Metabolomics manuscript No.
(will be inserted by the editor)

NormalizeMets: Assessing, selecting and implementing statistical methods for normalizing metabolomics data

Alysha M De Livera [1*†] · **Gavriel Olshansky** [2†] · **Julie A Simpson** [1#] · **Darren J Creek** [3#]

Received: date / Accepted: date

Abstract .

Introduction: In metabolomics studies, unwanted variation inevitably arises from various sources. *Normalization*, that is the removal of unwanted variation, is an essential step in the statistical analysis of metabolomics data. However, metabolomics normalization is often considered an imprecise science due to the diverse sources of variation and the availability of a number of alternative strategies that may be implemented.

Objectives: We highlight the need for comparative evaluation of different normalization methods and present software strategies to help ease this task for both data-oriented and biological researchers.

Methods: We present *NormalizeMets*- a joint graphical user interface within the familiar Microsoft Excel and freely-available R software for comparative evaluation of different normalization methods. The NormalizeMets R package along with the vignette describing the workflow can be downloaded from `https://cran.r-project.org/web/packages/NormalizeMets`. The Excel Interface and the Excel user guide are available on `https://metabolomicstats.github.io/ExNormalizeMets`.

Results: *NormalizeMets* allows for comparative evaluation of normalization methods using criteria that depend on the given dataset and the ultimate research question. Hence it guides researchers to assess, select and implement a suitable normalization method using either the familiar Microsoft Excel and/or freely-available R software. In addition, the package can be used for visualisation of metabolomics data using interactive graphical displays and to obtain end statistical results for clustering, classification, biomarker identification adjusting for confounding variables, and correlation analysis.

Conclusion: *NormalizeMets* is designed for comparative evaluation of normalization methods, and can also be used to obtain end statistical results. The use of freely-available R software offers an attractive proposition for programming-oriented researchers, and the Excel interface offers a familiar alternative to most biological researchers. The package handles the data locally in the user's own computer allowing for reproducible code to be stored locally.

1 Introduction

Every metabolomics experiment is subject to a component of unwanted variation inevitably arising from many potential sources. These include both experimental and biological sources (Sysi-Aho et al. 2007) such as sample preparation and storage, analysis of multiple batches, inter-instrument and inter-laboratory variation, and confounding biological variation due to the constitution of the biological samples (e.g., differing cell sizes, sample weight or volume), some of which are not easily measurable. The removal of this unwanted variation (referred to as *normalization*) thus forms an integral part of the statistical analysis of metabolomics data, and is required to alleviate problems of identifying false biomarkers, missing out on true biomarkers, artificial classification or clustering of the samples or metabolites (Gagnon-Bartsch et al. 2013).

During the last decade, various normalization methods have been used for normalizing metabolomics data. These approaches vary in terms of applicability and offer distinct strengths and weaknesses for different metabolomics experimental settings. As the results obtained from the statistical analysis of metabolomics data often depend on the normalization approach employed, it is vital to choose the optimal normalization method given the experimental design, dataset in hand and the research question of interest (De Livera et al. 2015). This process can be cumbersome and is often considered a 'grey area' (Roessner et al. 2011). The need for comparative evaluation of

[1] Biostatistics Unit, Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, VIC 3800, Australia [2] Department of Mathematics & Statistics, The University of Melbourne, VIC 3800, Australia [3] Drug Delivery, Disposition and Dynamics, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, VIC 3052, Australia, [†,#] Equal contribution.

normalization methods has been well-acknowledged in a recently published paper (Li et al. 2017), which introduces NOREVA- a web-based software with a convenient graphical user interface. NOREVA includes over twenty normalization methods, and allows for visualisation and comparison between different normalization methods. Here, we present a joint graphical user interface within Microsoft Excel and R software which handles the data locally in the user's own computer and allows comparison of over twelve widely-used normalization methods. While the use of freely-available R software for this task is an attractive proposition for programming-oriented researchers, the Excel interface offers a familiar alternative to most biological researchers. NormalizeMets allows for reproducible code to be stored locally either as R scripts or VBA scripts. The package can also be used for visualisation of metabolomics data using interactive graphical displays and to obtain end statistical results for clustering, classification, biomarker identification adjusting for confounding variables, and correlation analysis.

2 Description

2.1 Requirements and availability

R software (version 3.4.3 or higher) can be downloaded freely from `https://cran.r-project.org/`. For Excel users, Microsoft Excel (2016) is required on Windows 7 or higher. The NormalizeMets R package along with the vignette describing the workflow can be downloaded from `https://cran.rproject.org/web/packages/NormalizeMets`. The Excel Interface and the Excel user guide are available on `https://metabolomicstats.github.io/ExNormalizeMets`.

2.2 Getting started

The NormalizeMets workflow is described in detail in the NormalizeMets Vignette, which can be assessed by typing `browseVignettes("NormalizeMets")` in R or simply clicking on `Manual` button in Excel.

2.3 Input data format

The input data format consists of three parts: (i) `featuredata` which is the metabolomics data matrix containing all metabolite peak intensities (or concentrations). Unique sample names must be provided as row names and unique metabolite names as column names, (ii) `metabolitedata` contains metabolite-specific information in a separate dataframe. These information can include, but is not limited to, designation of metabolites as internal/external standards, or positive/negative controls. Metabolite names need to be provided as row names, and (iii) `sampledata` is a dataframe that contains sample-specific information. These information can include sample type, order of analysis, factors of interest and other sample-specific data relevant to the analysis. Unique sample names need to be provided as row names.

2.4 Methods

The package allows for initial processing of the data, such as log transforming, handling missing values using most popular methods (e.g., the k-th nearest neighbour algorithm (`knn`), replacing by half the minimum (`replace`)), and visualization using interactive graphical displays.

Over twelve normalization methods are presented in this package and these are divided into four categories, as those which use (i) internal, external standards and other quality control metabolites (`NormQcmets` function) (ii) quality control samples analysed periodically throughout the batch (`NormQcsamples` function), (iii) scaling methods (`NormScaling` function), and (iv) combined methods (`NormCombined` function).

The `NormQcmets` function includes the `is` method which uses a single standard (Gullberg et al. 2004), the `ccmn` (cross contribution compensating multiple internal standard) method (Redestig et al. 2009), the `nomis` (normalization using optimal selection of multiple internal standards) method (Sysi-Aho et al. 2007), and the remove unwanted variation methods (Gagnon-Bartsch et al. 2013) applied to metabolomics using `ruv2` (De Livera et al. 2012), `ruvrand` and `ruvrandclust` (De Livera et al. 2015). The `NormQcsamples` function implements the `rlsc` (quality control sample based robust locally estimated scatterplot smoothing) signal correction method described by (Dunn et al. 2011). The scaling normalization methods (Scholz et al. 2004, Wang et al. 2003) included in the package are normalization to a total sum (`sum`), median (`median`) or mean (`mean`) peak intensity (or concentration) of each sample, and the method `ref` normalizes the metabolite abundances to a specific reference vector such as the sample weight or volume. As some of the metabolomics normalization methods are unable to accommodate the overall unwanted variation component alone, a combination of normalization methods are sometimes employed in practice. The function `NormCombined` allows for this. By default, the function performs the `rlsc` signal correction method followed by the `median` scaling normalization method.

The criteria for assessing and selecting a normalization method have been described in detail by (De Livera et al. 2012, 2015) and (Gagnon-Bartsch et al. 2013). Four visualization approaches are available in the package:

(i) interactive volcano plots (`CompareVolcanoPlots`) for exploring the impact of the normalization methods on positive and negative control metabolites for identifying biomarkers, (ii) interactive relative log abundance plots (`CompareRlaPlots`) and principal component plots (`ComparePcaPlots`) for examining the unwanted variation component removed by the normalization methods, the normalized datasets or the residuals obtained from a fitted linear model designed for identifying biomarkers, (iii) histograms (`CompareHist`) for exploring the distribution of p-values obtained for biomarker identification or correlation analysis, or comparing correlation densities, (iv) clustering (`Dendrogram`) and classification (`SvmFit`) accuracies of the samples with known grouping structure, and (v) consistency of results from multiple analytical platforms (`VennPlot`).

In addition, the package can be used to obtain end statistical results for clustering, classification, biomarker identification with adjustment for confounders, and correlation analysis.

3 Example Analyses

Four example datasets are included in the package for demonstration purposes: `mixdata` (Redestig et al. 2009), `Didata` (Kirwan et al. 2014), `UVdata` (De Livera et al. 2012) and `alldata-eg` (De Livera et al. 2015). The analyses using these datasets are demonstrated in the `NormalizeMets` Vignette. See Figure 1 and Figure 2 for example screenshots.

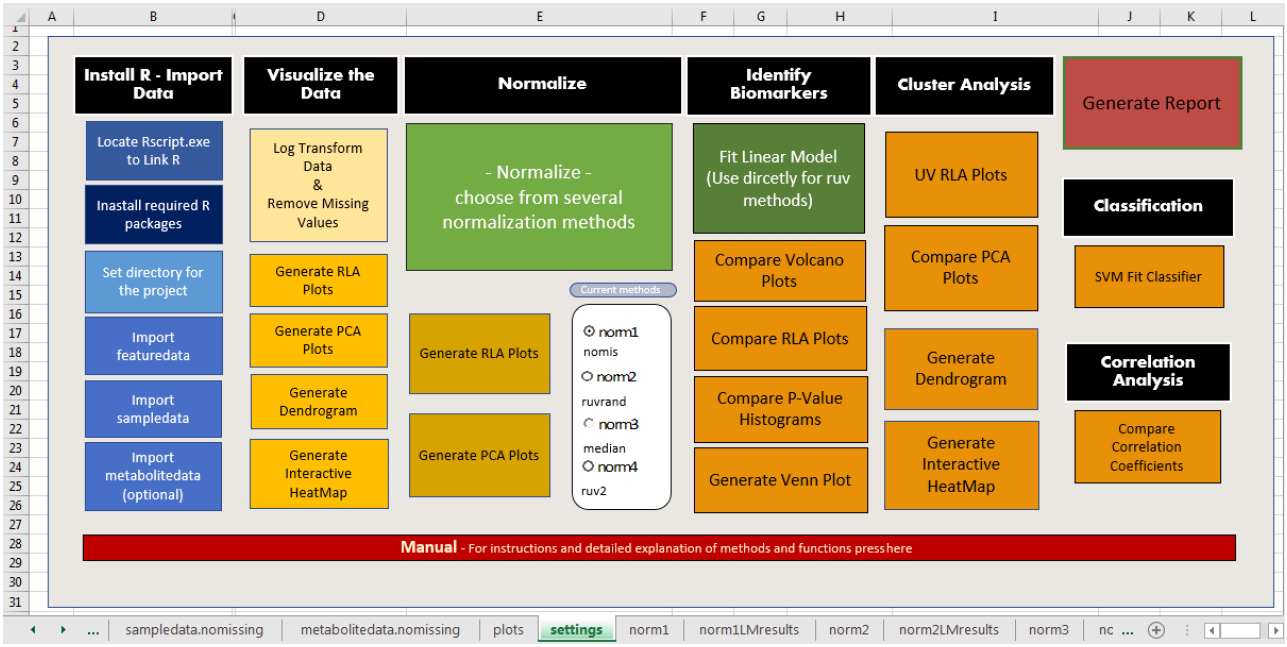

Fig. 1 Excel graphical user interface for the NormalizeMets package displaying only some of the included normalization methods.

4 Conclusion

With the increasing popularity of large-scale metabolomics studies, assessing and choosing a suitable normalization method is becoming ever more important. The selection of the method should depend on the experimental design, dataset in hand and the research question of interest. `NormalizeMets` builds on several useful software packages to help ease this task for both data-oriented and biological researchers, and complements the existing user-friendly metabolomics tools such as `NOREVA` (Li et al. 2017), `MetaboAnalyst` (Xia et al. 2012) and `IDEOM` (Creek et al. 2012). Due to the capabilities of the NormalizeMets R package, it will also serve as an updated version of both the 'metabolomics' package for R (De Livera et al. 2012) and the 'MetNorm' package for R (De Livera et al. 2015) which are two of the software packages currently being used by the general metabolomics community (Spicer et al. 2017) as well as in the IDEOM package (Creek et al. 2012) for downstream statistical analyses.

Acknowledgements

Professor Terry Speed, Walter and Eliza Hall Institute of Medical Research

Fig. 2 (a) Comparison for identifying biomarkers associated with age in a population cohort study. (a) (i) The volcano plots are produced by the CompareVolcanoPlots function. (a) (ii) The histograms are produced by the CompareHist function; (b) The correlation coefficient histograms produced by the CompreHist function for correlation analysis of the same population cohort; (c) Comparison for cluster analysis in a designed dataset. (c) (i) The Rla plots are produced by the CompareRlaPlots function. (c) (ii) The principal components analysis plots are produced by the ComparePcaPlots function.

Funding

Julie A Simpson is supported by a National Health and Medical Research Council (NHMRC) Senior Research Fellowship (1104975). Alysha M De Livera is supported by The University of Melbourne Research Fellowship. Darren J Creek is supported by a National Health and Medical Research Council (NHMRC) Career Development Research Fellowship (1088855).

Conflict of Interest

Authors have no conflict of interest to declare.

References

Creek, D. J., Jankevics, A., Burgess, K. E. V., Breitling, R. & Barrett, M. P. (2012), 'IDEOM: an Excel interface for analysis of LC-MS-based metabolomics data.', *Bioinformatics (Oxford, England)* **28**(7), 1048–9.

De Livera, A. M., Aho-Sysi, M., Jacob, L., Gagnon-Bartch, J., Castillo, S., Simpson, J. & Speed, T. P. (2015), 'Statistical methods for handling unwanted variation in metabolomics data', *Analytical chemistry* **87**(7), 3606–3615.

De Livera, A. M., Dias, D. A., De Souza, D., Rupasinghe, T., Pyke, J., Tull, D., Roessner, U., McConville, M. & Speed, T. P. (2012), 'Normalizing and integrating metabolomics data.', *Analytical chemistry* **84**(24), 10768–76.

Dunn, W. B., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J. D., Halsall, A., Haselden, J. N., Nicholls, A. W., Wilson, I. D., Kell, D. B. & Goodacre, R. (2011), 'Procedures

for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry.', *Nature protocols* **6**(7), 1060–1083.

Gagnon-Bartsch, J. A., Jacob, L. & Speed, T. P. (2013), 'Removing unwanted variation from high dimensional data with negative controls', *Berkeley: Tech Reports from Dep Stat Univ California* pp. 1–112.

Gullberg, J., Jonsson, P., Nordström, A., Sjöström, M. & Moritz, T. (2004), 'Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry.', *Analytical biochemistry* **331**(2), 283–95.

Kirwan, J., Weber, R., Broadhurst, D. & Viant, M. (2014), 'Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control', *Scientific Data* **1**, 1–13.

Li, B., Tang, J., Yang, Q., Li, S., Cui, X., Li, Y., Chen, Y., Xue, W., Li, X. & Zhu, F. (2017), 'Noreva: normalization and evaluation of ms-based metabolomics data', *Nucleic Acids Research* pp. W162–W170.

Redestig, H., Fukushima, A., Stenlund, H., Moritz, T., Arita, M., Saito, K. & Kusano, M. (2009), 'Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data.', *Analytical chemistry* **81**(19), 7974–7980.

Roessner, U., Nahid, A., Chapman, B., Hunter, A. & Bellgard, M. (2011), *Metabolomics- The Combination of Analytical Biochemistry, Biology, and Informatics*, Vol. 1, second edn, Elsevier B.V.

Scholz, M., Gatzek, S., Sterling, A., Fiehn, O. & Selbig, J. (2004), 'Metabolite fingerprinting: detecting biological features by independent component analysis.', *Bioinformatics (Oxford, England)* **20**(15), 2447–54.

Spicer, R., Salek, R. M., Moreno, P., Cañueto, D. & Steinbeck, C. (2017), 'Navigating freely-available software tools for metabolomics analysis', *Metabolomics* **13**(9), 106.

Sysi-Aho, M., Katajamaa, M., Laxman, Y. & Oresic, M. (2007), 'Normalization method for metabolomics data using optimal selection of multiple internal standards.', *BMC bioinformatics* **8**, 93.

Wang, W., Zhou, H., Lin, H., Roy, S., Shaler, T. A., Hill, L. R., Norton, S., Kumar, P., Anderle, M. & Becker, C. H. (2003), 'Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards.', *Analytical chemistry* **75**(18), 481848–26.

Xia, J., Mandal, R., Sinelnikov, I. V., Broadhurst, D. & Wishart, D. S. (2012), 'MetaboAnalyst 2.0–a comprehensive server for metabolomic data analysis', *Nucleic Acids Research* pp. 1–7.