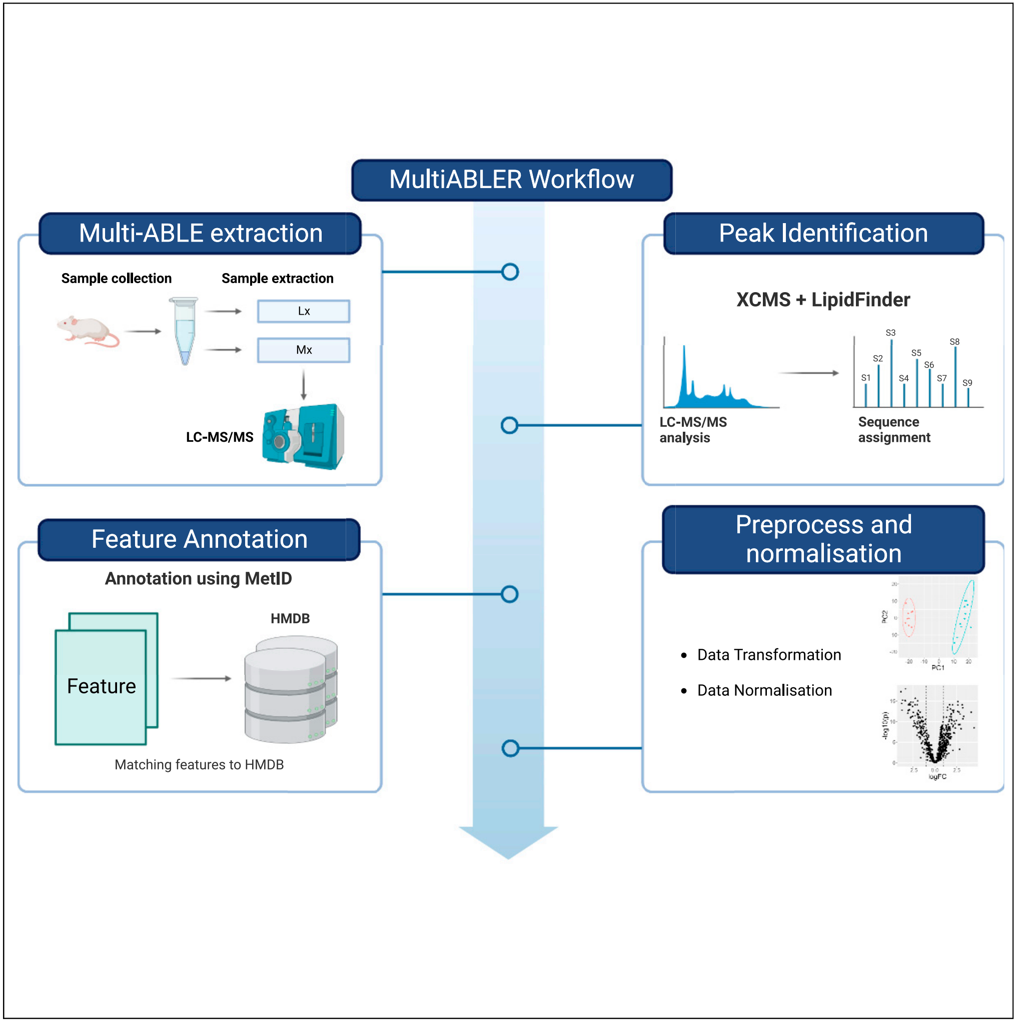
iScience

CellPress OPEN ACCESS

Article

Integrative processing of untargeted metabolomic and lipidomic data using MultiABLER



Ian C.H. Lee,
Sergey Tumanov,
Jason W.H. Wong,
Roland Stocker,
Joshua W.K. Ho

jwkho@hku.hk

Highlights

A unified and simple upstream LC-MS/MS data processing and analysis pipeline

Concurrent data analysis for metabolomic and lipidomic data from Multi-ABLE

Lee et al., iScience *26*, 106881
June 16, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.isci.2023.106881

CellPress
OPEN ACCESS

Article

Integrative processing of untargeted metabolomic and lipidomic data using MultiABLER

Ian C.H. Lee,[1,2] Sergey Tumanov,[3,4] Jason W.H. Wong,[1,5] Roland Stocker,[3,6] and Joshua W.K. Ho[1,2,5,7,*]

SUMMARY

Mass spectrometry (MS)-based untargeted metabolomic and lipidomic approaches are being used increasingly in biomedical research. The adoption and integration of these data are critical to the overall multi-omic toolkit. Recently, a sample extraction method called Multi-ABLE has been developed, which enables concurrent generation of proteomic and untargeted metabolomic and lipidomic data from a small amount of tissue. The proteomics field has a well-established set of software for processing of acquired data; however, there is a lack of a unified, off-the-shelf, ready-to-use bioinformatics pipeline that can take advantage of and prepare concurrently generated metabolomic and lipidomic data for joint downstream analyses. Here we present an R pipeline called MultiABLER as a unified and simple upstream processing and analysis pipeline for both metabolomics and lipidomics datasets acquired using liquid chromatography-tandem mass spectrometry. The code is available via an open-source license at https://github.com/holab-hku/MultiABLER.

INTRODUCTION

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a common analytical tool for untargeted metabolomic and lipidomic studies in biomedical research. LC-MS/MS data analysis typically involves raw data pre-processing, feature annotation, and statistical analysis.[1] There are many existing software and algorithms to perform different steps in the analysis pipeline, including XCMS,[2] OpenMS,[3] MS-DIAL,[4] LipidFinder,[5] and MZmine2[6] for MS data processing, LipidMS,[7] LipidMatch from LipidMatch Flow,[8] METLIN,[9] and metID[10] for feature annotation, ProteoMM[11] and CRMN[12] for data normalization, and limma[13] and MetaboAnalyst[14] for statistical analysis and data visualization. However, these tools often operate as standalone programs for specific tasks such that there is a lack of a single workflow from raw data to statistical analysis. In particular, there are different normalization methods available to prepare the data. These include baseline approaches such as median normalization and quantile normalization, classification-based approaches such as EigenMS, and internal-based approaches such as NOMIS and CRMN. Some existing bioinformatics pipelines try to tackle this problem, such as a metabolomic analysis pipeline on KNIME[15] with OpenMS; however, that platform is not commonly used by biomedical researchers, and the OpenMS pipeline does not incorporate lipidomic annotations. LipidMatch Flow can perform upstream lipidomic processing; however, it is limited to an outdated R version (R V3.3.3). MS-DIAL and MetaboAnalyst provide raw spectra processing and statistical analysis; however, they lack the feature annotation function for either lipidomic or metabolomic data. Finally, multi-omics integration programmes such as mixOmics[16] and Paintomics[17] provide multivariate methods for downstream integration and statistical analysis but require a processed feature table as input. There are also commercially available pipelines by the mass spectrometer vendors; however, these are usually only able to process data generated by specific instruments. The lack of a simple, easily reproducible, and unified workflow that can run on R and python for both LC-MS/MS lipidomic and metabolomic research in particular makes it difficult to compare results from studies that used a dual metabolomic and lipidomic extraction method.

With the advancement in sample extraction techniques, researchers have developed concurrent omic sample extraction methods to study the metabolomic and lipidomic profile from a single sample.[18–20] By performing dual extraction using a specific solvent mixture, researchers can extract sufficient hydrophilic metabolites and hydrophobic lipids from the aqueous and organic phase for LC-MS/MS analyses in a single extraction. This greatly reduces the volume of samples required for the experiment design and allows researchers to investigate the profile of tiny samples such as the atherosclerotic plaque in a mouse or tissue

[1] School of Biomedical Sciences, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

[2] Laboratory of Data Discovery for Health Limited ($D^2$4H), Hong Kong Science Park, Hong Kong SAR, China

[3] Heart Research Institute, 7 Eliza Street, Newtown, NSW 2042, Australia

[4] Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia

[5] Centre for PanorOmic Sciences, LKS Faculty of Medicine, The University of Hong Kong, Pokfulam, Hong Kong SAR, China

[6] School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW 2006, Australia

[7] Lead contact

*Correspondence: jwkho@hku.hk

https://doi.org/10.1016/j.isci.2023.106881

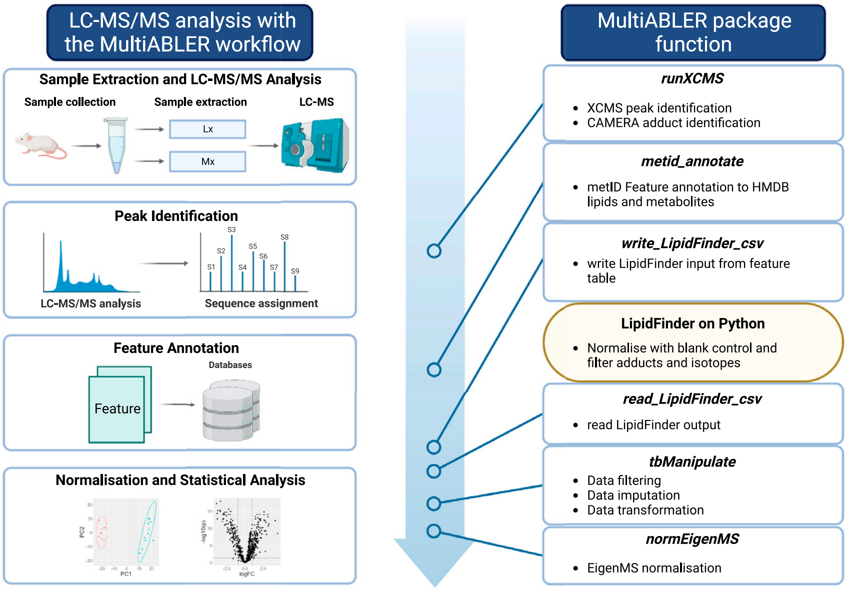

LC-MS/MS analysis with the MultiABLER workflow
Sample Extraction and LC-MS/MS Analysis
Sample collection
Sample extraction
LC-MS
Lx
Mx
Peak Identification
S1
S2
S3
S4
S5
S6
S7
S8
S9
LC-MS/MS analysis
Sequence assignment
Feature Annotation
Databases
Feature
Normalisation and Statistical Analysis
PC1
PC2
logFC
MultiABLER package function
runXCMS
• XCMS peak identification
• CAMERA adduct identification
metid_annotate
• metID Feature annotation to HMDB lipids and metabolites
write_LipidFinder_csv
• write LipidFinder input from feature table
LipidFinder on Python
• Normalise with blank control and filter adducts and isotopes
read_LipidFinder_csv
• read LipidFinder output
tbManipulate
• Data filtering
• Data imputation
• Data transformation
normEigenMS
• EigenMS normalisation

Figure 1. Schematic overview of MultiABLER

Multi-ABLE is a sample collection method that enables simultaneous generation of proteomics (not shown), lipidomic (Lx), and metabolomics (Mx) data from small amounts of tissues. After sample collection, omics data are collected using liquid chromatography-tandem mass spectrometry (LC-MS/MS). The raw mass spectrometry (MS) data are subjected to peak identification using XCMS. Peaks are annotated against HMDB and LipidMaps using metID. LipidFinder is used to normalize and filter the data against the solvent blank control, and the feature table is normalized using EigenMS and subsequently median normalization. Peak identification, feature annotation, and feature table normalization are all packaged in the MultiABER package that enables the lipidomic and metabolomic data to be processed and analyzed jointly. Figure created with BioRender.com.

biopsy.[19,20] The Multi-ABLE method developed by Talib et al.[19] is a sample extraction method using a barocycler to generate a single lysate of the biological tissue suitable for subsequent proteomic, lipidomic, and metabolomic analyses and multi-omic data generation. By pressurizing the tissue sample in the barocycler, sample liquification can be achieved in a very small volume of buffer (between 10 and 50 μL), and proteomic, metabolomic, and lipidomic extraction can be done using aliquots of the same sample. Therefore, there is a greater need in the informatics part of the metabolomic pipeline to have a simple and unified processing pipeline to accompany the data acquisition method and perform the data analysis for the concurrently extracted omic data.

Here, we present MultiABLER as a data analysis workflow for metabolomic and lipidomic data analysis to accompany the Multi-ABLE extraction described by Talib et al.[19] We compare the performance of different normalization methods to evaluate which is best suited for the analysis pipeline. The workflow is implemented in the MultiABLER package available on GitHub. The package integrates with LipidFinder for raw data processing to perform the MultiABLER workflow and provides the functionality to normalize and analyze the LC-MS/MS data. The schematic overview for the pipeline is shown in Figure 1. The data analysis pipeline is reproducible for both metabolomic and lipidomic studies, providing an easy and fair comparison for metabolomic and lipidomic data collected using the Multi-ABLE extraction.

RESULTS

Evaluation of the normalization methods

To decide which normalization method was most appropriate for the MultiABLER pipeline, we analyzed the lipidomic profiles of liver tissue from 19 apolipoprotein gene knockout (*Apoe*$^{-/-}$) mice wild-type (WT) for the biliverdin reductase a gene (*Brva*$^{+/+}$) and 16 *Apoe*$^{-/-}$*Brva*$^{-/-}$ Bvra double gene knockout (DKO) mice collected by Chen et al.[21] Lipidomic data were collected on LC-MS/MS in positive mode. The raw data were first processed using the *runXCMS* function from MultiABLER and then cleaned using LipidFinder, using the parameters described in Table S1. Features present in at least three samples were retained and imputed the data using half of the global minimum. The data were finally log transformed using base 2.

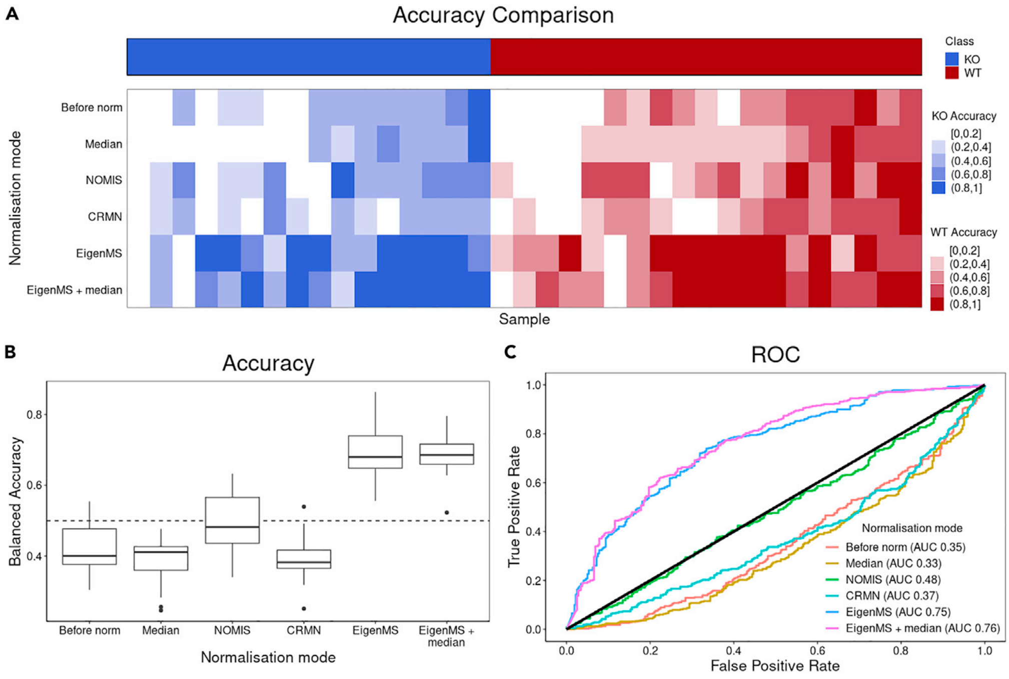


Figure 2. Evaluation of different normalization methods

(A–C) Re-sampled k-fold cross-validation was used to classify liver cells according to WT/KO label measured by (A) balanced accuracy for each sample by 5-fold cross-validation, (B) balanced accuracy for all samples in each repetition by 5-fold cross-valildation, and (C) area under the receiver operating characteristic (ROC) curve.

To decide which normalization method should be used for the MultiABLER pipeline, we compared the classification performance of median normalization, EigenMS, and the internal standard-based methods CRMN and NOMIS with the data before normalization. The performance of the normalized data was measured by performing 5-fold cross-validations on each dataset using ClassifyR (Figure 2). Without using any normalization, the performance was very low (median balanced accuracy [BA] = 0.40; area under the receiver operating characteristic (ROC) curve [AUC] = 0.35). Except for median normalization, all normalization showed an increase in performance compared with the data before normalization. Comparing AUC and the BA, EigenMS performed the best (median BA = 0.68; AUC = 0.75). Although internal standard was used in the experiments, both NOMIS and CRMN had a lower performance compared with EigenMS (NOMIS: median BA = 0.48, AUC = 0.48; CRMN: median BA = 0.38, AUC = 0.37). Furthermore, using median normalization after EigenMS normalization showed a performance similar to using EigenMS alone (median BA = 0.69; AUC = 0.76). Running median normalization after EigenMS normalization allowed all features across different omics to share the same median for fair comparison. Based on the comparison, EigenMS and median normalization were implemented in the MultiABLER pipeline.

Application case study

Multi-omics investigation of arterial lesions in a mouse model of atherosclerosis

To examine the functionality of MultiABLER, we analyzed the metabolomic and lipidomic profile of arterial tissue from 9 WT (*Apoe*$^{+/+}$) and 11 Apoe$^{-/-}$ mice collected and analyzed using the multi-ABLE method.[19] Metabolomic and lipidomic data were collected on LC-MS/MS in both positive and negative mode. Raw data were converted into mzML files using MSConvert. The data were then processed and annotated using the MultiABLER pipeline, using the parameters described in Table S1. Briefly, the raw LC-MS/MS data were processed with XCMS and filtered with LipidFinder. The lipidomic and metabolomic peaks were annotated with metID. Features were then filtered and normalized. Features with less than 3 representations were filtered out from the data, and the missing values were imputed using half of the global minimum value. The data were log transformed using base 2, and EigenMS and median normalization were used to normalize the feature table. In total, 1,650 unique lipids and 1,314 metabolites were found in the data (Table S2).

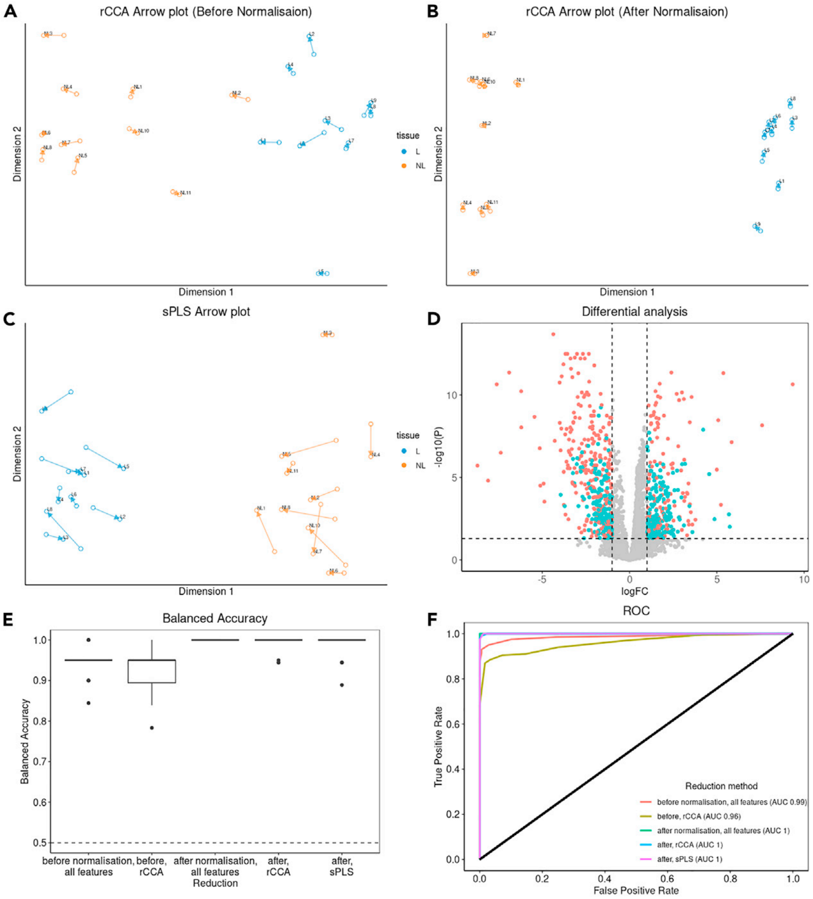


Figure 3. Application of the MultiABLER pipeline to jointly analyze metabolomics and lipidomic data from mouse arterial tissue

(A-B) rCCA analysis before and after the normalization. Arrow plot was used to visualize the agreement of the metabolomic and lipidomic data. The empty circle of the arrowhead indicates the sample in the space associated with the lipidomic components, and the arrowhead indicates the location of the sample associated with the metabolomic components L: lesion; NL: non-lesion.

(C) sPLS analysis after normalization. Arrow plot is used to visualize the agreement of the metabolomic and lipidomic data.

(D) Differential analysis using limma model. Vertical dotted line indicates logFC at −1 and 1. Horizontal dotted line indicates the p value at 0.05. logFC: log-fold change; Lx: Lipidomic; Mx: Metabolomic.

(E and F) Comparison of different analysis methods to the data using ClassifyR. 5-fold cross-validation was performed using differentially expressed features identified before and after normalization (all features), sPLS components identified after normalization, and rCCA canonical variates identified before and after normalization. The performance of each method was assessed using the balanced accuracy and ROC.

To evaluate the performance of the pipeline, we performed differential analysis to identify differentially expressed lipids and metabolites, sparse partial least squares (sPLS) projection to produce a multivariate model, and regularized canonical correlation analysis (rCCA) to identify correlation between the lipidomics and metabolomics. Differential analysis was performed using limma, while sPLS and rCCA analyses were performed using mixOmics. rCCA demonstrated that running the normalization increased the correlation of lipidomics and metabolomics data (Figures 3A and 3B). sPLS analysis using the normalized data also identified two distinct networks of lipids and metabolites (Figure S1). Using a limma model, 903 lipids and 601 metabolites were identified as differentially expressed between arterial tissue of *Apoe*$^{+/+}$ and *Apoe*−/− mice (adjusted $p < 0.05$), with 414 lipids and 334 metabolites having a log-fold change (base 2) > 1 (Figure 3D, Table S2). Pathway analysis using MetaboAnalyst showed that glycerophospholipid metabolism was affected between lesion and non-lesion tissue. The results demonstrated the ability of MultiABLER to process lipidomic and metabolomic data and to produce statistical validated results for biological interpretation. To verify the results of the differential analysis and multivariate analysis, cross-validation was used to evaluate the different methods. Differentially expressed features, sPLS latent components, and rCCA canonical variates were used to classify the samples in 5-fold cross-validation, and the performance was measured using the balanced accuracies and AUC. Cross-validation classification result showed that the normalization we used in MultiABLER improved the BA and AUC in both differential expression and rCCA analysis (Figures 3E and 3F). Computational time for each step in the pipeline is included in Table S3.

DISCUSSION

MultiABLER provides a simple analytic framework for metabolomic and lipidomic data collected from the Multi-ABLE extraction and LC-MS/MS. The pipeline is light weighted and provides clear steps to analyze concurrent metabolomic and lipidomic data. Moreover, as the pipeline is highly modularized, MultiABLER can also be used to analyze individual metabolomic and lipidomic profiles. However, MultiABLER currently does not provide functions to analyze proteomic data, which is another important output from the MultiABLER extraction method. Integration for proteomic analysis is planned for future development.

Using the output of MultiABLER, multiple statistical analyses can be performed on the normalized feature table of the metabolomic data. Cross-validation results in different downstream integrated analyses showed that the feature table generated from MultiABLER is suitable for various statistical analyses, including differential expression analysis, sPLS, and rCCA. The use of dimensionality reduction methods can help identify potential underlying latent variables in the model. Cantini et al.[22] have demonstrated how different joint dimensionality reduction (jDR) methods such as RGCCA, MOFA, and intNMF can improve performance in multi-omic studies involving genomics, transcriptomics, and proteomics in classification and clustering, biomarkers identification, and biological annotation identification. Our results demonstrate that the output of MultiABLER can be used in different contexts for different downstream analysis applications.

Limitations of the study

In our analyses, we identified EigenMS to be the best suited normalization method for MultiABLER. As such, other data manipulation and normalization methods, such as k-nearest neighbours (kNN) imputation, mean normalization, and quantile normalization,[23–25] were not implemented in MultiABLER. These methods may be implemented into MultiABLER in the future to expand its flexibility to suite different LC-MS/MS experimental design. Finally, a recent package TidyMass[26] was published on R with a similar intent to simplify the process for LC-MS/MS analyses. While TidyMass has a strong focus on object-oriented pipeline and covers a large number of existing methods for LC-MS/MS data analysis, MultiABLER provides a specific workflow for Multi-ABLE-generated metabolomic and lipidomic analysis.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.106881.

ACKNOWLEDGMENTS

We thank members in Ho Lab, Wong Lab, and Stocker Lab for their critical comments. This work was supported in part by AIR@InnoHK administered by Innovation and Technology Commission of Hong Kong to JWKH, as well as National Health & Medical Research Council of Australia Program Grant 1052616 and Senior Principal Research Fellowship 1111632 to RS.

AUTHOR CONTRIBUTIONS

JWKH, RS, and ICHL conceived the ideas for the analytical pipeline and supervised the project. ICHL developed the method, implemented the R packages, and performed all the analyses. JWHW provided critical support in development and evaluation of the bioinformatics method. ST and RS developed the Multi-ABLE experimental method and produced the experimental data used in this study. ICHL and JWKH wrote the first draft of the manuscript. All authors read, contributed to, and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: August 19, 2022
Revised: March 13, 2023
Accepted: May 11, 2023
Published: May 13, 2023

REFERENCES

1. Stanstrup, J., Broeckling, C.D., Helmus, R., Hoffmann, N., Mathé, E., Naake, T., Nicolotti, L., Peters, K., Rainer, J., Salek, R.M., et al. (2019). The metaRbolomics toolbox in bioconductor and beyond. Metabolites *9*, E200. https://doi.org/10.3390/metabo9100200.

2. Smith, C.A., Want, E.J., O'Maille, G., Abagyan, R., and Siuzdak, G. (2006). XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal. Chem. *78*, 779–787. https://doi.org/10.1021/ac051437y.

3. Röst, H.L., Sachsenberg, T., Aiche, S., Bielow, C., Weisser, H., Aicheler, F., Andreotti, S., Ehrlich, H.-C., Gutenbrunner, P., Kenar, E., et al. (2016). OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat. Methods *13*, 741–748. https://doi.org/10.1038/nmeth.3959.

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

5. Alvarez-Jarreta, J., Rodrigues, P.R.S., Fahy, E., O'Connor, A., Price, A., Gaud, C., Andrews, S., Benton, P., Siuzdak, G., Hawksworth, J.I., et al. (2021). LipidFinder 2.0: advanced informatics pipeline for lipidomics discovery applications. Bioinformatics *37*, 1478–1479. https://doi.org/10.1093/bioinformatics/btaa856.

6. Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. (2010). MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinf. *11*, 395. https://doi.org/10.1186/1471-2105-11-395.

7. Alcoriza-Balaguer, M.I., García-Cañaveras, J.C., López, A., Conde, I., Juan, O., Carretero, J., and Lahoz, A. (2019). LipidMS: an R package for lipid annotation in untargeted liquid chromatography-data independent acquisition-mass spectrometry lipidomics. Anal. Chem. *91*, 836–845. https://doi.org/10.1021/acs.analchem.8b03409.

8. Koelmel, J.P., Kroeger, N.M., Ulmer, C.Z., Bowden, J.A., Patterson, R.E., Cochran, J.A., Beecher, C.W.W., Garrett, T.J., and Yost, R.A. (2017). LipidMatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. BMC Bioinf. *18*, 331. https://doi.org/10.1186/s12859-017-1744-3.

9. Smith, C.A., O'Maille, G., Want, E.J., Qin, C., Trauger, S.A., Brandon, T.R., Custodio, D.E., Abagyan, R., and Siuzdak, G. (2005). METLIN: a metabolite mass spectral database. Ther. Drug Monit. *27*, 747–751. https://doi.org/10.1097/01.ftd.0000179845.53213.39.

10. Shen, X., Wu, S., Liang, L., Chen, S., Contrepois, K., Zhu, Z.-J., and Snyder, M. (2022). metID: an R package for automatable compound annotation for LC−MS-based data. Bioinformatics *38*, 568–569. https://doi.org/10.1093/bioinformatics/btab583.

11. Karpievitch, Y. V., Stuart, T., & Mohamed, S. (2021). ProteoMM: multi-dataset model-based differential expression proteomics analysis platform (1.10.0). Bioconductor version: release (3.13). https://doi.org/10.18129/B9.bioc.ProteoMM.

12. Redestig, H., Fukushima, A., Stenlund, H., Moritz, T., Arita, M., Saito, K., and Kusano, M. (2009). Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. Anal. Chem. *81*, 7974–7980. https://doi.org/10.1021/ac901143w.

13. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

14. Pang, Z., Chong, J., Zhou, G., de Lima Morais, D.A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.É., Li, S., and Xia, J. (2021). MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Res. *49*, W388–W396. https://doi.org/10.1093/nar/gkab382.

15. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. (2009). Knime - the Konstanz information miner: version 2.0 and beyond. SIGKDD Explor. Newsl. *11*, 26–31. https://doi.org/10.1145/1656274.1656280.

16. Lê Cao, K.A., González, I., and Déjean, S. (2009). integrOmics: an R package to unravel relationships between two omics datasets. Bioinformatics *25*, 2855–2856. https://doi.org/10.1093/bioinformatics/btp515.

17. Liu, T., Salguero, P., Petek, M., Martinez-Mira, C., Balzano-Nogueira, L., Ramšak, Ž., McIntyre, L., Gruden, K., Tarazona, S., and Conesa, A. (2022). PaintOmics 4: new tools for the integrative analysis of multi-omics datasets supported by multiple pathway databases. Nucleic Acids Res. *50*, W551–W559. https://doi.org/10.1093/nar/gkac352.

18. Villaret-Cazadamont, J., Poupin, N., Tournadre, A., Batut, A., Gales, L., Zalko, D., Cabaton, N.J., Bellvert, F., and Bertrand-Michel, J. (2020). An optimized dual extraction method for the simultaneous and

accurate analysis of polar metabolites and lipids carried out on single biological samples. Metabolites *10*, 338. https://doi.org/10.3390/metabo10090338.

19. Talib, J., Hains, P.G., Tumanov, S., Hodson, M.P., Robinson, P.J., and Stocker, R. (2019). Barocycler-based concurrent multiomics method to assess molecular changes associated with atherosclerosis using small amounts of arterial tissue from a single mouse. Anal. Chem. *91*, 12670–12679. https://doi.org/10.1021/acs.analchem.9b01842.

20. Yu, H., Villanueva, N., Bittar, T., Arsenault, E., Labonté, B., and Huan, T. (2020). Parallel metabolomics and lipidomics enables the comprehensive study of mouse brain regional metabolite and lipid patterns. Anal. Chim. Acta *1136*, 168–177. https://doi.org/10.1016/j.aca.2020.09.051.

21. Chen, W., Tumanov, S., Fazakerley, D.J., Cantley, J., James, D.E., Dunn, L.L., Shaik, T., Suarna, C., and Stocker, R. (2021). Bilirubin deficiency renders mice susceptible to hepatic steatosis in the absence of insulin resistance. Redox Biol. *47*, 102152. https://doi.org/10.1016/j.redox.2021.102152.

22. Cantini, L., Zakeri, P., Hernandez, C., Naldi, A., Thieffry, D., Remy, E., and Baudot, A. (2021). Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. Nat. Commun. *12*, 124. https://doi.org/10.1038/s41467-020-20430-7.

23. Moorthy, K., Jaber, A.N., Ismail, M.A., Ernawan, F., Mohamad, M.S., and Deris, S. (2019). Missing-values imputation algorithms for microarray gene expression data. Methods Mol. Biol. *1986*, 255–266. https://doi.org/10.1007/978-1-4939-9442-7_12.

24. De Livera, A.M., Olshansky, M., and Speed, T.P. (2013). Statistical analysis of metabolomics data. Methods Mol. Biol. *1055*, 291–307. https://doi.org/10.1007/978-1-62703-577-4_20.

25. Bolstad, B.M., Irizarry, R.A., Åstrand, M., and Speed, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics *19*, 185–193. https://doi.org/10.1093/bioinformatics/19.2.185.

26. Shen, X., Yan, H., Wang, C., Gao, P., Johnson, C.H., and Snyder, M.P. (2022). TidyMass an object-oriented reproducible analysis framework for LC–MS data. Nat. Commun. *13*, 4365. https://doi.org/10.1038/s41467-022-32155-w.

27. Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., et al. (2019). Welcome to the tidyverse. J. Open Source Softw. *4*, 1686. https://doi.org/10.21105/joss.01686.

28. Adusumilli, R., and Mallick, P. (2017). Data conversion with ProteoWizard msConvert. Methods Mol. Biol. *1550*, 339–368. https://doi.org/10.1007/978-1-4939-6747-6_23.

29. Tautenhahn, R., Böttcher, C., and Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS. BMC Bioinf. *9*, 504. https://doi.org/10.1186/1471-2105-9-504.

30. Prince, J.T., and Marcotte, E.M. (2006). Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal. Chem. *78*, 6140–6152. https://doi.org/10.1021/ac0605344.

31. Fahy, E., Alvarez-Jarreta, J., Brasher, C.J., Nguyen, A., Hawksworth, J.I., Rodrigues, P., Meckelmann, S., Allen, S.M., and O'Donnell, V.B. (2019). LipidFinder on LIPID MAPS: peak filtering, MS searching and statistical analysis for lipidomics. Bioinformatics *35*, 685–687. https://doi.org/10.1093/bioinformatics/bty679.

32. Karpievitch, Y.V., Nikolic, S.B., Wilson, R., Sharman, J.E., and Edwards, L.M. (2014). Metabolomics data normalization with EigenMS. PLoS One *9*, e116221. https://doi.org/10.1371/journal.pone.0116221.

33. Strbenac, D., Mann, G.J., Ormerod, J.T., and Yang, J.Y.H. (2015). ClassifyR: an R package for performance assessment of classification with applications to transcriptomics. Bioinformatics *31*, 1851–1853. https://doi.org/10.1093/bioinformatics/btv066.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited Data		
LC-MS lipidomic dataset containing 35 mouse liver tissue sample collected by Chen et al. (2021), https://doi.org/10.1016/j.redox.2021.102152	EMBL-EBI Biostudies Database	https://doi.org/10.1016/j.redox.2021.102152 Biostudies accession nubmer: S-BSST1038
LC-MS lipidomic and metabolomic dataset containing 19 arterial tissue sample collected by Talib et al. (2019), https://doi.org/10.1021/acs.analchem.9b01842	EMBL-EBI Biostudies Database	https://doi.org/10.1021/acs.analchem.9b01842 Biostudies accession nubmer: S-BSST1039
Software and Algorithms		
LipidFinder	Github repository	https://github.com/ODonnell-Lipidomics/LipidFinder
MultiABLER	Github repository	https://github.com/holab-hku/MultiABLER

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be provided by the lead contact, Joshua W. K. Ho (jwkho@hku.hk).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- The data used in this study was derived from mouse liver samples collected by Chen et al.[21] and mouse arterial tissue collected by Talib et al.[19] The liver data is available at EMBL-EBI BioStudies (accession number EBI-BioStudies: S-BSST1038) and arterial tissue data is available at EMBL-EBI BioStudies (accession number EBI-BioStudies: S-BSST1039) (see key resources table).
- The workflow is implemented as an R package named MultiABLER which is available at the GitHub repository (https://github.com/holab-hku/MultiABLER). A tutorial to run the analysis is also included in the repository (https://htmlpreview.github.io/?https://github.com/holab-hku/MultiABLER/blob/main/tutorials/tutorial.html). A video tutorial is available on YouTube (https://youtu.be/7qnLvJaVU-I). Information on how to access the GitHub repository is provided in the key resources table. The packages used are: R packages (xcms V3.20.0, CAMERA V1.54.0, metid V1.2.25, ProteoMM V1.16.0, crmn V0.0.21, ClassifyR V3.3.10, limma V3.54.0), Python packages (LipidFinder V2.0.2).
- Any additional information required to reanalyse the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This study represents computational research and does not utilize experimental models.

METHOD DETAILS

To generate an easily reproducible workflow for Mult-ABLE, we first compared the different normalization methods for the data. Median normalization, CRMN, NOMIS, and EigenMS were compared and chosen for their easy implementation or availability from established packages in R. The lipidomic profile of liver tissue collected by Chen et al.[21] was used to evaluate the normalization methods. After identifying the best normalization method, the workflow was implemented in R with the package MultiABLER. The functionality of the package was demonstrated using the lipidomic and metabolomic profile of data from arterial tissue collected by Talib et al.[19]

Overview of the MultiABLER bioinformatics workflow

MultiABLER enables metabolomic and lipidomic data to be processed and analyzed in a common set of software packages. The overall workflow is shown in Figure 1. The MultiABLER workflow includes data pre-processing using XCMS, normalization and filtering against blank samples using LipidFinder, feature annotation using metID, and data normalization using EigenMS followed by median normalization. XCMS, metID and EigenMS have been implemented in the MultiABLER package, together with support functions to automatically read and write LipidFinder input and output. As such, all functions in the MultiABLER package are single-lined, and compatible for both lipidomic and metabolomic analyses, thus providing the user with a concise and unified workflow between metabolomic and lipidomic data, without the need of looking for individual programmes for each steps in the workflow, among a plethora of programmes. The functions and output of MultiABLER also follow the tidyverse[27] format and use the tidyverse packages, allowing easy data frame manipulation. These steps are described in detail in the following sections. An R workflow and tutorial on each function and their input and output can also be found at (https://htmlpreview.github.io/?https://github.com/holab-hku/MultiABLER/blob/main/tutorials/tutorial.html), and a video tutorial can be found at (https://youtu.be/7qnLvJaVU-I).

Data pre-processing

MSConvert[28] was first used to convert raw LC-MS/MS data into mzML files. XCMS was then used to identify and align the LC-MS/MS peaks. The MultiABLER package performed XCMS analysis by the *runXCMS* function. Briefly, *runXCMS* performed XCMS peak detection using the centWave algorithm,[29] alignment of chromatograms based on retention time usig the Obiwarp method,[30] chromatographic peak grouping based on peak density, and CAMERA adduct annotation. The output of *runXCMS* is an aligned and adduct-annotated feature table.

Feature annotation

Feature annotation is performed using metID provided by the *metid_annotate* function. MultiABLER includes two databases for metabolomic and lipidomic data. HMDB V5.0 is obtained using metID. HMDB entries that are found in LipidMaps Structure database (LMSD) are labeled lipidome, and the remaining non-lipid entries are grouped as metabolome. Both databases are included in the MultiABLER package. Based on the database, metID uses the mass-to-charge (*m*/*z*) ratio to annotate the features.

Data processing and normalization using MultiABLER

As the XCMS output includes mass spectrometry artifacts, contaminants and adducts, and XCMS lacks a function to filter these adducts and normalize the data against solvent blanks and quality control samples, LipidFinder[31] was used to filter the sample data for contaminants and normalize it to the controls. Because LipidFinder is implemented in Python, this was the only step of the pipeline that requires an external program that was not packaged in MultiABLER. The *peakfilter* module of LipidFinder was used to normalize the solvent blank controls and filter the artifacts and contaminants. The MultiABLER package provides read and write functions for producing LipidFinder compatible input and reading LipidFinder output back into R.

After the data was processed by LipidFinder, the feature table was imported back into R and MultiABLER was used to continue processing the data, including feature filtering, data imputation and log transformation. Features with a large number of missing values (set by the user) were removed, and the missing values were imputed using half of the global minimum. Finally, the data was log-transformed using base 2. Normalization was then performed using EigenMS[32] and median normalization. Finally, for features found in both the positive and negative mode from LC-MS/MS, the feature row with the highest median across the sample was used as the representation of that annotated features. Each step of the data manipulation and normalization was modularized in MultiABLER and made available as a function.

QUANTIFICATION AND STATISTICAL ANALYSIS

The normalized feature table was used for different statistical analyses. To analyze the output, we performed differential analysis using limma, sparse partial least squares projection (sPLS), and regularized canonical correlation analysis (rCCA) using mixOmics.[16] To build the limma model, mouse arterial tissue was grouped into non-lesion (NL) and lesion (L) containing tissue to build the design matrix. The limma model was then used to perform differential analysis. For sPLS projection, annotated lipidomic and metabolomic profiles were set as canonical variables and used to identify the PLS components in the data. Arrow plots and cluster image maps of the model were generated to visualize the results using mixOmics. rCCA was performed using the shrinkage approach to calculate the ridge penalties. Arrow plots were used to visualize results using mixOmics.

Evaluation using k-fold cross-validation

To examine the potential suitability of different normalization approaches for MultiABLER, and to evaluate the statistical analysis results from limma and mixOmics, we used 5-fold cross-validation to fit the data to a classification model and compared their classification performance using ClassifyR.[33] Briefly, each dataset underwent 5-fold cross-validation which was randomly repeated 20 times. Within each round of 5-fold cross-validation, internal feature selection was performed by selecting the top n features using ranked t-test, where n was tuned automatically by ClassifyR. The selected features were used to train a model using Diagonal Linear Discriminant Analysis (DLDA). The 5-fold cross-validation results were summarized using balanced accuracy and the area under the receiver operating characteristic (AUC) scored.

ADDITIONAL RESOURCES

MultiABLER tutorial: https://htmlpreview.github.io/?https://github.com/holab-hku/MultiABLER/blob/main/tutorials/tutorial.html.

MultiABLER YouTube tutorial: https://youtu.be/7qnLvJaVU-I.