Analysis of RNA expression on different dietary regimes in obese subjects

Network Based Data Analysis

Ossanna Vittoria

June 10, 2023

1 Abstract

Overweight and obesity are growing health problems worldwide. One of the most effective strategies to lose weight is energy restriction (ER): restriction of food intake without malnutrition. Still, in literature is not clear the effect that different intake of macronutrients has on the overall condition. In this project I exploit data from a recent publication that tackles this lack of knowledge [1] and dig into the difference in expression in two dietary conditions (with high and standard protein ratio) along with the identification of enriched GO terms. From my analysis, it seems that the main distinction between these two lifestyles is summarized in a significant difference in the behaviour of RNA genes (binding and metabolism), along with elements liked to KRAS, an oncogene involved in several cancer developments.

2 Introduction

Overweight and obesity are growing health problems worldwide. The most effective strategy to reduce weight is energy restriction, which has been shown to be beneficial in disease prevention and it reduces chronic inflammation. The organ most extensively affected during energy restriction diets is white adipose tissue. Recent studies suggest that reducing the protein quantity of a diet contributes to the beneficial effects by energy restrictive diet, but studies on humans have been showing less consistent results with respect to animal tests. Based on previous studies and this meta-analysis, no definitive conclusion can be drawn on the effect of protein versus other macronutrient ratios in an energy restrictive (ER) diet on markers of metabolic health.
The objective of this project is to assess changes in gene expression between a high-protein diet and a normal protein diet during ER. This has been conducted through data available from Van Bussel et al. [1]: in this dataset we have 22 obese senior healty individuals that conducted an energy restrictive diet either with standard and high amount of protein. For further details about the collection of the samples please refer to the publication cited above.
The approach that this project follows consists in tree main steps: first I performed an exploratory analysis of the data along with PCR and conducted feature selection. Second, several classification methods have been tested in order to distinguish at best samples coming from high protein ratio and standard one. Last, from the best performing method, I extracted a list of probes that the algorithm identified as most informative in order to perform classification. This list has been used for over representation analysis and network based data analysis with different frameworks. To sum up, I analyzed the results we get from the last analysis and looked for validation of the results in literature, along with a comparison with the findings of the dataset's authors.
All scripts and data are freely available in this GitHub repository.

3 Materials and methods

3.1 Dataset

The data used for this study have been selected from a publication from *Van Bussel et al.*. published in 2017 [1] and are available on GEO (*Gene Expression Omnibus*) with accession number `GSE84046`. This dataset consists in 44 samples and 33297 RNA expression data, coming from an Affymetrix microarray technology. Among the 44 samples we can distinguish in:

- 22 control samples (C), taken from 22 obese subjects. These individuals will undergo an energy restrictive diet (restriction of 25%), either with standard (ER_SP) or high (ER_HP) amount of proteins.
- 12 samples (ER_SP) taken from the same subjects after 12 weeks of standard protein level diet (0.9 $\frac{g}{kg}$ per day).
- 10 samples (ER_HP) taken from the same subjects after 12 weeks of high protein level diet (1.7 $\frac{g}{kg}$ per day).

Subcutaneous adipose tissue biopsies were collected in all the three cases, before and after each diet. Original data already went through a process of *log* normalization, the corresponding boxplot is reported in Figure 7 in the Attachments.

3.2 Methods for analysis

This project exploited several advanced methods in order to get to meaningful results. Most of them are based on machine learning, both supervised and unsupervised. In this section will follow briefly methods used in the project along with packages and libraries.

At first, I conducted PCA and several clustering methods. Results from initial PCA led me to a filtering of the dataset for the downstream analysis. This task has been performed with the function prcomp from the library stats. Unsupervised methods for clustering have been exploited, such as K-means and hierarchical clustering. This analysis has been performed through the library stats, the number of clusters have been defines as $k = 3$ since, as explained before, I am dealing with a dataset with three classes. For the latter clustering method, I exploited several ways of filling the distance matrix. For supervised learning we mean approaches in which we are trying to perform some task starting from a vector of features along with the label. In this case we are dealing with discrete labels and the task will be classification. The labels used for classification are high-proteins diet and standard-protein diet, for some analysis also the control has been included. The fist method tested was random forest: this analysis has been done by using R libraries caret and randomForest, in order to get training and test. This procedure has been performed with $seed = 2000$ and the number of trees in the forest equal to 1000. Other 4 classification methods have been trained in the same manner on the same subset of data. These four methods are: LDA, rScudo, RIDGE and LASSO regression.
Starting from the best performing method for classification, I extracted a list of the top 200 probes that the method defines of importance when performing the classification task. I selected RIDGE and Random Forest as the two most accurate methods I experimented. Starting from these selections, I conducted Over Representation Analysis with gProfiler. Secondly, I used the same list of data (coming from RIDGE regression) for Network Based Analysis through pathfindR and STRING. While for the first one we should get functional annotation of the selected genes (therefore identifying over represented pathways or GO terms enriched in the list given as input), the second class of methods are supposed to expand the networks in order to find interactor outside the list or validate the one found by standard analysis. As last, I conducted a research in literature about the finding and a comparison against the published paper from the author of the dataset.
For the implementation details and further information about the methods, please refer to the script provided in the GitHub repository mentioned in the introductory section.

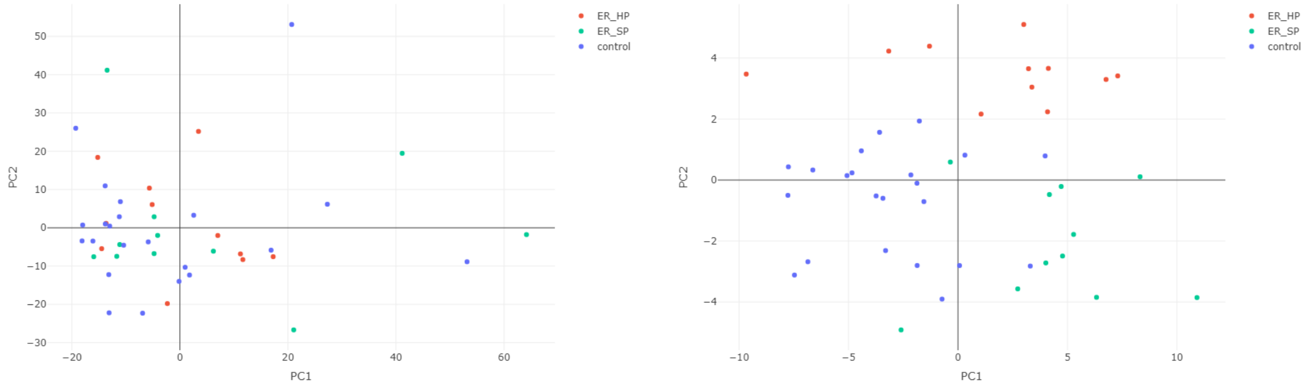
4 Results

4.1 Principal Component Analysis

As first approach for exploratory analysis of the dataset, I performed a standard Principal Component Analysis (PCA) with all the probes present in the table. I divided samples in three groups: C, ER_HP and ER_SP. In Figure 1a we can see he results: it is clear that no evident clustering is present exploiting this raw approach. For this reason, I deiced to perform a filtering on the data: I performed a standard *t-test* for every pair of groups (C vs ER_SP, ER_SP vs ER_HP, ER_HP vs C) and I kept only the probes that have been considered significant in at least one of the comparison (*p-value* below 0.01). By doing this filtering, I am keeping only genes that look significantly different between at least two groups. This filtering criteria should allow me to reduce significantly the amount of probes involved in the microarray without reducing too much the dataset. Since the next steps consists also in a selection of significant genes starting from this filtering, I avoided being to stringent on the selection. For the same reason, I decided not to go for more advanced method nor multiple hypothesis correction because in this way the test would have produced highly selected results. This procedure produced a second dataset consisting in 44 sample and 1259 probes that passed the filtering test. I performed the PCA analysis on this second with improved results (Figure 1b): in this plot is it possible to distinguish more clearly the region in which the three groups are concentrated. Still, the boundary region is blended: if no label was given, we could not distinguish clusters of any kind.
Every following analysis have been therefore performed using this filtered dataset.

4.2 Clustering

The project proceeds with several method of unsupervised learning, in particular for clustering. I first performed this task using K-means algorithm and then some trials with hierarchical clustering.



(a) PCA with whole dataset

(b) PCA with filtered dataset

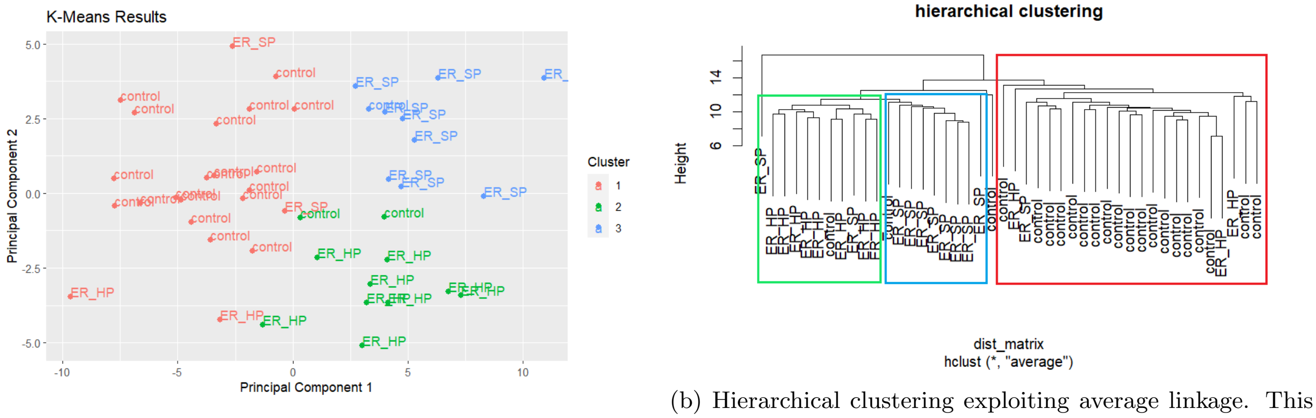
Figure 1: PCA plots with filtered and whole datasets. In the case of the whole dataset there no distinguishable cluster of data, while for the filtered data we see a better separation of the samples.

K-means

The results for this approach are reported in Figure 2a. Since the boundaries created within the PCA space do not exactly separate the samples, the results from this analysis are really dependent on the seed for the random initialization. Anyhow, with seed = 2900 I achieved a satisfying results, aside from the evident outliers present in the dataset.

Hierarchical clustering

The second approach consists in hierarchical clustering. Results for average linkage are reported in Figure 2b, while single and complete linkage are reported in attachments files (Figure 8). From the reported plot we do not clearly have a distinction between the groups starting with the first two branch separations, but looking at a deeper level we can notice that this method performs classification quite good, with exception for some outliers (also present before for PCA and K-means).



(a) K-means clustering using seed = 2900. The labels assigned to each point in the 2D space correspond to the ground truth labels.

(b) Hierarchical clustering exploiting average linkage. This method clusters the three groups after performing some initial branching. Aside from some outliers, we see that the three groups segregate quite well.

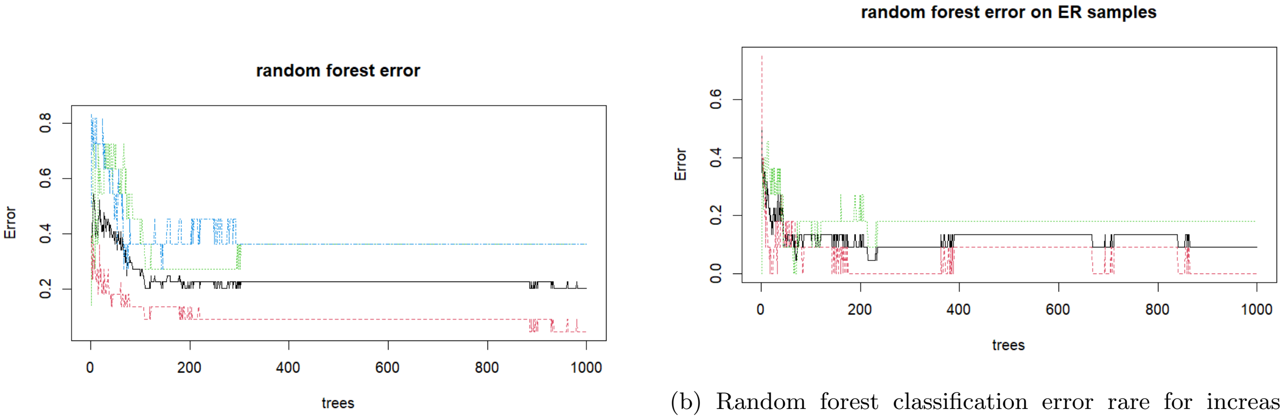
Figure 2: Clustering results

4.3 Supervised analysis

The project proceeds with several method of supervised learning, in particular Random Forest, rScudo, LDA, RIDGE and LASSO regression. Some of the results will be used for further analysis.

Random Forest

In this project I trained a random forest with increasing number of trees on the filtered dataset, graphical visualization of the error rates are reported in Figure 3a. In this plot we can see that, even if the error rates reaches a stable value after about 300 trees, the error rate for two groups are high, close to the toss of a coin. This is probably due to the fact that the control group is over-represented with respect to the two groups of the two diets. This reasoning led me to the idea of performing this training based only on the two diets (therefore ER_HP and ER_SP) in order to have a better balance between the numbers of samples. Graphical outcomes of this analysis are reported in Figure 3b. From this plot we can see that the error gets stable around 200 trees, but the two error rates are lower than 20%, implying an average accuracy around 90% (black line).



(a) Random forest classification error rate for increasing number of trees. This approach reaches a plateau but has low accuracy for two out of three groups.

(b) Random forest classification error rare for increasing number of trees considering only samples coming from the energy reduction diets' groups. This approach reaches a plateau and gets more accurate results within lower error rate for both groups.

Figure 3: Random forest error rates with increasing number of trees.

Other classification methods

Other 4 classification methods have been trained in the same manner on the same subset of data. These four methods are LDA, RIDGE regression, LASSO regression and rScudo.

The training of these methods, along with random forest, have been carried on in the same way and performance based on accuracy have been evaluated. In Figure 4, we can compare the accuracy and the variance of the trained models: the best performing algorithms are RIDGE regression and Random Forest training, equally reaching 0.95 accuracy. For all the the methods, variance is high, probably due to the low number of samples per group (10 and 12). rScudo results scored around 0.6 of accuracy, reaching a result close to random with parameters $nTop = nBottom = 25$ and $N = 0.4$. Results of training and testing through rScudo are reported in the attachment section (Figure 9). Still, parameters reported in the text look like the best performing when manually analyzing the network structure.

4.4 Network based analysis

As explained in the methods section, I selected the best performing methods for classification and extracted a list of 200 probes that those methods considered most informative in order to conduct their classification task. As Figure 4 is showing, RIDGE and Random Forest are performing both very well on the test data, therefore I extracted the list for both of them and used (eventually cut to different length) for the following analysis.

Over Representation Analysis

Tools for performing ORA (Over Representation Analysis) are DAVID and gProfiler, of these two, only the latter had the power to interpret the probe identification numbers of my data (linked to the microarray platform adopted). I was led to an obligated choice of using gProfiler for this step. I therefore performed this analysis on both a list of 200 from Random Forest and from RIDGE. As we can clearly see, the results from the over

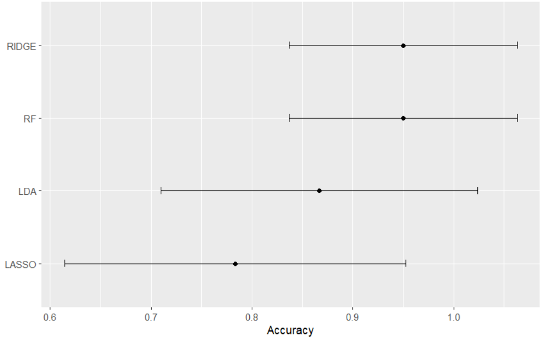
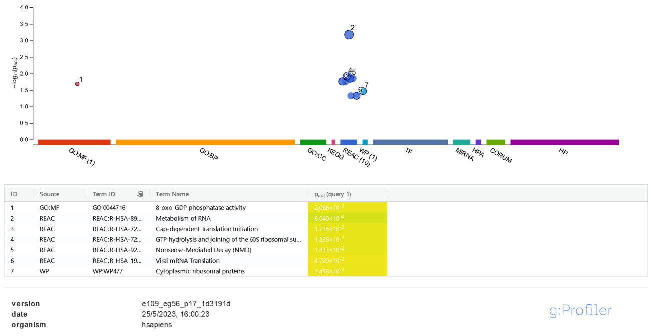


Figure 4: Accuracy differences for classification methods using cross validation available on `caret`. Even if the maximum value for the accuracy is = 1, RIDGE and Random Forest get above considering the variance. This problem is likely correlated to the low number of samples available in this study.

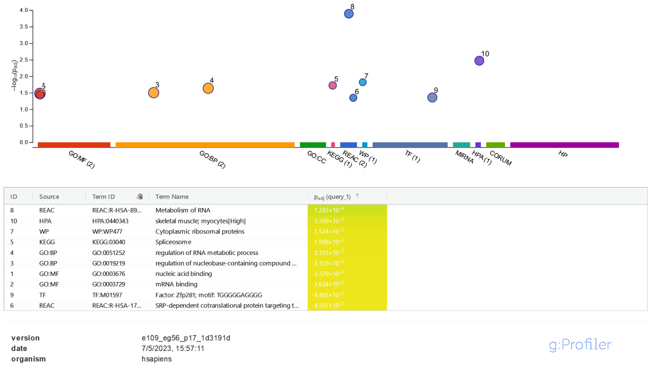
representation analysis led us to very few results: in both cases the lower p-value that we get is associated to RNA metabolism, and items from the biological process found for the RIDGE set of probes confirm the presence of terms associated to RNA regulation. The correlation of RNA metabolism and high protein intake has been already confirmed by Muñoz-Martínez et al [2]. Even if the right platform definition for this dataset was available on gProfiler, it seems that many probes ids correspond to a ”None” value. I tried also to convert the probes ids through `biomaRt` library and use HUGO nomenclature as input for gProfiler but no significant change has been reported. Since the list coming from RIDGE classification look richer in items found by ORA, I decided to use its list for the following steps.



ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0044716	8-oxo-GDP phosphatase activity	2.066×10^{-2}
2	REAC	REAC:R-HSA-89...	Metabolism of RNA	6.640×10^{-4}
3	REAC	REAC:R-HSA-72...	Cap-dependent Translation Initiation	1.735×10^{-2}
4	REAC	REAC:R-HSA-72...	GTP hydrolysis and joining of the 60S ribosomal su...	1.236×10^{-2}
5	REAC	REAC:R-HSA-92...	Nonsense-Mediated Decay (NMD)	1.433×10^{-2}
6	REAC	REAC:R-HSA-19...	Viral mRNA Translation	4.729×10^{-2}
7	WP	WP:WP477	Cytoplasmic ribosomal proteins	3.418×10^{-2}

version	e109_eg56_p17_1d3191d
date	25/5/2023, 16:00:23
organism	hsapiens

(a) ORA results for the top 200 genes used from random forest for classification. We can see from this plot that we get very few results, the majority coming from Reactome (pathways) database.



ID	Source	Term ID	Term Name	p_{adj} (query_1) ↑
8	REAC	REAC:R-HSA-89...	Metabolism of RNA	1.283×10^{-4}
10	HPA	HPA:0440343	skeletal muscle; myocytes[High]	3.399×10^{-3}
7	WP	WP:WP477	Cytoplasmic ribosomal proteins	1.524×10^{-2}
5	KEGG	KEGG:03040	Spliceosome	1.908×10^{-2}
4	GO:BP	GO:0051252	regulation of RNA metabolic process	2.315×10^{-2}
3	GO:BP	GO:0019219	regulation of nucleobase-containing compound ...	3.159×10^{-2}
1	GO:MF	GO:0003676	nucleic acid binding	3.329×10^{-2}
2	GO:MF	GO:0003729	mRNA binding	3.654×10^{-2}
9	TF	TF:M01597	Factor: Zfp281; motif: TGGGGGAGGGG	4.405×10^{-2}
6	REAC	REAC:R-HSA-17...	SRP-dependent cotranslational protein targeting t...	4.507×10^{-2}

version	e109_eg56_p17_1d3191d
date	7/5/2023, 15:57:11
organism	hsapiens

(b) ORA results for the top 200 genes used from RIDGE for classification

Figure 5: ORA analysis for both random forest and RIDGE lists of important probes.

Network based analysis

As next and last step I performed Network based analysis trough `STRING` and `pathfindR`. These tools allow us to discover new potential interactors or enriched pathways via including neighbours of the list of genes that we give them as input. This is done by considering links present in literature and performing the analysis on an enriched list. Since the set of genes will be expanded by the tools, I considered subsets of different length of the list (100, 150 and 200), still, the most relevant results for pathfindR have been obtained with the length of 200. For this reason I maintained the same list for all approaches to increase coherence and allow a more fair comparison.

Starting from pathfindR results in Figure 6, we see that the most interesting results (with low p-value) are not present in the ORA performed with gProfiler. Nevertheless, it is known from literature that the condition of obesity is linked to the level of the maturation of the oocyte though progesterone levels [3]. From this insight, we can hypothesize that one of the two diets, either with standard or high protein ratio, are modifying this peculiar behaviour.

From the network analysis of STRING and pathfindR (reported in the attachments as Figures 10 and 11) we can see that in both cases we can identify a cluster of enriched terms orbiting around KRAS, a GTPase that is known to play an important role in the regulation of cell proliferation, also involved in promoting oncogenic events [4]. The network produced by STRING found a big cluster of factors related to RNA binding proteins, RNA modifiers and proteasome regulatory subunits. This cluster is far more evident than the one involving KRAS and it is also shared by the over representation analysis results above.
From a comparison of the pathways found by pathfindR and the links found by STRING focused on KRAS, I found that `HBEGF` is involved the pathway `hsa05219`, linked to development of bladder cancer. This item is the only one in common between first and secondary neighbours of KRAS in the STRING network and the pathways found by pathfindR.

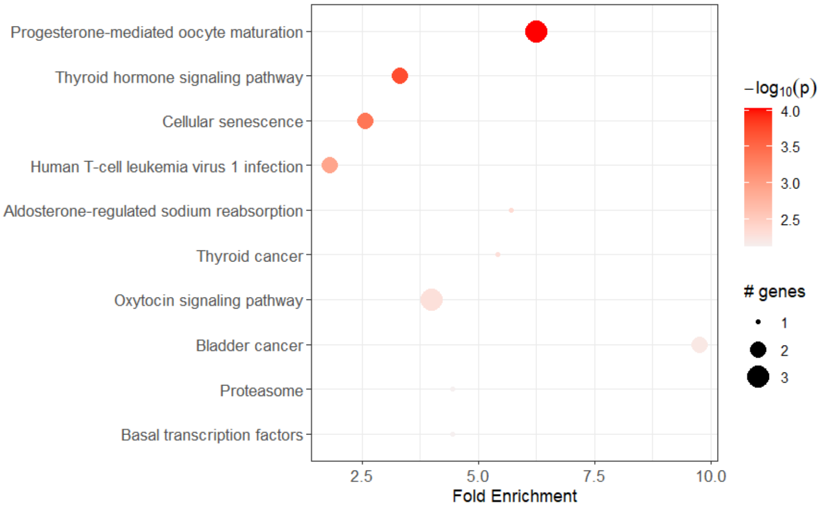


Figure 6: PathfindR dotplot, from this graph we can see the terms that seems to be enriched in the expanded network from the analysis with pathfindR.

5 Discussion

This project set its focus into finding association and differences in gene expression between obese individuals following an energy reductive diet either with normal or high levels of proteins. The best performing methods for distinction of these groups are RIGDE regression and Random Forest, the latter performs as good as the first exploiting around 200 trees. Even if the training of these method have been done through cross validation, results could be biased since the number of samples available is as a matter of fact low. Results with this number of samples could be highly linked to the data rather than the cohort condition: in this settings outliers will have a big impact on the analysis, still, I decided not to discard those samples in order not to further reduce the data available.
From the network analysis we clearly saw from more approaches that the probes selected from the classifier are enriched of terms related to RNA binding and RNA modification. Recent studies have discovered that post-transcriptional regulation, mainly mediated by RNA-binding proteins (RBPs), also plays a crucial role in excessive fat accumulation in adipose cells [2] [5]. Related to these last publications, I looked into the RBPs that they analyse and searched for a correspondence in my lists of genes, still, no matches have been found. In the original study was found, by doing a Differential Expression Analysis, that RNA metabolism was enriched with respect to the control condition [1]. From this point of view, we can assess that one of the two diets taken into account could be beneficial the behaviour and the level of activity of these protein. Still, at this point we are only sitting in front on speculation and I suggest deeper analysis. From the expansion of the network we also obtained an interesting p-values concerning interactors with KRAS protein, which is known from literature to have a large nutritional attributable risk of mutation when involved in several types of cancers [4]. Still it is not clear at this point why a difference in macro nutrients in the diet should be linked to KRAS in these subjects. Since this latter result comes from enriched networks, I suggest to carry further analysis to assess an actual possible relevance of this link.

References

[1] I P Van Bussel, E M Backx, C P De Groot, M Tieland, M Müller, and L A Afman. The impact of protein quantity during energy restriction on genome-wide gene expression in adipose tissue of obese humans. *International Journal of Obesity*, 41(7):1114–1120, 2017.

[2] K Kita, S Matsunami, and J Okumura. Relationship of protein synthesis to mRNA levels in the muscle of chicks under various nutritional conditions. *J. Nutr.*, 126(7):1827–1832, July 1996.

[3] Scott H Purcell and Kelle H Moley. The impact of obesity on egg quality. *J. Assist. Reprod. Genet.*, 28(6):517–524, June 2011.

[4] Stelzer G, Rosen R, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, Iny Stein T, Nudel R, Lieder I, Mazor Y, Kaplan S, Dahary, D, Warshawsky D, Guan - Golan Y, Kohn A, Rappaport N, Safran M, and Lancet D. The genecards suite: From gene data mining to disease genome sequence analyses. *Bioinformatics*, 2016.

[5] Pengpeng Zhang, Wenyan Wu, Chaofeng Ma, Chunyu Du, Yueru Huang, Haixia Xu, Cencen Li, Xiaofang Cheng, Ruijie Hao, and Yongjie Xu. RNA-binding proteins in the regulation of adipogenesis and adipose function. *Cells*, 11(15):2357, July 2022.

Attachments

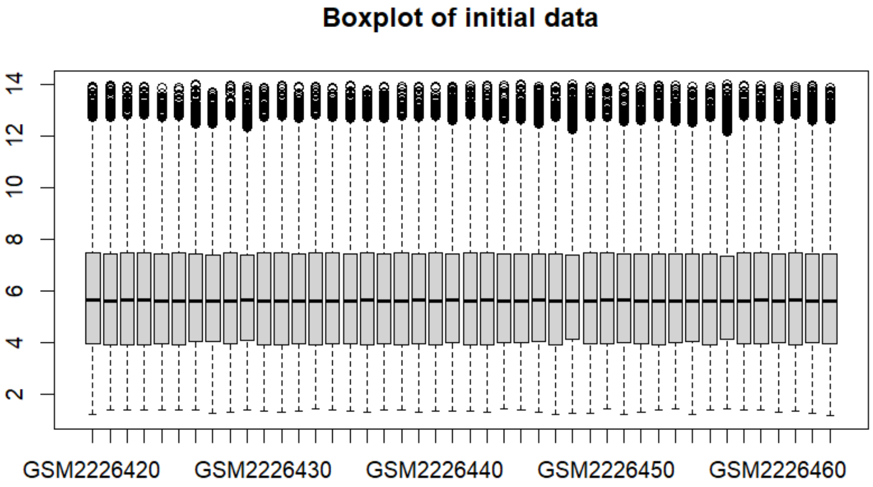
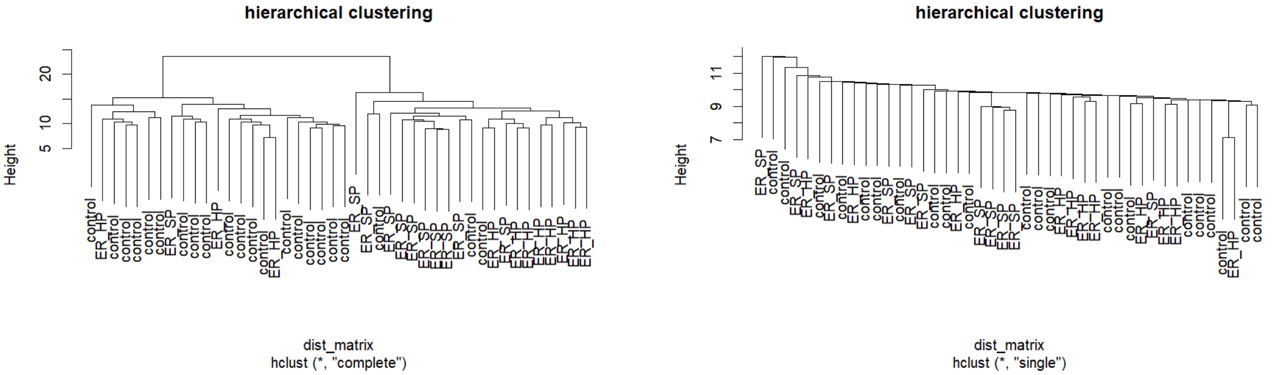


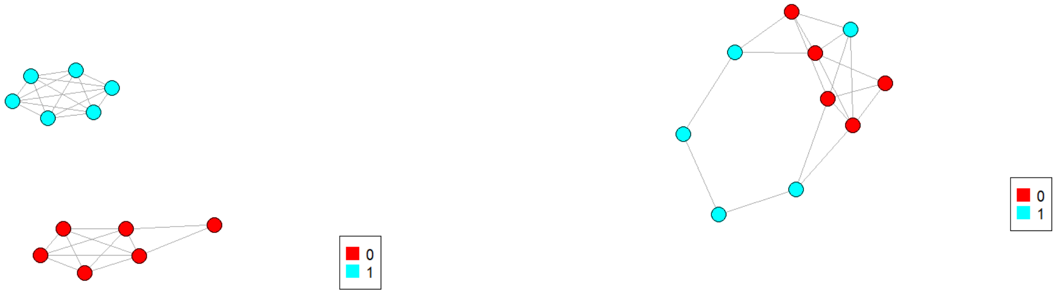
Figure 7: Boxplot of initial values per sample (x-axis), without further processing. Since median and variance looked well aligned, no further process of normalization have been done.



(a) Hierarchical clustering exploiting complete linkage.

(b) Hierarchical clustering exploiting single linkage.

Figure 8: Hierarchical clustering results



(a) Training network obtained from Scudo

(b) Test network obtained from Scudo. Even if samples in training have been perfectly separated, test looks not clearly classified. Accuracy is 0.6.

Figure 9: rScudo performance in training and test

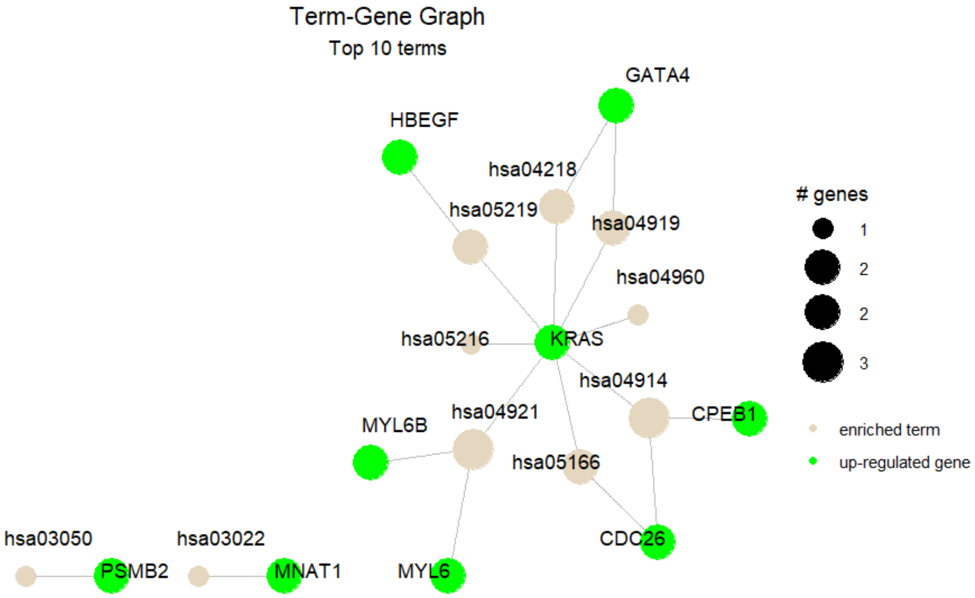


Figure 10: pathfindR network of interactions. This network have been achieved giving as input the list of the 200 most important genes for the RIDGE classification.

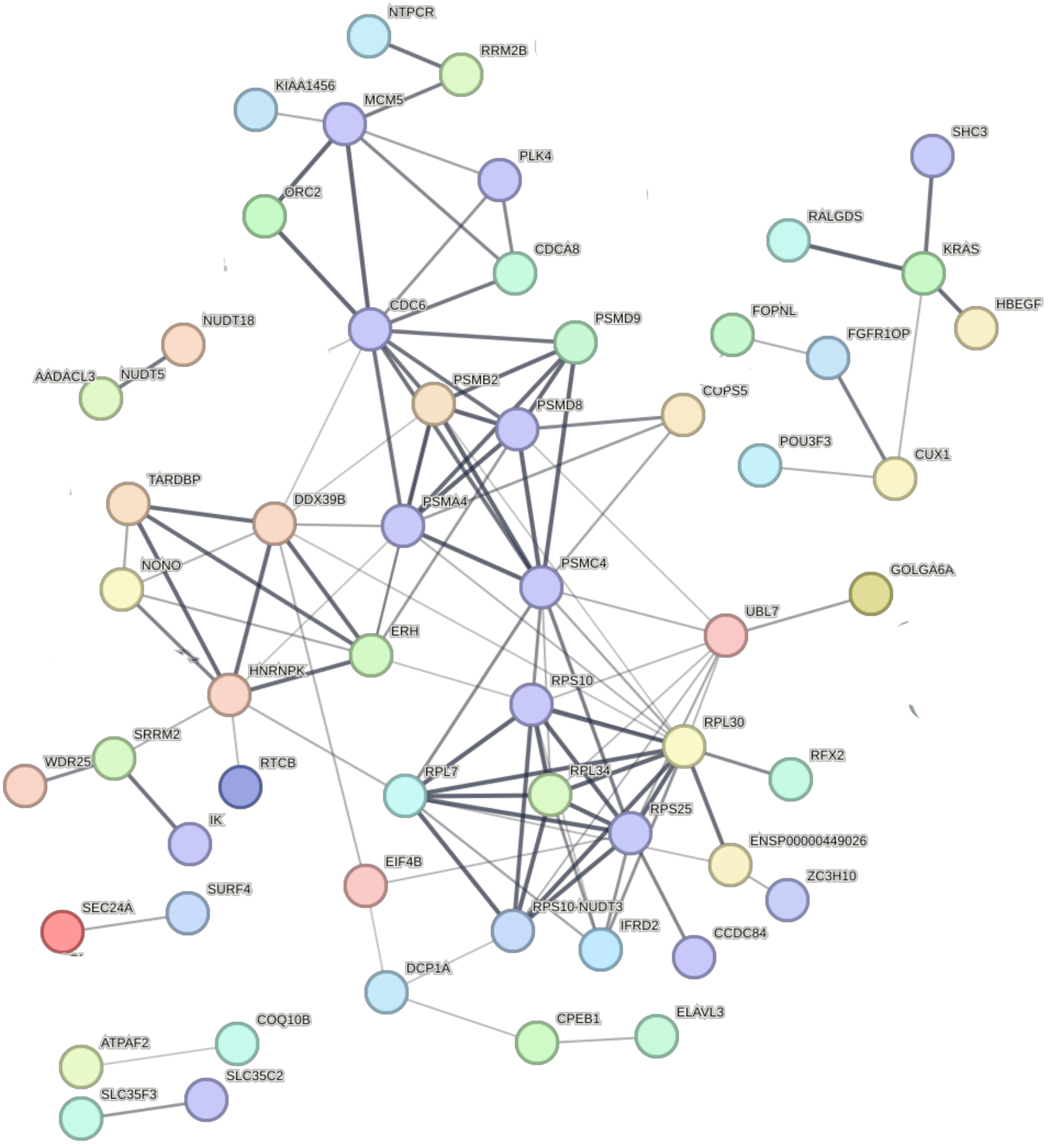


Figure 11: STRING network of interactions, the links with a wider stroke means a higher confidence level. This network have been achieved giving as input the list of the 200 most important genes for the RIDGE classification. We can clearly see a cluster of nodes in the middle, representing mostly proteins related to RNA modification.