




Data and text mining

# BIODICA: a computational environment for Independent Component Analysis of omics data

Nicolas Captier <sup>1,2,3,4,\*</sup>, Jane Merlevede<sup>1,2,3</sup>, Askhat Molkenov<sup>5</sup>, Ainur Seisenova<sup>5</sup>, Altynbek Zhubanchaliyev<sup>5</sup>, Petr V. Nazarov <sup>6</sup>, Emmanuel Barillot<sup>1,2,3</sup>, Ulykbek Kairov<sup>5</sup> and Andrei Zinovyev <sup>1,2,3,\*</sup>

<sup>1</sup>Institut National de la Santé et de la Recherche Médicale (INSERM), U900, F-75005 Paris, France, <sup>2</sup>Institut Curie, PSL Research University, F-75005 Paris, France, <sup>3</sup>MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, F-75006 Paris, France, <sup>4</sup>Laboratoire d'Imagerie Translationnelle en Oncologie, Institut Curie, INSERM U1288, PSL Research University, 91400 Orsay, France, <sup>5</sup>National Laboratory Astana, Center for Life Sciences, Nazarbayev University, Nur-Sultan 010000, Kazakhstan and <sup>6</sup>Multomics Data Science Research Group, Department of Cancer Research & Bioinformatics Platform, Luxembourg Institute of Health, L-1445 Strassen, Luxembourg

\*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 3, 2022; revised on March 29, 2022; editorial decision on March 31, 2022; accepted on April 4, 2022

## Abstract

**Summary:** We developed BIODICA, an integrated computational environment for application of independent component analysis (ICA) to bulk and single-cell molecular profiles, interpretation of the results in terms of biological functions and correlation with metadata. The computational core is the novel Python package *stabilized-ica* which provides interface to several ICA algorithms, a stabilization procedure, meta-analysis and component interpretation tools. BIODICA is equipped with a user-friendly graphical user interface, allowing non-experienced users to perform the ICA-based omics data analysis. The results are provided in interactive ways, thus facilitating communication with biology experts.

**Availability and implementation:** BIODICA is implemented in Java, Python and JavaScript. The source code is freely available on GitHub under the MIT and the GNU LGPL licenses. BIODICA is supported on all major operating systems. URL: <https://sysbio-curie.github.io/biodica-environment/>.

**Contact:** [nicolas.captier@curie.fr](mailto:nicolas.captier@curie.fr) or [andrei.zinovyev@curie.fr](mailto:andrei.zinovyev@curie.fr)

## 1 Introduction

The recent progress of high throughput omics technologies has made molecular data more accessible and has fostered the development of many computational analyses to exploit the rich information they offer. Such analyses require efficient tools to handle the high dimensionality of these data and reveal the underlying biological processes.

Independent component analysis (ICA) is a statistical and computational method which aims to represent observed signals as linear mixtures of independent latent factors. ICA has been successfully applied to omics data with the hypothesis that observed molecular profiles result from linear combinations of unobserved biological and technical processes (Liebermeister, 2002). In particular, it has been shown to extract interpretable and reproducible components and has stood out from other popular methods like principal component analysis (PCA) or non-negative matrix factorization (NMF) (Sompairac *et al.*, 2019).

Here, we present BIODICA, a complete computational environment for a user-friendly application of ICA to omics data. It encompasses a

set of tools to extract and interpret reproducible independent components, using methods that already proved to be successful in multiple studies (Aynaud *et al.*, 2020; Biton *et al.*, 2014) (Fig. 1).

## 2 Materials and methods

### 2.1 Stabilization procedure for extracting reproducible components

The computational core of BIODICA is the Python package *stabilized-ica*. It implements a stabilization procedure which addresses the variability of the solutions of ICA algorithms when run multiple times (Himberg and Hyvarinen, 2003). When applied to transcriptomics data, not only did this procedure provide a quantification of the significance of the independent components but it also extracted more reproducible ones than standard ICA (Cantini *et al.*, 2019). Besides, it allowed the development of an approach for selecting the optimal number of independent components to extract from omics data (Kairov *et al.*, 2017), which is also available in BIODICA.

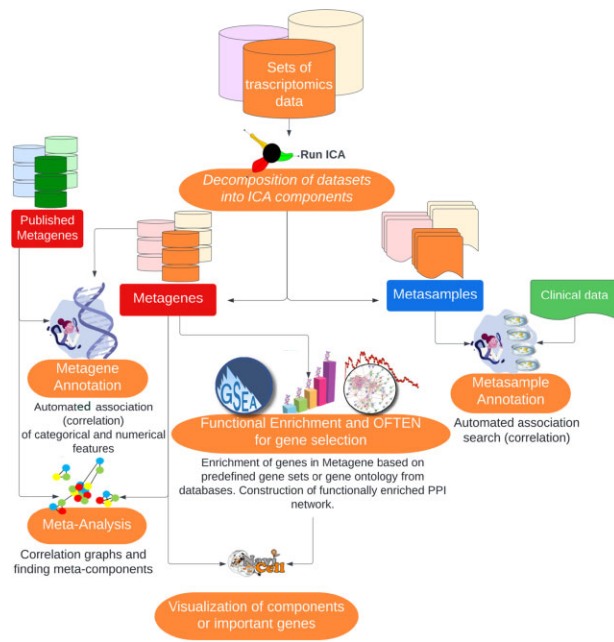


Fig. 1. BIODICA workflow

## 2.2 Biological annotation of extracted components

BIODICA provides a unique toolbox to help the biological interpretation of the extracted components, combining different annotation and visualization methods which already proved their usefulness (Kondratova et al., 2019; Teschendorff et al., 2007). Several knowledge-based annotation methods are proposed, such as functional enrichment analysis using ToppFun (Chen et al., 2009), Gene Set Enrichment Analysis (Subramanian et al., 2005) or network-based enrichment analysis using known graphs of protein–protein interactions (Kairov et al., 2012). BIODICA also integrates an insightful visualization tool to project the independent components on comprehensive maps of molecular interactions using NaviCell (Bonnet et al., 2015).

## 2.3 Studying inter-datasets reproducibility of extracted components

BIODICA provides a tool, based on application of mutual nearest neighbors (MNN), to match the components extracted from several independent omics datasets. Studying the reproducibility of independent components across multiple datasets may help distinguishing biological signals that are specific to a particular disease/data type or technical biases that are specific to particular conditions (Biton et al., 2014; Cantini et al., 2019).

## 3 Implementation

BIODICA comes with a user-friendly graphical user interface called BIODICA Navigator, providing non-experienced users a no-code access to all the BIODICA functionalities. It facilitates the communication with biology experts, producing sortable and interactive HTML-based reports. The interface has been designed and validated in several studies, including a study of Ewing sarcoma at single-cell level (Aynaud et al., 2020).

## 4 Future developments

For now, ICA-based blind deconvolution has been mainly applied to transcriptomics data. However, other omics technologies are now often incorporated into biological research and could also benefit

from the ICA analysis proposed by BIODICA. A few studies recently used ICA to deal with DNA methylation data (e.g. Meunier et al., 2021). We plan to foster the application of such methodologies to other omics data by adding new tools to BIODICA to interpret and exploit the independent components, taking into account the specificities of each technology. We also plan to integrate multi-omics analysis tools in BIODICA, building on the recent efforts that have been made for the integration of multiple omics layers (Teschendorff et al., 2018).

## Funding

The development of BIODICA was financially supported 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) and by the European Union’s Horizon 2020 program [826121, iPC project]. This work was also a part of the TIPIT project (Towards an Integrative approach for Precision ImmunoTherapy) funded by Fondation ARC call «SIGN’IT 2020—Signatures in Immunotherapy» and the IMMUCan project which has received funding from the Innovative Medicines Initiative 2 Joint Undertaking [821558]. The present study was supported by the research grants of the Ministry of Education and Science of the Republic of Kazakhstan [AP09058660], CRP NU [021220CRP222] ‘Identification of a long non-coding RNA (lncRNA) and microRNA in ESCC’. P.V.N. was supported by the Luxembourg National Research Fund [C17/BM/11664971/DEMICS].

*Conflict of Interest:* none declared.

## References

- Aynaud, M.-M. et al. (2020) Transcriptional programs define intratumoral heterogeneity of Ewing sarcoma at single-cell resolution. *Cell Rep.*, **30**, 1767–1779.e6.
- Biton, A. et al. (2014) Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Rep.*, **9**, 1235–1245.
- Bonnet, E. et al. (2015) NaviCell Web Service for network-based data visualization. *Nucleic Acids Res.*, **43**, W560–W565.
- Cantini, L. et al. (2019) Assessing reproducibility of matrix factorization methods in independent transcriptomes. *Bioinformatics*, **35**, 4307–4313.
- Chen, J. et al. (2009) ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. *Nucleic Acids Res.*, **37** (Suppl. 2), W305–W311.
- Himberg, J. and Hyvarinen, A. (2003) Icaasso: software for investigating the reliability of ICA estimates by clustering and visualization. In: *2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No. 03TH8718)*, Toulouse, France, pp. 259–268.
- Kairov, U. et al. (2012) Network analysis of gene lists for finding reproducible prognostic breast cancer gene signatures. *Bioinformatics*, **8**, 773–776.
- Kairov, U. et al. (2017) Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics*, **18**, 712.
- Kondratova, M. et al. (2019) A multiscale signalling network map of innate immune response in cancer reveals cell heterogeneity signatures. *Nat Commun.*, **10**, 4808.
- Liebermeister, W. (2002) Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, **18**, 51–60.
- Meunier, L. et al. (2021) DNA methylation signatures reveal the diversity of processes remodeling hepatocellular carcinoma methylomes. *Hepatology*, **74**, 816–834.
- Sompairac, N. et al. (2019) Independent component analysis for unraveling the complexity of cancer omics datasets. *Int. J. Mol. Sci.*, **20**, 4414.
- Subramanian, A. et al. (2005) Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 15545–15550.
- Teschendorff, A.E. et al. (2007) Elucidating the altered transcriptional programs in breast cancer using independent component analysis. *PLoS Comput. Biol.*, **3**, e161.
- Teschendorff, A.E. et al. (2018) Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biol.*, **19**, 76.