Suitable reporting for the reproducible research: an added value in the analysis of proteomics data

Eugenio Del Prete[1,2], Angelo Facchiano[2], Aldo Profumo[3], Claudia Angelini[4], Paolo 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)

Introduction

Computational reproducibility is a fundamental issue in omic studies because of the complex and high-dimensional nature of data. The **analysis of proteomics** data needs to exploit multistep workflows including pre-processing, elaboration, statistical validation, interpretation and presentation. The availability of source code increases the quality of research in terms of transparency and knowledge transfer. Moreover, it allows other researchers to **reproduce the results** in a local system, make a comparison among the results and re-use code for analyzing different datasets.

Methods

GeenaR is an extension of Geena 2 still under development. Fig.1 shows the workflow developed for the **analysis of proteomics data** (mass spectra).

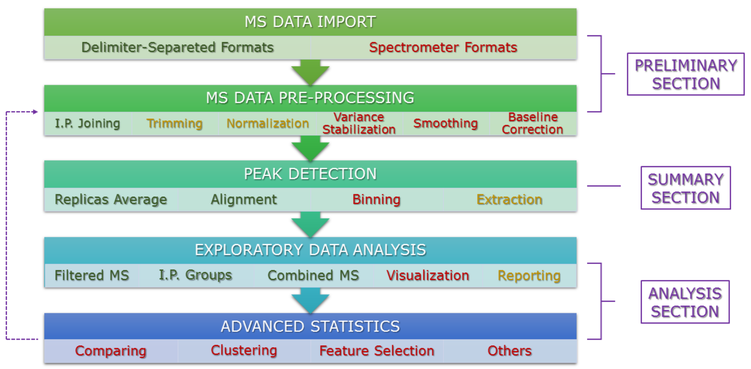


Fig. 1. *GeenaR analysis workflow. Green: already available; Yellow: under extension with R packages; Red: novel tools under development with R packages.*

Reproducible research is implemented using the following R features: the ***R-Markdown*** language, the ***knitr*** package and the ***spin*** function.

Results

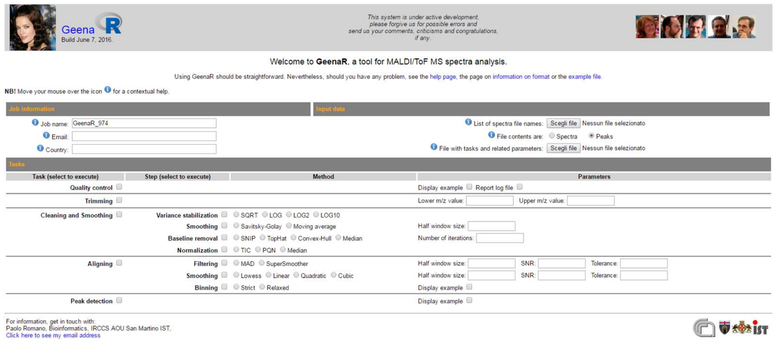
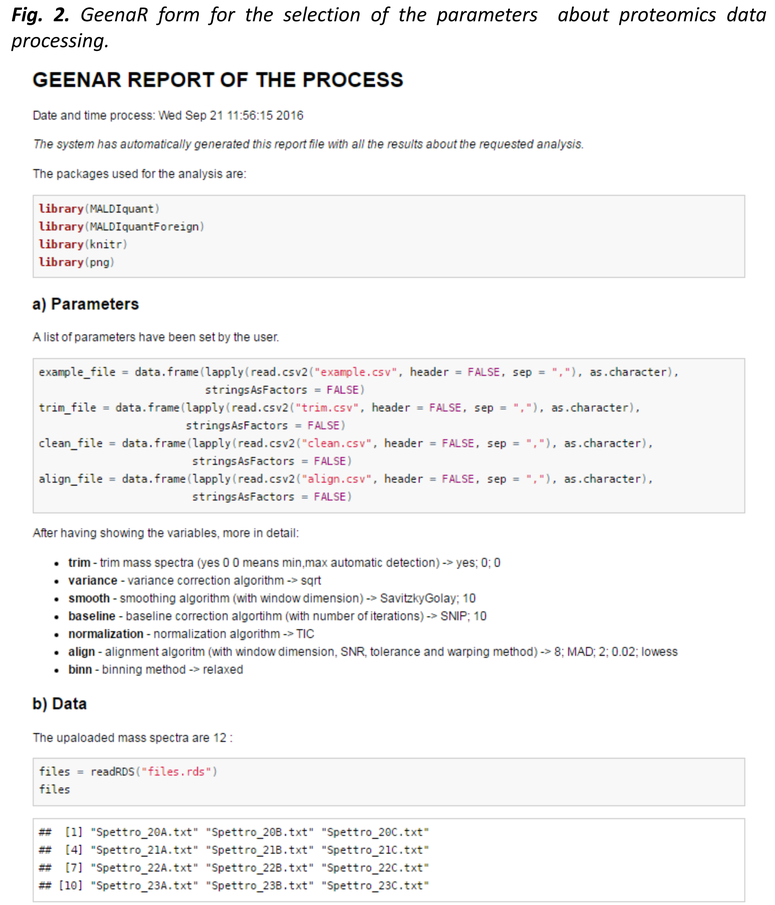


Fig. 2. *GeenaR form for the selection of the parameters about proteomics data processing.*



GEENAR REPORT OF THE PROCESS

Date and time process: Wed Sep 21 11:56:15 2016

The system has automatically generated this report file with all the results about the requested analysis.

The packages used for the analysis are:

```
library(MALDIquant)
library(MALDIquantForeign)
library(knitr)
library(png)
```

a) Parameters

A list of parameters have been set by the user.

```
example_file = data.frame(lapply(read.csv2("example.csv", header = FALSE, sep = ","), as.character),
                          stringsAsFactors = FALSE)
trim_file = data.frame(lapply(read.csv2("trim.csv", header = FALSE, sep = ","), as.character),
                       stringsAsFactors = FALSE)
clean_file = data.frame(lapply(read.csv2("clean.csv", header = FALSE, sep = ","), as.character),
                        stringsAsFactors = FALSE)
align_file = data.frame(lapply(read.csv2("align.csv", header = FALSE, sep = ","), as.character),
                        stringsAsFactors = FALSE)
```

After having showing the variables, more in detail:

- **trim** - trim mass spectra (yes 0 0 means min,max automatic detection) -> yes; 0; 0
- **variance** - variance correction algorithm -> sqrt
- **smooth** - smoothing algorithm (with window dimension) -> SavitzkyGolay; 10
- **baseline** - baseline correction algortihm (with number of iterations) -> SNIP; 10
- **normalization** - normalization algorithm -> TIC
- **align** - alignment algoritm (with window dimension, SNR, tolerance and warping method) -> 8; MAD; 2; 0.02; lowess
- **binn** - binning method -> relaxed

b) Data

The upaloaded mass spectra are 12 :

```
files = readRDS("files.rds")
files
```

```
##  [1] "Spettro_20A.txt" "Spettro_20B.txt" "Spettro_20C.txt"
##  [4] "Spettro_21A.txt" "Spettro_21B.txt" "Spettro_21C.txt"
##  [7] "Spettro_22A.txt" "Spettro_22B.txt" "Spettro_22C.txt"
## [10] "Spettro_23A.txt" "Spettro_23B.txt" "Spettro_23C.txt"
```

Fig. 3. *GeenaR automatic report with the information requested by the user. The a) section is generated in accordance with the choices made in the form shown in Fig. 1.*

Discussions

It is important to underline strongly that **reproducible research is not an optional**, but a fundamental component of a good computational practice, which becomes essential in computational biology. The possibility to reproduce exactly an experiment, from the beginning to the end, improves the robustness of results and it leaves a trail about how a particular result can be produced, with a view to **simplify the knowledge transfer**, even among researchers with different backgrounds.

Supplementary Materials

An example R script and the associated MD and HTML reports can be downloaded from: ***http://bioinformatics.hsanmartino.it/geenar/docs/nettab2016_suppl_materials.zip***

References

[1] Peng RD. Reproducible research in computational science. *Science*, 334:6060, 1226-1227, 2011.
[2] Romano P, Profumo A, Rocco M, et al. Geena 2, improved automated analysis of MALDI/TOF mass spectra. *BMC Bionformatics*, 17(Suppl 4):61, 2016.
[3] Del Prete E, Facchiano A, Profumo A, et al. GeenaR: a flexible approach to pre-process, analyse and compare MALDI-ToF mass spectra. *Conference Proceedings, V Congresso Gruppo Nazionale di Bioingegneria*, 2016.
[4] Russo F, Righelli D, Angelini C. Advantages and limits in the adoption of reproducible research and R-tools for the analysis of omic data. *Computational Intelligence Methods for Bioinformatics and Biostatistics, Lecture Notes in Bioinformatics*, 9874, 245-258, 2016.
[5] Xie Y. Dynamic Documents with R and knitr. *The R Series*. CRC Press, 29, 2015.

Contacts

Email: eugenio.delprete@{isa.cnr.it, unibas.it}, angelo.facchiano@isa.cnr.it, c.angelini@na.iac.cnr.it, {aldo.profumo, paolo.romano}@hsanmartino.it