Representation and quantification Of Module Activity from omics data with rROMA

Matthieu Cornet[1,2,3,#], Matthieu Najm[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.

Abstract

In many analyses of high-throughput data in systems biology, calculating the activity of a set of genes rather than focusing on the differential expression of individual genes has proven to be efficient and informative. Here, we present the rROMA software package for fast and accurate computation of the activity of gene sets with coordinated expression. We applied rROMA to cystic fibrosis, highlighting biological mechanisms potentially involved in the establishment and progression of the disease and the associated genes. Source code and documentation are available at https://github.com/sysbio-curie/rROMA.

Contact: loredana.martignetti@curie.fr

1 Introduction

A signaling pathway can be defined as a set of genes or proteins that interact to pass on information from the exterior to the interior of a cell. In many diseases, some pathways are altered and the deregulations may come from one or several genes in this pathway, which may differ from one patient to another. Quantification of pathway activity via numerical scores using high throughput measurement of gene and protein expression is widely applied to transform the gene-level data into interpretable gene sets that can reveal biological heterogeneity across samples.

Here we present rROMA, a user-friendly and interactive R implementation of the ROMA algorithm that quantifies the activity of a gene set (or module) by its first weighted principal component (PC) (Martignetti et al, 2016). Here, gene sets or modules referred to canonical pathways annotated by domain experts or derived by external databases. In ROMA, the co-variance across samples of the genes composing the module is interpreted as the result of the action of a hidden factor on the expression of the genes. This setting corresponds to the simplest uni-factor linear model of gene expression regulation (Schreiber and Baumann, 2007).

The rROMA package provides novel functionalities for the analysis, visualization and reporting of active/inactive modules. It allows to highlight two possible configurations of the expression of the genes of a module under the effect of a regulatory factor: we called *shifted* module the case where the genes of the module are collectively displaced to one side with respect to the center of the distribution (Fig. 1A), while *overdispersed* module is the case where the amount of variance explained by the first PC of a given gene set is significantly higher than expected in the global gene expression distribution and the genes are found dispersed on both sides of the center of the global gene expression distribution (Fig. 1B).

Moreover, the algorithm implements different ways to identify outlier genes that can significantly influence the results. Finally, several functions for differential analysis and graphical visualization of the results are provided.

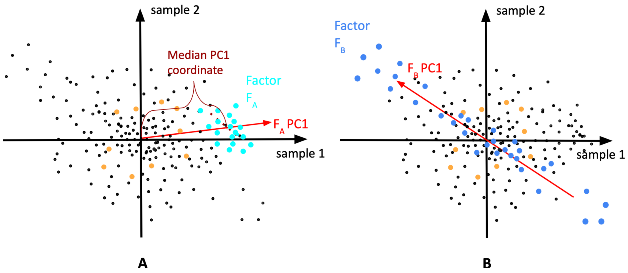

Fig. 1. Representation of genes in the case of two samples. Each dot represents one gene, its horizontal (resp. vertical) value corresponding to its expression in sample 1 (resp. sample 2). Genes associated with Factor A are plotted in light blue and the corresponding PC1 direction is plotted in red (A). This represents a shifted pathway, as assessed by a median of gene projections onto PC1 direction far from the origin of the distribution. Genes associated with Factor B are plotted in dark blue (B) and the corresponding PC1 direction is plotted in red. This represents an overdispersed pathway, as the PC1 is well aligned with the dots' distribution. Larger dots correspond to the genes with the highest score, highlighting their importance for the pathway. 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.

2 Implementation

rROMA is distributed as an open-source R software package and is available on GitHb: www.github.com/sysbio-curie/rROMA. A detailed vignette to reproduce all the analyses presented in this paper is also available.

3 Results

We applied rROMA to investigate the activity of pathways in airway epithelial cells from cystic fibrosis (CF) patients and from 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 GSE 176121 (Rehman et al, 2021).

Here, the Molecular Signature Database MsigDB hallmark gene set collection (Liberzon et al, 2015) was selected to ease interpretation. However, to provide a more complete view of the biological processes involved in a study, rRoma can be applied using multiple pathway databases. rRoma was run by specifying the pathway database to use and the expression matrix to analyze, as shown in the accompanying vignette.

Once rRoma has finished running the analysis, activated pathways can be determined by looking at the *ModuleMatrix* output. Pathways with a *ppv Median Exp* lower than a certain threshold were deemed as *shifted*, while those with a *ppv L1* lower than this threshold were *overdispersed*. Pathway activity across samples can be plotted with the *Plot.Genesets.Samples* function, and the top contributing genes in each pathway are determined by visualizing gene weights with the function *PlotGeneWeight*. Boxplot of the activity scores based on predefined groups can also be plotted for differential analysis. In this vignette, all highlighted pathways behaved significantly differently in CF patients versus healthy donors.

In our example, 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. In the context of CF, this pathway has been extensively studied, and its deregulation is a well known CF phenotype (for a review, see Strandvik, 2010). This illustrates the ability of rROMA to retrieve dysregulations from the transcriptomic data.

The *apical_surface* pathway is more difficult to interpret. This might actually highlight differences arising during cell differentiation, and it could thus be related to cell culture rather than to the disease. Finally, the *coagulation* pathway, the only overdispersed pathway, seems to be linked to one specific gene with a very high associated weight: *gelsolin* (*GSN*). We observe that *GSN* is by far the top contributing gene to the activity score of the *COAGULATION* pathway. *Gelsolin* has been reported as playing a role for CFTR activation (Cantiello 2001, Vasconcellos et al 1994). Overall, in this case study, rRoma highlighted a relevant mechanism in the context of CF, a potential bias due to cell culture, and an interesting gene which could be further investigated.

Many hyperparameters can be specified and changed to modify rRoma speed, precision, or behavior regarding outliers. Details about all available hyperparameters are described in the vignette. The computational time required to run the algorithm typically depends on the number of studied pathways and their relative size. It also depends on whether parallelization is enabled. Regarding the example discussed here, the algorithm ran in approximately 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. Note that parallelization was not used but would have increased the speed of the analysis if used.

In summary, this work indicates that rROMA is capable of identifying genetic pathways contributing to disease-associated transcriptomics enabling a clearer interpretation of results from a biological point of view, which allows interpreting cellular changes in a more holistic and functional way.

Acknowledgements

The authors would like to thank Pierre Gestraud for insightful discussions about the R implementation.

Funding

This work was supported by the European Commission's Horizon 2020 Program, H2020-SC1-DTH-2018-1, “iPC individualizedPaediatricCure” (ref. 826121) and by 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.

Conflict of Interest: none declared.

References

Cantiello H, (2001) Role of actin filament organization in CFTR activation. *Pflügers Archiv* **443**,S75–S80
Favia M. *et al*, (2019) An Intriguing Involvement of Mitochondria in Cystic Fibrosis. *J Clin Med.,* **6**;8(11):1890
Liberzon A. *et al*. (2015) The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst.,* **23**;1(6):417-425
Martignetti L. *et al*. (2016) ROMA: Representation and Quantification of Module Activity from Target Expression Data. *Front Genet*., **19**;7:18
Rehman T *et al*. (2021) Inflammatory cytokines TNF-α and IL-17 enhance the efficacy of cystic fibrosis transmembrane conductance regulator modulators. *J Clin Invest* **16**;131(16)
Schreiber A.W. and Baumann U. (2007) A framework for gene expression analysis. *Bioinformatics,* **15**;23(2):191-7
Strandvik B. (2010) Fatty acid metabolism in cystic fibrosis. P*rostaglandins Leukot Essent Fatty Acids* **83**(3):121-9
Vasconcellos C.A. *et al*, (1994) Reduction in viscosity of cystic fibrosis sputum in vitro by gelsolin. *Science*, **18**;263(5149):969-71