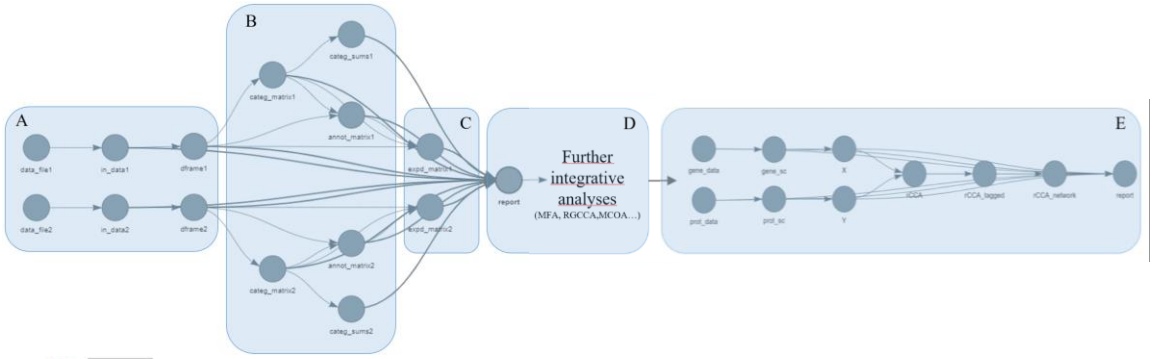
An R pipeline using the "*targets*" package for Multi-Omics Integrative Analyses

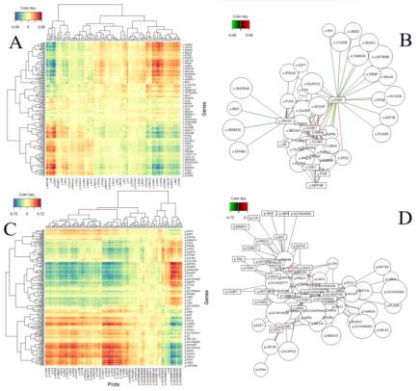
Ferran Brianso[1], *Alex Sánchez-Pla*[2]

[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]. Several tools have been implemented as frameworks for the development of bioinformatic pipelines [Leipzig, 2017; Wratten, 2021], in most cases conforming complex workflow managers, such as Galaxy [usegalaxy.org] and NextFlow [www.nextflow.io]. These platforms, however, require a certain degree of technical knowledge in order to be configured, often away from the standard technical level of biomedical researchers and analysts. In the work presented here, we show a pipeline for the integrative analysis of multi-omics data, implemented with the *targets* package [books.ropensci.org/targets] that, unlike most pipeline toolkits, which are language agnostic or Python-focused, allows data scientists and researchers to work entirely within R. The example presented here performs an integrative multi-omics analysis, combining protein and gene expression data (from public sources such as TCGA) with biological annotations (GO, pathways or custom annotations), and applying distinct dimension reduction techniques (Regularized Canonical Correlations Analysis and Multiple Factor Analysis) implemented in proven reliable packages (mixOmics[mixomics.org/] and FactoMineR[factominer.free.fr/], respectively).





Above: *Targets* workflow of the steps implemented in the multi-omics integrative pipeline: 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. Then, a first R markdown report is rendered to show steps and main results of the annotate-and-expand process, and this output is used for further integrative analyses (**D**). In this case, the Regularized Canonical Correlations Analysis carried out with mixOmics (**E**) that ends up with a second markdown report.

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 RCCA 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).

 Keywords: Workflow Pipelines; Multi-Omics; Integrative Analysis.