BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 17 2014, pages 2524–2526
doi:10.1093/bioinformatics/btu305

Systems biology

Advance Access publication May 2, 2014

MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments

Meena Choi[1], Ching-Yun Chang[1], Timothy Clough[1], Daniel Broudy[2], Trevor Killeen[2], Brendan MacLean[2] and Olga Vitek[1,3,*]

[1]Department of Statistics, Purdue University, West Lafayette, IN, [2]Department of Genome Sciences, University of Washington, Seattle, WA 98195 and [3]Department of Computer Science, Purdue University, West Lafayette, IN 47907, USA

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: MSstats is an R package for statistical relative quantification of proteins and peptides in mass spectrometry-based proteomics. Version 2.0 of MSstats supports label-free and label-based experimental workflows and data-dependent, targeted and data-independent spectral acquisition. It takes as input identified and quantified spectral peaks, and outputs a list of differentially abundant peptides or proteins, or summaries of peptide or protein relative abundance. MSstats relies on a flexible family of linear mixed models.

Availability and implementation: The code, the documentation and example datasets are available open-source at www.msstats.org under the Artistic-2.0 license. The package can be downloaded from www.msstats.org or from Bioconductor www.bioconductor.org and used in an R command line workflow. The package can also be accessed as an external tool in Skyline (Broudy *et al.*, 2014) and used via graphical user interface.

Contact: ovitek@purdue.edu

Received on September 21, 2013; revised on March 29, 2014; accepted on April 28, 2014

1 INTRODUCTION

Quantitative mass spectrometry-based proteomics is a technology of growing importance in biological and clinical research. Modern proteomic workflows are complex and diverse, and a statistical approach is required to reduce bias and inefficiencies, distinguish the systematic variation from random artifacts and maximize the reproducibility of the results (Käll and Vitek, 2011). Although some of these goals can be achieved with standard statistical methods, or with methods developed for other technologies such as gene expression microarrays, this is typically insufficient. Specialized methods and software are needed to characterize the stochastic properties of the data in a way that reflects the details of sample preparation and spectral acquisition, and to implement the statistical analysis workflows in a way that is useful for both experimentalists and statisticians.

MSstats is an open-source R-based package that provides such statistical functionalities for relative quantification of proteins and peptides using a flexible family of linear mixed models. For some special cases, the methods and the implementation were previously described (Chang *et al.*, 2012; Clough *et al.*, 2012; Surinova *et al.*, 2013). Here we present MSstats 2.0, a package that integrates the methodology across several mass spectrometric workflows and data acquisition strategies, contains new functionalities for model-based analyses, provides example datasets, enables interoperability with the existing popular computational tools and facilitates their use by both statistical and proteomic communities.

*To whom correspondence should be addressed.

2 DESCRIPTION

2.1 Applicability

MSstats 2.0 is applicable to multiple types of sample preparation, including label-free workflows, workflows that use stable isotope-labeled reference proteins and peptides and workflows that use fractionation. It is applicable to liquid chromatography coupled with mass spectrometry (LC-MS) in data-dependent acquisition (DDA, or shotgun) mode, targeted selected reaction monitoring (SRM) and data-independent acquisition (DIA, or Sequential Windowed data independent Acquisition of the Total High-resolution Mass Spectra (SWATH-MS)). It is applicable to experiments that make arbitrary complex comparisons of experimental conditions or times.

2.2 Statistical functionalities

MSstats 2.0 takes as input identified and quantified spectral peaks from multiple mass spectrometry runs and performs three analysis steps. The first step, *data processing and visualization*, transforms and normalizes the intensities of the peaks. It then generates workflow-specific and customizable numeric summaries and plots such as in Figure 1 for data visualization and quality control.

The second step, *statistical modeling and inference*, relies on the lm and lmer functionalities in R but customizes the model to the specific experiment and to each protein. The implementation automatically detects the experimental design (e.g. group comparison or time course, presence of labeled reference peptides or proteins) from the data. It then reflects the experimental design, the type of spectral acquisition strategy and the scope of conclusions (e.g. restricted to the subjects or expanded to the underlying populations) and fits an appropriate linear-mixed model.

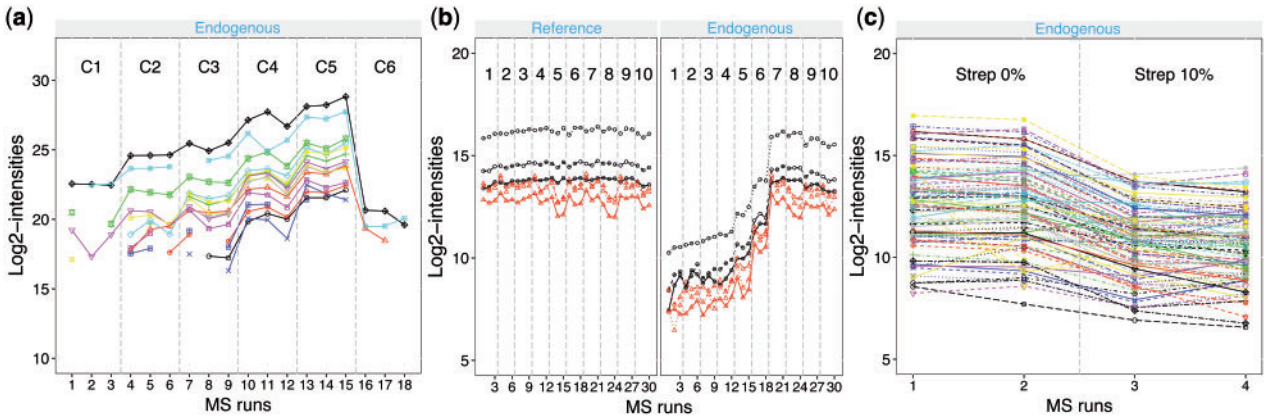

Fig. 1. Visualization of one representative protein in a DDA, an SRM and a DIA experiment. Colors and shapes represent peptides, and multiple line types of a same color and shape represent the fragments of the peptide. Vertical lines separate times or conditions. (**a**) Protein alcohol dehydrogenase-yeast spiked into a complex background in six concentrations. (**b**) Protein ACH1, at 10 times points after a stress. (**c**) Protein FabG of Streptococcus, with 0 and 10% human plasma added

As an example, Figure 1 shows representative profile plots for one protein in experiments from three spectral acquisition strategies. Spectral features in label-free DDA experiments are peptides; they have a relatively large variation and many missing values. Spectral features in label-based experiments are grouped into endogenous and reference measurements. Spectral features in SRM and DIA experiments are fragments, grouped into peptides. SRM experiments often present few missing values. DIA experiments have the largest number of spectral features, which differ in quality and variation. MSstats 2.0 takes all these differences into account. The users have choices of viewing the between-feature interferences as systematic deviations or a random noise, specifying constant or feature-specific variation, and imputing the missing values or filtering out poor quality features. MSstats 2.0 contains tools for model-based diagnostics to help specify the appropriate model.

The model is used to detect differentially abundant proteins or peptides, or to summarize the protein or peptide abundance in a single biological replicate or condition (that can be used, e.g. as input to clustering or classification). MSstats 2.0 automatically calculates the necessary linear combinations of model terms, and produces model-based summaries for each experimental workflow (e.g. estimates of abundance in each subject or condition, log-fold changes between the compared conditions, the associated estimates of variation and the degrees of freedom).

The third step, *statistical experimental design*, views the dataset being analyzed as a pilot study, uses its variance components and calculates the minimal number of replicates necessary to achieve a pre-specified statistical power.

2.3 Interoperability with existing computational tools

MSstats 2.0 is designed as a link between researchers with and without statistical background. Proteomic practitioners (the primary audience of the package) have a limited familiarity with R, and in the past this has hindered a broad adoption of R-based implementations. Now MSstats 2.0 is available as an external tool within Skyline (Broudy *et al.*, 2014), a popular Graphical user interface (GUI) tool for quantitative proteomics with >1100 registered users. The external tool support within Skyline manages MSstats installation, point-and-click execution, parameter collection in Windows forms and output display. Skyline manages the annotations of the experimental design and the processing of raw data. It outputs a custom report that is fed as a single stream input into MSstats. This design buffers proteomics users from the details of the R implementation, while enabling rigorous statistical modeling. MSstats 2.0 also benefits from inclusion in Skyline community resources such as message boards, support in tutorials and examples of publicly available datasets.

Alternatively, MSstats takes as input data in a tabular format, produced by any spectral processing tool such as SuperHirn, MaxQuant, Progenesis, MultiQuant, OpenMS or OpenSWATH.

For statistics experts, MSstats 2.0 satisfies the interoperability requirements of Bioconductor and takes as input data in the MSnSet format (Gatto and Lilley, 2012). The command line-based workflow is partitioned into a series of independent steps, which facilitate the development and testing of alternative statistical approaches.

3 CONCLUSIONS

MSstats 2.0 enables both the generality and the flexibility of the statistical analysis of a number of quantitative proteomic workflows. Its implementation as an external tool within Skyline introduces rigorous statistical methodology to a broader proteomic community. Its Bioconductor implementation facilitates the development of new statistical methodology, such as work with additional novel data acquisition strategies and labeling types.

Funding: The work was supported in part by the NSF CAREER award 1054826 and the NSF SI2-SSE award 1047962 (to O.V). The authors thank Hannes Röst from ETH Zürich for making available the S. Pyogenes dataset, supported by the grant ETH-30 11-2.

Conflict of Interest: none declared.

REFERENCES

Broudy,D. *et al*. (2014) A framework for installable external tools in Skyline. *Bioinformatics*, **30**, 2521–2523.

Chang,C. *et al*. (2012) Protein significance analysis in selected reaction monitoring (SRM) measurements. *Mol. Cell. Proteomics*, **11**, M111.014662.

Clough,T. *et al*. (2012) Statistical protein quantification and significance analysis in label-free LC-MS experiments with complex designs. *BMC Bioinformatics*, **13** (**Suppl. 16**), S6.

Gatto,L. and Lilley,K.S. (2012) MSnBase-an R/Bioconductor package for isobaric tagged mass spectrometry data visualization, processing and quantitation. *Bioinformatics*, **28**, 288–289.

Käll,L. and Vitek,O. (2011) Computational mass spectrometry–based proteomics. *PLoS Comput. Biol.*, **7**, e1002277.

Surinova,S. *et al*. (2013) Automated selected reaction monitoring data analysis workflow for large-scale targeted proteomic studies. *Nat. Protoc.*, **8**, 1602–1619.