Clustering and Classification of Leukocyte Gene Expression in Cystic Fibrosis

PORRECA ANNAMARIA[1], RECCHIUTI ANTONIO[1], BORRELLI PAOLA[1], DI NICOLA MARTA[1]

[1] Department of Medical, Oral and Biotechnologies Sciences, "G. d'Annunzio" University, Chieti-Pescara.

Introduction

Cystic fibrosis (CF) is the most common hereditary disease in the Caucasian population. It is caused by genetic mutations of the CFTR (Cystic fibrosis transmembrane conductance regulator) gene, which result in unbalanced chloride secretion and chronic inflammation in mucosal tissues. Lungs are the primary organs affected in people with CF, who develop a persistent and non-resolving inflammatory status that leads to progressive loss of respiratory function[1]. Hence, it is crucial to understand pathways that may influence immune cells towards pathological or protective functions and determine if and how they can be regulated in response to treatments. To this end, we compared single-cell RNA transcriptomics in leukocytes derived from mice with CFTR F508del/F508del mutations (KO), wild-type mice (V), and mice treated with resolvin D2 (DKO), an investigation molecule with anti-inflammatory properties. Thus, this research proposal results in a context of high-dimensionality data. In high-dimensional datasets, the abundance of explanatory variables could make classical data exploration and analysis procedures less effective, both in unsupervised and supervised learning. For this reason, interpreting results is often challenging with high-dimensional data. For this reason, we need to use machine learning as a powerful tool for Classification.

Objective

We sought to identify the most relevant genes regulating leukocyte functions in CF and their modification following RvD2 treatment by applying a Random Forest algorithm to classify different mice conditions based on the 36,603 genes of leukocytes.

Methods

The "Seurat" R package was used to normalise the scRNA-seq data. scRNASeq samples were prepared using the 10X Genomics technology and sequenced with an Illumina sequencer. Reads were mapped to the mouse genome and count at features identified using the Cell Ranger for Single Cells gene Expression software. A Seurat data object was created from the Cell Ranger outputs on a per-sample basis. High-quality cellular barcodes are selected using a mixture of both adaptive and selective thresholds. To pass quality control in a standard analysis, a cellular barcode must match all three of the following criteria:

1) The number of unique molecular identifiers (UMIs) must be within three median absolute deviations (MADs) of the population median.
2) The number of expressed genes must be within three MADs of the population median.
3) The percentage of reads mapping to mitochondrial genes must be under 15%.

Samples were integrated for joint analysis. We use scaled data because it was z-score transformed, so the highly expressed genes will not dominate the model. The data matrix is 4154 cells x 36,603 genes. Due to the high dimensionality, genes with zero variance were

eliminated from the dataset, and the dataset result was 4154 cells x 4363. Removing correlated features, the final collection of features is 4325. Thus, because these data are lower correlated, we decided to use the 4363 not to lose biological information. A preliminary analysis with unsupervised UMAPs was carried out to understand the presence of common patterns between the different studies or groups using the Elbow method to determine the number of clusters. A machine learning algorithm was used to understand which genes are most important for group classification: random forest. First, the original dataset was split into training and testing using 80% of the samples to make up the training and the remaining part (20%) the test dataset. The random forest was constructed with 300 trees, removing missing data and using mtry=$\sqrt{(4363)}$. The relative importance of the variables was assessed using the Mean Decrease Accuracy Index, and the classifier performance is reported in terms of the confusion matrix and AUC with a 95% confidence interval (CI).

Results

UMAP visualization identified 13 clusters of leukocytes in mice lungs belonging to the following main families: neutrophils, macrophages, T cells and B cells. Neutrophils and macrophages in their clusters showed heterogeneity in the UMAP visualization, indicating the existence of cellular subtypes that differ based on their gene expression (Panel A).

Panel B shows the importance of the first 20 genes features. Figure 2 reports the gene's important features. In particular, the Mean Decrease Accuracy identified the following features involved in biological processes:

1. Xist (Inactive X specific transcripts, ENSMUSG00000086503) is a long non-coding RNA involved in cell differentiation that promotes inflammation in chronic obstructive pulmonary disease [2].
2. S100A9 (S100 calcium binding protein A9/calgranulin B, ENSMUSG00000056071) is a protein released by neutrophils. It has been identified in sputum from people with CF and is involved in the regulation of the inflammatory response and its resolution by suppressing cytokines[3].
3. Srgn (Serglycin, ENSMUSG00000020077) encodes for a proteoglycan implicated in the adhesion and migration of leukocytes to inflamed tissues[4].

Panel C reports the Confusion Matrix and AUC (95%CI) for Random Forest on the test set. The AUC is 0.831 (0.8041-0.8562). An AUC between 0.7 and 0.9 indicated good diagnostic efficacy [5].

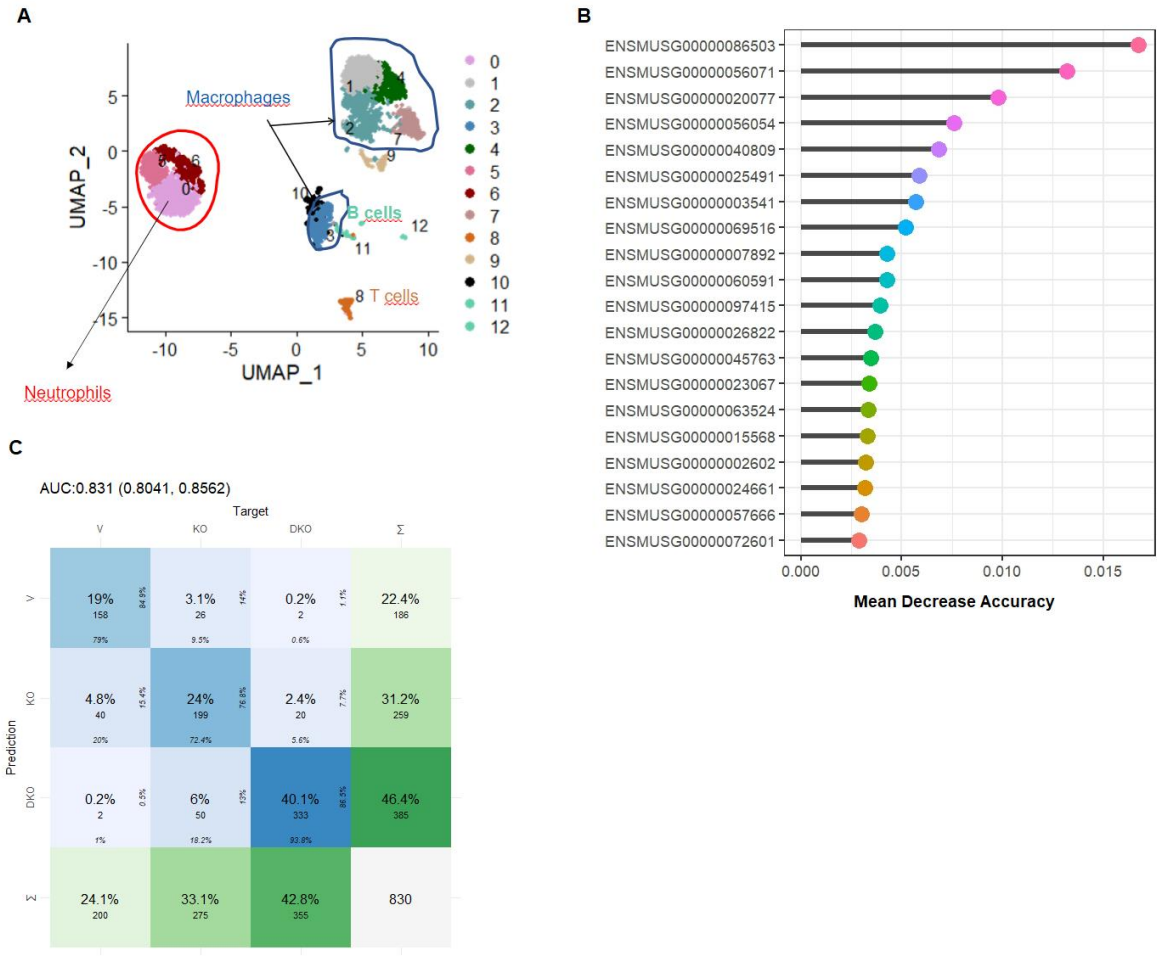

Figure 1. Unsupervised UMAP and Sample Contribution to Clusters. The number of clusters determined by the elbow method suggests 13 clusters.

Conclusions

Hence, our machine-learning approach identified genes involved in key steps of the inflammatory response that can serve as classifiers of the functions of leukocytes in CF lungs.

References

1. Miller, P.W., Hamosh, A., Macek Jr, M., Greenberger, P.A., MacLean, J., Walden, S.M., Slavin, R.G., Cutting, G.R., Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Gene Mutations in Allergic Bronchopulmonary Aspergillosis. Am. J. Hum. Genet, 1996, 59, 45.

2. Wang, W., Min, L., Qiu, X., Wu, X., Liu, C., Ma, J., Zhang, D., Zhu, L., Biological Function of Long Non-Coding RNA (LncRNA) Xist. Front. Cell Dev., Biol., 2021, 9.
3. Sreejit, G., Abdel Latif, A., Murphy, A.J., Nagareddy, P.R., Emerging Roles of Neutrophil-Borne S100A8/A9 in Cardiovascular Inflammation. Pharmacol. Res. 2020, 161, 105212, doi:10.1016/j.phrs.2020.105212.
4. Korpetinou, A., Skandalis, S., Labropoulou, V., Smirlaki, G., Noulas, A., Karamanos, N., THEOCHARIS, A. Serglycin: At the Crossroad of Inflammation and Malignancy., Front. Oncol., 2014, 3.
5. DeLong, E.R., DeLong, D.M., Clarke-Pearson, D.L., Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach., Biometrics, 1988, 44, 837–845.