Differential gene expression analysis of muscle cells from patients with and without type 2 diabetes

Network-based data analysis report.

Alessandro Pilli

September 2023

**Abstract**

We conducted an analysis of differential gene expression by employing both supervised and unsupervised statistical learning techniques. This investigation aimed to discern alterations in gene-level expression when comparing patients with type 2 diabetes mellitus to healthy subjects. The dataset used contains mRNA expression data throughout human muscle cell differentiation (myoblasts and myotubes) for both healthy and diabetic subjects. In total, the dataset is composed of 52 samples divided into 2 groups (26 subjects per group). The study resulted in the identification of processes activated by the muscle cells in response to pathological stress and of specific genes, in the different development stages, likeATP6, COX4I1, and AQR, that are possible candidates as biomarkers, whereas RAB8A and HSP90AA1may represent potential drug targets.

**Introduction**

Type 2 diabetes (T2D) is a serious and common chronic disease resulting from a complex inheritance-environment interaction along with other risk factors such as obesity and a sedentary lifestyle. Type 2 diabetes and its complications constitute a major worldwide public health problem, affecting almost all populations in both developed and developing countries with high rates of diabetes-related morbidity and mortality [1].

Its presence often leads to a decline in the differentiation capacity of myoblasts and progressive loss of muscle mass, which in turn results in the deterioration of skeletal muscle function.  However, effective therapies against skeletal muscle diseases are unavailable [2].

Due to its widespread prevalence, it becomes imperative to employ quantitative methodologies for investigating occurrences at the molecular and genetic level. Accordingly, we conducted an assessment of gene expression utilizing DNA microarray technology, encompassing a multitude of genes. This analysis is aimed to identify the genes that undergo deregulation within the muscular framework in patients with T2D. Identification of gene expression variations across diverse samples, may lead to pinpoint potential biomarkers associated with the pathology, as well as potential targets for novel drug development at the molecular level. Accomplishing this, involves the utilization of statistical classification techniques, functional enrichment analysis, and network-based approaches.

Inizio modulo

**Methods**

The dataset was retrieved from the Gene Expression Omnibus with GEO accession GSE166502 [3], containing oligonucleotide microarray according to the platform Illumina HumanHT-12 V4.0 expression beadchip.  
This dataset refers to undifferentiated muscle stem cells (myoblasts) isolated from human vastus lateralis muscle biopsies and to in vitro differentiated myoblasts. To induce differentiation, myoblasts were incubated for 7 days in two specific differentiation media, until cells were fully differentiated, as determined by visual confirmation of myotube formation.  
The 52 analyzed samples were obtained from patients affected by type 2 diabetes (n=26, “T2D”) and from normal glucose tolerant subjects (n=26, “NGT”). Further information about sampling and the microarray procedure can be found on the previously cited webpage of GEO [3].

Most of the analysis were carried out using RStudio [4]. Before performing procedures that involved randomness, *set.seed* function was always set to the value 1 and used to obtain reproducible results. Upon extraction of the data matrix, an initial set of 17,011 genes was identified, encompassing all 52 samples (excluding a solitary row containing NA values). The sole preprocessing step involved the utilization of the *scale* function from the R programming base, which facilitated normalization (the boxplot of the dataset can be seen in Supplementary Figure 1).

This project exploited a range of sophisticated methodologies to derive significant outcomes. The majority of these methodologies take advantage of machine learning techniques, encompassing both supervised and unsupervised approaches. A concise overview of the methods employed in the project will be presented, along with the corresponding packages and libraries utilized, in the following section.

The entire procedure was carried out to compare healthy subjects vs diabetic patients regarding to myoblasts and myotubes separately. Therefore, the initial step was to divide the dataset into the two parts of interest.

These unsupervised techniques were used: principal component analysis, k-means clustering on scaled data to mitigate the impact of outliers, and hierarchical clustering (all facilitated by the base stats library [5]).

A feature selection was done by performing a *T-test row by row* pairwise each group filtering for p-value < 0.05, in this way we restricted the number of significant genes in the two datasets.

Before conducting k-means and hierarchical clustering, the Elbow Method was employed to ascertain the most suitable cluster count. This method involves computing the Within-Cluster-Sum of Squared Errors (WSS) for varying cluster counts (k ranging from 1 to 10) and identifying the point at which the decline in WSS starts to plateau [6]. Analysis of the plot revealed the optimal cluster count to be k=2. For the hierarchical clustering process, the Euclidean distance metric and complete linkage method were selected.

Following identification of the optimal cluster count, the dataset has been divided into training and validation subsets using the *createDataPartition* function from the caret library [7], where 75% of the samples are allocated to the training set.

The supervised methods encompassed the utilization of random forest (via the randomForest library [8]) and linear discriminant analysis. These techniques were performed with repeated cross-validation (10-fold, repeated 10 times), employing the "Accuracy" metric (from the "MASS" library [9]). Additionally, Lasso and Ridge regressions were conducted using the glmnet library [10] under the "binomial" family setting. For both segments of the experiment, a suitable shrinkage parameter lambda was identified as 0.4 (Supplementary Figure 2).

The Signature-based Clustering for Diagnostic Purposes (SCUDO) was done utilizing the rSCUDO library [11]. Here, the dataset was split with 75% of samples allocated to the training set. The model was trained using *scudoTrain*", validation was performed using *scudoTest*, and classification was carried out using *scudoClassify*. The latter included parameter values nTop=25, nBottom=25, and N=0.5, necessary for generating a connected graph. The resultant network was visualized using the Cytoscape tool [12] (Supplementary Figure 3).

Functional enrichment analysis was conducted using the g:Profiler tool [13]. This involved applying Bonferroni correction and setting the p-value threshold to 0.05. The gene list employed in this analysis was derived from the Ridge regression results, with gene names initially extracted from the experiment's metadata.

Furthermore, Network-based Analysis was undertaken utilizing the pathfinder function from the pathfindR library [14], with the parameter "iterations" set to 5. This procedure aimed to identify genes displaying up or down regulation between groups. The input data frame for pathfindR encompassed gene names and p-values (obtained from T-tests). To expand the gene networks, we utilized the online resource String [15].

**Results**

The principal component analysis (PCA) wasn't performing well before cleaning the data (Supplementary Figure 4) but afterward, a good distinction between the two groups, in the 2D space, in both experiments was observed (Fig.1). Actually, the first 2 dimensions explain almost 40% of the variance within the myoblast subset, and 37% of the variance in the myotube subset.

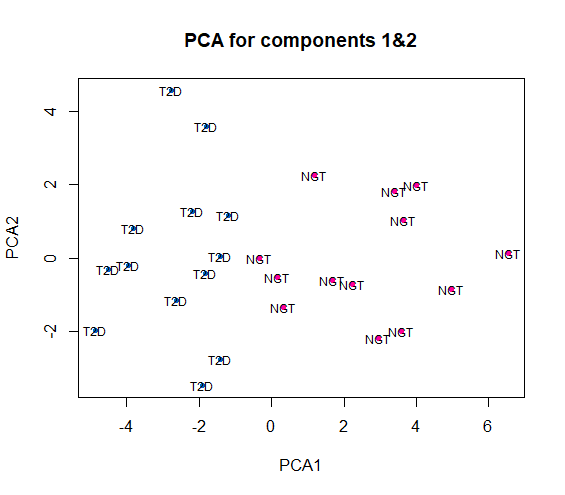
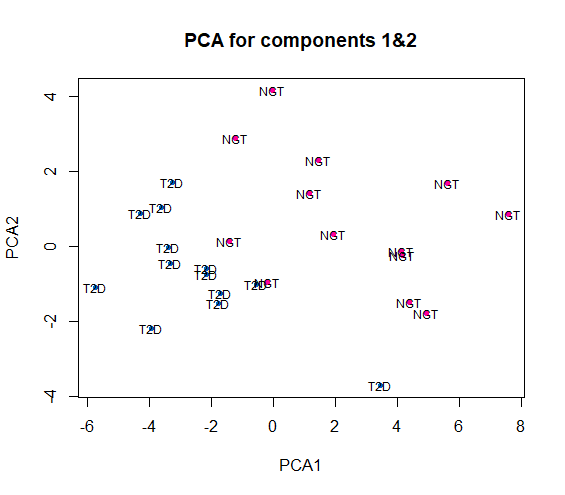


Fig.1: **PCA of the two sub-datasets.** The filtering highlighted the presence of two groups since the two colors are spread apart in both sub-sets

**Myoblast**

**Myotube**

The k-means performed well in the myoblast section, with only one misclassification, but it was a bit more imprecise regarding the myotube, misclassifying 5 samples (Fig.2).

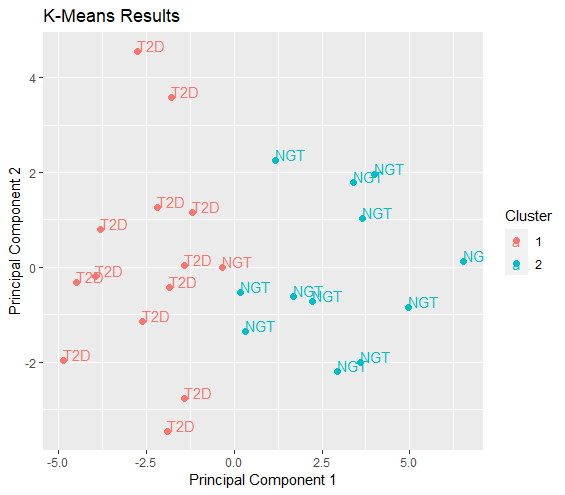
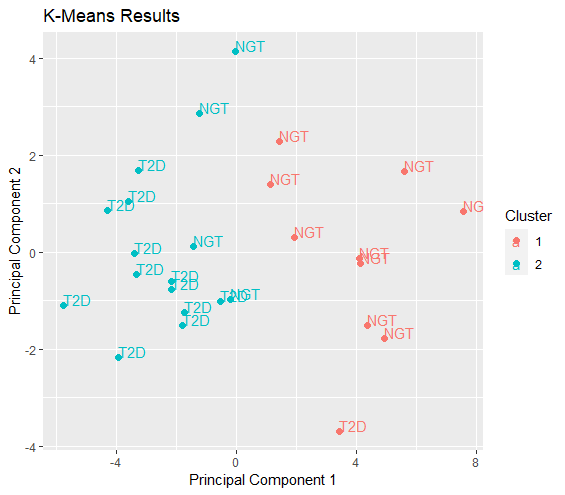


Fig.2: **K-mean of the two sub-datasets.** Erroneously classified samples are circled in red.

**Myotube**

**Myoblast**

As well as the k-means, the hierarchical clustering didn't separate efficiently the groups of the myotube section, leading to the construction of an erroneous group as reported below in Fig.3.

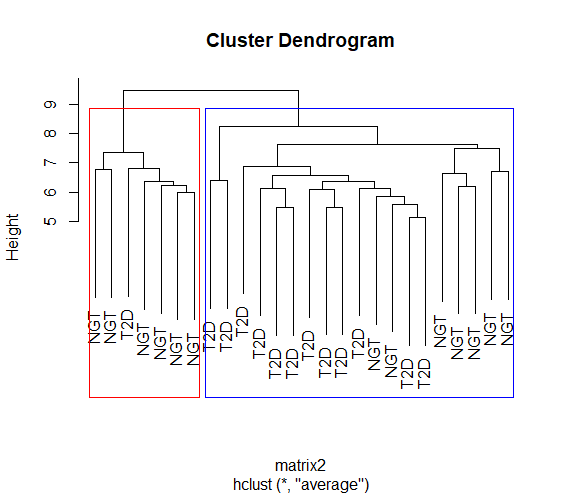
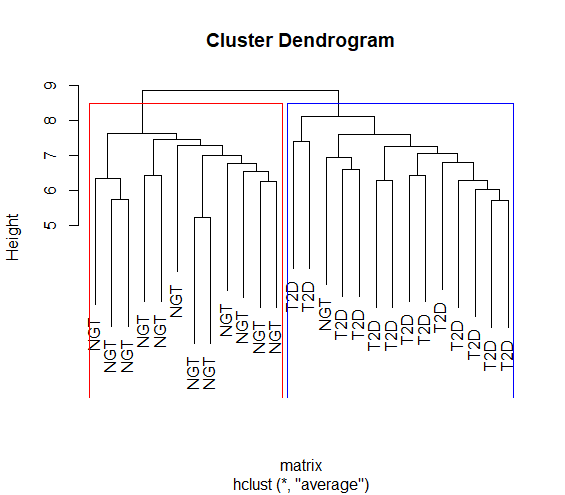


Fig.3: **Hierarchical clustering.** Within the myotube, the NGT are wrongly clustered, forming a small subcluster visible on the left of the corresponding dendogram (red square).

**Myoblast**

**Myotube**

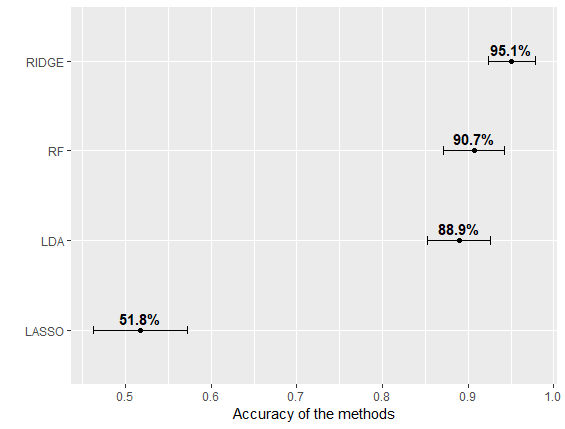
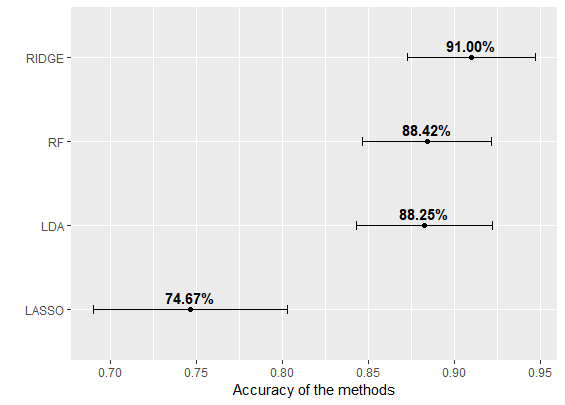
From the random Forest (RF) analysis, we retrieved that only the first 200 genes seem to be relevant in the myoblast part, and only 150 in the myotube, however this model is not the best performing one (the mean error between the two groups can be seen in Supplementary Figure 5).

Performing Linear Discriminant Analysis (LDA), we obtain a slightly worse separation between the groups (data not reported).

Ridge regression is the model that performs the best group separation and has the highest accuracy for both experiment parts (95.1% myoblast, 91 % myotube)), whereas Lasso regression performed badly (all the accuracy values are reported in Fig.4).

Finally, SCUDO analysis, yielded a good result with an accuracy of 91%, but it was heavily influenced by the seed and changed drastically to lower levels trying different ones (data not shown).

Fig.4: **Accuracy plot.** Average accuracy for each supervised method performed.



**Myoblast**

**Myotube**

By using Ridge regression, we retrieved a list of the top 200 genes per importance per sub-dataset and on these we performed the functional enrichment analysis via gprofiler which returned many main terms, the most significant have been highlighted below in the Manhattan plots (Fig.5).

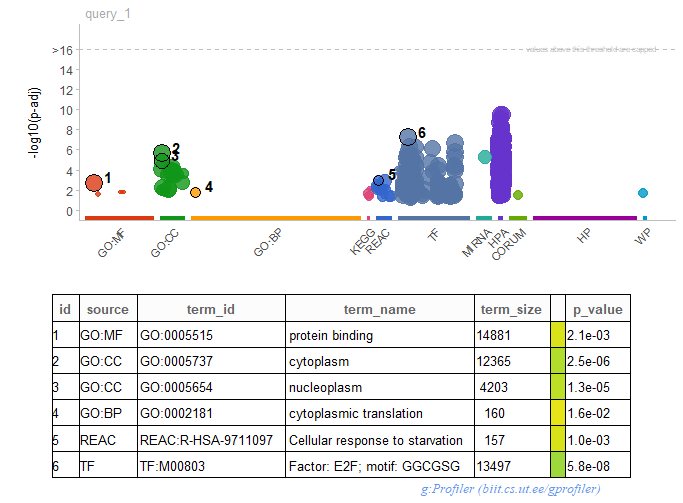
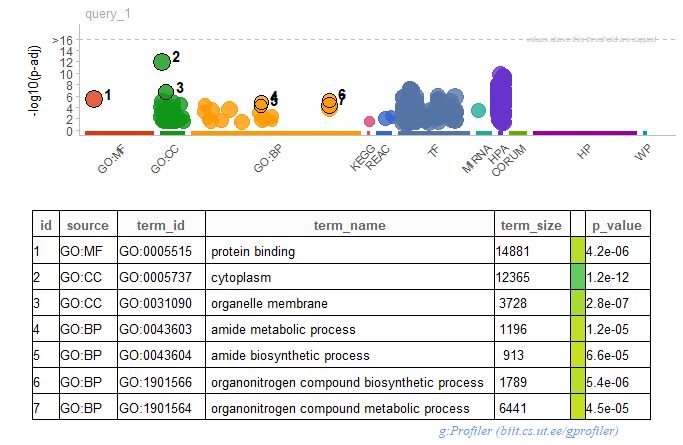


Fig.5: **Manahattan plots.** Identification of specific terms via functional enrichment analysis.

**Myoblast**

**Myotube**

In both the experimental parts, we can observe that a big issue is linked to the molecular function (GO:MF) related to protein binding (Fig.5) that is probably altered by the pathology.   
Focusing on the myoblast section the most significant biological process (GO:BP) involves organitrogen compounds, and in particular amides that in muscles contribute to energy storage (by creatine) and buffering (by carnosine).   
Looking, instead, at the myotube part we can observe that the only biological process (GO:BP) reported is the cytoplasmic translation (Fig.5). We also highlighted a transcription factor (TF) related to the E2F family, which is a big family of TFs linked generally to the promotion of cell cycle and cell proliferation. In literature, it has been demonstrated that, in mice, E2F1 and E2F2 activity negatively controls the growth of mature pancreatic cells and is necessary for the maintenance of differentiated pancreatic phenotypes in the adult [16].

It is probably important to note the presence, in the myotube section, of the term “Cellular response to starvation”, which is expected to be present in a situation where the energy supply is lacking due to the pathology.  
And lastly, in both myoblast and myotube, these processes are allocated in the cytoplasm (GO:CC).

In the subsequent and final phase, we carried out Network-based analysis using both pathfindR and STRING. These tools enabled us to unveil novel potential interactors and enriched pathways by encompassing neighboring genes associated with the input gene list. This process involves exploitation of connections, documented in the scientific literature, and execution of analyses on an augmented gene list.

Starting from pathfindR results (Fig.6) on the myoblast side, we can observe that the most interesting results are linked to the nucleocytoplasmic transport, ribosome, and protein processing holding a higher significance with respect to the others. However, they are not particularly linked to the previous results from gprofiler (fig.5), and the part that takes in account the myotubes result in a nonspecific collection of pathological linked terms.

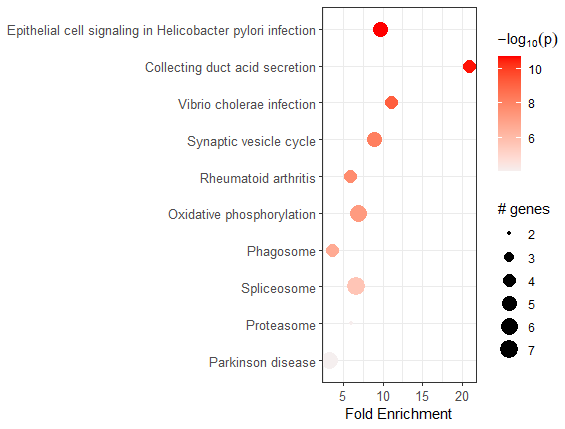
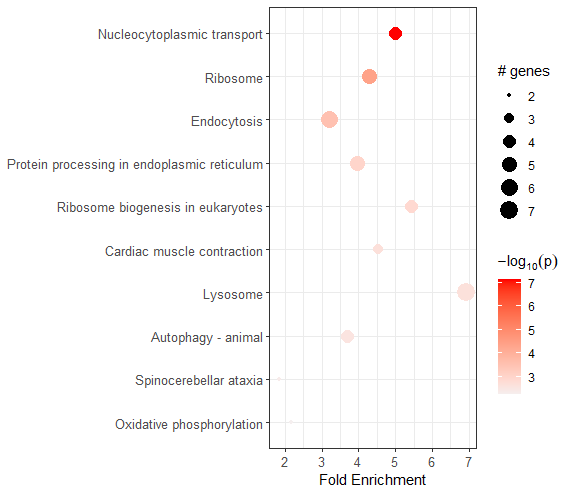


Fig.6: **pathfindR results.**

**Myoblast**

**Myotube**

One of the main terms, that stands out from pathfinR analysis, is Oxidative phosphorylation, present in both experimental parts and more significantly in the myotube which represent already developed cells. This could be due to the presence of the pathology, that limits the amount of supplementary energy resources, essential for the proper functioning of adult muscle cells.

Via the online tool STRING we analyzed the upregulated genes resulting from the pathfindR analysis, to identify potentially enriched network (Fig.7).

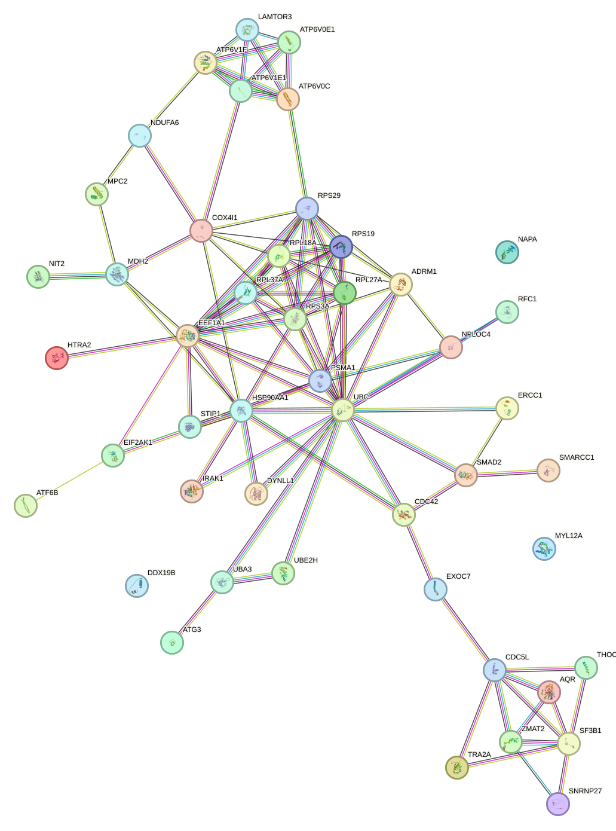
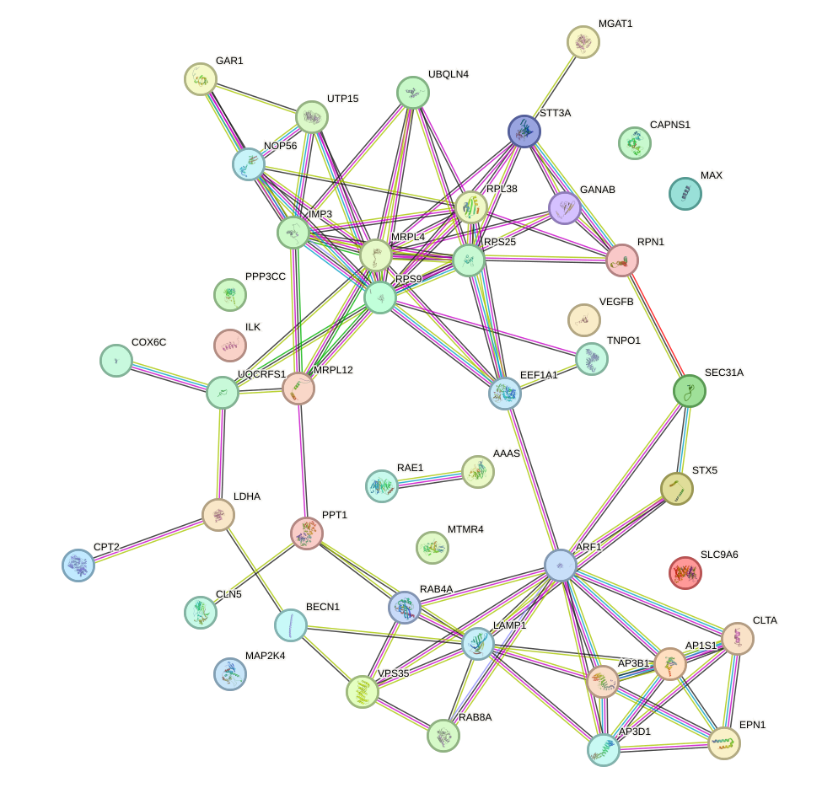


Fig.7: **String output.** Network reporting the connection between proteins.

**Myoblast**

**Myotube**

1. **3**

**4**

**2 5**

From STRING analysis the resulting network showed two main clusters (1 and 2) regarding the myoblast, cluster 1 including mainly ribosomal proteins, and cluster 2 containing molecules involved in protein sorting and transport via clathrin and non-clathrin mechanisms which finds a reference in the previous result with pathfindR (Fig.6) about the ribosome the nucleocytoplasmic transport and the protein processing since some of the proteins like AP3B1, AP3D1, AP1S1 are also involved in protein sorting in the Golgi network.   
Looking at the myotube network, three main clusters (3-4-5) have been identified; number 3, a small cluster of V-type ATPase subunit, number 4 mainly ribosomal proteins, and number 5 a small cluster of protein involved in the pre-mRNA splicing.

Analyzing these clusters, we found specific genes already reported in literature.  
From cluster 1:

* **VPS35**: The study demonstrated that high glucose regulates the Golgi apparatus stress through NLRP3/VPS35/Golph3/Vimentin pathway. [17]
* **RAB8A**: The first article about this gene refers to a different type of diabetes,Gestational Diabetes Mellitus (GDM), the data suggests that miR-30d-5p expression is down-regulated in placental tissue from GDM patients and affects trophoblast cell functions by targeting RAB8A, which may provide new insight into the pathogenesis of GDM, andwith a more aimed study RAB8A may also be related to type 2 diabetes [18]. The second article states that RAB8A regulates lipid uptake and storage in skeletal muscle so RAB8A deficiency decreases long-chain fatty acids entry into skeletal muscle and inhibits lipid droplets fusion in muscle cells. Consequently, blood lipid levels are elevated, and can be directly related to diabetes since one of the main predisposing factors linked to the pathology is obesity. [19]
* **CLTA**: gene polymorphisms of CTLA, including CTLA-4+49A/G and CTLA4-318C/T, are important predictors of type 2 diabetes mellitus. CTLA-4 may be a susceptibility gene for the pathology. [20]

From cluster 3:

* **ATP6:** mutation of m.9053G>A at the *ATP6* gene was found in patients with type 2 diabetes mellitus. The molecular docking suggests that water binding on the proton translocation channel in the S167N mutant was different from the wild type. The result of this study is hoped to be useful in the development of a new genetic marker for type 2 diabetes. [21]

From cluster 4:

* **HSP90AA1:** Fifteen hub targets of 1,25(OH)2D against COVID-19 associated with diabetes mellitus were identified, including EGFR, PIK3R1, PIK3CA, STAT3, MAPK1, ESR1, HSP90AA1, LCK, MTOR, IGF1, AR, NFKB1, PIK3CB, PTPN1, and MAPK14. Although HSP90AA1 is an already known molecular target, further investigations are needed to evaluate its efficacy in the case of type 2 diabetes not associated with SARS-COV-2 infection. [22]
* **COX4I1:** Cytochrome oxidase (COX) dysfunction is associated with mitochondrial oxidative stress. In the study, the authors determined the association between COX expression, obesity, and type 2 diabetes, and concluded that COX4I1 depression is related to insulin resistance and type 2 diabetes in obesity. In peripheral blood monocytes, it may be a useful diagnostic biomarker. [23]

From cluster 5:

* **AQR:** The authors previously demonstrated AQR as a susceptibility gene for type 2 diabetes mellitus (T2DM) and showed that it was increased in multiple tissues in animal models of T2DM or metabolic syndrome.In this paper, they demonstrated that the knockdown of the gene PLAU rescued senescence-related phenotypes, endothelial cell activation, and inflammation in models induced by AQR. These findings, for the first time, indicate that AQR/PLAU is a critical signaling axis in the modulation of endothelial cell senescence, revealing a novel link between hyperglycemia and vascular dysfunction. This may impact the muscle structure, limiting the quantity of energetic available supply.[24]

**Conclusion**  
Regarding the statistical methodologies employed, it can be observed that, on the whole, the supervised techniques yield superior results compared to the unsupervised ones since the classification resulted with an overall higher accuracy. Notably, Ridge regression stands out as the method demonstrating the highest accuracy.  
Regarding the specific processes we can say that a probable reaction, to diabetes, can be the overexpression of ribosomes and ATP producing processes to sustain the adult myotube in response to the lack of nutrients, and in case of persistence of the problem this may also lead to events that will lead to deterioration of the muscle and, later, skeletal structure.   
Particular genes both in myoblast and myotubes have been retrieved, identified as up-regulated by pathfindR and, since already known in the literature, are possible valid candidates as biomarkers to evaluate the presence of type 2 diabetes possibly in the early stages or even before the development of the pathology.Prior to considering these genes as potential biomarkers, an in-depth exploration of their molecular mechanisms and regulatory pathways is warranted.  
As a final note, all the genes found in our analysis are not reported in the published article from which the original dataset was derived [25], this could be due to the different statistical approaches and tools.

**References**

[1] Wu Y, Ding Y, Tanaka Y, Zhang W. “Risk factors contributing to type 2 diabetes and recent advances in the treatment and prevention.” Int J Med Sci. 2014 Sep 6;11(11):1185-200. doi: 10.7150/ijms.10001. PMID: 25249787; PMCID: PMC4166864.

[2] Jiang X, Feng N, Zhou Y, Ye X, Wang R, Zhang J, Cui S, Ji S, Chen Y, Zhu S. Slc2a6 regulates myoblast differentiation by targeting LDHB. Cell Commun Signal. 2022 Jul 18;20(1):107. doi: 10.1186/s12964-022-00915-2. PMID: 35850889; PMCID: PMC9290262.

[3] GEO. url: <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE166502>

[4] “RStudio Team. RStudio: Integrated Development Environment for R”. In: (2023).

[5] “R Core Team. R: A Language and Environment for Statistical Computing”. In: (2023).

[6] Purnima Bholowalia and Arvind Kumar. “EBK-means: A clustering technique based on elbow  
method and k-means in WSN”. In: International Journal of Computer Applications (2014).

[7] Max Kuhn. “Classification and Regression Training”. In: (2023).

[8] Andy Liaw and Matthew Wiener. “Classification and Regression by randomForest”. In: R news  
(2003).

[9] W. N. Venables and B. D. Ripley. “Modern Applied Statistics with S. Springer”. In: (2002).

[10] Trevor Hastie. “An introduction to glmnet”. In: (2023).

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

[12] G.-M.; Chen Ge B.-K.; Hu. “MSClustering: A Cytoscape Tool for Multi-Level Clustering of Biolog-  
ical Networks”. In: (2022). doi: 10.3390/ijms232214240.

[13] gprofiler. url: <https://biit.cs.ut.ee/gprofiler/gost>

[14] Sezerman OU Ulgen E Ozisik O. “An R Package for Comprehensive Identification of Enriched  
Pathways in Omics Data Through Active Subnetworks”. In: (2019). doi: 10.3389/fgene.2019.  
00858.

[15] String. url: <https://string-db.org/>

[16] Iglesias A, Murga M, Laresgoiti U, Skoudy A, Bernales I, Fullaondo A, Moreno B, Lloreta J, Field SJ, Real FX, Zubiaga AM. Diabetes and exocrine pancreatic insufficiency in E2F1/E2F2 double-mutant mice. J Clin Invest. 2004 May;113(10):1398-407. doi: 10.1172/JCI18879. PMID: 15146237; PMCID: PMC406522.

[17] Li Y, Zhang H, Long W, Gao M, Guo W, Yu L. Inhibition of NLRP3 and Golph3 ameliorates diabetes-induced neuroinflammation *in vitro* and *in vivo*. Aging (Albany NY). 2022 Nov 15;14(21):8745-8762. doi: 10.18632/aging.204363. Epub 2022 Nov 15. PMID: 36378718; PMCID: PMC9699760.

[18] Zhang L, Li K, Tian S, Wang XQ, Li JH, Dong YC, Xia HF, Ma X. Down-regulation of microRNA-30d-5p is associated with gestational diabetes mellitus by targeting RAB8A. J Diabetes Complications. 2021 Aug;35(8):107959. doi: 10.1016/j.jdiacomp.2021.107959. Epub 2021 May 24. PMID: 34052089.

[19] Chen Q, Rong P, Xu D, Zhu S, Chen L, Xie B, Du Q, Quan C, Sheng Y, Zhao TJ, Li P, Wang HY, Chen S. Rab8a Deficiency in Skeletal Muscle Causes Hyperlipidemia and Hepatosteatosis by Impairing Muscle Lipid Uptake and Storage. Diabetes. 2017 Sep;66(9):2387-2399. doi: 10.2337/db17-0077. Epub 2017 Jul 10. PMID: 28696211.

[20] Ren D, He L, Pang X. Association of CLTA-4 Gene Polymorphisms with Diabetes Mellitus: A Study Based on the Han Population of Northern China. Diabetes Metab Syndr Obes. 2022 Sep 3;15:2705-2712. doi: 10.2147/DMSO.S374451. PMID: 36091483; PMCID: PMC9451031.

[21] Permana Maksum I, Saputra SR, Indrayati N, Yusuf M, Subroto T. Bioinformatics Study of m.9053G>A Mutation at the *ATP6* Gene in Relation to Type 2 Diabetes Mellitus and Cataract Diseases. Bioinform Biol Insights. 2017 Sep 12;11:1177932217728515. doi: 10.1177/1177932217728515. PMID: 28932107; PMCID: PMC5598796.

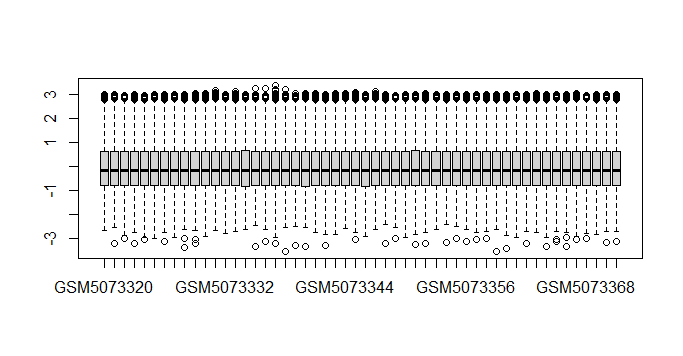
[22] Zeng F, Xu Y, Tang C, Yan Z, Wei C. Integrated bioinformatics and *in silico* approaches reveal the biological targets and molecular mechanisms of 1,25-dihydroxyvitamin D against COVID-19 and diabetes mellitus. Front Nutr. 2022 Dec 2;9:1060095. doi: 10.3389/fnut.2022.1060095. PMID: 36532549; PMCID: PMC9755336.

[23] Van der Schueren B, Vangoitsenhoven R, Geeraert B, De Keyzer D, Hulsmans M, Lannoo M, Huber HJ, Mathieu C, Holvoet P. Low cytochrome oxidase 4I1 links mitochondrial dysfunction to obesity and type 2 diabetes in humans and mice. Int J Obes (Lond). 2015 Aug;39(8):1254-63. doi: 10.1038/ijo.2015.58. Epub 2015 Apr 14. PMID: 25869607.

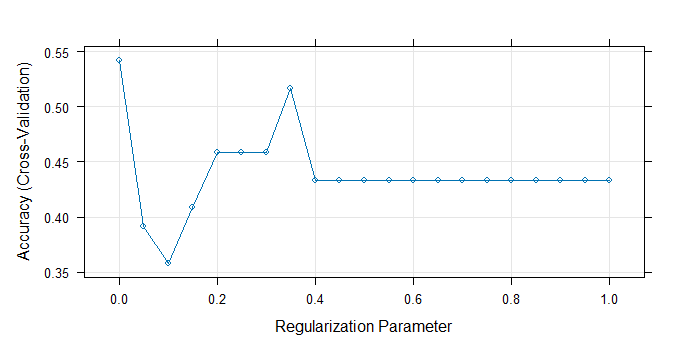
[24] Wan Y, Liu Z, Wu A, Khan AH, Zhu Y, Ding S, Li X, Zhao Y, Dai X, Zhou J, Liu J, Li Y, Gong X, Liu M, Tian XL. Hyperglycemia Promotes Endothelial Cell Senescence through AQR/PLAU Signaling Axis. Int J Mol Sci. 2022 Mar 7;23(5):2879. doi: 10.3390/ijms23052879. PMID: 35270021; PMCID: PMC8911151.

[25] Davegårdh C, Säll J, Benrick A, Broholm C, Volkov P, Perfilyev A, Henriksen TI, Wu Y, Hjort L, Brøns C, Hansson O, Pedersen M, Würthner JU, Pfeffer K, Nilsson E, Vaag A, Stener-Victorin E, Pircs K, Scheele C, Ling C. VPS39-deficiency observed in type 2 diabetes impairs muscle stem cell differentiation via altered autophagy and epigenetics. Nat Commun. 2021 Apr 23;12(1):2431. doi: 10.1038/s41467-021-22068-5. PMID: 33893273; PMCID: PMC8065135.

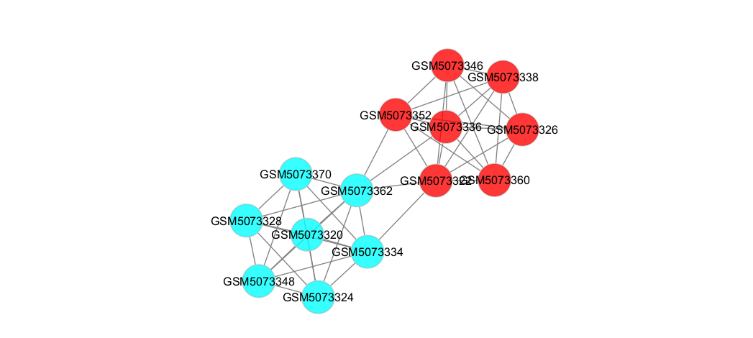
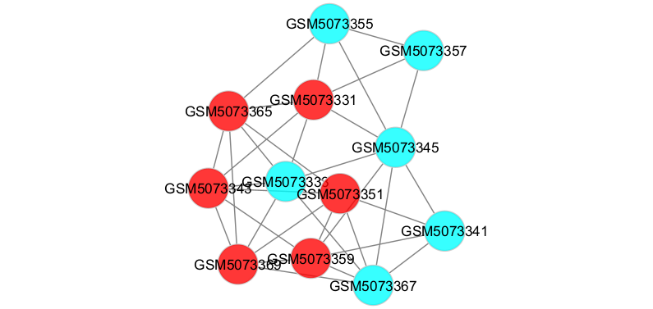
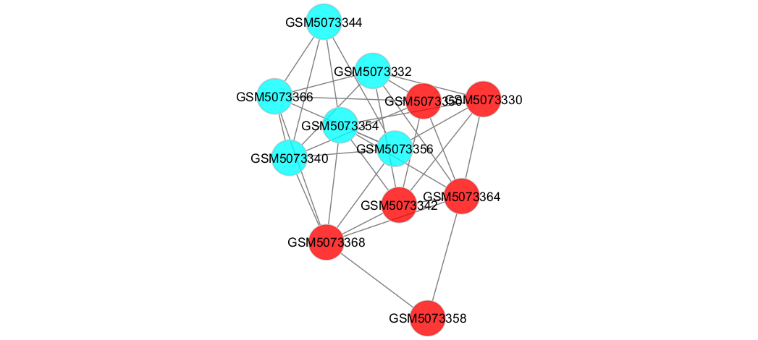
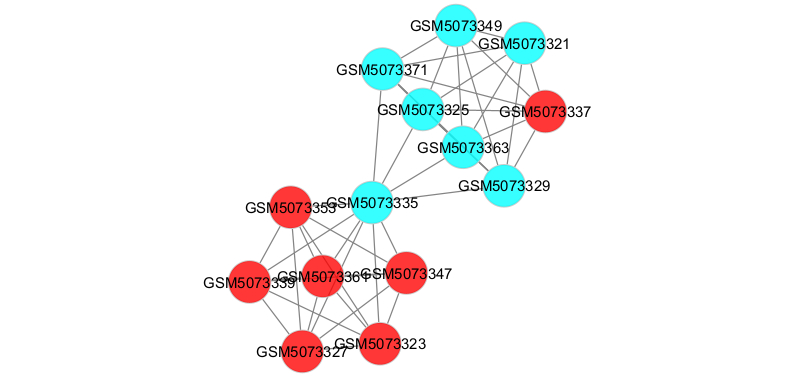
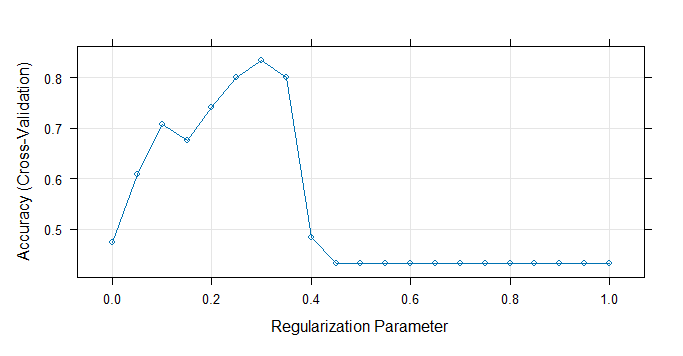
**Supplementary materials**



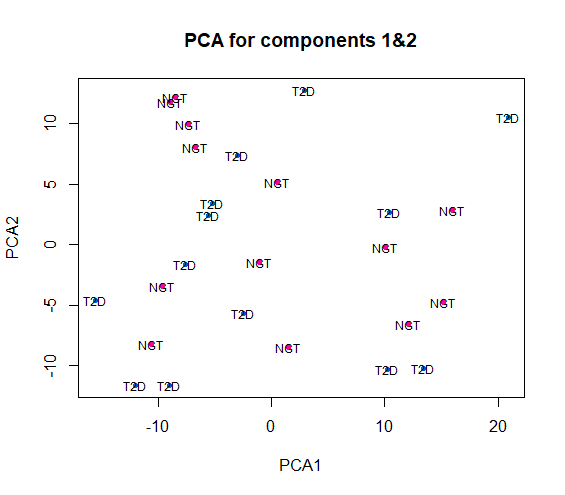
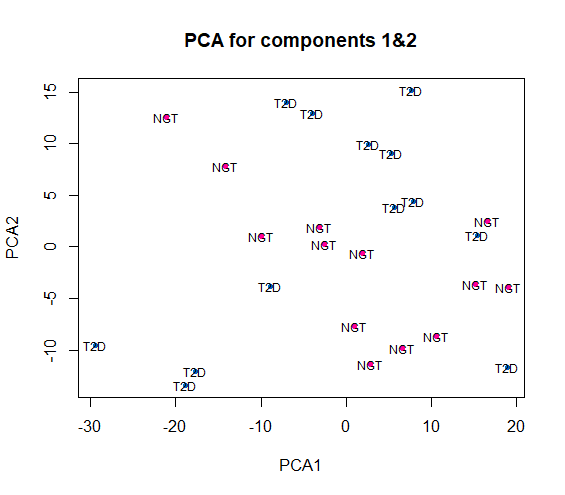
Sup.Fig1: **Boxplot**. Boxplot of the complete dataset after performing the scale function.



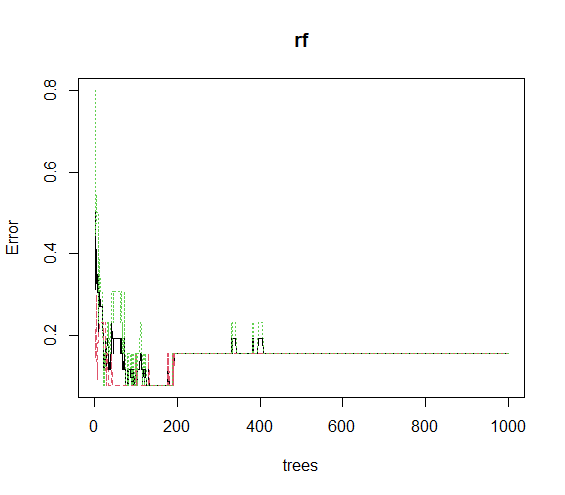
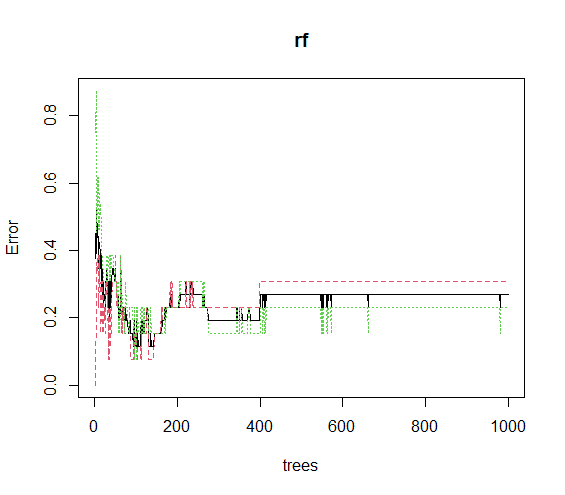
Sup.Fig2: **Lasso shrinkage parameter**. The optimal value of lambda for the two experimental parts, *0.4* for both before plateau.



Sup.Fig3: **Scudo results**. Visualization of scudo results through Cytoscape, on the left side the myoblast part on the right the myotube, on top the training part on the bottom the resulting testing.



Sup.Fig4: **PCA of unfiltered data**. PCA of the two sub-datasets regarding myoblast and myotube before performing the feature selection



Sup.Fig5: **Random Forest error**. Each color represents a group the black line is the mean.

Random forest (myoblast)

Random forest (myotube)