bioRxiv preprint doi: https://doi.org/10.1101/2024.06.20.599545; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

MargheRita: an R package for LC-MS/MS SWATH metabolomics data analysis and confident metabolite identification based on a spectral library of reference standards

Ettore Mosca[1]*, Marynka Ulaszewska[2], Zahrasadat Alavikakhki[3], Edoardo Niccolò Bellini[2], Valeria Mannella[2], Gianfranco Frigerio[2], Denise Drago[2], Annapaola Andolfo[2]*

[1]Institute of Biomedical Technologies, National Research Council, Segrate (Milan), Italy
[2]ProMeFa, Proteomics and Metabolomics Facility, Center for Omics Sciences (COSR), IRCCS San Raffaele Scientific Institute, Milan, Italy
[3]Università degli Studi di Milano, Milan, Italy

*corresponding authors

Short Structured Abstract

Untargeted metabolomics by mass spectrometry technologies generates huge numbers of metabolite signals, requiring computational analyses for post-acquisition processing and databases for metabolite identification. Web-based data processing solutions frequently include only a part of the entire workflow thus requiring the use of different platforms. The R package “margheRita” enhances fragment matching accuracy and addresses the complete workflow for metabolomic profiling in untargeted studies based on liquid chromatography (LC) coupled with tandem mass spectrometry (MS/MS), especially in the case of data-independent acquisition, where all MS/MS spectra are acquired with high quantitative accuracy.

Availability and Implementation: source code and documentation are available at https://github.com/emosca-cnr/margheRita.

Contact: ettore.mosca@itb.cnr.it, andolfo.annapaola@hsr.it

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.20.599545; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Metabolomics is paving the way for comprehensive studies of the low molecular weight molecules within an organism (Schrimpe-Rutledge et al., 2016). Metabolites are regulated by both biological and environmental factors and, therefore, provide great potential to connect genotype and phenotype. However, the exploitation of untargeted mass spectrometry (MS)-based metabolomics is slowed down by fundamental challenges in metabolite identification (Schrimpe-Rutledge et al., 2016). Despite several computational approaches exist in the field (Stanstrup et al., 2019), validation of retention times (rt) – the time needed by a molecule to pass through a chromatographic column – and MS/MS fragmentation data with a reference standard is almost always required for confident metabolite identification (Schrimpe-Rutledge et al., 2016). This is especially true for untargeted studies that rely on SWATH-MS technology (Bonner & Hopfgartner, 2019), which detects, in a data independent way, tens of thousands of features, i.e., mass-to-charge ratio (m/z) and rt pairs. Existing computational tools cover only parts of the workflow required to translate raw LC-MS/MS SWATH data into metabolite abundance data. We developed "margheRita", an R package that provides functionalities for the entire workflow of MS-based metabolomics data analysis and metabolite identification, supported by an original library of reference standards **(Figure 1a)** for given chromatographic methods (**Supplementary Methods)**.

Implementation

The software "margheRita" is implemented as an R package (R Core Team, 2022). It requires feature-level data (rt, m/z, MS/MS spectra and abundance across samples) and sample-level data in text format. Feature-level data files can be obtained from raw files through the freely available software MS-DIAL (Tsugawa et al., 2015), which performs data extraction, peak picking, peak alignment, SWATH spectral deconvolution and peak identification, supporting multiple technologies and vendors. Sample-level data files must contain a series of annotations related to the experimental design, like injection order, experimental batch and condition.

Data are structured as a particular list, the "mRList", which is the shared input/output of all functions that operate on data. To support interoperability with other packages with a focus on metabolomics, mRList objects can be reorganized as a "MetaboSet" object, which is used by "notame" (Klåvus et al., 2020), or as a "PomaSummarizedExperiment" object, which is used by "POMA" (Castellano-Escuder et al., 2021).

The package provides a series of functions to import data from MS-DIAL files and to perform quality control, filtering, imputation and normalization. These functions are tailored to metabolomics data, like the "heatscatter" visualization of m/z vs rt values, the filtering of features by m/z values or by coefficient of variation in samples relatively to quality controls, and the probabilistic quotient normalization (Dieterle et al., 2006).

The package provides an original metabolite library **(Supplementary Methods)** that supports validated annotation ("Level 1", see below), in both positive and negative polarity, for 4 chromatographic columns, namely BEH Amide hydrophilic interaction liquid chromatography (HILIC), Reverse-Phase (RP)-C18 with two different gradients (RPLong and RPShort), zic-pHILIC (pZIC), RP-

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.20.599545; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

C8 (LipC8) and contains about 2,000 MS/MS spectra obtained through data dependent acquisition. Moreover, margheRita supports the use of MS-DIAL reference libraries, which enable putative identification ("Level 2" or "Level 3"), for large number of molecules (~$10^{\wedge 5}$).

The scoring of an association between a feature f_i and a library metabolite m_j, is based on the following four quantities, where "*" indicates the quantities of library metabolites, while $k \in \{1, 2, \ldots, n_k\}$ and $l \in \{1, 2, \ldots, n_l\}$ indicate MS/MS fragments of f_i and m_j respectively:

(i) rt error $\varepsilon_t(i,j) = |t(i) - t^*(j)|$, where t is rt;

(ii) ppm (part per million) error $\varepsilon_{m_z}(i,j) = \frac{|m_z(i) - m_z^*(j)|}{m^*(j)} \cdot 10^6$ where m_z is m/z;

(iii) ppm error of MS/MS fragments: $\varepsilon_{m_z}(i,k;j,l) = \left(\frac{|m_z(i,k) - m_z^*(j,l)|}{m_z^*(j,l)}\right)$;

(iv) relative intensity error of MS/MS fragments: $\varepsilon_I(i,k;j,l) = \left(\frac{|I(i,k) - I^*(j,l)|}{I^*(j,l)}\right)$, where I is fragment peak relative intensity.

A match between f_i and m_j exists when the errors lie below the corresponding thresholds $(\alpha_t, \alpha_m, \alpha_I)$ and is classified as:

(i) "Level 1" (supported by rt, m/z and MS/MS), if there exist (i,j,k,l) such that $\varepsilon_t(i,j) < \alpha_t$, $\varepsilon_m(i,j) < \alpha_m$, and $(\varepsilon_I(i,k;j,l) < \alpha_I) \wedge (\varepsilon_m(i,k;j,l) < \alpha_m)$;

(ii) "Level 2" (supported by m/z and MS/MS), if there exist (i,j,k,l) such that $\varepsilon_m(i,j) < \alpha_m$, and $(\varepsilon_I(i,k;j,l) < \alpha_I) \wedge (\varepsilon_m(i,k;j,l) < \alpha_m)$;

(iii) "Level 3a" (supported by rt and m/z), if there exist (i,j) such that $\varepsilon_t(i,j) < \alpha_t$ and $\varepsilon_m(i,j) < \alpha_m$;

(iv) "Level 3b" (supported by m/z), if there exist (i,j) such that $\varepsilon_m(i,j) < \alpha_m$.

Besides the level label, every match between f_i and m_j is supplied with a number of quantitative and qualitative scores that summarize various aspects of the supporting evidence: classification of rt $\varepsilon_t(i,j)$ in {"super", "acceptable", "unacceptable"}; classification of $\varepsilon_m(i,j)$ in {"super", "acceptable", "suffer", "unacceptable"}; number of matching MS/MS peaks; the ratio between the number of matching MS/MS peaks and the number of metabolite MS/MS peaks; whether the whole feature f_i is part of its set of MS/MS fragments or not. These scores are used to filter the list of associations between features and metabolites. Indeed, a series of one-to-many and many-to-one associations between f_i and m_j could arise, especially when considering large sets of features and reference metabolites. Therefore, we designed an algorithm that, leveraging the classification (Level 1, Level 2, Level 3a and Level 3b), the errors (ε_t, ε_{m_z} and ε_I) and quantitative and qualitative scores, resolves the vast majority of the one-to-many and many-to-one associations, retaining only those supported by the strongest evidence **(Supplementary Methods)**, thus providing a very clear, immediate and selective annotation of the features.

Lastly, the statistical significance of pathway enrichment in a list of metabolites (represented by PubChem (Kim et al., 2023) identifiers) can be subject to over representation analysis and metabolite set enrichment analysis (i.e., gene set enrichment analysis applied to metabolites), by means of a wrapper function that takes advantage of the R package clusterProfiler (Wu et al., 2021).

Assessment of margheRita metabolite identification

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.20.599545; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To assess the performance of metabolite identification in margheRita we set up two experiments (**Supplementary Methods**): the first, "Standards", containing a dataset with a panel of 33 standard metabolites that could serve as ground truth; the second, "urine", to simulate a "real life dataset" through LC-MS/MS (RP-C18, short gradient, SWATH analysis, both polarities) derived from human urine samples. In both the experiments, we compared the results of margheRita metabolite identification with the annotations provided by MS-DIAL for the same features. In Standards dataset, margheRita recovered 89% and 82% of the metabolites in, respectively, positive and negative polarity, mostly supported by Level 1 evidence (as expected). We can speculate that margheRita did not reach the 100% because of the possible discrepancies between data extraction, peak picking, peak alignment and peak identification performed by MS-DIAL versus SCIEX-OS (SCIEX, 2023), which is the SCIEX software used for the spectral library. In Standards dataset, on the other hand, MS-DIAL correctly annotated the 12% (positive) and 27% (negative) of the corresponding features, mostly at Level 2 **(Figure 1b, Supplementary Tables ST5,7-8)**. In urine dataset, margheRita found 386 (of which 135 at Level 1) and 394 metabolites (of which 106 at Level 1) in, respectively, positive and negative polarity, while MS-DIAL provided correct annotation for approximately half of such features (43% and 55% respectively) **(Figure 1b, Supplementary Tables ST6,9-10)**.

Conclusion

The R package "margheRita" provides preprocessing, metabolite identification and post-acquisition analysis of LC-MS/MS SWATH metabolomics data. It also provides an original "Level 1" metabolite library thus improving the performance of metabolite identification in SWATH-based metabolomics. An overview of its capabilities is available at https://emosca-cnr.github.io/margheRita.

Data Availability Statement

The data underlying this article are available in Zenodo at https://doi.org/10.5281/zenodo.11243781.

Figure 1: Overview of margheRita. a) The main functionalities of our R package. **b)** Number of metabolites found by "margheRita" in comparison with MS-DIAL v 4.7 at various levels of confidence in "Standards" and "urine" experiments, in positive (+) and negative (-) polarity; NA: annotation not available; see Supplementary Tables ST5-6.

References

Bonner, R., & Hopfgartner, G. (2019). SWATH data independent acquisition mass spectrometry for metabolomics. *TrAC Trends in Analytical Chemistry*, *120*, 115278. https://doi.org/10.1016/j.trac.2018.10.014

Castellano-Escuder, P., González-Domínguez, R., Carmona-Pontaque, F., Andrés-Lacueva, C., & Sánchez-Pla, A. (2021). POMAShiny: A user-friendly web-based workflow for metabolomics and proteomics data analysis. *PLOS Computational Biology*, *17*(7), e1009148. https://doi.org/10.1371/journal.pcbi.1009148

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.20.599545; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic Quotient Normalization as Robust Method to Account for Dilution of Complex Biological Mixtures. Application in ^{1}H NMR Metabonomics. *Analytical Chemistry*, *78*(13), 4281–4290. https://doi.org/10.1021/ac051632c

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., & Bolton, E. E. (2023). PubChem 2023 update. *Nucleic Acids Research*, *51*(D1), D1373–D1380. https://doi.org/10.1093/nar/gkac956

Klåvus, A., Kokla, M., Noerman, S., Koistinen, V. M., Tuomainen, M., Zarei, I., Meuronen, T., Häkkinen, M. R., Rummukainen, S., Farizah Babu, A., Sallinen, T., Kärkkäinen, O., Paananen, J., Broadhurst, D., Brunius, C., & Hanhineva, K. (2020). "Notame": Workflow for Non-Targeted LC–MS Metabolic Profiling. *Metabolites*, *10*(4), 135. https://doi.org/10.3390/metabo10040135

R Core Team. (2022). R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. R Foundation for Statistical Computing.

Schrimpe-Rutledge, A. C., Codreanu, S. G., Sherrod, S. D., & McLean, J. A. (2016). Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *Journal of the American Society for Mass Spectrometry*, *27*(12), 1897–1905. https://doi.org/10.1007/s13361-016-1469-y

SCIEX. (2023). *SCIEX OS* (3.3). https://sciex.com.

Stanstrup, J., Broeckling, C., Helmus, R., Hoffmann, N., Mathé, E., Naake, T., Nicolotti, L., Peters, K., Rainer, J., Salek, R., Schulze, T., Schymanski, E., Stravs, M., Thévenot, E., Treutler, H., Weber, R., Willighagen, E., Witting, M., & Neumann, S. (2019). The metaRbolomics Toolbox in Bioconductor and beyond. *Metabolites*, *9*(10), 200. https://doi.org/10.3390/metabo9100200

Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., & Arita, M. (2015). MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nature Methods*, *12*(6), 523–526. https://doi.org/10.1038/nmeth.3393

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., & Yu, G. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, *2*(3), 100141. https://doi.org/10.1016/j.xinn.2021.100141

a

Data extraction

- Peak picking
- Peak alignment
- MS/MS SWATH deconvolution

MS-DIAL

Preprocessing

bioRxiv preprint doi: https://doi.org/10.1101/2024.06.21.599988; this version posted June 22, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

- Filtering
- Imputation
- Normalization
- Quality control

Metabolite Annotation

- Level 1 library of reference standards for LC-MS/MS SWATH (HILIC, RP-C8, zic-pHILIC, RP-C18 (RPLong and RPShort gradient)
- Feature-metabolite association scoring & filtering
- Support of external reference standards libraries

Other features

- Input: MS-DIAL text files, tab-delimited files
- Internal data structure («mList»)
- Interoperability with other packages for metabolomics
- Principal Component Analysis
- Statistical tests of metabolite level variation
- Enrichment analysis of metabolic pathways
- Source code: https://github.com/emosca-cnr/margheRita
- Documentation: https://emosca-cnr.github.io/margheRita

b