GUI for the msmsEDA package

Label-free SpC LC-MS/MS exploratory data analysis

Josep Gregori, Alex Sanchez, and Josep Villanueva
Vall Hebron Institute of Oncology &
Statistics Dept. Barcelona University
josep.gregori@gmail.com

January 23, 2014

1 Introduction

Label-free proteomic analysis provides a flexible alternative to the more labour intensive labelled experiments, where all samples to be compared are early mixed. In this setting each sample is processed and analysed separately. The counterpart is high risk of bias due to non controlled factors affecting differently one condition than the other [1, 2, 3, 4]. This requires carefully planned and designed experiments, to minimize confounding and bias as much as possible. Nevertheless, given the sample sizes attainable, not all factors potentially contributing to the results may be controlled. And this brings almost unavoidably to the expression of batch effects, mainly when the period of collecting and analysing the different samples spans a long period of time [5, 6].

Label-free experiments in all '*omics*' benefit of an exploratory data analysis (EDA) to: i) assess the quality of the collected data for each sample, ii) explore the putative influence of confounding factors, and iii) identify eventual outliers to be excluded. The result of a careful EDA informs about the potential of the collected data and gives clues on how to better exploit it.

The graphical user interface (GUI) described in this document uses the functions in the freely available Bioconductor R package `msmsEDA` [7] for exploratory data analysis of label-free protein differential expression experiments with spectral counts (SpC) [1, 8].

This document does not describe the methods but the uses of the GUI, and some background on label-free SpC differential proteomics is supposed. For further help see the `msmsEDA` manual and vignettes at

http://www.bioconductor.org/packages/release/bioc/html/msmsEDA.html.

A companion document "*msmsEDA and msmsTests: R/Bioconductor packages for spectral count label-free proteomics data analysis*" by the same authors, is an extra source of help. The document describes the context and the challenges in proteomics biomarker discovery, and the use of the functions in these packages to solve them. A set of results illustrate the advantages in the use of blocking factors when batch effects are evidenced, and in the use of the post-test filter to improve the reproducibility of DEP lists.

2 The msmsEDA based GUI

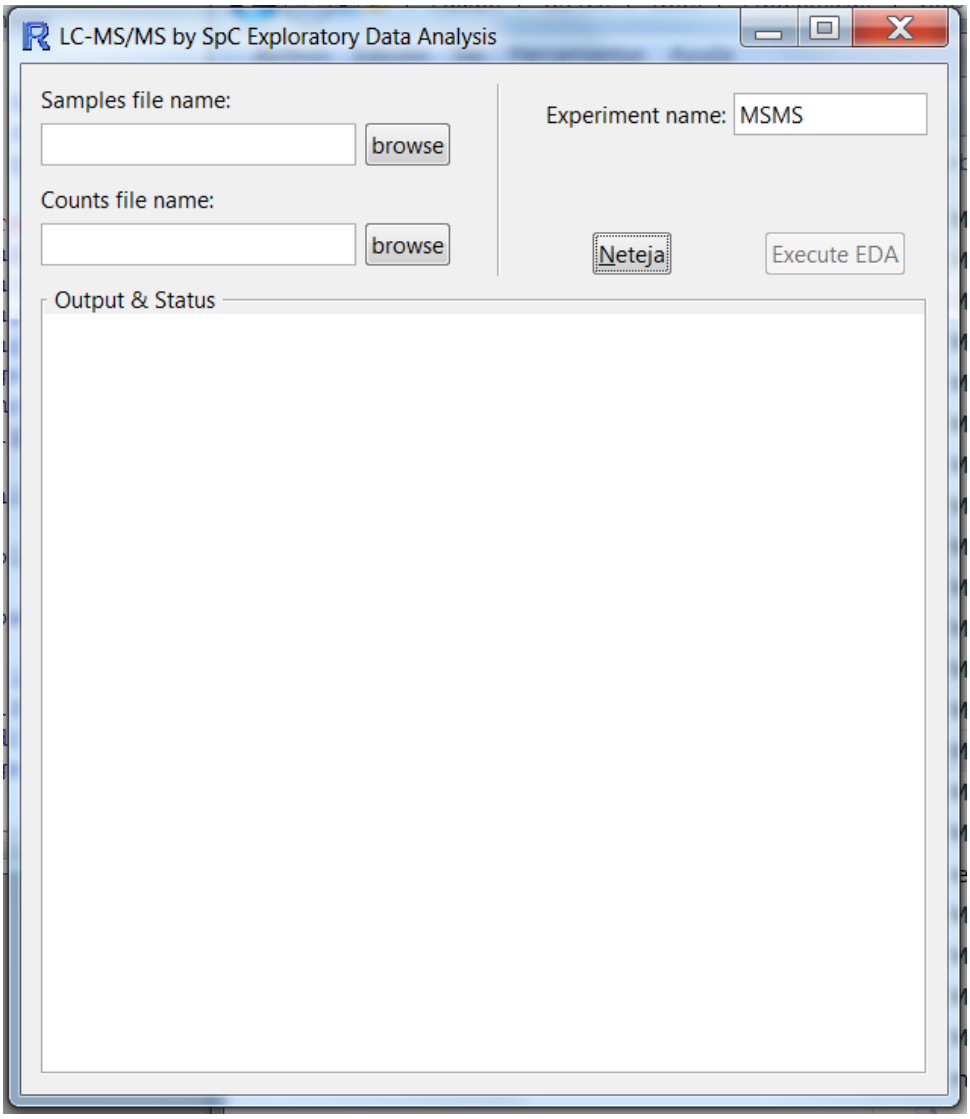

Figure 1: The graphical user interface.

Figure 1 shows the appearance of the `msmsEDA` GUI when just started its execution. Just four actions are required to obtain the results of an EDA:

1. **Give an experiment name**

   This name will be used as root for the different file names with figures, tables and reports produced with the results.

2. **Select a file with the samples description**

   This file must be a tab delimited text file with different columns and a header. The header of the first column must be 'Samples' and give the IDs of each sample. Subsequent columns will be interpreted as factors with the corresponding levels of each sample. The header of each factor column is the factor name. An extra column with header 'offsets' may be used and will be interpreted as normalization factors.

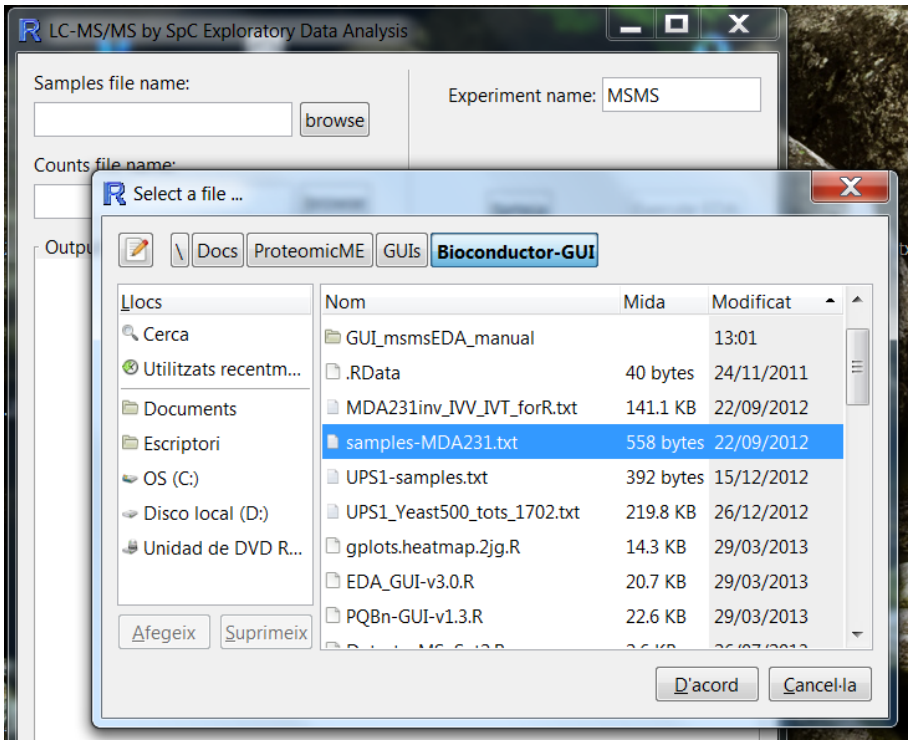

Figure 2: Select a samples description file.

Once selected, the file will be read and its contents displayed for the user to check it.

3. **Select a file with the SpC table, and proteins description.**

   This file must be a tab delimited text file with a header. A column with header 'Accession' with the proteins IDs is required, and will be used to name the proteins in the results. Each sample in the samples description table must have a column in this file, whose header must be the same ID. Extra columns in the SpC table are harmless and will be ignored. An optional column with header 'Proteins' will be interpreted as a textual description of each protein, including a field of the form "GN=[A-Z0-9_]*" from where the gene name will be taken.

The column names in the SpC table file will be displayed in the GUI output control for the user to check.

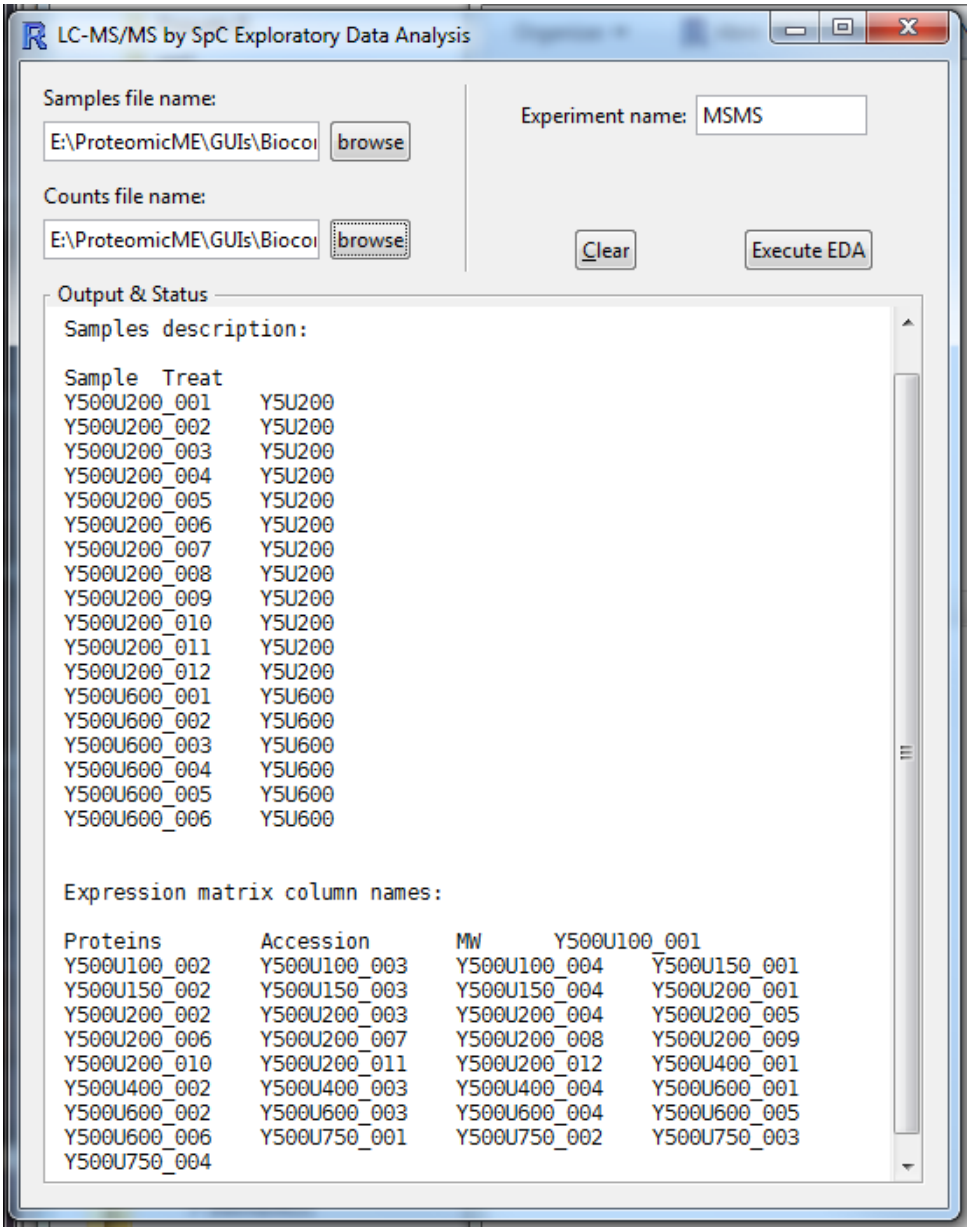

Figure 3: Output after selecting the samples description file, and the SpC matrix file.

4. **Check everything and click on the 'Exec' button**

   When both the samples description table and the SpC table have been loaded, the 'Exec' button becomes enabled because all relevant information is already available. At this point everything is ready and after rechecking all info we may click on the 'Exec' button to start the computations.

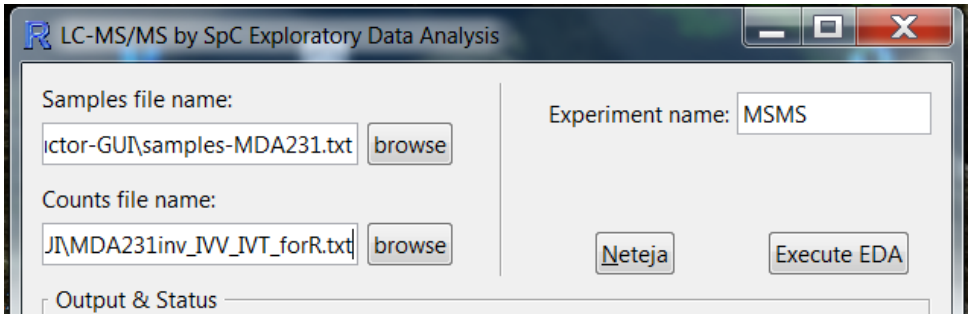

Figure 4: Click on the execute button to start the computation.

The first step in the computation is checking that all parameters conform. If they don't an error message is given and the computation is stopped.

3 The results

The development of the computation may be followed in the output control of the GUI, where messages are sent at the completion of each step.

A number of files are produced with the results. These files have names starting by the experiment name and followed by descriptive names and extensions. An HTML web page is produced as an index to the files with the EDA results, with a short description of each of them. This file is dubbed *exp_name.html*, where exp_name is the name given in the 'Experiment name' edit control, and contains links to each of the generated files.

- **A text file with statistic summaries**

  A text file of name ending with '_EDA.txt' with statistic summaries:

  - A summary of SpC distribution statistics by sample, at each step: raw, normalized, and batch corrected.
  - The contribution of the first four PC to the total variance, at each step: raw, normalized, and batch corrected.
  - The sample size normalization factors scaled to the median.
  - The sample Bio normalization factors scaled to the mean, when given.
  - The sample size + Bio normalization factors scaled to the mean, when given.
  - With the final dataset, a summary of distribution statistics of the residual dispersion.

- **Normalizing divisors barplots**

  A pdf file with a barplot showing the relative size of the normalizing divisors by sample. Multiples pdf files may be generated.

  - The sample size divisors in a pdf file of name ending with "-SizeDivsBarplots.pdf".

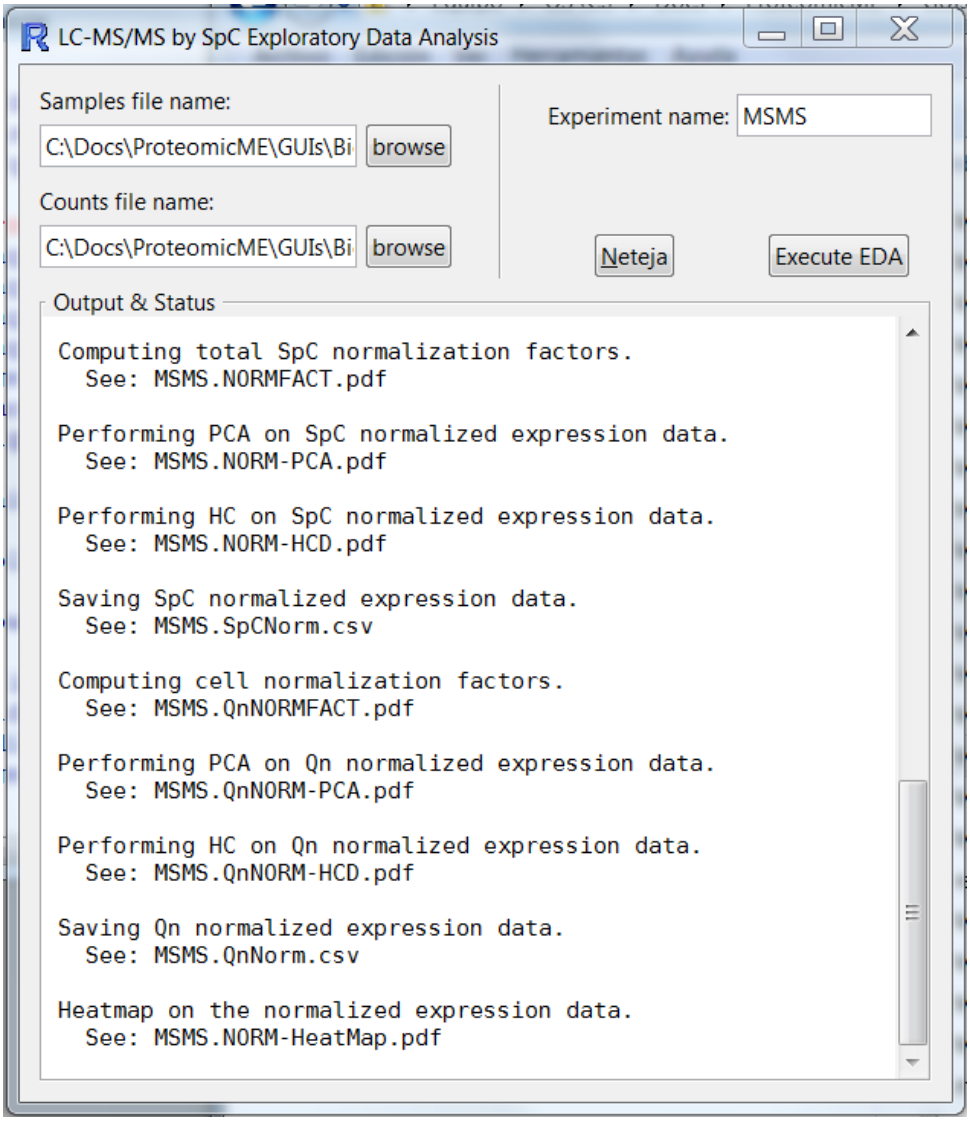

Figure 5: Output control showing the development of the EDA

- The divisors provided in the 'offsets' column in the samples description file, in a pdf file of name ending with "-BioDivsBarplots.pdf"
- The divisors obtained as the product of the sample sizes and the 'offsets' in the samples description file, in a pdf file of name ending with "-BioSzDivsBarplots.pdf"

- **A tab delimited file with the normalized SpC matrix**

  A csv tab delimited file with the SpC matrix resulting of a normalization. Multiples files may be generated depending of the settings.

  - A sample size normalized SpC file of name ending with "-SizeNorm.csv" when no offsets are given in the samples description file.
  - A biological plus sample size normalized SpC file of name ending with "-BioSzNorm.csv" when offsets are provided in the samples description file.

msmsEDA Graphical User Interface

EXPLORATORY DATA ANALYSIS RESULTS - Output files index

FILE / LINK	Contents
MDA231_EDA.txt	**Text file with statistic summaries at each step**
MDA231-DispPlots.pdf	PDF file with plots Final SpC matrix **Residual dispersion plots**
MDA231-BatchCorrected-PCA.pdf	PDF file with plots Normalized + batch corrected SpC matrix **Principal components plot**
MDA231-BatchCorrected-HCD.pdf	PDF file with plots Normalized + batch corrected SpC matrix **Hierarchical clustering plot**
MDA231-BatchCorrected-HeatMap.pdf	PDF file with plots Normalized + batch corrected SpC matrix **Heatmap**

Figure 6: HTML page with the list of EDA generated files

  - A normalized and batch corrected SpC file of name ending with "-BatchCorrected.csv", when a second factor is given in the samples description file, which is interpreted as a blocking factor.

- **Principal Components scatterplots**

  A pdf file with a scatterplot on the first two PC with samples colored as per treatment factor level. Multiples pdf files may be generated.

  - Using the raw SpC matrix, a pdf file of name ending with "-RAW-PCA.pdf".
  - Using the normalized SpC matrix, a pdf file of name ending with "-SizeDivs-PCA.pdf" or "-BioSzNorm-PCA.pdf".
  - Using the batch corrected SpC matrix, a pdf file of name ending with "-BatchCorrected-PCA.pdf".

- **Hierarchical clustering dendrograms**

  A pdf file with a dendrogram with branches colored as per treatment factor level. Multiples pdf files may be generated.

  - Using the raw SpC matrix, a pdf file of name ending with "-RAW-HCD.pdf".
  - Using the normalized SpC matrix, a pdf file of name ending with "-SizeDivs-HCD.pdf" or "-BioSzNorm-HCD.pdf".
  - Using the batch corrected SpC matrix, a pdf file of name ending with "-BatchCorrected-HCD.pdf".

- **Heatmaps**

  Two pdf files, one with a heatmap in an A4 page, and an expanded heatmap with 3mm tall rows allowing to identify all features, and with the SpC superposed on each cell. Again, multiples pdf files may be generated.

  - Using the raw SpC matrix, two pdf files of names ending with "-RAW-HeatMap.pdf" and "-RAW-ExpandedHeatmap.pdf".
  - Using the normalized SpC matrix, two pdf files of names ending with "-SizeDivs-HeatMap.pdf" and "-SizeDivs-ExpandedHeatmap.pdf", or "-BioSzNorm-HeatMap.pdf" and "-BioSzNorm-ExpandedHeatmap.pdf".
  - Using the batch corrected SpC matrix, two pdf files of names ending with "-BatchCorrected-HeatMap.pdf" and "-BatchCorrected-ExpandedHeatmap.pdf".

- **Informative features scatterplot**

  A pdf file with a scatterplot of the means of each of the two levels of the treatment factor, in a control vs treatment plot. Multiples pdf files may be generated.

  - Using the raw SpC matrix, a pdf file of name ending with "-RAW-factor-Scatterplot.pdf".
  - Using the normalized SpC matrix, a pdf file of name ending with "-SizeDivs-factor-Scatterplot.pdf" or "-BioSzNorm-factor-Scatterplot.pdf".
  - Using the batch corrected SpC matrix, a pdf file of name ending with "-BatchCorrected-factor-Scatterplot.pdf".

- **Plots of residual dispersion**

  A pdf file of name ending with "-DispPlots" with a density plot of the residual dispersion values of fitting a linear model with the provided two level treatment factor. And a scatterplot of means vs residual variances.

4 Again

The clicking on the usual icon on the top right of the window will terminate the GUI. In case that a new computation is required an option is to click on the 'Clear' button which will reset all variables. Then start anew by selecting a samples description file and a SpC matrix file.

References

[1] Neilson K.A., et al,. *Less label, more free: approaches in label-free quantitative mass spectrometry.* Proteomics, 2011, **11**, 535-53.

[2] Sandin, M. et al., *Generic workflow for quality assessment of quantitative label-free LC-MS analysis.* Proteomics, 2011, **11**, 1114-1124.

[3] Zhu, W. et al., *Mass spectrometry-based label-free quantitative proteomics*. J. Biomed. Biotechnol., 2010, **2010**, Article ID 840518, 6 pages http://dx.doi.org/10.1155/2010/840518

[4] Patel, V. J., et al., *A comparison of labeling and label-free mass spectrometry-based proteomics approaches*. J. Proteome Res., 2009, **8**, 3752-3759.

[5] Scherer, A. (Ed.), *Batch Effects and Noise in Microarray Experiments: Sources and Solutions*. Wiley Series in Probability and Statistics, John Wiley & Sons, 2009

[6] Gregori J., et al., *Batch effects correction improves the sensitivity of significance tests in spectral counting-based comparative discovery proteomics*, Journal of Proteomics, 2012, **75**, 3938-3951

[7] Gregori, J. et al., *msmsEDA: LC-MS/MS Exploratory Data Analysis*.
R package version 1.1.1.
http://www.bioconductor.org/packages/release/bioc/html/msmsEDA.html

[8] Mallick P., & Kuster B., *Proteomics: a pragmatic perspective.* Nat. Biotechnol., 2010, **28**, 695-709.