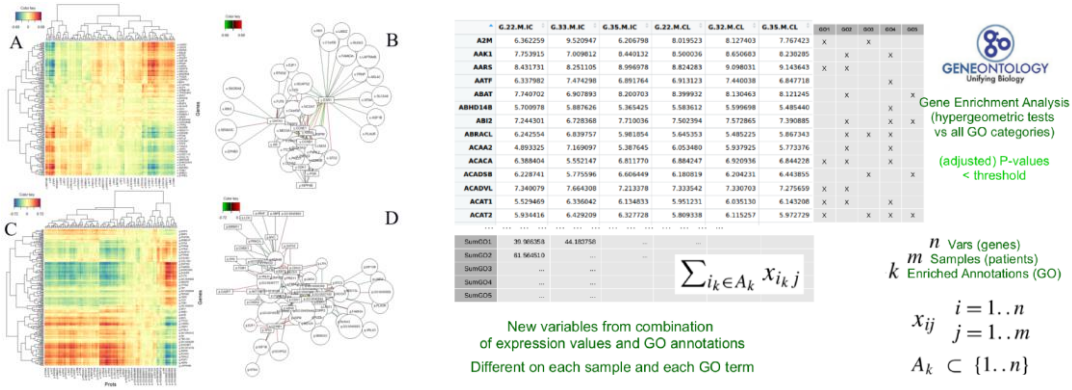
Integrative Analysis of Multi-Omics Data with Addition of Biological Knowledge

Ferran Briansó[1], *Alex Sánchez-Pla*[2]

[1]ferran.brianso@gmail.com, [2]asanchez@ub.edu

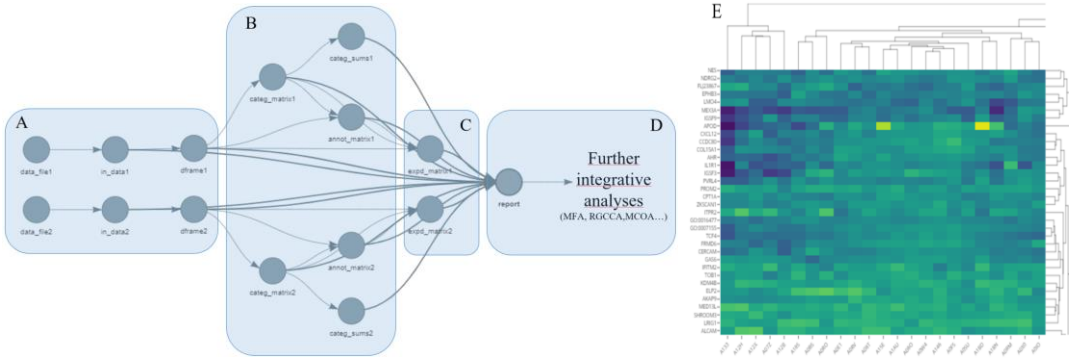
Department of Genetics, Microbiology and Statistics, University of Barcelona

Integrative analysis of multiple omics data allows, not only for the combination of heterogeneous data, but also for the combined use of biological data to extract information that could not be unveiled by the separated analysis of each of the original data types [Gomez-Cabrero, 2014]. One common approach to omics integration is using dimension reduction methods, which are also helpful for data visualization [Meng, 2016]. There is however one point that may be improved: the difficulty in interpreting results from the biological point of view [Yamada, 2021]. In the work presented here, biological annotations, such as GO Terms, Gene Sets or custom annotations, are combined with numerical values, such as protein or gene expression, using multiple factor analysis or related techniques, allowing to improve interpretability and providing better biomedical insights.



Left: Some of the results of an analysis of 150 samples from TCGA. Heat maps (**A, C**) and association networks (**B, D**) resulting from the integration by Regularized Canonical Correlations Analysis with mixomics R package. Performed with the original data sets (**A, B**) or using data expanded with biological annotations to Gene Ontology (**C, D**), so adding some GO terms to the features from each source, where outputs contain higher level of information (higher density in both type of plots). **Right:** Representation of the process of integrating those annotated GO terms as new variables, new rows, as proposed by Busold et al. in 2005.

An R package with the methods developed, and a pipeline using the *targets* package, have been implemented to facilitate reproducibility and application to biomedical research.



Workflow of the steps implemented in the annotation and expansion of omics data: A couple of 'omics-derived input data sets (e.g. pre-processed gene expression and protein abundance matrices) are converted to R data frames (**A**). Annotations are created, or loaded, as additional objects (**B**), one for each given input matrix, and used to expand these original data, to end up with a pair of data frames (**C**) containing the starting values plus the average expression/abundance values of the features related to each annotation as new variables. Finally, an R markdown report is rendered to show steps and main results of the process, and the output is used for further integrative analyses (**D**). Snapshot (**E**) of one of the heat maps created to show the expanded matrices resulting from the analysis.

 Keywords: Multi-Omics; Integrative Analysis; Biological Annotations.