DECOMICS, a shiny application for unsupervised cell type deconvolution and biological interpretation of bulk omic data

Slim Karkar, Ashwini Sharma, Carl Herrmann, Yuna Blum, Magali Richard

DECOMICS, a shiny application for unsupervised cell type deconvolution and biological interpretation of bulk omic data

Slim Karkar[1,*], Ashwini Sharma[2], Carl Herrmann[2], Yuna Blum[3,*,⊕], Magali Richard[4,*,⊕]

[1]IBGC, UMR 5095, University of Bordeaux, CNRS, Bordeaux Bioinformatic Center, Bordeaux, France, 2 Health Data Science Unit, Medical

[2] Faculty Heidelberg and BioQuant, Heidelberg, Germany, 3 IGDR (Institut de Genetique et Developpement de Rennes), UMR 6290, ERL

[3] U1305, Equipe Labellisée Ligue Nationale contre le Cancer, Univ Rennes, CNRS, INSERM, 35000, Rennes, France

[4] TIMC, UMR 5525, Univ. Grenoble Alpes, CNRS, F-38700, Grenoble, France

[*]Corresponding authors : slim.karkar@u-bordeaux.fr ; magali.richard@univ-grenoble-alpes.fr ; yuna.blum@univ-rennes1.fr

[⊕]Equal co-last contributors

keywords: R shiny, deconvolution, unsupervised approaches, transcriptomic data, methylome, tissue heterogeneity

Abstract

Unsupervised deconvolution algorithms are often used to estimate cell composition from bulk tissue samples. However, applying cell type deconvolution and interpreting the results remains a challenge, even more without prior training in bioinformatics. We propose here a tool for estimating and identifying cell type composition from bulk transcriptomes or methylomes. DECOMICS is a shiny-web application dedicated to unsupervised deconvolution approaches of bulk omic data. It provides (i) a variety of existing algorithms to perform deconvolution on the gene expression or methylation-level matrix, (ii) an enrichment analysis module to aid biological interpretation of the deconvolved components, based on enrichment analysis, and (iii) some visualisation tools. Input data can be downloaded in csv format and pre-processed in the web application (normalisation, transformation and feature selection). The results of the deconvolution, enrichment and visualisation processes can be downloaded.

Availability and implementation: DECOMICS is an R-shiny web application that can be launched (i) directly from a local R session using the R package available here https://gitlab.in2p3.fr/Magali.Richard/decomics (either by installing it locally, or via a virtual machine and a Docker image that we provide); or (ii) in the Biosphere - IFB Clouds Federation for Life Science (French Institute of Bioinformatics), a multi-cloud environment scalable for high performance computing: https://biosphere.france-bioinformatique.fr/catalogue/appliance/193/, which requires creating an account and logging onto the platform

Introduction

Identification of the cell composition contributing to bulk molecular signals is a major challenge in molecular analysis in various applications such as cancer[1]. The development of in silico deconvolution methods has made it possible to revisit existing bulk omic data from large patient cohorts with regard to intra-sample heterogeneity, and thus to compare sample cell composition with available clinical annotations such as treatment response. Both supervised[2] and unsupervised[3–7] deconvolution methods have been proposed in the literature. Supervised methods estimate the component proportions using known cell-type reference matrices, whereas unsupervised methods estimate both the reference profiles and the component proportions, without prior knowledge except for the number of components to be considered. Supervised approaches are therefore limited by the quality of the reference signatures, while unsupervised approaches present difficulties in interpreting the inferred components and estimating the number of components to be

considered.

An intriguing advantage of unsupervised methods is that, unlike supervised methods, they can identify new cell populations or populations that would not have been taken into account a priori. Unsupervised deconvolution approaches have recently been used to identify radiogenomic signatures to predict prognosis of colorectal cancer[8] , to identify cellular compartment in unknown tumoral samples[9], or to infer clinical outcomes in melanoma patients[10] .

Applying unsupervised deconvolution approaches to patient cohorts is now possible thanks to recent advances in high-throughput sequencing that have generated an enormous amount of transcriptomic data as well as numerous methylome data. This is particularly true in the field of oncology, where more and more biological samples are being sequenced to help with patient stratification and prognosis. However, analyzing this type of data requires professional coding skills, which are rarely available to clinicians.

In order to appeal large data analysis to a wider audience, user-friendly alternatives have been devised, including the development of the R-shiny package, which enables interactive web applications to be built using the R statistical and data mining software[11]. If web applications exist to apply supervised algorithms on clinical datasets[12], exploration of unsupervised algorithms have been so far limited to bioinformaticians with computing skills and do not provide guidance for biological interpretation of their outputs.

Here we propose DECOMICS, a user-friendly shiny interactive web application designed to perform unsupervised deconvolution on transcriptomic and methylome (DNAm) data. Six

different unsupervised methods are implemented in DECOMICS, including the most commonly used (ICA and NMF), and more recent algorithms (CDSeq[4], debCAM[3] and PREDE[5], EDec[7]). Our tool also provides guidance during the process and helps with the biological interpretation of the results, which should be of significant interest for both bioinformaticians and clinicians.

Software description

DECOMICS is a Shiny application available as an R package that can be built from source (GitLab access) or used online through the biosphere cloud of the IFB (Institut Français de Bioinformatique). DECOMICS workflow is illustrated in Figure 1. It includes a Guide section, that can be used as a material and methods, and two main modules: the deconvolution module and a biological interpretation module (Figure 1A-B).

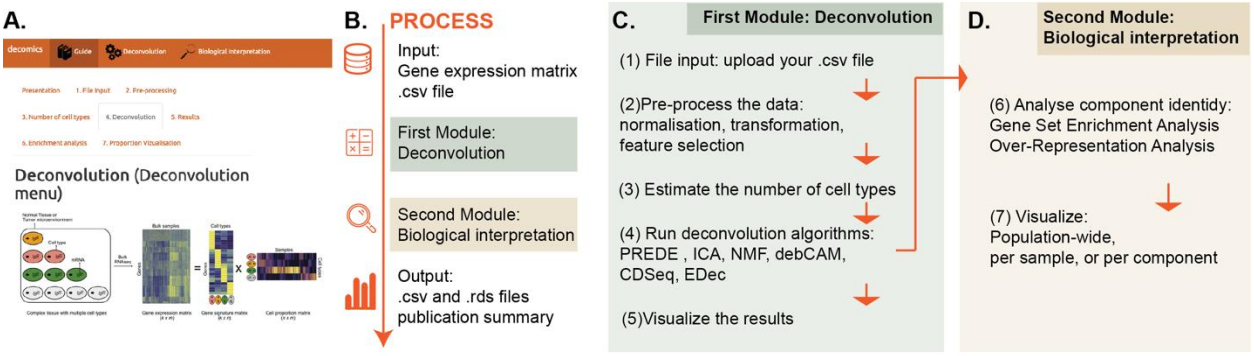


Figure 1: An overview of DECOMICS workflow to perform deconvolution of omics data. A. Screenshot of the application. B. Summary of the DECOMICS process. C. Workflow of the module ‘Deconvolution’. D. Workflow of the module ‘Biological interpretation’.

Deconvolution module

The deconvolution module (Figure 1C) is used to load the data, carry out the pre-processing, estimate the number of components, run the deconvolution and visualize the results. **File input (1)** requires a .csv file containing the omic data (either gene expression or DNAm) with samples in columns and features in row. Gene expression can be in the form of raw counts, processed counts or processed gene expression in the case of microarray-based technologies. We offer basic pre-processing features for gene expression data in step 2. DNAm data should be provided in the form of □-values. Basic **Pre-Processing (2)** of the gene expression data can be achieved within the application, including normalization adapted to RNA-seq data (Read-per-million or DESeq2) and transformation (log2 or pseudoLog). Pre-processing also includes the option to select a subset of features. Specifically, one can choose the top 1,000 or 5,000 gene expressions, or the top 10,000 or 20,000 □-values, based on the highest coefficients of variation. The **Number of components (3)** to infer (corresponding to the deconvoluted components) have to be estimated by the user. In general, the optimal number of components can be identified through various methods, including Cattell's rule applied to Principal Component Analysis (PCA) eigenvalues[13], the Minimum Description Length (MDL)[3], bootstrapping techniques[14], and cross-validation methods[21]. However, in a prior benchmark analysis, we found that different methods yielded comparable results[5]. For the sake of clarity, we have chosen to present a single method in the application: a guidance plot based on PCA eigenvalues. Then **Deconvolution (4)** is run by one of the six unsupervised algorithms provided in the application. Depending on the type of omic data

provided, a subset of algorithms is available: (i) Independent Component Analysis, Non-negative Matrix Factorization, CDSeq, debCAM and PREDE for gene expression data, and (ii) ICA, NMF, debCAM and EDec for DNAm data. Finally, deconvolution **Results (5)** can be visualized by (i) a 'cell-type' signature heatmap displaying the 5 top markers of each component, and (ii) a 'cell-type' proportion heatmap of each component.

Biological interpretation module

The biological interpretation module (Figure 1D) performs an enrichment analysis on the components estimated by the chosen deconvolution algorithm and displays the deconvoluted proportion matrix. First, we propose an **Enrichment analysis (6)** step, to help with the biological interpretation of the components identified by unsupervised approaches. It proposes to perform Enrichment analysis (Gene Set Enrichment analysis[16] or Over-representation analysis[17]) using various biological databases: 'CellMatch'[18], a reference database derived from various resources and other reference ones (GO, GTEx, KEGG, Reactome, Tissue Cell Types, Cancer Cell Types and Cancer Cell lines).

Second, a **Proportion Visualisation (7)** section offers the possibility to visualize the full component distribution for a single sample, or the distribution of a single component throughout the total cohort.

Material and methods

Deconvolution Algorithms

Unsupervised deconvolution problem applied to omic bulk data consists in solving an equation of form $X = T * A$ where T and A are jointly inferred from X. This is achieved by estimating the mixture of K cell-types, present in different proportion in each sample (cell-type proportion matrix A). Therefore, X can be described as a combination of cell-type specific molecular profiles (cell-type specific gene expression matrix T). The specifics of each existing unsupervised deconvolution algorithm and the reasons for choosing to include them or not in the DECOMICS application are presented in Supplementary Table 1. In DECOMICS, we provide six different algorithms to run deconvolution:

- ICA: Independent Component Analysis is a blind source separation algorithm that decomposes signal into statistically independent components. In DECOMICS ICA deconvolution is run using fastICA ($R\ CRAN$) and the Deconica[19] R packages. By default, 30 significant gene markers are selected to get component scores, using the "weighted.mean" summary metric.

- NMF[20]: In non-negative matrix factorization approach, the molecular profile matrix X is factorized into two matrices A and T, with the property that all three matrices have no negative elements. DECOMICS uses the $R\ CRAN$ NMF package with method ="snmf/r". The estimated A matrix is constrained to sum the proportion to 1, and T is computed as

$T = A^{-1}X$ using the *ginv* inverse function from MASS R package. Finally, all negative values for T are set to 0.

- CDSeq[4] aims at simultaneously estimating A and T matrices using a probabilistic model based on latent Dirichlet allocation (LDA). DECOMICS uses the R implementation of the CDseq method CDSeqR with the following parameters: beta = 0.5, alpha = 5, mcmc_iterations = 300. Reduction factor is computed to avoid expression values $> 10^5$, block numbers and gene block size are computed $s.t.$, a block do not exceed 10^3 genes.

- debCAM[3] stands for deconvolution by Convex Analysis of Mixtures. This method uses a geometric approach to identify a solution to the NMF problem in the simplex space. Thus, the proposed solution for A is always a proportion matrix. In DECOMICS, the function CAM is called from the debCAM R packages using the following empirical parameters: cluster.num is computed to be 5 times greater than the number of expected components, and dim.rdc set to divide the number of input genes by a tenth.

- PREDE[5] is a method that offers the possibility to conduct partial reference-based deconvolution method solved via an iterative Quadratic Programming procedure. In DECOMICS, PREDE function is used from PREDE R package, with the following parameters: W1 = NULL (which corresponds to a complete deconvolution approach), type = "GE", iters = 100 and rssDiffStrop = 1e-5.

- EDec-step1[7] estimates both average component methylation profiles and component proportions using an iterative constrained matrix factorization algorithm. This algorithm identifies cell type-specific methylation profiles and constituent cell type proportions by

minimizing the Euclidean distance between the reconstituted and original mixed methylation matrices. In DECOMICS, we use the EDec::run_edec_stage_1 function with the parameters max_its = 2000 and rss_diff_stop = 1e-10.

Gene Set Enrichment Analysis

In order to biologically characterize each of the unsupervised components identified, biological enrichment analyses are performed. For each component, the first step consists in ranking the genes according to their coordinates on the component, in order to identify the most contributing genes of the component. For methylation data, CpG coordinates are aggregated at the gene level, taking the maximum value observed for CpGs of the same gene. This approach is a way of considering a gene as strongly contributing to the component if it has at least one strongly contributing CpG. In a second step, an enrichment analysis is performed either based on Gene Set Enrichment Analysis (GSEA)[16] or Over-Representation Analysis (ORA)[17]. If the coordinates contain sufficient non-duplicate values (threshold set at 30% by default) to enable reliable ordering of the values, a GSEA analysis is performed using the fgsea R package[21]; otherwise, an ORA analysis is performed, taking as gene selection the top 20% of the component's most contributing genes and as gene universe all the genes available in the user’s dataset.

Various biological databases can be queried. DECOMICS includes the CellMatch database restricted to the human species and cell types with at least 3 marker genes per cell type. After filtering, it provides marker genes for 120 different cell types across 103 normal

tissues and 26 tumoral tissues. DECOMICS also includes the latest versions of the following biological databases provided on the Enrichr [22] tool website:

- *Cancer Cell Line Encyclopedia* (967 terms)
- *CellMarker Augmented 2021* (1097 terms)
- *GO Biological Process 2023* (5407 terms)
- *GO Cellular Component 2023* (474 terms)
- *GO Molecular Function 2023* (1147 terms)
- *GTEx Tissues V8 2023* (511 terms)
- *KEGG 2021 Human* (320 terms)
- *MSigDB Oncogenic Signatures* (189 terms)
- *NCI 60 Cancer Cell Lines* (93 terms)
- *Reactome 2022* (1818 terms)

Previously to the enrichment analysis, components obtained from ICA-based method are subjected to a reorientation, as proposed in the deconica R package based on the hypothesis that the highest absolute values of a component weight should be positive

Availability

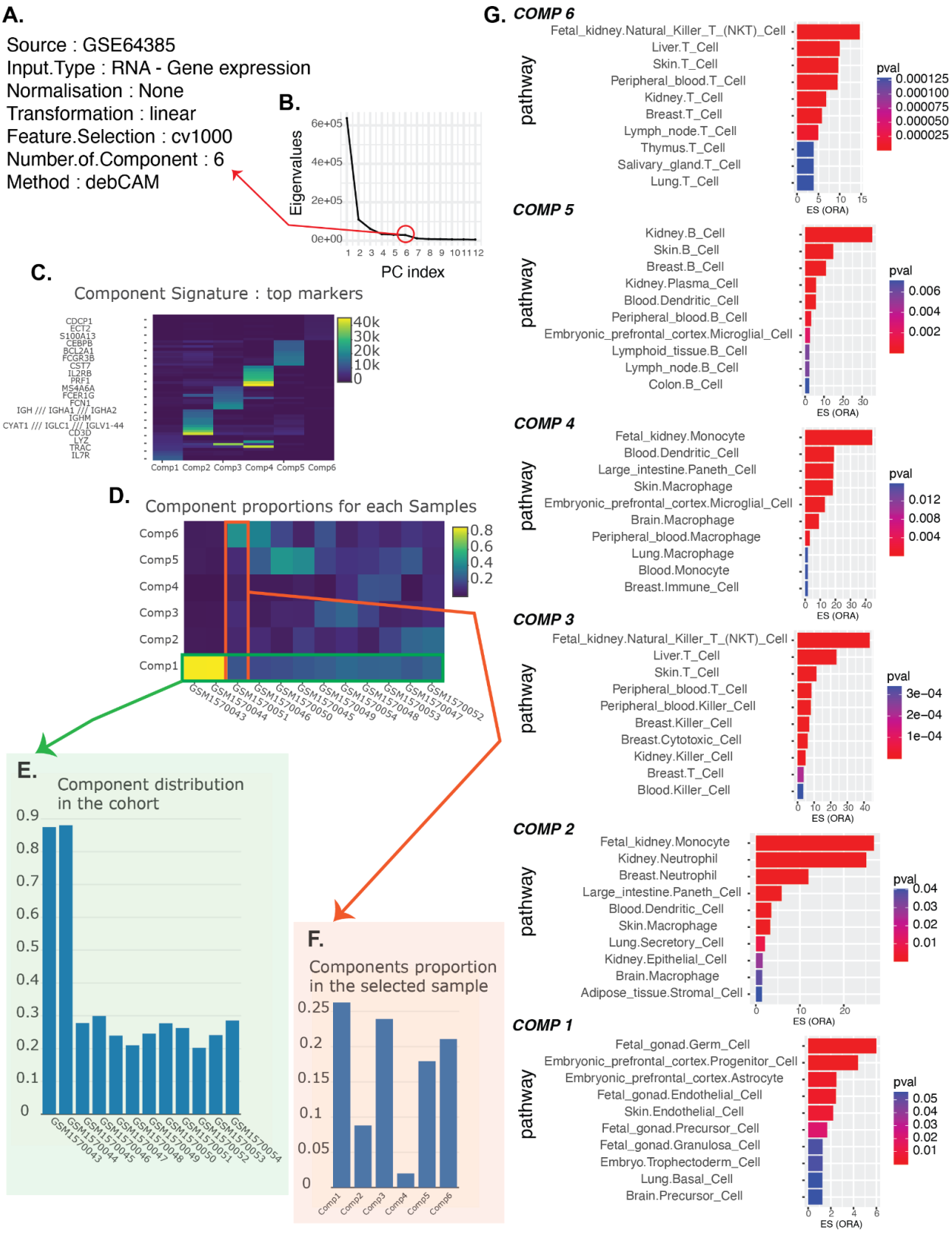
Installation instructions can be found on DECOMICS gitlab webpage: https://gitlab.in2p3.fr/Magali.Richard/decomics. There are three installation options: (i) full local installation, (ii) running locally a virtual machine (VM), or (iii) using the Biosphere-IFB cloud. Local installation requires several packages to be loaded. To help users, we propose a conda recipe on the DECOMICS gitlab webpage. To offer the possibility to run DECOMICS on a local VM, we built a docker container. The user simply needs to install docker on his machine and to launch the image we provide. Finally, DECOMICS is deployed on the Biosphere portal (searchable through the RAINBio catalogue). To use the clouds of IFB-Biosphere, users need to create an account and get membership of an active group (more information can be found here : https://ifb-elixirfr.github.io/biosphere/signin). Then users can deploy and connect to VM using the web interface (tutorial here: https://ifb-elixirfr.github.io/biosphere/vm_connect).

Application and results

We have provided two use cases to illustrate the DECOMICS pipeline. The first use case is based on gene expression data (GSE64385[23]) dataset is available for download in .csv format from the DECOMICS application. This dataset consists of a mixture of six cell types: HTC116, neutrophils, natural killer cells, monocytes, B cells, and T cells. As shown in Figure 2, the unsupervised components identified by DECOMICS are highly consistent with the constituent cell types. In this example, the deconvolution algorithm employed is

debCAM and the functional enrichment analyses was performed using the CellMatch database. The second use case utilizes DNAm profiles from reconstituted mixtures of six purified immune cells derived from human blood samples (GSE77797[24]). This example further showcases the efficacy of the DECOMICS pipeline for DNAm unsupervised deconvolution and biological interpretation of the unsupervised components. Detailed information on this use case is provided in Supplementary Figure 1, and the corresponding .csv dataset is available for download from the DECOMICS application.

In the DECOMICS pipeline, the input data consist of simple count tables in .csv format, which can be uploaded directly to the application. Example input files demonstrating the input format are available for download from the application. The pre-processed data (.csv), deconvolution results (.rds), top 100 contributing genes of each component (.csv), estimated typical gene expressions for components (.csv), and proportion estimates (.csv) can also be downloaded from the application. Additionally, enrichment analyses can be downloaded in the form of a .csv table containing the enrichment scores and P-values for the queried database.


A.
Source : GSE64385
Input.Type : RNA - Gene expression
Normalisation : None
Transformation : linear
Feature.Selection : cv1000
Number.of.Component : 6
Method : debCAM
B.
Eigenvalues
PC index
C.
Component Signature : top markers
D.
Component proportions for each Samples
E.
Component distribution in the cohort
F.
Components proportion in the selected sample
G.
COMP 6
COMP 5
COMP 4
COMP 3
COMP 2
COMP 1
pathway
pval
ES (ORA)

242

243

Figure 2: Illustration of DECOMICS based deconvolution of gene expression data. **A.** Parameters used in the DECOMICS application are described. **B.** Scree plot illustrating the selection process for the number of components to be deconvolved (module 1, step 3: Number of cell types). **C-D.** Deconvolution results are presented: (**C**) Component signatures are plotted, and (**D**) a heatmap showing component proportions (module 1, step 5: Results). **E-F**. Visualization of component distribution: (**E**) Distribution of a specific component across the cohort, and (**F**) distribution of all components within a given sample (module 2, step 7: Proportion visualization). **G**. Enrichment plot displaying the enrichment analysis for each component (module 2, step 6: Enrichment analysis) ORA indicates if an Overrepresentation Analysis has been performed, GSEA indicates if a Gene Set Enrichment Analysis has been performed for a given sample. Pval corresponds to the adjusted p-values of the Enrichment Score (ES) after correction for multiple testing.

Discussion

Despite their advantages over supervised methods, the use of unsupervised deconvolution methods is not trivial, as it requires a priori knowledge of the number of cell populations to be considered, as well as a biological interpretation of the estimated components. Here we provide a user-friendly interactive web-application to perform unsupervised deconvolution by assisting the user in the choice of the number of components and biological interpretation of the results.

Significant efforts have been made by colleagues to integrate unsupervised deconvolution into analysis pipelines, including determining the number of components to infer and interpreting the biological data[25,26]. However, executing these pipelines in their entirety necessitates the installation of R packages and the use of command line interfaces, which our web application avoids. Additionally, these pipelines offer a limited selection of deconvolution methods, restricted to those developed by the authors of the integrative pipeline. In this work, we propose an unbiased approach that incorporates several deconvolution methods.

DECOMICS stands out for its ease of use, speed, and comprehensive documentation. It is designed to be accessible to users without expertise in R or biostatistics. With its integrated biological interpretation feature, users can seamlessly perform both deconvolution and interpretation within the same platform, eliminating the need for external tools. Moreover, DECOMICS is highly adaptable, allowing for easy updates and redeployment as new reference-free methods or enrichment analysis databases become available. This flexibility ensures that DECOMICS remains at the cutting edge, capable of incorporating the latest technical and methodological advancements.

Competing interests

No competing interest is declared.

Authors contributions and statements

S.K., Y.B and M.R. conceived the project. S.K., A.S., C.H., Y.B and M.R. contributed to code development and results analysis. S.K., Y.B. and M.R. wrote the manuscript.

Acknowledgements

This work is a contribution of the EIT Health program COMETH. It has been partially supported by MIAI @ Grenoble Alpes (ANR-19-P3IA-0003), and by the French Agency for National Research (CauseHet // ANR-22-CE45-0030). Finally, it has also been carried out with financial support from ITMO Cancer of Aviesan within the framework of the 2021-2030 Cancer Control Strategy, on funds administered by Inserm (ACACIA project AAP-MIC-2021). We thank E. Amblard, H. Barbot, F. Pittion and L. Lamothe for testing the application.

Bibliography

1. Nguyen, H., Nguyen, H., Tran, D., Draghici, S. & Nguyen, T. Fourteen years of cellular deconvolution: methodology, applications, technical evaluation and outstanding challenges. *Nucleic Acids Res.* **52**, 4761–4783 (2024).
2. Avila Cobos, F., Alquicira-Hernandez, J., Powell, J. E., Mestdagh, P. & De Preter, K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat.*

Commun. **11**, 5650 (2020).

3. Chen, L. *et al*. debCAM: a bioconductor R package for fully unsupervised deconvolution of complex tissues. *Bioinforma. Oxf. Engl.* **36**, 3927–3929 (2020).
4. Kang, K., Huang, C., Li, Y., Umbach, D. M. & Li, L. CDSeqR: fast complete deconvolution for gene expression data from bulk tissues. *BMC Bioinformatics* **22**, 1–12 (2021).
5. Qin, Y. *et al*. Deconvolution of heterogeneous tumor samples using partial reference signals. *PLoS Comput. Biol.* **16**, e1008452 (2020).
6. Sompairac, N. *et al*. Independent Component Analysis for Unraveling the Complexity of Cancer Omics Datasets. *Int. J. Mol. Sci.* **20**, 4414 (2019).
7. Onuchic, V. *et al*. Epigenomic Deconvolution of Breast Tumors Reveals Metabolic Coupling between Constituent Cell Types. *Cell Rep.* **17**, 2075–2086 (2016).
8. Zhong, M.-E. *et al*. CT-based radiogenomic analysis dissects intratumor heterogeneity and predicts prognosis of colorectal cancer: a multi-institutional retrospective study. *J. Transl. Med.* **20**, 1–12 (2022).
9. Peng, X. L., Moffitt, R. A., Torphy, R. J., Volmar, K. E. & Yeh, J. J. De novo compartment deconvolution and weight estimation of tumor samples using DECODER. *Nat. Commun.* **10**, 4729 (2019).
10. Nazarov, P. V. *et al*. Deconvolution of transcriptomes and miRNomes by independent component analysis provides insights into biological processes and clinical outcomes of melanoma patients. *BMC Med. Genomics* **12**, 1–17 (2019).
11. Jia, L. *et al*. Development of interactive biological web applications with R/Shiny. *Brief.*

Bioinform. **23**, bbab415 (2022).

12. Li, T. *et al.* TIMER2.0 for analysis of tumor-infiltrating immune cells. *Nucleic Acids Res.* **48**, W509–W514 (2020).
13. Cattell, R. B. The Scree Test For The Number Of Factors. *Multivar. Behav. Res.* **1**, 245–276 (1966).
14. Houseman, E. A. *et al.* Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics* **17**, 259 (2016).
15. Lutsik, P. *et al.* MeDeCom: discovery and quantification of latent components of heterogeneous methylomes. *Genome Biol.* **18**, 1–20 (2017).
16. Mootha, V. K. *et al.* PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
17. Goeman, J. J. & Bühlmann, P. Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* **23**, 980–987 (2007).
18. Shao, X. *et al.* scCATCH: Automatic Annotation on Cell Types of Clusters from Single-Cell RNA Sequencing Data. *iScience* **23**, 100882 (2020).
19. Czerwinska, U. UrszulaCzerwinska/DeconICA: DeconICA first release. Zenodo https://doi.org/10.5281/zenodo.1250070 (2018).
20. Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC Bioinformatics* **11**, 367 (2010).
21. Korotkevich, G. *et al.* Fast gene set enrichment analysis. 060012 Preprint at https://doi.org/10.1101/060012 (2021).
22. Chen, E. Y. *et al.* Enrichr: interactive and collaborative HTML5 gene list enrichment

analysis tool. *BMC Bioinformatics* **14**, 128 (2013).

23. Becht, E. *et al*. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. *Genome Biol*. **17**, 1–20 (2016).
24. Koestler, D. C. *et al*. Improving cell mixture deconvolution by identifying optimal DNA methylation libraries (IDOL). *BMC Bioinformatics* **17**, 1–21 (2016).
25. Li, Z. & Wu, H. TOAST: Improving reference-free cell composition estimation by cross-cell type differential analysis. *Genome Biol*. **20**, (2019).
26. Scherer, M. *et al*. Reference-free deconvolution, visualization and interpretation of complex DNA methylation data using DecompPipeline, MeDeCom and FactorViz. *Nat. Protoc*. **15**, 3240–3263 (2020).

Supplementary material

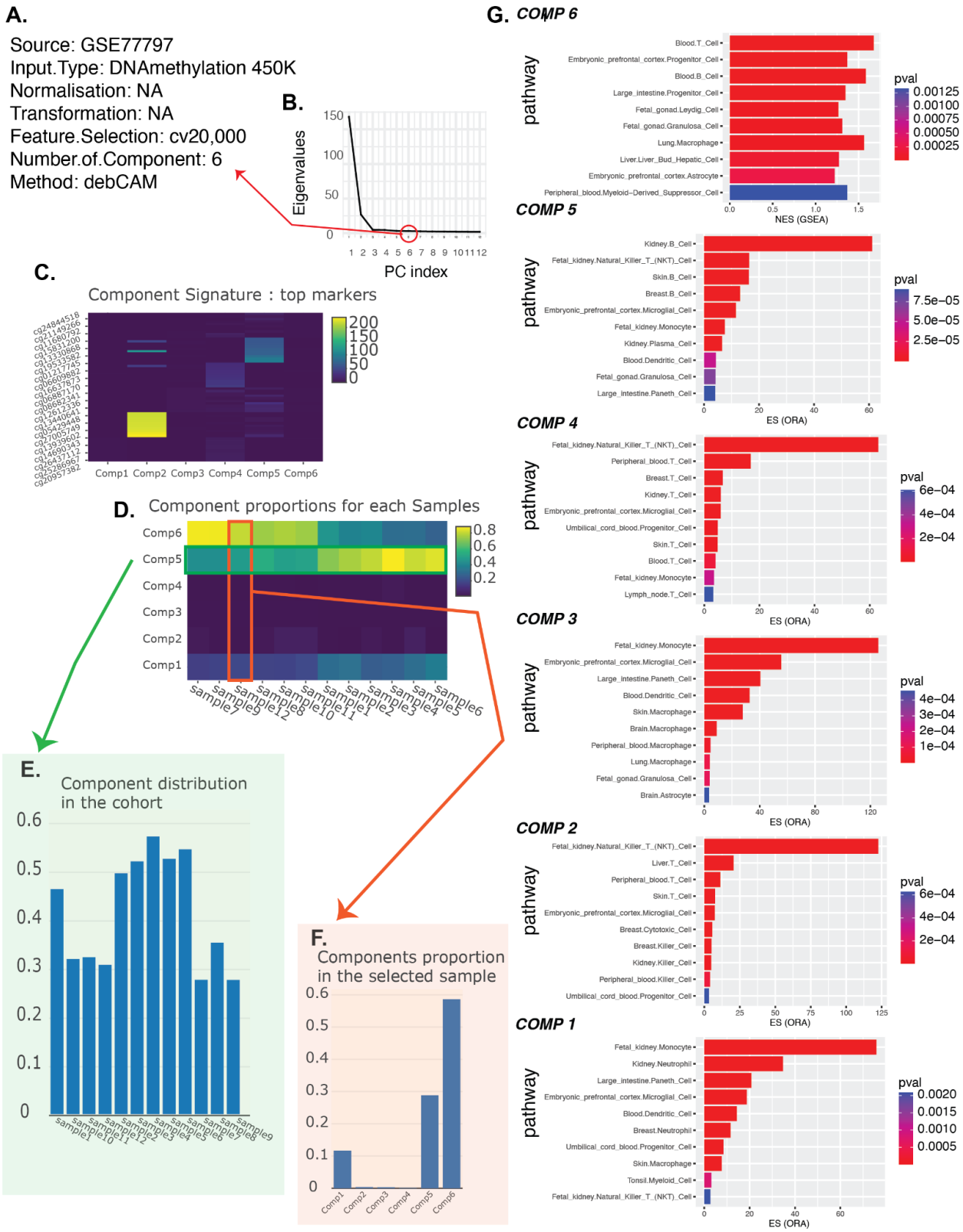

A.
Source: GSE77797
Input.Type: DNAmethylation 450K
Normalisation: NA
Transformation: NA
Feature.Selection: cv20,000
Number.of.Component: 6
Method: debCAM
B.
Eigenvalues
PC index
C.
Component Signature : top markers
Comp1 Comp2 Comp3 Comp4 Comp5 Comp6
D.
Component proportions for each Samples
Comp6
Comp5
Comp4
Comp3
Comp2
Comp1
E.
Component distribution in the cohort
F.
Components proportion in the selected sample
G.
COMP 6
COMP 5
COMP 4
COMP 3
COMP 2
COMP 1
pathway
pval
NES (GSEA)
ES (ORA)

Figure S1: Illustration of DECOMICS based deconvolution of DNA methylation data. **A.** Parameters used in the DECOMICS application are described. **B.** Scree plot illustrating the selection process for the number of components to be deconvolved (module 1, step 3: Number of cell types). **C-D.** Deconvolution results are presented: (**C**) Component signatures are plotted, and (**D**) a heatmap showing component proportions (module 1, step 5: Results). **E-F**. Visualization of component distribution: (**E**) Distribution of a specific component across the cohort, and (**F**) distribution of all components within a given sample (module 2, step 7: Proportion visualization). **G**. Enrichment plot displaying the enrichment analysis for each component (module 2, step 6: Enrichment analysis) ORA indicates if an Overrepresentation Analysis has been performed, GSEA indicates if a Gene Set Enrichment Analysis has been performed for a given sample. Pval corresponds to the adjusted p-values of the Enrichment Score (ES) after correction for multiple testing.

Methods	Approach	Omic type	Implementation	Date of publication	DOI	Nb of citations	OS compatibilities	Package accessibility	Imports	Depends	Methods URL	
A	ICA	both	R	2023	10.32614/CRAN.package.fastICA	NA	NA	cran		R (≥ 4.0.0)	https://cran.r-project.org/web/packages//fastICA/index.html	Yes
IF	NMF	both	R	2010	doi:10.1186/1471-2105-11-367	640	NA	cran	graphics, stats, stringr (≥ 1.0.0), digest, grid, grDevices, gridBase, colorspace, RColorBrewer, foreach, doParallel, ggplot2, reshape2, Biobase, codetools, BiocManager	R (≥ 3.0.0), methods, utils, registry, rngtools (≥ 1.2.3), cluster	https://cran.r-project.org/web/packages//NMF/index.html	Yes
Com	NMF + constrains	DNAm	R	2017	10.1186/s13059-017-1182-6	64	Only binary release accessible for MacOS, docker image for windows	github	RcppEigen (>= 0.3)	R (>= 3.2.0), Rcpp, pracma, gtools, gplots, parallel, RUnit, RnBeads	https://github.com/lutsik/MeDeCom	No (probl
AM	Convex analysis	both	R	2020	10.1093/bioinformatics/btaa205	9	NA	bioconductor	methods, rJava, BiocParallel, stats, Biobase, SummarizedExperiment, corpcor, geometry, NMF, nnls, DMwR2, pcaPP, apcluster, graphics	R (>= 3.5)	https://www.bioconductor.org/packages/release/bioc/html/debCAM.html	Yes
/RefFreeCell x	NMF	DNAm	R	2016	10.1186/s12859-016-1140-4	74	NA	removed from cran in 2021	NA	NA	NA	No, no lo (but simila
step1	NMF, equivalent to RefFreeEWAS	DNAm	R	2016	10.1016/j.celrep.2016.10.057	51	NA	github	quadprog, gtools, stats, clue, utils	R (>= 3.1.2)	https://github.com/BRL-BCM/EDec	Yes
DE	NMF	RNA	R	2020	10.1371/journal.pcbi.1008452	5	NA	github		R (>= 3.5.0), matrixStats, stats, utils, quadprog, gtools	https://github.com/Xiaoqizheng/PREDE/blob/master/DESCRIPTION	Yes (only
eqR	Probabilistic model	RNA	R	2021	10.1186/s12859-021-04186-5	3	NA	github	Rcpp (>= 1.0.3), MASS, foreach, doParallel, dirmult, RcppThread, iterators, parallel, grDevices, Seurat, ggplot2, magrittr, dplyr, rlang, Matrix, matrixStats, ggpubr, pheatmap, harmony	R (>= 3.6.0)	https://github.com/kkang7/CDSeq_R_Package	Yes
sal	SISAL (geometric algorithm)	DNAm	R	2021	10.1093/bioinformatics/btaa930	4	NA	bioconductor	stats, methods, SummarizedExperiment, corpcor, doParallel, parallel, ggplot2, tidyr, Ggally	R (>= 3.6), EpiDISH, limma, nnls, quadprog	https://bioconductor.org/packages/release/bioc/html/TOAST.html	No, beca generate output, ra selected
eth	NMF	DNAm	R	2022	10.1186/s12859-022-04893-7	1	NA	not a package			https://github.com/hedingqin/PRMeth	No (no R
M3	Convex analysis	both	R	2024	10.1093/bioinformatics/btae107	0	NA	github	methods, debCAM, stats, Biobase, SummarizedExperiment, corpcor, geometry, NMF, nnls, graphics, geometry	R (>= 3.6)	https://github.com/ChiungTingWu/CAM3/	

Table S1: Descriptive table of existing unsupervised deconvolution algorithm. This table provides a brief overview of each algorithm, highlighting their methodologies and key features.