Università degli Studi di Trento

Network Based Data Analysis Final Report
Cluster & Enrichment Analysis in PCOS Samples

Alberto Lupatin

CiBio, 13/06/2024

1 Abstract

Polycystic ovary sindrome (PCOS) is one of the greatest disorders affecting a significant percentage of women of reproductive age worldwide. Traditional approaches to study PCOS have been focused on the clinical symptoms and the individual biomarkers; however, these methods often lack in interpretability and do not allow to compare several sample in one study.
In this study we analyze the sample clustering and the differential gene expression of 29 Vastus lateralis muscle samples from 16 control women and 13 obese women with PCOS. To perform cluster analysis, both supervised and unsupervised learning methods such as K-means clustering, hierical clustering, or random forest were used. Results showed that supervised learning performed better compared to unsupervised learning. Functional analysis performed on a set of genes with very low p-value after applying a *t-test*, revealed some already correlated pathways, but further studies about them need to be done. Furthermore, enrichment analysis was performed on the same genes. Results indicate new opportunities for researches, since new pathways, such as taurine metabolism, have been found.

2 Introduction

2.1 State of the Art

PolyCystic Ovary Syndrome (PCOS) is one of the most common hormone disorders affecting about one in seven reproductive-aged women worldwide and approximately 6 million women in the United States (U.S.). PCOS can be a significant burden to those affected and is associated with an increased prevalence of mental health disorders such as depression, anxiety, eating disorders, and postpartum depression [5].

PCOS's key features are hyperandrogenism, ovulatory dysfunction, and polycystic ovarian morphology, with excessive androgen production by the ovaries. First PCOS's sympthoms in women were were already noted in 1921, when Doctors Achard and Thiers reported a possible linkage between carbohydrate metabolism and hyperandrogenism, called: "*the diabetes of bearded women (diabete des femmes a barbe)*". It was only in 1980 that the correlation between PCOS and basal and glucose-stimulated hyperinsulinemia was demonstrated [1]. In particular, Burghen and colleagues noted significant positive linear correlations between insulin and androgen levels and suggested that this might have etiological significance.

2.2 Aim of the Study

This report aims to analyze the dataset *GSE6798* provided by Skov *et al.*. In particular, by examining the gene expression in skeletal muscle from obese and hyperinsulinemic women, researchers are looking for a possible explanation for the increased risk of type two diabetes in women with PCOS compared to healthy individuals.

3 Methods

3.1 Experiment Design

The dataset is composed of 29 Vastus lateralis muscle samples, collected from obese (Body Mass Index ≥ 25) women with ($n = 16$) or without ($n = 13$) PCOS. The two groups were similar for these parameters:

- Age
- Body Mass Index (BMI)
- Skeletal muscle insulin sensitivity, measured by euglycemic-hyperinsulinemic clamp
- Metabolism activity, measured by indirect calorimetry

Gene expression was obtained from the specimens through the HG-U133 Plus 2.0 expression array from Affymetrix.

3.2 Principal Component Analysis

Principal Component Analysis (PCA) is a machine learning method used to simplify a large dataset into a smaller set while still maintaining significant patterns and trends. Even though it sacrifices some accuracy, the advantage of having a smaller dataset containing just the important variables dramatically helps the data analysis process.

PCA was performed in R through *prcomp*, contained in *stats v. 4.3.3* R package. This function performs a PCA on the given data matrix and returns the results as an object of class *prcomp*. PCAs were then plotted through *plot* function in *base v. 4.3.3*.

3.3 Clustering

Clustering aims to automatically split the sample based on their relative similarity.

All the methods used are based on a distance function, which is one of the methods to estimate similarity. In particular, the distance function uses mathematical distance to assess the similarity among samples.

In this report, it will be performed:

- **Unsupervised Learning**: the data given as input is unlabeled and the algorithm is free to discover patterns without restrictions
- **Supervised Learning**: the input dataset is labeled by the experimenter so that the algorithm can learn the relationships between the input and the output.

3.3.1 K-means

The K-means clustering is an unsupervised learning method. Its objective is to divide the set of n observations into k clusters, with each observation assigned to the cluster whose mean is close. Geometrically it looks at centroids: point which is representative of the center of the cluster in the space. The algorithm starts with a random estimate of the centroid and then uses other samples to refine its position. In output, it will create a data structure that describes to which group each sample is assigned to

The number of clusters has to be chosen according to the experiment design. Since in this study two groups of patients are compared, the clusters found will be two.

K-means clustering is a very fast and efficient algorithm. However, its main weaknesses regard the accuracy of the centroid position prediction and the sensitivity to outliers.

K-means clustering was performed using *kmeans* R function while plotting of the results was done with *plot* R function. Data regarding sample type was added to the plot to better analyze the plot. Both functions are present in *stats v. 4.3.3* package.

3.3.2 Hierical Clustering

Hierical Clustering (HC) is an unsupervised clustering method that exploits information contained in the distance matrix to build a dendrogram.

Unlike, K-means clustering, to perform HC the distance matrix has to be calculated first. This step was done in R thanks to *dist* function with default parameters. Then, clustering was performed through *hclust* by setting two different methods: average distance and McQuitty. Lastly, results were plotted with *plot* function and the two main clusters were divided with the *rect.hclust* function. All the functions cited are present in *stats v. 4.3.3* R package.

3.3.3 Random Forest

Unlike K-means and HC, Random Forest (RF) is a supervised learning method composed of multiple decision trees. Each tree seeks to find the best split to subset the data by exploiting information given in input. Data is the process through an ensemble learning named bagging. It consists of training multiple models independently on random subsets of the data and aggregating their predictions through voting or averaging. RF classifier is an extension of bagging, which adds another level of randomness called feature randomness. As a result, RF reduces the risk of overfitting, bias, and overall variance, resulting in more precise predictions.

Generally, a Random Forest classifier is a commonly used machine learning algorithm that combines the output of multiple decision trees to reach a single result.

To perform RF in R, *randomForest* function was exploited. To optimize the function's performance, the number of trees to grow was set to 600. This decision was made after assessing how the accuracy of the prediction was affected by each tree's growth. Figure 1 shows that the accuracy did not decrease after 600 trees. The function *randomForest* is contained in *randomForest v. 4.7-1.1* R package.

3.3.4 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is another supervised learning method. In particular, it is focused on:

- Maximising the **between-class distance**: the distance between the centroids of different classes
- Minimising the **Within-class distance**: accumulated distance of an instance to the centroid of its class

To evaluate LDA performance ROC curve has been calculated. The ROC (Receiver Operating Characteristic) curve is a graph showing the performance

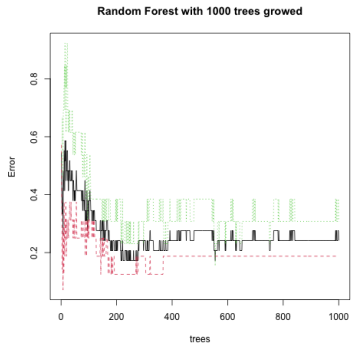


Figure 1: Level of error for the increasing number of trees in the ensemble. The green line represents the error of the control patients, the red line the error of the affected patients, and the black line represents the average between the two.

of a classification model at all classification thresholds. It plots the specificity (or the true negative rate) on the X-axis and the sensitivity (or the true positive rate) on the Y-axis calculated for every threshold.

Firstly, a subset of the initial dataset was isolated. In particular, a *t-test* was performed based on the samples' metadata. Data regarding the probes with a $p\text{-}value < 0.01$ were extracted.

Then, LDA was performed in R through *lda* function contained in package *MASS v. 7.3-60.0.1*. To further evaluate LDA performance, the dataset was split into training and testing. LDA arguments were:

1. Function $\rightarrow$ *Affected* $\sim$.: name of the last column of the input data frame indicating whether the sample is a control or a PCOS one. It specifies to the function which variables have to guess (*Affected*) and which ones can be used to do the guessing (everything but Affected).

2. Data $\rightarrow$ the input data frame

3. Prior $\rightarrow$ 0.5, 0.5. This is the prior probability of class membership.

4. Subset $\rightarrow$ the training dataset

ROC curve was performed in R through *plot.roc* function contained into *pRoc v. 1.18.5* package.

3.3.5 Linear Regression Analysis

Linear Regression Analysis (LRA) is used to predict the value of a variable based on the value of another variable. Notably, LRA estimates the coefficients of the linear equation, involving one or more independent variables that best

predict the value of the dependent variable. It fits a straight line or surface that minimizes the discrepancies between predicted and actual output values.

A known issue of LRA is overfitting, which occurs when an algorithm fits too closely or even exactly to its training data, resulting in a model that can't make accurate predictions or conclusions from any data other than the training data. To partially solve this issue, there are 3 main methods: Lasso, Ridge regression and Elastic net regulatization.

In our study, the first method was used thanks to the *train* function, in *caret v. 6.0-94* R package, by adding the following parameters:

- Function → *Affected* ∼.: refer to LDA methods
- Method → glmnet
- Family → binomial
- alpha → 1 so that glmnet performs lasso penalty
- control → 10 fold cross-validation
- Metric → Affected: name of the last column, refer to LDA methods

Furthermore, a comparison among Lasso, RF, and Caret was performed and plotted through *ggplot* function, in *ggplot2 v. 3.5.1* R package.

3.3.6 Caret

Since RF, LDA and LRA cannot be directly compared as the performance's values are not relatable, to analyze the differences between the two methods, *Caret* package was used.

Caret uses a more sophisticated approach when it comes to training and testing. In particular, it does a n-fold cross-validation, gets a robust estimate of the results, and plots the latter. This pipeline is used for every supervised clustering method, so the results are comparable.

Firstly, we chose the parameters of the *function* by exploiting the *trainControl* function with the following parameters:

- Method → cv: set the resampling method to cross-validation
- Number → 10: to specify a 10-fold cross-validation
- Repeats → 5: repeat the 10-fold cross-validation 5 times

Secondly, the *train* function was exploited for LDA, RF, and LRA by setting the same Function parameter as the LDA, the metric as *"Accuracy"*, and the train control set in the previous step. Thirdly, results from both methods were merged with *resample* function. Lastly, results were plotted through *ggplot* function.

Every function but *ggplot* was from *caret v. 6.0-94* R package; *ggplot* function is contained into *ggplot2 v. 3.5.1* R package.

3.4 Functional Analysis

Functional analysis was performed through the Database for Annotation, Visualization, and Integrated Discovery (DAVID) online tool. This database provides a comprehensive set of functional annotation tools that allows a biological understanding of large list of genes.

The input dataset was chosen by intersecting the top 200 differentially expressed genes extracted from the *t-test* (see LDA) and the RF. The resulting 189 genes were uploaded in DAVID site. In addition, the *Homo sapiens* specie was selected and the *Human Genome U133 Plus 2 Array* background was chosen. Furthermore, everything but gene ontology and pathways was deselected in order to limit the search results. Results were then ordered by the *Benjamini* value.

3.5 Enrichment Analysis

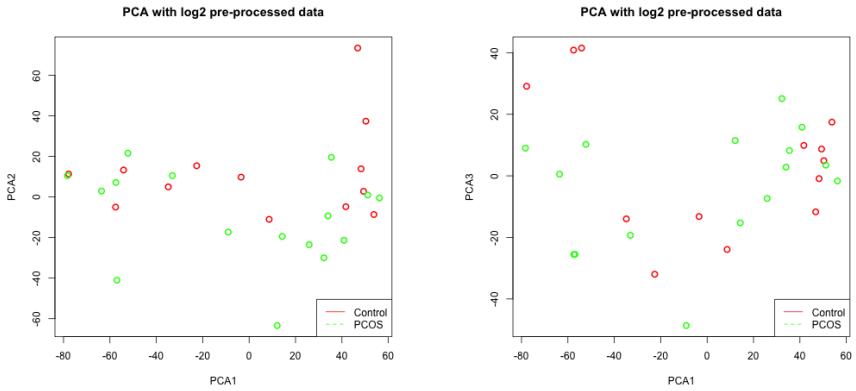
Gene enrichment analysis was performed using the EnrichNet online tool and using the 189 genes extracted in the previous paragraph (see Functional Analysis). Results regarding the possible pathways were downloaded and imported in R, where have been plotted through *ggplot* function in *ggplot2 v. 3.5.1.*

4 Results

4.1 Principal Component Analysis

In figure 2, PCA is shown. According to both graphs, two clear clusters cannot be found. However, there seems to be some separation in both graphs, even though it does not seem to be correlated with the metadata.

In figure 2, two Principal Component Analysis (PCA) graphs are presented to visualize the distribution of samples based on their principal components. Two distinct clusters are not clearly defined. This suggests that the ability to catch inter-sample variability by the PCA was not sufficient to separate the dataset into two defined clusters.



(a) PCA 1 vs PCA 2 *(b) PCA 1 vs PCA 3*

Figure 2: Two Plots representing the first three main PCAs in different combinations. Colors representing metadata were added: red dots illustrate control samples while green dots indicate samples with PCOS.

However, in plot 2a representing PCA 1 and PCA 2, there seems to be some degree of separation between the left and the right part of the graph, with some samples occupying the center. Indeed, these separations do not correlate with metadata, as both sample types are mixed. This lack of correlation indicates that the variation detected by the PCA is due to some factors that are not provided by the researchers. As a consequence, we can infer that with additional metadata, such as the participants' physical activity level or obesity status (e.g. obese or morbidely obese, years of obesity), it would be reasonable to find some correlation.

4.2 Clustering

4.2.1 K-means

Results from K-means clustering show an overall great separation of the samples in the two groups, thus revealing a more distinct separation from the two samples' types.
According to the plot in figure 3:

- Cluster number 1 on the left side of the graph consists mainly of control samples. Merely 4 PCOS patients have been included in this cluster.
- Cluster number 2, located on the right side of the graph, is primarily composed of PCOS samples, including only 2 control samples.

Clearly, results show a strong correlation of Cluster 1 with control samples and of Cluster 2 with PCOS patients.

The marked separation between the two clusters is a sufficient indicator that K-means clustering is a great tool for separating samples between control and PCOS.

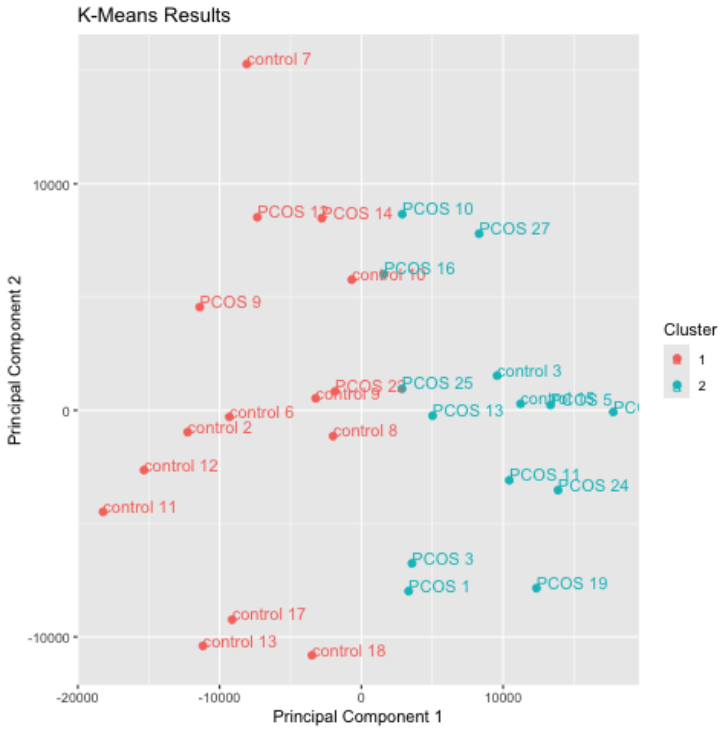


Figure 3: K-means clustering plotted through PCAs. Colors represent the two clusters found by the algorithm. Labels with samples' names have been added

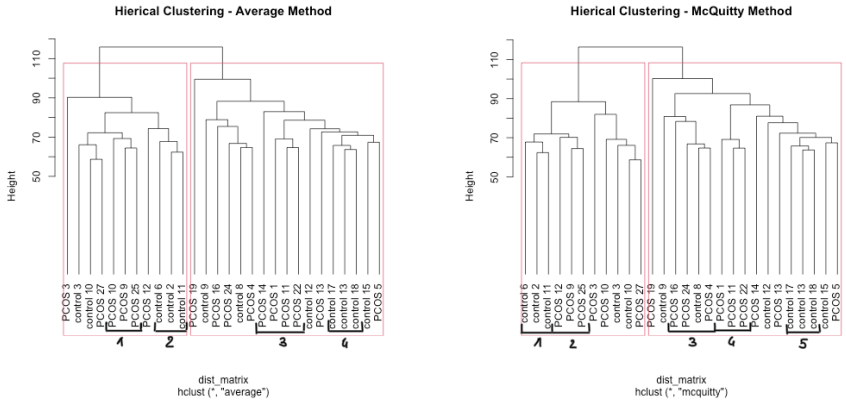
4.2.2 Hierical Clustering

When HC was applied the first time, sample *control 7* was interfering with the results, since it was being grouped alone in a single cluster. This goes to show that the HC method is sensible to outliers. For this reason, this sample was removed before performing HC.

To make results comparable, the new dataset without the outlier was used to perform HC with both clustering methods.

In figure 4, results are shown. Even though the two clusters found are not perfect, both methods correctly correlated some samples' distance with their type (control or PCOS). For instance, inner clusters number 1 and 2 in both graphs are composed of the same samples. Overall, HC exploiting the McQuitty algorithm, in figure 4b, correctly identified 5 inner clusters of samples, while the Average algorithm, in figure 4a, only identified 4 clusters. Nonetheless, we can

infer that both methods are comparable for this study case, and that, generally, HC is a good clustering method to enhance the differences between the two samples' types.



(a) Hierical Clustering with Average clustering method

(b) Hierical Clustering with McQuitty clustering method

Figure 4: Two Plots representing HC: each branch of the tree represents a sample The two clusters were selected automatically by the algorithm. The inner samples highlighted at the bottom were manually drawn.

4.2.3 Random Forest

Moving to the unsupervised clustering methods, RF was performed.

Figure 5 represents how important is a gene regarding its classification. Notably, the vast majority of the genes seem to be not relevant according to the RF. Only 1234 genes out of more than 50000 have an importance value different from zero. In other words, these results indicate that these 1234 genes are predicted by the RF to be the most predictive for the outcome variable, while other genes don't contribute to the model's performance or are noisy, hence they have zero importance.

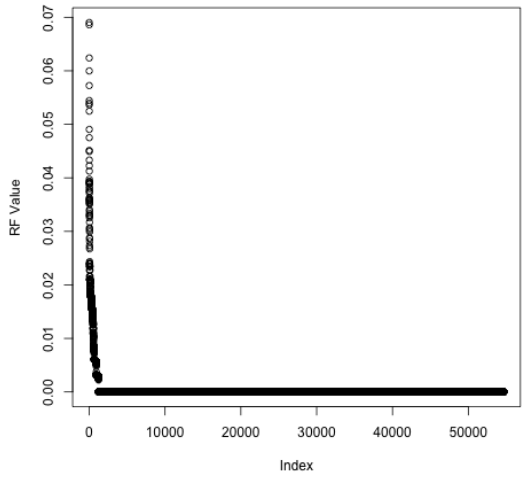


Figure 5: Representation of the values computed by RF for each gene. In the X axis, the indexes of the genes are reported, while the Y axis represents the gene's relative importance for the classification.

To sum it up, this result from the RF analysis indicates that the dataset has a highly relevant 1000 genes which likely are responsible for the pathways correlated with the disease onset or development.

Finally, the list of the first most relevant probe names was exported to perform some further testing. Figures 6 and 7 represent, respectively, a screenshot of a portion of the most relevent probes' and genes' list.

```
> head(top200, n = 10)
 [1] "242587_at"   "238896_at"   "1562903_at"  "244185_at"   "239057_at"   "239783_at"   "241415_at"   "215549_x_at"
```

Figure 6: List of the 200 top representative probes computed by RF

```
> head(top200_g, n=8)
[1] "C4orf33" "GPR78"   "SH2D1B"  "BRICD5"  "SETMAR"  "SMIM11"  "ITGA11"  "ACY3"
```

Figure 7: List of the 200 top representative genes computed by RF

4.2.4 LDA

To evaluate LDA performance in the training and testing part, the projections of the LDA algorithm results were plotted. In the upper part of figure 8, the classification of the control group is represented, while in the lower part the classfication of samples with PCOS is shown.

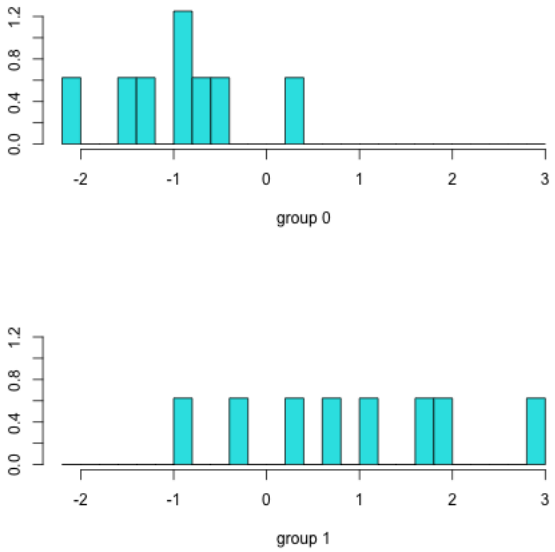


Figure 8: Graphical representation of the LDA results. The upper bar plot represents the control samples while the bar plot below represents the affected samples. The origin of the X-axis is considered as the point of the division between the two groups.

Overall, LDA performed a great classification with only a few outliers. For instance, in the bar plot depicting the control sample, only one sample has a positive value, therefore not taking part in the main group; on the other hand, every other sample is correctly collected in the same group as it has all negative values. Analogously, in the other bar plot almost every sample has a positive value, while merely one sample has a negative value, thus not taking part in this group.

In figure 9 another plot showing the results is plotted. In this case, the black line (indicating the value 0 of the test before) is the threshold that splits the two samples' groups. As described in the previous graph, in both groups a single sample is miscalculated.

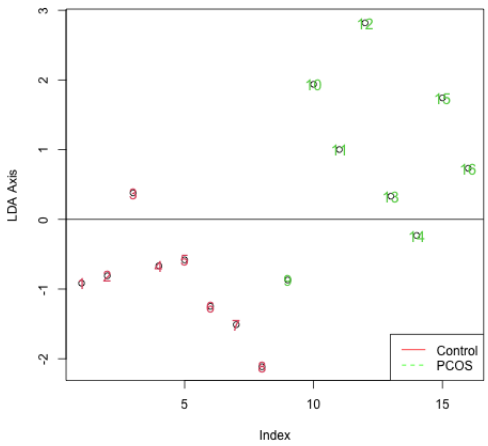


Figure 9: Alternative graphical representation of the LDA results. To enhance visualization, a black line has been added where the Y axis is 0. Also, colors and labels have been added.

Lastly, to test the accuracy of LDA the area under the ROC curve has been calculated and plotted (figure 10). Compared to a random model, LDA performs significantly better. The area under the experimental curve is significantly higher when compared to a random model. Specifically, figure 10 shows that the true positive rate is good (greater than 0.8) while maintaining a good false discovery rate (0.8).

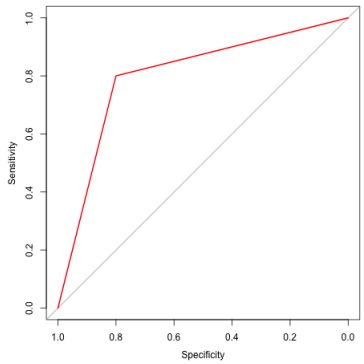


Figure 10: Plot of the area under the ROC curve, where the X axis is the false positive rate (or specificity) while the Y axis represents the true discovery rate (or sensitivity). The red bar represents the LDA result which can be compared to the grey bar, representing a random model.

4.2.5 Linear Regression Analysis

Before analyzing the performance of LRA, its efficiency when using different regularization factors, or lambda, was plotted. According to the plot in figure 11, the best lambda value is around -1, since it minimizes the binomial difference.

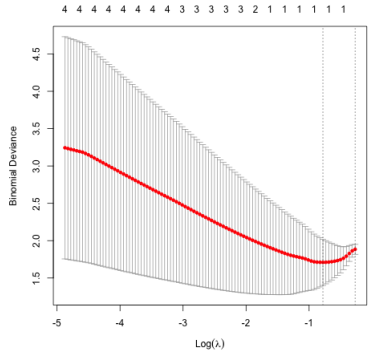


Figure 11: Plot of the binomial deviance (Y-axis) after performing LRA at different lambda values (X-axis). The red line represents the median while the area between the two dotted grey lines represents the lambda value in which the binomial deviance is at its lowest.

4.2.6 Caret

To make results from RF, LDA, and LRA comparable, Caret was performed.

Results in figure 12 show that RF and LRA are the best methods for supervised learning of this particular dataset. In particular, these two methods have the same mean (0.79) while LDA has a 6% worse accuracy (0.73). Overall, all three methods showed good accuracy, making them valid tools to perform supervised clustering.

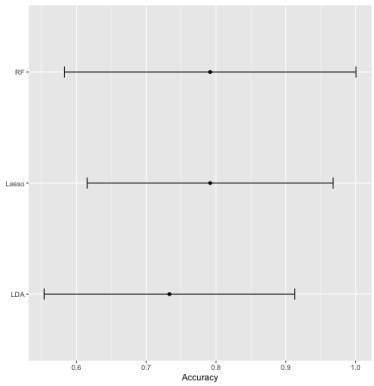


Figure 12: Plot showing the mean of the accuracy in the X axis of RF, LDA, and LRA methods, labeled in the Y axis.

4.3 Functional Analysis

While results of the functional analysis seems too general overall, the second result in figure 13 provides some room for discussion.

	Category	Term	Count	PValue	Genes	List.Total	Fold.Enrichment	Bonferroni	Benjamini
1	GOTERM_MF_DIRECT	GO:0005524~ATP binding	26	0.000299691288703253	236858_S_AT, 210024_S_AT, 220246_AT, 225611_AT, 225999_AT, 208941	136	2.1609488158743	0.0865442404459209	0.0905067691883824
2	WIKIPATHWAYS	WP4541:Hippo Merlin signaling dysregulation	7	0.000735080752641324	207172_S_AT, 202647_S_AT, 1554819_A_AT, 215305_AT, 210844_X_AT, 2(	74	6.33543289051764	0.176451370622008	0.194796399449951
3	REACTOME_PATHWAY	R-HSA-1236394~Signaling by ERBB4	5	0.00140785590807363	200633_AT, 202647_S_AT, 200665_S_AT, 228554_AT, 211373_S_AT	88	10.1204937304075	0.701862449138893	0.426741022391117
4	REACTOME_PATHWAY	R-HSA-9682385~FLT3 signaling in disease	4	0.00162299561381943	200633_AT, 202647_S_AT, 206674_AT, 202778_S_AT	88	16.7711039861039	0.752237786714362	0.426741022391117
5	REACTOME_PATHWAY	R-HSA-9683701~Translation of Structural Proteins	4	0.00179931581574465	200633_AT, 201998_AT, 238896_AT, 208761_S_AT	88	16.1927899686852	0.78711527954101 5	0.426741022391117
6	REACTOME_PATHWAY	R-HSA-8866652~Synthesis of active ubiquitin: roles of E1 and E	4	0.00198715260717633	200633_AT, 201343_AT, 210024_S_AT, 222657_S_AT	88	15.6530303030303	0.818892108341596	0.426741022391117
7	GOTERM_MF_DIRECT	GO:0005515~protein binding	114	0.00395800896099803	236858_S_AT, 201998_AT, 207159_X_AT, 222216_S_AT, 208031_S_AT, 21(	136	1.13714731268586	0.698109378529916	0.597659353110702

Figure 13: Screenshot of a table showing the top seven terms resulted to be annotated with the input set of genes

The Hippo pathway, a highly conserved pathway that regulates organ size control, plays an important role in govering ovarian physiology and pathology. In particular, the Hippo merlin signaling plays a key role in follicole growth and activation. Disregulation of Hippo pathway contributes to loss of follicolar homeostasis and reproductive disorders such as PCOS [2]. In addition, the inhibition of Hippo signal transduction provides a treatment for ovarian disorders, although the underlying mechanism remains elusive [4].

4.4 Enrichment Analysis

From the enrichment analysis in figure 14, we can see that the most enriched pathways are the taurine and hypotaurine metabolism. With the current state of the art, a direct correlation between taurine and PCOS has not been found yet. However, the supplementation of taurine showed promising effects in PCOS management, in particular, protective effects against diabetes and cardiovascular disease have been proven [3] [6].

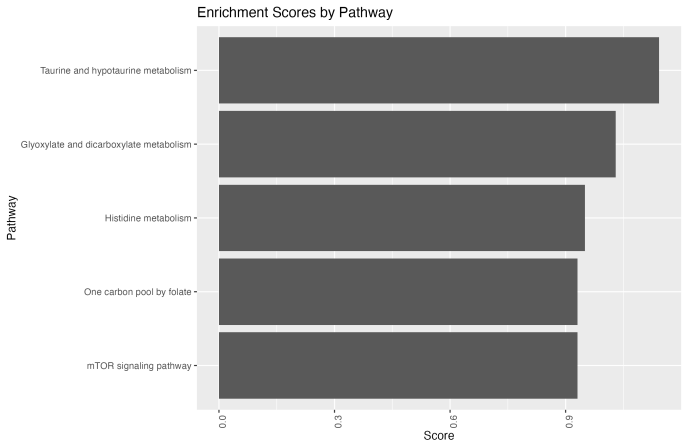


Figure 14: Plot showing the top five metabolic pathways (in the Y-axis) with the highest enrichment score (X-axis)

5 Discussion

To sum it up, this report provides some insight regarding the mechanism underlying PCOS. By finding the optimal clustering method, we can enhance the samples analysis and grouping steps in terms of elapsed time and accuracy. The latter value gives a great overall view when comparing all the methods. In fact, *Caret* function showed that RF and LRA achieve slightly higher accuracy compared to the other methods. Subsequently, functional analysis with *DAVID* tools revealed some already known pathways in this field, even though the direct correlation with PCOS have still to be unveiled. In addition, enrichment analysis showed some novelty regarding the possible pathway involved into PCOS pathogenesis and development. These new pathways could be further analyzed, both in *in silico* and *in vitro* to test the correlation with the disease and, eventually, to develop some new drugs which could help in the disease managment.

Despite some graphs are not provided probably due to some incompatibility with the dataset, some known issues that have been addressed are:

- **Optimization of clustering methods**: due to the dataset specificity, finding the optimal clustering methods with the optimal parameter was not easy to perform.
- **Metadata integration**: since only the sample type was given (control or with PCOS), some other inferences could not be made, thereby leaving the clustering analysis with unanswered questions
- **Enrichment Validation**: clearly, finding some novelty in a well-known topic such as PCOS is always great news. However, the correlation be-

tween the newly found pathways and the disease need to be further tested in order to prevent the research of false positives.

Results from this report could indicate that these methods could be used not only for the drug discovery and development but also for the personalized medicine approaches. Indeed, by targeting specific molecular pathways, associated with the individual patients' disease, some new protocols can be developed that are more effective and have less side-effects. Results also indicate that this type of analysis could be exploited to build predictive tools that could accelerate the diagnosis by looking, for instance, at the gene expression level.

References

[1] G. A. Burghen, J. R. Givens, and A. E. Kitabchi. Correlation of hyperandrogenism with hyperinsulinism in polycystic ovarian disease. *J Clin Endocrinol Metab*, 50(1):113–116, Jan 1980.

[2] K. L. Clark, J. W. George, E. Przygrodzka, M. R. Plewes, G. Hua, C. Wang, and J. S. Davis. Hippo Signaling in the Ovary: Emerging Roles in Development, Fertility, and Disease. *Endocr Rev*, 43(6):1074–1096, Nov 2022.

[3] Hariprasath Gopalakrishnan, S Sakila, K Kumari, and Subramaniam Sethupathy. Taurine supplementation improves insulin sensitivity and lipid profile in pcos women. 7:208–210, 01 2018.

[4] S. P. Wang and L. H. Wang. Disease implication of hyper-Hippo signalling. *Open Biol*, 6(10), Oct 2016.

[5] S. Yadav, O. Delau, A. J. Bonner, D. Markovic, W. Patterson, S. Ottey, R. P. Buyalos, and R. Azziz. Direct economic burden of mental health disorders associated with polycystic ovary syndrome: Systematic review and meta-analysis. *Elife*, 12, Aug 2023.

[6] T. A ZYKOVA, L. V ULEDEVA, A. V STRELKOVA, and L. B KOPTYAEVA. Clinical and metabolic effects of taurine in reproductive age women with polycystic ovarian syndrome. *Obstetrics and Gynecology*, (1):89–93, 2013.