Representation and quantification Of Module Activity from omics data with rROMA

Matthieu Najm[1,2,3#], Matthieu Cornet[1,2,3#], Luca Albergante[1,2,3], Andrei Zinovyev[1,2,3], Isabelle Sermet-Gaudelus[4,5,6], Véronique Stoven[1,2,3], Laurence Calzone[1,2,3] and Loredana Martignetti [1,2,3 *]

[1] INSERM U900, 75428 Paris, France, [2] Center for Computational Biology, Mines ParisTech, PSL Research University, 75006 Paris, France, [3] Institut Curie, PSL Research University, 75248 Paris, France, [4] Faculté de Médecine, Université de Paris, Paris, France, [5]Institut Necker Enfants Malades, INSERM U1151, Paris, France, [6]AP-HP. Centre - Université Paris Cité; Hôpital Necker Enfants Malades, Centre de Référence Maladie Rare - Mucoviscidose, Paris, France.

[#] These authors contributed equally to this work;
[*] To whom correspondence should be addressed: loredana.martignetti@curie.fr

Abstract
The efficiency of analyzing high-throughput data in systems biology has been demonstrated in numerous studies, where molecular data, such as transcriptomics and proteomics, offers great opportunities for understanding the complexity of biological processes.
One important aspect of data analysis in systems biology is the shift from a reductionist approach that focuses on individual components to a more integrative perspective that considers the system as a whole, where the emphasis shifted from differential expression of individual genes to determining the activity of gene sets.
Here, we present the rROMA software package for fast and accurate computation of the activity of gene sets with coordinated expression. The rROMA package incorporates significant improvements in the calculation algorithm, along with the implementation of several functions for statistical analysis and visualizing results. These additions greatly expand the package's capabilities and offer valuable tools for data analysis and interpretation. It is an open-source package available on github at: www.github.com/sysbio-curie/rROMA.
Based on publicly available transcriptomic datasets, we applied rROMA to cystic fibrosis, highlighting biological mechanisms potentially involved in the establishment and progression of the disease and the associated genes. Results indicate that rRoma can detect disease-related active signaling pathways using transcriptomic data. The results notably identified a significant mechanism relevant to cystic fibrosis, raised awareness of a possible bias related to cell culture, and uncovered an intriguing gene that warrants further investigation.

Contact: loredana.martignetti@curie.fr

Introduction

The use of high-throughput molecular techniques, such as transcriptomics and proteomics, is becoming increasingly easier with the improvement of data acquisition tools, leading to a drastic decrease in the costs associated with such analyses. This allows for precise measurement of the molecular profiles of biological systems at several levels. However, the amount of data produced during such experiments is very important and requires the use of dedicated softwares and algorithms

to analyze them. Moreover, the ability to interpret the data in terms of biological processes becomes a crucial issue. Dedicated analyses are needed to synthesize and transform the data into valuable biological information [1].

A commonly employed approach in genomics involves comparing measurements at the individual gene or protein level to identify distinctive markers indicative of specific disease states (biomarkers) or genes that play a causal role in the studied disease [2].

Nonetheless, in numerous systemic diseases, the disruption of a signaling pathway can arise from distinct genes within that pathway, and these gene alterations may vary from one patient to another. For example, in cancer It has become apparent in recent years that the same pathways are affected by defects in different genes and that the molecular profiles of patient samples are more similar at the pathway level than at the individual gene level [3].
Therefore, quantification of gene set activity from transcriptomic or proteomic measurements is now widely used to transform gene-level data into associated sets of genes representing biological processes [4, 5, 6]. By employing gene set-based approaches in the analysis of omics data, it becomes possible to capture valuable biological insights that would otherwise remain undetectable when solely focusing on individual genes.
In this study, we developed an algorithm, implemented as an R package called *rRoma*, which was designed to quantify the activity of sets of genes characterized by their involvement in a common functional role. These gene sets can thus correspond to genes with the same functional activities, genes regulated by the same motifs, genes belonging to the same signaling pathway or genes forming a group of frequently co-expressed genes. The underlying hypothesis of rRoma is to assess the activity of a gene set by determining the maximum amount of one-dimensional variance, which is represented by the first principal component (PC1) derived from the gene within the set. This quantity is proportional to the influence of a single latent factor on the gene expression within the gene set and reflects the variability of this factor's activity across the studied samples. This setting corresponds to the uni-factor linear model of gene expression regulation [7].

Naive quantification of the activity of a gene set often consists in calculating the mean or median of the expression of the gene set in each sample. Alternatively, it may rely on a single marker gene that represents the overall activity of the gene set. In contrast, *rRoma* differs from these approaches by its ability to effectively model scenario where individual genes within the set do not contribute equally to its activity. This is particularly relevant when some genes have a more significant impact than others, such as transcription factors downstream of signaling pathways.
Furthermore, rRoma is well-suited for cases where some genes exhibit a correlation of opposite sign with the activity of the gene set. In such situations, the first component of rRoma can capture this effect, whereas a simple averaging approach would not yield effective results.

Other more complex gene set quantification methods have already been proposed to compute the activity of a gene set by calculating the first principal component of the expression matrix restricted to the genes in the gene set [8]. In the study by Bild et al [9], a similar strategy was exploited to define activity of several cancer-related pathways on a large collection of human cancer transcriptomes. In another study by Fan et al. [10], the authors suggested the notion of an *overdispersed* pathway in the context of single-cell transcriptomic analysis. Other methods have been developed to estimate the

activity scores of gene sets in individual samples, such as the extension to a single sample of GSEA (ssGSEA) [11] or OncoFinder [6]. Our algorithm expands the repertoire of existing methods by introducing unique functionalities that are increasingly relevant in various contexts of systems biology. One distinguishing feature of the algorithm is that it computes a p-value that denotes the significance of the gene set's activity. This reflects the probability of obtaining the observed activity for a specific gene set by chance. The rROMA algorithm uses a random gene set procedure to generate a null distribution for the L_1 amount of variance explained by the PC1 and calculates the p-value by comparing the observed L_1 to the null distribution. Usually, a p-value threshold of 0.05 is employed to determine the significance of the gene set's activity.

In addition, the algorithm estimates the statistical significance of the distribution of samples along the first component for a gene set in two ways: it distinguishes between *shifted* and *over-dispersed* gene sets. A *shifted* set of genes corresponds to the situation where the median expression of all the genes in the gene set is significantly different from the one of all the genes studied, i.e. that the gene set shows a particularly high expression in at least one sample (see Fig 1A). *Over-dispersion* of a gene set corresponds to the situation where the amount of variance explained by PC1 calculated for only the genes in that gene set is significantly greater than the variance of a randomly selected set of genes of the same size. Thus, overdispersion means greater variability in a set of genes among the considered samples (see Fig 1B). The fact that rROMA distinguishes these two situations is particularly useful because, in many cases, the activity of a gene set does not correspond to overdispersion of the module in the global gene expression space but to a shift of the genes in a particular direction. Therefore, analysis of shifted gene sets can highlight findings that would not be identified with overdispersion analysis alone.

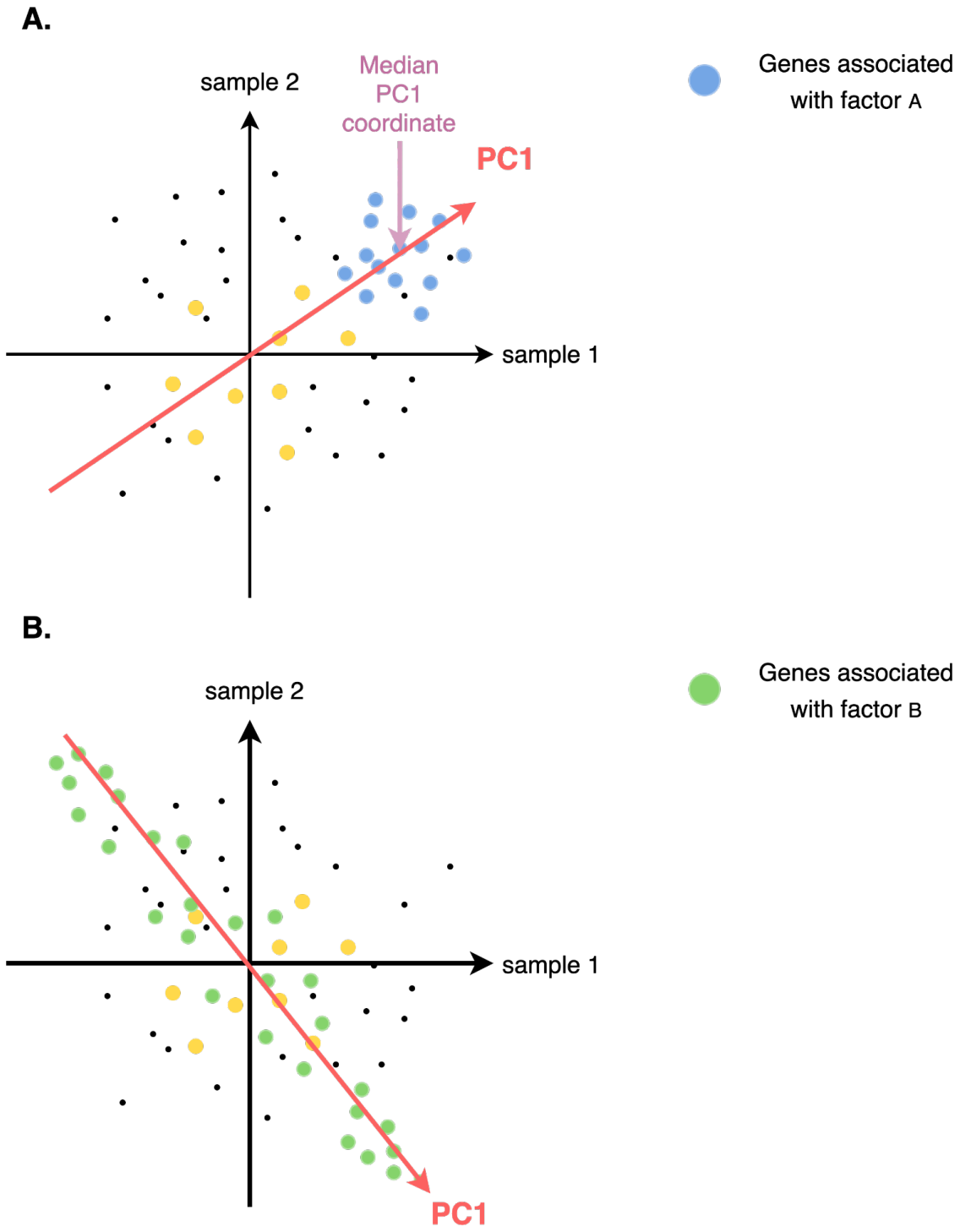

Figure 1
Representation of gene sets in the case of two samples. Each dot represents one gene, its horizontal (resp. vertical) value corresponding to its expression in sample 1 (resp. sample 2). Genes associated with latent Factor A are plotted in blue and the corresponding PC1 direction is plotted in red (A). This example corresponds to a shifted pathway, as assessed by a median of gene projections onto PC1 direction far from the origin of the distribution. Genes associated with latent Factor B are plotted in green (B) and the corresponding PC1 direction is plotted in red. This example corresponds to an

overdispersed pathway, as the PC1 is well aligned with the dots' distribution. Genes in yellow are neither overdispersed nor shifted, as PC1 explains a relatively small fraction of variance (not represented on the figure) and the median of projections onto PC1 is close to the origin for this group of genes.

Importantly, the algorithm provides the activity level of the gene set for each individual sample and does not require a predefined labeled classification of the samples into various conditions or groups. These activity values can be subsequently compared, bringing to light the heterogeneity present in the dataset in relation to the analyzed gene sets. This may be useful to define groups of samples or patients, when such stratification is unknown.

The rROMA algorithm is specifically designed to handle situations where genes within a gene set do not equally contribute to its activity. The unifactor model underlying rROMA presumes that an unobserved factor (i.e. a latent factor) acts on the gene expression variables observed in the gene set, and that this action is characterized by the calculated weights. The weights indicate the strength and direction of effect, and can have opposite signs, as in the case in which a transcription factor has an activating action on some genes of the set and a repressor action on others. Identification of the genes displaying the stronger contribution to a gene set's activity allows data analysis in the context of network modeling. Indeed, the output of rROMA, in particular the top weighted genes of significantly active gene sets, can be interpreted as nodes comprising genes and proteins of importance in the system under investigation, which can be used to construct mathematical models.

The package rRoma is an evolution of the algorithm ROMA, a program originally developed in Java [12] . The new rROMA package incorporates significant improvements in the calculation algorithm, along with several functions for statistical analysis and graphical visualization of results. These additions greatly expand the package's capabilities and offer valuable tools for data analysis and interpretation. It is an open-source package available on github at: www.github.com/sysbio-curie/rROMA.

As an example of its interest to study complex diseases, we applied rROMA to publicly available transcriptomic datasets in the context of Cystic Fibrosis (CF). CF is a genetic disease caused by mutations in a single gene, the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene, coding for the CFTR protein that functions as a chloride channel in various epithelial cells, including airways epithelial cells. Absence of a functional CFTR protein causes further functional deregulations in various biological pathways, leading to various deleterious phenotypes observed in CF patients that cannot be related to the chloride channel function in a direct manner. Analysis of deregulations at the pathways level may provide a better understanding of the molecular basis of some of the CF disease phenotypes.
Therefore, we applied rROMA to investigate pathway activities in airway epithelial cells from CF patients and from healthy donors. The analysis shows that rRoma can identify biological pathways associated with diseases from transcriptomic data, allowing both a clearer interpretation of high-throughput data from a biological point of view, and the interpretation of molecular changes in a functional way. Results highlighted a relevant mechanism in the context of CF, a potential bias due to cell culture, and an interesting gene that could be studied further. The analysis workflow is

schematized in Figure 2. A detailed vignette to reproduce all the analysis presented here is also available on the gitHub containing the source code of the algorithm.

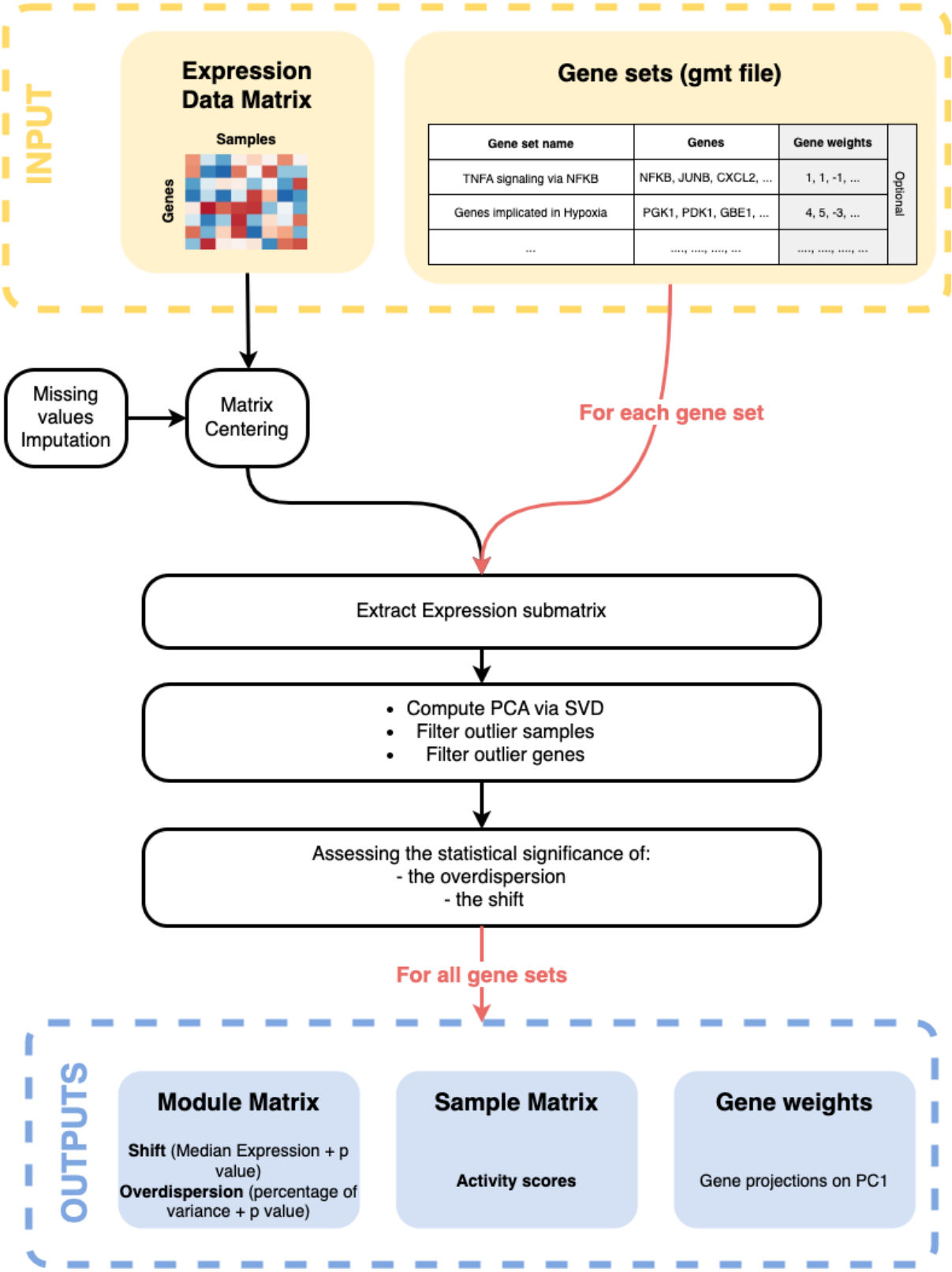

Figure 2
Schematic diagram illustrating the workflow of the rROMA algorithm.

Methods

First Principal Component and the Simplest Uni-Factor Linear Model of Gene Expression Regulation

The main idea of rRoma is based on the simplest uni-factor linear model of gene regulation in which it is assumed that the expression of a gene *G* in sample *S* is proportional to the activity of one latent biological factor *F* (which can be a transcription factor or any other endogenous or exogenous factor affecting gene expression) in sample *S* with positive or negative (response) coefficient. Within this model, the expression of a gene *G* in a sample *S* is proportional to the activity of a factor *F*, so that we can write:

Expression(gene *G*, sample *S*) $\approx \alpha^F_G$ Activity ^F_S

where α^F_G is the response coefficient of the gene *G* to the factor *F* and Activity ^F_S is the activity of factor *F* in the sample *S*. These two values can be easily determined by considering the first component of the PCA of the genes of the considered gene set in the space of the samples. In this case, the vector containing the activities of the different samples corresponds to the first eigenvector, i.e. the first column of the weight matrix. The vector containing the response coefficients of the different genes in the considered gene set corresponds to the projection of the genes onto the first component.
As many other applications of PCA, rRoma uses singular value decomposition to speed up the calculation. However, to be valid, the use of this method requires centering the data beforehand.
We define a dataset made up of *n* individuals, described by *p* numerical variables, which we represent by a matrix X of *n* x *p* size. In this matrix, the *i* th column corresponds to the vector x_i of observations of the variable *i*. The PCA consists in determining the eigenvalues and the eigenvectors of the covariance matrix C of X. This can be written as:

$$C = (X-X_{mean})\,(X-X_{mean})^T/(n-1)$$

Where X_{mean} contains the mean values of x_i for each column. When data are centered, this equation can be simplified as $C= XX^T/(n-1)$. Therefore, we have a symmetric matrix that can be diagonalized. If we denote C as the covariance matrix of X, there exists a diagonal matrix L and a matrix W such that $C = WLW^T$. Following the SVD theorem, any matrix X can be written as the product of a diagonal matrix *S* and two matrices *U* and *V* such that: $X = USV^T$. By analogy, the eigenvectors of the X matrix are obtained from the V matrix, and its eigenvalues are proportional to the diagonal of the S matrix. The computation of the SVD is therefore equivalent to computing the covariance matrix of X. The former is also much faster, and there are algorithms that allow it to accelerate even more by focusing only on the first columns of V.
rRoma uses the R irlba package [13], which allows to focus on the first principal component, by computing only the first column of V and the first value of S. However, if the data are not centered, the simplification is no longer valid, and we must now consider the fact that the mean is not zero. The rRoma algorithm therefore starts by centering the data of the global matrix.

Pre-processing of data for rROMA analysis

The input format for gene or protein expression for rRoma is a tab-delimited text file with columns corresponding to biological samples, and rows corresponding to genes or proteins. The first row is assumed to contain the sample identifiers while the first column is assumed to contain the non-redundant gene or protein names. If the data table contains missing values, they can be imputed using an approximation of the data matrix with missing values by a full matrix of lower rank. To do this, the user must specify the rank of the approximate full matrix that he wants to use. Then, the principal components are calculated up to the specified rank, using an algorithm capable of working with missing data [14]. This PCA decomposition is used to construct the full approximate matrix of lower rank, from which the missing values of the original data are imputed. In the rest of the algorithm, the full imputed matrix is used.

Orientation of the PC1

In standard PCA, all components are calculated with an undefined sign of orientation: there is mirror symmetry, which makes it difficult to determine whether a given set of genes is over- or under-activated. In *rRoma*, several methods exist to solve this ambiguity.
If knowledge exists about the role of a gene in a given collection, we recommend using it by associating a sign with the effect of the gene in the collection: negative for an inhibitor, and positive for an activator, for example. *rRoma* then uses the information about these signs to choose the orientation that maximizes the number of genes associated with a positive sign whose projection in PC1 is positive, and the number of genes associated with a negative sign whose projection in PC1 is negative.
Although less efficient, other methods of orienting the first component exist in the case where there is no a priori knowledge about the genes in the reference gene sets. The most efficient method consists in considering only the genes associated with the most extreme weights according to the first component (according to a percentage defined by a modifiable hyperparameter), and then summing these weights. If the result is negative, then the orientation of the PC1 is reversed. The principal behind this method is to orient the first component according to the most contributing genes for the gene set studied.

Filtering of outlier samples

Measurements may have been performed incorrectly in some samples and keeping them may lead to erroneous results from *rRoma*. By default, in the algorithm, no sample filtering is performed, as the matrix in input is assumed to contain only correct samples.
However, sample checking can be activated by a hyperparameter, and two filtering steps are then performed. First, the algorithm ensures that a similar number of genes is detected in all samples, and the allowed difference threshold is defined manually. Then, the samples are projected in the gene space, and rROMA ensures that no sample is too far from the others. The number of PCs used to perform the filtering and the allowed difference threshold are defined by two hyperparameters. It is also possible to perform only one of the two filtering steps.

Filtering of outlier genes

The calculation of PC1 can be affected by the presence of an outlier gene in the dataset. This outlier could indeed artificially affect the PC1. In order to increase the robustness of the PC1 calculation, we

use in rRoma the "leave-one-out" cross-validation approach [15]. This method works as follows: first, for each gene in the considered gene set, we calculate the percentage of variance L_1 explained by PC1 when the gene is removed from the dataset. Each gene is then associated with a L_1 value. In a second step, the distribution of these L_1 values is centered and reduced to obtain z-scores. In a third step, all genes whose associated z-score above a threshold value set in the algorithm are considered as outliers and removed from the analysis. The idea behind this method is to identify genes that have too much impact on the PC1 on their own: if the percentage of variance explained by PC1 increases significantly in the absence of a single gene, this means that this gene does not follow the alignment of all the others. It is then considered as an outlier gene.

When a gene is considered as outlier for a given gene set, it is only removed from that gene set. This default behavior can be modified by a hyperparameter, so that genes are completely removed from all analyses, as soon as they are considered outliers in at least one gene set. However, in some cases, these genes may still be of interest for analysis. Additional analysis steps for these genes are therefore available in rRoma.

In particular, the greater the number of gene sets analyzed containing a given gene, the more likely it is that the gene will be considered as an outlier in at least one gene set, and therefore be removed from the analysis. In order to avoid such abusive withdrawals, a Fisher test is performed. This consists of comparing the average proportion of collections in which the genes in the analysis are considered as outliers to this same proportion for a particular gene. If the proportion of outliers is close to the average proportion for all the genes, then it is no longer considered an outlier. Conversely, if the proportion of aberrations is significantly higher (threshold determined by a modifiable hyperparameter), then it is still considered an outlier.

Instead, a gene present only in a small number of gene sets may be important for the understanding of these gene sets. In such cases, it is not desirable to remove it, even if it is considered an outlier. Thus, if a gene is present in less than a defined number of gene sets (set by a modifiable hyperparameter), then it is not considered an outlier, regardless of the results of the “leave-one-out” approach for it.

Finally, it is possible that a significant proportion of genes are considered outliers for a given gene set. However, removing too many genes can totally distort the detectable behavior for a collection. A final filter therefore exists to limit the maximum proportion of genes that can be considered as aberrant for a given gene set and removed from the analysis (the proportion is set by a modifiable hyperparameter). In this way, only genes with the most extreme leave-one-out variations are effectively considered outliers for this gene set.

Interpretation of results

The core of the *rRoma* algorithm starts once these pre-processing steps, involving imputing missing values, removing outlier samples, centering the data at the global level, and then removing outlier genes from each gene set have been performed. Each gene set is then analyzed separately.

First, the algorithm computes the first principal component of the genes in the gene set in the sample space. Two measures of interest are then considered: the percentage of variance explained by the first component alone, and the average expression value of the genes projected onto this first component.

The same measures are then performed for a hundred of randomly generated gene sets of same size, which will constitute the reference null distribution. As explained in the previous section, the use of

SVD is only justified if the data are centered. If this condition is not met, the median projection on the first component will not be zero. For this reason, the average value of the projection of the genes in the gene set under study onto PC1 is compared to the values obtained for the random collections of genes that constitute the null distribution and are therefore assumed to be centered. These values can be positive or negative, but since we are interested in the deviation from the center, the absolute values are considered. If less than 5% (value defined by a tunable hyperparameter) of the values obtained in the null distribution are lower than the one obtained for the gene set under study, the latter is said to be *shifted* (ppv Median Exp <0.05). This means that the average expression of the genes in the gene set is different from the average expression of all genes for at least one sample. In such a case, additional analyses are needed to determine precisely the origin of this shift. If the collection is not shifted, then the centered data assumption is considered valid. We are then interested in the percentage of variance explained by the first component. If less than 5% (value defined by a modifiable hyperparameter) of the values obtained in the null distribution are lower than the one obtained for the gene set under study, then the latter is said to be *overdispersed* (ppv L1 <0.05).

Analysis of a shifted gene set

For a *shifted* gene set, the measurement of sample activity is particularly important. It determines which samples are responsible for the shift of the gene set. If the samples were already separated into several conditions prior to the analysis, it is possible to verify that this separation into groups is indeed responsible for the shift, by checking that the activity scores are significantly different between the conditions. Conversely, if the conditions are not known a priori, it is possible to determine new groups by performing a hierarchical clustering analysis on the gene sets that are shifted, and thus determine potential new groups of interest from the analysis.

Analysis of an over-dispersed gene set

The analyses mentioned above for the case of shifted gene sets are also valid for *over-dispersed* gene sets. But in the case of the latter, the analysis of gene weights also becomes interesting.

The genes associated with the highest weights are the driving force in the activity scores of the gene sets. They summarize the information of the gene sets, which can be particularly useful, especially for interpretations using systems biology approaches.

The analysis of the sign of the genes' weights is also particularly interesting. For example, gene sets containing both activators and inhibitors can be highlighted by rRoma by being overdispersed, and the associated genes highlighted. Such gene sets would not be detected as *overdispersed* by methods based on the average expression of genes in the samples.

Optimization of the calculation of null distributions

In practice, it is often necessary to test many different gene sets available in large reference databases such as KEGG [16] or MSigDB [17]. Estimating the null distribution for each set of genes can lead to very time-consuming calculations. rRoma does not compute the significance scores of overdispersion and shift for all gene sets, but approximates them on a predefined grid of values, depending on the size of the considered gene set. Indeed, since these two values are dependent on the size of the gene set, it is not possible to use the same null distribution for all gene sets. In order to rapidly estimate the

importance of the over-dispersion and shifting scores, rROMA constructs null distributions for a representative list of gene set sizes. These are selected to be uniformly distributed in the logarithmic scale between the minimum and maximum size of the reference database. For a given gene set, the null distribution that is closest in size in the log scale is then chosen.

Case study

rROMA identifies active signaling pathways in CF tissues compared to healthy donors

We applied rROMA to investigate the activity of pathways in airway epithelial cells from CF patients and healthy donors. More precisely, we compared the transcriptomes of primary cultures of airway epithelial cells from patients (N=6) with those of healthy controls (N=6), based on RNAseq data publicly available in the NCBI's GEO database, under the accession ID GSE176121 [18]. rRoma was run by specifying the pathway database to use and the expression matrix to analyze, as shown in the accompanying vignette.

Here, the Molecular Signature Database MsigDB hallmark gene set collection [19] was used, a gene set collection of 50 gene sets specifically curated to represent core biological processes and pathways that are commonly dysregulated in cancer. However, to provide a more complete view of the biological processes involved in a study, rRoma can be applied with different reference databases.

The results of rRoma highlight pathways that are provided in the *ModuleMatrix* output. Pathways with a *ppv Median Exp* lower than a given threshold were deemed as *shifted*, while those with a *ppv L1* lower than this threshold were *overdispersed*. The *Plot.Genesets.Samples* function allows for the visualization of activity scores for significantly shifted and overdispersed pathways across samples, in the form of a heatmap representation (see Figure 4). rROMA identified two shifted pathways, APICAL_SURFACE which is found to have higher activity scores in healthy controls than in CF patients, and FATTY_ACID_METABOLISM which has higher scores in CF patients than in controls, and an overdispersed pathway, COAGULATION, with higher activity scores in healthy controls than in CF patients.

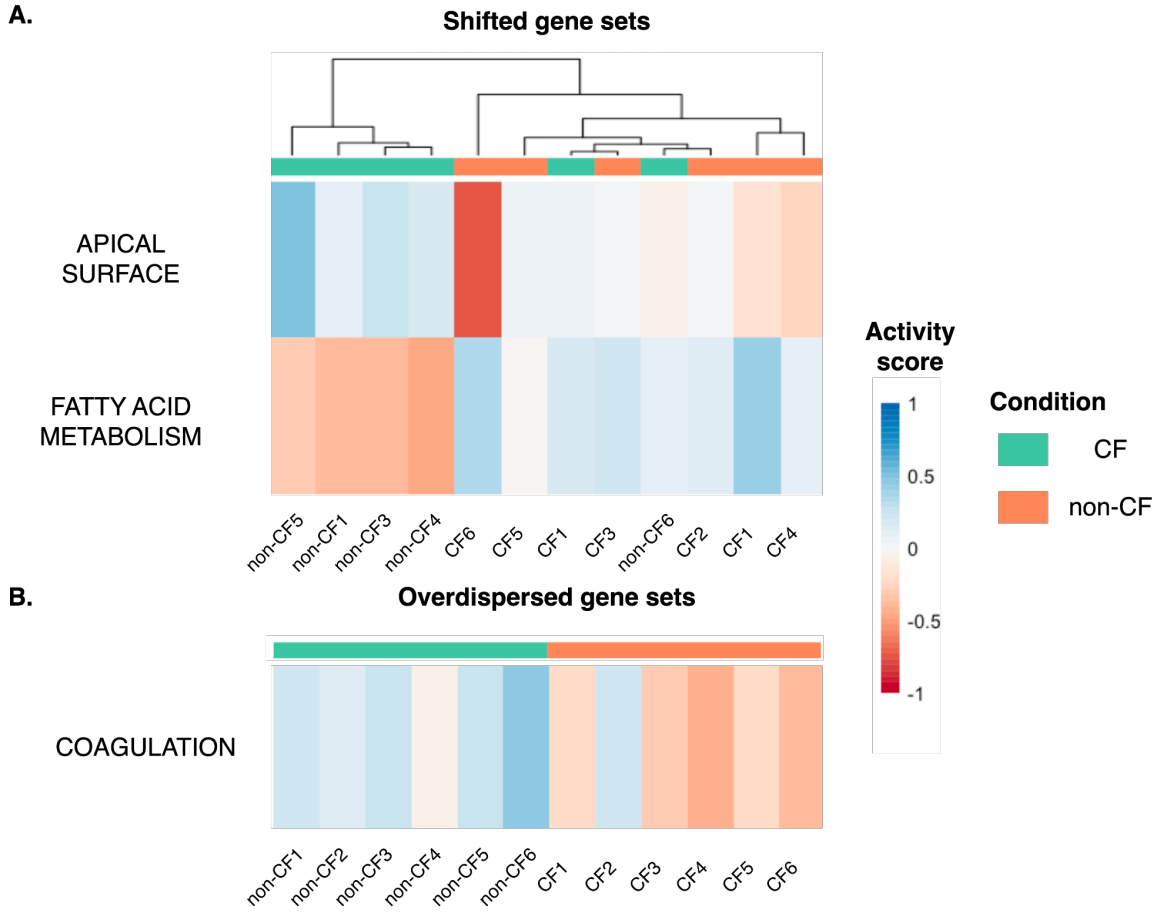

Figure 3
Heatmap of activity scores for gene sets identified as significantly shifted (A) or significantly overdispersed (B) in GSE176121 dataset. Samples are in columns, gene sets are in rows. Horizontal sidebar color encodes true class labels.

When sample groups have been pre-defined, as the two CF or control groups in our case, these groups can be compared based on the activity scores of the gene sets observed in the samples belonging to the two groups. Boxplot of the activity scores based on predefined groups can help differential analysis. In our study, shifted and overdispersed pathways behaved significantly differently in CF patients versus healthy controls, as shown in Figure 3. Alternatively, when the groups are not predefined, analyzing the shifted and overdispersed pathways can reveal clusters of samples exhibiting similar pathway activity.

The analysis of top contributing genes in each pathway also provides crucial information. The weights assigned to each gene in the PC1 vector allow to identify the key genes that most contribute to variations in pathways activities as those with the higher weights. For each pathway, gene weights are provided by the PlotGeneWeight function.
For example, plotting the gene weights for the COAGULATION pathway (see Figure 4) highlighted the GSN gene, encoding the protein *Gelsolin*, as the highest contributor to the activity score. Notably,

Gelsolin has been previously reported to play a role for CFTR activation [20,21] , and to promote mucus fluidification in CF [22].

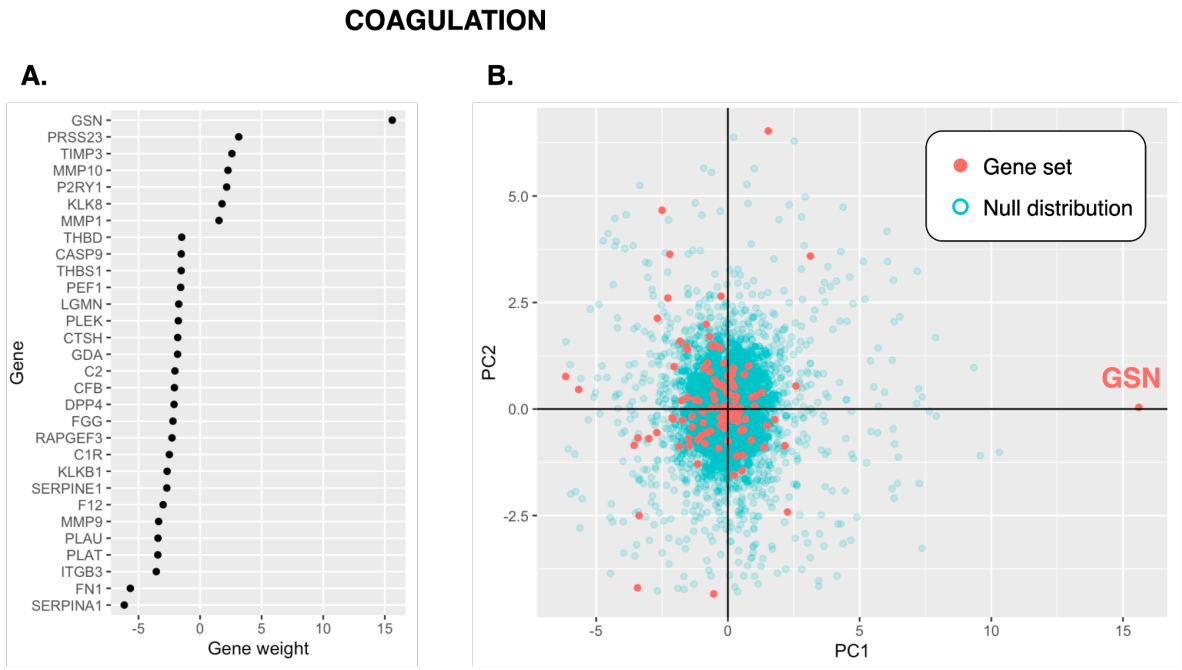

Figure 4
Plots illustrating the contribution of genes to the COAGULATION gene set activity score. The weights in panel A indicate the gene projections on PC1, limited to the genes that have the greatest contribution to the observed variation in the COAGULATION gene set. In panel B the genes of the COAGULATION gene set are represented in the PCA space. Red dots are genes from the gene set, blue dots show randomly selected genes used to generate a null distribution.

Finally, many hyperparameters can be specified and changed to modify rRoma speed, precision, or behavior regarding outlier detection. Details about all available hyperparameters are described in the vignette accompanying the software. The computational time required to run the algorithm typically depends on the number of considered pathways and their relative sizes. It also depends on whether parallelization is enabled. In the present work, the algorithm ran in about 3 minutes and 15 seconds on a MacBook Pro equipped with a 2,6 GHz Intel Core i7 6 cores processor. A single 60 genes pathway took roughly 5 seconds to be analyzed. Parallelization was not used, but this would have increased the speed of the analysis.

rROMA estimates cell type abundances from bulk transcriptomic data

In addition, our case study also demonstrates the ability of rROMA to investigate cell type abundance based on bulk transctiptomic data. Saint-Criq et al. [23] investigated the impact of two differentiation media on primary cultures of CF and non-CF airway epithelial cells, as determined by transcriptomic data. They built a gene signature for each cell type using the 50 most expressed markers derived from a single-cell RNA sequencing (scRNAseq) dataset [24]. They observed a significant overexpression of

genes belonging to the signature of the secretory cell subtype in cultures grown in one of the media (referred to as UNC), compared to the other medium (referred to as SC). Conversely, gene markers of the ciliated subtype were overexpressed in primary cultures grown using the SC medium compared to UNC medium.

We applied rROMA to the Saint-Criq RNA seq dataset to estimate cell type abundance in CF and non-CF samples. More precisely, the Plasschaert signature [24] for each cell types was used and gene reference gene sets, and the activity scores of these gene sets across the samples were represented in the form of a heat map. Samples were found to be clustered according to the differentiation medium in which they were grown, and our results confirmed the higher abundance of the ciliated cell subtype in UNC medium and higher abundance of the secretory cell subtype in SC medium (Figure 5A). We repeated the rROMA analysis using an alternative signature [25]. In contrast with the Plasschaert signature, the Okuda signature includes the most differentially expressed genes in each cell type, encompassing both overexpressed and underexpressed genes. As illustrated in figure 5B, rROMA analysis using a reference gene set corresponding to this alternative signature consistently revealed the same relative abundances of secretory and ciliated subtypes between UNC and SC growing media, as observed with the initial signature.
Thus, rROMA allows us to clearly highlight differences in cell-type abundances, facilitating the use of gene signatures that contain both upregulated and downregulated genes and thus potentially more accurate.

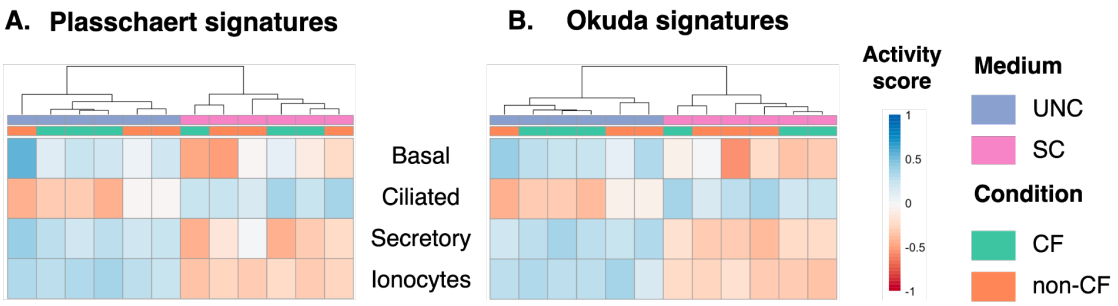

Figure 5.
Heatmap of rROMA scores obtained for Plasschaert (A) and Okuda (B) signatures of cell types in the Saint-Criq RNA seq dataset. Samples are in columns, gene sets corresponding to cell types are in rows. Horizontal sidebar color encodes true class labels.

Discussion

Quantifying the activity of biologically related gene sets is a commonly employed approach to extract valuable biological insights from high-throughput data. The use of gene sets as aggregated variables from molecular data enables the capture of biological information that may not be detectable when solely focusing on individual genes. To address this challenge, we introduced the rROMA algorithm. Based on a gene expression data matrix, this algorithm implements a linear model of gene regulation and efficiently and reliably quantifies the activity of gene sets by computing the first principal component (PC1), while also evaluating the statistical significance of this approximation.

We applied rROMA to CF transcriptomic datasets, highlighting some biological mechanisms potentially involved in the initiation or progression of the disease, and their associated genes. In our study, out of the 50 hallmark pathways tested, 3 were significantly active: FATTY_ACID_METABOLISM, APICAL_SURFACE, and COAGULATION. The FATTY_ACID_METABOLISM pathway has significantly different activity scores between CF patients and healthy donors. This pathway has been extensively studied in CF, and essential fatty acid deficiency is a well-known CF phenotype (for a review, see Strandvik, 2010). The APICAL_SURFACE can be related to another well-known hallmark of CF, i.e. a perturbation of airway surface secretory mucus content. Finally, the COAGULATION pathway, the only overdispersed pathway in our study, seems to be highlighted due to one specific gene with a very high associated weight, that is by far the most contributing gene to the activity score of this pathway: *gelsolin* (*GSN*). *Gelsolin* has been reported as playing a role for CFTR activation [20, 21], which suggests that the role of this gene in CF disease may be interesting to study in more detail. The goal of this use case was not to undertake a detailed systems biology approach of CF, which is beyond the scope of the present paper. In particular, it would require including additional transcriptional dataset to take more samples into account and increase the statistical power of our analyses, and to test several reference databases of gene sets.
However, overall, this case study illustrates that rROMA is able to identify disease-associated pathway deregulations from transcriptomic data, allowing a more comprehensive and functional interpretation of the data. It is also a versatile tool that can shed light on various biological questions such as highlight the key genes driving these deregulations, identify clusters of samples, study samples' cell-type composition, or other cellular changes in a broader biological perspective.

Acknowledgements
The authors would like to thank Pierre Gestraud for insightful discussions about the R implementation. This work was done with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm and with support from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute). MC was supported by Fondation pour la Recherche Médicale (FRM), and MN was supported by Vaincre La Mucoviscidose (VLM) grant number20190502488 and by Fondation Dassault Systèmes.

Conflict of Interest
None declared.

Author Contribution Statement

M.N., M.C., A.Z., L.A. and L.M.: conceptualization, methodology M.N. and M.C.: data curation, data analyses, formal analysis. L.M. and M.N.: Writing-original draft I.S.,V.S. and L.C.: scientific support, editing and content review.

References

[1] Hawkins, R. D., Hon, G. C., & Ren, B. (2010). Next-generation genomics: an integrative approach. *Nature Reviews Genetics*, ***11*(7),** 476-486.

[2] Barillot, E., Calzone, L., Hupe, P., Vert, J. P., & Zinovyev, A. (2012). *Computational systems biology of cancer*. CRC Press.

[3] Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics*, **11(12)**, 843-854.

[4] Levine, D. M., Haynor, D. R., Castle, J. C., Stepaniants, S. B., Pellegrini, M., Mao, M., & Johnson, J. M. (2006). Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome biology*, **7(10)**, 1-17.

[5] Ramos-Rodriguez, R. R., Cuevas-Diaz-Duran, R., Falciani, F., Tamez-Pena, J. G., & Trevino, V. (2012). COMPADRE: an R and web resource for pathway activity analysis by component decompositions. *Bioinformatics*, **28(20)**, 2701-2702.

[6] Buzdin, A. A., Zhavoronkov, A. A., Korzinkin, M. B., Venkova, L. S., Zenin, A. A., Smirnov, P. Y., & Borisov, N. M. (2014). Oncofinder, a new method for the analysis of intracellular signaling pathway activation using transcriptomic data. *Frontiers in genetics*, **5**, 55.

[7] Schreiber, A. W., & Baumann, U. (2007). A framework for gene expression analysis. *Bioinformatics*, **23(2),** 191-197.

[8] Tomfohr, J., Lu, J., & Kepler, T. B. (2005). Pathway level analysis of gene expression using singular value decomposition. *BMC bioinformatics*, **6,** 1-11.

[9] Bild, A. H., Potti, A., & Nevins, J. R. (2006). Linking oncogenic pathways with therapeutic opportunities. *Nature reviews cancer*, **6(9)**, 735-741.

[10] Fan, J., Salathia, N., Liu, R., Kaeser, G. E., Yung, Y. C., Herman, J. L., ... & Kharchenko, P. V. (2016). Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. *Nature methods*, **13(3)**, 241-244.

[11] Barbie, D. A., Tamayo, P., Boehm, J. S., Kim, S. Y., Moody, S. E., Dunn, I. F., ... & Hahn, W. C. (2009). Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature*, **462(7269),** 108-112.

[12] Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., & Zinovyev, A. (2016). ROMA: representation and quantification of module activity from target expression data. *Frontiers in genetics*, **7**, 18.

[13] Baglama, J. (2016). IRLBA: Fast Partial Singular Value Decomposition Method.

[14] Gorban, A. N., & Zinovyev, A. Y. (2010). Principal graphs and manifolds. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques* (pp. 28-59). IGI Global.

[15] Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* **2**, 1-758. New York: springer.

[16] Kanehisa, M. (2002). The KEGG database. In *'In Silico'Simulation of Biological Processes: Novartis Foundation Symposium* **247***,* 91-103. Chichester, UK: John Wiley & Sons, Ltd.

[17] Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27(12)**, 1739-1740.

[18] Rehman, T., Karp, P. H., Tan, P., Goodell, B. J., Pezzulo, A. A., Thurman, A. L., ... & Welsh, M. J. (2021). Inflammatory cytokines TNF-α and IL-17 enhance the efficacy of cystic fibrosis transmembrane conductance regulator modulators. *The Journal of clinical investigation*, **131**(16).

[19] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The molecular signatures database hallmark gene set collection. *Cell systems*, **1(6)**, 417-425.

[20] Cantiello, H. (2001). Role of actin filament organization in CFTR activation. *Pflügers Archiv*, **443**, S75-S80.

[21] Vasconcellos, C. A., Allen, P. G., Wohl, M. E., Drazen, J. M., Janmey, P. A., & Stossel, T. P. (1994). Reduction in viscosity of cystic fibrosis sputum in vitro by gelsolin. *Science*, **263(5149)**, 969-971.

[22] Bucki, R., Cruz, K., Pogoda, K., Eggert, A., Chin, L., Ferrin, M., ... & Janmey, P. A. (2015). Enhancement of Pulmozyme activity in purulent sputum by combination with poly-aspartic acid or gelsolin. *Journal of Cystic Fibrosis*, **14(5)**, 587-593

[23] Saint-Criq, V., Delpiano, L., Casement, J., Onuora, J. C., Lin, J., & Gray, M. A. (2020). Choice of differentiation media significantly impacts cell lineage and response to CFTR modulators in fully differentiated primary cultures of cystic fibrosis human airway epithelial cells. *Cells*, **9(9)**, 2137.

[24]Plasschaert, L. W., Žilionis, R., Choo-Wing, R., Savova, V., Knehr, J., Roma, G., ... & Jaffe, A. B. (2018). A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte. *Nature*, **560(7718)**, 377-381.

[25] Okuda, K., Dang, H., Kobayashi, Y., Carraro, G., Nakano, S., Chen, G., ... & Boucher, R. C. (2021). Secretory cells dominate airway CFTR expression and function in human airway superficial epithelia. *American journal of respiratory and critical care medicine*, **203(10)**, 1275-1289.