Benchmarking of multi-omics joint dimensionality reduction (DR) approaches for cancer study

Laura Cantini[1], Pooya Zakeri[2], Aurélien Naldi[1], Denis Thieffry[1], Elisabeth Remy[3] and Anaïs Baudot[2,4]

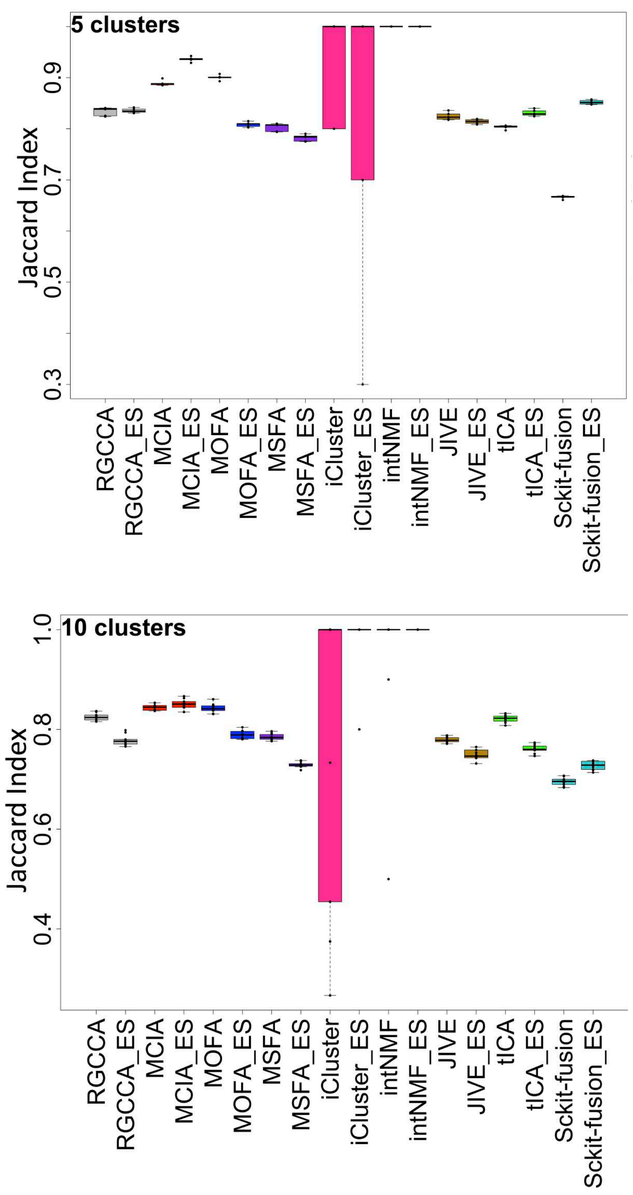
[1]Computational Systems Biology Team, Institut de Biologie de l'Ecole Normale Supérieure, CNRS UMR8197, INSERM U1024, Ecole Normale Supérieure, Paris Sciences et Lettres Research University, 75005 Paris, France.

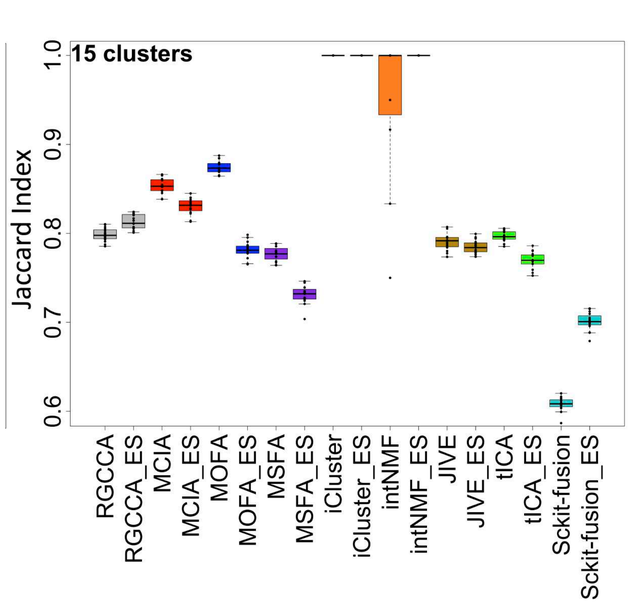
[2]Aix Marseille Université, INSERM, MMG, CNRS, 13005 Marseille, France.
[3]Aix Marseille University, CNRS, Centrale Marseille, I2M, Marseille, France.
[4]Barcelona Supercomputing Center (BSC), 08034 Barcelona, Spain.

Why multi-omics in cancer study? High-dimensional data are now a standard in biology, particularly in cancer biology, where national and international consortia have profiled thousands of patients for multiple molecular assays ("multi-omics"). Moreover, multi-omics single-cell data are emerging, opening to the need of approaches able to deal with their volume superior to the one of bulk data.

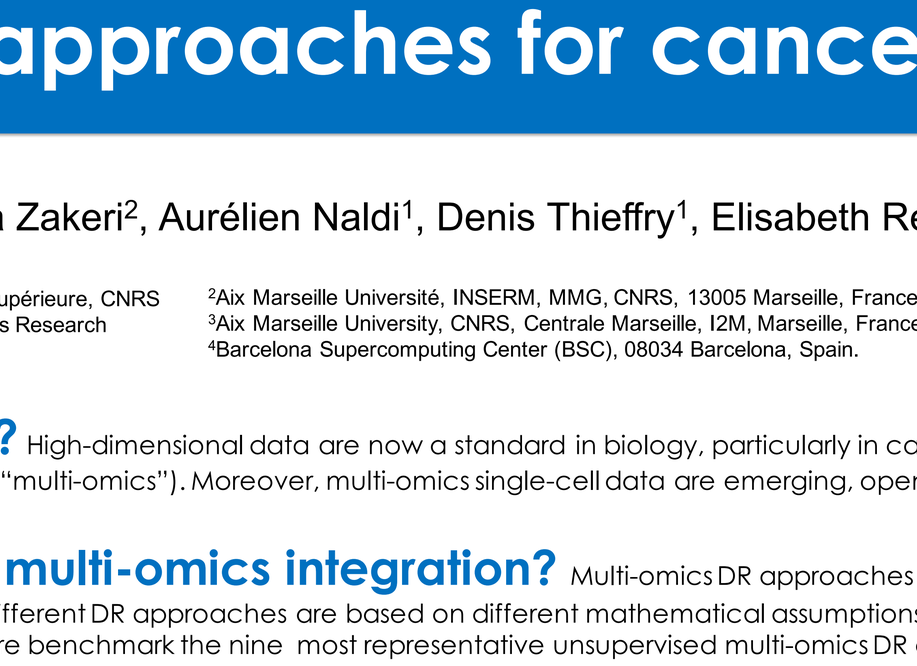
Why Dimensionality Reduction in multi-omics integration? Multi-omics DR approaches jointly decompose omic datasets into low-dimensional spaces while preserving most of their original information content. Different DR approaches are based on different mathematical assumptions, which makes difficult to chose which method to prioritize depending on the particular problem under analysis. We here benchmark the nine most representative unsupervised multi-omics DR approaches [1-9].

Simulated omics

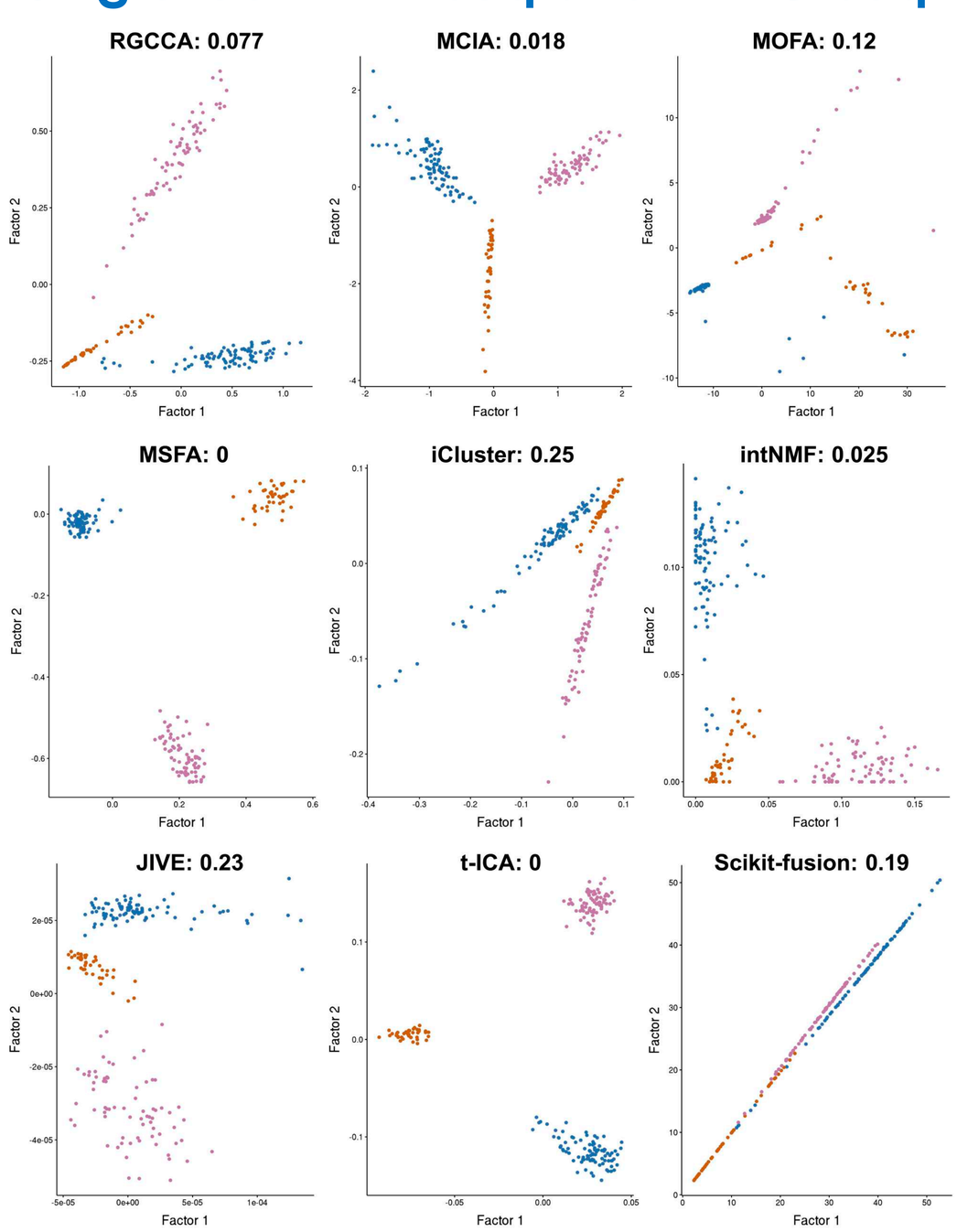




Jaccard-index between the clusters imposed in the simulated data and those retrieved by the different methods. Each plot refers to a different number of clusters: 5, 10, 15. For each method (e.g. RGCCA) both its performances on heterogeneous and equally-sized clusters are reported (denoted as RGCCA and RGCCA_ES, respectively).

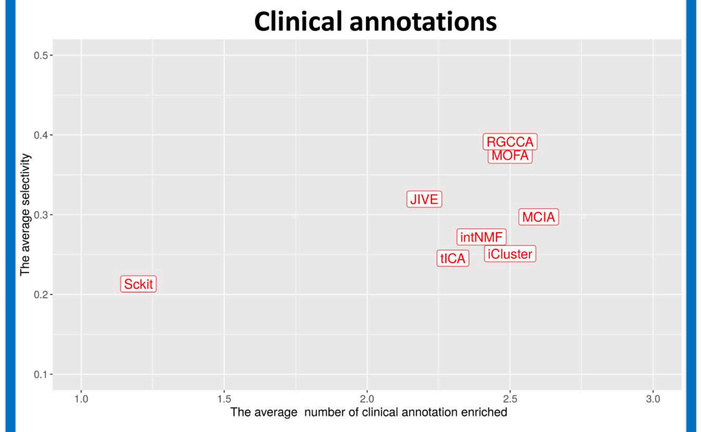


Single-cell RNA-seq and ATAC-seq

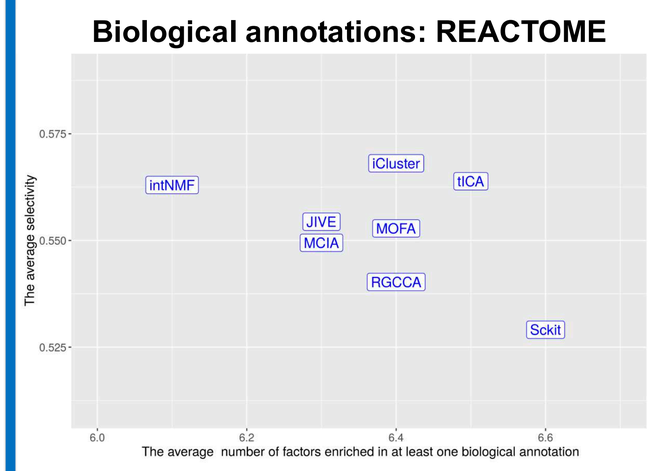


Scatterplots of Factor 1 and 2 are reported for each methodology. The colors denote the cancer cell line of origin: pink for K562, orange for Hela and blue for HCT. For each method, the C-index, evaluating the quality of the obtained clusters, is also reported.

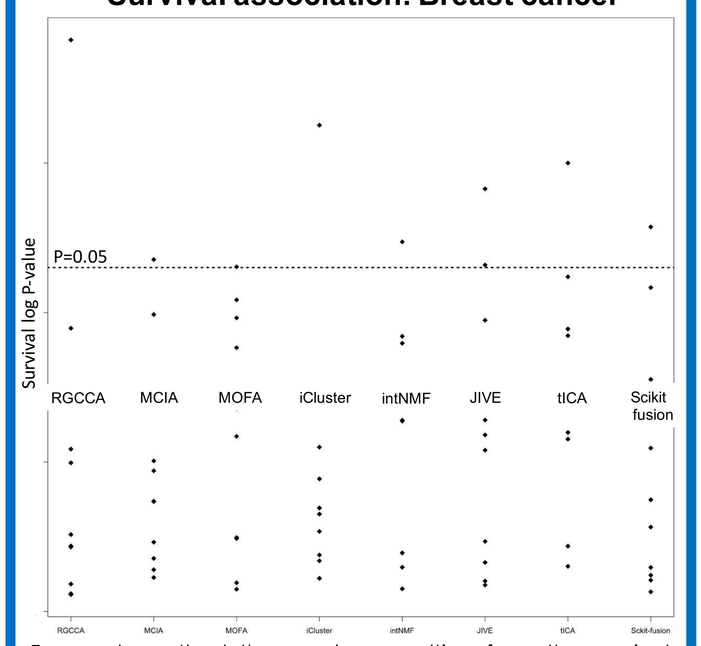
TCGA cancer omics



For each method, the number of their associated clinical annotations is compared with their selectivity. The across-cancer average behavior is here summarized.



For each method, the number of factors enriched in at least a biological annotation is compared with their selectivity. The across-cancer average behavior is here summarized.



For each method the p.values resulting from the survival analysis applied to its factors are reported.

Do you want to compare your method with those here tested?

Do you want to compare the existing methods on your data?

Use our Jupyter notebook

https://github.com/ComputationalSystemsBiology/momix-notebook

REFERENCES:

[1] Tenenhaus, A., and Tenenhaus, M. *Psychometrika* (2011).
[2] Meng, C., et al. *BMC bioinformatics* (2014).
[3] Argelaguet, R., et al. *Molecular systems biology* (2018).
[4] De Vito, R., et al. *Biometrics* (2019).
[5] Chalise, P., and Brooke L. F. *PloS one* (2017).
[6] Shen, R., et al. *Bioinformatics* (2009).
[7] Lock, E. F., et al. *The annals of applied statistics* (2013).
[8] Teschendorff, A. E., et al. *Genome biology* (2018).
[9] Žitnik, M. and Blaž Z., *IEEE transactions on pattern analysis and machine intelligence* (2014).

ABOUT THE AUTHOR:

Laura Cantini
CNRS Researcher
I am passionate about **Computational Biology, Network Theory**.
Hope you enjoyed the poster and do not hesitate to contact me!

Paris, France
laura.cantini@ens.fr
/profile/Laura_Cantini
lcan88