# A WORKFLOW TO INTEGRATE PRE-PROCESSING, ANALYSIS AND COMPARISON OF MALDI-TOF MASS SPECTRA IN GEENAR

E. Del Prete (1,2), A. Facchiano (2), A. Profumo (3), C. Angelini (4), and P. Romano (3)

(1) Dipartimento di Scienze, Università della Basilicata, Viale dell'Ateneo Lucano 10, 85100, Potenza (Italy)
(2) Istituto di Scienze dell’Alimentazione, CNR, Via Roma 64, 83100 Avellino (Italy)
(3) IRCCS AOU San Martino IST, Largo Rosanna Benzi 10, 16132 Genova (Italy)
(4) Istituto per le Applicazioni del Calcolo, CNR, Via Pietro Castellino 111, 80131 Napoli (Italy)

## MOTIVATION

Many large-scale proteomics studies have been performed in the last years, and this field of investigation is expanding up. If the analysis of any single spectrum can be performed by tools already made available along with the mass spectrometry (MS) instrumentation, comparison of spectra on a large scale represents a complex aspect of the analysis and interpretation of the study.

Recently, we developed Geena 2 [1], a tool for the automation of different steps in the MALDI/ToF MS data analysis. Integration of further tools can be performed, in order to improve some aspects of the whole workflow: the input of more data formats, the implementation of new algorithms for data cleaning, the graphical visualization and the reporting of the results, the use of advanced statistics for the comparison of mass spectra. For this motivations, we are now developing GeenaR, a new robust web tool for pre-processing, analysing, visualizing and comparing a set of MALDI-ToF mass spectra. The aim of this work is the presentation of on-going developments. The first results will be presented at the conference.

## METHODS

GeenaR is being written in PHP, Perl (from Geena 2) and R languages. The R packages used are *MALDIquant* and *MALDIquantForeign* [2] for mass spectra pre-processing and analysis, *OrgMassSpecR* [3] for mass spectra comparison, *dendextend* [4] and *pvclust* [5] for clustering, and *sda* [6] and *crossval* [7] for variable selection. The system is being implemented in a LAMP (Linux, Apache, MySQL, PHP) environment. Proper interfaces between PHP on one side and perl and R on the other are then implemented.

## RESULTS

The aim of GeenaR is to provide to the users a wider range of statistical methods and graphical results, without making it more difficult to use for researchers with little expertise in programming. In order to achieve this goal, we have taken advantage of the availability of several packages, written in R language [8], for mass spectrometry statistical that are going to be integrated in the system. The complete pipeline of GeenaR (see Figure 1) includes some features already available in Geena 2 plus others under development thanks to the integration of the R environment. In fact, an original set of heuristic algorithms is already available in Geena 2. In particular, they are the identification of isotopic peaks by taking into account molecular weight of signals and the related trend of abundances; the normalization on the basis of a reference standard molecule; the peak selection by means of a threshold line, built by linearly interpolating values provided for given m/z values; the alignment, by

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2255v1 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016

selecting the nearest peaks, within a limited m/z difference, in the different mass spectra. By means of some R packages [9], GeenaR adds new statistical methods which are highly relevant for mass spectra analysis, like e.g., square root transformation for the variance stabilization, Savitsky-Golay filter for the smoothing, Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP) algorithm for the baseline correction, Total Ion Current (TIC) method for the normalization, Local Weighted Scatterplot Smoothing (LOWESS) technique for the alignment, Median Absolute Deviation (MAD) method for the peak detection, cosine correlation as similarity measure, Linear Discriminant Analysis (LDA) method for the variable selection.

The proposed pipeline is complete since it starts from the spectra generated by spectrometers, pre-processes raw data, extracts relevant information and executes advanced statistical analyses. The adoption of an extended statistical control supports in-silico validation of data, reproducibility of analyses and accuracy of results.

## ACKNOWLEDGMENTS

This work is partially supported by the Flagship InterOmics Project (PB.P05, funded and supported by the Italian Ministry of Education, University and Research and Italian National Research Council organizations).

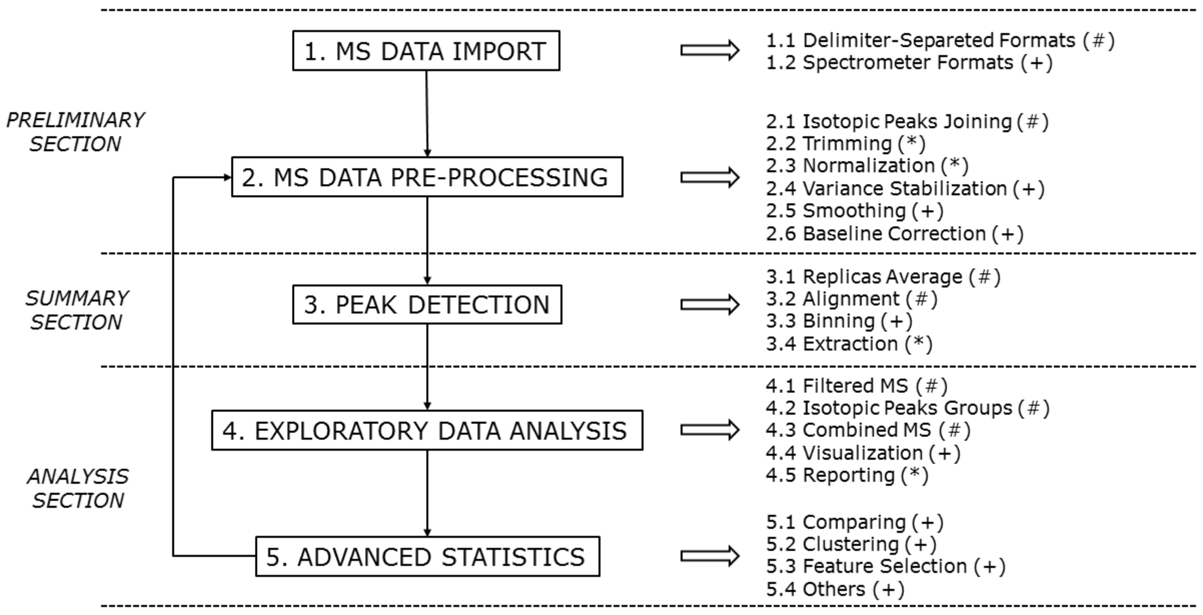

**Figure 1: GeenaR analysis workflow.** The pipeline is divided in three main sections: preliminary, summary and analysis. Each section incorporates different tools denoted with three distinct symbols: (#) tools already available in Geena 2; (*) tools available in Geena 2 and that will be extended in GeenaR by means of novel algorithms; (+) tools currently under development and/or implementation, that will be included in GeenaR. The feedback arrow on the left represents a mass spectrometry data re-processing, useful in order to refine results after possible unexpected evidences.

## REFERENCES

[1] P. Romano, A. Profumo, M. Rocco, R. Mangerini, F. Ferri, and A. Facchiano, “Geena 2, improved automated analysis of MALDI/TOF mass spectra,” *BMC Bionformatics*, 17(Suppl 4):61, pp. 247-269, 2016.

[2] S. Gibb, and S. Strimmer, “MALDIquant: a versatile R package for the analysis of mass spectrometry data”, *Bioinformatics*, 28(17), pp. 2270-2271, 2012.8.

[3] N. G. Dodder, and K. M. Mullen, “OrgMassSpecR: organic mass spectrometry,” R package version 0.4-4, 2014. URL: http://CRAN.R-project.org/package=OrgMassSpecR.

[4] T. Galili, “dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering,” *Bioinformatics*, 31(22), pp. 1-3, 2015.

[5] R. Suzuki, and H. Shimodaira, “pvclust: hierarchical clustering with p-values via multiscale bootstrap resampling,” R package version 2.0-0, 2015. URL: http://CRAN.R-project.org/package=pvclust.

[6] M. Ahdesmaki, V. Zuber, S. Gibb, and K. Strimmer, “sda: shrinkage discriminant analysis and CAT score variable selection,” R package version 1.3.7, 2015. URL: http://CRAN.R-project.org/package=sda.

[7] K. Strimmer, “crossval: generic functions for cross validation,” R package version 1.0.3, 2015. URL: http://CRAN.R-project.org/package=crossval

[8] R Core Team, “R: a language and environment for statistical computing,” R Foundation for Statistical Computing, Vienna, Austria, 2014. URL: http://www.R-project.org/.

[9] R. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, et al., “Bioconductor: open software development for computational biology and bioinformatics”, *Genome Biology*, 5, 2004. URL: https://www.bioconductor.org/.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2255v1 | CC BY 4.0 Open Access | rec: 6 Jul 2016, publ: 6 Jul 2016