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 and proteomic 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

Keywords

Pathway activity, Principal Component Analysis, Transcriptomics, Proteomics

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 software 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

53 diseases, the disruption of a signaling pathway can arise from distinct genes within that pathway,
54 and these gene alterations may vary from one patient to another. For example, in cancer It has
55 become apparent in recent years that the same pathways are affected by defects in different genes
56 and that the molecular profiles of patient samples are more similar at the pathway level than at the
57 individual gene level [3]. Therefore, quantification of gene set activity from omics measurements is
58 now widely used to transform gene-level data into associated sets of genes representing biological
59 processes [4]. By employing gene set-based approaches in the analysis of omics data, it becomes
60 possible to capture valuable biological insights that would otherwise remain undetectable when
61 solely focusing on individual genes.
62 In this study, we developed an algorithm, implemented as an R package called rROMA, which was
63 designed to quantify the activity of sets of genes characterized by their involvement in a common
64 functional role. Gene set approaches have become very popular to summarize individual molecular
65 measurements into more interpretable pathways and biological processes. Gene sets can thus
66 correspond to genes with the same functional activities, genes regulated by the same motifs, genes
67 belonging to the same signaling pathway, target genes of a transcription factor or genes forming a
68 group of frequently co-expressed genes. The underlying hypothesis of rROMA is to assess the
69 activity of a gene set by determining the maximum amount of one-dimensional variance, which is
70 represented by the first principal component (PC1) derived from the gene within the set. This
71 quantity is proportional to the influence of a single latent factor on the gene expression within the
72 gene set and reflects the variability of this factor's activity across the studied samples. This setting
73 corresponds to the uni-factor linear model of gene expression regulation [5].
74 The strength of our method lies in its ability to calculate a p-value, a statistical measure that serves
75 as a crucial indicator of the gene set activity's significance. This p-value enables the prioritization of
76 biological processes that are pertinent to the underlying conditions in question. Another unique
77 feature of the rROMA algorithm and to our knowledge not implemented in any other algorithm
78 available in the literature is its ability to estimate 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. 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 will highlight findings that would not be identified with overdispersion analysis alone.

We applied rROMA to examine pathway activities in airway epithelial cells of individuals with Cystic Fibrosis (CF) and healthy individuals, shedding light on the molecular processes behind CF phenotypes. Moreover, we illustrate the use of rROMA in estimating cell type abundance through bulk transcriptomic data analysis. Here, rROMA enables the clear estimation in cell-type proportions, enhancing the potential accuracy of gene signatures that encompass both upregulated and downregulated genes. Finally, we show application of rROMA to a mass-spectrometry-based breast cancer proteomic dataset, showing its capability of identifying significantly dysregulated pathways in clinical breast cancer subtypes.

Results

Review of existing methods

Over the past decades, numerous tools have emerged to perform gene set analysis from omics data to explore pathway activity. Sample-wise pathway activity computation and global pathway enrichment scoring are two distinct approaches used in the analysis of biological pathways. The primary goal of sample-wise pathway activity analysis is to assess the extent to which a particular biological pathway is active or dysregulated in each individual sample. Methods for sample-wise analysis provide a quantitative measure of pathway activity for each pathway in each sample that can then be interpreted in terms of pathway activity profiles. Global pathway enrichment scoring, on the other hand, condenses the information from all samples into a single enrichment score, providing an overview of the pathway's significance or activity across the study. This approach is

105 commonly used in over-representation analysis (ORA) of differentially expressed genes [6], gene set
106 enrichment analysis (GSEA) [7] and similar methods that provide an aggregated score of pathway
107 enrichment between two biological conditions.
108 A simple method for sample-wise pathway activity quantification consists of calculating the mean or
109 median of the expression of genes annotated in a given pathway in each sample. Alternatively, one
110 can rely on the expression of a single marker gene that represents the overall activity of the pathway
111 in different samples. A more sophisticated gene set quantification method, called PLAGE (Pathway
112 Level Analysis of Gene Expression) was introduced by Tomfohr and colleagues [8] to compute the
113 activity of a gene set by calculating the first principal component (PC1) of the expression matrix
114 restricted to the genes in the gene set. In the study by Bild et al [9], this PCA-based strategy was
115 exploited to define activity of several cancer-related pathways on a large collection of human cancer
116 transcriptomes. This PCA-based method is efficient in modeling situations in which individual genes
117 within the pathway display different expression patterns across samples, either due to their unequal
118 contributions to pathway activation or to their involvement encompassing both upregulation and
119 downregulation. This is particularly relevant when some genes have stronger or opposite expression
120 variation than others, such as transcription factors. However, the PLAGE algorithm has some
121 limitations. Specifically, it does not provide a statistical assessment of whether the observed activity
122 in a particular pathway is statistically significant or merely occurring by random fluctuations.
123 Moreover, the standard PCA calculation implemented in PLAGE lacks the capability to identify
124 pathway activation when dealing with a specific gene configuration, which is referred to here as a
125 *shifted* gene set.
126 Many other methods have been developed to estimate pathway activity from transcriptomic data. A
127 recent comprehensive and systematic comparison of existing computational tools evaluates
128 different approaches [10]: (i) ranking-based methods such as single sample GSEA (ssGSEA) [11],
129 GSVA [12], (ii) PCA-based methods, (iii) pathway topology-based methods such as PARADIGM [13]
130 and many others. In this study, PCA-based methods have been shown to perform best in preserving

131 characteristics of the original gene expression data while transforming the transcriptome data at the
132 pathway level whereas ranking-based methods show high robustness against noise. In Table 1, we
133 reported a feature-wise comparison of rROMA to existing bulk-based pathway analysis tools. This
134 table comprises methods that quantify pathway activity on a sample-wise basis, considering only the
135 expression matrix. Other methods exist that necessitate further input such as prior knowledge or
136 inferred networks to infer pathway activity, such as PARADIGM [13], VIPER [14], PROGENy [15].
137 Recently, multiple algorithms for pathway activity analysis of single cell transcriptomic data have
138 been proposed, including PCA-based methods such as PAGODA [16] and MAYA [17]. Benchmark
139 studies [18,19] comparing their performances in interpreting sparse and noisy single cell
140 transcriptomic datasets in terms of functional pathways suggest that bulk-based pathway analysis
141 tools can be applied to single cell datasets, partially outperforming dedicated single-cell tools.
142

Method	Type of algorithm	Overdispersed gene-sets	Shifted gene-sets	Statistical significance	A priori fixed gene weights	Outlier detection	Code availability	Ref
z-score	Standardized expression		X				R	[20]
GSVA	Ranking-based		X				R	[12]
ssGSEA	Ranking-based		X				Java, R	[11]
PLAGE	PCA-based	X					R	[8]
GO-PCA	PCA-based	X		X			Python	[21]
PCGSE	PCA-based		X	X			R	[22]
rROMA	PCA-based	X	X	X	X	X	R	
Pathifier	Principal curve	X	X				Java, R	[23]

143
144 Table 1. Feature-wise comparison of rROMA to existing tools

145

146 **Implementation of rROMA**

147 The rROMA algorithm expands the repertoire of existing PCA-based methods for pathway activity
148 quantification by introducing unique functionalities, notably (i) a statistical assessment of pathway
149 activity measurement, enabling the determination of the biological relevance of the results and
150 facilitating their prioritization; (ii) a PCA computation with a fixed center which allows to detect two

151 different configurations of pathway genes in the global gene expression space, referred to as *shifted*
152 and *over-dispersed* gene sets.
153 The package rROMA is an evolution of the algorithm ROMA, a program originally developed in Java
154 [24]. The new rROMA software, developed in R language, incorporates significant improvements
155 including an automatic procedure to not only distinguish between *shifted* and *over-dispersed* gene
156 sets but also automatically detect outlier genes that can affect the PCA computation. Moreover, the
157 new software includes several novel functions for user-friendly downstream statistical analysis and
158 graphical visualization of results that are documented in a software accompanying vignette. These
159 additions greatly expand the package's capabilities and offer valuable tools for data analysis and
160 interpretation. The rROMA software is open-source and available on github at:
161 www.github.com/sysbio-curie/rROMA. The workflow of the rROMA algorithm is schematized in
162 Figure 1. A detailed vignette to reproduce an example of analysis is available on the github
163 containing the source code of the algorithm.
164 The algorithm requires as input a genome-wide expression data matrix and a gene set annotation
165 file in GMT (Gene Matrix Transposed) format. Optionally, it is possible to incorporate prior
166 knowledge about the significance of certain genes within a pathway by initially assigning gene
167 weights in a given pathway within the GMT file. These "adjusted weights" will then substitute the
168 weights acquired from the PC1 calculation for those specific genes.
169 The rROMA algorithm starts by imputing missing values in the data matrix and centering values, if
170 not done beforehand. Then, each gene set is analyzed separately. A multistep procedure is
171 implemented for (i) extracting the expression submatrix corresponding to a given gene set, (ii)
172 quantifying the PC1-based activity values robust to outlier samples and outlier genes and (iii)
173 assessing the statistical significance of the activity values.
174 First, the algorithm extracts the expression submatrix corresponding to a given gene set. On this
175 submatrix, it computes the PCA in the sample space via the SVD algorithm [25]. Two measures of
176 interest are then considered: the percentage of variance explained by the PC1 (referred to as *L1*) and

177 the median value of the gene projections onto PC1 (referred to as *Median Exp*). Based on these two
178 values, pathway activity significance is estimated in two distinct ways through the definition of
179 *shifted* and *overdispersed* gene sets (Figure 2). A *shifted* set of genes corresponds to the case where
180 the median expression (*Median Exp*) of all the genes in the gene set is significantly different from the
181 one of all the genes studied, i.e. that the gene set shows a particularly high expression in at least one
182 sample (Figure 2A). *Over-dispersion* of a gene set corresponds to the situation where the amount of
183 variance explained by PC1 (*L1*) calculated for only the genes in that gene set is significantly greater
184 than the variance of a randomly selected set of genes of the same size. Thus, overdispersion means
185 greater variability in a set of genes among the considered samples (Figure 2B). An automatic
186 procedure for the identification of outlier genes and samples is applied to increase robustness of the
187 PC1 computation (see Methods).
188 The same measures *L1* and *Median Exp* are computed for *N* randomly generated gene sets of the
189 same size, which will constitute the reference null distribution. The *Median Exp* value is compared to
190 the distribution of *Median Exp* values obtained for the random gene sets. If less than 5% (value
191 defined by a tunable hyperparameter) of the values obtained in the null distribution are lower than
192 the one obtained for the gene set under study, the latter is said to be *shifted* (ppv Median Exp
193 <0.05). If less than 5% (value defined by a modifiable hyperparameter) of the *L1* values obtained in
194 the null distribution are lower than the one obtained for the gene set under study, then the latter is
195 said to be *overdispersed* (ppv L1 <0.05). The ppv L1 and ppv Median Exp values are corrected for
196 multiple testing by the Benjamini and Hochberg method [26].
197 As output, the rROMA software generates an R object including multiple results: (i) a module matrix
198 detailing *L1* and *Median Exp* values and their corresponding *p*-values and *q*-values for each pathway,
199 (ii) a sample matrix that presents the activity level of each pathway across samples derived from PC1
200 and (iii) a gene weight list for each gene set, which reports the gene projections in the PC1 space.
201 For a *shifted* gene set, the measurement of sample activity level is particularly important. It
202 determines which samples are responsible for the shift of the gene set. If the samples were already

203 separated into several conditions prior to the analysis, it is possible to verify that this separation into
204 groups is indeed responsible for the shift, by checking that the activity levels are significantly
205 different between the conditions. Conversely, if the conditions are not known *a priori*, it is possible
206 to determine new groups by performing a hierarchical clustering analysis on the pathways that are
207 shifted, and thus determine potential new groups of interest from the analysis.
208 The analyses mentioned above for the case of shifted gene sets are also valid for *over-dispersed*
209 gene sets. However, in the case of the latter, the analysis of gene weights, meaning the loadings
210 from the PC1, also becomes interesting: the genes associated with the highest weights are the
211 driving force in the activity scores of the gene sets. They summarize the information of the gene
212 sets, which can be particularly useful for assessing the importance of some genes. They can further
213 be used as an input for systems biology approaches.
214 The analysis of the sign of the gene weights is also particularly relevant. For example, gene sets
215 containing both activators and inhibitors can be highlighted by rROMA by being overdispersed, and
216 the associated genes highlighted. Such gene sets would not be detected as overdispersed by
217 methods based on the average expression of genes in the samples.
218 **Uncovering active pathways in Cystic Fibrosis with rROMA**
219 We applied rROMA to investigate the activity of pathways in airway epithelial cells from CF patients
220 and healthy donors. More precisely, we compared the transcriptomes of primary cultures of airway
221 epithelial cells from patients (N=6) with those of healthy controls (N=6), based on RNAseq data
222 publicly available in the NCBI's GEO database, under the accession ID GSE176121 [27]. rROMA was
223 run by specifying the pathway database to use and the expression matrix to analyze, as shown in the
224 accompanying vignette. Here, the Molecular Signature Database MsigDB hallmark gene set
225 collection [28] was used, a gene set collection of 50 gene sets specifically curated to represent core
226 biological processes and pathways that are commonly dysregulated in cancer. However, to provide a
227 more complete view of the biological processes involved in a study, rROMA can be used with
228 different reference databases.

229 The results of rROMA highlight pathways that are provided in the *ModuleMatrix* output. Pathways
230 with a *ppv Median Exp* lower than a given threshold were deemed as *shifted*, while those with a *ppv*
231 *L1* lower than this threshold were *overdispersed*. The *Plot.Genesets.Samples* function allows for the
232 visualization of activity scores of significantly shifted and overdispersed pathways across samples, in
233 the form of a heatmap representation (Figure 3). rROMA identified two shifted pathways,
234 APICAL_SURFACE which is found to have higher activity scores in healthy controls than in CF
235 patients, and FATTY_ACID_METABOLISM which has higher scores in CF patients than in controls, and
236 an overdispersed pathway, COAGULATION, with higher activity scores in healthy controls than in CF
237 patients.
238 When sample groups have been pre-defined, as CF or control groups in our case, these groups can
239 be compared based on the activity scores of the gene sets observed in the samples specific to the
240 two groups. Boxplots of the activity scores based on predefined groups can facilitate the
241 interpretation and complement a typical differential analysis. In our study, shifted and overdispersed
242 pathways behaved differently in CF patients versus healthy controls, as shown in Figure 3.
243 Alternatively, when the groups are not predefined, analyzing the shifted and overdispersed
244 pathways can reveal clusters of samples exhibiting similar pathway activity.
245 The analysis of top contributing genes in each pathway also provides crucial information. The
246 weights assigned to each gene in the PC1 vector allow to identify the key genes that most contribute
247 to variations in pathway activities as those with the higher weights. For each pathway, gene weights
248 are provided by the PlotGeneWeight function. For example, plotting the gene weights for the
249 COAGULATION pathway (see Figure 4) highlighted the GSN gene, encoding the protein *Gelsolin*, as
250 the highest contributor to the activity score. Notably, *Gelsolin* has been previously reported to play a
251 role for CFTR activation [29,30], and to promote mucus fluidification in CF [31].
252 Finally, many hyperparameters can be specified and changed to modify rROMA speed, precision, or
253 behavior regarding outlier detection. Details about all available hyperparameters are described in
254 the vignette accompanying the software. The computational time required to run the algorithm

255 typically depends on the number of considered pathways and their relative sizes. It also depends on
256 whether parallelization is enabled. In the present work, the algorithm ran in about 3 minutes and 15
257 seconds on a MacBook Pro equipped with a 2,6 GHz Intel Core i7 6 cores processor. A single analysis
258 60 pathways took roughly 5 second. Parallelization was not used, but this would have increased the
259 speed of the analysis.

260 **Comparison with state-of-art methods**

261 We compared rROMA with four sample-wise pathway quantification methods implemented in the
262 GSVA package [12] that cover different types of algorithms, namely GSVA, PLAGE, ssGSEA and z-
263 score. We conducted a comparison of these methods using the RNAseq data from the same
264 transcriptomic study (GSE176121) on airway epithelial cells. We selected the samples to simulate
265 two distinct analysis scenarios. In the first scenario, as previously described, we compared CF and
266 healthy control samples. Our goal was to explore a pathway database to identify pathways that
267 could differ between these two conditions without any a priori hypotheses of the specific pathways
268 involved. In the second scenario, we compared the gene expression data of six healthy control
269 samples before and after treatment with IL17+ TNFα. In this case, our aim was to assess whether the
270 different methods would effectively reveal the activation of the IL17 and TNFα pathways.
271 Although only rROMA provides a statistical significance of the pathway activity scores, for the other
272 methods significance with respect to a phenotype can be evaluated using conventional statistical
273 models. Therefore, for each method we compared the distributions of the pathway activity scores
274 between the two groups with a Student t-test and considered dysregulated pathways between the
275 two groups as those with adjusted p-values for the statistical test less than 0.05.
276 In the first scenario, the GSVA, ssGSEA, and z-score methods did not reveal any pathways as
277 dysregulated between CF and non-CF samples based on the statistical test. In contrast, PLAGE
278 identified dysregulation in 45 out of 50 pathways, i.e. a large number of pathways which prevents
279 meaningful interpretation with respect to the IL17+TNFα treatment. However, rROMA detected

280 eight pathways with significant activity differences between CF and healthy samples, shedding light
281 on potential biological mechanisms that could play a role in the disease.
282 In the second scenario, the TNFα signaling pathway was found significantly active in the samples
283 treated with IL17+TNFα compared to the control ones with all methods. We then compared the
284 ranking of this pathway among the significant results for the different methods. The results obtained
285 for the different methods are provided in Supplementary Tables 1-10. The TNFα pathway's ranking
286 among significantly dysregulated pathways based on the t-statistics was as follows: 5 out of 24 for
287 GSVA, 5 out of 27 for ssGSEA, 7 out of 17 for the z-score and 18 out of 45 for PLAGE. In contrast, the
288 rROMA algorithm highlights TNFα signaling pathway as the top significantly shifted pathway, thus
289 providing the most consistent result with the expected output. Among the rROMA results in this
290 scenario, we pointed out an example in which the P53 pathway exhibits significant overdispersion,
291 with target genes being affected in opposite directions by the transcription factor (Supplementary
292 Figure 1).
293 All results obtained for the two scenarios with the different methods are provided in Supplementary
294 Tables 1-10 and can be reproduced using the codes publicly available in the following github
295 repository: https://github.com/sysbio-curie/rRoma_comp.git
296 **rROMA estimates cell type abundances from bulk transcriptomic data**
297 Furthermore, we showcase how rROMA can be utilized to explore cell type abundance using bulk
298 transcriptomic data. Saint-Criq et al. [32] investigated the impact of two differentiation media on
299 primary cultures of CF and non-CF airway epithelial cells, as determined by transcriptomic data. They
300 built a gene signature for each cell type using the 50 most expressed markers derived from a single-
301 cell RNA sequencing (scRNAseq) dataset [33]. They observed a significant overexpression of genes
302 belonging to the signature of the secretory cell subtype in cultures grown in one of the media
303 (referred to as UNC), compared to the other medium (referred to as SC). Conversely, gene markers
304 of the ciliated subtype were overexpressed in primary cultures grown using the SC medium
305 compared to UNC medium.

306 We applied rROMA to the Saint-Criq RNA seq dataset to estimate cell type abundance in CF and non-
307 CF samples. More precisely, the Plasschaert signature [33] for each cell type was used and gene
308 reference gene sets, and the activity scores of these gene sets across the samples were represented
309 in the form of a heat map. Samples were found to be clustered according to the differentiation
310 medium in which they were grown, and our results confirmed the higher abundance of the ciliated
311 cell subtype in UNC medium and higher abundance of the secretory cell subtype in SC medium
312 (Figure 5A). We repeated the rROMA analysis using an alternative signature from Okuda and
313 colleagues [34]. In contrast with the Plasschaert signature, the Okuda signature includes the most
314 differentially expressed genes in each cell type, encompassing both overexpressed and
315 underexpressed genes. As illustrated in figure 5B, rROMA analysis using the Okuda signature
316 consistently revealed the same relative abundances of secretory and ciliated subtypes between UNC
317 and SC growing media, as observed with the Plasschaert signature. Thus, rROMA allows us to clearly
318 highlight differences in cell-type abundances, facilitating the use of gene signatures that contain
319 both upregulated and downregulated genes and thus potentially more accurate.

320 **Functional analysis of proteomic data with rROMA**

321 Finally, we applied rROMA to study a mass-spectrometry-based proteomic dataset of 9771 proteins
322 quantified in 122 treatment-naive primary breast cancers [35]. Individual samples were
323 characterized by their activity levels computed by rROMA for 50 pathways in the MsigDB hallmark
324 gene set collection [28]. We identified 34 shifted and 6 overdispersed pathways with p-value lower
325 than 0.05. The samples were clustered based on their activity profiles across the significant
326 pathways, using unsupervised hierarchical clustering and Euclidean distance.

327 Overall rROMA achieved satisfactory results with the proteomic dataset. Unsupervised clustering
328 divided breast cancer samples into six clusters (Figure 6), showing a higher heterogeneity of the
329 disease compared to the clinical stratification based on five intrinsic PAM50 subtypes [36] (Figure 6).
330 Cluster 1 predominantly contains basal tumors, whereas the Luminal B tumors are mainly associated
331 with cluster 5. These two clusters are characterized by high activation of TNFα signaling via NFK-β,

332 epithelial to mesenchymal transition and inflammatory pathways. These observations are consistent
333 with known facts about inflammatory and basal breast cancer [37,38]. Luminal A tumors are
334 segmented mainly in two clusters, namely cluster 4 and cluster 6, both showing high activation of
335 estrogen response, which is consistent with the fact that ER is the defining and driving factor in
336 luminal breast cancer.
337
338 **Discussion**
339 Quantifying the activity of biologically related gene sets is a commonly employed approach to
340 extract valuable biological insights from high-throughput data. The use of gene sets as aggregated
341 variables from molecular data enables the capture of biological information that may not be
342 detectable when solely focusing on individual genes. To address this challenge, we introduced the
343 rROMA algorithm, a PCA-based approach for quantifying pathway activity. Based on a gene
344 expression data matrix, this algorithm implements a linear model of gene regulation and efficiently
345 and reliably quantifies the activity of gene sets by computing the first principal component (PC1),
346 while also evaluating the statistical significance of this approximation.
347 Importantly, the algorithm provides the activity level of the gene set for each individual sample and
348 does not require a predefined labeled classification of the samples into various conditions or groups.
349 These activity levels can be subsequently compared, bringing to light the heterogeneity present in
350 the dataset in relation to the analyzed gene sets. This may be useful to define groups of samples or
351 patients when such stratification is unknown. Furthermore, the algorithm capability of identifying
352 *shifted* and *over-dispersed* gene sets is very peculiar compared to existing methods.
353 We applied rROMA to CF transcriptomic datasets, highlighting some biological mechanisms
354 potentially involved in the initiation or progression of the disease, and their associated genes. In our
355 study, out of the 50 hallmark pathways tested, 3 were significantly active:
356 FATTY_ACID_METABOLISM, APICAL_SURFACE, and COAGULATION. The FATTY_ACID_METABOLISM
357 pathway has significantly different activity scores between CF patients and healthy donors. This

358 pathway has been extensively studied in CF, and essential fatty acid deficiency is a well-known CF
359 phenotype (for a review, see [39]). The APICAL_SURFACE can be related to another well-known
360 hallmark of CF, i.e. a perturbation of airway surface secretory mucus content. Finally, the
361 COAGULATION pathway, the only overdispersed pathway in our study, seems to be highlighted due
362 to one specific gene with a very high associated weight, that is by far the most contributing gene to
363 the activity score of this pathway: gelsolin (GSN). Gelsolin has been reported as playing a role for
364 CFTR activation [29, 30], which suggests that the role of this gene in CF disease may be interesting to
365 study in more detail. The goal of this use case was not to undertake a detailed systems biology
366 approach of CF, which is beyond the scope of the present paper. In particular, it would require us to
367 include additional transcriptional dataset to take more samples into account and increase the
368 statistical power of our analyses, and to test several reference databases of gene sets. Overall, this
369 case study illustrates that rROMA can identify disease-associated pathway deregulations from
370 transcriptomic data, allowing a more comprehensive and functional interpretation of the data. It is
371 also a versatile tool that can shed light on various biological questions such as highlight the key
372 genes driving these deregulations, identify clusters of samples, study samples' cell-type composition,
373 or other cellular changes in a broader biological perspective. Indeed, the output of rROMA, in
374 particular the top weighted genes of significantly active pathways, can be interpreted as nodes
375 comprising genes and proteins of importance in the system under investigation, which can be used
376 as inputs to construct mathematical models.
377 Finally, we showcased the application of rROMA for functional analysis of proteomic data. In this
378 study, rROMA was applied to a proteomic dataset consisting of 9771 proteins from 122 treatment-
379 naive primary breast cancer samples. The results showed that rROMA performed well with the
380 proteomic dataset, revealing distinct clusters of breast cancer samples and higher disease
381 heterogeneity compared to clinical stratification based on the PAM50 subtypes. Furthermore, the
382 pathways identified as active by rROMA within these clusters are consistent with established
383 knowledge regarding signaling pathways relevant to the subtypes, such as high activation of TNFα

signaling via NFK-β, epithelial to mesenchymal transition and inflammatory pathways in the basal subtype and high activation of estrogen response in the luminal subtype.

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:

$$\text{Expression(gene } G\text{, sample } S) \approx \alpha^{F}_{G}\ \text{Activity}^{F}_{S} \qquad (1)$$

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.

The uni-factor 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.

409 As many other applications of PCA, rROMA uses SVD [25] to speed up the calculation. rROMA uses
410 the R irlba package [40, 41], which allows to focus on the first principal component, by computing
411 only the first column of V and the first value of S. However, if the data are not centered, the
412 simplification is no longer valid, and we must now consider the fact that the mean is not zero. The
413 rROMA algorithm therefore starts by centering the data of the global matrix.

414 **Pre-processing of data for rROMA analysis**

415 The input format for gene or protein expression for rROMA is a tab-delimited text file with columns
416 corresponding to biological samples, and rows corresponding to genes or proteins. The first row is
417 assumed to contain the sample identifiers while the first column is assumed to contain the non-
418 redundant gene or protein names.

419 If the data table contains missing values, they can be imputed using an approximation of the data
420 matrix with missing values by a full matrix of lower rank. To do this, the user must specify the rank of
421 the approximate full matrix that he wants to use. Then, the principal components are calculated up
422 to the specified rank, using an algorithm capable of working with missing data [42]. This PCA
423 decomposition is used to construct the full approximate matrix of lower rank, from which the
424 missing values of the original data are imputed. In the rest of the algorithm, the full imputed matrix
425 is used.

426 **Orientation of the PC1**

427 In standard PCA, all components are calculated with a sign ambiguity: there is mirror symmetry,
428 which makes it difficult to determine whether a given set of genes is over- or under-activated. In
429 rROMA, several methods exist to solve this ambiguity.

430 If knowledge exists about the role of a gene in a given gene set, we recommend using it by
431 associating a sign with the effect of the gene in the gene set: negative for an inhibitor, and positive
432 for an activator, for example. rROMA then uses the information about these signs to choose the
433 orientation that maximizes the number of genes associated with a positive sign whose projection in

434 PC1 is positive, and the number of genes associated with a negative sign whose projection in PC1 is
435 negative.
436 Although less efficient, other methods of orienting the first component exist in the case where there
437 is no a priori knowledge about the genes in the reference gene sets. The most efficient method
438 consists in considering only the genes associated with the most extreme weights according to the
439 first component (according to a percentage defined by a modifiable hyperparameter), and then
440 multiplying these weights by the expression level of the corresponding genes. If the result is
441 negative, then the orientation of the PC1 is reversed. The principal behind this method is to orient
442 the first component to maximize the positive correlation between the expression of genes and the
443 PC weight.
444 **Filtering of outlier samples**
445 Measurements may have been performed incorrectly in some samples and keeping them may lead
446 to erroneous results from rROMA. By default, in the algorithm, no sample filtering is performed, as
447 the matrix in input is assumed to contain only correct samples. However, outlier sample detection
448 can be activated by a hyperparameter, and a filtering procedure in two steps is then applied based
449 by (i) checking the samples for a similar number of detected genes and (ii) PCA computation and
450 projection of samples in the gene space for identifying data points (samples) that deviate
451 significantly from the data distribution. Here, the number of PCs used to perform the filtering and
452 the distance threshold for defining outliers are defined by two hyperparameters. It is also possible to
453 apply only one of the two filtering steps.
454 **Filtering of outlier genes**
455 The calculation of PC1 can be affected by the presence of an outlier gene in the dataset. This outlier
456 could indeed artificially affect the PC1. To increase the robustness of the PC1 calculation, we use in
457 rROMA the "leave-one-out" cross-validation approach [43], working as follows:
458

459 I. for each gene in the selected gene set, the method calculates the percentage of
460 variance explained by the first principal component (PC1) when that gene is
461 removed from the dataset. This value is denoted as *L1*, and each gene is associated
462 with its *L1* value;
463 II. the distribution of these *L1* values is normalized by centering and scaling to obtain z-
464 scores for each gene;
465 III. those genes whose associated z-scores exceed a specified threshold value are
466 identified as outliers and removed from the analysis.
467
468 The idea behind this method is to identify genes that have too much impact on the PC1 on their
469 own: if the percentage of variance explained by PC1 increases significantly in the absence of a single
470 gene, this means that this gene does not follow the alignment of all the others. It is then considered
471 as an outlier gene.
472 When a gene is considered an outlier for a given gene set, it is only removed from that gene set. This
473 default behavior can be modified by a hyperparameter, so that genes are completely removed from
474 all analyses, as soon as they are considered outliers in at least one gene set. However, in some cases,
475 these genes may still be of interest for analysis. Additional analysis steps for these genes are
476 therefore available in rROMA. In particular, the greater the number of gene sets analyzed containing
477 a given gene, the more likely it is that the gene will be considered as an outlier in at least one gene
478 set, and therefore be removed from the analysis. To avoid such abusive withdrawals, a Fisher test is
479 performed. This consists of comparing the average proportion of collections in which the genes in
480 the analysis are considered as outliers to this same proportion for a particular gene. If the proportion
481 of outliers is close to the average proportion for all the genes, then it is no longer considered an
482 outlier. Conversely, if the proportion of aberrations is significantly higher (threshold determined by a
483 modifiable hyperparameter), then it is still considered an outlier.

484 Instead, a gene present only in a small number of gene sets may be important for the understanding
485 of these gene sets. In such cases, it is not desirable to remove it, even if it is considered an outlier.
486 Thus, if a gene is present in less than a defined number of gene sets (set by a modifiable
487 hyperparameter), then it is not considered an outlier, regardless of the results of the "leave-one-
488 out" approach for it.
489 Finally, it is possible that a significant proportion of genes are considered outliers for a given gene
490 set. However, removing too many genes can totally distort the detectable activity for a gene set. A
491 final filter therefore exists to limit the maximum proportion of genes that can be considered as
492 aberrant for a given gene set and removed from the analysis (the proportion is set by a modifiable
493 hyperparameter). In this way, only genes with the most extreme leave-one-out behavior are
494 effectively considered outliers for this gene set.
495 The rROMA algorithm uses PCA calculations at different stages of its workflow, each serving a
496 specific purpose. These stages include imputing missing values, identifying outlier samples and
497 genes, and computing pathway activity. For missing value imputation, rROMA utilizes the iterative
498 PCA method implemented in the *mice* package [44] on the global expression matrix, imputing any
499 missing values within the dataset. To detect outlier samples, PCA is performed on the global
500 expression matrix with imputed missing values. Samples are treated as observations, and genes
501 serve as variables in this space, allowing for the identification of observable sample outliers. To
502 identify outlier genes and calculate pathway activity, PCA is carried out within a submatrix that
503 exclusively contains genes related to the specific pathway of interest. In this context, genes are
504 treated as observations, and samples serve as variables for PCA analysis.
505 **Optimization of the calculation of null distributions**
506 The p-value is computed as the probability of obtaining the observed activity measure for a specific
507 gene set by chance. The rROMA algorithm uses a random gene set procedure to generate a null
508 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 practice, it is often necessary to test many different gene sets available in large reference databases such as KEGG [45] or MSigDB [46]. Estimating the null distribution for each set of genes can lead to very time-consuming calculations. In the rROMA algorithm, there is an option to avoid calculating overdispersion and shift significance scores for all gene sets using the *ApproxSample* parameter. When this parameter is set to zero, the calculation is conducted for the null distribution without any approximation for all gene sets. Instead, when the *ApproxSample* parameter is different from zero, these scores can be approximated based on a predefined grid of values, which depends on the size of the gene set under consideration. 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. 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.

Data availability

The data utilized in this study has been downloaded from the NCBI GEO database (GEO accession ID GSE176121).

Code availability

The computational analyses were performed in R 4.2.2 and all the code is available under a GNU General Public License V3 in a GitHub repository at the following url: https://github.com/sysbio-curie/rROMA.

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 Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm (ITMO MIC project ModPhosphoNet) 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 number 20190502488 and by La Fondation Dassault Systèmes.

Competing Interests

None declared.

Author Contribution Statement

M.N. and M.C. contributed equally to this work. M.N., M.C., A.Z., L.A. and L.M.: conceptualization, methodology implementation 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., Next-generation genomics: an integrative approach, Nature Reviews Genetics, **11(7),** 476-486. (2010)

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

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

[4] Levine, D. M. et al., Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. Genome biology, **7(10)**, 1-17, (2006)

560 [5] Schreiber, A. W., & Baumann, U., A framework for gene expression analysis, Bioinformatics,
561 **23(2),** 191-197, (2007)
562 [6] Puthier D., and van Helden J., Statistics for Bioinformatics - Practicals - Gene enrichment
563 statistics, https://dputhier.github.io/ASG/practicals/go_statistics_td/go_statistics_td_2015.html ,
564 (2015)
565 [7] Subramanian, T. et al., Gene set enrichment analysis: A knowledge-based approach for
566 interpreting genome-wide expression profiles, Nat Genet, **34**, 267-273, (2003)
567 [8] Tomfohr, J., Lu, J., & Kepler, T. B., Pathway level analysis of gene expression using singular value
568 decomposition. BMC bioinformatics, **6,** 1-11, (2005)
569 [9] Bild, A. H., Potti, A., & Nevins, J. R., Linking oncogenic pathways with therapeutic opportunities.
570 Nature reviews cancer, **6(9)**, 735-741, (2006)
571 [10] Lim, S., Lee, S., Jung, I., Rhee, S., & Kim, S., Comprehensive and critical evaluation of
572 individualized pathway activity measurement tools on pan-cancer data. Briefings in bioinformatics,
573 **21(1)**, 36-46, (2020)
574 [11] Barbie DA, Tamayo P, et al., Systematic RNA interference reveals that oncogenic KRAS-driven
575 cancers require TBK1, Nature, **462**:108-112, (2009)
576 [12] Hänzelmann, S., Castelo, R., & Guinney, J., GSVA: gene set variation analysis for microarray and
577 RNA-seq data, BMC bioinformatics, **14**, 1-15, (2013)
578 [13] Vaske, C. J. et al., Inference of patient-specific pathway activities from multi-dimensional cancer
579 genomics data using PARADIGM, Bioinformatics, **26(12)**, i237-i245, (2010)
580 [14] Alvarez, M. J. et al. Network-based inference of protein activity helps functionalize the genetic
581 landscape of cancer, Nature genetics, **48(8)**, 838, (2016)
582 [15] Schubert, M. et al., Perturbation-response genes reveal signaling footprints in cancer gene
583 expression, Nature communications, **9(1)**, 20, (2018)
584 [16] Fan, J. et al., Characterizing transcriptional heterogeneity through pathway and gene set
585 overdispersion analysis, Nature methods, **13(3)**, 241-244, (2016)

586 [17] Landais, Y., & Vallot, C., Multi-modal quantification of pathway activity with MAYA, Nature
587 Communications, **14(1)**, 1668, (2023)
588 [18] Holland, C. H. et al., Robustness and applicability of transcription factor and pathway analysis
589 tools on single-cell RNA-seq data, Genome biology, **21**, 1-19, (2020)
590 [19] Zhang, Y. et al., Benchmarking algorithms for pathway activity transformation of single-cell RNA-
591 seq data, Computational and structural biotechnology journal, **18**, 2953-2961, (2020)
592 [20] Lee, E., Chuang, H. Y., Kim, J. W., Ideker, T., & Lee, D., Inferring pathway activity toward precise
593 disease classification, PLoS computational biology, **4(11)**, e1000217, (2008)
594 [21] Wagner, F., GO-PCA: an unsupervised method to explore gene expression data using prior
595 knowledge, PloS one, **10(11)**, e0143196, (2015)
596 [22] Frost, H. R., Li, Z., & Moore, J. H., Principal component gene set enrichment (PCGSE), BioData
597 mining, **8(1)**, 1-18, (2015)
598 [23] Drier, Y., Sheffer, M., & Domany, E. Pathway-based personalized analysis of cancer. Proceedings
599 of the National Academy of Sciences, **110(16)**, 6388-6393, (2013)
600 [24] Martignetti, L., Calzone, L., Bonnet, E., Barillot, E., & Zinovyev, A., ROMA: representation and
601 quantification of module activity from target expression data, Frontiers in genetics, **7**, 18, (2016)
602 [25] Golub, G. H., & Reinsch, C., Singular value decomposition and least squares solutions, Handbook
603 for Automatic Computation: Volume **II**: Linear Algebra, 134-151, Springer Berlin Heidelberg, (1971)
604 [26] Benjamini, Y., & Hochberg, Y. Controlling the false discovery rate: a practical and powerful
605 approach to multiple testing, Journal of the Royal statistical society: series B (Methodological), **57(1),**
606 289-300, (1995)
607 [27] Rehman, T. et al. Inflammatory cytokines TNF-α and IL-17 enhance the efficacy of cystic fibrosis
608 transmembrane conductance regulator modulators, The Journal of clinical investigation, **131**(16),
609 (2021)
610 [28] Liberzon, A. et al. The molecular signatures database hallmark gene set collection, Cell systems,
611 **1(6)**, 417-425, (2015)

612 [29] Cantiello, H. Role of actin filament organization in CFTR activation, Pflügers Archiv, **443**, S75-S80,
613 (2001)
614 [30] Vasconcellos, C. A. et al. Reduction in viscosity of cystic fibrosis sputum in vitro by gelsolin,
615 Science, **263(5149)**, 969-971, (1994)
616 [31] Bucki, R. et al., Enhancement of Pulmozyme activity in purulent sputum by combination with
617 poly-aspartic acid or gelsolin, Journal of Cystic Fibrosis, **14(5)**, 587-593, (2015)
618 [32] Saint-Criq, V. et al., Choice of differentiation media significantly impacts cell lineage and
619 response to CFTR modulators in fully differentiated primary cultures of cystic fibrosis human airway
620 epithelial cells, Cells, **9(9)**, 2137, (2020)
621 [33] Plasschaert, L. W. et al., A single-cell atlas of the airway epithelium reveals the CFTR-rich
622 pulmonary ionocyte, Nature, **560(7718)**, 377-381, (2018)
623 [34] Okuda, K. et al. Secretory cells dominate airway CFTR expression and function in human airway
624 superficial epithelia, American journal of respiratory and critical care medicine, **203(10)**, 1275-1289,
625 (2021)
626 [35] Krug, K., et al., Proteogenomic landscape of breast cancer tumorigenesis and targeted therapy,
627 Cell, **183(5)**, 1436-1456, (2020)
628 [36] Parker, J. S. et al., Supervised risk predictor of breast cancer based on intrinsic subtypes, Journal
629 of clinical oncology, **27(8)**, 1160, (2009)
630 [37] Fouad, T. M., Kogawa, T., Reuben, J. M., & Ueno, N. T., The role of inflammation in inflammatory
631 breast cancer, Inflammation and cancer, 53-73, (2014)
632 [38] Sarrió, D. et al., Epithelial-mesenchymal transition in breast cancer relates to the basal-like
633 phenotype, Cancer research, **68(4)**, 989-997, (2008)
634 [39] Strandvik, B. Fatty acid metabolism in cystic fibrosis, Prostaglandins, leukotrienes and essential
635 fatty acids, **83(3)**, 121-129, (2010)
636 [40] Baglama, J., IRLBA: Fast Partial Singular Value Decomposition Method, (2016)

[41] Tsuyuzaki, K., Sato, H., Sato, K., & Nikaido, I., Benchmarking principal component analysis for large-scale single-cell RNA-sequencing, Genome biology, **21(1)**, 1-17, (2020)

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

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

[44] Van Buuren, S., & Groothuis-Oudshoorn, K., Mice: Multivariate imputation by chained equations in R. Journal of statistical software, **45**, 1-67, (2011)

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

[46] Liberzon, A. et al Molecular signatures database (MSigDB) 3.0. Bioinformatics, **27(12)**, 1739-1740, (2011)

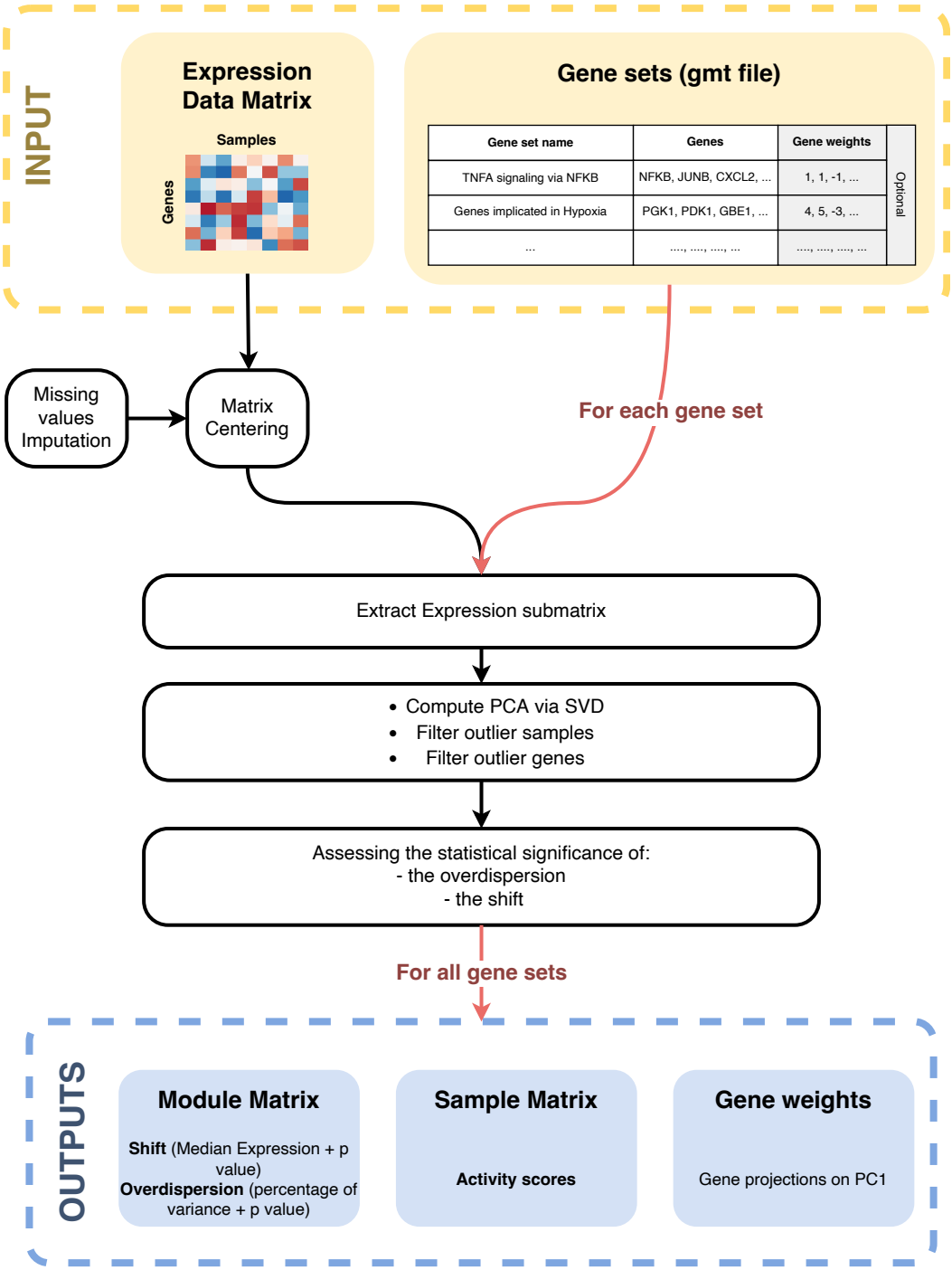
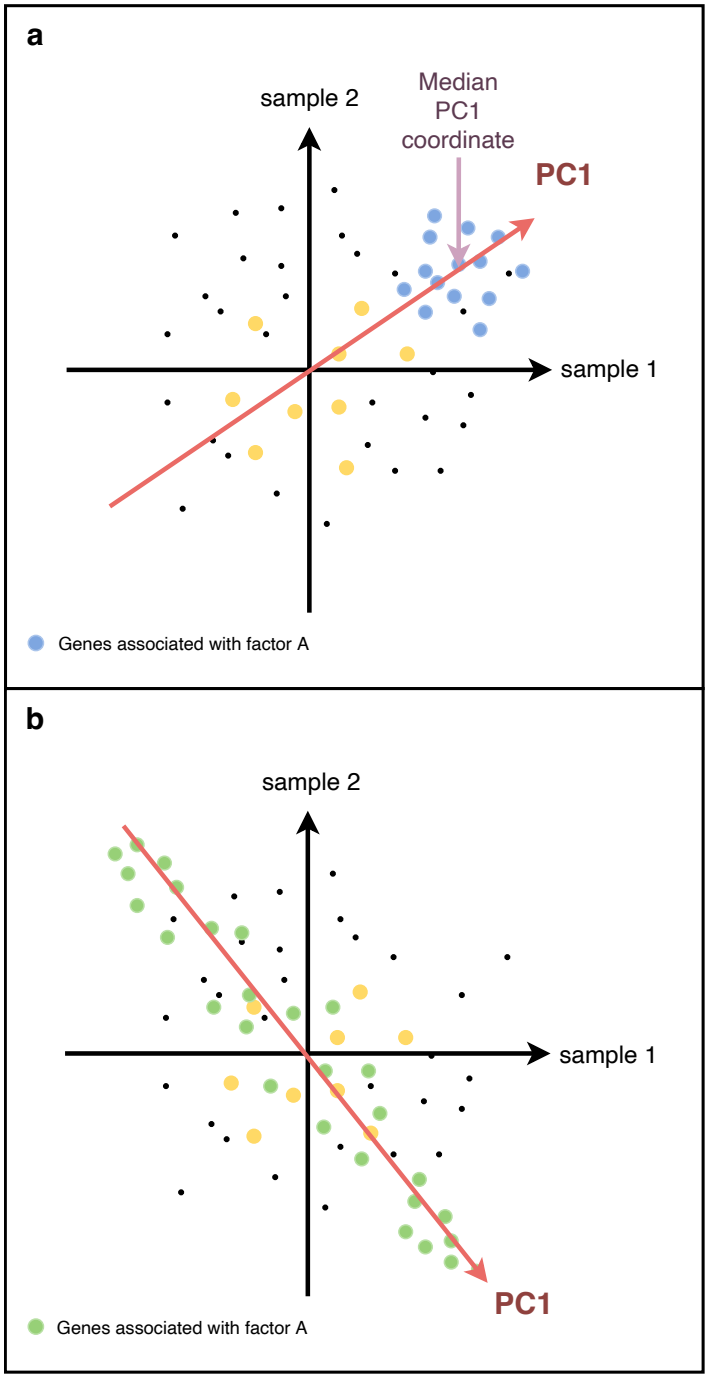


Figure 1. rROMA workflow. Schematic diagram illustrating the workflow of the rROMA algorithm.



651

652 **Figure 2. Representation of shifted and overdispersed gene sets.** Representation of gene sets in the

653 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.

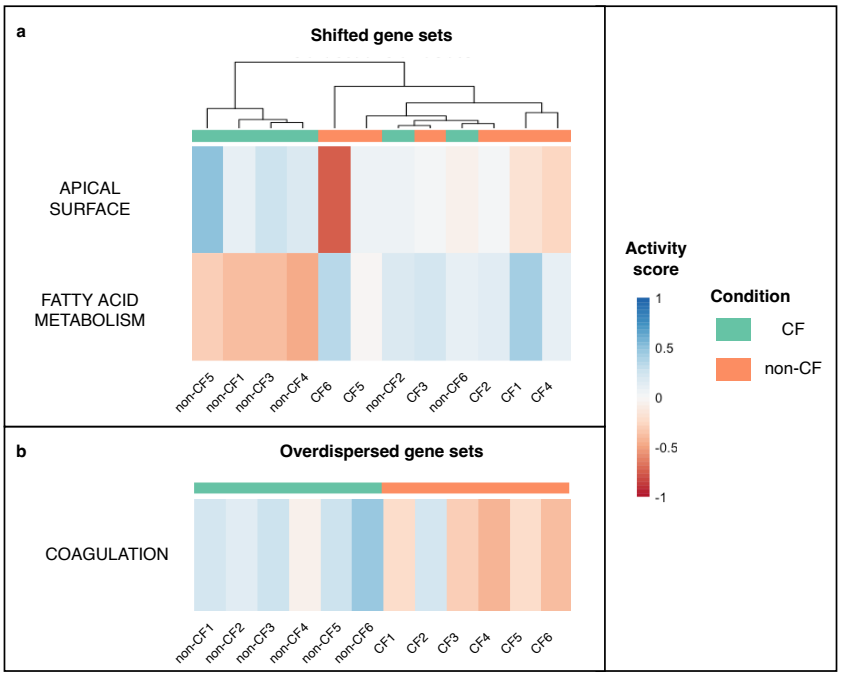


Figure 3. Active pathways in Cystic Fibrosis. Heatmap of activity scores for gene sets identified as significantly shifted (a) or significantly overdispersed (b) in GSE176121 dataset. Samples are in columns, gene sets are in rows. Horizontal sidebar color encodes true class labels.

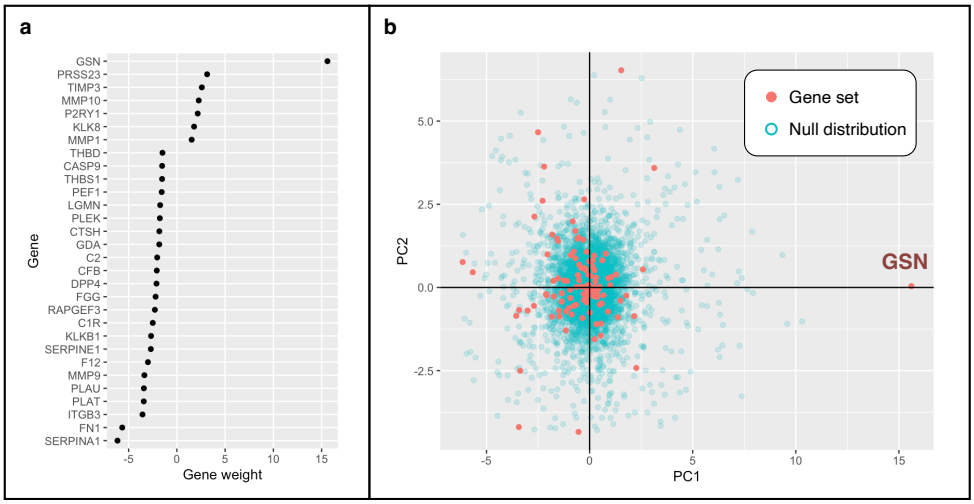
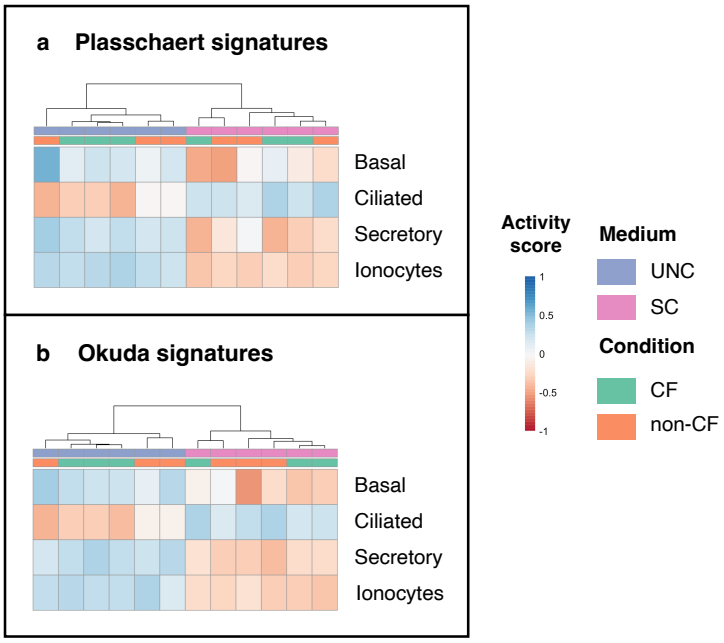


Figure 4. Gene weights for the COAGULATION pathway. Plots illustrating the contribution of genes to the COAGULATION gene set activity score. The weights in panel (a) indicate the gene projections on PC1, limited to the genes that have the greatest contribution to the observed variation in the COAGULATION 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.



677

678 **Figure 5. Estimated cell type abundances from CF transcriptomic data.** Heatmap of rROMA scores

679 obtained for Plasschaert (A) and Okuda (B) signatures of cell types in the Saint-Criq RNA seq dataset.

680 Samples are in columns, gene sets corresponding to cell types are in rows. Horizontal sidebar color

681 encodes true class labels.

682

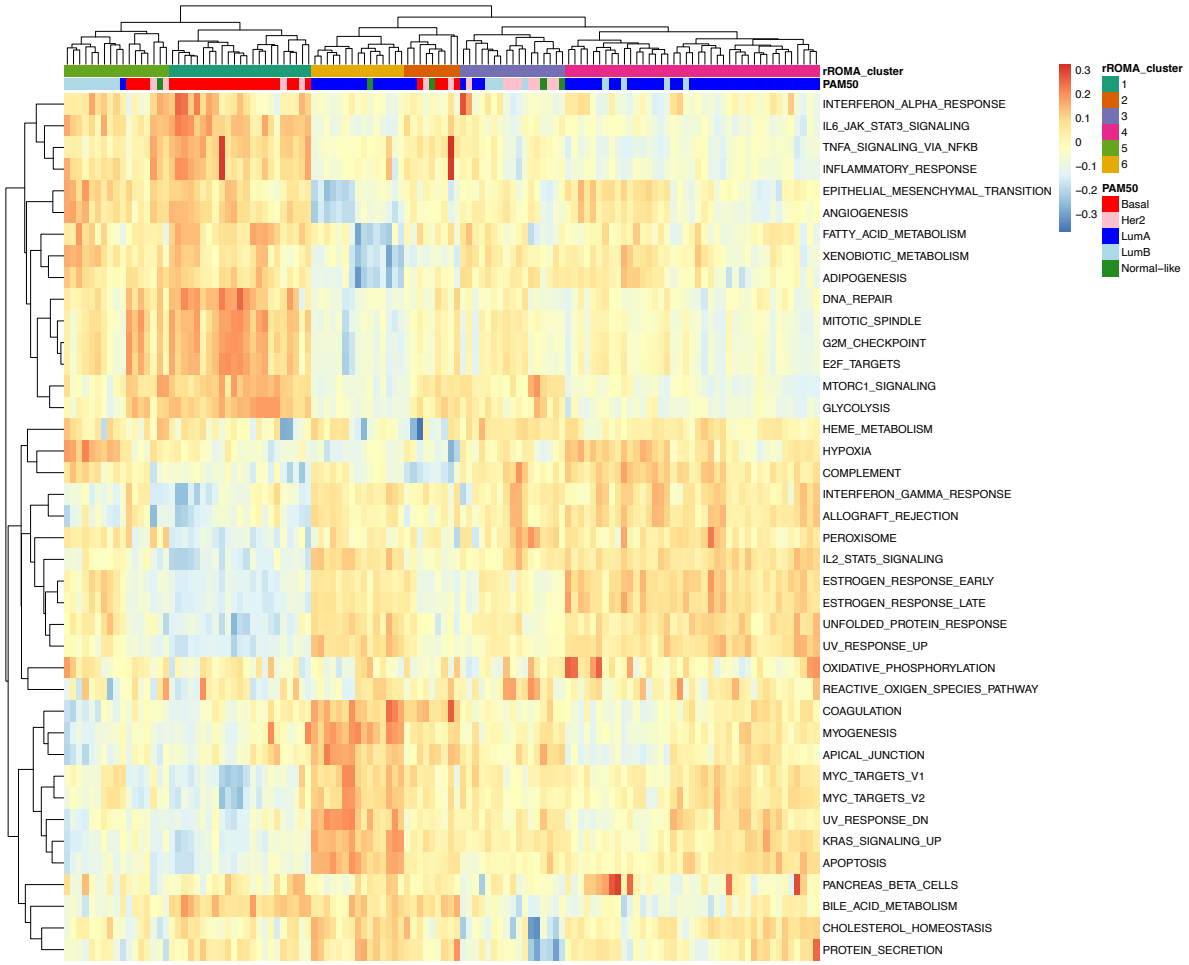


Figure 6. Functional analysis of proteomic data. Clustering of pathway activity level matrix for the rROMA significantly active pathways. The color bars indicate the cluster assignment based on rROMA pathway activity profiles (top) and the PAM50 subtype classification (bottom).