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

Nicolas Captier, Jane Merlevede, Askhat Molkenov, Ainur Seisenova, Altynbek Zhubanchaliyev, Petr Nazarov, Emmanuel Barillot, Ulykbek Kairov, Andrei Zinovyev

BIODICA: a computational environment for Independent Component Analysis

Nicolas Captier[1236]*, Jane Merlevede[123], Askhat Molkenov[4], Ainur Seisenova[4], Altynbek Zhubanchaliyev[4], Petr V. Nazarov[5], Emmanuel Barillot[123], Ulykbek Kairov[4] and Andrei Zinovyev[123]*

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

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; 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 (Biton *et al.*, 2014; Aynaud *et al.*, 2020).

2. 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.

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 (Teschendorff *et al.*, 2007; Kondratova *et al.* 2019)[1]. Several knowledge-based annotation methods are proposed, such as functional enrichment analysis using ToppFun (Chen *et al.*, 2009)[1], Gene Set Enrichment Analysis (Subramanian *et al.*, 2005)[1], or network-based enrichment analysis using known graphs of protein-protein interactions (Kairov *et al.*, 2012)[1]. BIODICA also integrates an insightful visualization tool to project the independent components on comprehensive maps of molecular interactions using NaviCell (Bonnet *et al.*, 2015)[1].

2.3. Studying inter-data sets 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 data sets. Studying the reproducibility of independent components across multiple data sets 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).

For more advanced users, the stabilized-ica package can be used as a standalone tool to integrate some of the functionalities of BIODICA into Python analysis pipelines.

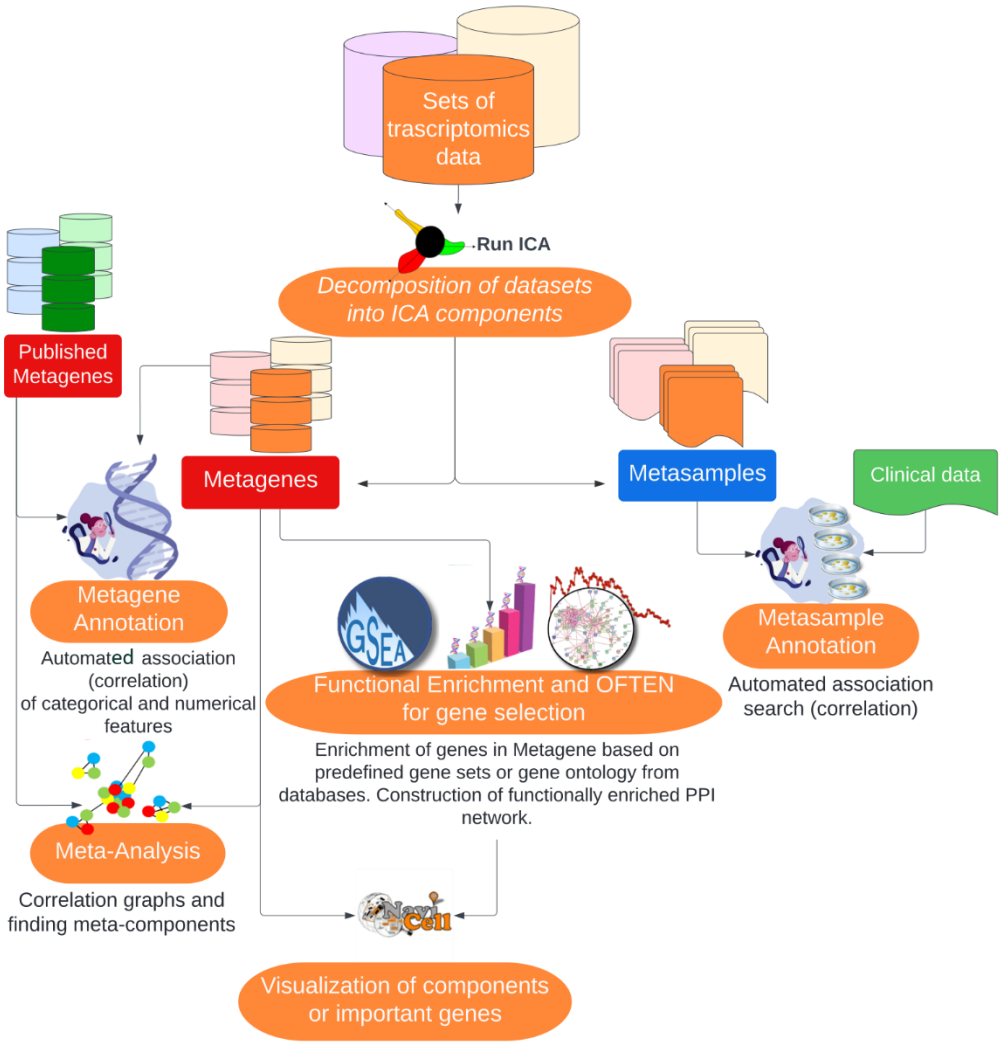
4. Future developments

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

[1] These references are available at https://sysbio-curie.github.io/biodica-environment/docs/references/

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).

Fig. 1. BIODICA workflow



5. Acknowledgements

The development of BIODICA has been financially supported by 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 European Union’s Horizon 2020 program (grant No. 826121, iPC project). This work is 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 (grant No. 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 grant 021220CRP222 “Identification of a long non-coding RNA (lncRNA) and mi-croRNA in ESCC”. Petr Nazarov was supported by the Luxembourg National Research Fund (C17/BM/11664971/DEMICS).

References

Aynaud, M.-M., Mirabeau, O., Gruel, N., Grossetête, S., Boeva, V., Durand, S., Surdez, D., Saulnier, O., Zaïdi, S., Gribkova, S., Fouché, A., Kairov, U., Raynal, V., Tirode, F., Grünewald, T. G., Bohec, M., Baulande, S., Janoueix-Lerosey, I., Vert, J.-P., Barillot, E., Delattre, O., and Zinovyev, A. (2020). Transcriptional programs define intratumoral heterogeneity of ewing sarcoma at single-cell resolution. Cell Reports, 30(6), 1767–1779.e6.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S., Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., and Radvanyi, F. (2014). Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Reports,9(4), 1235–1245.

Cantini, L., Kairov, U., de Reyniès, A., Barillot, E., Radvanyi, F., and Zinovyev, A. (2019). Assessing reproducibility of matrix factorization methods in independent transcriptomes. Bioinformatics,35(21), 4307–4313.

Himberg, J. and Hyvarinen, A. (2003). Icasso: software for investigating the reliability of ica estimates by clustering and visualization. In2003 IEEE XIII Workshop on Neural Networks for Signal Processing (IEEE Cat. No.03TH8718), pages 259–268.

Kairov, U., Cantini, L., Greco, A., Molkenov, A., Czerwinska, U., Barillot, E., and Zinovyev, A. (2017). Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics,18(1), 712.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. Bioinformatics,18(1), 51–60.

Meunier, L., Hirsch, T. Z., Caruso, S., Imbeaud, S., Bayard, Q., Roehrig, A., Couchy, G., Nault, J.-C., Llovet, J. M., Blanc, J.-F., Calderaro, J., Zucman-Rossi, J., and Letouzé, E. (2021). Dna methylation signatures reveal the diversity of processes remodeling hepatocellular carcinoma methylomes. Hepatology, 74(2), 816–834.

Sompairac, N., Nazarov, P. V., Czerwinska, U., Cantini, L., Biton, A., Molkenov, A., Zhumadilov, Z., Barillot, E., Radvanyi, F., Gorban, A., Kairov, U., and Zinovyev, A. (2019b). Independent component analysis for unraveling the complexity of cancer omics datasets. International Journal of Molecular Sciences, 20(18)

Teschendorff, A. E., Jing, H., Paul, D. S., Virta, J., and Nordhausen, K. (2018). Tensorial blind source separation for improved analysis of multi-omic data. Genome Biology, 19(1), 76.