Classification Methods and Enrichment Analysis for Gene Expression Signatures of Respiratory Viral Infections.

Nicolo’ Merzi

1. Abstract

The aim of this project was to try different classification techniques in order to distinguish between three types of acute respiratory viral infections (Rhinovirus, Respiratory Syncytial Virus and Influenza A) given gene expression signatures and identify genes most affected by these diseases. Some of the developed models show good results in classifying healthy individual vs affected, indicating that Acute Respiratory Infections (ARI) induce changes in human gene expression, which can be used to better understand these types of disease and help the diagnostic process.

2. Introduction

2.1 Dataset

The dataset used to train the models is a collection of three different experiments [1] where healthy volunteers were inoculated with intranasal influenza, respiratory syncytial virus or rhinovirus. Peripheral blood samples were drawn in order to extract gene expression for a total of 113 samples:

Respiratory viral infection	Control group	Affected group	Total
Rhino	19	20	39
RSV	20	20	40
Influenza A	17	17	34

2.2 Problems and Known Issues

Classification problems are common and there exist a numerous collection of different algorithms that can address them. When dealing with biological data, and in particular, gene expression signatures, we are dealing with unique and specific types of datasets, where the number of features is significantly greater than the number of samples. This can make some classification algorithms not feasible, due to both the algorithm design and from a computational perspective.

Another problem that presented with this dataset is a clear presence of batch effect, due to the different timing and settings of the three experiments. As we can see from *Figure 1,* the PCA projection of the samples is clearly divided based on the different experiments. We tried two different approaches: in one case we addressed the problem of batch effect by using the function “RemoveBatchEffect()” from the R package “Limma” [2], and proceeded with the classification considering all three batches together. In the other case we considered each batch separately.

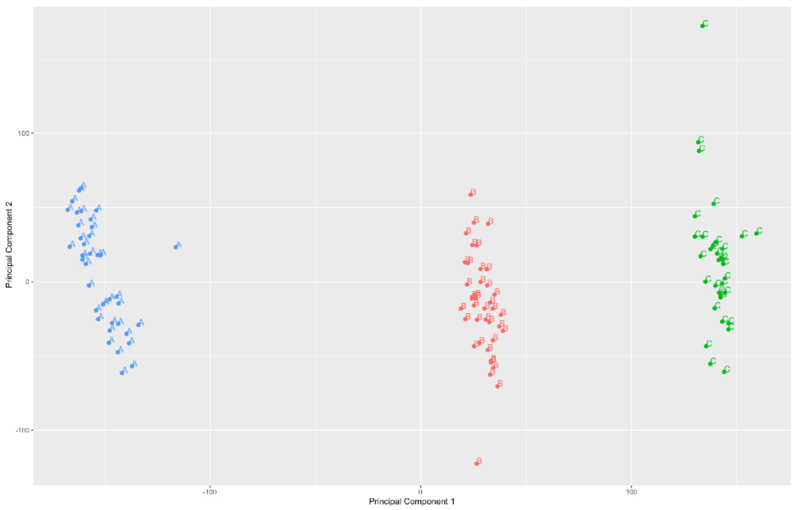


Figure 1, PCA projection of all the samples. Each color corresponds to a different batch. Clear batch effect.

3. Methods

3.1 Unsupervised methods

First, we tried to see if it was possible to obtain a good classification using unsupervised methods, in particular K-means clustering. The results show that, even though we were able to remove the batch effect, unsupervised classification is not enough to produce satisfying results, even when considering each batch separately. As we can see from *Figure 2,* K-means (K = 4) was not able to correctly distinguish between the three viruses and the control group.

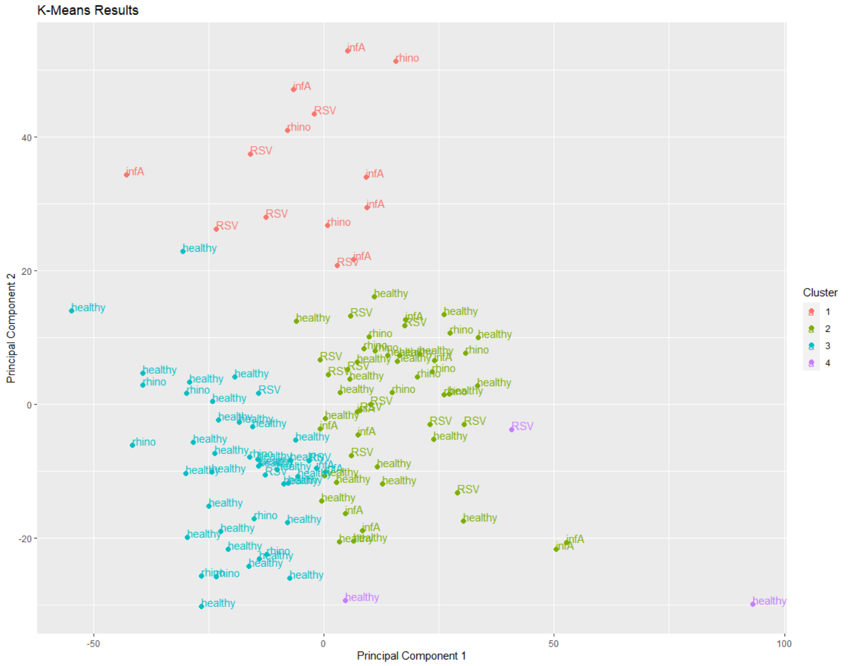


Figure 2, K-means (K = 4) clustering plotted on PCA after removing Batch Effect

The same unsatisfactory results were obtained when considering each batch separately. Only for the batch of Influenza A, k-means (K = 2) was able to produce a decent unsupervised classification.

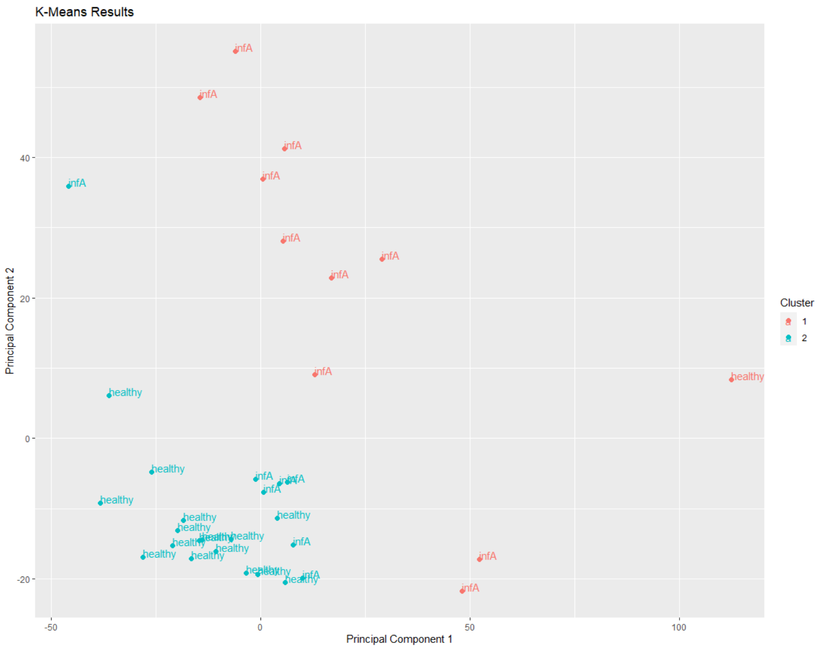


Figure 3, K-means (K = 2) plotted on PCA for Influenza A batch

3.2 Supervised methods

Since unsupervised methods were proved insufficient for classification purposes, the next logical step is trying supervised methods.

3.2.1 Random Forest

We started by using Random Forest and learning an ensemble of decision trees (ntrees = 1000), which would provide a classification based on majority voting. Random Forest was applied considering all three batches together, after removing the batch effect. The model produced unsatisfactory results with a classification accuracy of just 0.59. The model would classify most points as belonging to the control group, probably due to the overrepresentation of the control group when considering batches together. Instead of classifying between three different diseases and the control group, we tried to just learn decision trees capable of recognizing if the person was healthy or sick. This approach significantly improved the accuracy of the model, obtaining a correct classification rate of 0.79.

When dealing with each batch separately, Random Forest performs better. After cross-validating each model with a 10-fold cross validation procedure, we obtained the following accuracies on the test sets:

• Accuracy on Rhino batch:	0.9
• Accuracy on RSV batch:	0.8
• Accuracy on Influenza A batch:	0.78

With Random Forest we are also able to extract the top genes in order of importance, in other words, the genes that influence the classification the most. We can see that for every batch, only 300-400 genes are useful out of over 20.000.

3.2.2 Linear Discriminant Analysis

The second method we tried was Linear Discriminant Analysis (LDA). Before training the model, we also did feature selection by performing a t-test and keeping only the genes that would score a p-value below 0.1. This resulted in only considering approximately 5000-6000 genes instead of over 20.000. Applying LDA on the three batches together did not improve the accuracy of the Random Forest model. For this reason, instead of trying to distinguish between each virus and the control group, we again grouped the diseases and classified only based on healthy vs sick. This resulted in a significant boost with an accuracy of almost 0.8 from 0.56. Based on this, we can say that the three respiratory viral infection share similarities regarding their effect on gene expression when compared against healthy individuals.

When considering each batch separately, LDA is the best model with an almost perfect accuracy. After a 10-fold cross validation procedure we obtained the following accuracies on the test sets:

• Accuracy on Rhino batch:	1
• Accuracy on RSV batch:	1
• Accuracy on Influenza A batch:	0.78

3.2.3 Lasso & Ridge Regression

The third model is a particular form of linear regression that uses Ridge and Lasso regularization in order to reduce model complexity and multi-collinearity, which is especially problematic when dealing with datasets with a large number of features. While Ridge regularization shrinks the effect of some of the coefficients, Lasso regularization can also set them to zero, effectively making Lasso a powerful feature selection tool. When considering the three batches together and classifying based on healthy vs sick we obtain similar results as LDA with an accuracy of 0.79, using alpha = 0.1 and lambda = 0.197.

Good results were obtained also when considering each batch separately. After a 10-fold cross validation procedure we obtained the following accuracies on the tests sets:

• Accuracy on Rhino batch:	1	alpha = 1	lambda = 0.245
• Accuracy on RSV batch:	0.9	alpha = 0.55	lambda = 0.240
• Accuracy on Influenza A batch:	0.78	alpha = 0.1	lambda = 0.277

3.2.4 SCUDO

For the last classification model, we trained SCUDO: a ranked based signature method, which only considers the *n* most and least expressive genes signatures for every sample. After that, it computes the distance between all possible pairs, resulting in a graph network that maps the similarity of genes. After an initial feature selection using the Wilcoxon rank sum test and using 0.1 as cutoff, we are left with around 6000 gene signatures. The model was trained considering the 15 most and least expressive genes. *Figure 4* shows the resulting network graph of the train and test set.

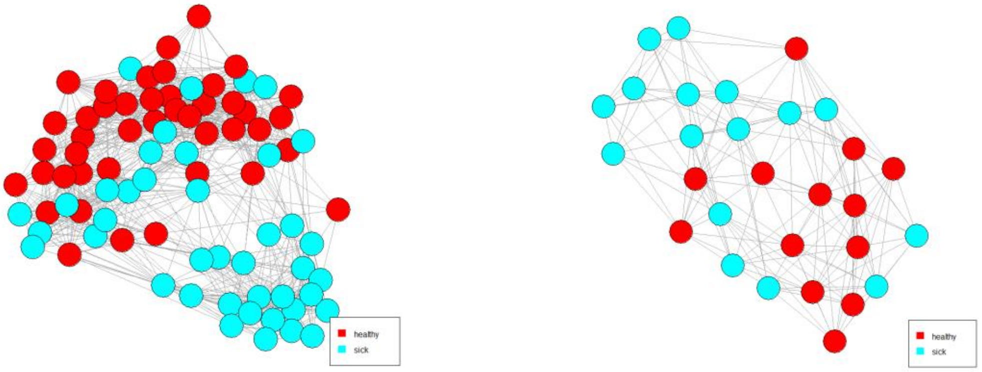


Figure 4, SCUDO all three batches. Left: train network. Right: test network.

The model performs slightly worse than the previous models, with an accuracy of around 0.76 on the test set.

When considering each batch separately we obtain better results:

- On the Rhino batch, feature selection leaves us with around 4000 genes. Considering the top and bottom 30 gene expressions, the accuracy obtained on the test set is 0.9

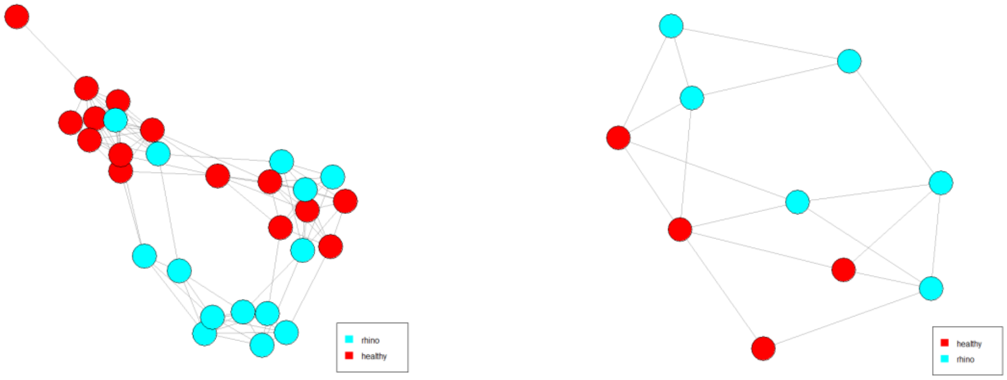


Figure 5, SCUDO Rhino batch. Left: train network. Right: test network

- On the RSV batch, feature selection leaves us with around 5000 genes. Considering the top and bottom 30 gene expressions, the accuracy obtained on the test set is 0.8

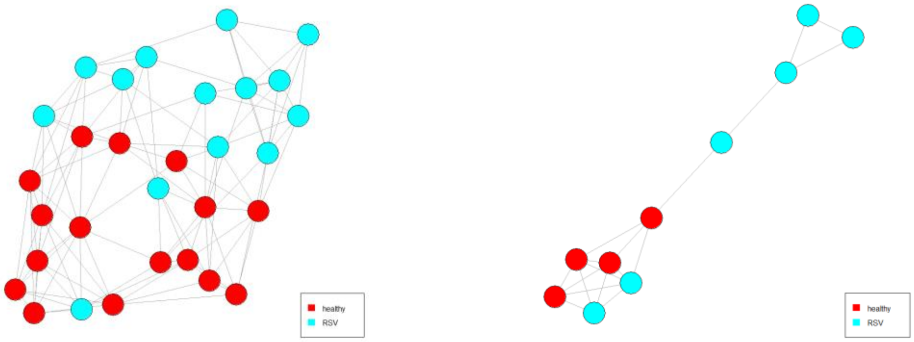


Figure 6, SCUDO RSV batch. Left: train network. Right: test network

- On the Influenza A batch, feature selection leaves us with around 8000 genes. Considering the top and bottom 80 gene expressions, the accuracy obtained on the test set is 0.89

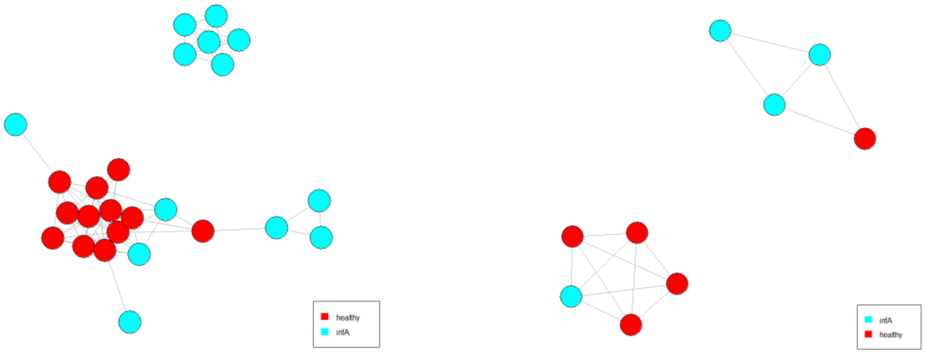


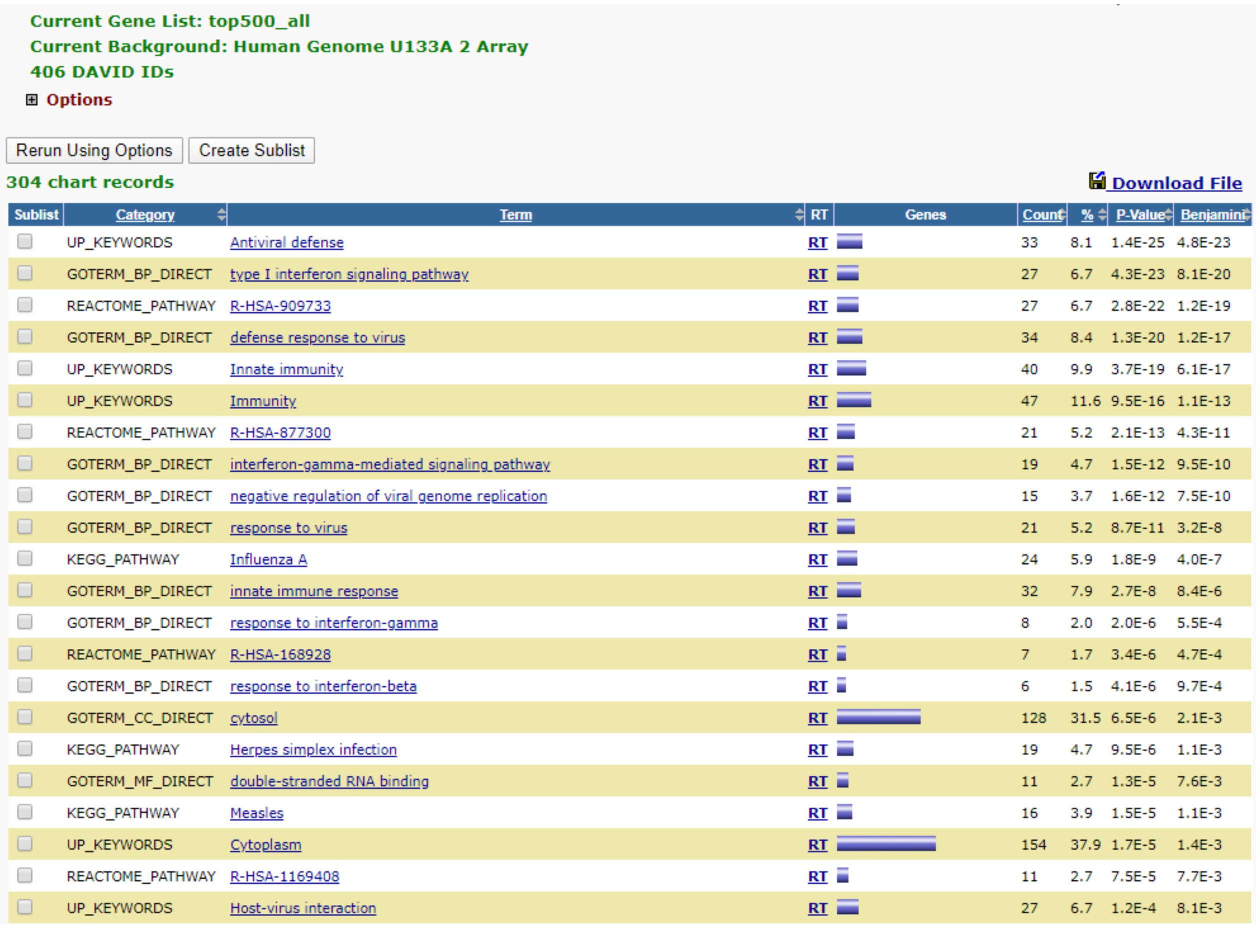
Figure 7, SCUDO Influenza A batch. Left: train network. Right: test network

3.3 Functional Enrichment Analysis

Now that we have trained our models, we are left with a list of features (genes) used by the models to classify samples. The next step is to perform a functional enrichment analysis to see if these genes are enriched in known functions, complexes and pathways. In other words, we want a biological interpretation of the genes used to distinguish between healthy and sick.

In order to do this, we take the top 500 genes based on the t-test we performed with LDA, since it was the best model in terms of accuracy. The software used to perform the analysis is *David* [3], an online bioinformatic resource which aims to provide functional interpretation of large lists of genes. *David* takes as input a list of genes and returns a list of Gene Ontology terms associated to the biological process, cellular components and molecular functions of the genes in input.

Figure 8 shows the most prominent GO terms associated with the top 500 genes considering all three batches together. As we can see, the top terms are all related to the Immune System and the host response to virus, which makes sense considering we are dealing with viral infections. The Benjamini adjusted p-value indicates that these results are statistically significant.



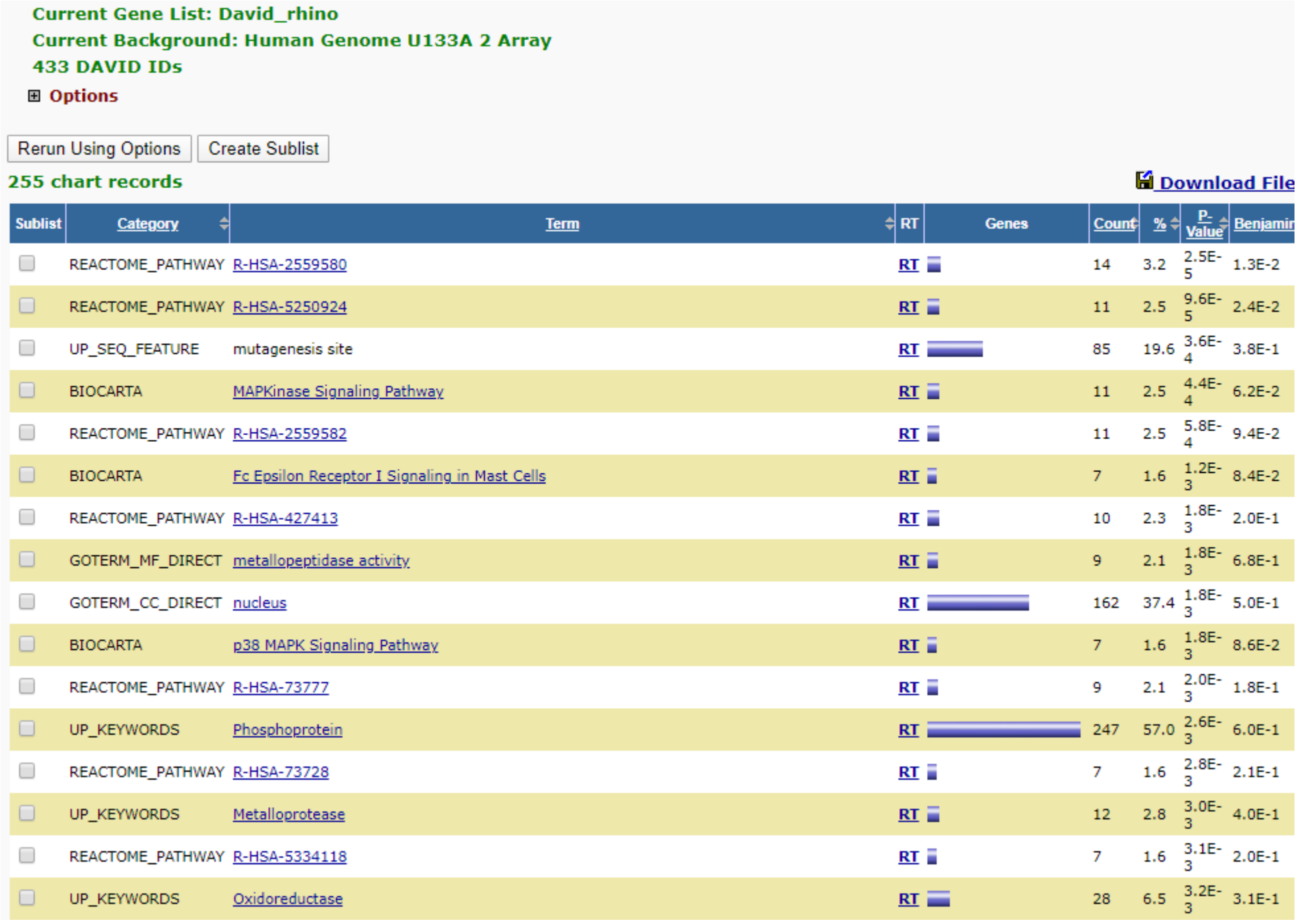
Current Gene List: top500_all
Current Background: Human Genome U133A 2 Array
406 DAVID IDs
⊞ Options

304 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
□	UP_KEYWORDS	Antiviral defense	RT		33	8.1	1.4E-25	4.8E-23
□	GOTERM_BP_DIRECT	type I interferon signaling pathway	RT		27	6.7	4.3E-23	8.1E-20
□	REACTOME_PATHWAY	R-HSA-909733	RT		27	6.7	2.8E-22	1.2E-19
□	GOTERM_BP_DIRECT	defense response to virus	RT		34	8.4	1.3E-20	1.2E-17
□	UP_KEYWORDS	Innate immunity	RT		40	9.9	3.7E-19	6.1E-17
□	UP_KEYWORDS	Immunity	RT		47	11.6	9.5E-16	1.1E-13
□	REACTOME_PATHWAY	R-HSA-877300	RT		21	5.2	2.1E-13	4.3E-11
□	GOTERM_BP_DIRECT	interferon-gamma-mediated signaling pathway	RT		19	4.7	1.5E-12	9.5E-10
□	GOTERM_BP_DIRECT	negative regulation of viral genome replication	RT		15	3.7	1.6E-12	7.5E-10
□	GOTERM_BP_DIRECT	response to virus	RT		21	5.2	8.7E-11	3.2E-8
□	KEGG_PATHWAY	Influenza A	RT		24	5.9	1.8E-9	4.0E-7
□	GOTERM_BP_DIRECT	innate immune response	RT		32	7.9	2.7E-8	8.4E-6
□	GOTERM_BP_DIRECT	response to interferon-gamma	RT		8	2.0	2.0E-6	5.5E-4
□	REACTOME_PATHWAY	R-HSA-168928	RT		7	1.7	3.4E-6	4.7E-4
□	GOTERM_BP_DIRECT	response to interferon-beta	RT		6	1.5	4.1E-6	9.7E-4
□	GOTERM_CC_DIRECT	cytosol	RT		128	31.5	6.5E-6	2.1E-3
□	KEGG_PATHWAY	Herpes simplex infection	RT		19	4.7	9.5E-6	1.1E-3
□	GOTERM_MF_DIRECT	double-stranded RNA binding	RT		11	2.7	1.3E-5	7.6E-3
□	KEGG_PATHWAY	Measles	RT		16	3.9	1.5E-5	1.1E-3
□	UP_KEYWORDS	Cytoplasm	RT		154	37.9	1.7E-5	1.4E-3
□	REACTOME_PATHWAY	R-HSA-1169408	RT		11	2.7	7.5E-5	7.7E-3
□	UP_KEYWORDS	Host-virus interaction	RT		27	6.7	1.2E-4	8.1E-3

Figure 8, GO terms associated with top genes considering all batches

When considering each batch separately, the GO terms are again related to the Immune System, interferons (group of signaling proteins made and released by host cells in response to the presence of several viruses [4]) and host response to virus, but the results do not have the same statistical significance probably due to the lower number of samples when considering each batch separately.

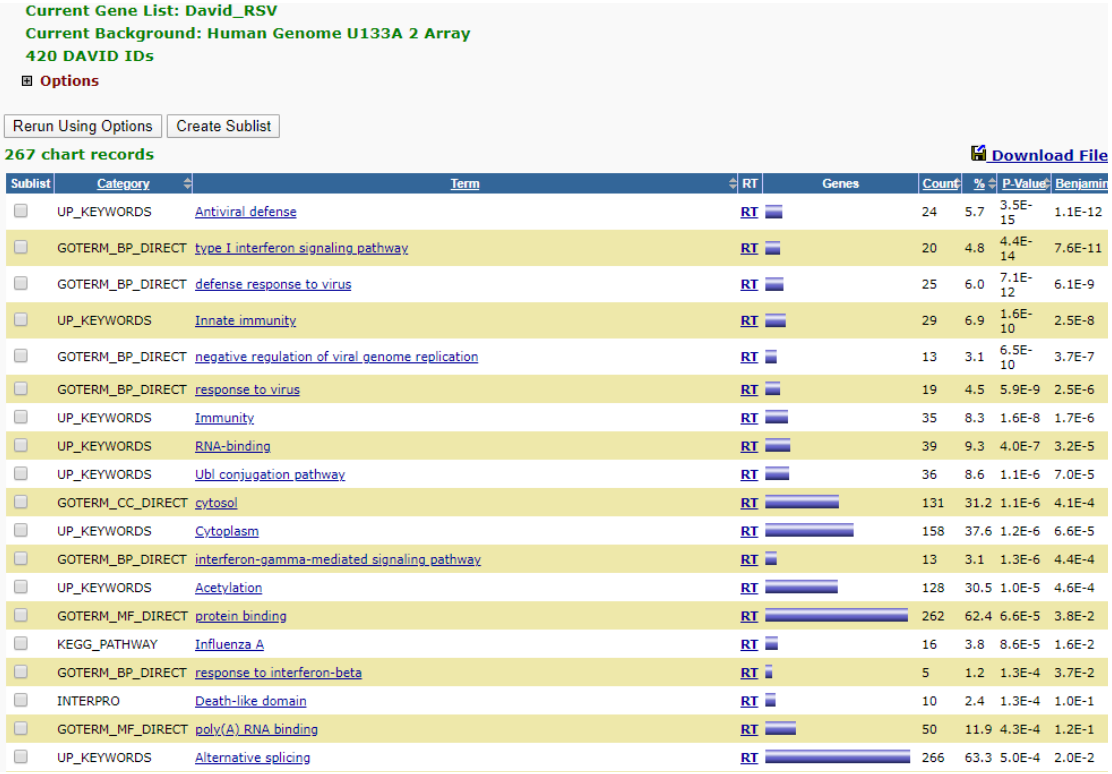


Current Gene List: David_rhino
Current Background: Human Genome U133A 2 Array
433 DAVID IDs
⊞ **Options**

255 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamir
☐	REACTOME_PATHWAY	R-HSA-2559580	RT		14	3.2	2.5E-5	1.3E-2
☐	REACTOME_PATHWAY	R-HSA-5250924	RT		11	2.5	9.6E-5	2.4E-2
☐	UP_SEQ_FEATURE	mutagenesis site	RT		85	19.6	3.6E-4	3.8E-1
☐	BIOCARTA	MAPKinase Signaling Pathway	RT		11	2.5	4.4E-4	6.2E-2
☐	REACTOME_PATHWAY	R-HSA-2559582	RT		11	2.5	5.8E-4	9.4E-2
☐	BIOCARTA	Fc Epsilon Receptor I Signaling in Mast Cells	RT		7	1.6	1.2E-3	8.4E-2
☐	REACTOME_PATHWAY	R-HSA-427413	RT		10	2.3	1.8E-3	2.0E-1
☐	GOTERM_MF_DIRECT	metallopeptidase activity	RT		9	2.1	1.8E-3	6.8E-1
☐	GOTERM_CC_DIRECT	nucleus	RT		162	37.4	1.8E-3	5.0E-1
☐	BIOCARTA	p38 MAPK Signaling Pathway	RT		7	1.6	1.8E-3	8.6E-2
☐	REACTOME_PATHWAY	R-HSA-73777	RT		9	2.1	2.0E-3	1.8E-1
☐	UP_KEYWORDS	Phosphoprotein	RT		247	57.0	2.6E-3	6.0E-1
☐	REACTOME_PATHWAY	R-HSA-73728	RT		7	1.6	2.8E-3	2.1E-1
☐	UP_KEYWORDS	Metalloprotease	RT		12	2.8	3.0E-3	4.0E-1
☐	REACTOME_PATHWAY	R-HSA-5334118	RT		7	1.6	3.1E-3	2.0E-1
☐	UP_KEYWORDS	Oxidoreductase	RT		28	6.5	3.2E-3	3.1E-1

Figure 9, GO terms associated to top genes of Rhino batch

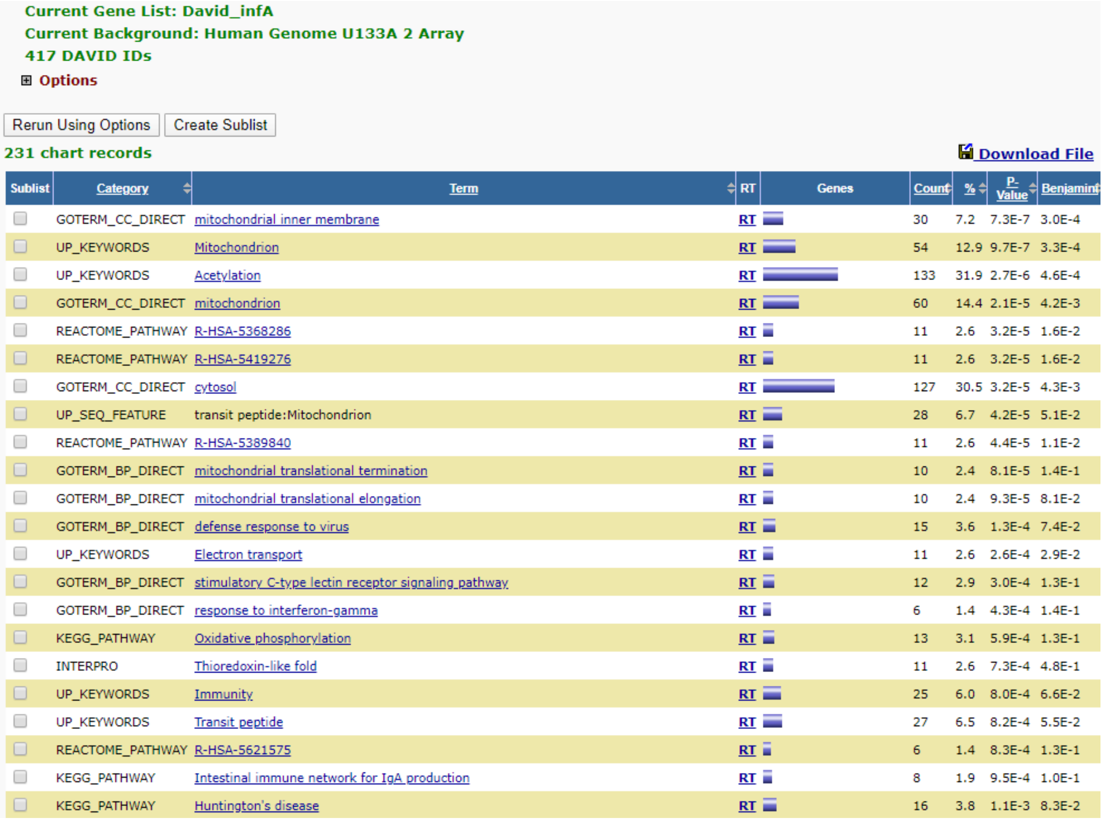


Current Gene List: David_RSV
Current Background: Human Genome U133A 2 Array
420 DAVID IDs
⊞ **Options**

267 chart records

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	UP_KEYWORDS	Antiviral defense	RT		24	5.7	3.5E-15	1.1E-12
☐	GOTERM_BP_DIRECT	type I interferon signaling pathway	RT		20	4.8	4.4E-14	7.6E-11
☐	GOTERM_BP_DIRECT	defense response to virus	RT		25	6.0	7.1E-12	6.1E-9
☐	UP_KEYWORDS	Innate immunity	RT		29	6.9	1.6E-10	2.5E-8
☐	GOTERM_BP_DIRECT	negative regulation of viral genome replication	RT		13	3.1	6.5E-10	3.7E-7
☐	GOTERM_BP_DIRECT	response to virus	RT		19	4.5	5.9E-9	2.5E-6
☐	UP_KEYWORDS	Immunity	RT		35	8.3	1.6E-8	1.7E-6
☐	UP_KEYWORDS	RNA-binding	RT		39	9.3	4.0E-7	3.2E-5
☐	UP_KEYWORDS	Ubl conjugation pathway	RT		36	8.6	1.1E-6	7.0E-5
☐	GOTERM_CC_DIRECT	cytosol	RT		131	31.2	1.1E-6	4.1E-4
☐	UP_KEYWORDS	Cytoplasm	RT		158	37.6	1.2E-6	6.6E-5
☐	GOTERM_BP_DIRECT	interferon-gamma-mediated signaling pathway	RT		13	3.1	1.3E-6	4.4E-4
☐	UP_KEYWORDS	Acetylation	RT		128	30.5	1.0E-5	4.6E-4
☐	GOTERM_MF_DIRECT	protein binding	RT		262	62.4	6.6E-5	3.8E-2
☐	KEGG_PATHWAY	Influenza A	RT		16	3.8	8.6E-5	1.6E-2
☐	GOTERM_BP_DIRECT	response to interferon-beta	RT		5	1.2	1.3E-4	3.7E-2
☐	INTERPRO	Death-like domain	RT		10	2.4	1.3E-4	1.0E-1
☐	GOTERM_MF_DIRECT	poly(A) RNA binding	RT		50	11.9	4.3E-4	1.2E-1
☐	UP_KEYWORDS	Alternative splicing	RT		266	63.3	5.0E-4	2.0E-2

Figure 10, GO terms associated to top genes of RSV batch



Current Gene List: David_infA
Current Background: Human Genome U133A 2 Array
417 DAVID IDs
⊞ **Options**

231 chart records

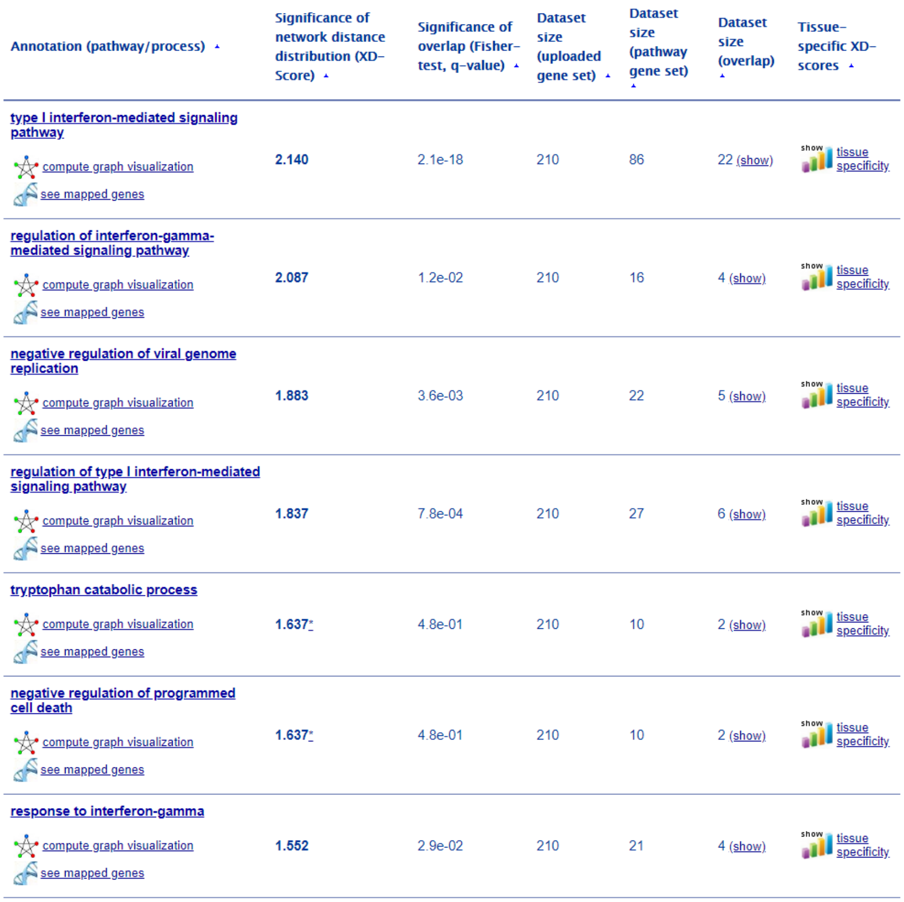
Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
☐	GOTERM_CC_DIRECT	mitochondrial inner membrane	RT		30	7.2	7.3E-7	3.0E-4
☐	UP_KEYWORDS	Mitochondrion	RT		54	12.9	9.7E-7	3.3E-4
☐	UP_KEYWORDS	Acetylation	RT		133	31.9	2.7E-6	4.6E-4
☐	GOTERM_CC_DIRECT	mitochondrion	RT		60	14.4	2.1E-5	4.2E-3
☐	REACTOME_PATHWAY	R-HSA-5368286	RT		11	2.6	3.2E-5	1.6E-2
☐	REACTOME_PATHWAY	R-HSA-5419276	RT		11	2.6	3.2E-5	1.6E-2
☐	GOTERM_CC_DIRECT	cytosol	RT		127	30.5	3.2E-5	4.3E-3
☐	UP_SEQ_FEATURE	transit peptide:Mitochondrion	RT		28	6.7	4.2E-5	5.1E-2
☐	REACTOME_PATHWAY	R-HSA-5389840	RT		11	2.6	4.4E-5	1.1E-2
☐	GOTERM_BP_DIRECT	mitochondrial translational termination	RT		10	2.4	8.1E-5	1.4E-1
☐	GOTERM_BP_DIRECT	mitochondrial translational elongation	RT		10	2.4	9.3E-5	8.1E-2
☐	GOTERM_BP_DIRECT	defense response to virus	RT		15	3.6	1.3E-4	7.4E-2
☐	UP_KEYWORDS	Electron transport	RT		11	2.6	2.6E-4	2.9E-2
☐	GOTERM_BP_DIRECT	stimulatory C-type lectin receptor signaling pathway	RT		12	2.9	3.0E-4	1.3E-1
☐	GOTERM_BP_DIRECT	response to interferon-gamma	RT		6	1.4	4.3E-4	1.4E-1
☐	KEGG_PATHWAY	Oxidative phosphorylation	RT		13	3.1	5.9E-4	1.3E-1
☐	INTERPRO	Thioredoxin-like fold	RT		11	2.6	7.3E-4	4.8E-1
☐	UP_KEYWORDS	Immunity	RT		25	6.0	8.0E-4	6.6E-2
☐	UP_KEYWORDS	Transit peptide	RT		27	6.5	8.2E-4	5.5E-2
☐	REACTOME_PATHWAY	R-HSA-5621575	RT		6	1.4	8.3E-4	1.3E-1
☐	KEGG_PATHWAY	Intestinal immune network for IgA production	RT		8	1.9	9.5E-4	1.0E-1
☐	KEGG_PATHWAY	Huntington's disease	RT		16	3.8	1.1E-3	8.3E-2

Figure 11, GO terms associated to top genes of Influenza A batch

4. Network-based Enrichment Analysis

We are now interested in observing if these gene sets are supported by already known interactions. To achieve this, we make use of *EnrichNet*, a network-based enrichment analysis method to identify functional associations between user-defined gene or protein sets and cellular pathway [5]. The gene set is mapped to a database of interest where genes pairwise association are assessed by graph-based statistic. The output is a ranking table of pathways or processes with association scores.

Figure 12 shows the results of *EnrichNet* using Gene Ontology as annotation database and the same top 500 genes found with the feature selection of LDA considering all three batches. From the table below we can see that the statistically significant pathway and processes found by the analysis are related to the immune system response to viral infections. These results confirm that the classification models we have trained before make use of genes associated with the host response to viral infection in order to distinguish between healthy and sick individuals. Viral infections do not change gene expressions per se, but rather they trigger the host's immune system to defend against viruses, which in turns activate particular genes.



Annotation (pathway/process) ▲	Significance of network distance distribution (XD-Score) ▲	Significance of overlap (Fisher-test, q-value) ▲	Dataset size (uploaded gene set) ▲	Dataset size (pathway gene set) ▲	Dataset size (overlap) ▲	Tissue-specific XD-scores ▲
type I interferon-mediated signaling pathway compute graph visualization see mapped genes	2.140	2.1e-18	210	86	22 (show)	show tissue specificity
regulation of interferon-gamma-mediated signaling pathway compute graph visualization see mapped genes	2.087	1.2e-02	210	16	4 (show)	show tissue specificity
negative regulation of viral genome replication compute graph visualization see mapped genes	1.883	3.6e-03	210	22	5 (show)	show tissue specificity
regulation of type I interferon-mediated signaling pathway compute graph visualization see mapped genes	1.837	7.8e-04	210	27	6 (show)	show tissue specificity
tryptophan catabolic process compute graph visualization see mapped genes	1.637*	4.8e-01	210	10	2 (show)	show tissue specificity
negative regulation of programmed cell death compute graph visualization see mapped genes	1.637*	4.8e-01	210	10	2 (show)	show tissue specificity
response to interferon-gamma compute graph visualization see mapped genes	1.552	2.9e-02	210	21	4 (show)	show tissue specificity

Figure 12, Results of EnrichNet on top 500 genes (all batches).

5. Classification Results Summary

	Batch:			
Algorithm:	**ALL** (healthy vs sick)	**Rhino**	**RSV**	**Influenza A**
Random Forest				
Accuracy	0.79	0.9	0.8	0.78
Sensitivity	0.77	1	0.75	0.8
Specificity	0.81	0.83	0.83	0.75
LDA				
Accuracy	0.79	1	1	0.78
Sensitivity	0.64	1	1	0.8
Specificity	0.93	1	1	0.75
Lasso/Ridge				
Accuracy	0.79	1	0.9	0.78
Sensitivity	0.69	1	1	0.8
Specificity	0.87	1	0.83	0.75
SCUDO				
Accuracy	0.76	0.9	0.8	0.89
Sensitivity	0.84	1	1	0.8
Specificity	0.68	0.83	0.67	1

6. Conclusions

We trained different models to define gene expression patterns that are characteristic of response to respiratory viral infection. The results show that gene expression signatures can be effectively used in order to classify respiratory viral infection with a good degree of accuracy.

In addition, we cross-validated each model on different test sets to measure the goodness and generalization capability of the models. We trained a model on a train set of one batch (i.e. Rhino batch) and tested it on the test set of another batch (i.e. Influenza A batch). The models were still able to correctly classify healthy vs sick individuals with a good level of accuracy even though they were trained and tested on different viral infections.

This shows that the host response to viral infection behave similarly despite the differences between diseases. We have seen that we are able to classify viral infections based on gene expressions directly related to the immune system of the individual and the response of the immune system to viral infections. This can be a problem if we are dealing with previously immunocompromised patients since they have a reduced ability to fight infections [6]. A lower response of the immune system to fight off an infection would make it harder for our classification methods to correctly assess the individual’s health status.

In conclusion, gene expression signatures can be an effective tool for diagnostic testing and to better understand the nature of both infections and the host response to infections.

References

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE17156

[2] https://www.rdocumentation.org/packages/limma/versions/3.28.14/topics/removeBatchEffect

[3] https://david.ncifcrf.gov/home.jsp

[4] https://www.sciencedirect.com/topics/neuroscience/interferon

[5] https://ico2s.org/servers/enrichnet.html

[6] https://www.cancer.gov/publications/dictionaries/cancer-terms/def/immunocompromised