Down syndrome and congenital heart disease: RNA-seq reveals differentially expressed cardiac tissue related genes

Adriano Voltolini[1], Mario Lauria[2]

[1]Department of Cellular, Computational and Integrative Biology, University of Trento, Povo (TN), Italy

[2]The Microsoft Research - University of Trento Centre for Computational and Systems Biology, Rovereto (TN), Italy

INTRODUCTION

Down syndrome (DS), also known as trisomy 21, is the most common chromosome anomaly in humans. About 44.3% of people with DS are diagnosed with a type of Congenital Heart Disease (CHD), which is the main cause of death in this population during the first two years of life.
Three of the most common heart conditions seen in children with DS are atrioventricular septal defect, patent ductus arteriosus, and tetralogy of Fallot. The molecular causes for these cardiac defects are not yet entirely understood.

The objective of this study is to identify the molecular causes of CHD in people with DS through a RNA-seq approach.

MATERIALS AND METHODS:

Dataset:
The E-MTAB-10604 RNA-seq dataset[3] describes RNAs extracted from peripheral blood mononuclear cells (PBMCs) of people with DS, and of controls.

PBMCs separation was performed by using the Ficoll-Paque (Ficoll Plaque PLUS - GE Healthcare Life Sciences, Piscataway, USA) and the RNA was extracted using the TRIzol reagent (TRIzol Reagent, Invitrogen Life Tecnologies, Carlsband, CA, USA), according to the instructions provided by the manufacturer.

For RNA extraction, each whole blood sample (approx. 10 ml) was processed within 2 hours of collection and the RNA was stored at -80 °C.

Indexed libraries were prepared from 1 μg/ea purified RNA with TruSeq Stranded mRNA (Illumina) Library Prep Kit, according to the manufacturer's instructions.

Libraries were sequenced using an Illumina NextSeq 550 Dx System (Illumina) in a 2x75 paired-end format.

The read-count for the genes of interest was normalized, considering all genes expressed in the samples, using DESeq2, with standard parameter. using the median of ratio, to perform the differential expression analysis. In particular the counts were divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene.

Principal Component Analysis:

PCA was done using the built-in R function prcomp. The plot for the first 3 principal components was obtained using the package scatterplot3d.

K-means Clustering:

K-means clustering was done using the built-in R function kmeans and the obtained model was plotted using the fviz_cluster function of the R package factoextra. In this method, a K equal to 2 was chosen since it was expected to see two main clusters, one of controls and one of affected.

Hierarchical Clustering:

Hierarchical clustering was done using the built-in R function hclust and the obtained model was plotted using the fviz_dend function of the R package factoextra. In this method, a K equal to 2 was chosen since it was expected to see two main clusters, one of controls and one of affected.

Supervised learning:

Initial gene filtering was done with the R package genefilter. Genes with no statistically significant mean difference between the two groups have been ignored in all the supervised methods used (p-value > 0.1).

For all the supervised methods, the R package caret was used. All the models have been evaluated through 13-fold cross validation repeated 3 times.

The AUC-ROC curve plot was obtained using the pROC package, while the heatmaps have been built using the pheatmap, grid and RColorBrewer packages.

For SCUDO, an additional sample validation was done with optimal parameters in order to obtain the network diagram, using the igraph package.

DAVID

An initial functional enrichment analysis on the 200 most important genes for LDA classification was performed using DAVID (database for annotation, visualization and integrated discovery). The tool managed to identify 187 of the input genes. We used the default categories plus UP_TISSUE for tissue expression and CHROMOSOME. In this study we considered only the results with FDR below 0.05.

STRING

STRING was used to build a protein-protein interaction network on the peptides encoded by the top 200 RNAs, by setting the confidence to 0.5 and adding no more than 10 interactors. The obtained network was then exported to Cytoscape and the results of the DAVID analysis were color-coded onto it.

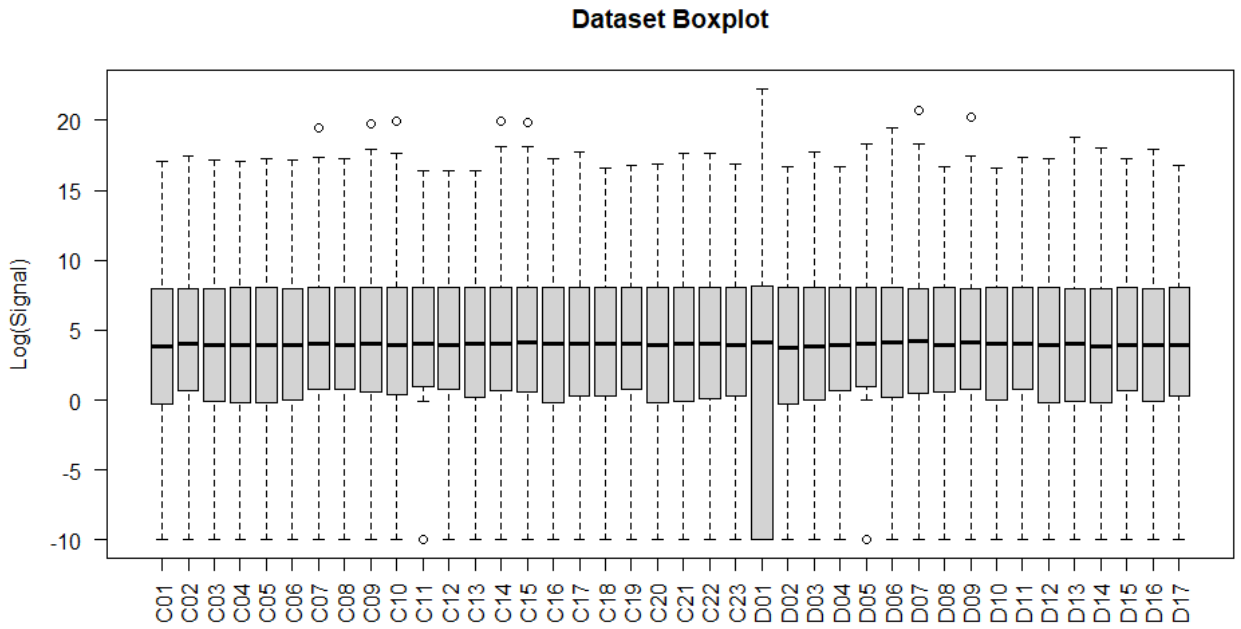
ENRICHNET

A tissue-specific network-based enrichment analysis was performed on the same 200 genes using Enrichnet and the Gene Ontology database. The tool managed to recognize 150 of the input genes. The results were exported in csv format and were filtered in R for pathways with cardiac tissue specific XD-scores above 2.5 (only AV-node and heart tissues have been considered).

PATHFINDR

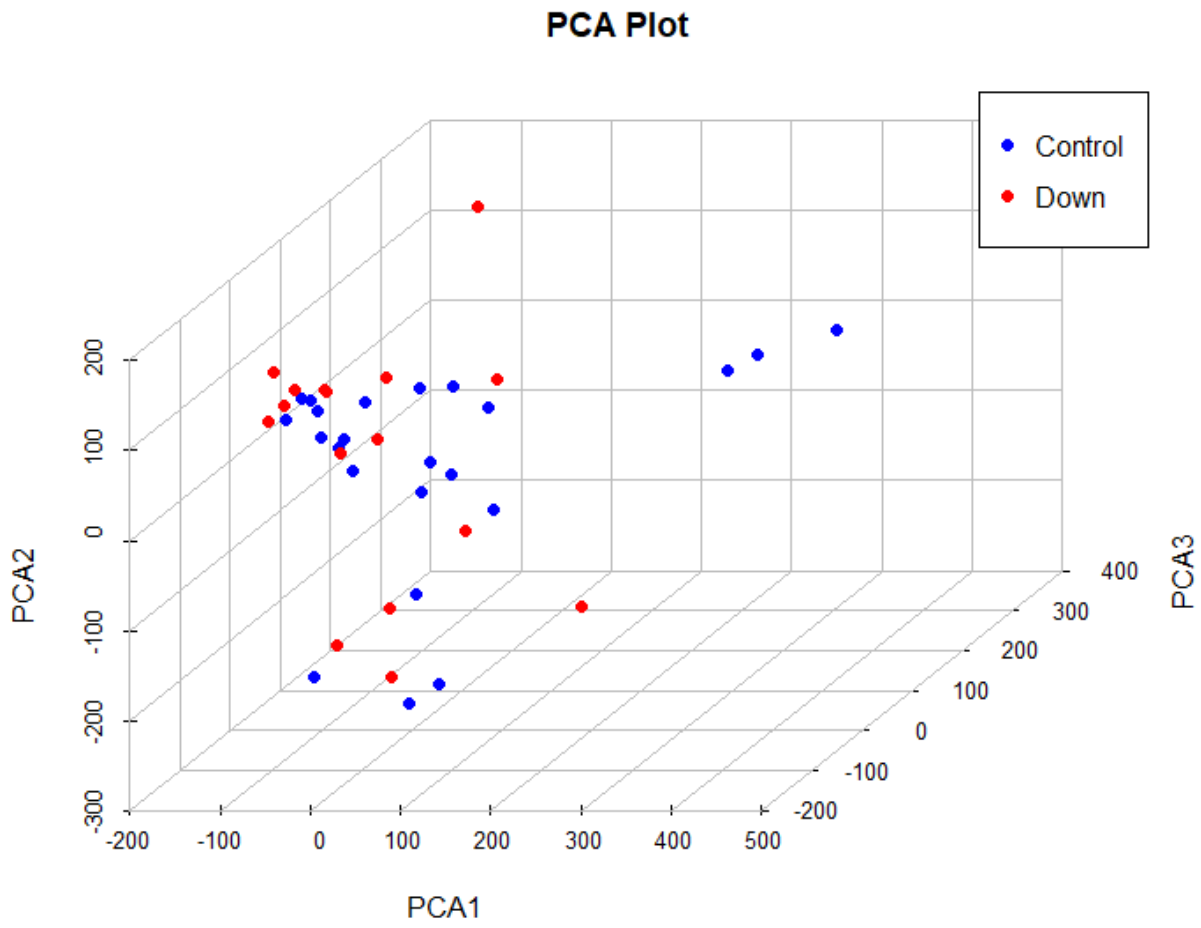
The entire dataset was then analyzed using PathfindR, another enrichment analysis tool, together with the KEGG database. Out of the 26201 genes given as input, only 168 had low enough adjusted p-value and known interactions.

Exploratory Data Analysis



The boxplot confirms that the dataset has been normalized using the median of ratio. However, it reveals that the D01 sample contains one outlier, which was later identified to be the signal data of the HBB RNA. This result could be caused by contamination of the D01 sample. A possible solution to this problem would be to recover all the raw data of the study, remove or adjust the read counts of HBB and do the normalization again on the entire dataset. However, this solution would be computationally expensive and time consuming, so the step taken was to just ignore the D01 sample.

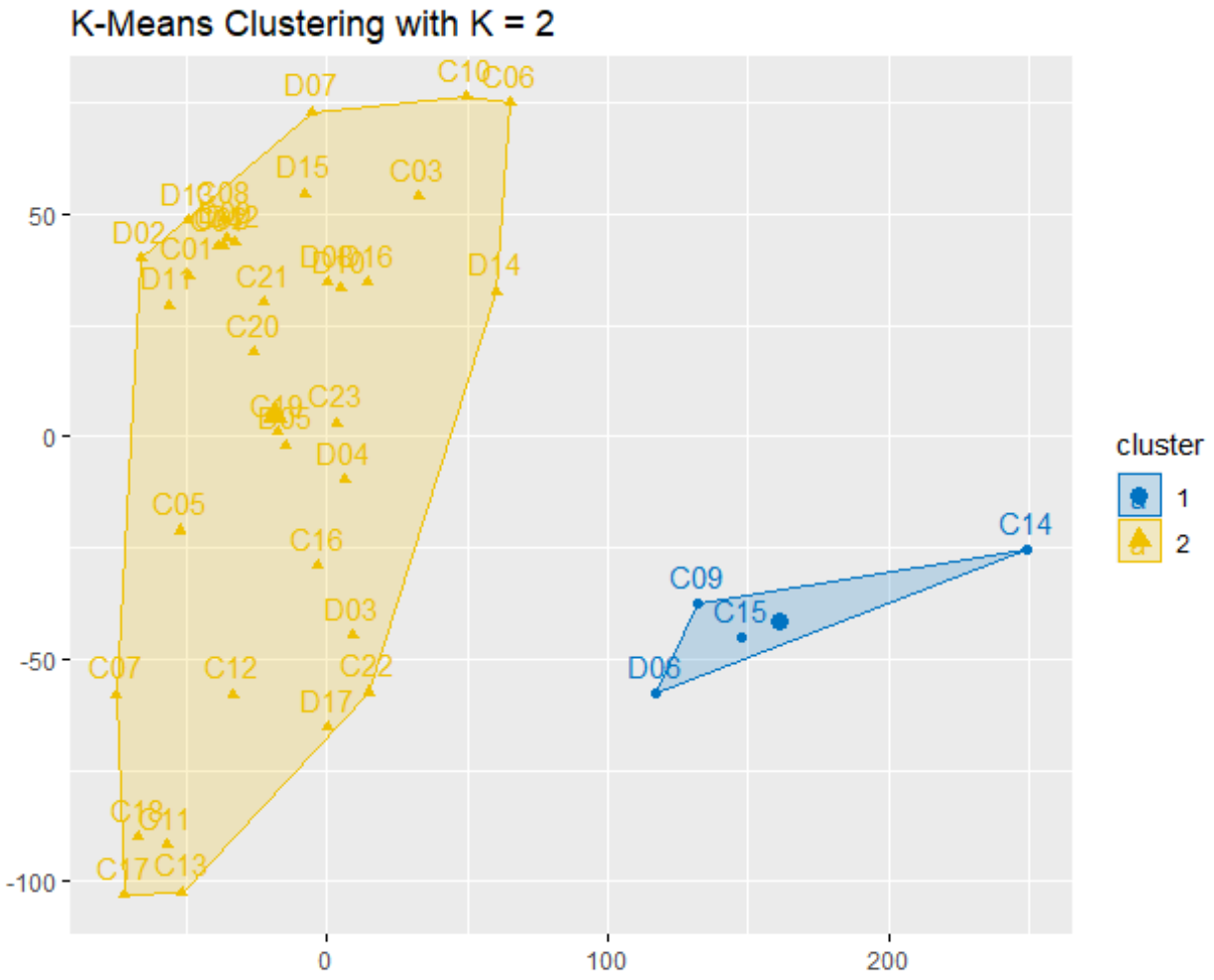
UNSUPERVISED LEARNING

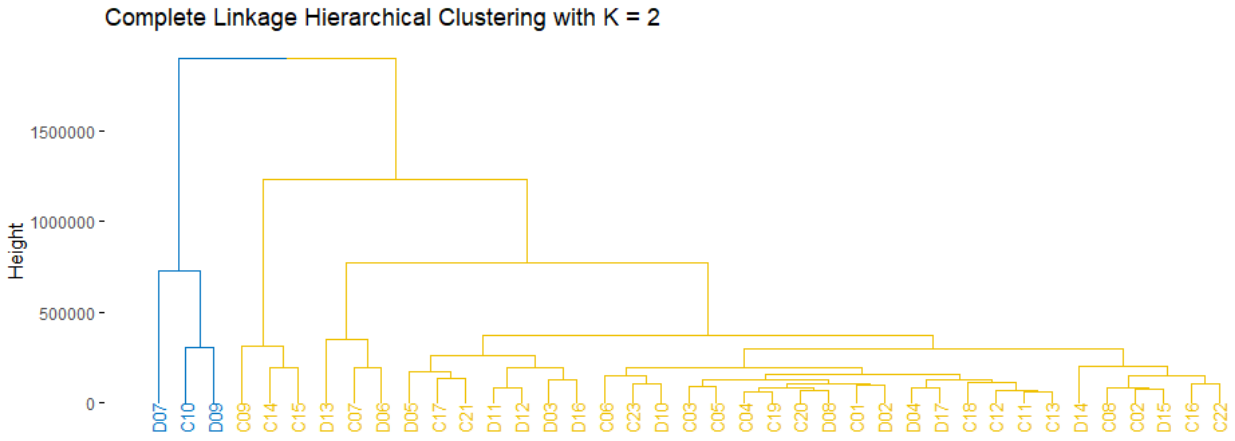


The plot resulting from the principal component analysis (PCA) does not show a clear division between the Control and the Down Syndrome groups.

By looking at the summary of the PCA, it can be observed that the first 3 principal components describe only 16.8% of the total variance combined:

```
Importance of components:
                            PC1       PC2       PC3
Standard deviation     132.18261 116.13227 101.30101
Proportion of Variance   0.07137   0.05509   0.04192
Cumulative Proportion    0.07137   0.12646   0.16837
```

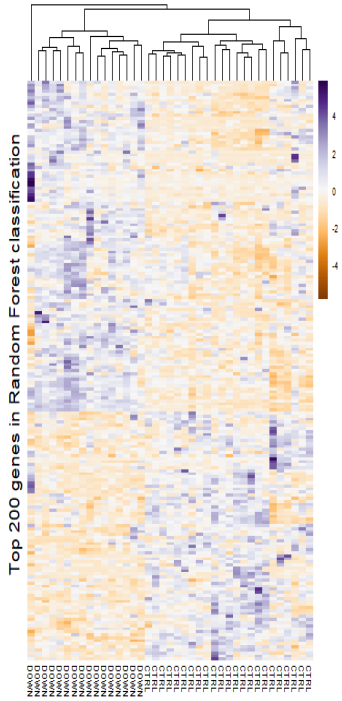


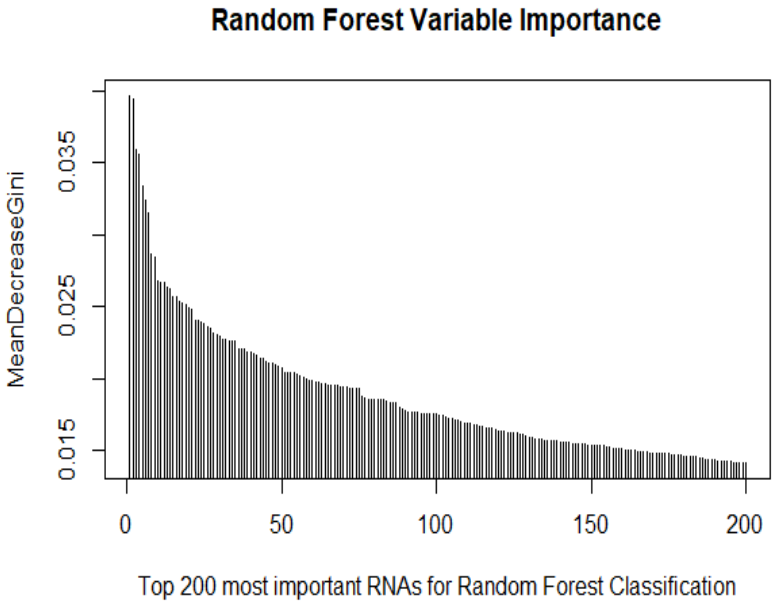


K-Means and Hierarchical Clustering also show that there is no clear distinction between the two expected groups.

SUPERVISED LEARNING

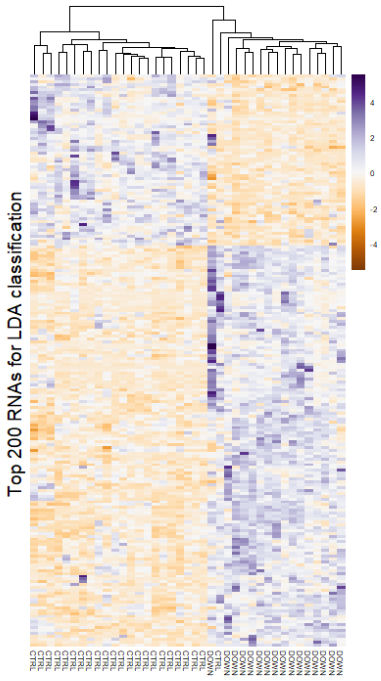
Random Forest

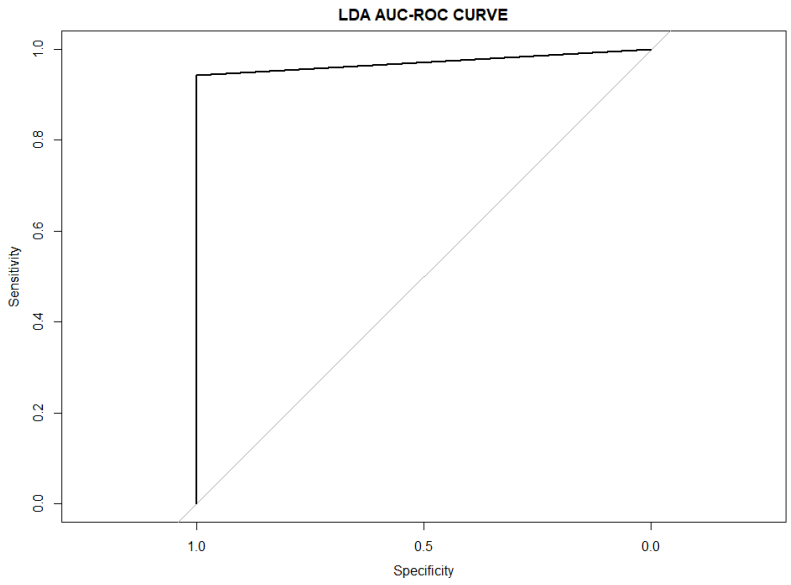




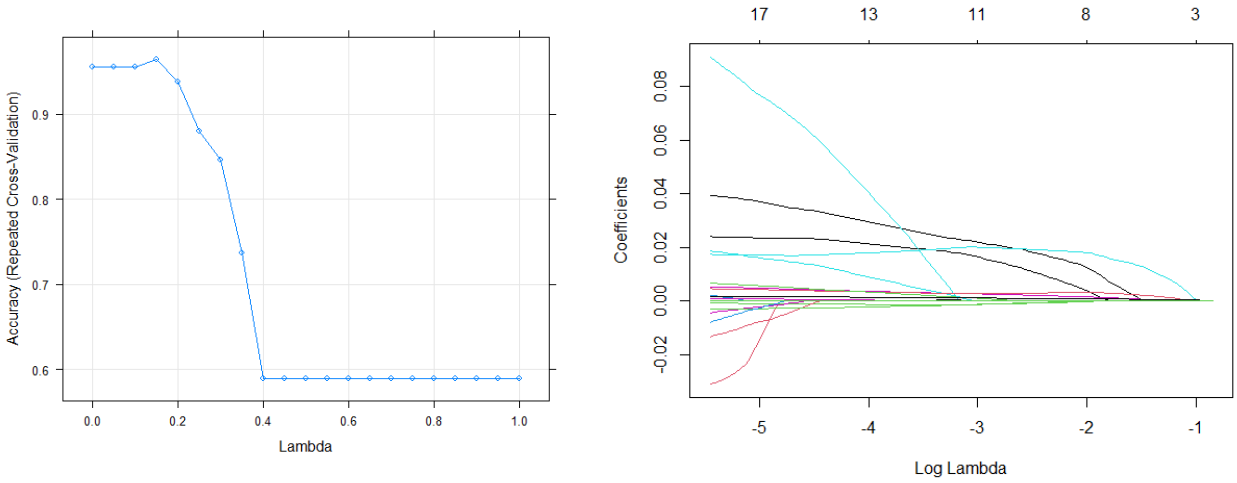
The highest accuracy in Random Forest with 1000 trees was observed when the mtry parameter, which is the number of variables randomly sampled at each split, was set to 2.

Linear Discriminant Analysis





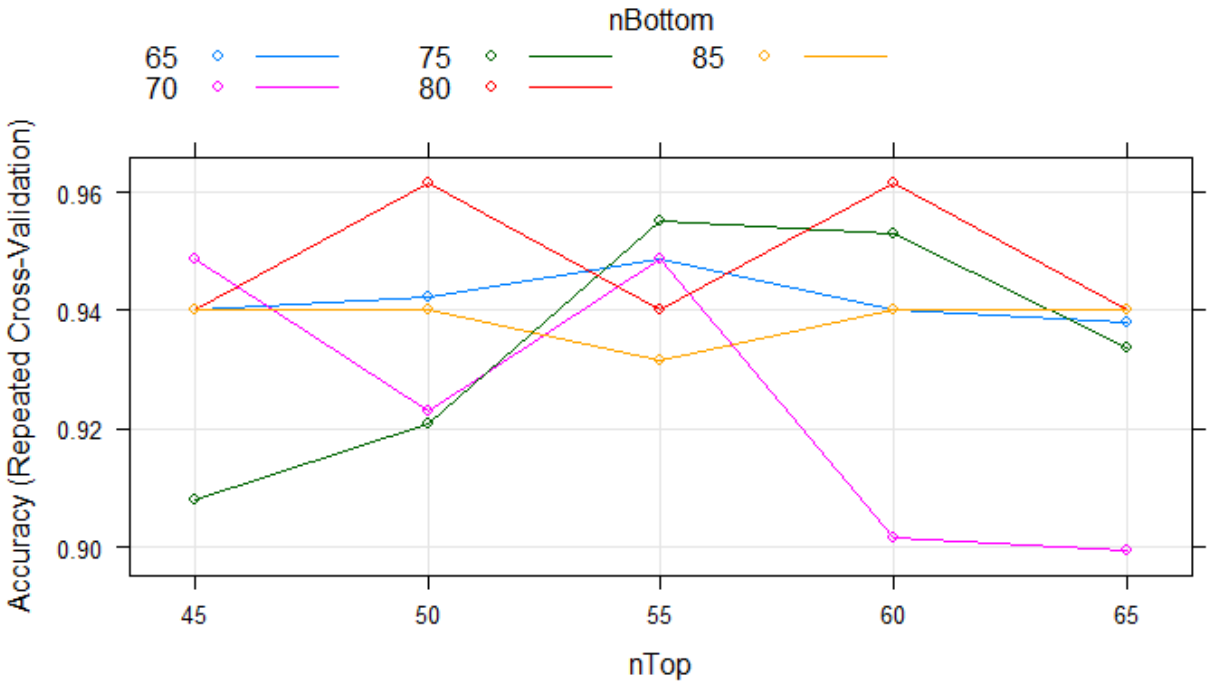
Lasso



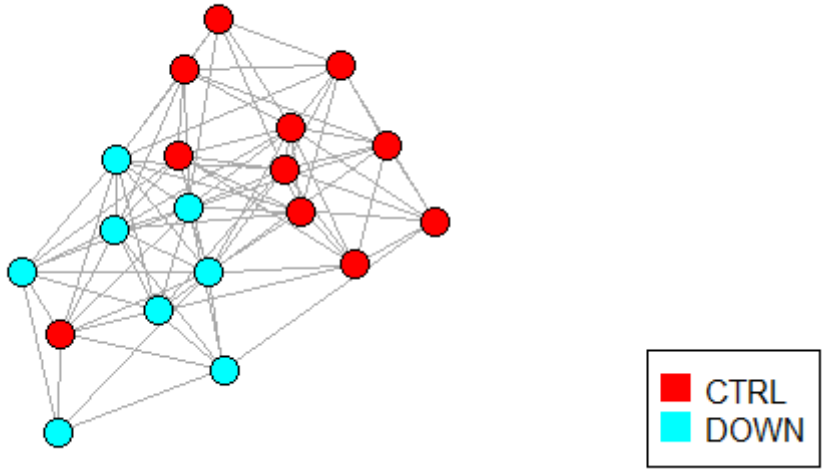
The highest accuracy in lasso was observed when the λ parameter was set to 0.15 (Log λ = -1.9). Using this value, only 6 variables have a coefficient different from zero: CHAF1B, VSIG2, CRYZL1, UBXN10, IFNAR2, and UBE2G2.

It is important to note that 4 out of 6 of these genes are located on chromosome 21. This confirms that the two groups can be distinguished by looking at an imbalance in chromosome 21 transcripts.

RScudo



Using repeated cross validation we obtained the optimal parameters for classifying our data using SCUDO: nTop = 50, nBottom = 80. The N parameter was increased to 0.5 to reduce the number of warnings and increase the accuracy of the model.



Model	Parameters	Accuracy	Kappa
Random Forest	mtry = 2, ntree = 1000	0.970	0.944
Linear Discriminant Analysis		0.970	0.944
Lasso	λ = 0.15	0.957	0.918
SCUDO	nTop = 50, nBottom = 80, N = 0.5	0.962	0.918

All the considered models have been able to classify the data with high accuracy. In the next steps of the study involving functional analysis, it was arbitrarily chosen to analyze the 200 most important genes for LDA classification.

FUNCTIONAL ENRICHMENT ANALYSIS

David

David recognized 187 out of the 200 RNAs given as input. The data was analyzed using the default annotation categories plus UP_TISSUE, for tissue expression. The statistically significant results are the following:

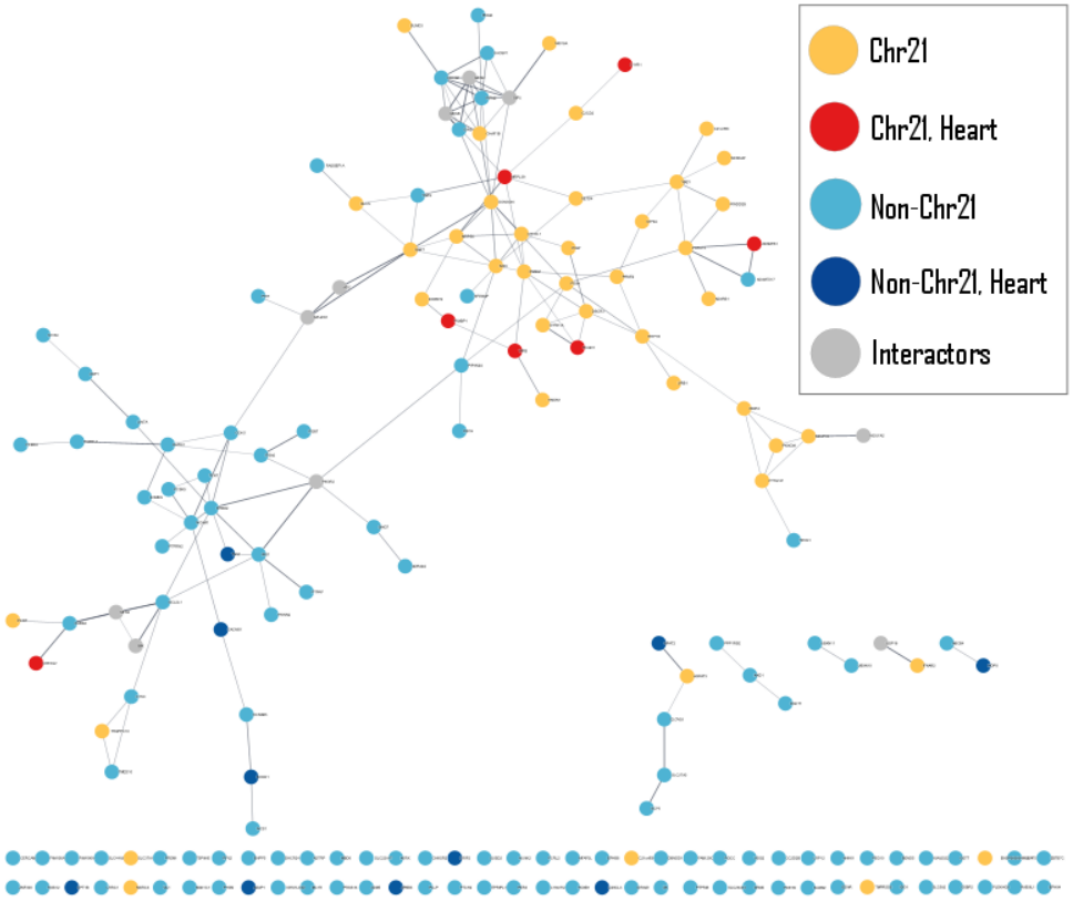
Category	Term	Count	P-value	FDR
UP_TISSUE	Heart	17	1,8E-4	2,0E-2
UP_KW_DOMAIN	Transmembrane helix	64	2,0E-4	2,7E-3
UP_KW_DOMAIN	Transmembrane	64	3,0E-4	2,7E-3

The proteins of the 17 genes expressed in the heart are shown in the following protein-protein interaction network as red (for chr21 proteins) and dark blue (for non-chr21 proteins).

NETWORK ANALYSIS

STRING

168 proteins encoded by the top 200 genes were analyzed using a protein-protein interaction network built with STRING. Confidence was set to 0.5 and no more than 10 interactors have been used.



72 singletons. We can observe two main clusters, one of which is mainly populated by chromosome 21 proteins.

Enrichnet

We again analyzed the top 200 genes in LDA classification using Enrichnet and the Gene Ontology database.

150 input genes have been taken into consideration by the tool. This analysis revealed 21 affected pathways with tissue-specific XD scores above 2.5 (either atrioventricular node or heart tissue XD-scores have been considered).

A selection of these processes are shown in the following table:

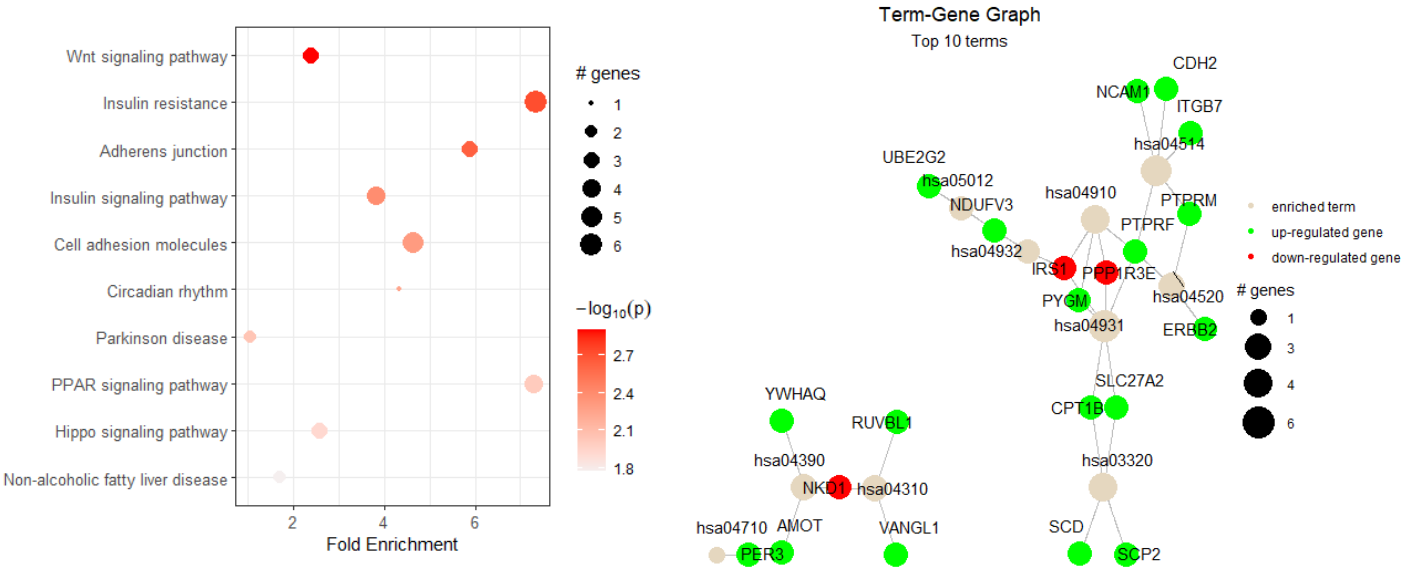
Annotation	AV-node XD-score	Heart XD-score
heart trabecula formation	2.94	-0.07
calcium-dependent cell-cell adhesion	0.94	4.43
striated muscle cell differentiation	1.44	4.43
porphyrin-containing compound metabolic process	4.44	-0.07
positive regulation of NF-kappaB import into nucleus	2.94	-1.07
blood vessel morphogenesis	0.94	2.93
heterophilic cell-cell adhesion	0.63	4.43
positive regulation of muscle cell differentiation	0.58	8.93

positive regulation of MAPK cascade	0.26	8.93
regulation of protein localization	0.69	2.93
muscle cell differentiation	0.44	8.93
regulation of Rho protein signal transduction	0.24	4.43
Cell-cell adhesion	0.20	2.93

This analysis demonstrates that both muscle cells differentiation and cell-cell adhesion processes have been affected in the cardiac tissues of people with Down Syndrome.

PathfindR

the PathfindR tool was used in conjunction with the KEGG database. Out of the 26201 initial genes, after p-value filtering and the removal of genes with no known interactions, the final number of genes in input was 168.



The Wnt signaling pathway (hsa04310) has a key role in embryonic development. In recent studies, it has been shown that Wnt activity is important for early precardiac mesoderm differentiation, but must be inhibited in subsequent steps for cardiomyocyte differentiation to proceed.

It's interesting to note that both the Enrichnet and the PathfindR analyses show that cell adhesion processes and cell differentiation pathways have been affected.

DISCUSSION

In this study we attempted to understand the molecular causes of CHD in people with DS through a RNA-seq data analysis approach. We first fitted our data using different supervised learning methods and we identified the 200 most important RNAs for correct sample classification in LDA. We analyzed them using STRING and DAVID, through which we discovered a set of 17 differentially expressed cardiac tissue specific genes and highlighted how they interact with one another. We also analyzed them with Enrichnet together with the Gene Ontology database, and discovered that muscle cell differentiation and cell-cell adhesion processes have been affected in cardiac tissues of people with DS. We confirmed this hypothesis by doing another analysis using PathfindR in conjunction with the KEGG database, which showed similar results of affected cell-cell adhesion processes and differentiation pathways.

We are aware that the importance of this study is limited by the use of data coming from PBMCs rather than from actual cardiac tissue. But nonetheless we managed to obtain potentially useful information regarding CHD in people with DS while avoiding important ethical issues.

Additional studies regarding cell-cell adhesion and muscle cell differentiation during cardiac tissue development are needed in order to confirm the findings of this study.

References:

[3] A transcriptome analysis study with the focus on the role of long non-coding RNAs in Down syndrome subjects. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10604/

Acknowledgements:

We thank Genomix4Life S.r.l. for publishing the dataset used in this study.