BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 1 2014, pages 133–134
doi:10.1093/bioinformatics/btt589

Systems biology Advance Access publication October 29, 2013

MSPrep—Summarization, normalization and diagnostics for processing of mass spectrometry–based metabolomic data

Grant Hughes[1,*], Charmion Cruickshank-Quinn[2], Richard Reisdorph[2], Sharon Lutz[1], Irina Petrache[3], Nichole Reisdorph[2], Russell Bowler[4] and Katerina Kechris[1]

[1]Department of Biostatistics and Informatics, University of Colorado School of Public Health, Aurora, CO, [2]Department of Immunology, National Jewish Health Center, Denver, CO, [3]Department of Medicine, Indiana University School of Medicine, Indianapolis, IN and [4]Department of Pulmonary Medicine, National Jewish Health Center, Denver, CO, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Although R packages exist for the pre-processing of metabolomic data, they currently do not incorporate additional analysis steps of summarization, filtering and normalization of aligned data. We developed the MSPrep R package to complement other packages by providing these additional steps, implementing a selection of popular normalization algorithms and generating diagnostics to help guide investigators in their analyses.

Availability: http://www.sourceforge.net/projects/msprep

Contact: grant.hughes@ucdenver.edu

Supplementary Information: Supplementary materials are available at *Bioinformatics* online.

Received on August 22, 2013; revised on October 7, 2013; accepted on October 8, 2013

1 INTRODUCTION

The primary goal of the MSPrep package is to prepare metabolomics data for advanced statistical analysis by automating the processing of datasets and generating diagnostic graphs. The initial processing of Liquid Chromatography coupled with Mass Spectrometry (LCMS) data is covered by a variety of software packages provided by instrument manufacturers and a number of open source packages such as xMSAnalyzer (Uppal *et al.*, 2013), XCMS (Smith *et al.*, 2006) and MzMine (Pluskal *et al.*, 2010). While these manage the initial data pre-processing steps of peak detection, chromatogram building, alignment and quantification, they often lack functions for further processing. We designed the MSPrep package to complement existing software by providing additional processing tools and statistical and graphical tools for evaluation of different methods. As there are no universally accepted procedures, the package provides implementation of a variety of novel and previously published methods. The primary functions of the MSPrep package are summarization of replicates, filtering, imputation of missing data, normalization and/or batch effect adjustment and dataset diagnostics.

2 MATERIALS AND METHODS

Input: Three files are required for MSPrep: Aligned LCMS abundance/intensity data, a dataset linking subject ID to LCMS run and a clinical dataset containing unique subject identifiers and any outcomes, phenotypes and batch number for normalization and PCA purposes.

Summarization/Averaging: The first processing step is summarization of technical replicates, three replicates required per subject/sample. MSPrep provides options to remove erroneous data and to reduce the effect of extreme observations. The user specifies a cutoff for the coefficient of variation (CV), calculated by dividing the standard deviation of the replicates by the average, yielding a measure for magnitude of the variation between replicates. The summarization routine summarizes each compound by subject (or sample) and returns a single observation per compound per subject. Only abundances that are found in at least two of three replicates are kept. If CV is below the user-specified level, the average of the replicates is used. If the CV is above the specified level and found in exactly two of three replicates, the summarization is not used and the observation is left blank. If the compound was found in all three replicates but with unacceptable CV, the median is used as the summarization measure. This approach removes potential erroneous data. We have found that most compounds with high CV have two consistent and one extreme observation. Using the median reduces the effect of the extreme observation.

Filtering: The resulting summarized dataset contains all compounds with one observation per subject (or sample). The next processing step filters the data to only compounds found in a user-specified percentage of subjects.

Missing Data: There are three primary modes of missing data in metabolomics datasets and each mode has different implications for subsequent analysis; therefore, different imputation routines and statistical methods are required and three are offered in the MSPrep package. The three modes are truly not present, present below the detectable limit of the instrument and absent owing to error in pre-processing algorithms. The MSPrep package implements three methods of managing missing data: (i) No imputation assumes the mode of missing is true zeros and therefore assigns the missing values as zeros. This dataset could be useful for PCA analysis, cluster analysis and methods that account for clustering at zero. Unless a stringent filter is applied, normalization routines may have poor performance, as most have assumptions about underlying distributions that are not valid with zero clustered data. (ii) The second option assumes missing compounds were below the detectable limit and imputes a value of one half of the minimum observed value for that compound (Xia *et al.*, 2009). (iii) The final method is a call to the Bayesian PCA (BPCA) imputation algorithm (Oba *et al.*, 2003) from the PCAMethods R package (Stacklies *et al.*, 2007) and assumes that the compound is present but failed to be accurately detected. This

*To whom correspondence should be addressed.

algorithm estimates the missing value by a linear combination of principal axis vectors, where the parameters of the model are identified by a Bayesian estimation method and is not sensitive to the quantity of missing data.

Normalization: There are five options for normalization: Median (Wang *et al.*, 2003), Quantile (Bolstad *et al.*, 2003), Cross-Contribution Compensating Multiple Standard Normalization (CRMN) (Redestig *et al.*, 2009), Surrogate Variable Analysis (SVA) (Leek *et al.*, 2007) and Removal of Unwanted Variation (RUV) (Gagnon-Bartsch *et al.* 2012). Median, quantile and CRMN all result in an adjusted (normalized) dataset. RUV and SVA each estimate a matrix of unobserved factors of importance using different methods of supervised factor analysis. The unnormalized, median and quantile adjusted data are adjusted with ComBat (Johnson *et al.*, 2007), an empirical Bayes batch effect correction algorithm for the removal of potential batch effects. For RUV and CRMN, users can either specify compounds to use as controls or data driven controls will be estimated if no negative controls exist (DeLivera *et al.*, 2012).

Results will vary based on biological and technical differences in experiments. Users are encouraged to try different processing steps and MSPrep produces two documents to assist in comparing the methods. Examples of these are available online in the Supplementary Materials. The first contains distributional histograms of raw and log transformed abundances by all three imputation methods. The second can be used to compare normalization methods and contains color-coded PCA plots for a categorical phenotype or covariate and numbering for batches, box plots by subject (or sample, cell type, etc), box plots by batch and box plots by the phenotype or covariate. Examples of both output files are in the Supplementary Material.

3 RESULTS

An operator difference dataset was generated by the Reisdorph Mass Spectrometry laboratory at National Jewish Health. Three technicians performed all steps of sample prep for profiling of a base human plasma sample containing six spiked in control compounds at concentrations of 1X, 2X and 4X and two negative controls at 1X in all samples. Pre-processing was performed in Agilent's Mass Hunter software.

For summarization, the CV cutoff was set at 0.50. There were 23 compounds (0.1% of compounds) that were present in all replicates but above this cutoff. Of these, all but one fit the pattern of two consistent observations and one outlier. The data were filtered at 80% present in all subjects, which resulted in 891 compounds. The BPCA imputation was used to maintain a Gaussian distribution and allows application and evaluation of all normalization methods.

Owing to the design, we expect compounds to be consistently measured between subjects with the only detected differences occurring in the six compounds that were spiked into the datasets. Thus, all other differences are due to the batch/operator effect. As shown in Table 1, the SVA normalization method was the most effective at controlling false-positive rate while reducing the operator effect for this dataset, based on a linear model of the log2 abundances. While the operator effect is still apparent in the PCA plot of the SVA adjusted data in Supplementary Figure S2, the lower percentages and RMSE in Table 1 represents an improvement over the other methods.

Table 1. Percentage of compounds found significant by normalization method and RSME for detection of 4× spiked compounds

Method	None	Median	Quant	Quant Combat	SVA	RUV	CRMN
Spike	6.0%	7.4%	5.6%	22.4%	6.6%	12.1%	15.4%
Operator	8.0%	11.4%	10.8%	0.0%	3.7%	6.7%	5.8%
RSME	1.42	1.31	1.34	1.34	1.26	1.42	1.69

4 CONCLUSION

The MSPrep package is designed to help investigators prepare their datasets for analysis, while reducing the amount of manual processing of datasets and generating useful diagnostic aids. Our package complements existing pre-processing methods such as xMSAnalyzer (Uppal *et al*, 2013), and the output is in a format ready for input to leading software such as MetaboAnalyst (Xia *et al.*, 2009) to perform clustering and other downstream analyses. The SVA routine provided the best performance when applied to the dataset presented, but it is important to note the strengths and weaknesses of each routine, and, in our experience, no single method is best in all circumstances. We encourage users to examine the diagnostic plots generated by the package to compare different methods for their datasets.

Funding: NIH/NHLBI P20 HL113445.

Conflict of Interest: none declared.

REFERENCES

Bolstad,B.M. *et al.* (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

DeLivera,A.M. *et al.* (2012) Normalizing and integrating metabolomic data. *Anal. Chem*, **84**, 10768–10776.

Gagnon-Bartsch,J.A. *et al.* (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

Johnson,W.E. *et al.* (2007) Adjusting batch effects in microarray expression data using Empirical Bayes methods. *Biostatistics*, **8**, 118–127.

Leek,J.T. *et al.* (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, **3**, , e161.

Oba,S. *et al.* (2003) A Bayesian missing value estimation for gene expression profile data. *Bioinformatics*, **19**, 2088–2096.

Pluskal,T. *et al.* (2010) MZ Mine 2: Modular framework for processing, visualizing and analyzing mass spectrometry based molecular profile data. *BMC Bioinformatics*, **11**, 395.

Redestig,H. *et al.* (2009) Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal. Chem.*, **81**, 7974–7980.

Smith,C. *et al.* (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal. Chem.*, **78**, 779–787.

Stacklies,W. *et al.* (2007) pcaMethods—a bioconductor package providing PCA methods for incomplete data. *Bioinformatics*, **23**, 1164–1167.

Uppal,K. *et al.* (2013) xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC Bioinformatics*, **14**, 15.

Wang,W. *et al.* (2003) Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.*, **75**, 4818–4826.

Xia,J. *et al.* (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucleic Acids Res*, **37**, W652–W660.