**‘Omics data analysis pipeline**

Andrej Blejec,

Kristina Gruden &

Maja Zagorščak

National Institute of Biology (NIB)

ADAPT: Accelerated Development of multiple-stress tolerant potato

Horizon 2020 [H2020-SFS-2019-2, Grant Agreement Number 862858]

Abstract

This protocol describes multi-omics data integration and modelling and systems biology analysis and visualization pipeline in R

Keywords: multiOmics, multi-omics, integrative omics, panOmics, pan-omics

‘Omics data analysis pipeline

# Prior to data analysis and integration, some prerequisites should be met.

# Prerequisites and protocol

**Data management framework**

All data should be annotated with detailed metadata, including *Phenodata* –a master sample description table, and preferably structured according to the [pISA-tree](https://github.com/NIB-SI/pISA-tree) data management framework (1). Each measurement should be stored as a separate data file, including SampleName (and/or SampleID) and measurements only, preferably as tab-separated text file. If using Excel as primary measurement storage format, prior to analyses, export each data set as tab separated text file. Avoid merged cells and Excel calculations in exported files under any cost.

**Expected measurements**

Omics' strategies include: Hormonomics, Transcriptomics, Proteomics (non-targeted), Metabolomics, Phenomics and more. Measurements can include single or multiple genotypes with single or multiple tissues under single or combine abiotic or biotic stressors. Preferential measurements are time-series measurements.

**Analysis steps**

### Step 1: Exprimental design master table

Design *Phenodata*, a master experimental design table describing samples for analysis, prior to sample collection according to good data management practice. Store [Phenodata](_p_Omics/_I_Omics/phenodata_20221001.txt) at [*Investigation*](_p_Omics/_I_Omics) level. Define relative path of *Phenodata* in [\_INVESTIGATION\_METADATA.TXT](_p_Omics/_I_Omics/_INVESTIGATION_METADATA.TXT), as well as in [\_ASSAY\_METADATA.TXT](_p_Omics/_I_Omics/_S_multiOmics/_A_multiOmics-integration-R/_ASSAY_METADATA.TXT). Phenodata must contain SampleName (and/or SampleID) column, which will be utilized to combine measurements with sample descriptions.

### Step 2: Data preprocessing and overall inspection

Prior to analyses, it is expected that the analyst conducted data preprocessing and overall inspection, which might include: i) detection of outliers and faulty measurements, ii) data transformation, iii) interpolation, iv) extrapolation and, v) imputation. For qPCR imputation suggestions see (2). For other steps see suggested packages in the <README.md> file of this repository.

**Step 3: Statistical analysis of individual omics data layers**

This step focuses on various within-level correlations calculations and visualization, calculation of correlation differences, multidimensional scaling, t-tests and log2FC calculations. Example of input data, consisting of three Omics’ levels can be found within [*Assay*](_p_Omics/_I_Omics/_S_multiOmics/_A_multiOmics-integration-R) input directory. Experimental design can be inspected from 01\_Experimental-design-and-days-of-tissue-sampling.xlsx file, stored at [*Investigation*](_p_Omics/_I_Omics) level. In this scenario, SampleName column was created from condition, time point, and plant replicate number. Plants were exposed to Heat stress (H, later defined as ‘Stress’); measurements were taken at days: 1, 7, 8, and 14; which is denoted under Treatment and SamplingDay columns of *Phenodata*. For statistical analysis of individual omics data layers prepare data in similar manner, and use scripts from *Assay* \_A\_multiOmicsStat-R. Main packages and functions are listed in the <README.md> file of this repository.

**Step 4: Correlation based network inference within each omics level**

Correlation-based network can be constructed using ‘Leave-One-Out’ or [Lioness](https://bioconductor.org/packages/release/bioc/vignettes/lionessR/inst/doc/lionessR.html) approach from all Omics’ levels. Since they are highly depended on thresholds we advise to use thresholding approach from *Assay* \_A\_multiOmics-differential-networks-Py.

**Step 5: Integration across different omics datasets**

This step focuses on Canonical Correlation Analysis and N-Integration Discriminant Analysis with DIABLO using mixOmics (3). Prepare data in a similar manner and run master script from *Assay* \_A\_DiABLO-R. Main packages and functions are listed in the <README.md> file of this repository. Read more about DIABLO at <https://mixomics.org/mixDIABLO/>.

**Step 6: Integration of data with prior knowledge**

Log2FC values from Step 3, or standard differential expression analysis, can be visualised in prior knowledge network in a differential network context using [Cytoscape](https://cytoscape.org/). Cystoscape manual is available at <https://manual.cytoscape.org/en/stable/>. Example of manually created network for this specific data set can be found within the [input](_p_Omics/_I_Omics/_S_multiOmics/_A_multiOmics-integration-R/input) directory of the [*Assay*](_p_Omics/_I_Omics/_S_multiOmics/_A_multiOmics-integration-R)\_A\_multiOmics-differential-networks-Py. For more prior knowledge network for plant species see <https://skm.nib.si/biomine/> and <https://skm.nib.si/downloads/>.

Footnotes

1References Petek, M., Zagorščak, M., Blejec, A. et al. pISA-tree - a data management framework for life science research projects using a standardised directory tree. Sci Data 9, 685 (2022). <https://doi.org/10.1038/s41597-022-01805-5>

2Baebler, Š., Svalina, M., Petek, M. et al. quantGenius: implementation of a decision support system for qPCR-based gene quantification. BMC Bioinformatics 18, 276 (2017). <https://doi.org/10.1186/s12859-017-1688-7>

3Rohart F, Gautier B, Singh A, Lê Cao KA. mixOmics: An R package for 'omics feature selection and multiple data integration. PLoS Comput Biol. 2017 Nov 3;13(11):e1005752. doi: 10.1371/journal.pcbi.1005752. PMID: 29099853; PMCID: PMC5687754.