Recent developments in MSstats ecosystem: a collection of statistical methods for general scalable quantitative analysis of proteomic experiments

Devon Kohler[1], Mateusz Staniak[4], and Olga Vitek[1]

[1]Khoury College of Computer Science, Northeastern University, Boston, MA, USA
[4]University of Wrocław, Wrocław, Poland

The *MSstats* ecosystem is a family of open-source R/Bioconductor packages implementing statistical methods for quantitative mass spectrometry-based proteomic experiments. Here we review its recent developments, as well as advances in previously available methods and implementations.

MSstatsPTM is a new package for experiments studying post-translational modifications (PTMs). The package includes statistical methods to accurately detect relative changes in PTMs, removing confounding with changes in the global protein and revealing modifications of interest that are entirely masked by changes in the unmodified protein on an experimental level. The methods provides high flexibility in terms of modeling, accounting for outliers/missing values, and handling complex designs and robust implementations. The package is applicable to a variety of experimental designs and acquisition methods, including data-dependent acquisitions (DDA) that are label-free or tandem mass tag (TMT)-based. The package includes plotting functionalities for side by side comparisons between the effect of the modification vs global protein. The package is open source and is available on Bioconductor and Github.

MSstatsLiP is a new package for experiments using limited proteolysis-mass spectrometry (LiP-MS), a technique that uses limited proteolysis to detect differences in protein structural changes on a proteome-wide scale. Structural changes are detected by identifying significant changes in LiP peptide abundance across conditions. Changes in LiP abundance may be confounded with changes in the global protein, masking the true change in protein structure. *MSstatsLiP* removes this confounding, allowing the user to identify proteins whose structure is truly altered. The package produces barcode plots to analyze LiP peptide coverage over the entire protein sequence. The package is available on Bioconductor and Github.

In addition to new packages, there have been several notable advancements to the older packages in the *MSstats* ecosystem. *MSstats*, a package for detecting differentially abundant proteins in label-free experiments, is being updated to process large experimental datasets that do not fit into a standard computers memory. As the largest datasets are generated by DIA experimets, MSstats aims to improve its support for results of analysis performed with popular tools. We added a DIA-NN converter to the GitHub version of MSstats and MSstatsConvert packages. We also updated the existing Spectronaut converter to handle datasets analyzed with the latest version of the program. In particular, Spectronaut files larger than memory can be processed using functions provided in MSstatsBig package which is still in development. This package provides tools to greatly reduce the size of Spectronaut output and reshape it into the MSstats format. We plan to add more functionalities for effective work with large datasets, and welcome community suggestions, beta-testing and feedback.

MSstatsTMT, a package for detecting differentially abundant proteins in experiments with TMT labeling, now includes novel statistical framework to model time series experiments, where biological replicates are repeatedly measured over multiple conditions or times. The time series model in *MSstatsTMT* enables more accurate estimation of all the relevant sources of variation, allowing users to better identify deferentially abundant proteins which may have otherwise been missed. The model includes functionality to

account for design complexities, such as unbalanced designs, and automatically adjusts to the specific experimental design, easing implementation by the user. The methods are implemented in the newest version of *MSstatsTMT*, available on both Bioconductor and Github.

To increase the usability of the methods in the *MSstats* package family, we have recently released *MSstatsShiny*, an R-Shiny graphical user interface (GUI) integrated with the *MSstats*, *MSstatsTMT*, and *MSstatsPTM*. The GUI is designed to increase the generalizability of these packages, opening up the methods to the wider proteomics community, including researchers who cannot code in R. The GUI supports reproducible research through automatically tracking the user's analysis selections, and providing them an R script that can recreate their analysis. *MSstatsShiny* is available for local installation on Github, as well as on the cloud at `http://www.msstatsshiny.com`. The cloud based version is designed for smaller datasets, under 250 MB, while the local installation can handle any dataset that fits into the memory on the local computer.

Finally, to help foster the *MSstats* community of users we have created the *MSstats* Google Group where users can share their experiences, difficulties, solutions, and suggestions.

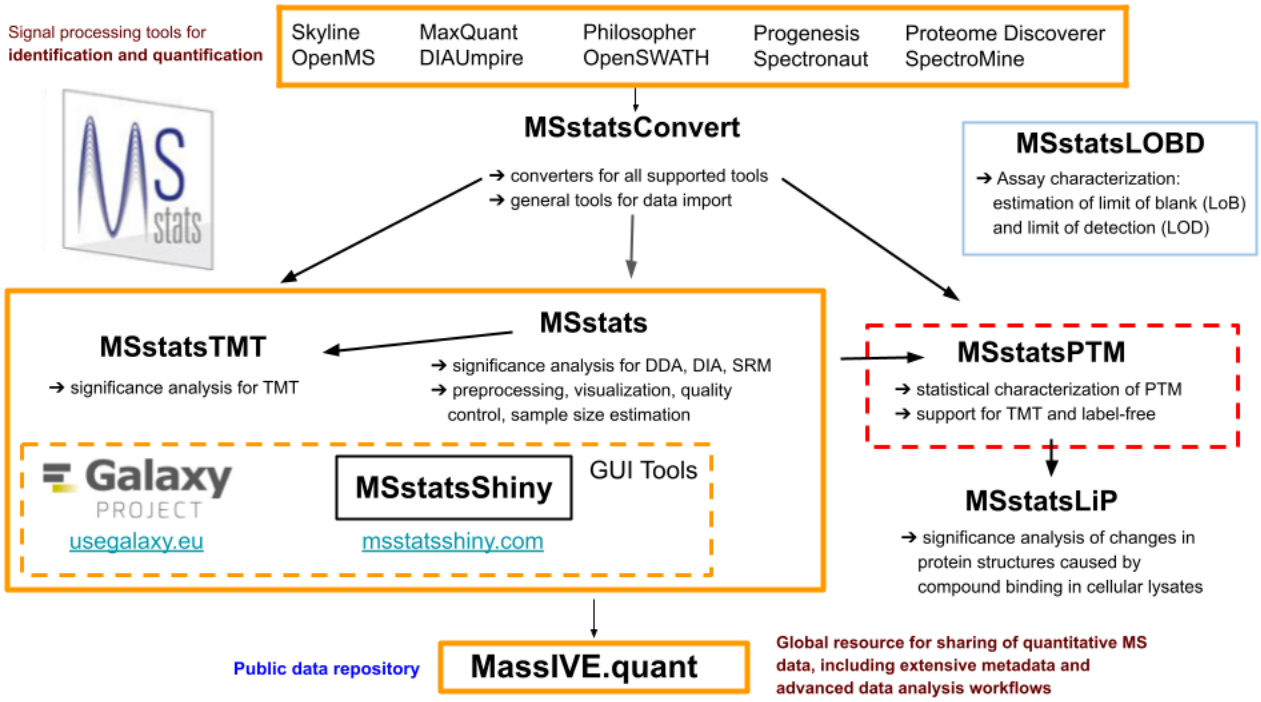

Figure 1: An overview of the *MSstats* ecosystem, showing all available packages and infrastructure.

Bio

Devon Kohler is a PhD student in Olga Vitek's lab at Northeastern University's Khoury College of Computer Science in Boston, MA. His primary research areas are in the application of statistical and causal inference to biological systems. Prior to this he received his Masters in Data Science from Northeastern.

Mateusz Staniak is a PhD student at University of Wrocław, Poland. His research focuses on statistical methods for quantitative proteomics. He is the main developer of MSstats package. He received his Masters in Mathematics from University of Wrocław.

Olga Vitek is Professor in Khoury College of Computer Sciences at Northeastern University. Her group develops statistical methods and software for studies of biomolecular systems.