The Journal of Open Source Software

Taming your metabolic datasets with MeDUSA

Laura Ann Castaneda [1*], **Eric Hetzel**[2*], **and Ahmed Ali** [1¶]

1 Leiden University, Leiden, Netherlands **2** Independent Researcher, Netherlands ¶ Corresponding author
* These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals
Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

Over ten trillion cells are hard at work in the human body(Bianconi, 2013) and there can be significant heterogeneity amongst them affecting biological development, disease progression, and treatment response(Zhang & Vertes, 2018).

One technique to categorize this heterogeneity is single cell analysis using mass spectrometry. However this technique introduces new challenges. One of which is the small sample volume, which limits the possibilities for separation and thus makes the data non-compatible with traditional analysis pipelines. We introduce the R package `MeDUSA` for Metabolomic Direct-infusion Untargeted Single-cell Analysis.

`MeDUSA` is a start-to-finish analysis package allowing metabolomics researchers to focus on analytical content rather than R proficiency. `MeDUSA` handles the suggested workflow by Southam (Southam et al., 2017) from data extraction to filtering without a chromatogram, carrying the data through statistical analysis and identification of features for biological interpretation.

Statement of Need

Due to the small volume of a single cell, direct infusion(DI), nano-electrospray ionization (nESI) is a highly suitable technique. However liquid chromatography mass spectrometry (LC-MS) is significantly more common than DI mass spectrometry, therefore most software is being developed for separation chromatography. For instance, XCMS offers filtering and statistical analysis with visualization, however, all of the filtering is reliant on the presence of a chromatogram(Smith et al., 2006). Similarly, MetaboAnalyst has an impressive interface that allows for metabolomic pathway analysis, but the full potential is only unlocked with a chromatogram(Pang, 2021). Even paid subscription based software such as Thermo Fisher Scientific's Compound Discoverer is designed to align and filter peaks based on a chromatogram(Cooper & Yang, 2024). Furthermore, the mentioned software does not allow the user to define preprocessing methods such as centroiding or alignment. In contrast, the modularity of `MeDUSA` is built specifically for direct infusion data, and it lays the framework for method specification. The modularity also offers the user the ability to bypass functions, introduce external functions to filter, and process the data as they see fit, an option that other software does not offer. Currently, if researchers want this level of modularity and customization designed for single cell data, they must write their own scripts which requires proficiency in a programming language and a significant time allotment. Therefore the metabolomic community needs a software option that will enable complete, modular, and customizable processing of mass spectra without a chromatogram.

The Journal of Open Source Software

Description

The goal of MeDUSA is to provide a toolset that is modular, customizable, and user friendly. There are five major sections along this standard flow as shown in Figure 1.

Modularity is achieved by using standardized interoperable data-objects. This allows the user to choose any collection and order of functions as they see fit. See the README for a list and description of the standard objects (https://github.com/laura-hetzel/MeDUSA#data-structures).

Customization is achieved by being greatly parameterized and leveraging modularity. This allows the user to dial in their variables, such as thresholds, aggregation methods, tolerances, etc. The user may also interrupt the suggested flow to perform any custom logic to their needs, and reintroduce their updated data into the MeDUSA flow, so long as the data structure is maintained. The three primary data-objects have the same structure and are differentiated by name for human readability along the suggested flow. However, they are technically interchangeable, thus increasing customization.

User-friendliness is achieved with "magic" functions, readability, and containerization. The magic functions leverage suggested parameter values and simplify many functions within each of the five major sections. These methods can allow a user to go from mzML files to a list of compounds in few commands. Figure 2 illustrates a detailed list of functions, and how the magic functions simplify them. To ease readability, files and methods are prepended with the expected input data-object type. This naming convention helps users identify what methods are available to them at different stages of the suggested flow. Containerization manages dependencies and provides HMDB and lipid data for convenient m/z to compound mapping.

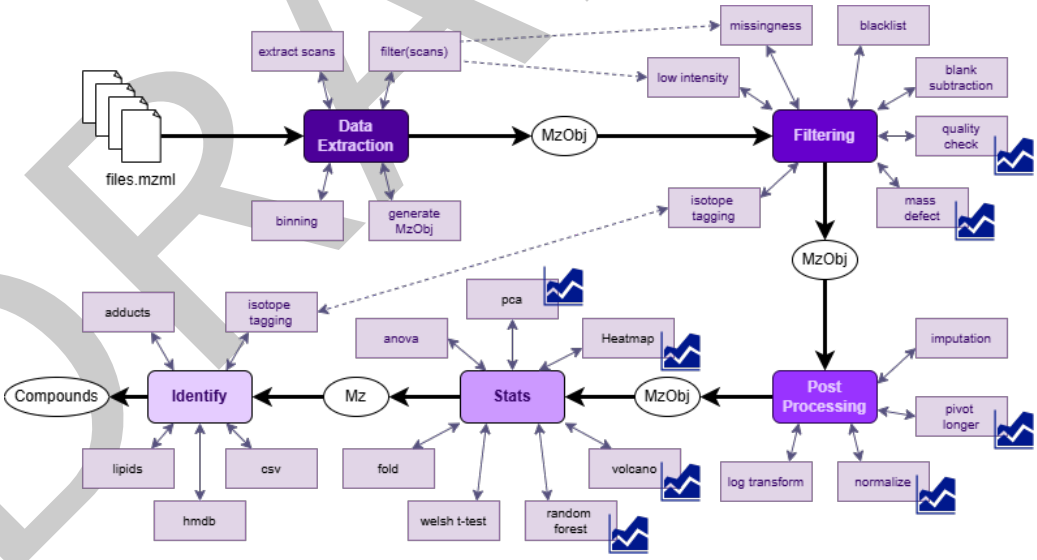

Figure 1: Map of the five sections of MeDUSA and the capabilities of each section. The bold arrows indicate a suggested workflow. The dashed arrows indicate references. The circled text indicated the object data type. The plot symbol indicates the function may output a plot.

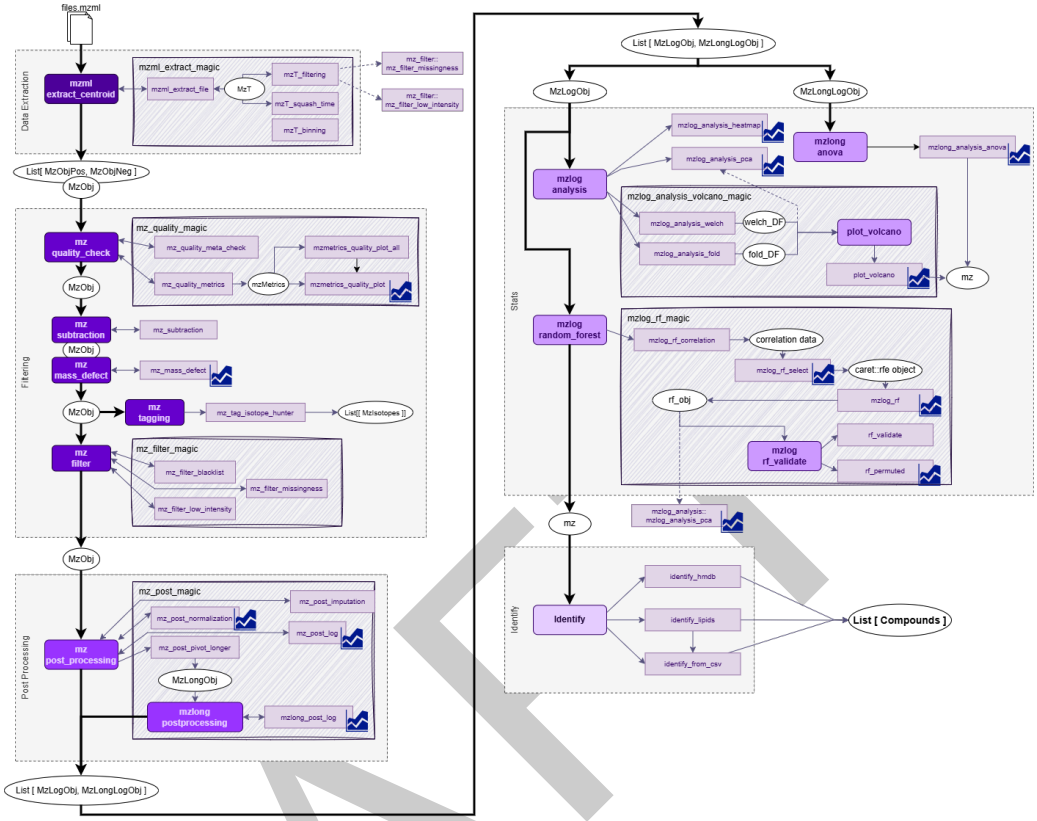

Figure 2: Detailed suggested flow and function map. Note the "magic" functions which aggregate similar functions for user ease. The bold arrows indicate a suggested workflow. The circled text indicated the object data type. The plot symbol indicates the function may output a plot.

Research projects using the software

Current research projects relating to single cell metabolic profiling of the cell cycle of FUCCI cells, stem cell differentiation, hypoxic organoids, and metastatic organoids would benefit from this package. The projects currently use two different commercially available software as well as R scripts that lack robustness for the data analysis; shifting the analysis to MeDUSA will enable timely and reliable results. Upon the release of the package, it will be implemented lab wide for live single cell metabolomics.

Acknowledgements

This research was funded in part by the European Union's Horizon 2020 research and innovation program under the Marie Skladowska-Curie grant agreement number 861196.

Rosa (species = Canis Familiaris, breed = Catahoula Leopard) was essential to the mental well-being and creativity of the authors, and therefore essential to the project.

We would like to thank Alida Kindt and Thomas Hankemeier for their support and direction on this project.

References

Bianconi, E. et al. (2013). An estimation of the number of cells in the human body. *Ann. Hum. Biol.*, *40*. https://doi.org/10.3109/03014460.2013.807878

Cooper, B., & Yang, R. (2024). An assessment of AcquireX and Compound Discoverer software 3.3 for non-targets metabolomics. *Sci. Rep..*, *14*. https://doi.org/10.1038/s41598-024-55356-3

Pang, Z. et. al. (2021). MetaboAnalyst 5.0: narrowingthe gap between raw spectra and functional insights. *Nucleic Acids Res.*, *49*. https://doi.org/10.1093/nar/gkab382

Smith, C. A., Want, E. J., O'Maille, G., & Siuzdak, G. (2006). XCMS: Processing Mass Spectormetry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.*, *78*. https://doi.org/10.1021/ac051437y

Southam, A. D., Weber, R. J. M., Engel, J., Jones, M. R., & Viant, M. R. (2017). A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics. *Nat. Protoc.*, *12*. https://doi.org/10.1038/nprot.2016.156

Zhang, L., & Vertes, A. (2018). Single-Cell Mass Spectrometry Approaches to Explore Cellular Heterogeneity. *Angew. Chem. Int.*, *57*. https://doi.org/10.1002/anie.201709719