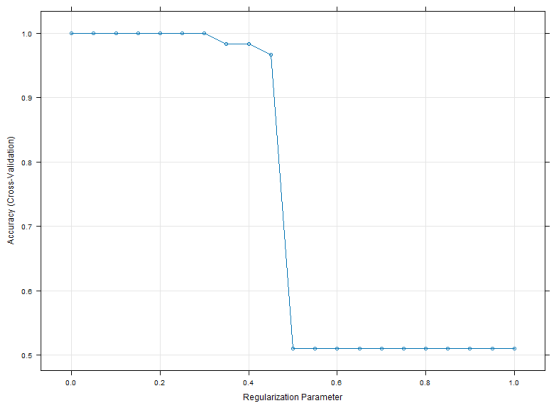
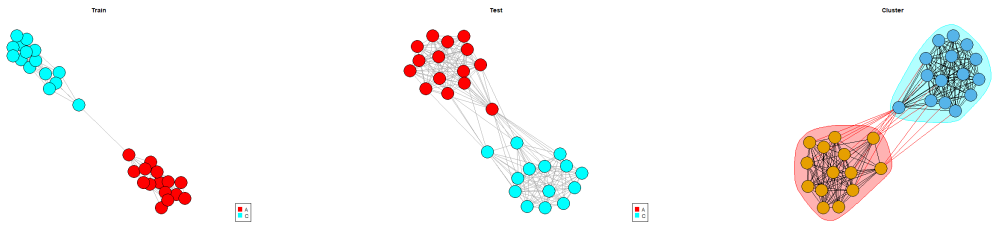


Figure 5: Possible regularization parameters for the LASSO method.

2.3.4 Scudo

The final supervised model we fit to classify our data is the SCUDO model, a rank-based method useful for getting robust results that mitigates the effect of batching.
After dividing randomly our dataset into a training and a test set, we try different values for the number of keys to take into consideration. We decide to keep 250 keys for both the upper and the lower signatures as this combination gives the best results: we can see the network that comes from the training data in Figure 6a and the network that comes from the validation data in Figure 6b. If we try to cluster this last graph into two subgraphs we get the results in Figure 6c: we notice that we get a perfect classification. So we can conclude that also with the SCUDO procedure we get a perfect classification method.



(a) Scudo Network from the training data

(b) Scudo Network from the validation data

(c) Clustering of the validation data

Figure 6: Scudo results.

2.3.5 Comparison between the results

We can finally compare the classification techniques we applied to our data by looking at their accuracy scores, which are summarized in Figure 7. It is interesting to observe that all the methods led to a perfect accuracy score of 1: this means that all the models are able to capture the differences between the adenocarcinoma tissue and the NTL one.
In the next part of our analysis we will try to extract biological information from these results: we decide to focus only on the Random Forest ones because, as stated before, this approach is able to provide a variable importance selection in the genes we are studying.

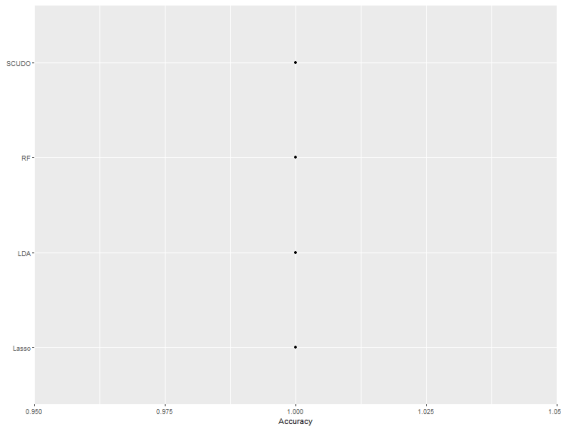


Figure 7: Comparison of the accuracy scores of the supervised methods.

2.4 Functional Analysis

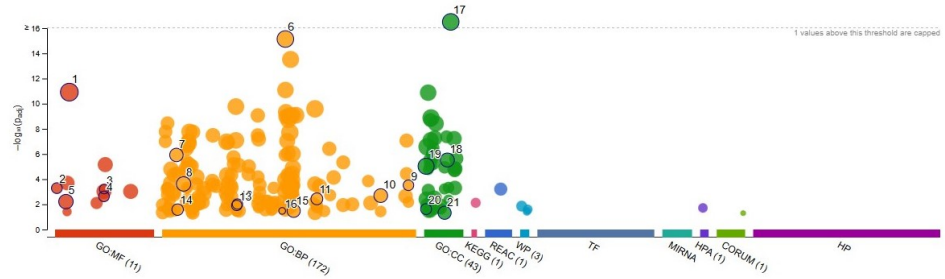
Our aim is now to gain deeper biological information on the adenocarcinoma tissue from the results we got in the previous part. We can do this by following the functional enrichment analysis approach. First, we will have a look at the results got with standard instruments, such as `DAVID` and `g:Profiler`. Then we will try to enrich our conclusions by including pre-existing knowledge with the network-based analysis, and we will make use of the `STRING` and `EnrichNet` tools.

2.4.1 g:Profiler

We perform a functional enrichment analysis on our data using the `g:Profiler` platform: after trying different number of genes, we decide to analyze what happens if we keep the 400 most important ones that we inferred using the Random Forest approach. In Figure 8 we can see the most relevant pathways detected by this tool: we notice that the first two, "cell periphery" and "positive regulation of biological processes" (and also the third one, "protein binding") present a low p-value, which is a sign that they are extremely significant in the process we are studying.
These results should be compared with the pathways identified by the DAVID platform (which are included in the Appendix). However, as with this tool we did not get significant results, we decide to test a different approach, the network-based analysis, which is able to include previous knowledge by adding a network structure to the proteins. We will perform this with the `STRING` and the `EnrichNet` online platforms.

2.4.2 STRING

We can study protein-to-protein interactions by using the `STRING` platform: first of all, we can study what happens if we use again the 400 most important genes detected by the Random Forest method, without adding interactors. The `STRING` tool detects the following most relevant pathways: "positive regulation of biological process", "cell periphery", "plasma membrane", and "whole body" tissue expression, which are in line with the result got with the `g:Profiler` platform.
As we are interested in gaining more biological information, we decide to try different settings: we increase the number of interactors in the first shell, and meanwhile we decrease



ID	Source	Term ID	Term Name	p_{adj} (query_1)
17	GO:CC	GO:0071944	cell periphery	7.363×10^{-18}
6	GO:BP	GO:0048518	positive regulation of biological process	7.270×10^{-16}
1	GO:MF	GO:0005515	protein binding	1.208×10^{-11}
7	GO:BP	GO:0003013	circulatory system process	1.189×10^{-6}
18	GO:CC	GO:0070161	anchoring junction	2.998×10^{-6}
19	GO:CC	GO:0005576	extracellular region	9.363×10^{-6}
8	GO:BP	GO:0006468	protein phosphorylation	2.389×10^{-4}
9	GO:BP	GO:2000379	positive regulation of reactive oxygen species met...	3.050×10^{-4}
2	GO:MF	GO:0001540	amyloid-beta binding	5.159×10^{-4}
3	GO:MF	GO:0038187	pattern recognition receptor activity	6.103×10^{-4}
10	GO:BP	GO:1903047	mitotic cell cycle process	2.023×10^{-3}
4	GO:MF	GO:0038024	cargo receptor activity	2.325×10^{-3}
11	GO:BP	GO:0070374	positive regulation of ERK1 and ERK2 cascade	3.654×10^{-3}
5	GO:MF	GO:0004888	transmembrane signaling receptor activity	6.077×10^{-3}
12	GO:BP	GO:0032731	positive regulation of interleukin-1 beta production	9.459×10^{-3}
13	GO:BP	GO:0032655	regulation of interleukin-12 production	1.216×10^{-2}
20	GO:CC	GO:0005581	collagen trimer	2.396×10^{-2}
14	GO:BP	GO:0003158	endothelium development	2.636×10^{-2}
15	GO:BP	GO:0051338	regulation of transferase activity	3.262×10^{-2}
16	GO:BP	GO:0048014	Tie signaling pathway	3.323×10^{-2}
21	GO:CC	GO:0045177	apical part of cell	4.698×10^{-2}

Figure 8: Functional Enrichment Analysis performed using the `g:Profiler` online platform on the 400 most important genes from the Random Forest classification method. The platform was able to recognize 347 genes.

the number of genes that we take into consideration in order to reduce the complexity of the network. After several trials, the best balance between biological relevance and network interpretability was reached when we used the 50 most important genes detected by the random forest and a maximum of 20 interactors in the first shell. In particular, our analysis highlights that the following pathways are extremely significant for our problem: "PDGF/VEGF domain, and Vascular Endothelial Growth Factor", "Constitutive Signaling by Aberrant PI3K in Cancer", "VEGF ligand–receptor interactions, and Tie signaling", "Leukocyte transendothelial migration", "Signaling by VEGF", "VEGFA–VEGFR2 signaling pathway", "VEGF signaling pathway" and "Hippo signaling regulation pathways". These pathways are generally related to angiogenesis and blood vessel formation: this is a sign that vascular-driven tumor growth plays a key role in lung adenocarcinoma. Full details on the pathways detected by the `STRING` tool are reported in the Appendix.

2.4.3 EnrichNet

At last we can have a look at the results from the network-based analysis performed using the `EnrichNet` platform: after trying different settings, we decide to study what happens when we keep the 200 most important genes that come from the Random Forest method, enriched using the Gene Ontology Biological Process pathways. Running the tool we get a (relatively) high number of significant pathways, which correspond to pathways with an high XD-score and a small Fisher-q value: to reduce them we check if they have an high tissue-specificity value in correspondence of tissues that are related to lung adenocarcinoma, e.g. lung, tumor and similar. With this requirement we get that the most significant pathways are the following ones: "vasoconstriction", "mammary gland involution", "neg-

ative regulation of cytokine-mediated signaling pathway" and "vesicle organization". We notice that we got different pathways with respect to the previous approaches. In addition, these pathways are not conventionally linked to tumor growth: this could be an interesting starting point for further investigations and new therapies approaches.

3 Multi-Omics Analysis

The final step of our analysis will be to combine and integrate through a multi-omics analysis information from DNA methylation data (that can be retrieved again from the Gene Expression Omnibus GEO, accession number GSE32867 [2]) to the results we got using the expression profiling array data. We will focus our attention only on the two groups of patients that we introduced in Section 2 as we need an overlapping in the samples.
We combined results from the two datasets using two approaches: the Canonical Correlation Analysis (CCA) and the Partial Least Squares Analysis (PLS). In the following section we will discuss results from the last method as it is more robust to high dimensional data, but results from CCA can be retrieved in the Appendix.

3.1 Partial Least Squares Analysis

We proceed by applying a Partial Least Squares approach to compare the expression profiling by array and the DNA methylation data. The Variables Plot that comes from this method is shown in Figure 9a: in particular, expression profiling array variables are in blue, and the methylation ones are in orange. We know from the theory behind the PLS method that variables in two different groups that are on the opposite sides of the unit circle are positively correlated, while variables from the two groups that are on the same side of the circle are negatively correlated. First of all, to gain interpretable results, we link the methylation sites to their correspondent gene: however, as we know that it is important to have the precise sites for a detailed result, we report them in the Appendix. Then we group the negatively correlated genes in Table 1 and the positively correlated ones in Table 2 (the specific function of each gene can be retrieved in the Appendix). These tables could be used as an investigative basis for further studies: in particular, it could be useful to look at the methylation sites correlated to the expression of the MELK, EPB41L3 and TCF21 genes (which are related to tumors and cancers) and to the expression of genes correlated to the methylation sites of the RB1, CDKN1C and IGF2 genes, which are again cancer-linked.
Finally we will have a look at the arrow plot, which is shown in Figure 9b: we see that the arrows that connect the two data (expression profiling by array and DNA methylation) relative to the same patient are relatively long. This is a sign that the two omics that we are investigating are not correlated, and so we should take into consideration other omics to deepen our analysis. However we notice that the two groups of tissues that we are taking into consideration (adenocarcinoma and NTL) are well separated along the PLS components, so the combination of these two omics is useful to study a possible classification method.

Expression Genes	**Gene linked to the correlated methylation site**
S1PR1, STX11, TCF21, PEBP4, TM6SF1, CPA3, TCF21, LINC02538, FABP4, None	RPTOR, PAX8, CDKN1C, IGF2

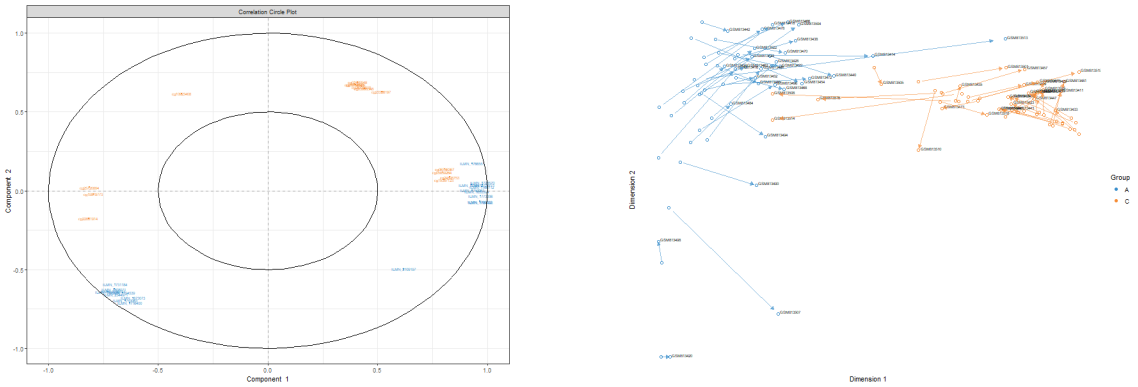
Table 1: Gene expression and methylation sites variables negatively correlated from the sPLS technique (gene symbols)

Expression Genes	**Gene linked to the correlated methylation site**
CENPF, WDR62, TROAP, FOXM1, MELK, KIF15, None, KIF14, EPB41L3, FOXM1	IFNA17, KCNK16, CYP2E1, TCF7L2
S1PR1, STX11, TCF21, PEBP4, TM6SF1, CPA3, TCF21, LINC02538, FABP4, None	DLGAP2, RB1, GATA4
CENPF, WDR62, TROAP, FOXM1, MELK, KIF15, None, KIF14, EPB41L3, FOXM1	GABRA2, APOE

Table 2: Gene expression and methylation sites variables positively correlated from the sPLS technique (gene symbols)

4 Conclusions

In this document we performed a re-analyses of data collected by Selamat et al. (2012, [3]): we showed that unsupervised techniques (such as PCA and clustering) are able to distinguish adenocarcinoma tumor samples from the non-tumor ones. Supervised learning approaches (Random Forest, LDA, LASSO and SCUDO) showed a perfect ability in classifying these groups and all achieved a 100% accuracy when tested. A functional enrichment analysis performed on the most significant genes detected by the random forest technique suggested that the most relevant pathways linked to this cancer are "cell periphery", "positive regulation of biological processes" and "protein binding", which are coherent with the results in [3]. As a matter of fact, in Selamat et al. the identified significant pathways are the ones related to cell differentiation, epithelial to mesenchymal transition, RAS and WNT signaling pathways and cell cycle regulation. With respect to the original article, our analysis showed a possible connection of adenocarcinoma tumor growth to angiogenesis and blood vessel formation through the network-based functional analysis, which could be an interesting and unexplored new point of view. In addition, a comparison between the two omics (gene expression and DNA methylation) showed possible correlations between genes identified by the two platforms.
Interestingly, as suggested by the analysis of the k-means clustering results when $k = 3$,



(a) Variables Plot from the PLS analysis.

(b) Arrow Plot from the PLS analysis

Figure 9: Results from the PLS analysis on the two omics.

there is a possibility that inside the adenocarcinoma cancer there are two different subgroups. Selamat et al. delved into this topic, trying to understand if this division is linked to the smoking status of the patient: even if this hypothesis seems reasonable, their conclusion was that it should be investigated by further studies.
As stated at the beginning of this document, lung adenocarcinoma is one of the most widespread form of cancers. In our analysis we were able to identify genes whose mutation is highly linked to the onset of this disease, and we validated and extended the results obtained by Selamat et al. In addition, our conclusions may serve as a starting point for further and more focused studies, and they could be helpful in developing therapeutic interventions.

References

[1] Daniele Capelli. *Data Analysis and Exploration Code*. `https://github.com/dcapelli02/Data-Analysis-and-Exploration`. Accessed: July 15, 2025. 2025.

[2] Ite A. Laird-Offringa and collaborators. *DNA methylation and gene expression in lung adenocarcinoma*. `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32867`. NCBI Gene Expression Omnibus, accession GSE32867. 2012.

[3] Suhaida A. Selamat et al. "Genome-scale analysis of DNA methylation in lung adenocarcinoma and integration with mRNA expression". In: *Genome Research* 22.7 (2012), pp. 1197–1211. DOI: `10.1101/gr.132662.111`.

R Packages Version

We used `R` version 4.4.3 (2025-02-28 ucrt) and the following packages to derive our analysis.

- `GEOquery`: version 2.74.0
- `useful`: version 1.2.6.1
- `randomForest`: version 4.7.1.2
- `geneFilter`: version 1.88.0

- `caret`: version 7.0.1
- `pRoc`: version 1.18.5
- `rScudo`: version 1.22.0
- `igraph`: version 2.1.4
- `ggplot2`: version 3.5.2
- `mixOmics`: version 6.30.0

A Additional Results

A.1 K-means clustering

In Figure 10 we plot the results of a 2-means clustering to our dataset. Our patients are well separated by the model: this is a sign that the two PCA components are good summaries of our variables, and even without knowing the labels a priori we would obtain a good classification.

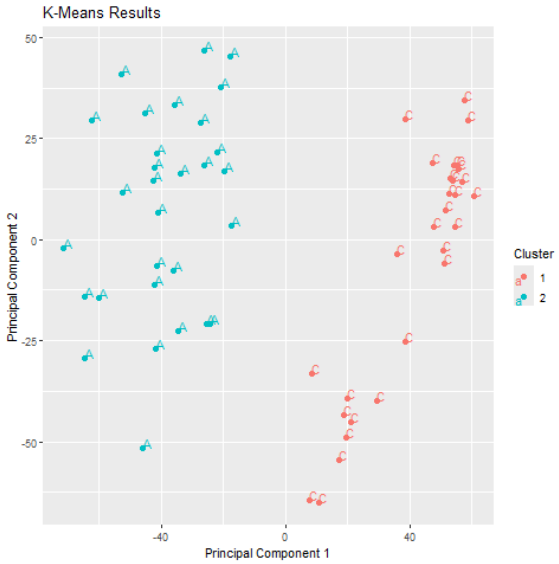
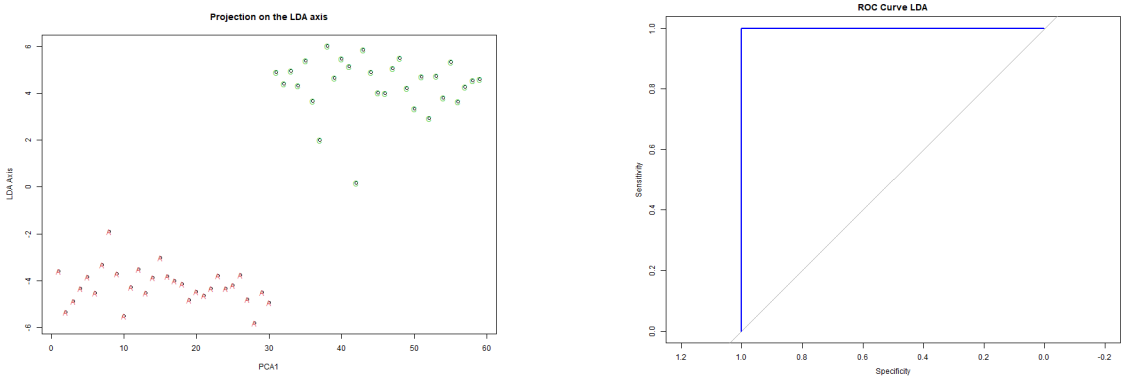


Figure 10: 2-Means Clustering of the data

A.2 LDA

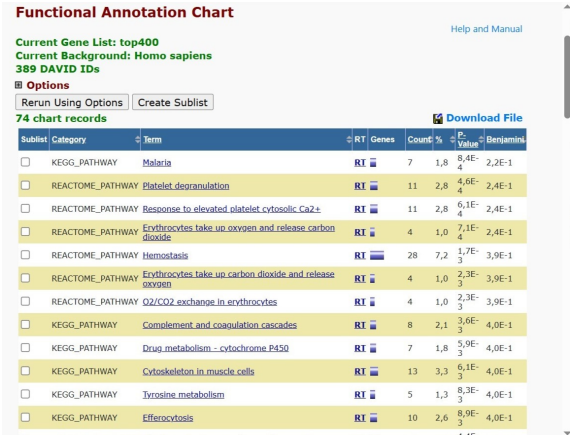
In Figure 11a we can see the LDA projection plot: the two groups are perfectly separated and they do not overlap. In Figure 11b we find the ROC curve for this method, which is the ideal one.



(a) LDA projection axis

(b) ROC curve for the LDA method

Figure 11: Results for the LDA method.



Functional Annotation Chart

Help and Manual

Current Gene List: top400
Current Background: Homo sapiens
389 DAVID IDs

Options

Rerun Using Options | Create Sublist

74 chart records

Download File

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	KEGG_PATHWAY	Malaria	RT		7	1,8	8,4E-4	2,2E-1
	REACTOME_PATHWAY	Platelet degranulation	RT		11	2,8	4,6E-4	2,4E-1
	REACTOME_PATHWAY	Response to elevated platelet cytosolic Ca2+	RT		11	2,8	6,1E-4	2,4E-1
	REACTOME_PATHWAY	Erythrocytes take up oxygen and release carbon dioxide	RT		4	1,0	7,1E-4	2,4E-1
	REACTOME_PATHWAY	Hemostasis	RT		28	7,2	1,7E-3	3,9E-1
	REACTOME_PATHWAY	Erythrocytes take up carbon dioxide and release oxygen	RT		4	1,0	2,3E-3	3,9E-1
	REACTOME_PATHWAY	O2/CO2 exchange in erythrocytes	RT		4	1,0	2,3E-3	3,9E-1
	KEGG_PATHWAY	Complement and coagulation cascades	RT		8	2,1	3,6E-3	4,0E-1
	KEGG_PATHWAY	Drug metabolism - cytochrome P450	RT		7	1,8	5,9E-3	4,0E-1
	KEGG_PATHWAY	Cytoskeleton in muscle cells	RT		13	3,3	6,1E-3	4,0E-1
	KEGG_PATHWAY	Tyrosine metabolism	RT		5	1,3	8,3E-3	4,0E-1
	KEGG_PATHWAY	Efferocytosis	RT		10	2,6	8,9E-3	4,0E-1

Figure 12: Functional Enrichment Analysis performed using the `DAVID` online platform on the 400 most important genes from the Random Forest classification method.

A.3 DAVID Functional Enrichment Analysis

In Figure 12 we see the results of a functional enrichment analysis performed using the `DAVID` online platform on the 400 most important genes we got from the Random Forest model: we notice that we get a small Benjamini p-value, and so this results cannot be considered particularly significant.

A.4 STRING

In Figures 13 - 19 we see additional results from the `STRING` network-based analyses.

A.5 EnrichNet

We report in Figure 20 the most important pathways identified by the `EnrichNet` platform, in Figure 21 the XD-scores versus Fisher q-values plot and in Figure 22 an example of a tissue specificity matrix.

A.6 PLS

In Table 3 and 4 we report the positively and negatively correlated variables detected by the PLS technique in terms of IDs identified by the platforms. In Table 5 we report the interpretation of the gene codes expressed in Tables 1 and 2.

A.7 CCA

We show in this section results from the CCA method applied to our datasets: the Variables Plot is shown in Figure 23a while the Arrow Plot in Figure 23b. We notice that we get better results in terms of the arrows length with respect to the PLS approach: however we decided to study this last approach because of its robustness to high dimensional data.

Figure 13: Visual representation of the network coming from the network-based analysis of the 400 most important genes from the Random Forest method performed with the `STRING` online platform.

Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0048518	Positive regulation of biological process	100 of 6207	0.24	0.61	7.01e-07
GO:0048856	Anatomical structure development	85 of 5117	0.26	0.58	1.01e-05
GO:0048522	Positive regulation of cellular process	87 of 5584	0.23	0.51	9.63e-05
GO:0032502	Developmental process	87 of 5657	0.22	0.49	0.00014
GO:0051716	Cellular response to stimulus	94 of 6357	0.21	0.48	0.00014
					(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0005515	Protein binding	98 of 7242	0.17	0.34	0.0080
GO:0038024	Cargo receptor activity	7 of 82	0.97	0.36	0.0418
GO:0001540	Amyloid-beta binding	7 of 84	0.96	0.36	0.0418

Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0071944	Cell periphery	97 of 6015	0.24	0.63	2.49e-07
GO:0005886	Plasma membrane	87 of 5544	0.23	0.55	1.31e-05
GO:0031226	Intrinsic component of plasma membrane	41 of 1788	0.4	0.65	3.07e-05
GO:0070821	Tertiary granule membrane	8 of 73	1.08	0.86	0.00032
GO:0070820	Tertiary granule	11 of 164	0.86	0.79	0.00032
					(more ...)

Reference Publications (PubMed)

publication	(year) title	count in network	strength	signal	false discovery rate
PMID:34030460	(2021) Integrated Single-Cell Atlas of Endothelial Cells of the Huma...	26 of 225	1.1	2.35	1.99e-13
PMID:36067196	(2022) A feature selection-based framework to identify biomarkers ...	15 of 36	1.66	3.0	8.31e-13
PMID:31681566	(2019) Integrated Network Analysis Reveals FOXM1 and MYBL2 as...	16 of 108	1.21	1.68	4.48e-08
PMID:36101460	(2022) Identifying General Tumor and Specific Lung Cancer Biomar...	15 of 104	1.2	1.51	3.16e-07
PMID:37414529	(2023) Differential roles of FOXC2 in the trabecular meshwork and ...	13 of 73	1.29	1.46	9.53e-07
					(more ...)

Local Network Cluster (STRING)

cluster	description	count in network	strength	signal	false discovery rate
CL:17382	VEGF ligand-receptor interactions, and Tie signaling pathway	6 of 15	1.64	1.06	0.00012
CL:17330	Constitutive Signaling by Aberrant PI3K in Cancer, and VEGF ligand-...	7 of 61	1.1	0.61	0.0043
CL:6608	Mixed, incl. Regulation of mitotic sister chromatid segregation, and...	6 of 51	1.11	0.52	0.0099
CL:10427	Leukotriene biosynthesis, and Cysteinyl leukotriene receptor	3 of 6	1.74	0.41	0.0342
CL:15007	Cavin family, and Caveolar macromolecular signaling complex	3 of 6	1.74	0.41	0.0342

Figure 14: Patwhays identified by the `STRING` platform in the setting of Figure 13

.

Reactome Pathways

pathway	description	count in network	strength	signal	false discovery rate
HSA-202733	Cell surface interactions at the vascular wall	9 of 139	0.85	0.43	0.0199

WikiPathways

pathway	description	count in network	strength	signal	false discovery rate
WP5171	Leukotriene metabolic pathway	4 of 13	1.52	0.53	0.0113

Tissue Expression (TISSUES)

tissue	description	count in network	strength	signal	false discovery rate
BTO:0001489	Whole body	156 of 13099	0.11	0.49	1.78e-06
BTO:0000042	Animal	170 of 15148	0.09	0.47	1.78e-06
BTO:0000763	Lung	33 of 1395	0.41	0.57	0.00038
BTO:0000141	Bone marrow	18 of 528	0.57	0.58	0.0013
BTO:0003718	Vasculature	4 of 9	1.68	0.75	0.0018

(more ...)

Subcellular Localization (COMPARTMENTS)

compartment	description	count in network	strength	signal	false discovery rate
GOCC:0005886	Plasma membrane	66 of 3535	0.31	0.63	6.68e-06
GOCC:0071944	Cell periphery	70 of 3860	0.3	0.62	6.68e-06
GOCC:0016020	Membrane	86 of 5715	0.21	0.5	0.00010
GOCC:0045121	Membrane raft	12 of 181	0.86	0.86	0.00012
GOCC:0031226	Intrinsic component of plasma membrane	25 of 851	0.5	0.66	0.00016

(more ...)

Annotated Keywords (UniProt)

keyword	description	count in network	strength	signal	false discovery rate
KW-1015	Disulfide bond	62 of 3338	0.31	0.61	1.09e-05
KW-0325	Glycoprotein	72 of 4386	0.25	0.55	3.55e-05
KW-0037	Angiogenesis	10 of 131	0.92	0.87	0.00015
KW-0732	Signal	56 of 3277	0.27	0.5	0.00029
KW-1003	Cell membrane	56 of 3277	0.27	0.5	0.00029

(more ...)

Figure 15: Patwhays identified by the STRING platform in the setting of Figure 13 (continuation).

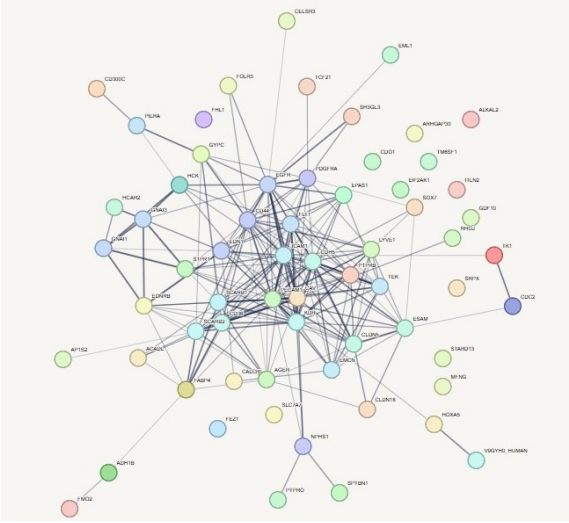
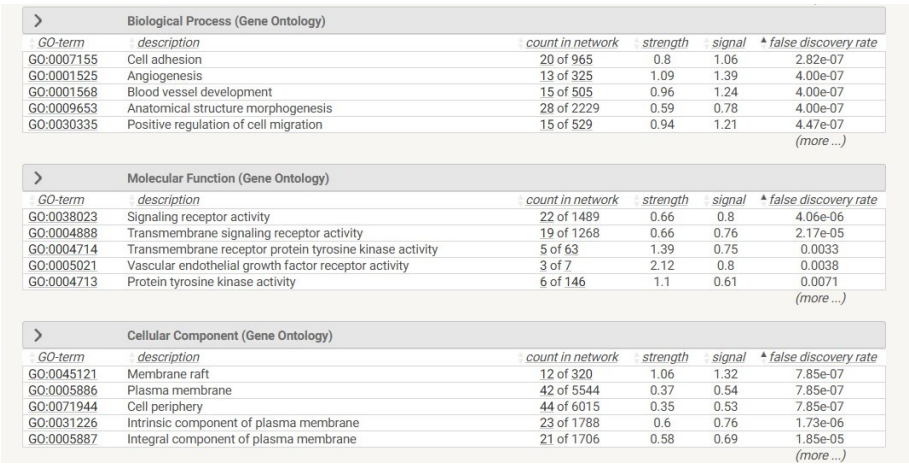


Figure 16: Visual representation of the network coming from the network-based analysis of the 50 most important genes coming from the Random Forest method and 20 interactors performed with the STRING online platform.



Biological Process (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0007155	Cell adhesion	20 of 965	0.8	1.06	2.82e-07
GO:0001525	Angiogenesis	13 of 325	1.09	1.39	4.00e-07
GO:0001568	Blood vessel development	15 of 505	0.96	1.24	4.00e-07
GO:0009653	Anatomical structure morphogenesis	28 of 2229	0.59	0.78	4.00e-07
GO:0030335	Positive regulation of cell migration	15 of 529	0.94	1.21	4.47e-07

(more ...)

Molecular Function (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0038023	Signaling receptor activity	22 of 1489	0.66	0.8	4.06e-06
GO:0004888	Transmembrane signaling receptor activity	19 of 1268	0.66	0.76	2.17e-05
GO:0004714	Transmembrane receptor protein tyrosine kinase activity	5 of 63	1.39	0.75	0.0033
GO:0005021	Vascular endothelial growth factor receptor activity	3 of 7	2.12	0.8	0.0038
GO:0004713	Protein tyrosine kinase activity	6 of 146	1.1	0.61	0.0071

(more ...)

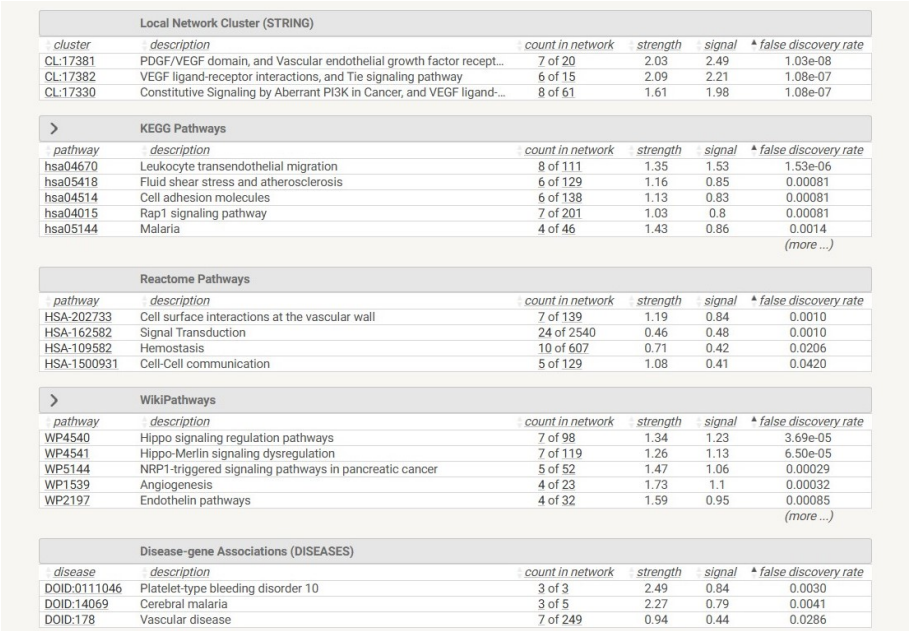
Cellular Component (Gene Ontology)

GO-term	description	count in network	strength	signal	false discovery rate
GO:0045121	Membrane raft	12 of 320	1.06	1.32	7.85e-07
GO:0005886	Plasma membrane	42 of 5544	0.37	0.54	7.85e-07
GO:0071944	Cell periphery	44 of 6015	0.35	0.53	7.85e-07
GO:0031226	Intrinsic component of plasma membrane	23 of 1788	0.6	0.76	1.73e-06
GO:0005887	Integral component of plasma membrane	21 of 1706	0.58	0.69	1.85e-05

(more ...)

Figure 17: Patwhays identified by the `STRING` platform in the setting of Figure 16

.



Local Network Cluster (STRING)

cluster	description	count in network	strength	signal	false discovery rate
CL:17381	PDGF/VEGF domain, and Vascular endothelial growth factor recept...	7 of 20	2.03	2.49	1.03e-08
CL:17382	VEGF ligand-receptor interactions, and Tie signaling pathway	6 of 15	2.09	2.21	1.08e-07
CL:17330	Constitutive Signaling by Aberrant PI3K in Cancer, and VEGF ligand-...	8 of 61	1.61	1.98	1.08e-07

KEGG Pathways

pathway	description	count in network	strength	signal	false discovery rate
hsa04670	Leukocyte transendothelial migration	8 of 111	1.35	1.53	1.53e-06
hsa05418	Fluid shear stress and atherosclerosis	6 of 129	1.16	0.85	0.00081
hsa04514	Cell adhesion molecules	6 of 138	1.13	0.83	0.00081
hsa04015	Rap1 signaling pathway	7 of 201	1.03	0.8	0.00081
hsa05144	Malaria	4 of 46	1.43	0.86	0.0014

(more ...)

Reactome Pathways

pathway	description	count in network	strength	signal	false discovery rate
HSA-202733	Cell surface interactions at the vascular wall	7 of 139	1.19	0.84	0.0010
HSA-162582	Signal Transduction	24 of 2540	0.46	0.48	0.0010
HSA-109582	Hemostasis	10 of 607	0.71	0.42	0.0206
HSA-1500931	Cell-Cell communication	5 of 129	1.08	0.41	0.0420

WikiPathways

pathway	description	count in network	strength	signal	false discovery rate
WP4540	Hippo signaling regulation pathways	7 of 98	1.34	1.23	3.69e-05
WP4541	Hippo-Merlin signaling dysregulation	7 of 119	1.26	1.13	6.50e-05
WP5144	NRP1-triggered signaling pathways in pancreatic cancer	5 of 52	1.47	1.06	0.00029
WP1539	Angiogenesis	4 of 23	1.73	1.1	0.00032
WP2197	Endothelin pathways	4 of 32	1.59	0.95	0.00085

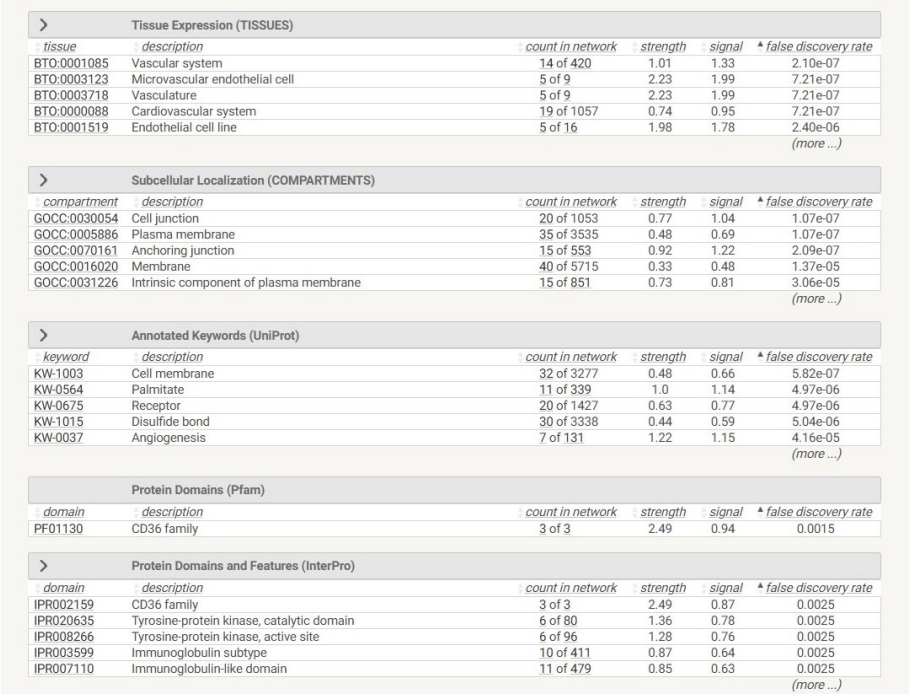
(more ...)

Disease-gene Associations (DISEASES)

disease	description	count in network	strength	signal	false discovery rate
DOID:0111046	Platelet-type bleeding disorder 10	3 of 3	2.49	0.84	0.0030
DOID:14069	Cerebral malaria	3 of 5	2.27	0.79	0.0041
DOID:178	Vascular disease	7 of 249	0.94	0.44	0.0286

Figure 18: Patwhays identified by the `STRING` platform in the setting of Figure 16 (continuation)

.



Tissue Expression (TISSUES)

tissue	description	count in network	strength	signal	false discovery rate
BTO:0001085	Vascular system	14 of 420	1.01	1.33	2.10e-07
BTO:0003123	Microvascular endothelial cell	5 of 9	2.23	1.99	7.21e-07
BTO:0003718	Vasculature	5 of 9	2.23	1.99	7.21e-07
BTO:0000088	Cardiovascular system	19 of 1057	0.74	0.95	7.21e-07
BTO:0001519	Endothelial cell line	5 of 16	1.98	1.78	2.40e-06

(more ...)

Subcellular Localization (COMPARTMENTS)

compartment	description	count in network	strength	signal	false discovery rate
GOCC:0030054	Cell junction	20 of 1053	0.77	1.04	1.07e-07
GOCC:0005886	Plasma membrane	35 of 3535	0.48	0.69	1.07e-07
GOCC:0070161	Anchoring junction	15 of 553	0.92	1.22	2.09e-07
GOCC:0016020	Membrane	40 of 5715	0.33	0.48	1.37e-05
GOCC:0031226	Intrinsic component of plasma membrane	15 of 851	0.73	0.81	3.06e-05

(more ...)

Annotated Keywords (UniProt)

keyword	description	count in network	strength	signal	false discovery rate
KW-1003	Cell membrane	32 of 3277	0.48	0.66	5.82e-07
KW-0564	Palmitate	11 of 339	1.0	1.14	4.97e-06
KW-0675	Receptor	20 of 1427	0.63	0.77	4.97e-06
KW-1015	Disulfide bond	30 of 3338	0.44	0.59	5.04e-06
KW-0037	Angiogenesis	7 of 131	1.22	1.15	4.16e-05

(more ...)

Protein Domains (Pfam)

domain	description	count in network	strength	signal	false discovery rate
PF01130	CD36 family	3 of 3	2.49	0.94	0.0015

Protein Domains and Features (InterPro)

domain	description	count in network	strength	signal	false discovery rate
IPR002159	CD36 family	3 of 3	2.49	0.87	0.0025
IPR020635	Tyrosine-protein kinase, catalytic domain	6 of 80	1.36	0.78	0.0025
IPR008266	Tyrosine-protein kinase, active site	6 of 96	1.28	0.76	0.0025
IPR003599	Immunoglobulin subtype	10 of 411	0.87	0.64	0.0025
IPR007110	Immunoglobulin-like domain	11 of 479	0.85	0.63	0.0025

(more ...)

Figure 19: Patwhays identified by the `STRING` platform in the setting of Figure 16 (continuation)

.

Annotation (pathway/process)	Significance of network distance distribution (XD-Score)	Significance of overlap (Fisher-test, q-value)	Dataset size (uploaded gene set)	Dataset size (pathway gene set)	Dataset size (overlap)	Tissue-specific XD-scores
vasoconstriction compute graph visualization see mapped genes	1.958*	0.21	150	13	3 (show)	show tissue specificity
mammary gland involution compute graph visualization see mapped genes	1.682*	0.45	150	10	2 (show)	show tissue specificity
negative regulation of cytokine-mediated signaling pathway compute graph visualization see mapped genes	1.682*	0.45	150	10	2 (show)	show tissue specificity
morphogenesis of a branching structure compute graph visualization see mapped genes	1.518*	0.45	150	11	2 (show)	show tissue specificity
vesicle organization compute graph visualization see mapped genes	1.518*	0.45	150	11	2 (show)	show tissue specificity
heart trabecula formation compute graph visualization see mapped genes	1.518*	0.45	150	11	2 (show)	show tissue specificity

Figure 20: List of the first pathways identified by the `EnrichNet` platform in the settings of Section 2.4.3.

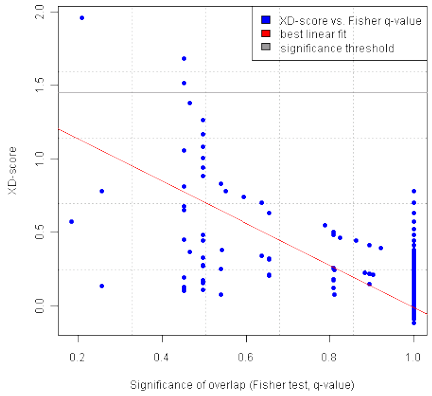


Figure 21: Plot of the XD scores combined with the Fisher q-value that comes from the `EnrichNet` analysis performed in Section 2.4.3.

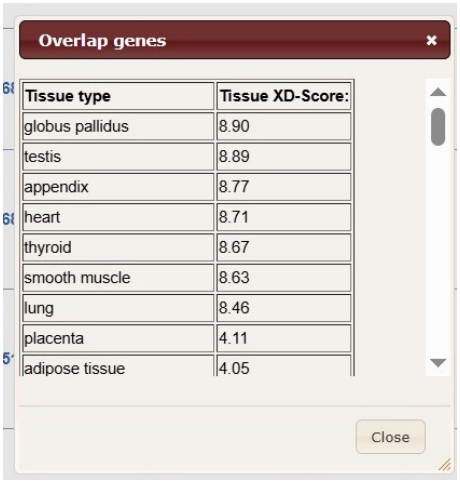
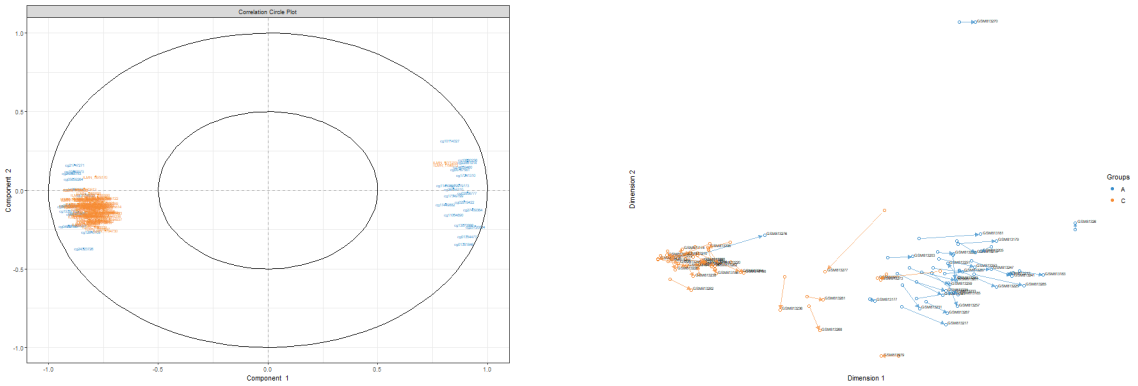


Figure 22: Example of a Tissue Specificity matrix (in this particular case the tissue specificity matrix of the Mammary Gland Involution pathway).



(a) Variables Plot from the CCA analysis.

(b) Arrow Plot from the CCA analysis

Figure 23: Results from the CCA analysis on the two omics.

Expression Profiling by Array	**Methylation**
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg01074640
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg01669948
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg03998348
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971, ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg10190509
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg19979773
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg22881914
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg23326197
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg25720804
ILMN_1664516, ILMN_1673073, ILMN_1700337, ILMN_1716400, ILMN_1731184, ILMN_1753063, ILMN_1794539, ILMN_1808071, ILMN_2109197, ILMN_2344971	cg26738010

Table 3: Gene expression and methylation variables positively correlated through the PLS.

Expression Profiling by Array	**Methylation**
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg05859264
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg06766367
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg15387123
ILMN_1653504, ILMN_1720771, ILMN_1728570, ILMN_1734472, ILMN_1750961, ILMN_1766551, ILMN_1766712, ILMN_1766832, ILMN_1773006, ILMN_1784493	cg24898753

Table 4: Gene expression and methylation variables negatively correlated through the PLS.

Gene Symbol	Full Name	Function / Description
CENPF	Centromere Protein F	Cell division, mitosis
WDR62	WD Repeat Domain 62	Brain development, mitotic regulation
TROAP	Trophinin Associated Protein	Cell adhesion and proliferation
FOXM1	Forkhead Box M1	Transcription factor, cell cycle regulator
MELK	Maternal Embryonic Leucine Zipper Kinase	Cell proliferation, cancer-associated
KIF15	Kinesin Family Member 15	Mitotic spindle assembly
KIF14	Kinesin Family Member 14	Cytokinesis, cell proliferation
EPB41L3	Erythrocyte Membrane Protein Band 4.1 Like 3	Tumor suppressor
S1PR1	Sphingosine-1-Phosphate Receptor 1	Cell migration, immune signaling
STX11	Syntaxin 11	Vesicle trafficking in immune cells
TCF21	Transcription Factor 21	Mesenchymal development, tumor suppression
PEBP4	Phosphatidylethanolamine Binding Protein 4	Apoptosis, cell signaling
TM6SF1	Transmembrane 6 Superfamily Member 1	Lipid metabolism (less characterized)
CPA3	Carboxypeptidase A3	Mast cell protease, inflammation
LINC02538	Long Intergenic Non-Protein Coding RNA 2538	Non-coding RNA, regulatory function
FABP4	Fatty Acid Binding Protein 4	Fatty acid transport, inflammation

Gene Symbol	Full Name	Function / Description
IFNA17	Interferon Alpha 17	Antiviral immune response
KCNK16	Potassium Channel Subfamily K Member 16	Potassium ion transport
CYP2E1	Cytochrome P450 Family 2 Subfamily E Member 1	Drug metabolism, detoxification
TCF7L2	Transcription Factor 7 Like 2	Wnt signaling, diabetes-related
DLGAP2	DLG Associated Protein 2	Synaptic function, neuronal signaling
RB1	Retinoblastoma 1	Tumor suppressor, cell cycle control
GABRA2	GABA A Receptor Subunit Alpha 2	Neuronal inhibition, GABA receptor
GATA4	GATA Binding Protein 4	Transcription in heart and organ development
APOE	Apolipoprotein E	Lipid transport, Alzheimer's disease risk
RPTOR	Regulatory Associated Protein of MTOR Complex 1	Component of mTORC1, regulates cell growth in response to nutrients and stress
PAX8	Paired Box 8	Transcription factor involved in thyroid development and cancer
CDKN1C	Cyclin Dependent Kinase Inhibitor 1C	Cell cycle inhibitor, tumor suppressor
IGF2	Insulin Like Growth Factor 2	Growth factor involved in development and cancer

Table 5: Gene symbols, full names, and functional descriptions of genes in Tables 1 and 2