Developing computational approaches to enable high-throughput LC-MS-based global metabolomics

Zhiqiang Pang

Institute of Parasitology

McGill University, Montreal, Canada

August 2023

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

Contents

Abstract VI
Résumé VII
Acknowledgements IX
Contribution to Original Knowledge X
Contribution of Authors XII
List of Figures XV
List of Publications XVIII
List of Main Abbreviations XIX
Chapter 1: Introduction and literature review 1
1.1 Background 1
1.2 Overview of metabolome 1
1.3 Metabolomics and system biology 2
1.4 Metabolomics raw data processing and algorithms 4
1.4.1 LC-MS data processing 5
1.4.2 LC-MS/MS data processing 13
1.5 MS/MS reference libraries 17
1.6 Functional analysis of metabolomics data 20
1.7 Rationale and objectives 22
1.8 Outline of achievements 25
Preface to Chapter 2 27
Chapter 2: MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics 28
2.1 Abstract 29
2.2 Introduction 29
2.3 Results 32
2.3.1. Peak identification benchmark case study 34
2.3.2. Algorithm reliability benchmark case study 35
2.3.3. Overall workflow evaluation using a large-scale clinical dataset 36
2.4 Discussion 40
2.5 Conclusions 42
2.6 Materials and methods 42
2.6.1. Peak picking optimization 42
2.6.2. Adaptive batch effort correction 46

2.6.3. Mummichog2 for pathway activity prediction 47
2.6.4. Benchmark case studies 48
2.7 Supplementary materials 49
Preface to Chapter 3 59
Chapter 3: MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights 60
3.1 Abstract 61
3.2 Introduction 62
3.3 Overview of Metaboanalyst 5.0 workflow 65
3.4 Raw data processing 66
3.5 Functional analysis of MS peaks 69
3.6 Meta-analysis of global metabolomics data 70
3.7 Multi-omics integrative analysis 71
3.7 Extended knowledge bases 72
3.7.1 Compound database 72
3.7.2 Metabolite sets 72
3.7.3 Pathway libraries 73
3.8 Other features 73
3.8.1 Enhanced visualizations 73
3.8.2 Improved compound name matching 73
3.8.3 Automated batch effect correction 74
3.8.4 Merging technical replicates 74
3.8.5 Supporting new input formats 75
3.8.6 Streamlined data analysis 75
3.9 Implementation 75
3.10 Comparison with other web-based tools 76
3.11 Conclusion 77
Preface to Chapter 4 80
Chapter 4: MetaboAnalystR 4.0: towards a unified LC-MS workflow for global metabolomics 81
4.1 Abstract 82
4.2 Introduction 83
4.3 Results 85
4.3.1 General workflow 85
4.3.2 Benchmarking and validation 89

4.3.3 Characterizing performance of compound annotation with standard mixtures 90
4.3.4 Effects of reference spectral databases 92
4.3.5 Evaluating false discovery rate 92
4.3.6 Characterizing unique metabolome of different types of blood samples 95
4.3.7 Evaluation on quantitative performance with serial dilutions 97
4.3.8 Biological interpretation of COVID-19 metabolomics data 99
4.3.9 Computational performance assessment 103
4.4 Methods and materials 103
4.4.1 Chemicals 103
4.4.2 Sample preparation of bloods 104
4.4.3 Sample preparation of serial dilutions 105
4.4.4 LC-MS/MS analysis 105
4.4.5 MS/MS spectra reference library curation 107
4.4.6 DDA data deconvolution algorithm 110
4.4.7 SWATH-DIA data deconvolution algorithm 113
4.4.8 Spectra consensus of replicates algorithm 113
4.4.9 Reference library searching and scoring algorithms 114
4.4.10 Neutral loss searching 115
4.4.11 Result export 115
4.4.12 Decoy spectra generation and null evaluation 116
4.4.13 MetaboAnalystR usage 116
4.4.14 MS-DIAL/MS-FINDER usage 117
4.4.15 MZmine usage 117
4.4.16 XCMS usage 118
4.4.17 SIRIUS usage 118
4.4.18 Computational performance assessment 118
4.4.19 Integration of MS/MS results into mummichog algorithm 119
4.4.20 Interfacing with other tools 119
4.5 Discussion 120
4.6 Conclusion 121
4.7 Supplementary materials 122
Preface to Chapter 5 152
Chapter 5: Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets 153

5.1 Abstract 154
5.2 Introduction 155
5.3 Results 157
5.3.1. Summary of different datasets and their clinical characteristics 157
5.3.2. Processing and overview of individual datasets 158
5.3.3. Metabolic pathways changes in COVID-19 patient 161
5.3.4. Identification of metabolic hot spots in COVID-19 162
5.3.5. Metabolic changes between mild-to-moderate and severe COVID-19 164
5.3.6 Exploration of metabolic perturbations in fatal COVID-19 165
5.4 Discussion 167
5.5 Methods and materials 170
5.5.1 Data curation 170
5.5.2 Patient classification 170
5.5.3 Raw spectra processing 170
5.5.4 Statistical analysis 171
5.5.5 Metabolic pathway analysis and meta-analysis 171
5.5.6 Global metabolic network visualization 171
5.5.7 Cluster heatmap analysis 172
5.6 Conclusion 172
5.7 Supplementary materials 173
Chapter 6: General Discussion 190
6.1 Brief summary of this thesis 190
6.2 More discussions on chapters 2-5 191
6.3 Strengths 196
6.4 Limitations 198
6.5 Future directions 199
Chapter 7: Conclusions and Future works 201
References: 203

Abstract

Liquid chromatography coupled with high-resolution mass spectrometry (LC-MS) has become a workhorse in global metabolomics studies with growing applications across biomedical and environmental sciences. However, outstanding bioinformatics challenges for global metabolomics data processing remain critical barriers to the wider adoption of this technology. These challenges include optimal raw spectral MS processing, tandem MS (MS/MS) spectral deconvolution, and accurate function analysis of global metabolomics. This thesis aims to address these challenges by developing multiple bioinformatic tools and platforms, and bridging auto-optimized raw spectra data processing to yield accurate functional insights.

Firstly, an auto-optimized raw LC-MS spectra processing workflow was developed in an R package (MetaboAnalystR 3.0) and implemented in MetaboAnalyst (v5.0) website with a user-friendly interface. This workflow has been demonstrated by multiple case studies as highly efficient and fast. Secondly, an ultra-fast and auto-optimized data-dependent acquisition MS/MS data and data-independent acquisition data processing workflow was implemented in a later version of MetaboAnalystR package (v4.0). The performance of quantification and qualification of MetaboAnalystR for LC-MS and LC-MS/MS data processing has been benchmarked against other popular tools, showing significant improvements in multiple aspects. Finally, functional analysis for global metabolomics has been enhanced by integrating retention time and MS/MS-based identifications into *mummichog* algorithm to improve the accuracy. Functional meta-analysis has also been developed to enable integration of multiple metabolomics datasets. The performance has also been demonstrated through COVID-19 case studies.

Overall, this thesis describes how MetaboAnalystR and MetaboAnalyst can be utilized to bridge global metabolomics raw spectral processing to accurate biological insights.

Résumé

La chromatographie liquide couplée à la spectrométrie de masse à haute résolution (LC-MS) est devenue un outil indispensable dans les études de métabolomique globale avec des applications croissantes dans les domaines biomédicaux et environnementaux. Cependant, d'importants défis bioinformatiques dans le traitement des données de métabolomique globale restent des obstacles critiques à une adoption plus large de cette technologie. Ces défis incluent le traitement optimal des spectres de spectrométrie de masse bruts, la déconvolution spectrale MS/MS (MS/MS) en tandem et l'analyse précise des fonctions de métabolomique globale. Cette thèse vise à relever ces défis en développant plusieurs outils et plateformes bioinformatiques pour relier le traitement automatique des données de spectres bruts à des informations fonctionnelles précises.

Tout d'abord, un flux de travail de traitement de spectres LC-MS bruts auto-optimisé a été développé dans un paquet informatique R (MetaboAnalystR 3.0) et implémenté sur le site web de MetaboAnalyst (v5.0) avec une interface conviviale. Ce flux de travail a été démontré par de multiples études de cas comme étant très efficace et rapide. Deuxièmement, un flux de travail de traitement ultra-rapide et auto-optimisé des données de MS/MS dépendantes des données et des données d'acquisition indépendantes a également été implémenté dans une version ultérieure du paquet informatique MetaboAnalystR (v4.0). Les performances de quantification et de qualification de MetaboAnalystR pour le traitement des données LC-MS et LC-MS/MS ont été comparées à d'autres outils populaires, montrant des améliorations significatives dans de multiples aspects. Enfin, l'analyse fonctionnelle pour la métabolomique globale a été améliorée en intégrant le temps de rétention et les identifications basées sur MS/MS dans l'algorithme *mummichog* pour améliorer la précision. Une méta-analyse fonctionnelle a également été développée pour permettre

l'intégration de multiples ensembles de données de métabolomique. Les performances ont également été démontrées par des études de cas de COVID-19.

Dans l'ensemble, cette thèse décrit comment MetaboAnalystR et MetaboAnalyst peuvent être utilisés pour relier le traitement des données spectrales brutes de la métabolomique globale à des informations biologiques précises.

Acknowledgements

Most work of this thesis was performed at Institute of Parasitology, McGill University. It involves tremendous help and support from XiaLab, collaborators and families.

First of all, I would like to sincerely express my deep gratitude to my supervisor, Prof. Jianguo Xia for his expertise in metabolomics field, highly-patient guidance, valuable suggestions and continuous support. I have grown and developed generally, including not only advanced scientific techniques, but also broad academic vision, critical thinking and courage to explore unknown world. All of these points are essentials as a scientist. Besides, he has also spent a lot of time reviewing and revising my manuscript, and as a result, my writing skills have been improved in a great scale.

I also would like to thank all my committee members, Dr. Shuzhao Li and Dr. Stéphane Bayen. Their professional and precious advice helps me improve my research in many aspects. I also would like to show my great appreciation to all my thesis examiners and collaborators for your hard work and help in many aspects.

I am also greatly thankful to Xia Lab members: Dr. Guangyan Zhou, Dr. Jasmine Chong, Dr. Charles Viau, Dr. Peng Liu, Yao Lu, Dr. Orcun Hacariz, Dr. Jessica Ewald, Dr. Le Chang, Dr. Xue Gu, Dr. Othman Soufan, etc. Without your help, my PhD research trip would be much less efficient and loss a lot of fun.

Finally, but the most importantly, I am deeply grateful for the support and companion from my wife, my parents, my sister and my grandparents. Your support has also been a consecutive motivation for my entire PhD period. I cannot imagine how much loneliness and hard time I have to face myself without your encouragement and stay.

Contribution to Original Knowledge

The proposed methods in this thesis aim to solve several essential bottlenecks by developing tools and platforms for LC-MS/MS raw data pre-processing, features annotation and functional analysis. Herein, we have enabled all users to obtain optimal metabolomics data processing results in an ultra-fast way without laborious programming or parameters' optimization work by manual. All significant novel contributions to metabolomics field are highlighted.

1. MetaboAnalystR 3.0, an R package was developed to implement an auto-optimized workflow for global metabolomics data processing. This package addresses three key bottlenecks by providing the following features:
    - An efficient parameters' optimization method used for LC-MS raw spectral processing to obtain optimal peak profiling results automatically.
    - An adaptive batch effect correction method used for large-scale metabolomics datasets with multiple batches to minimize the influence from batch effects.
    - Functional analysis algorithm, *mummichog*, has been updated by annotating putative compounds based on both *m/z* and retention time information to increase the accuracy.
2. MetaboAnalyst 5.0, a website aiming to further narrow the gap from raw data to functional insights for global metabolomics based on high-resolution mass spectrometry by enabling a user-friendly interface and providing more functional utilities for metabolomics data interpretation. There are three main features developed:
    - A LC-MS Spectra Processing module which offers an easy-to-use and intuitive interface-based pipeline that can perform automated parameter optimization and resumable analysis to significantly lower the barriers to LC-MS spectra processing.

- An enhanced Functional Analysis module which expands the previous *mummichog* algorithm to allow users to intuitively select any peak groups of interest and evaluate their enrichment of functions as defined by metabolic pathways and metabolite sets.
- A functional meta-analysis method developed to combine multiple global metabolomics datasets for comprehensive functional insights.

3. MetaboAnalystR 4.0, an R package was updated by enabling streamlined MS/MS spectral deconvolution and compound annotation coupled with comprehensive spectral reference databases and offering a further enhanced sensitive functional interpretation:
    - An auto-optimized DDA data deconvolution workflow to remove contamination signals in chimeric spectra.
    - A highly efficient SWATH-DIA data deconvolution pipeline.
    - Comprehensive MS/MS databases curated from all public database that can support diverse application purposes.
    - Accurate functional activity prediction by integrating LC-MS and MS/MS results.
4. Comprehensive meta-analysis of multiple COVID-19 datasets was performed to demonstrate LC-MS raw spectral processing and functional analysis pipelines and reveal biological insights associated with the pathogenesis of COVID-19.
    - The efficacy of computational pipeline for raw spectra processing, functional analysis and functional meta-analysis has been demonstrated.
    - Extensive dysregulations of amino acids metabolism, damage to the oxygen transport in red blood cells, exhaustion of endogenous immune bioactive metabolites and the suppression of physiological processes are related to the progression of COVID-19.

Contribution of Authors

The entire work described here was completed under the supervision of Dr. Jianguo Xia. This thesis is written by Zhiqiang Pang and comprises of four academic manuscripts. Zhiqiang Pang is the primary author of all chapters in this thesis. Chapter 1 is completely finished by Zhiqiang Pang. It mainly contains a comprehensive literature review, introductions of the background, motivation and objectives of this thesis. Chapter 2, 3 and 5 are works that have been published in journal *Metabolites*, *Nucleic Acids Research*, and *Metabolites* in 2020, 2021, and 2021 respectively. Chapter 4 is being submitted to *Nature Communications*. Chapter 6 is a general conclusion. The last chapter, chapter 7 is a brief discussion of the whole thesis and indicate the future directions.

In addition, this thesis is also supported by another Publication in *Nature Protocol*, which has formalized the parameters' auto-optimization workflow for LC-MS raw spectral data processing. This is protocol publication and not listed as an individual chapter in this thesis. But the protocol has further consolidated the work of this Chapter 2 and 3. All author of the protocol has also been included in this thesis.

The following authors contributed to one or more of the manuscripts in this thesis:

Zhiqiang Pang (Z.P.); Jianguo Xia (J.X.); Guangyan Zhou (G.Z.); Jasmine Chong (J.C.); Shuzhao Li (S.L); Niladri Basu (N.S.); Reza Salavati1 (R.S.); Charles Viau (C.V.); Lei Xu (L.X.); Jessica Ewald (J.E.); Le Chang (L.C.); Orcun Hacariz (O.H.); David Anderson de Lima Morais (D.M.); Michel Barrette (M.B.); Carol Gauthier (C.G.); Pierre-Étienne Jacques (P.E.)

The contributions of all co-authors to the manuscripts (chapters) are as follows:

1. Manuscript 1 (Chapter 2): Conceptualization, J.X.; Data curation, Z.P.; Formal analysis, Z.P. and J.C.; Funding acquisition, J.X.; Methodology, Z.P., J.C., S.L., and J.X.; Supervision, J.X.; Writing, original draft, Z.P. and J.C.; Review & editing, J.X. and S.L.
2. Manuscript 2 (Chapter 3): Conceptualization, J.X. and Z.P.; data curation, Z.P.; formal analysis, Z.P., J.C., G.Z., L.C.; funding acquisition, J.X.; methodology, Z.P., D.M., M.B., C.G., and J.X.; supervision, J.X.; writing, original draft, Z.P., J.C., G.Z.; review and editing, J.X., and N.B.
3. Manuscript 3 (Chapter 4): Conceptualization, J.X. and Z.P.; data curation, Z.P., N.B. and C.V.; formal analysis, Z.P; funding acquisition, J.X.; methodology, Z.P., L.X., C.V. and J.X.; supervision, J.X. and R.S.; writing, original draft, Z.P.; review and editing, J.X., S.L. and N.B.
4. Manuscript 4 (Chapter 5): Conceptualization, J.X.; data curation, Z.P.; formal analysis, Z.P., G.Z. and J.C.; funding acquisition, J.X.; methodology, Z.P., G.Z., J.C. and J.X.; supervision, J.X.; writing, original draft, Z.P.; review and editing, J.X. and J.C.
5. Manuscript 5 (Protocol Manuscript). Not included as a chapter in this thesis. Z.P., J.E., N.B. and J.X. prepared the manuscript. Z.P., G.Z., J.E., L.C., O.H. and J.X. contributed to the development and testing of the MetaboAnalyst. All authors read and approved the final manuscript.

List of Tables

Table 2.1 Qualitative peak picking results of the different tools using different settings 34

Table 2.2 The pathway enrichment results (top 20, Crohn's disease vs. non-IBD) generated by *mummichog* v1.0.8 and v2.0. 39

Table 2.3 Batch effect correction methods available in MetaboAnalystR 3.0.....47

Table S2.1 Optimized Parameters Summary of All Datasets 50

Table S2.2 Clinical Characteristics Summary of IBD Subjects 50

Table S2.3 *Mummichog* (v.1) Pathways (Top 20) of non-optimized IBD data (CD vs. nonIBD)51

Table S2.4 *Mummichog* (v.2) Pathways of non-optimized IBD data (CD vs. nonIBD).....52

Table S2.5 *Mummichog* (v.1) Pathways (Top 20) of optimized IBD data (CD vs. nonIBD).....53

Table S2.6 *Mummichog* (v.2) Pathways (Top 20) of optimized IBD data (CD vs. nonIBD).....54

Table 3.1 Comparison of MetaboAnalyst (versions 3.0-5.0) with other web-based tools.....78

Table 4.1 Summary of identified compounds by different tools (DDA, ESI^+).....89

Table 4.2 Comparison of computational performance of different tools.....101

Table S4.1 Summary of compounds in different database options 143

Table S4.2 Summary of identified compounds by different tools (DDA, ESI^-) 143

Table S4.3 Summary of identified compounds by different tools (SWATH-DIA ESI^+) 143

Table S4.4 Summary of identified compounds by different tools (SWATH-DIA ESI^-) 143

Table S4.5 Pathway enrichment results of polar metabolites datasets from three tools (SWATH-DIA) 144

Table S4.6 Pathway enrichment results of polar metabolites datasets from three tools (DDA).146

Table S4.7 Pathway enrichment results of lipids datasets from three tools 147

Table S4.8 Demographics of all subjects involved in current study 148

Table S4.9 Chromatographic conditions, gradient procedure instrumental settings 148

Table S4.10 Parameters of mass spectrometers for both MS1 and MS/MS 149

Table S4.11 Design of SWATH-DIA for different modes for blood Samples.....150

Table 5.1 Summary of the seven datasets and the corresponding COVID-19 patient classifications 159

Table S5.1 Classification Standards for Different Severities of COVID-19 174

Table S5.2 Technical Information of all datasets included in this study 174

Table S5.3 Clinical Demographics Characteristics of All Samples 175

Table S5.4 Optimized parameters of all datasets for raw spectral processing 176

List of Figures

Figure 1.1 Overall design of the project presented in this thesis 24

Figure 2.1 MetaboAnalystR 3.0 provides an optimized workflow for global metabolomics 32

Figure 2.2 Time consumed by One Variable at A Time (OVAT), Isotopologue Parameter Optimization (IPO), MetaboAnalystR, and AutoTuner for parameter optimization on three different datasets 33

Figure 2.3 Assessment of the performance of different tools utilizing NIST 1950 serum dilution series 36

Figure 2.4 Performance evaluation using Inflammatory Bowel Disease (IBD) data 37

Figure 2.5 The selection process of regions of interest (ROIs) that are enriched for true peak signals 44

Figure S2.1 Bar plots of *mummichog* pathway enrichment results applied on Crohn's Disease patients versus nonIBD controls 55

Figure S2.2 Scatter plots of the *mummichog* pathway enrichment results applied on ulcerative colitis patients versus healthy controls 56

Figure S2.3 TICs of benchmark 1 (Known standard data) before (top) and after (bottom) optimization. 57

Figure S2.4 TICs of benchmark 2 (NIST series) before (top) and after (bottom) optimization ..57

Figure S2.5 TICs of benchmark 3 (IBD data) before (top) and after (bottom) optimization 58

Figure 3.1 Overview of MetaboAnalyst v5.0 workflows 65

Figure 3.2 Example outputs from several new features of MetaboAnalyst v5.0.....68

Figure 4.1 Implementation of MetaboAnalystR for LC-MS/MS data processing and biological interpretation 87

Figure 4.2 Validation of MetaboAnalystR with standard mixtures 90

Figure 4.3 Validation of MetaboAnalystR with standard mixtures and false discoveries 94

Figure 4.4 Comparison of chemical identification from different blood samples 96

Figure 4.5 Evaluation of quantitative and qualitative performance based on serial dilutions 98

Figure 4.6 Interpretation of biological insights of COVID-19 100

Figure S4.1 Workflow of spectral consensus of replicates 122

Figure S4.2 Validation results of MetaboAnalystR with simple standard mixtures in ESI^- mode 123

Figure S4.3 Analysis results of complicated standards mixture 124

Figure S4.4 Workflow to generate decoy spectra data (SWATH-DIA) 125

Figure S4.5 Summary of falsely identified compounds from null evaluations in ESI^- mode 125

Figure S4.6 PCA results of metabolomic profiles of four different modes analyzed by MetaboAnalystR 126

Figure S4.7 Heatmap of blood samples from C18 ESI^- mode 127

Figure S4.8 Heatmap of blood samples from HILIC ESI^+ mode 128

Figure S4.9 Heatmap of blood samples from HILIC ESI^- mode 129

Figure S4.10 Statistics of MS features detected by different tools. Features are classified as "Generic Features" and "Unique Features" 130

Figure S4.11 Summary of chemical classification of identified compounds from MetaboAnalystR 131

Figure S4.12 Evaluation of quantitative performance based on serial dilutions 132

Figure S4.13 Serial dilution heatmaps. Heatmaps of all MS features detected by MetaboAnalystR under different modes 133

Figure S4.14 Statics of compound identification by different tools 134

Figure S4.15 Scatter plot of pathway enrichment of datasets analyzed by MetaboAnalystR ...135

Figure S4.16 Scatter plot of pathway enrichment of datasets generated by MSDIAL / MSFinder 136

Figure S4.17 Scatter plot of pathway enrichment of datasets generated by MZmine / SIRIUS or XCMS / SIRIUS 137

Figure S4.18 Venn diagram of pathway analysis results from different datasets 138

Figure S4.19 Comparison of computational efficiency of different tools 139

Figure S4.20 Comparison of computational efficiency of different datasets 140

Figure S4.21 Steps involved in preparing a serial dilution 141

Figure S4.22 Statistics of compounds in pathway reference library 141

Figure S4.23 Schema of MS/MS reference library for MetaboAnalystR 142

Figure 5.1 The workflow diagram of our data curation process and analysis strategy 158

Figure 5.2 Overview of the separation patterns between COVID-19 and healthy controls (HCs) across the seven datasets 160

Figure 5.3 Pathway analysis and meta-analysis between COVID-19 and healthy controls (HC) across the seven datasets 161

Figure 5.4 Overview of potentially perturbed metabolites and extracted metabolic pathways based on the seven datasets 163

Figure 5.5 Metabolic pathway analysis and cluster heatmap analysis between mild-to-moderate (MM) and severe groups 164

Figure 5.6 Pathway analysis and cluster heatmap analysis between severe and fatal groups 166

Figure S5.1 The spearman correlation analysis on the onset time (days) with the metabolites in the significantly perturbed pathways 177

Figure S5.2 Cluster heatmap analysis between Covid and HC groups of Dataset A1 in negative and positive mode. 178

Figure S5.3 Cluster heatmap analysis between Covid and HC groups of Dataset A2 in negative and positive mode 179

Figure S5.4 Cluster heatmap analysis between Covid and HC groups of Dataset A3 in negative and positive mode 180

Figure S5.5 Cluster heatmap analysis between Covid and HC groups of Dataset C1 in negative and positive mode 181

Figure S5.6 Cluster heatmap analysis between Covid and HC groups of Dataset C2 and C3 ..182

Figure S5.7 Cluster heatmap analysis between Covid and HC groups of Dataset B1 in negative and positive mode 183

Figure S5.8 Overview of perturbed pathways in COVID-19 across datasets for comparison between MM (Mild-to-moderate) and Severe COVID-19 184

Figure S5.9 The metabolic pattern between MM and Severe of dataset C3.....185

Figure S5.10 The metabolic pattern between MM and Severe of dataset C1 and C2 186

Figure S5.11 Overview of perturbed pathways in COVID-19 across datasets for comparison between Severe and Fatal COVID-19.....187

Figure S5.12 The PRISMA Flow Diagram 188

Figure 6.1. Overview of MetaboAnalyst and MetaboAnalystR 195

List of Publications

1. **Pang Z**, Chong J, Li S, Xia J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. Metabolites. 2020;10(5):186.

2. **Pang Z**, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Research. 2021;49(W1):W388-W96.

3. **Pang Z**, Zhou G, Chong J, Xia J. Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets. Metabolites. 2021;11(1):44.

4. **Pang Z**, Zhou G, Ewald J, Chang L, Hacariz O, Basu N, et al. Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nature Protocols. 2022;17(8):1735-61.

5. Lu Y, **Pang Z**, Xia J. Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. Briefings in Bioinformatics. 2022;24(1).

6. Liu P, Ewald J, **Pang Z**, Legrand E, Jeon YS, Sangiovanni J, et al. ExpressAnalyst: A unified platform for RNA-sequencing analysis in non-model species. Nature Communications. 2023;14(1):2995.

7. Lu Y, Zhou G, Ewald J, **Pang Z**, Shiri T, Xia J. MicrobiomeAnalyst 2.0: comprehensive statistical, functional and integrative analysis of microbiome data. Nucleic Acids Research. 2023;51(W1):W310-W8.

8. Zhou G, **Pang Z**, Lu Y, Ewald J, Xia J. OmicsNet 2.0: a web-based platform for multi-omics integration and network visual analytics. Nucleic Acids Research. 2022;50(W1):W527-W33.

List of Main Abbreviations

Abbreviation	Full Spelling
LC	Liquid Chromatography
GC	Gas Chromatography
MS	Mass Spectrometry
MS/MS	Tandem Mass Spectrometry
HRMS	High-Resolution Mass Spectrometry
H-NMR	Proton Nuclear Magnetic Resonance
UPLC	Ultra-Performance Liquid Chromatography
HPLC	High-Performance Liquid Chromatography
Q/E	Q-Exactive
TOF	Time-of-Flight
ESI	Electrospray Ionization
NL	Neutral Loss
DDA	Data-Dependent Acquisition
DIA	Data-Independent Acquisition
SWATH	Sequential Window Acquisition of all Theoretical Mass Spectra
C18	Octadecylsilane
HILIC	Hydrophilic interaction chromatography
ROI	Regions of Interest
CWT	Continuous Wavelet Transform
ADAP	Automated Data Analysis Pipeline
IPO	Isotopologue Parameter Optimization
HMDB	Human Metabolome Database
KEGG	Kyoto Encyclopedia of Genes and Genomes
GNPS	Global Natural Product Social Molecular Networking
MINEs	Metabolic In silico Network Expansions
MoNA	Massbank of North America
T3DB	The Toxin and Toxin Target Database
API	Application Programming Interface
QC	Quality Control
SM	Standard Mixture
EIC	Extracted Ion Chromatogram
TIC	Total Ion Chromatogram
BPI	Base Peak Intensity
OVAT	One Variable at A Time
RT	Retention Time
EC	Empirical Compounds

NIST	National Institute of Standards and Technology
DoE	Design of Experiment
CV	Coefficient of Variation
RI	Reliability Index
PCA	Principal Component Analysis
PLS-DA	Partial Least Squares - Discriminant Analysis
OPLS-DA	Orthogonal Projections to Latent Structures Discriminant Analysis
QC-RLSC	Quality Control-robust LOESS signal correction
QS	Quality Score
GR	Gaussian Peaks Ratio
QcoE	Quality Coefficient
RCS	Retention time Correction Score
LIP	Low-Intensity Peaks
MSEA	Metabolite Set Enrichment Analysis
GSEA	Gene Set Enrichment Analysis
DSPC	Debiased Sparse Partial Correlation
SOM	Self-Organizing Map
SMILES	Simplified Molecular-Input Line-Entry System
PubChem CIDs	PubChem Compound Identifiers
PubChem SIDs	PubChem Substance Identifiers
TCA	Tricarboxylic Acid
WHO	World Health Organization
COVID-19	Coronavirus Disease 2019
SARS-CoV-2	Severe Acute Respiratory Syndrome Coronavirus 2
IBD	Inflammatory Bowel Disease
CD	Crohn’s Disease
HC	Healthy Controls
MM	Mild to Moderate
SLURM	Simple Linux Utility for Resource Management
BMI	Body Mass Index

Chapter 1: Introduction and literature review

1.1 Background

The emergence of various "omics" techniques has revolutionized many areas of systems biology (1). Owing to the advances of high-throughput sequencing technologies, genomics and transcriptomics have become crucial in deciphering the biological functions of genes and transcripts. Meanwhile, post-genomics biochemistries, such as proteomics and metabolomics, have focused on studying the systematic changes in molecules of different weight scales (2, 3). Specifically, proteomics investigates the interactions, function, composition, and structures of proteins and their cellular activities (4, 5), while metabolomics is a comprehensive study of small molecules or metabolites (3).

1.2 Overview of metabolome

The metabolome is the complete set of small molecule metabolites present in a biological system, such as cells, tissues, and organisms (6). These metabolites include sugars, amino acids, lipids, and other organic molecules that are involved in cellular processes such as energy production, signaling, and biosynthesis. The composition of human metabolome can vary depending on factors such as diet, genetics, environmental exposure and medications (7, 8).

The metabolome is a critical component of biological systems, with implications for understanding health and disease. It plays a vital role in cellular processes such as energy production, signaling, and biosynthesis, and changes in the metabolome have been linked to a variety of diseases and disorders, including cancer, diabetes, and neurological disorders (9, 10). In contrast to the genome, which provides information on potential biological events, the metabolome directly reflects what

has occurred in a biological system (11). In a brief summary, the metabolome represents a critical component of biological systems, with implications for understanding health and disease.

In recent years, exposomics has gained prominence as a means to investigate environmental exposures and their potential effects on biological individuals. Similarly, exposomics primarily focuses on the comprehensive range of low-weighted compounds and metabolites present in both biological and environmental samples, such as soil, water, and air (12). It offers direct insights into the metabolic activity of organisms and the overall state of ecosystems. Consequently, it has broad applications across various fields, including agriculture, environmental toxicology, and the study of environmental pollution and microbial interactions (13). By characterizing the exposome, researchers can gain a better understanding of the impact of environmental factors on ecosystems and develop strategies to preserve ecological balance and improve human health.

1.3 Metabolomics and system biology

Metabolomics is the study of the metabolome, which involves the identification and quantification of metabolites in biological samples using analytical techniques such as gas/liquid chromatography-mass spectrometry (GC-MS/LC-MS) and nuclear magnetic resonance spectroscopy (NMR) etc. The metabolome is complex and presents significant challenges for metabolomics research, including the large number of metabolites that need to be annotated, as well as the wide range of metabolite concentrations present in biological samples (14).

Metabolomics encompasses two primary approaches: targeted metabolomics and untargeted metabolomics (15, 16). Targeted metabolomics measures the levels of specific metabolites, typically using internal standards and specific mass spectrometry instruments for accurate quantitation. In contrast, untargeted metabolomics aims to measure all molecules that ionize within

a specific range of mass values, providing broader global coverage of the metabolome (15). While targeted metabolomics could offer better quantitation if it is designed for an absolute quantification, untargeted metabolomics has been more widely used and is now considered the main workhorse for metabolomics research. Thus, in this thesis, terms "metabolomics" and "global metabolomics" will refer specifically to untargeted metabolomics.

Metabolomics employs a range of analytical techniques to detect small molecules (<1,500 Da) and their interactions within a biological system, including blood (11, 17), urine (18), feces (19), sputum (20), and even tissues (21). LC-MS and GC-MS are the most commonly used techniques in metabolomics (22). While traditional proton nuclear magnetic resonance (H-NMR) based metabolomics usually show a higher quantitative capacity, MS-based platforms, particularly high-resolution MS (HRMS), offer consistently higher sensitivity and precision (23). GC-MS is typically used to separate volatile and thermally stable or easily derivatized compounds, which is its major limitation (24). In comparison, LC-MS displays wider applicability across almost all organic compounds in contrast to GC-MS, and has been implemented extensively, becoming a primary technique for metabolomics studies (25, 26). In addition to chromatogram-coupled MS, direct injection MS and flow injection MS have emerged as useful approaches for large-scale studies and have shown high consistency with LC-MS to some extent (27). These new techniques have expanded the scope of metabolomics research and allowed for the analysis of a broader range of metabolites in complex matrices of diverse samples.

In brief, metabolomics enables the systematic identification and quantitation of all metabolites present in a biological or environmental system, and is increasingly being used to comprehensively illustrate various events, including responses to disease and environmental stress (28, 29). When combined with other systems biology techniques, such as genomics, transcriptomics, and

proteomics, metabolomics can help reconstruct the functionalities of a system and further aid in deciphering the mechanisms or discover the biomarkers of various biological processes, including different diseases (30, 31).

1.4 Metabolomics raw data processing and algorithms

Despite the significant advancements in MS instrumentation that allow for high-throughput sample acquisition, high-quality data processing remains a challenge for untargeted metabolomics. Raw LC-MS data of a single sample consists of a series of MS and/or MSn spectra scans that form a three-dimensional entity comprising retention time (RT), *m/z*, and intensity, but is inhomogeneous in nature. The primary task in LC-MS data processing is to identify molecular ion traces (metabolic features) in this entity, followed by finding correspondences to other features within the same sample (e.g., isotopes, ion adducts) and across samples. Detection and quantification for each potential ion feature on this entity requires different algorithmic steps. The first and most important one is "peak picking" (32), which extracts and integrates the signals originating from each sample.

Different mass spectrometers produce spectra with distinct characteristics, making it challenging to identify the most effective data processing strategy. Poor signal-to-noise ratios for low-abundance metabolites, detector noise, and the presence of various peaks from isotopes, contaminants, and in-source degradation products can complicate peak picking (33). Over the past few decades, several computational algorithms have been developed and integrated into different software programs, such as XCMS (34, 35), MZmine (36-38), MS-DIAL (39, 40), OpenMS (41), apLCMS (42). Each algorithm initiates peak detection from different dimensions or perspectives. The computational performance and sensitivity of different tools are significant.

LC tandem mass spectrometry (LC-MS/MS) is a widely used platform in untargeted metabolomics, enabling the separation of thousands of metabolites and providing fragmentation patterns of LC-MS features/ions to identify chemical structures (32). MS/MS acquisition involves data-dependent acquisition (DDA) and data-independent acquisition (DIA). Multiple algorithms have been implemented to improve the accuracy of MS/MS-based compound identification. In the following sections, we provide a systematic and concise summary of the internal mathematical mechanisms of these algorithms for both LC-MS and LC-MS/MS data processing.

1.4.1 LC-MS data processing

1.4.1.1. From retention time dimension

A commonly used and easily understood approach for peak detection is based on the RT dimension, since analytes are eluted chromatographically from the column with time. Correspondingly, ion signals are acquired from the time domain. Thus, the raw data file is organized from the RT dimension as a series of scans in most open-source format of mass spectral data, such as mzML and mzXML (43).

The most conventional method for peak detection is the "*MatchedFilter*," which was first adopted by XCMS in 2006 (35). This method utilizes a mass slice-based peak detection and matching approach. Specifically, it divides the MS data into small slices along the *m/z* dimension (typically with 0.1 *m/z* width), and then attempts to overlap adjacent slices to generate clean signals. Next, a second-derivative Gaussian model is used to filter the slices and identify potential peaks across the entire RT domain of the slice. The zero-crossing points are used as the peak borders. To improve the precision for HRMS, a manually defined signal-to-noise ratio and intensity-weighted mean are used to generate mass slices (35).

Similarly, MS-DIAL also employs mass slices binning-based feature detection to extract the peak spots, which are the peaks being referred to (40). However, MS-DIAL applies a different approach than the "*MatchedFilter*" algorithm of XCMS. MS-DIAL uses a linearly weighted smoothing average model to reduce the noise of the data points. This process returns the detected peaks across the RT domain (44). Nonetheless, one major drawback of this method is determining an optimal bin size, which can significantly impact the peak shape (if too small) or cause features to be buried by adjacent high-intensity features (if too large) (45). Besides, computing speed is also a bottleneck for this method.

To address the limitations of the mass slices-based peak picking method, XCMS developed a new algorithm called *centWave*. This algorithm uses a combination of density-based detection of Regions of Interest (ROIs) in the *m/z* domain and a Continuous Wavelet Transform (CWT)-based approach for chromatographic peak resolution (45). Initially, a series of *m/z* lists are generated based on the *m/z* values of the first scan and then extended along the scan order, consistent with the RT domain, to generate all ROIs. Any ROIs that cannot be continuously extended are discarded. Next, a CWT model is applied as a stretchable wavelet model to adapt to different peak widths, with the help of various noise removal parameters. Finally, the peak centroid information is computed based on the weighted mean of ions.

The advantage of *centWave* is that it can adapt to different peak widths, providing a self-adjusted peak detection model. However, a major disadvantage is that dozens of parameters need to be manually defined, requiring experience to produce optimal results unless an additional software is used to optimize the parameters first (46). Another popular raw spectral processing software, MZmine, has also implemented *centWave*. However, the *centWave* used by MZmine differs slightly from the one used by XCMS (47). The difference lies mainly in the generation of ROIs

for CWT fitting. In a brief summary, XCMS adds the *m/z* points of the entire scan into a ROI if the *m/z* tolerance meets the criteria, while MZmine only considers the *m/z* points with near intensity. This modification seems to produce similar peak picking results, although not entirely consistent (47).

In addition to its *matchedFilter* and *centWave* algorithms, XCMS has also integrated *Massifquant* for peak picking. This algorithm uses a Kalman filter model to address the issue of heteroscedasticity in *m/z* variance (48). Essentially, a Kalman Gain is used to search for peaks across the RT domain while simultaneously evaluating all potential *m/z* centroids to avoid any missing. This approach solves the potential *m/z* merging issue of *centWave* when generating the ROIs list but tends to be oversensitive and return a lot of noise. Although it requires fewer parameters than *centWave*, this method is rarely used. *Massifquant* originated from *TracMass* (49), which has been upgraded and replaced by *TracMass2* (50). *TracMass2* uses a similar concept to ROIs and CWT, but it provides real-time graphical feedback to users for parameter exploration. However, the mathematical abstraction of TracMass2's parameters makes it difficult to understand their meaning from an analytical chemistry perspective.

1.4.1.2. From m/z dimension

RT-oriented approaches were initially developed for low-resolution mass spectrometry and have since evolved to adapt to HRMS. One specific strategy for HRMS peak detection from the *m/z* dimension is implemented in the R package, apLCMS (42). In detail, the approach taken by apLCMS involves breaking down the conventional data RT-based data structure and re-ordering all *m/z* centroids from highest to lowest (e.g., from 1,500 to 50). The differences in *m/z* values are then calculated between all adjacent *m/z* values, resulting in the differences between all *m/z* neighbors being made up of two distinct components. The first component is the differences

between *m/z* values in the same peak, which is caused by instrumental variation and tends to be quite small. The second component is the *m/z* differences between different peaks or noises, which are usually larger than the first component. A mixed distribution model is then used to separate these components in the following steps.

Once the first component is extracted, different *m/z* points can be easily grouped. Kernel density estimators are used in both the *m/z* and RT domains respectively to extract all potential peaks. A *run-filter* model is then applied to further check the continuity and filter noise. Finally, a kernel smoother and pseudo-likelihood model are used to determine the location of features and resolve potential overlapping issues between peaks with extremely similar *m/z* values and neighboring RTs. While this approach provides a unique and reasonable solution for HRMS, it is not compatible with data generated by relatively low-resolution MS spectrometers for MS feature detection, which is a major disadvantage. However, an automated and resumable processing workflow developed by apLCMS can reduce the heavy computational burden when users adjust the parameters of peak picking.

1.4.1.3. From intensity dimension

FeatureFinderMetabo (51) is a software tool developed and embedded in OpenMS (41) that differs from conventional RT and *m/z* domain-oriented approaches by focusing on the intensity domain. Initially, all *m/z* centroids are re-ordered based on intensity, from highest to lowest. Then, the *m/z* with the highest intensity is selected and extended alongside the RT domain by aggregating similar *m/z* values. A heteroscedastic model is used to estimate the *m/z* error in real-time and automatically adapt to the specific spectra, with the aggregated *m/z* values being excluded from the following steps. These steps are performed recursively until all *m/z* centroids are exhausted. This approach

provides a unique solution, but may be not suitable for low resolution MS data due to the estimation on *m/z* error and RT range.

Unlike the other algorithms, *FeatureFinderMetabo* fits the peaks with a local regression model with a polynomial of degree two rather than a Gaussian model. This smoothing approach does not make any assumptions about the shapes of chromatographic peaks and allows for high sensitivity in detecting low-intensity peaks. The peaks with similar *m/z* values and RTs will split from the minima between the two maximums. Finally, the mass traces or features detected from the intensity domain are kept for the following feature assembly process (described later).

Similar to the algorithm used in OpenMS, MZmine has developed an intensity domain-oriented approach called the Automated Data Analysis Pipeline (ADAP). This method starts with the most intense *m/z* centroid and extends along the RT in both directions (52). However, what significantly differentiates ADAP from OpenMS is the subsequent steps. Instead of using the LOWESS model to fit potential peaks, ADAP adopts CWT fitting, which is quite similar to *centWave*. Ridgeline detection is used to determine the boundaries and locations of all peaks. A noise estimation and S/N ratio strategy has also been developed to remove potential noise.

There are several other software tools available for peak picking, including commercial ones such as Compound Discovery, MarkerLynx, and Progenesis QI. Despite these commercial tools have been used for this task, they are not open-source, and their detailed functional mechanisms are unknown, making it impossible to fully evaluate their performance.

In summary, several algorithms have been developed over the past decades to handle peak picking in MS data analysis. Among these algorithms, *centWave* has shown compatibility towards diverse

MS platforms, although it requires careful manual optimization of numerous parameters to achieve an optimal peak picking result.

1.4.1.4. Peak Alignment and gap filling

Chromatographical deviation and mass drift are common issues in metabolomics analysis, often requiring grouping of peaks across the samples. To address this challenge, XCMS has developed a kernel density estimator-based "grouping" method (35), which uses a non-linear correction alignment approach to remove RT deviation. This method utilizes a local regression fitting approach, loess, which outperforms traditional linear correction methods and eliminates the need for internal standards, which can obscure real features (35).

Another alignment algorithm, *Obiwarp*, uses a dynamic time warping model to achieve optimal correction results in proteomics data alignment (53). MZmine initially developed a two-dimensional alignment method, called "*Join Aligner*", but it cannot handle non-linear deviation (36, 37). To address this limitation, a "Random Sample Consensus Aligner" was implemented in MZmine (36). However, MS-DIAL still uses the Join Aligner from MZmine for peak alignment, despite four versions of software evolution (39, 40). Besides, apLCMS using a kernel density estimator to align the peaks across the RT domain, while *TracMass2* is using a P-splines based clustering and warping method to handle this issue (50). Peak alignment for metabolomic data processing is missing in OpenMS. apLCMS and TracMass2 employ kernel density estimator (42) and P-splines based clustering and warping methods, respectively, to align peaks across the RT domain (50). Notably, peak alignment is currently absent in OpenMS for metabolomic data processing.

Overall, various algorithms have been integrated with different peak picking methods in different software tools to address the peak alignment issue. The quality of peak alignment plays a critical

role in subsequent data analysis steps, as it can affect the stability of peak groups and data point distribution. Suboptimal parameters or algorithms used in peak alignment may result in poor alignment quality, leading to high variability of peaks across different samples. Nevertheless, peak grouping coupled with the loess or *Obiwarp* method is generally capable of handling both linear and non-linear cases, and has been widely utilized by the metabolomics community.

In addition, metabolites with low concentrations may not be detected by peak picking algorithms or may not meet peak shape criteria, such as the Gaussian model, resulting in a zero entry for that peak in a particular sample in the feature table produced by peak picking and alignment steps. This loss of information could be significant as valuable data may be overlooked. To address this issue, gap filling methods have been developed, which recover weaker signals. Effective gap filling strategies have been implemented in apLCMS, XCMS, and MS-DIAL (35, 40, 42). Firstly, a RT boundary is established based on the picked peaks and alignment results. Next, signals within specific regions of the time domain are extracted. Finally, the peak with the most consistent characteristics to previously picked peaks is used to replace the zero entry (42). This simple yet effective strategy enables the recovery of valuable information that might have been missed due to low metabolite concentrations.

1.4.1.5 Peak Annotation and identification

The mass spectra obtained from LC-MS are complex due to the presence of numerous adducts, isotopologues, dimers, and fragments. Consequently, identification of the molecular ion, which is not always the peak with the highest intensity in the MS spectrum, can be challenging (24). Consequently, a large number of features, which can be several-fold higher due to the aforementioned factors, can be detected and aligned into a feature table. This often leads to an overestimation of the actual number of compounds present (54, 55). Although the percentage of

unique metabolites among all detected features varies widely between studies, ranging from 3% to over 50%, it is well-established that a tool capable of annotating potential relationships among the detected features is critical (56-58).

To address the challenge of annotating complex mass spectra generated by LC-MS, several algorithms have been developed and integrated into the peak pre-processing pipeline. One of the widely used algorithms is the CAMERA R package, which works in conjunction with XCMS for the annotation of adducts and isotopes by considering the ratio of peaks and feature similarities (54). Another newly developed R package, CliqueMS, has been reported as more powerful compared to CAMERA (55). While CliqueMS can directly communicate with the results from XCMS, its annotation mechanism is different from CAMERA. However, a major limitation of CliqueMS is its inability to process data from multiple samples.

In addition to CAMERA and CliqueMS, *FeatureFinderMetabo* from OpenMS provides an alternative solution for isotope annotation, which is based on millions of empirical and mimic mass traces instead of the compounds from the real world (51). But this approach is providing an alternative method to do the isotope annotation. Several other tools, like MS-FLO (59), MSClust (60), MetAssign as well as the algorithm in MZmine mainly rely on intensity correlation analysis.

There are several tools available for direct compound identification from MS1 level. For example, ProbMetab employs a Bayesian probabilistic model and knowledge database of biochemical reactions to achieve basic chemical identification (61). However, the limited size of the chemical database restricts the practical application of this approach. Another tool, xMSannotator, modularizes all features based on the intensity correlations among them and uses a kernel density estimator. The features are then matched with different compound databases, including HMDB,

KEGG, and T3DB, with the assistance of metabolic knowledge-based networks. Finally, a scoring rule is applied to evaluate the possibilities of all candidate compounds for a specific feature (62).

In summary, the aforementioned tools offer a range of functionalities to annotate and identify MS1 features, thereby reducing redundancy. However, in order to enhance accuracy and minimize mismatching results, chemical identification of MS features based on MS/MS spectrum, together with reference library is often necessary.

1.4.2 LC-MS/MS data processing

The primary challenge in LC-MS-based metabolomics is identifying the structural identities of the targeted features. Metabolic profiling can be performed in either MS1 or MS2 (MS/MS) mode. MS1 profiling provides a complete coverage on the entire metabolome, whereas MS2 captures structural information from the fragmentation pattern of features. Two commonly used MS/MS fragmentation approaches for LC-MS are data-dependent acquisition (DDA) and data-independent acquisition (DIA) (63, 64).

DDA is usually performed by acquiring the top intensity precursors (65). DDA method enables clear association between precursors and their fragmentation patterns. In contrast, DIA involves a series of acquisition cycles, each comprising a full scan plus fragmentation of all precursors in the subsequent MS/MS scan(s). The MS/MS scan can be a full range (all ion fragmentation mode, AIF) or a sequential window acquisition of all theoretical mass spectra (SWATH) (66). Due to the absence of a direct relationship between precursors and fragments, deconvolution on MS/MS spectra data is always required for DIA.

The utilization of MS/MS acquisition with the DIA method involves a series of cycles, which has been reported to yield superior quantitative precision and MS2 spectrum coverage compared to

other methods (67). Theoretically, DIA methods could potentially acquire all fragment ions for all precursors simultaneously in each cycle to enhance the coverage of detectable compounds (40). However, the SWATH-DIA spectral acquisition approach typically employs a wide isolation window (> 15*m/z*), which disrupts the association between precursors and their corresponding MS/MS fragments and may introduce some contaminants into the spectral data, resulting in what is known as a "Chimeric" spectrum (68). The process of relinking the precursors and their associated MS/MS fragmentation patterns is referred to as deconvolution.

In comparison to DIA, DDA spectral acquisition utilizes a narrow isolation window (e.g., 1*m/z*) to fragment a single precursor in each MS/MS scan. However, recent studies have revealed that over 50% of the acquired MS/MS spectra in DDA are still contaminated by other precursors, leading to the prevalence of chimeric spectra (69, 70). These chimeric spectra can result in a lack of matching or mismatch with the reference library, thereby hindering the identification of compounds. Furthermore, the metabolome coverage of DDA acquisition is limited unless iterative DDA acquisition is implemented (71). Nevertheless, the use of iterative DDA increases the acquisition times and sample consumption, rendering it unsuitable for rare samples.

Several algorithms have been developed to deconvolve multiplexed DIA spectra, with the objective of separating and identifying the constituent metabolites. These algorithms can be broadly classified into two categories, as described in a previous study (72): 1) spectral library-based deconvolution, which entails comparing the DIA spectrum to a reference library of known metabolites to facilitate identification and quantification of the constituent metabolites, and 2) *de novo* deconvolution, which involves fitting the MS/MS chromatogram to the precursor ions for identification and quantification of metabolites without reliance on a reference library.

In DIA metabolomics, spectral data deconvolution can be achieved using various tools. For instance, Specter utilizes a linear algebra approach to match mixtures of spectra against a spectral library (73). OpenSWATH is another widely-used tool for DIA data deconvolution in proteomics (74). DIAMetAlyzer (67), a tool derived from OpenSWATH, has been developed to extract MS/MS fragments from DIA based on the spectrum pattern of data-dependent acquisition (DDA). However, due to the limited coverage of DDA data over the metabolome, DIAMetAlyzer may not be able to provide complete coverage over the metabolome. Moreover, the potential contamination of the DDA spectral library has not been taken into account. Similarly to Specter, MetDIA matches MS/MS fragments with their precursors by referencing a spectral library (75), but it also considers the correlation of peaks to avoid over-fitting.

Furthermore, various tools have been developed to perform deconvolution of DIA metabolomics spectra by fitting MS/MS chromatography. The widely used tool in this field is MS-DIAL. Initially, MS-DIAL employed the MS2Dec algorithm to deconvolve multiplexed DIA spectra by selecting three model peaks to deconvolve the original chromatogram (40). However, this method is not universally applicable to all ion fragmentation (72). Subsequently, MS-DIAL adopted a correlation-based deconvolution approach, *CorrDec*, to address this limitation. It has been demonstrated that *CorrDec* outperforms MS2Dec (72). DecoMetDIA has also been developed to deconvolve multiplexed SWATH-DIA data. Different from MS-DIAL, DecoMetDIA selects multiple model peaks to automatically fit the original chromatography. DecoMetDIA has been shown to effectively enhance the coverage of metabolome. Nevertheless, DecoMetDIA is programmed in R, and the computing performance is the primary bottleneck (76).

DDA spectra has been found to be prevalently chimeric, necessitating deconvolution to match the reference library. Traditional methods of deconvolution for DDA spectra rely on the experimental

proportional changes of MS/MS fragments from the elution profiles of chromatography, including RT shifting and intensity changes within and across samples. Several algorithms have been developed to deconvolve chimeric spectra based on this principle (40, 68, 69, 77). Moreover, HERMES is a tool that has been developed from an experimental perspective to optimize the acquisition of DDA data and enhance the selectivity and sensitivity for MS/MS acquisition (78). Therefore, the chimeric issue of DDA data is not resolved. DecoID, on the other hand, was created to address the deconvolution of chimeric DDA data. This algorithm employs spectral records from a database to decompose an experimentally acquired spectrum. In brief, multiple candidate components of the chimeric spectrum are extracted and a LASSO regression model is then applied to deconvolve the spectrum into distinct components. However, critical parameters require manual curation, and the missing components from database may hinder the performance.

Following the deconvolution of MS/MS spectra, a critical step in retrieving the chemical information of spectra is MS/MS reference library searching. Several tools have been developed for this purpose. One of the most widely used software is MS-Finder, which initially determines the formula from accurate mass, isotope ratio, and product ion information. It then utilizes hydrogen rearrangement rules to annotate MS/MS fragments and score the outcomes (79). Additionally, CSI:FingerID, a technique that combines fragmentation tree computation and machine learning (80), has been implemented by SIRIUS (81). Subsequent versions of the software have enabled the prediction of accurate molecular formula and matching fragmentation pattern in greater depth. The comprehensive internal database and user-friendly interface facilitate the easy database searching. Nevertheless, the bottleneck of this tool is the need to optimize numerous parameters, while the RESTful web-based searching is also easily influenced by network traffic.

Several studies have demonstrated that compounds sharing similar chemical structures tend to exhibit similar MS/MS spectra. Consequently, a number of tools, such as GNPS (82), MetDNA (83), and NetID (84), have been developed to use spectral similarity for MS/MS-based compound identification. GNPS employs a molecular networking model to identify compounds. In this model, the ions or features are connected based on the similarity of the extracted ion chromatogram of MS features and MS/MS spectral data. The relationships of adductions and isotopes are usually annotated by MS feature association, while MS/MS similarity among ions is used to illustrate their chemical associations and differences. Unlike GNPS, MetDNA employs molecular networking in a different manner. It begins with a few seeds of MS/MS spectra, and the metabolic reaction network model allows these seeds to propagate recursively to create a network. This network is based on the reaction pairs (substrate-product metabolites) (83). This metabolic reaction-based molecular network significantly improves chemical identification. NetID, developed from GNPS and MetDNA, establishes a propagatable network based on abiotic and bio-transformation knowledge to substantially improve annotation in untargeted metabolomics datasets. The entire network is then optimized using an integer linear model to facilitate metabolite discovery.

1.5 MS/MS reference libraries

The recommended method for identifying compounds using tandem mass spectrometry is to compare the spectrum against an in-house standard spectral library, which is considered the "gold standard" approach (85). However, this method may not be applicable in many situations due to the limited size of in-house libraries. In such cases, using public data sources would be more helpful, even though it may result in lower accuracy. Several public MS/MS databases have been developed and widely used for MS/MS identification, including METLIN (86), NIST database, MoNA (87), Massbank (87), mzCloud (https://www.mzcloud.org/), HMDB (88) and GNPS (89).

In addition, several in silico MS/MS database prediction algorithms have been developed, such as MINEs (90) and LipidBlast (91). Furthermore, some databases offer information about compounds without MS/MS spectra, such as KEGG (92), LIPIDMAPS (93) and LipidBank (94).

MassBank (https://massbank.eu/MassBank/) is also a public MS/MS reference database, providing access to both EI-MS and ESI-MS spectra references. It comprises the NIST and RIKEN databases. The MassBank of North America (MoNA, https://mona.fiehnlab.ucdavis.edu) has significantly expanded upon MassBank, offering additional spectral references from Oliver Fiehn's lab. All of these databases are formatted as msp files, which can be readily redistributed for MS/MS matching.

The Human Metabolome Database (HMDB, https://hmdb.ca) is a comprehensive, well-curated collection of human metabolome data. HMDB offers a wealth of information, including MS and NMR spectra, and provides online searching and open-source downloading capabilities. The original database is formatted as an XML file, which can be easily parsed by different programming languages. The versatility of HMDB makes it applicable for various fields, such as metabolomics, lipidomics, and exposomics.

Global Natural Product Social Molecular Networking (GNPS, https://gnps-external.ucsd.edu/gnpslibrary) is a dynamic online platform that facilitates interactive MS/MS spectral analysis and matching. The platform is specifically designed for MS/MS data processing and chemical result matching. Multiple MS/MS reference libraries are provided by GNPS, including clinical databases, pesticides, and small molecular pharmacological compounds, among others. Data within the database are offered in various formats, including msp, mgf, and json.

Metabolic In silico Network Expansions (MINEs, https://minedatabase.mcs.anl.gov/) is a novel extension of existing metabolic databases, featuring a distinct focus on previously unknown, yet

biological reactions related compounds (90). To construct this database, the Biochemical Network Integrated Computational Explorer algorithm was employed in conjunction with specific reaction rules. Notably, MINEs functions as a complementary resource to other metabolic databases, providing predicted spectra that enable the confident identification of unknown peaks. MINEs database offers open-source Application Programming Interface (API) access and permits database downloading in the msp format, facilitating user and developer dissemination.

Lipids play an essential role in cellular function and pathogenesis of various diseases. To support lipidomics studies, several databases have been developed, including LIPID MAPS (93), LipidBank (94), and LipidBlast (91). LIPID MAPS database provides a comprehensive collection of lipids, complete with their chemical structures, as well as tools for structure drawing. However, it should be noted that no MS/MS spectra are included in the database. Similarly, LipidBank offers a diverse array of lipids, categorized into distinct classes. Unfortunately, the database does not provide direct access to MS/MS spectra, and chemical information can only be downloaded as csv files. LipidBlast is an in-silico MS/MS spectra database, consisting of over one hundred thousand lipids belonging to 26 distinct lipid classes. Curated from LIPID MAPS, LipidBlast predicts the spectra of lipids at various voltages of fragmentation. LipidBlast is downloadable in msp format.

The KEGG Compound database (https://www.genome.jp/kegg/compound/) constitutes an extensive collection of small molecules, biopolymers, and other chemical substances that participate in crucial biological processes and interact with genomic components. This database contains more than 20,000 biologically relevant compounds across different species. However, MS/MS spectra are absent. The KEGG Compound database is readily downloadable through the KEGGREST package via API services in the KGML format, which can be easily parsed.

Furthermore, it is noteworthy that there is another valuable MS/MS spectral database, METLIN, which stands as one of the largest experimental MS/MS databases, featuring over 700,000 chemical MS/MS and neutral loss data. It can be freely accessed by the public for general search and compound annotation. However, the entire database is not possible to be downloaded or used for other machine learning purposes. mzCloud is also a popular MS/MS spectral database. Regrettably, it is not open-source.

In summary, over the past decades, multiple open-source MS/MS reference databases have been published. Identification based on MS/MS spectral matching can be helpful in revealing the chemical identity of MS features to some extent and thus, identifying potential chemical markers. However, in most cases, identification based solely on MS/MS is not sufficient to confirm chemical structures, which can hinder the analysis of biological insights from metabolomics data.

1.6 Functional analysis of metabolomics data

Functional analysis of metabolomics data mainly refers to the interpretation to the perturbed biological pathways and processes based on metabolomic data (95, 96). Metabolites are direct products and readout of functional activities. After a set of metabolites were identified, a functional analysis is required to convert these raw lists of compounds into biological knowledge.

Metabolomics functional analysis aims to understand the functional significance of metabolites and their role in biological systems (11, 97). This can be achieved through various methods, such as pathway analysis, enrichment analysis, and network analysis. Pathway analysis involves mapping metabolites to known biochemical pathways, while enrichment analysis compares the abundance of metabolites in a given dataset to a reference database to identify overrepresented functions. Network analysis aims to uncover the relationships between metabolites and other

cellular components, such as proteins and genes, to gain insights into the underlying mechanisms of metabolic regulation (95). Overall, functional analysis of metabolomics data is critical to provides a comprehensive understanding of the metabolic processes involved in cellular function and disease (11), and can aid in revealing the biological stories behind the metabolomics dataset.

Pathway enrichment analysis is employed to establish associations between molecules and pathways, which represent collections of molecular entities that share a biological function. The analysis is based on existing knowledge of biological pathways, with metabolites being mapped into a set of pathways that represent different biological functions. The most commonly used method for pathway analysis is over-representation analysis (98). Additionally, topology-based analysis can also be utilized to uncover the perturbation of pathways (99). Another method is Metabolite Set Enrichment Analysis (MSEA), which was developed from Gene Set Enrichment Analysis (GSEA) - a technique used for transcriptomics data analysis. MSEA aids researchers in identifying and interpreting patterns of biological perturbation (100). Moreover, several other tools are available for handling the functional enrichment of metabolites from metabolomics (97).

However, functional analysis of untargeted metabolomics data faces a bottleneck due to the difficulty of comprehensively identifying all MS features without the use of standard libraries. Therefore, annotation-based pathway enrichment analysis results may be inaccurate or biased.

Recently, the *mummichog* algorithm was established as a state-of-the-art method to overcome the challenge of functional enrichment analysis in untargeted metabolomics by utilizing the feature list directly (101, 102). This algorithm predicts pathway enrichment based on the collective power of metabolites within perturbed pathways. Initially, all MS features are matched to potential empirical compounds based on their *m/z*, RT, and adducts information. Next, the significant empirical compounds (based on user-defined p-value cutoff) are permuted with all empirical

compounds, and the perturbations of pathways are evaluated based on the enrichment level of significant empirical compounds from the permutation test. Currently, *mummichog* utilizes *m/z* and RT information to propose empirical compounds, but improvements to expand the MS data dimension used by *mummichog* are still needed.

1.7 Rationale and objectives

LC-MS-based metabolomics data processing involves a series of procedures that convert raw mass signals into metabolic features. Among the available algorithms, *centWave* is considered one of the optimal choices for processing both low-resolution MS and HRMS data, although many parameters require optimization to obtain satisfying results. Deconvolution of MS/MS spectra is crucial to obtain high-quality spectra for reference library-based compound identification. Previously ignored, DDA data deconvolution is now recognized as crucial due to the prevalence of chimeric spectra in DDA. While DecoID has been proposed as a solution, manual optimization of a key parameter for linear regression is required, and a missing component can cause the deconvolution to fail. An efficient and auto-tuned deconvolution algorithm for DDA is still highly needed. Besides, although DecoMetDIA has been shown to perform well for deconvolution of DIA data, it has low computing performance, and an efficient algorithm is still required. Moreover, functional analysis is a critical step in revealing biological insights in scientific research. However, current algorithms either rely on accurate compound identification or do not sufficiently utilize the information from LC-MS/MS data.

Therefore, it is hypothesized that an automated parameters' optimization pipeline cooperating with *centWave* could obtain optimal feature lists for different scenarios without the need for manual intervention. Using an auto-tuned regression, together with a spectral similarity-based spectra prediction network may be helpful to deconvolve chimeric DDA data. Furthermore, a highly-

efficient Rcpp/C++-based framework for DIA spectral data deconvolution algorithm could enhance the computing efficiency and decompose multiplexed DIA data more effectively. Additionally, integrating the MS/MS identification results into *mummichog* algorithm could further improve the accuracy of functional prediction.

The primary objectives of my thesis are:

1. To develop an automated pipeline for optimizing parameters in *centWave*, with the goal of improving its performance. To achieve this, a Design of Experiment-based mathematical model will be constructed to identify the optimal parameter combinations. This optimization process will be based on regions of interest extracted from the entire spectrum, allowing for the development of an effective and efficient method for optimizing *centWave*.
2. To improve the efficiency and efficacy of MS/MS spectral deconvolution for both DDA and SWATH-DIA data. To accomplish this, a spectral similarity network-based spectrum prediction model will be developed to predict missing components for DDA deconvolution. A penalized elastic regression model will be applied to deconvolve chimeric spectra, with an auto-tuned matrix utilized to minimize deconvolution residue. Additionally, SWATH-DIA deconvolution will be improved through the implementation of an Rcpp/C++ framework, resulting in enhanced performance.
3. To enhance the functional analysis algorithm by integrating MS/MS results into *mummichog*. This will be accomplished by combining compound identification results from MS/MS, which will reduce the redundancy of empirical compound lists and improve the accuracy of functional analysis. These improvements to *mummichog* will lead to a more comprehensive and accurate results from function analysis.

4. To evaluate the performance of the auto-optimized MS feature processing pipeline. A comprehensive meta-analysis on multiple metabolomics datasets will be performed. This evaluation will ensure that the data processing pipeline is effective and reliable, and the integration of results from meta-analysis will be helpful to reveal biological insights.

The general design of this thesis is depicted in Figure 1.1.

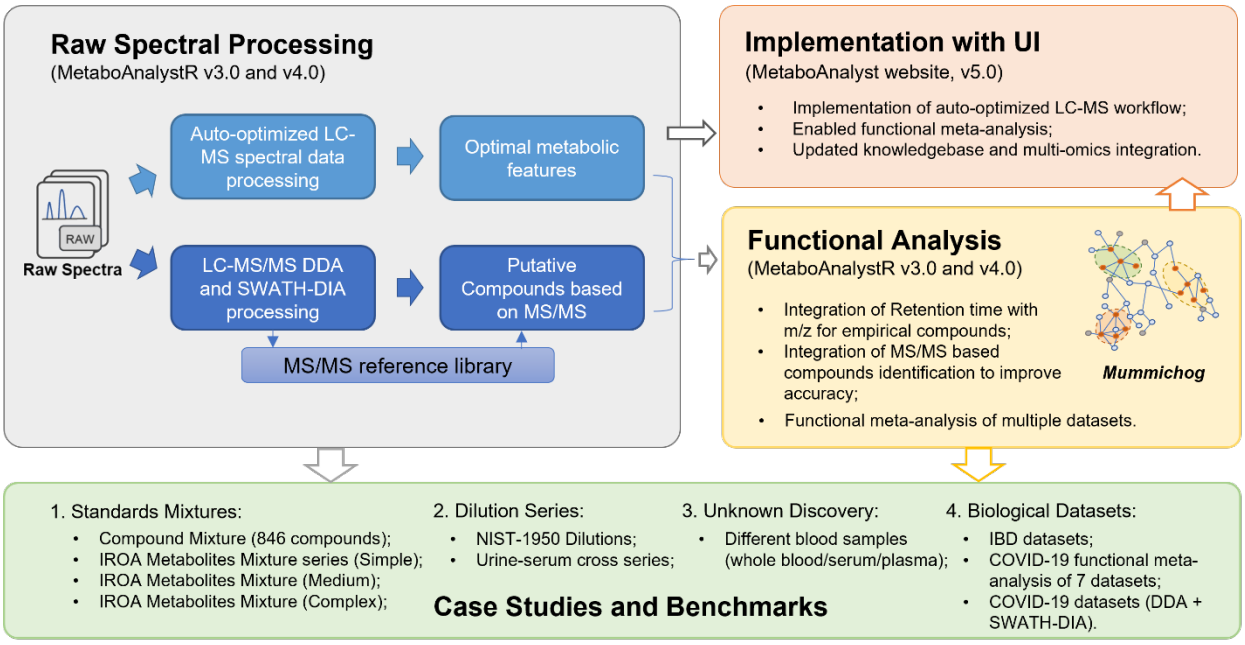


Figure 1.1. Overall design of the project presented in this thesis. The primary focus of this thesis is on advancing the development of auto-optimized LC-MS data processing and MS/MS data deconvolution workflows. This also involves the implementation of algorithms along with user interfaces and improvements to the functional analysis of global metabolomics. All of these functionalities have been introduced and integrated into various tools, including MetaboAnalystR version 3.0 and 4.0, as well as the MetaboAnalyst website version 5.0. To validate the effectiveness and performance of these tools, multiple datasets have been incorporated, and these tools have been benchmarked against other widely used software.

1.8 Outline of achievements

This section outlines the four main projects that have been undertaken to achieve the objectives of this dissertation.

1. MetaboAnalystR 3.0: an R package that has been updated by incorporating an auto-optimized parameter optimization workflow for the *centWave* algorithm. This version of MetaboAnalystR extracts specific regions of interest from spectra data and uses them to fine-tune parameters to an optimal state for detecting features in the entire spectra data in the following steps. The primary aim of this project was to accomplish Objective 1;
2. MetaboAnalyst 5.0: an updated version of website that now supports the processing of raw MS spectral data with the auto-optimized workflow. In this version, a user-friendly interface has been designed to enable users to process their raw spectral data online, using functions from MetaboAnalystR 3.0. Additionally, other features have also been incorporated. The primary objective of this project was also to achieve Objective 1;
3. MetaboAnalystR 4.0: an R package updated again to incorporate more functions for processing raw LC-MS/MS data. This version of MetaboAnalystR now supports DDA and SWATH-DIA raw spectral data processing and includes highly efficient data deconvolution, a comprehensive MS/MS reference library, and an enhanced *mummichog* algorithm that integrates MS/MS results into the workflow of empirical compounds generation. Several benchmark studies have been conducted, including standards validation, unknown compound discovery from whole blood samples, evaluation with serial dilutions, and two biological cases. This version was mainly developed to achieve Objectives 2 and 3;

4. Meta-Analysis of COVID-19 metabolomics datasets: analysis of seven COVID-19 metabolomics datasets were performed with MetaboAnalyst. In this study, raw spectral data was processed with the auto-optimized workflow. The function analysis was performed with *mummichog* algorithms. Results were integrated at the pathway level. Several significant pathways have been demonstrated to be associated with the pathogenesis of COVID-19. The main aim of this study was to achieve Objective 4.

The four achievements listed above correspond to Chapters 2 through 5 of this dissertation, respectively.

Preface to Chapter 2

This chapter presents an updated version of the MetaboAnalystR package that provides a streamlined workflow for processing LC-MS raw spectral data using an auto-optimized approach. The primary focus of this chapter is to address the issue of optimizing parameters for the *centWave* algorithm. The aim is to achieve Objective 1 successfully. In addition to this, several other functions have also been updated, with particular emphasis on *mummichog*. In summary, this version of MetaboAnalystR offers three primary functionalities: 1. an efficient workflow for auto-optimization of parameters for peak picking; 2. automated batch effect correction for large metabolomics datasets across multiple batches, and 3. more accurate pathway activity prediction using version 2 of the *mummichog* algorithm.

Chapter 2: MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics

Zhiqiang Pang [1], Jasmine Chong [1], Shuzhao Li [2] and Jianguo Xia [1,3] *

[1] Institute of Parasitology, McGill University, 21111 Lakeshore Road, Ste Anne de Bellevue, Quebec, H9X 3V9, Canada; zhiqiang.pang@mail.mcgill.ca (Z.P.); jasmine.chong@mail.mcgill.ca (J.C.)

[2] The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, CT 06032; shuzhao.li@jax.org

[3] Department of Animal Science, McGill University, 21111 Lakeshore Road, Ste Anne de Bellevue, Quebec, H9X 3V9, Canada

* Correspondence: jeff.xia@mcgill.ca; Tel.: +1-(514) 398-8668

This chapter has been published in Metabolites (Metabolites 2020, 10 (5), 186)

2.1 Abstract

Liquid chromatography coupled to high-resolution mass spectrometry platforms are increasingly employed to comprehensively measure metabolome changes in systems biology and complex diseases. Over the past decade, several powerful computational pipelines have been developed for spectral processing, annotation, and analysis. However, significant obstacles remain with regard to parameter settings, computational efficiencies, batch effects, and functional interpretations. Here, we introduce MetaboAnalystR 3.0, a significantly improved pipeline with three key new features: 1. efficient parameter optimization for peak picking; 2. automated batch effect correction; and 3. more accurate pathway activity prediction. Our benchmark studies showed that this workflow was 20~100X faster compared to other well-established workflows and produced more biologically meaningful results. In summary, MetaboAnalystR 3.0 offers an efficient pipeline to support high-throughput global metabolomics in the open-source R environment.

2.2 Introduction

Global or untargeted metabolomics is increasingly used to investigate metabolic changes of various biological or environmental systems in an unbiased manner (6, 103). Liquid chromatography coupled to high-resolution mass spectrometry (LC-HRMS) has become the main workhorse for global metabolomics (89, 104). The typical LC-HRMS metabolomics workflow involves spectra collection, raw data processing, statistical and functional analysis (105). A wide array of bioinformatics tools has been developed to address one or several of these steps (105, 106). Despite significant progress made in recent years, critical issues remain with regard to several key steps involved in the current metabolomics workflow.

The first issue is related to peak detection during raw spectra processing. Improving the ability to extract real compound signals and reduce noise is crucial to avoid noise inflation prior to statistical and functional analyses. Default parameters provided by common spectra processing tools are not applicable to all experiments (107), and misuse of parameters can lead to significant issues in data quality (108). To mitigate this issue, commercial tools such as Waters MassLynx™ and open-source software such as XCMS (35) and MZmine (36) allow users to specify multiple parameters to define LC-MS scan signals as chromatographic peaks. Although useful, such manual configuration assumes users are familiar with the experiments, which is often not the case. To facilitate the process, several tools and protocols have been developed for optimizing parameters for spectra processing (46, 109, 110). For instance, Isotopologue Parameter Optimization (IPO) is an R package designed to estimate the best parameters for XCMS (46). While the approach is effective, its stepwise optimization based on the entire spectra is very time consuming. IPO can often take days to weeks to compute the optimized parameters. Another recent tool is AutoTuner (110), which optimizes peak widths based on pre-defined extracted ion chromatograms (EIC). Despite being more computationally efficient than IPO, it may lead to potential errors due to unverified EICs used. Aside from these tools, Design of Experiment (DoE) strategies based on diluted samples provide a relative time-saving protocol for parameter optimization, but requires an extra series of diluted standards to be prepared (111). Another optimization strategy, One Variable at A Time (OVAT) (112), attempts to maintain the lowest coefficient of variation of peaks within a group, but this method takes even more computational time than IPO, in our experience.

The second issue is batch effect, which is commonly associated with large-scale clinical or population studies when samples are analyzed in different batches or across a long time period (113, 114). Over the course of spectral collection, chromatographic conditions can change and

baselines can drift (115). Besides, mass and intensity drifts are also quite common for LC-HRMS based metabolomics (116, 117). To address this issue, several types of batch correction methods have been developed based on quality control (QC) samples, QC metabolites, internal standards, matrix factorization, or location-scale normalization (118). These methods are based on different assumptions with their own advantages and limitations. Selecting a suitable batch correction method is critical, as it has a significant impact on downstream statistical and functional analysis.

Finally, biological interpretation of metabolomics data typically requires metabolites to be first identified prior to functional analysis. This process is very time consuming and remains a key bottleneck in global metabolomics (119, 120). The *mummichog* algorithm has introduced the concept of predicting pathway activity from ranked LC-MS peaks based on matching patterns of putatively annotated metabolites (101). The algorithm is available as Python scripts (121). To support the broad R user community, previous versions of MetaboAnalystR (105, 122) implemented *mummichog* v1.08. The recently released version 2 has added several improvements including the use of RT to refine the grouping of signals into empirical compounds (EC). The inclusion of RT will reduce false-positive annotations to increase the accuracy of pathway activity prediction.

Here, we introduce version 3.0 of MetaboAnalystR. Compared to its predecessor, version 3.0 has three key features: 1. efficient parameter optimization for spectral peak picking; 2. automatic selection of an optimal batch correction approach from 12 well-established methods; and 3. incorporation of RT coupled with updated pathway libraries for improved pathway activity prediction. The performances of these new features are assessed in the three case studies below.

2.3 Results

MetaboAnalystR 3.0 aims to provide an efficient pipeline to support end-to-end analysis of LC-HRMS metabolomics data in a high-throughput manner. This open-source R package is freely available at the GitHub repository (123). Detailed tutorials, manuals, example datasets, and R scripts are also included in the repository. The enhanced key points in the global metabolomics workflow of MetaboAnalystR 3.0 is summarized in Figure 2.1.

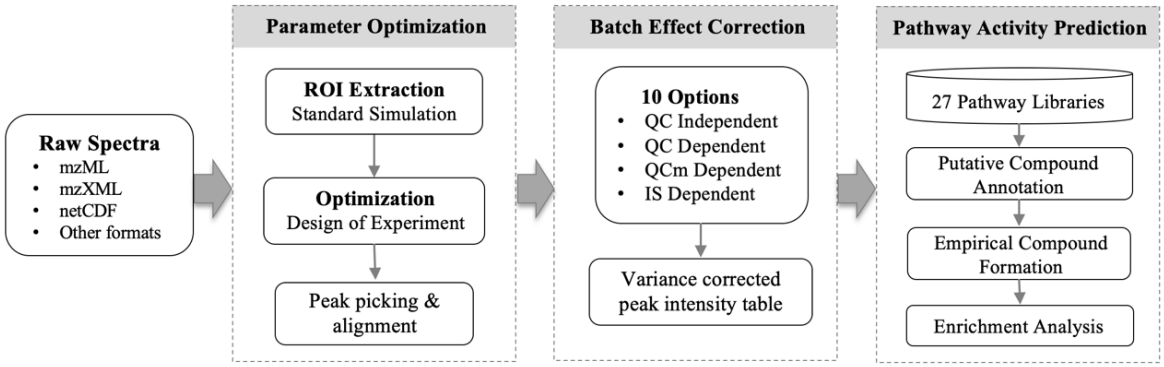


Figure 2.1. MetaboAnalystR 3.0 provides an optimized workflow for global metabolomics. (A) optimized peak picking, (B) automized batch effect correction, and (C) improved pathway activity prediction.

In comparison with other currently available parameter optimization tools, MetaboAnalystR 3.0 adopts an optimization strategy based on regions of interest (ROI) to avoid the time-consuming step of recursive peak detection using complete spectra. Briefly, the algorithm first scans the whole spectra across *m/z* and RT dimensions to select several ROIs that are enriched for real peaks. Second, these ROIs are then extracted as new synthetic spectra. Finally, a DoE model is used to optimize peak picking parameters based on the synthetic spectra (See Methods, 5.1. Peak Picking Optimization for more detail).

In this study, three benchmark datasets were used to evaluate the performance of MetaboAnalystR 3.0 including four standard mixture (SM) samples from a recent benchmark study (124), 12 standard reference materials samples from the National Institute of Standards and Technology (NIST), and 12 Quality Control (QC) samples from a large-scale metabolomics study on inflammatory bowel disease (IBD) (113). The overall time to complete the parameter optimization by the four different tools is shown in Figure 2.2. Compared to OVAT and IPO, there was a significant improvement in terms of speed for MetaboAnalystR 3.0. The CV based OVAT strategy took days to complete (>4 days for four samples), which is impractical for real-world datasets. Therefore, OVAT was not included in the case studies described in later sections.

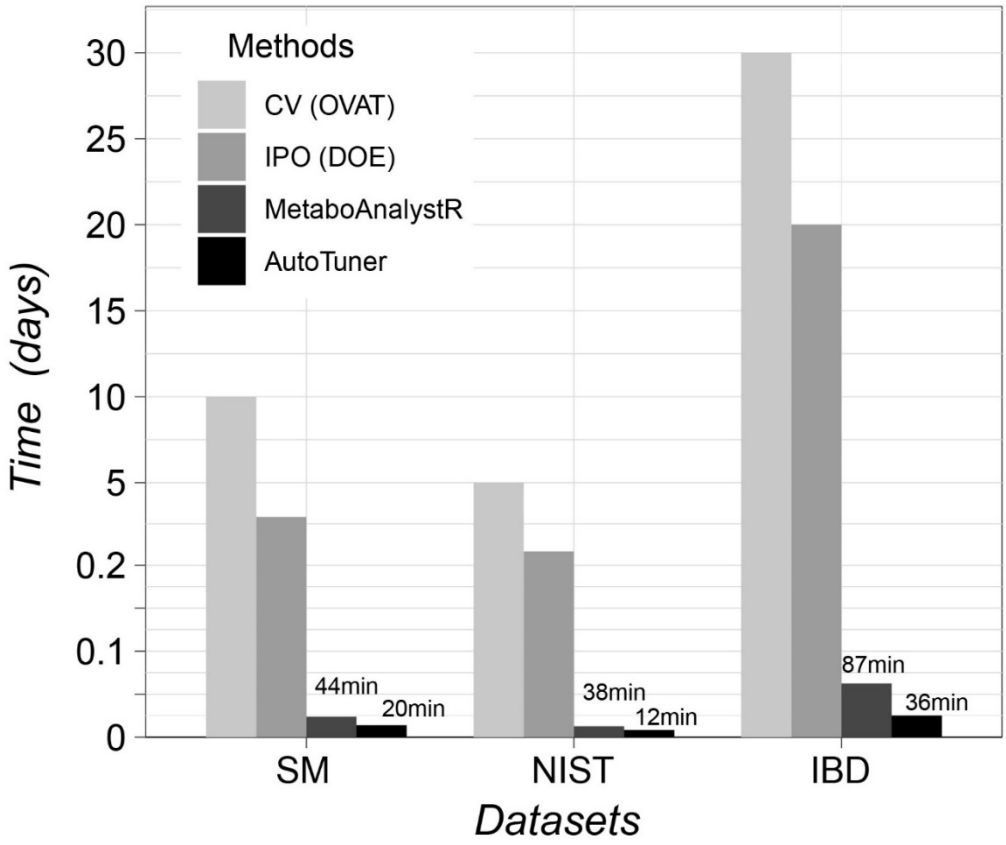


Figure 2.2. Time consumed by One Variable at A Time (OVAT), Isotopologue Parameter Optimization (IPO), MetaboAnalystR, and AutoTuner for parameter optimization on three

different datasets. The evaluations were performed on a desktop computer (Ubuntu 18.04.3 with an Intel® Core™ i7-4790 CPU and 32 GB of memory).

2.3.1. Peak identification benchmark case study

The performance of the optimized parameters for peak picking was evaluated with the SM samples consisting of 1100 common metabolites and drugs (124). The results of the raw data processing tools: (i) XCMS-Online with default parameters, XCMS R package (v3.8.2) with parameter optimization using (ii) IPO or (iii) AutoTuner, and (iv) MetaboAnalystR 3.0, are shown in Table 2.1.

Table 2.1. Qualitative peak picking results of the different tools using different settings.

Methods	Total Peaks	True Peaks	Quantified Consensus	Gaussian Peak Ratio
Default	16896	382	350	47.8%
IPO	24346	744	663	52.0%
AutoTuner	25517	664	603	40.5%
MetaboAnalystR 3.0	18044	799	754	64.4%

True peaks are peaks that match the targeted metabolomics results with *m/z* ppm <10 and RT difference <0.3 min. Qualified consensus refers to the peaks where the relative error of intensity ratio between the two groups is less than 50% compared with the actual concentration. Gaussian Peak Ratio is the ratio of peaks with shapes following a Gaussian distribution ($cor > 0.9$ and $P < 0.05$).

From Table 2.1, it is clear that the default parameters for XCMS are not optimal for this dataset. All parameter optimization tools (IPO, AutoTuner, and MetaboAnalystR 3.0) significantly improved the number of true peaks as well as peaks with consensus qualification. With regard to true peaks and quantified consensus peaks, MetaboAnalystR 3.0 increased 109.1% and 115.4%,

respectively, compared to the default XCMS. For IPO and AutoTuner, as the number of true peaks increased, so did the total number of peaks, indicating a potential inflation of noise. Meanwhile MetaboAnalystR 3.0 maintained a low total number of peaks (increase of 6.79% compared with default XCMS). In addition to the quantification of true peaks, we calculated the number of identified peaks following a *Gaussian* distribution. Peaks with a *cor* estimate over 0.9 and *P* value less than 0.05 are considered *Gaussian* Peaks. XCMS under different parameters (default, IPO and AutoTuner) displayed different performances on the peak simulation. Meanwhile, peaks picked by MetaboAnalystR 3.0 had the highest *Gaussian* Peaks ratio compared with other strategies.

2.3.2. Algorithm reliability benchmark case study

The reliability of MetaboAnalystR 3.0 and other tools/approaches were evaluated using the NIST SRM 1950 diluted serum series (125). The performance was assessed using the reliability index (RI) as defined by Zheng et al. (111). Briefly, peaks following the linearity in diluted series are considered to be reliable peaks, the higher the RI value, the better the data quality (126). *RI* is used to describe the general relative reliability of all identified peaks, while *Linear peaks* is the absolute count of peaks following linearity. The results from the four approaches are summarized in Figure 2.3.

As shown in Figure 2.3A, compared to the default (no optimization), IPO produces the best RI value (6252), however, at the cost of speed (316 minutes in total). Meanwhile MetaboAnalystR 3.0 has both good RI performance (5658) and acceptable speed (total of 49 minutes for optimization and data processing). AutoTuner is the fastest for optimization and data processing, but the improvement on RI is marginal. The number of peaks that meet the linearity ($P < 0.001$) are summarized in Figure 2.3B. MetaboAnalystR 3.0 produced the largest number of linear peaks compared to the other options.

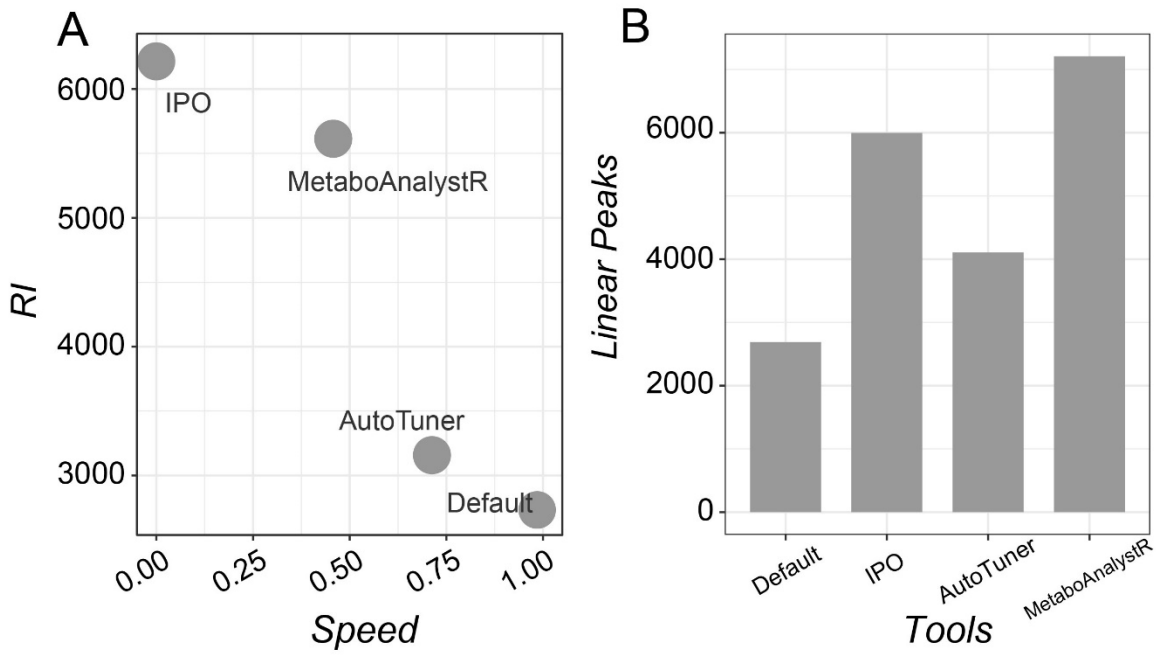


Figure 2.3. Assessment of the performance of different tools utilizing the NIST 1950 serum dilution series. (**A**) Reliability Index (RI) vs. processing speed for three optimization strategies compared to the default. X axis, speed, is the normalized time consumption to the default. (**B**) A bar graph showing the number of peaks with good linearity ($P < 0.001$).

2.3.3. Overall workflow evaluation using a large-scale clinical dataset

To evaluate the performance of the overall workflow, we applied the data processing pipeline on 545 clinical metabolomics samples obtained from the Inflammatory Bowel Disease (IBD) Multiomics Database (113). The dataset includes 58 QC samples assayed per every 20 patients' samples. The QCs are a pooled mixture of all patients' samples. Raw data processing identified a total of 8542 peak features using the optimized picking parameters compared to 6653 peaks with the default settings. The peak intensity tables were subjected to PCA and batch effect correction as shown in Figure 2.4.

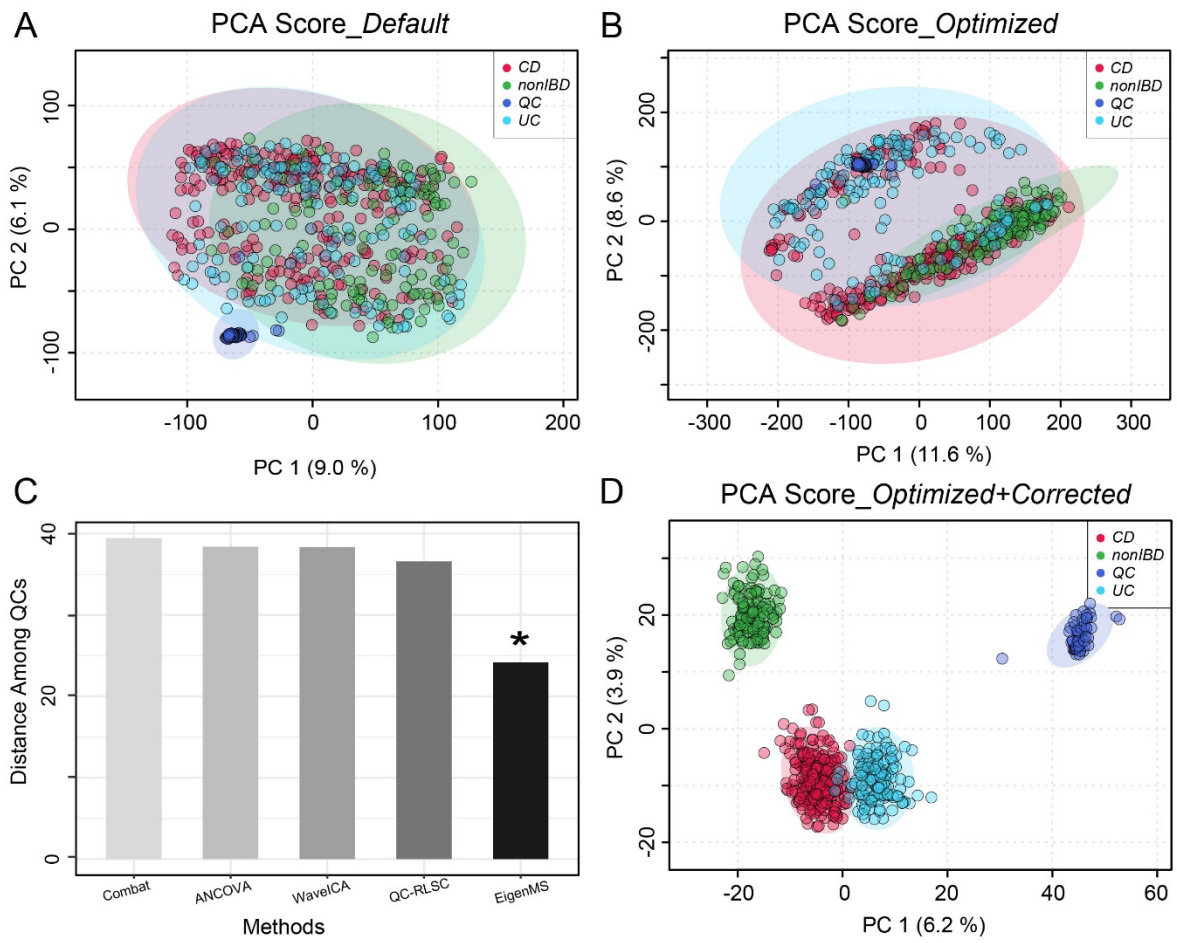


Figure 2.4. Performance evaluation using Inflammatory Bowel Disease (IBD) data. Principal Component Analysis (PCA) of peaks picked with (**A**) default parameters and (**B**) optimized parameters. (**C**) Performance of batch effect correction by different strategies. Among them, EigenMS behaved the best (indicated by *). (**D**) PCA of the optimized and batch corrected data.

Given that the QC samples are a homogenous mixture of all of the patients' samples, they are expected to locate in the center of the PCA as a tight cluster. However, this was not the case using the default parameters (Figure 2.4A). Using optimized parameters, these pooled QC samples were better mixed with the other samples (Figure 2.4B). However, both A and B showed systematic variations among these samples, suggesting batch effects in this large-scale study. In this case,

MetaboAnalystR3.0 applied batch effect correction with the Combat, Analysis of Covariance (ANCOVA), WaveICA, Quality Control-robust LOESS signal correction (QC-RLSC), and EigenMS methods, respectively. The PCA distances among all QC samples are summarized in Figure 2.4C, which indicates that the best correction was performed by EigenMS, a method based on singular value decomposition to detect and correct for systematic bias (127). After applying EigenMS, QCs were tightly clustered together and biological samples were clustered based on their biological origins (Figure 2.4D), providing strong evidence for the utility of the batch effect correction method selected by MetaboAnalystR 3.0.

Predicting pathway activities directly from LC-HRMS peaks can significantly accelerate biological discoveries in global metabolomics. We have previously implemented *mummichog* v1.08 within MetaboAnalystR 2.0. Now, MetaboAnalystR 3.0 has incorporated a major update of *mummichog* (v2.0) with RT integration. To demonstrate the improvements to biological interpretation stemming from both the optimized pre-processing steps and the updated *mummichog* algorithm, we applied both versions of the *mummichog* algorithm using the human BiGG and Edinburgh Model pathway library ("has_mfn") to compare the biological significance detected by the original pipeline (default peak parameters and non-corrected data, as shown in Figure S2.1) versus the optimized pipeline. For the Crohn's disease (CD) and non-IBD controls, a total of 3048 features were identified using the optimized pipeline and 2364 features using the non-optimized pipeline. For the non-optimized dataset, *mummichog* v. 1.08 identified no significant pathways (Gamma-adjusted P value < 0.05), while *mummichog* v. 2.0 identified 16 significantly different pathways (Tables S2.3 and S2.4). Similarly, for the optimized dataset, *mummichog* v. 1.08 identified only nine significantly perturbed pathways, whilst v. 2.0 identified 17 significantly perturbed pathways (Table 2.2). Evidently, *mummichog* version 2.0, with its integration of RT

information to group related *m/z* features into empirical compounds, reveals more biological insights than its predecessor. Moreover, *mummichog* results (both versions 1.08 and 2.0) for the optimized versus non-optimized dataset consistently identified differences in *Bile acid biosynthesis*, *Vitamin D metabolism,* and *Vitamin E metabolism* between CD patients and non-IBD controls. The details of the pathways identified are summarized in Tables S3–S6. Finally, both versions of *mummichog* algorithms also consistently identified a higher total number of pathways for the optimized dataset, versus the non-optimized dataset. This highlights the importance of data calibration to improve the detection of true biological signals. The other comparisons (ulcerative colitis vs. non-IBD control) showed similar results, as shown in Figure S2.2.

Table 2.2 The pathway enrichment results (top 20, Crohn's disease vs. non-IBD) generated by *mummichog* v1.0.8 and v2.0. Text in grey indicate pathways that are not significant (*P* value > 0.05).

***Mummichog* v1.0.8**		***Mummichog* v2.0**	
Pathways	***P* Value**	**Pathways**	***P* Value**
Bile acid biosynthesis	0.017199	Bile acid biosynthesis	0.011283
Vitamin D3 (cholecalciferol) metabolism	0.017526	Vitamin E metabolism	0.011321
Vitamin E metabolism	0.017966	Vitamin D3 (cholecalciferol) metabolism	0.014207
Carnitine shuttle	0.018084	Galactose metabolism	0.016026
Glycosphingolipid metabolism	0.021048	Glycerophospholipid metabolism	0.020464
De novo fatty acid biosynthesis	0.026554	Carnitine shuttle	0.021085
Keratan sulfate degradation	0.031317	Chondroitin sulfate degradation	0.025739
Fatty Acid Metabolism	0.032132	Vitamin B2 (riboflavin) metabolism	0.025739
N-Glycan Degradation	0.043912	Vitamin H (biotin) metabolism	0.025739
Phosphatidylinositol phosphate metabolism	0.053756	Fatty acid oxidation	0.025739

Hexose phosphorylation	0.069236	Omega-6 fatty acid metabolism	0.025739
Fatty acid activation	0.075044	Glycosphingolipid metabolism	0.041115
Limonene and pinene degradation	0.078492	Phosphatidylinositol phosphate metabolism	0.043604
Chondroitin sulfate degradation	0.082534	Hyaluronan Metabolism	0.04815
Glycosphingolipid biosynthesis - globoseries	0.082534	Putative anti-Inflammatory metabolites formation from EPA	0.04815
Saturated fatty acids beta-oxidation	0.082534	Electron transport chain	0.04815
Heparan sulfate degradation	0.082534	Heparan sulfate degradation	0.04815
Glycerophospholipid metabolism	0.09418	Sialic acid metabolism	0.061564
Starch and Sucrose Metabolism	0.13566	Vitamin A (retinol) metabolism	0.061564
Ascorbate (Vitamin C) and Aldarate Metabolism	0.14503	Saturated fatty acids beta-oxidation	0.061564

2.4 Discussion

The previous version (v2.0) of MetaboAnalystR provided an end-to-end workflow to process raw LC-HRMS metabolomics data (105). This new version (v3.0) has further enhanced three key steps of this workflow by focusing on efficient optimization for peak picking, improved batch effect correction, and more meaningful putative compound annotations for pathway analysis.

Parameter optimization remains a computational bottleneck in current raw LC-HRMS spectra data processing. Most tools rely on users to manually adjust the default parameters, which is inconvenient as users need to be very familiar with their MS instruments and experimental setup. The key concept of our optimization strategy is to use a subset of spectra based on multiple ROIs that are enriched for real peaks, instead of using complete spectra. These ROIs are selected based on the characteristics of the eluted compounds' peaks across the whole chromatogram to extract

peaks with wide *m/z* ranges (see Materials and Methods for more detail). The subsequent optimization is performed on peaks in these ROIs. One potential criticism we anticipate is the "bias" toward high-intensity peaks. We would like to point out that this is generally not the case, low intensity peaks are still sufficiently represented in these ROIs due to the sparse nature of LC-HRMS spectra (see Figure 2.5 in Materials and Methods). By focusing computational resources on real signals instead of noise, our approach has significantly accelerated the process for practical applications. Meanwhile, users can manually adjust the default *m/z* or RT window for selecting ROIs. The qualitative and quantitative efficacy of this approach have been demonstrated by two benchmark datasets. In particular, a significant improvement on the identification of true peak features has been observed using a known standards benchmark dataset (124). This resulted from the increased emphasis on the Gaussian fitting and peak group stability at the same time, rather than only focusing on the number of detected isotopes. The quantitative improvement of the parameters optimized by MetaboAnalystR 3.0 was also illustrated using the NIST SRM 1950 datasets. It should be noted that this data contains only two replicates for each concentration, which is a limiting factor for this validation.

Finally, the IBD data was first processed using the optimized parameters, followed by batch correction based on QC samples. The PCA revealed clear group patterns according to different IBD groups. Furthermore, more metabolic pathways were reported when using our optimized metabolomics workflow. The majority of these pathways are biologically meaningful according to previous studies including bile acid (127, 128), vitamin E (129), vitamin D3 (130, 131), galactose (132), glycerophospholipid (132), fatty acid (128, 133), and hyaluronan (134) metabolism pathways. Similarly, other comparisons between the different IBD groups also produced more perturbed metabolic pathways by our optimized workflow in MetaboAnalystR 3.0.

Using the IBD samples, we also compared the performances of the *mummichog* algorithm implemented in MetaboAnalystR 2.0 versus that in MetboAnalystR 3.0. The main difference between their implementations is that RT information is integrated when performing the putative compound annotation. This step moves pathway enrichment from the compound space to the empirical compound space formed by grouping co-eluting *m/z* features. Our results show that the new version improves both the number and quality of significant pathways that can be identified, as it identified perturbed pathways that are more consistent with IBD literature, as stated above.

2.5 Conclusions

MetaboAnalystR 1.0 provided the comprehensive statistical and functional analysis underlying the MetaboAnalyst web application, while MetaboAnalystR 2.0 equipped v1.0 with comprehensive raw LC-MS data processing and pathway activity prediction from MS peaks. MetaboAnalystR 3.0 has further enhanced three key aspects of the LC-MS data processing workflow including parameter optimization for peak picking, adaptive batch effect correction, and improved annotation of putative compounds for pathway activity prediction. MetaboAnalystR 3.0 represents our latest efforts toward developing an efficient pipeline for high-throughput global metabolomics.

2.6 Materials and methods

2.6.1. Peak picking optimization

The steps for parameter optimization include representative peaks extraction using the *PerformDataTrimming* function and parameter optimization based on the extracted peaks with the *PerformParamsOptimization* function. The concepts and mathematical details behind each function are provided below.

2.6.1.1. Extraction of Representative Peaks from Regions of Interest (ROIs)

The extraction of representative MS peaks is performed with the *PerformDataTrimming* function, which reads raw MS data of common formats (mzXML, mzML, etc.) into memory and extracts peaks using three strategies. The first strategy (default option) is named "*Standards Simulation Method*" (*ssm*). As its first step, at the *m/z* dimension, *ssm* divides the whole mass spectra into *m/z* bins and detects the signal intensity with a sliding window in parallel for all bins. The windows with the highest scan intensity sum within each bin will be retained, as shown in Figure 2.5A. Second, at the RT dimension, the sliding window method is used again to detect the scan signal intensity and returns the window with the highest values (Figure 2.5B). Synthetic spectra are created based on the returned ROIs defined by the two dimensions (*m/z* and RT). Peaks are extracted from the synthetic spectra to simulate standards across the whole *m/z* range (Figure 2.5C). These ROIs are enriched for true peaks, which are characterized by overall high-intensity signals distributed across the window. It is important to note that ROIs still contain a sufficient number of low-intensity signals for optimization, as shown in Figure 2.5D. The RT sliding window is also manually adjustable to cover different percentages (0, 100%] of RT dimension to further overcome the potential bias. If there are internal standards or quality control metabolites included within the user's samples, peaks with specific *m/z* and/or RT can be extracted or removed with the modes named "mz*_specific*" or "rt*_specific*".

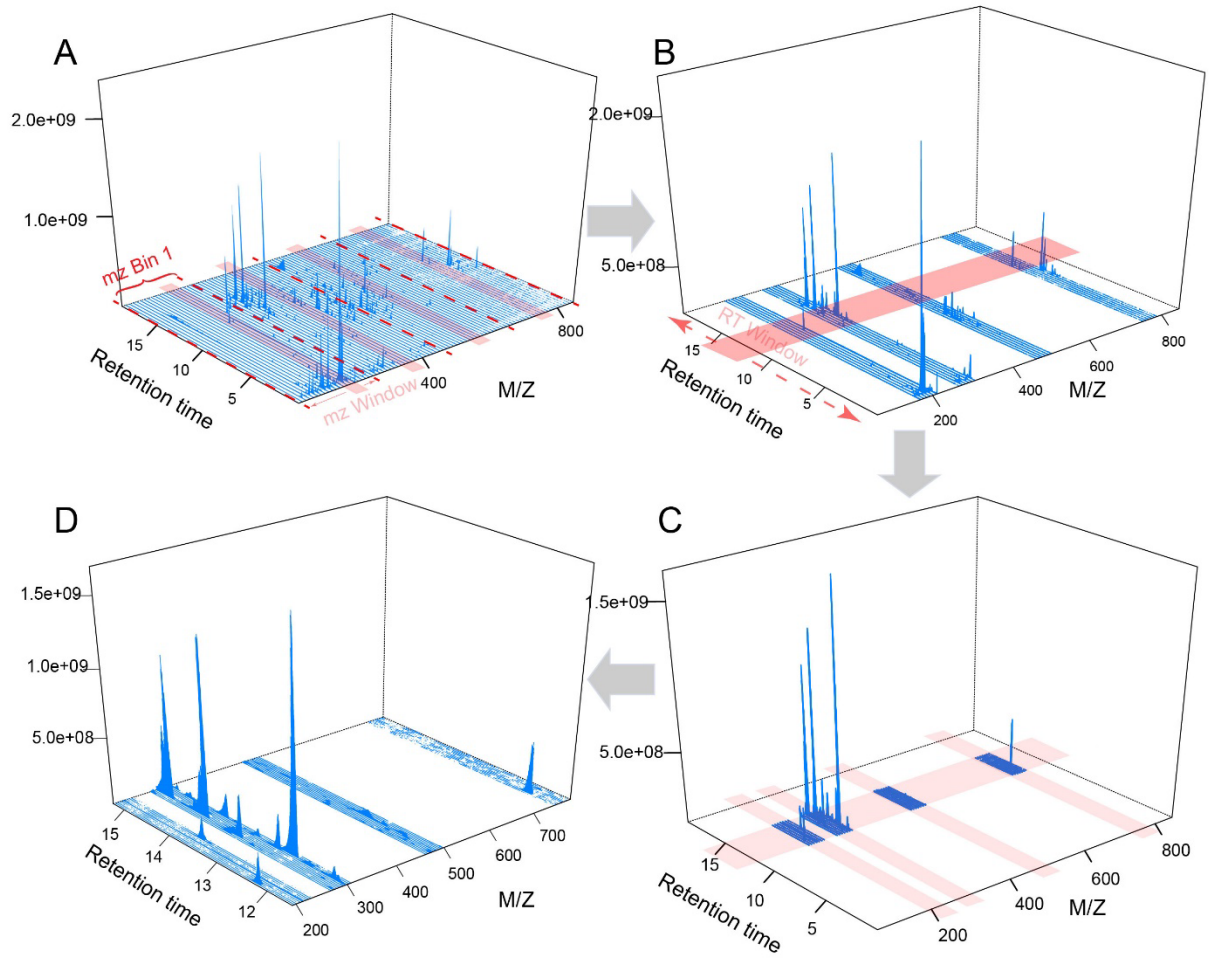


Figure 2.5. The selection process of regions of interest (ROIs) that are enriched for true peak signals. Red dashes in (**A**) represent the bin boundaries used for sliding windows' working to contain the most signal points. The whole spectrum is divided evenly into four bins. Four *m/z* windows (light red area) will slide within each bin respectively in parallel and select the window with the highest scan intensity sum in the retained *m/z* window. RT window (light red area) in (**B**) will slide across the entire RT dimension to get RT regions with the highest scan signal intensity. (**C**) The intersected MS scan signals from both the *m/z* and RT dimensions containing four ROIs. (**D**) The zoomed-in view of the ROIs (note low intensity peaks are still abundant).

2.6.1.2. Design of Experiment (DoE) Based Optimization

Once the representative peaks are obtained, the parameter optimization based on these peaks is performed with the *PerformParamsOptimization* function. The noise level (including *noise* and *prefilter* parameters) and the *m/z* variation (*ppm*) of a certain ROI is first evaluated with the kernel density estimator model developed by AutoTuner. Then, other detailed peak width and alignment parameters (*peak width min*, *peak width max*, *mzdiff*, *s/n_thershold* and *bandwidth*) are optimized with the DoE model based on the Box–Behnken method, as used by IPO. Unlike IPO, the optimization effects during the process is evaluated with the response variable, Quality Score (*QS*), defined below.

$$QS = \frac{RP^{3/2}}{'all\ peaks' - LIP} * GR^2 * QcoE$$

where *RP* is the reliable peaks and *LIP* is the low-intensity peaks, as defined by IPO according to the isotopes detected by CAMERA. Briefly, *RPs* refers to peaks with detectable isotopes. "*all peaks*" means all peaks detected including reliable and unreliable peaks. *LIP* refers to a group of peaks with the intensity of their isotopes too low (less than the average of the lowest 3% peak intensity in the spectra). Unlike IPO, the exponential factor for RP was lowered to 1.5 to reduce the sensitivity for peak picking and to avoid the inflation of noise. GR is the Gaussian peaks ratio. An exponential factor of 2 was empirically used to put more emphasis on the peak shape. QcoE is the quality coefficient. GR and QcoE are defined as below.

$$GR = \frac{Gaussian\ Peaks}{all\ peaks}$$

where *Gaussian* Peaks refer to the peaks that have shapes that follow the *Gaussian* distribution (*cor* estimate ≥ 0.9 and *P* value ≤ 0.05).

$$QcoE = norm(\mathrm{RCS}) + norm(\mathrm{GS}) + norm(\mathrm{CV})$$

where *RCS* is the RT correction score and *GS* is the grouping score and both are defined by IPO (46). Briefly, they are used to evaluate the RT shift and peak number within a peak group, respectively. Higher values of RCS and GS mean more stable and reliable peaks have been included and grouped as a peak feature. CV, the coefficient of variation, refers to the *CV* of peak intensity in a group, as described by Sascha K (112). This index highlights the importance of the peak intensity within a group. *RCS*, *GS*, and *CV* are normalized using the unit-based method. QcoE is further normalized to 0 to 1 and by weighted *RCS*, *GS*, and *CV* with 0.4, 0.4, and 0.2, respectively.

The *SetPeakParam* function provides initial parameters for different platforms including Ultra Performance Liquid Chromatography (UPLC)- Q-Exactive (Q/E) Orbitrap, UPLC- Quadrupole Time-of-Flight (Q/TOF), UPLC-Triple TOF (T/TOF), UPLC-Ion trap, UPLC-G2-S, High-performance liquid chromatography (HPLC)-Q/TOF, HPLC-Ion Trap, HPLC-Orbitrap, and HPLC- Single Quadrupole (S/Q). The best parameter combination is the one that produces the greatest number of reliable peaks, whose peak shapes follow a Gaussian distribution and show stable peak groups, as defined by the formula for Quality Score. The step is performed in parallel using multicores to accelerate the process.

2.6.2. Adaptive batch effort correction

Batch effect correction can be achieved with the updated *PerformBatchCorrection* function. All correction strategies are summarized in Table 2.3. At least three method candidates are available for all experimental designs. To identify the most suitable method for a given dataset, the correction results will be evaluated using PCA or the CCA model according to the gradient length

along the first axis of DCA analysis. If the value is over 3, PCA is an appropriate method, otherwise, CCA will be used (135). The results showing minimum inter-batch distances will be returned. QC-RLSC could be specified to adjust the signal drift.

Table 2.3. Batch effect correction methods available in MetaboAnalystR 3.0.

Categories	Methods
QC Sample Independent	Combat (136), WaveICA (118), Eigens MS (137)
QC Sample Dependent	QC-RLSC (114), ANCOVA (138)
QC Metabolite Dependent	RUV-random (139), RUV2 (140), RUVseq (141)
Internal Standards Dependent	NOMIS (142), CCMN (143)

2.6.3. Mummichog2 for pathway activity prediction

The R implementation of *mummichog* (101) was described in the previous version (105). *Mummichog* version 2 has incorporated RT in grouping ions and introduced the concept of empirical compounds (ECs). ECs are putative metabolites as measured by LC-HRMS, possibly containing a mixture of enantiomers, stereoisomers, and positional isomers that are not resolved by the instruments. Thus, ECs are similar to the "feature groups" referred by Mahieu and Patti (2017) (58). Whilst the Python version is available on GitHub as a separate project, our implementation in MetaboAnalystR 3.0 is as follows:

1) All *m/z* features are matched to potential compounds considering isotopes and adducts. Then, per compound, all matching *m/z* features are split into ECs based on whether they match within an expected RT window. By default, the RT window (in seconds) is calculated as the maximum RT * 0.02. This results in the initial EC list. Users can either customize the RT fraction (default is 0.02) or RT tolerance in general in the *UpdateInstrumentParameters* function (*rt_frac* and *rt_tol*, respectively).

2) ECs are merged if they have the same *m/z*, matched form/ion, and RT. This results in the merged empirical compounds list.

3) Primary ions are enforced (defined in the *UpdateInstrumentParameters* function [force_primary_ion]), only ECs containing at least one primary ion are kept. Primary ions considered are 'M+H[1+]', 'M+Na[1+]', 'M−H2O+H[1+]', 'M−H[−]', 'M−2H[2−]', 'M−H2O−H[−]', 'M+H [1+]', 'M+Na [1+]', 'M−H2O+H [1+]', 'M−H [1−]', 'M−2H [2−]', and 'M−H2O−H[1−]'. This produces the final EC list.

4) Pathway libraries are converted from "Compound" space to "Empirical Compound" space. This is done by converting all compounds in each pathway to all empirical compound matches. Then, the *mummichog*/GSEA algorithm works as before to calculate pathway enrichment.

5) To use the updated algorithm, set the version parameter in *SetPeakEnrichMethod* to "v2".

2.6.4. Benchmark case studies

2.6.4.1. Known Standards Mixture

The SM dataset produced by the HPLC-Q/E HF system consists of two samples with five replicates for each sample, as described by Li et al. 2018 (124). The global mass spectra were inspected with the *PerfromDataInspect* function. The extremely anomalous high-intensity dimethyl sulfoxide stock contaminant peak ([2*M+H] at *m/z* 157.035) was removed to avoid mistakenly overwhelming the parameter optimization process. The total ion chromatogram (TIC) of the data is shown in Figure S2.2. The parameter optimization was performed with HPLC-Q/E initial parameters based on two samples randomly selected from each group. The optimized parameters are provided in Table S2.1.

2.6.4.2. NIST-1950 Serum Diluted Series

The NIST 1950 serum dilution samples of 1, 0.2, 0.1, 0.05, and 0.025 were obtained from the MassIVE database (MSV0000083469). This dataset was generated by Pieter Dorrestein et al. using a Q Exactive Orbitrap (Thermo Fisher Scientific) in positive mode. Scanning *m/z* range was set between 133.0000 to 1981.0000 Thomson. The raw spectra were first converted to centroided mzXML format with ProteoWizard (v3.0.19073) msConvert (144). Parameter training was performed using the dilutions of 1 and 0.2 starting from the UPLC-Q/E default settings. TICs of the data are shown in Figure S2.3. The optimized parameters are provided in Table S2.1.

2.6.4.3. Clinical Inflammatory Bowel Disease Data

The Clinical IBD data was obtained from the Inflammatory Bowel Disease Multiomics Database (113). A large cohort of IBD patients were included for this study. The stool samples of CD (n = 266), UC (n = 144), and non-IBD (n = 135) were collected. The extraction and purification steps have already been described previously (128). The quality control (QC, n = 59) samples were also included. All clinical information from the samples is summarized in Table S2.2. The data format conversion and initial parameters were identical to the NIST dilution series above. The TICs of the data are shown in Figure S2.4. Parameter optimization was performed using four QC samples from each group randomly selected from the whole batch. The optimized parameters are provided in Table S2.1. The data analysis was finished with the whole MetaboAnalystR 3.0 workflow. Functional analysis was performed by integration with *mummichog*2 for the comparisons between different groups (cutoff of P value 2.0×10^{-6}2.0e-6).

2.7 Supplementary materials

Supplementary Materials: The following are available online at https://www.mdpi.com/2218-1989/10/5/186/s1, Figure S2.1: Bar plots of *mummichog* pathway enrichment results applied on Crohn's disease patients versus non-IBD controls, Figure S2.2: Scatter plots of the *mummichog* pathway enrichment results applied on ulcerative colitis patients versus healthy controls, Figure S2.3: TICs of benchmark 1 (known standard data) before and after optimization, Figure S2.4: TICs of benchmark 2 (NIST series) before and after optimization, Figure S2.5: TICs of benchmark 3 (IBD data) before and after optimization, Table S2.1: Optimized parameters summary of all datasets, Table S2.2: Clinical characteristics summary of IBD subjects, Table S2.3: *Mummichog* (v.1) pathways (Top 20) of non-optimized IBD data (CD vs. non-IBD), Table S2.4: *Mummichog* (v.2) pathways of non-optimized IBD data (CD vs. non-IBD), Table S2.5: *Mummichog* (v.1) pathways (Top 20) of optimized IBD data (CD vs. non-IBD), Table S2.6: *Mummichog* (v.2) pathways (Top 20) of optimized IBD data (CD vs. non-IBD).

Table S2.1. Optimized Parameters Summary of All Datasets

Data Cases	peakwidith	ppm	mzdiff	snthreshold	noise	prefilter	bw
Known Standards	(5.875,37)	1.84	0.0192	19.15	8061	(2,15046)	2
NIST Dilution	(6.25,14.75)	1.629	0.024	5	267	(2,601)	3
IBD	(8.125,15)	1.879	0.0012	7.5	1801	(2,3372)	5

Table S2.2. Clinical Characteristics Summary of IBD Subjects

Characteristics	CD	UC	nonIBD
Number	265	146	135
Gender (F/M)	116/149	93/53	61/74
Age of Diagnosis	20.54 ± 10.88	23.34 ± 14.39	/
Antibiotics/Yes	36	10	3
Chemotherapy/Yes	14	6	2
Immunosuppressants/Yes	61	19	0
Ileum ulcers	39.53%	0 %	0 %
Right Colon ulcers	13.95%	0 %	0 %
Transverse Colon ulcers	9.30 %	0 %	0 %
Left Colon ulcers	9.30 %	0 %	0 %
Rectum ulcers	11.62 %	0 %	0 %

Table S2.3. *Mummichog* (v.1) Pathways (Top 20) of non-optimized IBD data (CD vs. nonIBD)

Pathway Names	Pathway total	Hits total	Gamma-adjusted P value	Emp Hits	Empirical
De novo fatty acid biosynthesis	106	17	0.088229	0	0
Vitamin D3 (cholecalciferol) metabolism	16	11	0.10484	0	0
Porphyrin metabolism	43	12	0.10773	0	0
Bile acid biosynthesis	82	50	0.11527	0	0
Drug metabolism - cytochrome P450	53	26	0.14867	0	0
Aspartate and asparagine metabolism	114	27	0.1516	0	0
Tyrosine metabolism	160	59	0.17341	0	0
Leukotriene metabolism	92	36	0.17784	0	0
Vitamin A (retinol) metabolism	67	37	0.18074	0	0
Tryptophan metabolism	94	41	0.19228	0	0
C21-steroid hormone biosynthesis and metabolism	112	79	0.22105	0	0
Androgen and estrogen biosynthesis and metabolism	95	53	0.22638	0	0
Hyaluronan Metabolism	8	4	1	0	0
Glycolysis and Gluconeogenesis	49	13	1	0	0
Pyruvate Metabolism	20	5	1	0	0
Sialic acid metabolism	107	14	1	0	0
Chondroitin sulfate degradation	37	3	1	0	0
Linoleate metabolism	46	31	1	0	0
Galactose metabolism	41	22	1	0	0
Carnitine shuttle	72	26	1	0	0

Table S2.4. *Mummichog* (v.2) Pathways of non-optimized IBD data (CD vs. nonIBD)

Pathway Names	Pathway total	Hits total	Gamma-adjusted P value	Emp Hits	Empirical
Chondroitin sulfate degradation	1	1	0.000507	0	0
Omega-6 fatty acid metabolism	1	1	0.000507	0	0
De novo fatty acid biosynthesis	13	13	0.000516	5	0.05
Heparan sulfate degradation	2	2	0.000555	0	0
Hyaluronan Metabolism	3	3	0.000607	0	0
Ascorbate (Vitamin C) and Aldarate Metabolism	3	3	0.000607	0	0
Pentose and Glucuronate Interconversions	3	3	0.000607	0	0
Phosphatidylinositol phosphate metabolism	6	6	0.000787	0	0
Fatty acid activation	9	9	0.00101	24	0.24
Tryptophan metabolism	11	11	0.001186	0	0
Porphyrin metabolism	13	13	0.001387	61	0.61
Prostaglandin formation from arachidonate	14	14	0.001497	33	0.33
Glycerophospholipid metabolism	15	15	0.001615	41	0.41
Bile acid biosynthesis	17	17	0.001872	54	0.54
Arachidonic acid metabolism	19	19	0.002162	39	0.39
Leukotriene metabolism	19	19	0.002162	45	0.45

Table S2.5. *Mummichog* (v.1) Pathways (Top 20) of optimized IBD data (CD vs. nonIBD)

Pathway Names	Pathway total	Hits total	Gamma-adjusted P value	Emp Hits	Empirical
Bile acid biosynthesis	82	48	0.017199	0	0
Vitamin D3 (cholecalciferol) metabolism	16	10	0.017526	0	0
Vitamin E metabolism	54	35	0.017966	0	0
Carnitine shuttle	72	27	0.018084	0	0
Glycosphingolipid metabolism	67	20	0.021048	0	0
De novo fatty acid biosynthesis	106	18	0.026554	94	0.94
Keratan sulfate degradation	68	3	0.031317	0	0
Fatty Acid Metabolism	63	10	0.032132	36	0.36
N-Glycan Degradation	16	4	0.043912	8	0.08
Phosphatidylinositol phosphate metabolism	59	10	0.053756	22	0.22
Hexose phosphorylation	20	14	0.069236	0	0
Fatty acid activation	74	20	0.075044	97	0.97
Limonene and pinene degradation	10	6	0.078492	2	0.02
Chondroitin sulfate degradation	37	3	0.082534	0	0
Glycosphingolipid biosynthesis - globoseries	16	3	0.082534	3	0.03
Saturated fatty acids beta-oxidation	36	3	0.082534	34	0.34
Heparan sulfate degradation	34	3	0.082534	0	0
Glycerophospholipid metabolism	156	30	0.09418	4	0.04
Starch and Sucrose Metabolism	33	12	0.13566	0	0
Ascorbate (Vitamin C) and Aldarate Metabolism	29	9	0.14503	0	0

Table S2.6. *Mummichog* (v.2) Pathways (Top 20) of optimized IBD data (CD vs. nonIBD)

Pathway Names	Pathway total	Hits total	Gamma-adjusted P value	Emp Hits	Empirical
Bile acid biosynthesis	37	37	0.011283	21	0.21
Vitamin E metabolism	8	8	0.011321	0	0
Vitamin D3 (cholecalciferol) metabolism	4	4	0.014207	3	0.03
Galactose metabolism	7	7	0.016026	0	0
Glycerophospholipid metabolism	11	11	0.020464	8	0.08
Carnitine shuttle	14	14	0.021085	100	1
Chondroitin sulfate degradation	1	1	0.025739	0	0
Vitamin B2 (riboflavin) metabolism	1	1	0.025739	0	0
Vitamin H (biotin) metabolism	1	1	0.025739	0	0
Fatty acid oxidation	1	1	0.025739	0	0
Omega-6 fatty acid metabolism	1	1	0.025739	0	0
Glycosphingolipid metabolism	8	8	0.041115	64	0.64
Phosphatidylinositol phosphate metabolism	5	5	0.043604	23	0.23
Hyaluronan Metabolism	2	2	0.04815	0	0
Putative anti-Inflammatory metabolites formation from EPA	2	2	0.04815	51	0.51
Electron transport chain	2	2	0.04815	18	0.18
Heparan sulfate degradation	2	2	0.04815	0	0
Sialic acid metabolism	6	6	0.061564	0	0
Vitamin A (retinol) metabolism	6	6	0.061564	72	0.72
Saturated fatty acids beta-oxidation	6	6	0.061564	83	0.83

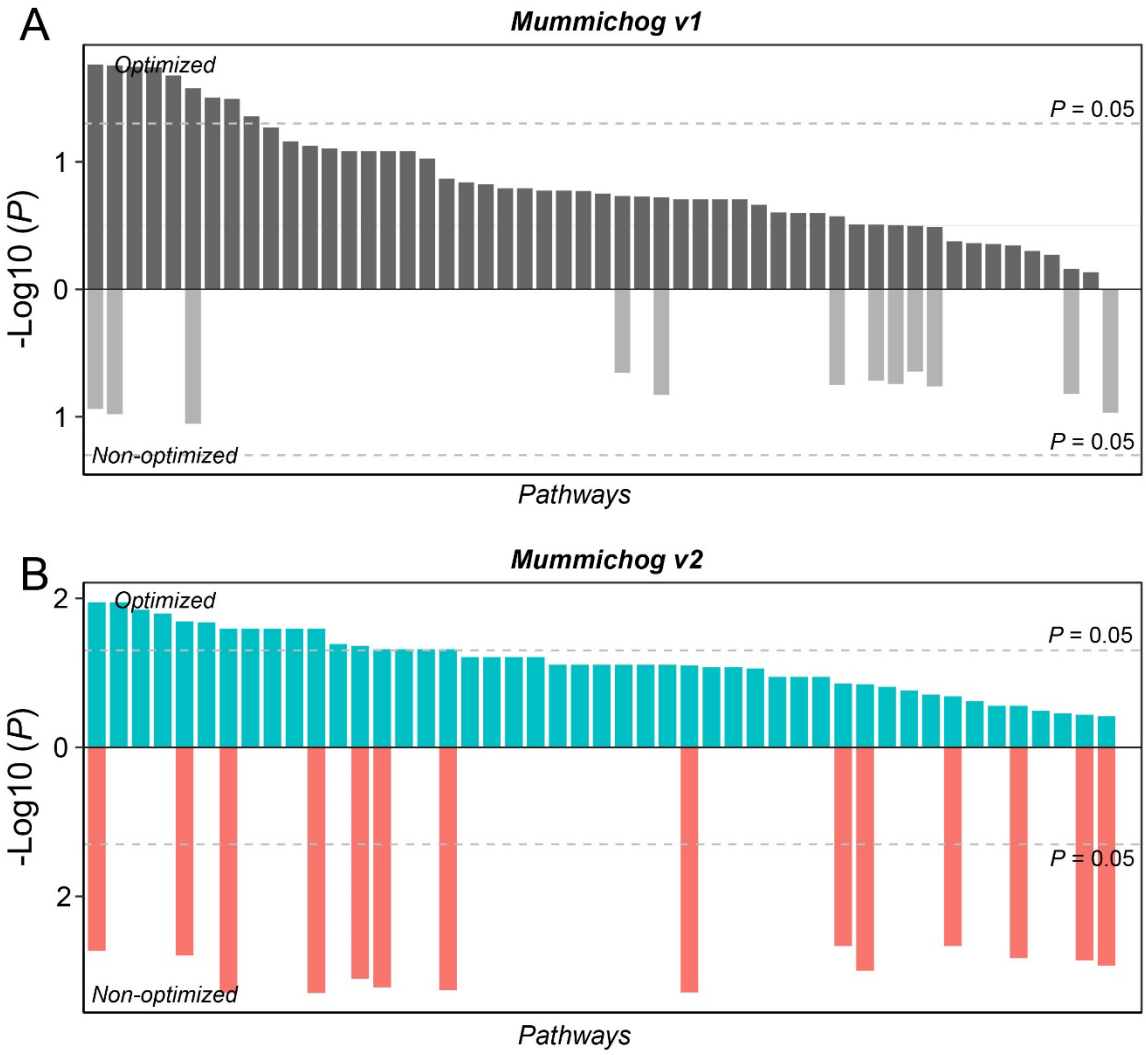


Figure S2.1. Bar plots of *mummichog* pathway enrichment results applied on Crohn's Disease patients versus nonIBD controls for: data generated using the default pre-processing parameters with *mummichog* v. 1.08 (A, above); using the optimized pre-processing parameters with *mummichog* v. 1.08 (A, below); using the default pre-processing parameters with *mummichog* v. 2.0 (B, above); and finally using the pre-processing optimized parameters with *mummichog* v. 2.0 (B, below). The light pink areas represent the pathways with P values < 0.05. For each panel, only the top 5 pathways are labeled.

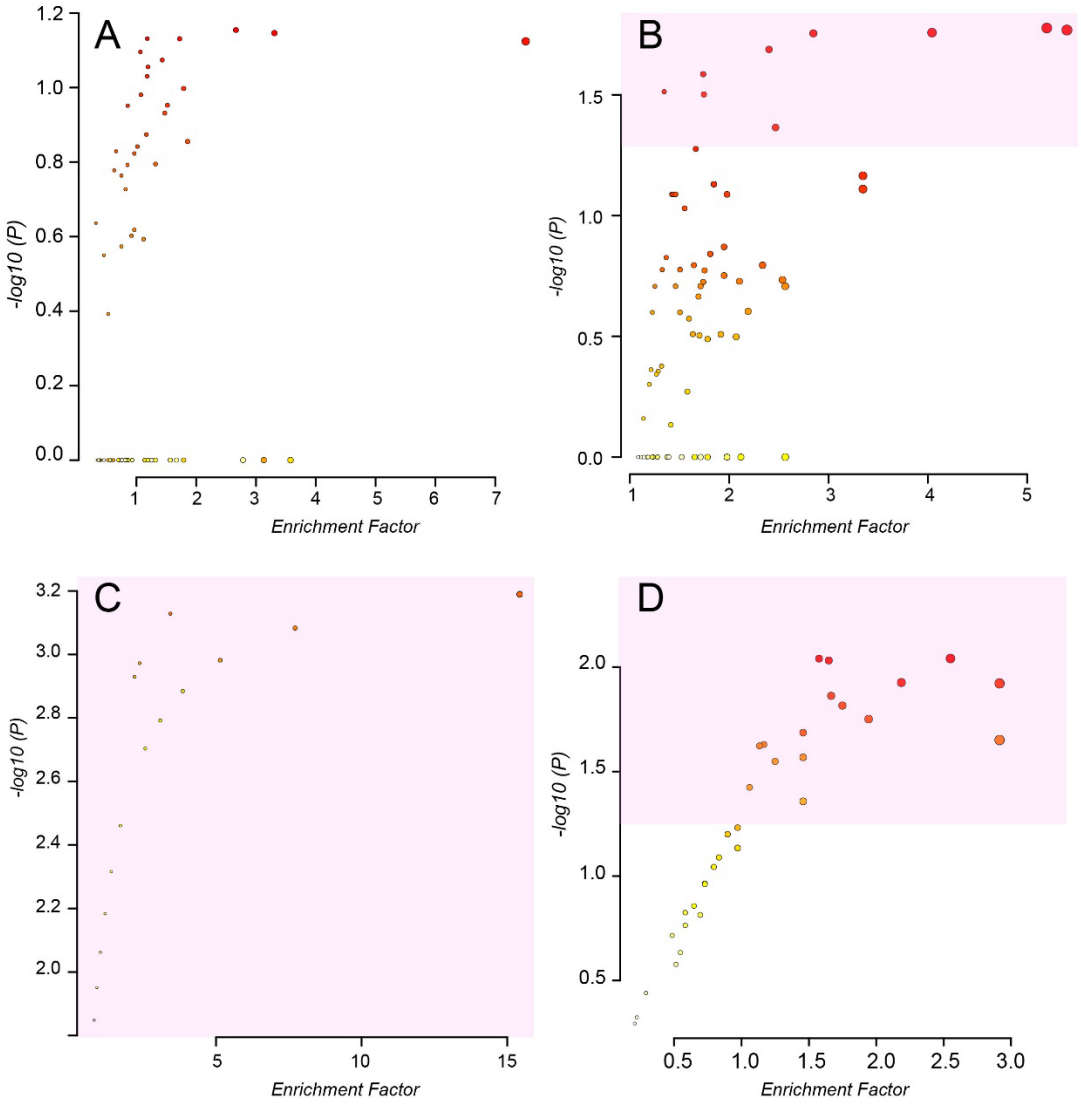


Figure S2.2. Scatter plots of the *mummichog* pathway enrichment results applied on ulcerative colitis patients versus healthy controls for: data generated using the default pre-processing parameters with *mummichog* v. 1.08 (A); using the optimized pre-processing parameters with *mummichog* v. 1.08 (B); using the default pre-processing parameters with *mummichog* v. 2.0 (C); and finally using the pre-processing optimized parameters with *mummichog* v. 2.0 (D). The light pink areas represent the pathways with *P* values < 0.05.

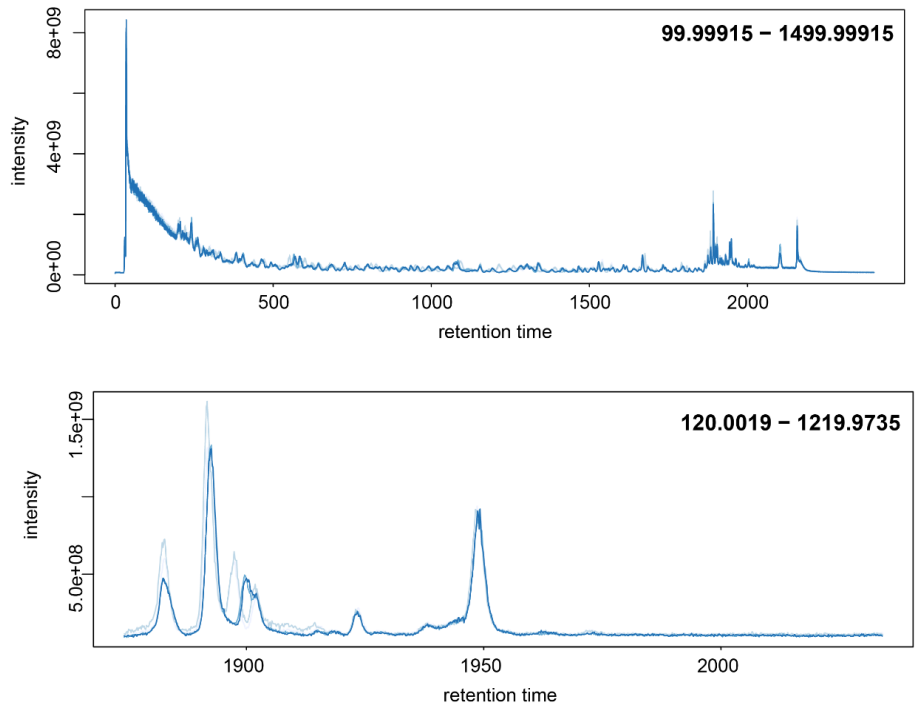


Figure S2.3. TICs of benchmark 1 (Known standard data) before (top) and after (bottom) optimization.

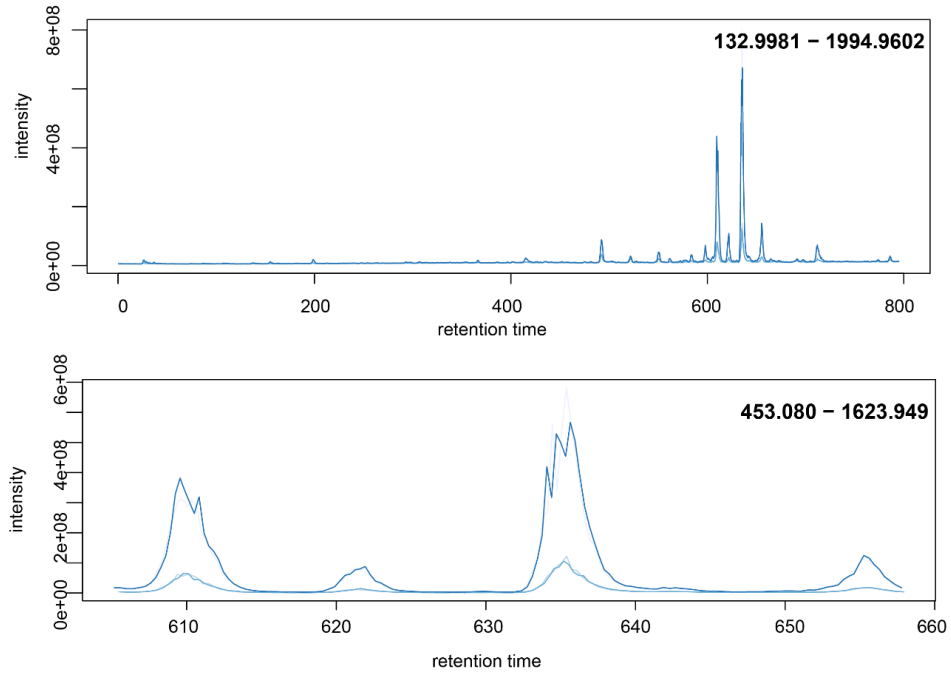


Figure S2.4. TICs of benchmark 2 (NIST series) before (top) and after (bottom) optimization.

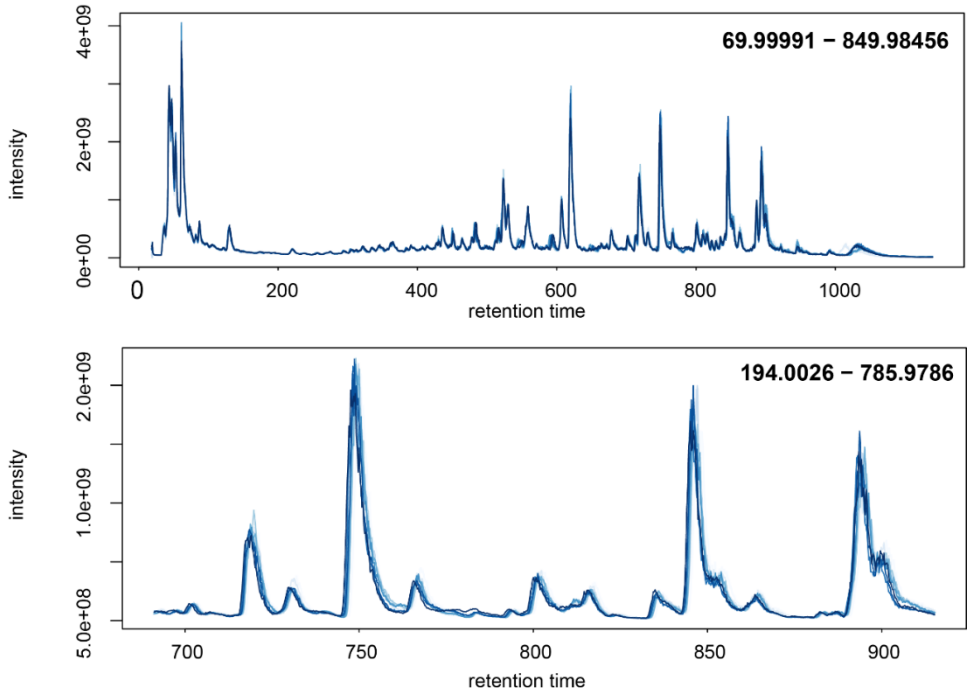


Figure S2.5. TICs of benchmark 3 (IBD data) before (top) and after (bottom) optimization.

Author Contributions: Conceptualization, J.X.; Data curation, Z.P.; Formal analysis, Z.P. and J.C.; Funding acquisition, J.X.; Methodology, Z.P., J.C., S.L., and J.X.; Supervision, J.X.; Writing, original draft, Z.P. and J.C.; Review & editing, J.X. and S.L.

Funding: This research was funded by Genome Canada, Génome Québec, U.S. National Institutes of Health (U01 CA235493), the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Canada Research Chairs (CRC) Program.

Acknowledgments: The authors truly appreciate the support from all members of the Xia lab.

Conflicts of Interest: The authors declare no conflicts of interest.

Preface to Chapter 3

This chapter provides an updated version of MetaboAnalyst, a popular web-based tool for metabolomics data analysis. The primary objective of this update is to integrate newly-developed functionalities from MetaboAnalystR 3.0 (Chapter 2) into the web-based platform, thereby enabling users to process their data through a user-friendly interface without the need for installation of R packages or programming languages on their local machine. The updated version facilitates online processing of raw spectral data through an auto-optimized approach, and enables functional analysis of one or multiple metabolomics datasets with the use of *mummichog* version 2. Additionally, this version implements a streamlined analysis workflow linking raw spectral data processing to functional insights, and several other updated features, including multi-omics integration. Overall, the purpose of this chapter is to enhance the tool's functionalities to meet the objective 1.

Chapter 3: MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights

Zhiqiang Pang [1], Jasmine Chong [1], Guangyan Zhou [1], Le Chang [2], David Anderson de Lima Morais [3], Pierre-Étienne Jacques [3], Shuzhao Li [4], Jianguo Xia [1,2,5 *]

[1] Institute of Parasitology, McGill University, Montreal, Quebec, Canada;

[2] Department of Human Genetics, McGill University, Montreal, Quebec, Canada;

[3] Département de Biologie, Université de Sherbrooke, Sherbrooke, Quebec, Canada;

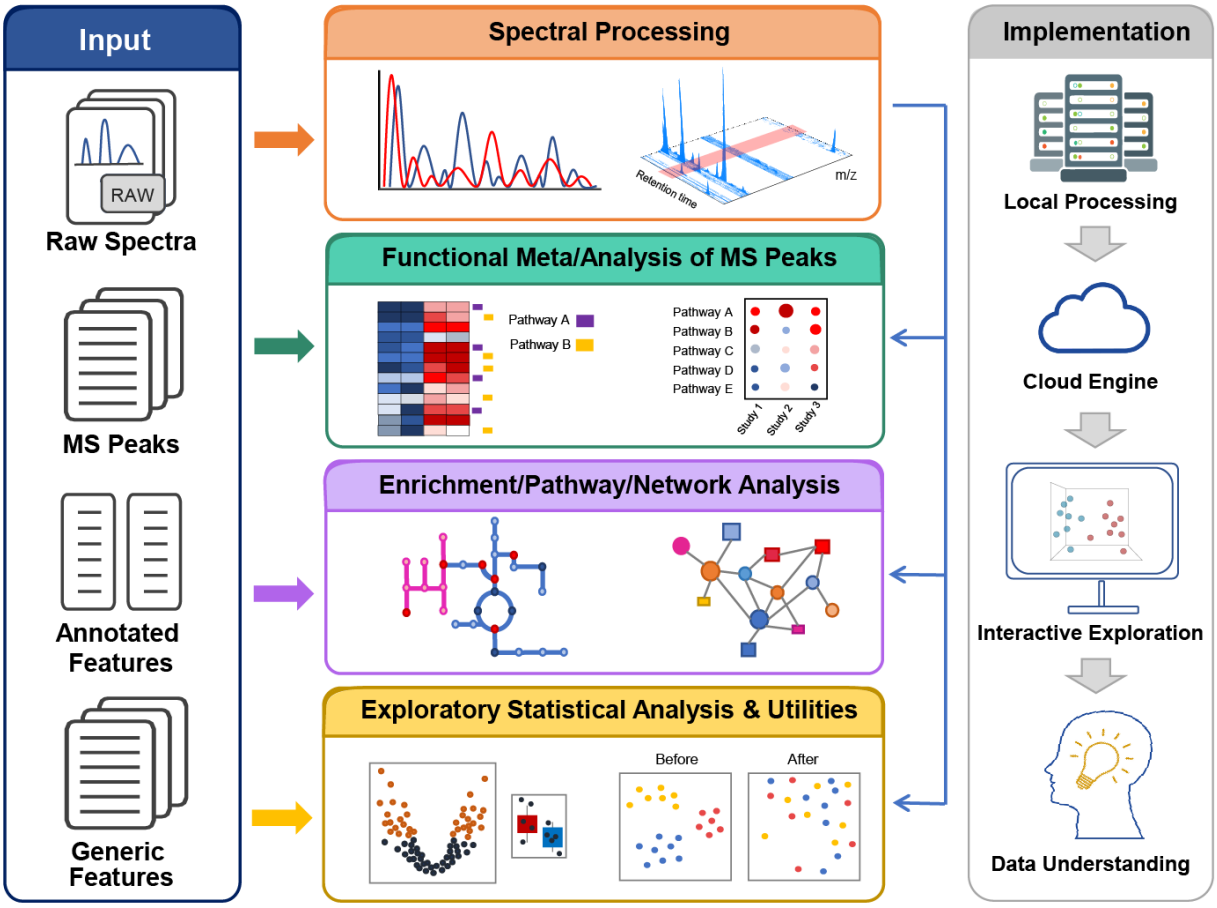
[4] The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, the United States of America

[5] Department of Animal Science, McGill University, Montreal, Quebec, Canada.

*Correspondence: jeff.xia@mcgill.ca (J.X.); Tel.: +1-(514) 398-8668 (J.X.)

This chapter has been published in Nucleic Acids Research (Nucleic Acids Research 2021, 40 (W1), 388)

3.1 Abstract



Since its first release over a decade ago, the MetaboAnalyst web-based platform has become widely used for comprehensive metabolomics data analysis and interpretation. Here we introduce MetaboAnalyst version 5.0, aiming to narrow the gap from raw data to functional insights for global metabolomics based on high-resolution mass spectrometry (HRMS). Three modules have been developed to help achieve this goal, including: (i) a LC–MS Spectra Processing module which offers an easy-to-use pipeline that can perform automated parameter optimization and resumable analysis to significantly lower the barriers to LC-MS1 spectra processing; (ii) a Functional Analysis module which expands the previous MS Peaks to Pathways module to allow users to intuitively select any peak groups of interest and evaluate their enrichment of potential

functions as defined by metabolic pathways and metabolite sets; (iii) a Functional Meta-Analysis module to combine multiple global metabolomics datasets obtained under complementary conditions or from similar studies to arrive at comprehensive functional insights. There are many other new functions including weighted joint-pathway analysis, data-driven network analysis, batch effect correction, merging technical replicates, improved compound name matching, etc. The web interface, graphics and underlying codebase have also been refactored to improve performance and user experience. At the end of an analysis session, users can now easily switch to other compatible modules for a more streamlined data analysis. MetaboAnalyst 5.0 is freely available at https://www.metaboanalyst.ca.

3.2 Introduction

Over the past two decades, metabolomics has contributed significantly to our understanding of metabolism across a broad spectrum of physiological and pathophysiological conditions (145, 146). It also plays a leading role in dissecting host-environment interactions (147) and has become an essential component in deep phenotyping for precision medicine (11, 148-150). As with other omics technologies, bioinformatics and analytics go hand-in-hand to enable high-throughput metabolomics data processing, analysis and interpretation. Among a wide array of bioinformatics tools developed for metabolomics (151, 152), MetaboAnalyst has been often listed among the popular choices together with XCMS (34, 153) and SIMCA-P (Umetric) etc. The first version (v1.0) of MetaboAnalyst was introduced over a decade ago, focusing on data normalization and statistical analysis (154). Since then, it has undergone continuous growth and co-evolves with metabolomics, encapsulated as milestone releases every three years. The v2.0 expanded to support functional analysis for targeted metabolomics (155). The v3.0 focused on translational biomarker analysis (156) and addressed the performance bottleneck by leveraging cloud computing and

modern web technologies (157). The v4.0 further improved on integrative and reproducible analysis (158), and began to support functional interpretation of global metabolomics data (101). With these successive releases, MetaboAnalyst has been steadily gaining and retaining users. According to Google Analytics, this web-based platform has processed over three million jobs submitted from >100 000 users worldwide in the past 12 months alone. For advanced users, the underlying functions have been released as the MetaboAnalystR package to permit more tailored data analysis and batch processing (102, 105, 122).

A key limitation of MetaboAnalyst was its limited support for global metabolomics especially with regards to raw data processing. With the consolidation of various protocols and availability of commercial kits developed for targeted metabolomics (159), global metabolomics based on high-resolution mass spectrometry (HRMS) has received growing attention (104, 147, 160). HRMS instruments such as Orbitrap or time-of-flight (TOF) systems can simultaneously measure a vast number of endogenous and exogenous compounds in a biological sample, providing unique information on an individual's metabolic phenotype, environmental exposures and associated biological responses. However, HRMS data processing is currently a labor-intensive task involving significant user input, as many parameters need to be empirically tuned in order to obtain satisfying results (107, 161). To democratize the power of HRMS to researchers beyond a few expert groups, we need to overcome two major hurdles - developing a self-tuning algorithm to enable automated parameter optimization and implementing a high-performance computational platform to deal with the big data challenges associated with raw data processing.

For most researchers, the peak tables obtained from raw data processing are not interpretable. The conventional approaches such as pathway or enrichment analysis require peaks to be identified first to gain functional insights (162). Therefore, it is necessary to enhance the support for

functional analysis directly based on peak tables. With the availability of public metabolomics repositories (163, 164), there is a growing interest in data mining and meta-analysis. However, the heterogeneity of global metabolomics datasets due to differences in analytical platforms and data processing parameters has posed significant challenges for this purpose. Addressing this need will greatly improve the value of global metabolomics datasets. Finally, improved support for lipidomics data, better integration with other 'omics' data, batch-effect correction, etc. have been among the common requests from the MetaboAnalyst users.

Here, we introduce MetaboAnalyst version 5.0, which represents our three years of effort to narrow the gap between raw HRMS spectra and functional insights since the release of the version 4.0. The key features of MetaboAnalyst v5.0 include:

1. A new module to support high-throughput, self-optimized LC-MS1 spectral processing.
2. A new module to allow meta-analysis of multiple global metabolomics datasets.
3. A weighted joint pathway analysis module for multi-omics integration, and a new function for data-driven network analysis (165).
4. Significantly updated and expanded underlying knowledge bases (species-specific pathway libraries and metabolite sets) for comprehensive functional analysis of both targeted and untargeted metabolomics.
5. Completely upgraded interactive graphics, refactored underlying codebase for improved performance and streamlined data analysis across compatible modules.
6. Other new features including support for mzTab 2.0-M (166) input format and importing data from the Metabolomics Workbench (163), as well as utility functions for automated batch correction and merging technical replicates.

The MetaboAnalyst v5.0 is freely available at https://www.metaboanalyst.ca. To accommodate computational demand, we have also set up two mirror sites hosted on high-performance computers dedicated for raw data processing. We have updated frequently asked questions (*FAQs*) and added seven new tutorials, which are easily accessible from the home page. The key features of MetaboAnalyst 5.0 are described below.

3.3 Overview of Metaboanalyst 5.0 workflow

In addition to supporting raw data processing for MS-based global metabolomics, MetaboAnalyst version 5.0 harmonizes workflows for both targeted and untargeted data analysis. As summarized in Figure 3.1, after proper data processing, all main inputs can be handled consistently within the framework of statistical analysis, functional analysis and meta-analysis with coherent interface design and navigation support. Altogether, these updates allow users to easily perform their analytical workflow and focus more on understanding their own data rather than how to operate the tool.

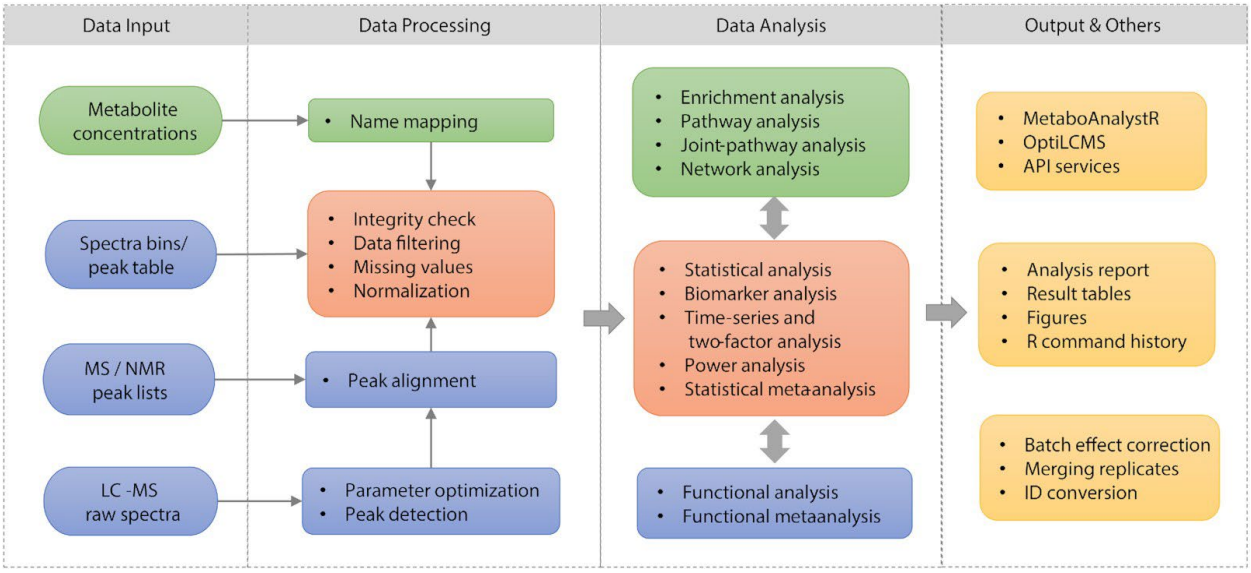


Figure 3.1. Overview of MetaboAnalyst v5.0 workflows. Steps for targeted metabolomics are indicated by boxes in green, steps for untargeted metabolomics are in blue, and those in orange can be used for both. Experienced users can use various utility functions or install the corresponding R packages (yellow boxes) to perform analysis beyond those pre-defined regular workflows.

3.4 Raw data processing

Over the past 15 years, XCMS and MZmine have evolved into the two most popular, open-source tools for HRMS raw data processing (35, 36). Both now use the *CentWave* algorithm for chromatographic peak detection (45). However, multiple parameters often need to be specified beforehand in order to obtain good results, which has caused challenges for its practical applications even for an experienced analyst. The XCMS Online platform has partially addressed the issue by offering several pre-optimized platform-specific parameters (34, 167). However, more refined parameter optimization is usually necessary, because chromatography can vary greatly between laboratories, and the spectral data are influenced by sample preparation and many configurations or conditions of mass spectrometers.

We have recently developed a self-tuning parameter optimization method for XCMS-based HRMS spectra processing and benchmarked its performance against other well-established approaches (102). The algorithm was initially developed as a component in MetaboAnalystR 3.0. Based on user feedback, we recently extracted and optimized the algorithm as an independent R package (OptiLCMS, https://github.com/xia-lab/OptiLCMS) to be embedded in other pipelines. This pipeline is designed to automatically identify the optimal parameters for a user-provided dataset in an efficient manner. Briefly, the 'automated optimization' pipeline will select multiple regions of interest (ROIs) across the whole spectra as the training spectra. Then, a design-of-experiment

(DoE) optimization will be executed to find out the combination of parameters with the most well-behaved peak shape and stable peak groups to be applied to whole dataset for peak detection. MetaboAnalyst v5.0 offers this pipeline via its user-friendly web interface to support both automated and manual parameters optimization to accommodate both regular and expert users. We have also developed a resumable workflow to accelerate data re-analysis after parameter update. To accommodate a wide range of spectral data qualities, we recently implemented a function to detect and exclude common background noises and experimental contaminants during the parameter optimization stage. Specifically, all *m/z* centroids from the whole spectrum will be extracted first and those *m/z* features appearing consecutively across half of the entire chromatogram will be excluded for parameters' optimization. Following peak detection and alignment, the annotation of adducts and isotopes is based on the CAMERA R package (54). The pipeline is now available as the new LC–MS Spectral Processing module in MetaboAnalyst v5.0.

Users can upload up to 200 data files in the supported open data formats (mzML, mzXML, netCDF or mzData). Since raw data processing is a time-consuming process, users can create and save a bookmark link after job submission. The link is used to check their job status and to retrieve the result. Alternatively, users can freely create accounts using their emails for better data management and communication. Registered users can create up to 10 projects, revisit or re-analyze their data later. When raw spectral processing is complete, users can visually inspect their results in an interactive 3D PCA plot (Figure 3.2A), as well as view total ion chromatogram (TIC) plots, base peak intensity (BPI) plots, RT correction results, etc. Furthermore, users can click any feature of interest to view its corresponding extracted ion chromatogram (EIC) plot. From the Results Download page, users can download all the processed data and peaks tables or start a new journey to other compatible modules.

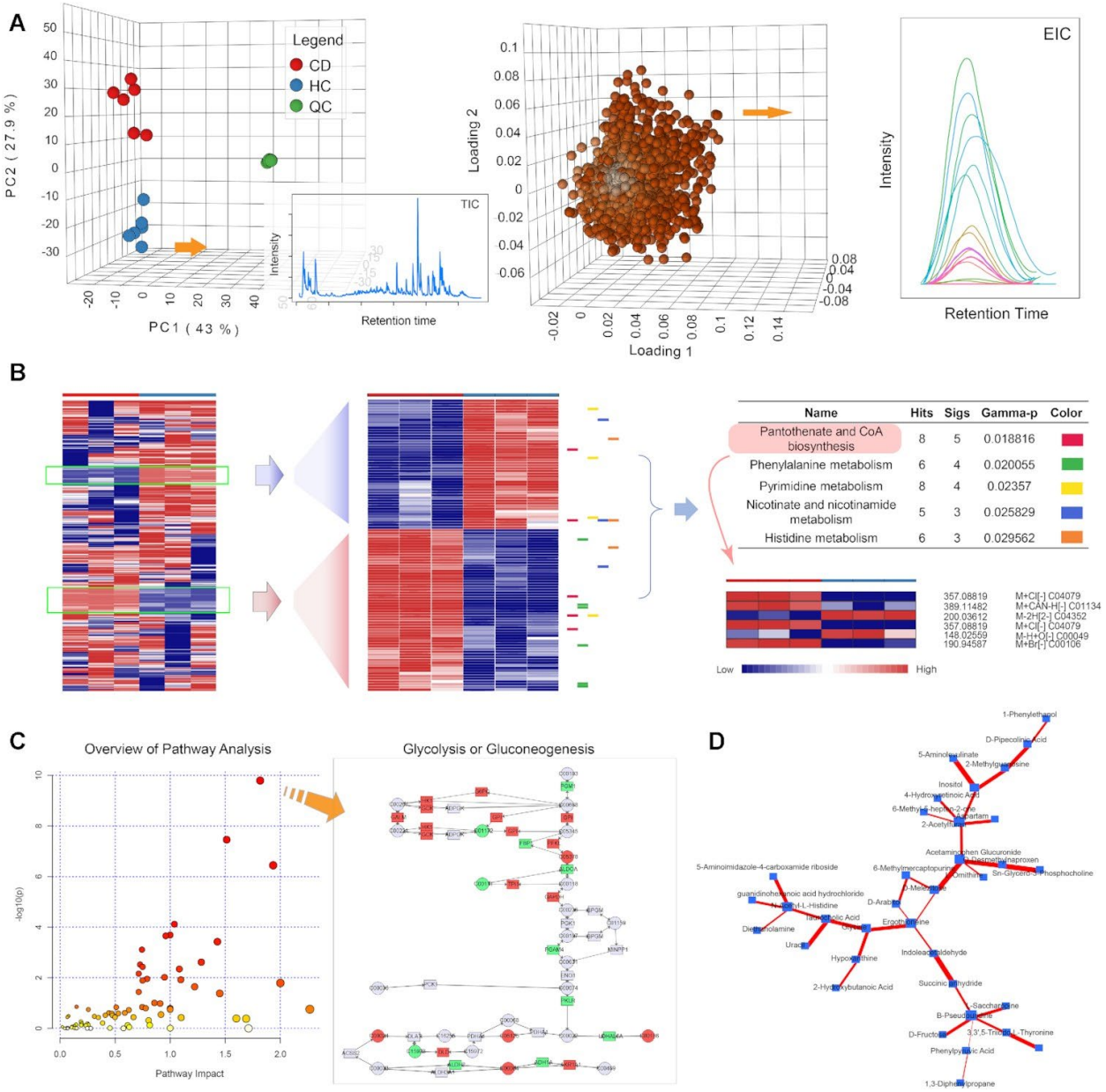


Figure 3.2. Example outputs from several new features of MetaboAnalyst v5.0. (A) Interactive PCA scores and loadings plots generated from the Raw Data Processing module. Users can click any samples or features to view their spectra; (B) Enrichment analysis of patterns detected in a peak table from the Functional Analysis module. Users can drag-select any patterns and test their enriched functions for further exploratory analysis; (C) An example output from the Joint-Pathway

Analysis module. Users can click any data points to view the underlying pathways; (D) An example output from the DSPC network analysis.

3.5 Functional analysis of MS peaks

It is now possible to directly translate a HRMS peak table into biological insights after raw data processing. MetaboAnalyst v4.0 first implemented the 'MS Peaks to Pathways' module based on the *mummichog* algorithm (101). Briefly, the algorithm first performs putative annotation of MS peaks considering different adducts and ion modes. These putative compounds are then mapped onto user selected pathway libraries for pathway activity prediction. The previous version (*mummichog* version 1) only considered the *m/z* dimension. In MetaboAnalyst v5.0, we have upgraded the algorithm to version 2 by integrating both *m/z* and RT dimensions to formulate empirical compounds, thereby further improving the accuracy of functional interpretation (102). Both versions of the *mummichog* algorithm are now available in MetaboAnalyst v5.0. The new interface also allows advanced users to customize the default adduct lists and currency metabolites - ubiquitous compounds such as water, oxygen, carbon dioxide, *etc*. (168).

The typical application of the *mummichog* algorithm is to predict pathway activities based on a list of MS peaks ranked based on t-tests. The concept can be generalized to test enrichment of any predefined function (i.e., metabolite sets) in any peak groups of interest (i.e., a cluster of similar peaks instead of significant peaks). Herein, we have implemented an interactive heatmap to allow users to perform functional analysis on any manually selected region of interest. In this case, the uploaded peak intensity table will be first visualized as an interactive heatmap (169). Users can perform cluster analysis with different methods, and then specify (via drag-select) one or more patterns of interest. The *mummichog* will be applied to predict enriched functions for the selected

peaks. From the result, users can click any function name (i.e., pathway or metabolite set) to see the corresponding features annotated beside the heatmap (Figure 3.2B).

3.6 Meta-analysis of global metabolomics data

It is notoriously challenging to integrate untargeted metabolomics data across different studies, because different extraction methods, chromatographic conditions and mass spectrometry platforms all lead to heterogeneity of HRMS data. This issue has precluded the use of untargeted metabolomics datasets for large-scale meta-analysis using conventional statistical methods (170). Some strategies have been proposed to resolve this issue before (171). To address this gap, we have developed a new module to enable researchers to perform functional meta-analysis of global metabolomics datasets.

Users can submit multiple peak intensity tables obtained from the same (or very similar) diseases or phenotypes of interest. The meta-analysis can be performed by pathway-level integration or by pooling peaks. If the studies are independent of each other (i.e., different samples) but interrogate more or less the same pathways, the integration should be performed at the pathway level. In this case, the pathway analysis will be first performed on each dataset and the final significant pathways will be identified based on the integrated p-values. The results can be visually explored in an interactive Venn diagram. In contrast, the peak pooling strategy aims to improve the metabolome coverage by combining complementary information obtained under different experimental conditions (i.e., compound extractions, chromatographic conditions, ion modes, etc.) from the same set of samples using the same or very similar MS instruments. The results can be visually explored in a KEGG metabolic network. The utility of the pathway-level integration has been demonstrated in our recent meta-analysis of COVID-19 global metabolomics datasets (29).

3.7 Multi-omics integrative analysis

Integrating data from different omics layers can provide greater resolutions to reveal mechanistic insights as compared to using a single omics profile. Multi-omics integration can be either data-driven based on multivariate statistics (172) or knowledge-driven based on known pathways or molecular interaction networks (173). In practice, both data-driven and knowledge-driven approaches can be further integrated to maximize information gain (174, 175).

Integrated pathway analysis of genes and metabolites was first launched as the 'Joint Pathway Analysis' module in MetaboAnalyst v3.0 by directly concatenating genes and metabolites into a single query (i.e., combining queries) followed by over-representation analysis. However, the results are often dominated by transcriptomics data which tends to yield many more significant features than metabolomics. To address this issue, we have added three new options for combining P-values from different tests of the same hypothesis (176), including one unweighted (Fisher's method) and two weighted approaches (Stouffer's Z-score method). The weights are the proportions of genes or metabolites within the combined universe (overall) or within individual pathways (pathway-level) (Figure 3.2C). Four types of pathway libraries are provided. Users can choose metabolic pathways or all pathways (including signaling pathways) for integrated analysis. The other two types - metabolic pathways (metabolite only) and all pathways (gene only) allow users to perform pathway analysis for individual omics data.

The integration of transcriptomics and metabolomics data can also be explored using the Network Analysis module. The knowledge-based network integration has been established since MetaboAnalyst v4.0. However, such an approach excludes the high volume of unannotated MS features detected by HRMS. We added the support for data-driven network analysis by implementing the well-established debiased sparse partial correlation (DSPC) algorithm (165).

Briefly, networks are created using a graphical LASSO model to compute the partial correlation coefficients and P-values for every pair of features in the dataset (177). The result can be visually explored as an interactive network with node size corresponding to node degrees and edge thickness based on the correlations between two connecting nodes (Figure 3.2D). The DSPC network is applicable to both targeted and global metabolomics and can be accessed from either Network Analysis or Statistical Analysis module.

3.7 Extended knowledge bases

The underlying knowledge bases within MetaboAnalyst have undergone significant updates to ensure that users' inputs can be identified correctly and accurately. These improvements are summarized as below.

3.7.1 Compound database

The compound databases, used by the Enrichment, Pathway, Joint-Pathway and Network Analysis modules have been enhanced by updating chemical identifiers from HMDB (178), KEGG (179), PubChem (180) and ChEBI (181). We have also expanded the database by including an additional 197,854 lipids from RefMet (182) and LIPID MAPS (183).

3.7.2 Metabolite sets

Metabolite sets, which are groups of metabolites with shared biological functions or collective behaviors, regulations or structures, are the backbone of the Enrichment Analysis module. To enhance these sets, we have added 44 metabolite sets related to disease signatures found in fecal samples, as well as 1571 metabolite sets identified by RefMet (182) and LIPID MAPS (183). These metabolite sets have also been transformed into appropriate libraries for Functional Analysis

module for global metabolomics. Users can now identify perturbations in organism-specific metabolic pathways or metabolite sets from raw spectra or peak lists.

3.7.3 Pathway libraries

Pathway libraries are used by the Pathway, Joint Pathway and Network Analysis Modules. All KEGG pathways libraries have been updated with the latest information from KEGG using their API (179). Additionally, we have added five new species (*Plasmodium vivax*, *Chlorella variabilis*, *Klebsiella pneumoniae*, *Klebsiella variicola* and *Streptococcus pyogenes*) based upon users' requests. The global KEGG metabolic network has also been updated to the latest version for the Network Analysis and Functional Analysis modules.

3.8 Other features

3.8.1 Enhanced visualizations

We have systematically updated the interactive plots across several modules (Enrichment, Pathway, Statistical and Biomarker Meta-Analysis), including synchronized 3D scatter plots for Principal Component Analysis (PCA) and Partial Least Squares - Discriminant Analysis (PLS-DA), as well as interactive volcano plots, bar plots, pie charts, and 2D scatter plots using the powerful Chart.js library (https://www.chartjs.org/). Furthermore, we have enhanced several publication quality graphics in the Statistical Analysis module such as box plots, K-means and self-organizing map (SOM) overview plots. Finally, users can now customize the colors and shapes of groups or samples in many important images.

3.8.2 Improved compound name matching

To provide better support for lipidomics data, we have implemented a smart name matching algorithm to improve the mapping from a user uploaded list or table of lipid names with our internal compound database. This algorithm considers common lipid abbreviations used by the LIPID MAPS classification system (184) as well as variations in punctuation marks used by different companies or databases. Compound synonyms for all metabolites in our internal compound database have been complemented from HMDB, PubChem and LIPID MAPS. This algorithm is used in all compatible modules within MetaboAnalyst v5.0.

3.8.3 Automated batch effect correction

The utility function for batch effect and signal drift correction has been updated with eight algorithms-EigenMS (137), QC-RLSC (114), ANCOVA (138), RUV-random (139), RUV2 (140), RUVseq (141), NOMIS (142) and CCMS (143) for correction based on either the data itself, QC samples or internal standards. The highlight for this update is the 'automated' design that can automatically identify and perform the optimal correction for the results (102). Users can upload the batches individually or as a merged table with all data together. All applicable correction methods will be executed, and the best results indicted by the distance among the batches will be returned.

3.8.4 Merging technical replicates

Technical replicates improve the stability and reproducibility of global metabolomics (176). However, averaging signals across the replicates may not be the best approach. We developed a new utility function to handle technical replicates in MetaboAnalyst v5.0. For a certain feature with multiple replicates, if the missing proportion in the replicates is over 1/3, the coefficient of variation (CV) of the feature in these replicates will be evaluated (185). If the CV is over 1.0, this

feature will be considered ‘highly variant’ with an assigned value of zero. A kernel density estimator is also available for users to smooth their data.

3.8.5 Supporting new input formats

The mzTab-M is a standard quantitative metabolomics data format (166). The latest version of this data format (version 2.0) is now supported by MetaboAnalyst v5.0. The Metabolomics Workbench is one of the most popular data repositories for metabolomics (163). We have added support to allow users to easily perform analyses on published datasets deposited in the Metabolomics Workbench. Users simply need to input the study ID of their preferred dataset. MetaboAnalyst will then retrieve the deposited data table for further statistics, functional enrichment, biomarker or network analysis.

3.8.6 Streamlined data analysis

A major effort in v5.0 is to refactor the underlying software architecture to enhance the modular structure and to improve the interoperability among different modules. With this update, modules can be developed and tested more independently, and users can now switch to other compatible modules at the end of each analysis, therefore creating their own custom pipelines.

3.9 Implementation

The web component of MetaboAnalyst v5.0 is implemented using the PrimeFaces framework (https://www.primefaces.org/). The core functions and graphics are executed using R (v4.0.2) and are freely available from the GitHub repositories as MetaboAnalystR (https://github.com/xia-lab/MetaboAnalystR) and OptiLCMS (https://github.com/xia-lab/OptiLCMS). The main site of MetaboAnalyst is hosted on a Google Cloud Server (with 64GB RAM and eight virtual CPUs with 2.6 GHz for each) for general data analysis except for the raw data processing module. To

accommodate the computing demand for raw data processing, we have set up two additional computing nodes located at the McGill Data Center and Compute Canada through a collaboration with the GenAP project (genap.ca), respectively, with 1TB RAM and 50TB of storage in total. These two websites are linked with the main site. Users can choose whether to register an account to manage their jobs. A maximum of 40GB data volume is allocated for each project (at most 10 projects for each registered user). The job submission and scheduling are based on the Simple Linux Utility for Resource Management (SLURM) system. During the upgrade to v5.0, we have made every possible effort to ensure backward compatibility with v4.0. For those who still need to access MetaboAnalyst v4.0, we have made it available as a Docker image (https://github.com/xia-lab/MetaboAnalyst_Docker).

3.10 Comparison with other web-based tools

Several web-based tools are available for metabolomics data analysis. Here we compared MetaboAnalyst v5.0 with these tools as well as the previous two versions (v4.0 and v3.0). The main features and characteristics of different tools are summarized in Table 3.1. Compared to the previous versions, the v5.0 has significantly enhanced many features and is distinctive in raw data processing and functional analysis for global metabolomics. Among other web-based tools, XCMS Online is well-known for raw data processing (34). MetaboAnalyst compares favorably with XCMS Online in several aspects including optimized raw data processing and downstream statistical and functional analysis, while XCMS Online excels in compound annotations based on the METLIN database (186). Among the remaining tools, Workflow4Metabolomics (W4M) (187) is a Galaxy-based workflow which uses the XCMS package for raw LC–MS data processing. The default workflow does not include a parameter optimization step, although experienced users can customize the pipeline to include IPO (46). In addition, W4M supports other types of raw

metabolomics data including GC-MS and NMR. The two other tools - 3Omics (188) and NOREVA (189) mainly focus on metabolomics data integration and normalization, respectively. MetaboAnalyst v5.0 remains the most comprehensive web-based platform that enables user-friendly and streamlined metabolomics data analysis and interpretation.

3.11 Conclusion

We have implemented a fully automated workflow to perform optimized peak detection, alignment and annotation tasks for LC–MS1 data generated in global metabolomics. The workflow can be easily accessed via the user-friendly web interface of MetaboAnalyst v5.0 or can be installed locally as an R package. We have also enhanced functional analysis by allowing biological interpretation directly from any peak groups or patterns of interest. The functional meta-analysis module further enables users to integrate heterogeneous global metabolomics datasets for improved understanding. We have also updated the compound databases and pathway libraries to enable comprehensive functional analysis for a wide range of species. During the process, we have consolidated the majority of modules in terms of interface, graphics and code architecture to improve user experience and performance. Overall, MetaboAnalyst v5.0 has addressed important gaps in the current metabolomics data processing and analysis pipeline. In the future, we aim to support more vendor data formats for raw spectral processing and to support spectral deconvolution based on tandem MS data.

Table 3.1. Comparison of MetaboAnalyst (versions 3.0-5.0) with other web-based tools. Symbols used for feature evaluations with '√' for present, '-' for absent, and '+' for a more quantitative assessment (more '+' indicate better support)

Tools Name	MetaboAnalyst			XCMS Online	W4M	3Omics	NOREVA
	5.0	4.0	3.0				
Raw Spectral Processing							
Parameter Optimization	+++	-	-	+	-	-	-
Supported Algorithms	+++	-	-	++	++	-	-
Resumable Analysis	+++	-	-	-	+	-	-
Compound Annotation	+	+	-	+++	++	-	-
Statistical Analysis							
Univariate	+++	++	++	+	+	-	+
Multivariate	+++	++	+	+	+++	-	++
Clustering	+++	+++	++	+	+	-	-
Power Analysis	√	√	√	-	-	-	-
Time-series Analysis	√	√	√	-	-	-	√
Biomarker Analysis	√	√	√	-	-	-	-
Biomarker Meta-analysis	√	√	-	-	-	-	-
Functional Analysis							
Function Analysis (MS peaks)	+++	++	-	++	-	-	-
Enrichment Analysis (compounds)	+++	+	+	-	-	++	-
Pathway Analysis	+++	++	+	-	-	++	-
Functional Meta-analysis	+++	-	-	++	-	-	-
Integrative Analysis							
Unbiased Joint Pathway	+++	+	+	-	-	+++	-
Knowledge-based Network	++	++	-	-	-	++	-
Correlation-based Network	++	-	-	-	-	-	-
Other Features							
Data Normalization	++	+	+	-	+	-	+++
Missing Value Estimation	√	√	√	-	-	-	√
Technical Replicates Merging	√	-	-	-	-	-	-

- XCMS online: https://xcmsonline.scripps.edu/
- Workflow4Metabolomics (W4M): https://workflow4metabolomics.usegalaxy.fr/
- 3Omics: https://3omics.cmdm.tw/
- NOREVA: http://idrblab.cn/noreva/

Funding: Genome Canada, Génome Québec, US National Institutes of Health [U01 CA235493]; Natural Sciences and Engineering Research Council of Canada (NSERC); Canada Research Chairs (CRC) Program; Calcul Québec and Compute Canada as well as the funding from CANARIE (in part); Fonds de la Recherche du Québec - Santé (FRQS). Funding for open access charge: Genome Canada.

Conflict of interest statement. None declared.

Preface to Chapter 4

This chapter describes a comprehensive update to MetaboAnalystR, an R package toolkit used for metabolomics data processing, statistical analysis and functional interpretation. This chapter update version of MetaboAnalystR from version 3 to version 4. The aim of this chapter is to achieve Objectives 2 and 3, providing an auto-optimized workflow for LC-MS/MS data processing and integrating MS/MS for functional analysis. There are four new features included by this version of MetaboAnalystR. 1) An auto optimized DDA data deconvolution workflow to clean chimeric spectra; 2) A highly efficient SWATH-DIA data deconvolution pipeline for SWATH-DIA data; 3) Comprehensive MS/MS databases supporting diverse application purposes; 4) More accurate functional analysis by integrating LC-MS and MS/MS results. In this update, the auto-optimized LC-MS pipeline in version 3 has been updated and incorporated to the new functionalities of LC-MS/MS data processing. In this chapter, MetaboAnalystR could handle both LC-MS and LC-MS/MS raw data processing conjunctively. Results from raw spectral processing can be used for function analysis directly. Overall, MetaboAnalystR could bridge raw spectral processing to functional insights.

Chapter 4: MetaboAnalystR 4.0: towards a unified LC-MS workflow for global metabolomics

Zhiqiang Pang[1], Lei Xu[1], Charles Viau[1], Yao Lu[2], Reza Salavati[1], Niladri Basu[1], and Jianguo Xia[1,2*]

[1] Faculty of Agricultural and Environmental Sciences, McGill University, Ste-Anne-de-Bellevue, Québec, Canada

[2] Department of Microbiology and Immunology, McGill University, Montreal, Québec, Canada

*Correspondence:

jeff.xia@mcgill.ca (J.X.);

Tel.: +1-(514) 398-8668 (J.X.)

4.1 Abstract

Liquid chromatography – high-resolution mass spectrometry (LC-HRMS) has played a significant role in advancing metabolomics and exposomics. However, it remains challenging to perform data analysis especially in terms of raw spectra processing, compound identification and functional interpretation. A flexible yet comprehensive pipeline is urgently needed. Here we introduce MetaboAnalystR 4.0 as a unified workflow to address three computational bottlenecks in current LC-HRMS workflow: 1) an auto-optimized feature detection and quantification module for LC-MS1 spectra processing; 2) an efficient MS2 spectra deconvolution and compound identification module for both data-dependent or data-independent acquisition; and 3) a sensitive functional interpretation module integrating LC-MS1 and MS2 results. MetaboAnalystR 4.0 comes with a large collection of reference spectra databases and knowledge libraries to allow large-scale local processing. In benchmarking and case studies with other well-established platforms, MetaboAnalystR 4.0 has identified > 10% more high-quality MS1 and MS2 features; it has also significantly increased true positive rate of identification (> 40%) without increasing false positives; finally, pathway enrichment analysis integrating LC-MS1 and MS2 spectra from COVID-19 datasets has produced results that are better aligned with the literature report. MetaboAnalystR 4.0 represents a significant step toward a unified workflow for LC-MS based global metabolomics in the open-source R environment.

4.2 Introduction

Liquid chromatography - mass spectrometry (LC-MS) has been the main analytical workhorse for untargeted (global) metabolomics and exposomics (83). To facilitate quantitative analysis and compound identification, LC-MS experiments are typically conducted with MS full scans coupled with tandem MS or MS/MS using data-dependent acquisition (DDA) or data-independent acquisition (DIA) methods (67, 190). DDA acquires MS/MS spectra by fragmentation of precursor ions selected using a relatively narrow MS/MS isolation window (e.g., 1 *m/z*). Although DDA spectra are directly linked to precursors, recent studies show that over 50% of them are 'chimeric' and need to be deconvolved before searching any reference database (190). DIA usually fragments all ions in a wider *m/z* range (e.g., >15 *m/z*) with multiple cycles to improve the coverage on the metabolome. SWATH-MS (sequential window acquisition of all theoretical fragment ion spectra mass spectrometry) is a common DIA approach for both metabolomics and proteomics (191). Spectral deconvolution is essential to relink precursors with fragment ions in DIA.

Raw data processing in LC-MS based metabolomics starts with MS and MS/MS feature detection. Putative compound identifications are performed by matching m/z values and retention times of MS1 features, as well as their associated MS2 patterns against reference spectral databases. The process often returns more than one candidate and requires further time-consuming, manual curation before one can perform functional interpretation. It has been shown recently that functional activities can be reliably predicted based on the global patterns of putative identification results despite uncertainties at individual compound level (101, 192).

Several powerful algorithms and tools have been developed to process both LC-MS and MS/MS spectra using different strategies (35, 38, 39, 51, 76, 81, 89). DIA usually utilizes either spectral library or extracted chromatogram to deconvolve multiplexed spectra (39, 67, 76). *Pseudo*-MS2

based deconvolution could reconstruct MS/MS *de novo* without depending on pre-defined database, thereby improving the identification and coverage over unknowns (76). MS-DIAL is the first to deconvolve DIA by linearly decomposing the chromatographic profiles with three neighboring model peaks to reconstruct MS/MS spectra of precursors (40). DecoMetDIA further extended this approach by using multiple model peaks-based deconvolution (76). However, the computational cost of DecoMetDIA is high, making it unsuitable for high-throughput analysis. Similar to DIA, deconvolution of DDA data is also necessary (193). The group who reported prevalent contamination in DDA also presented a solution - DecoID (190) to remove the contamination by using a spectral library-assisted linear regression model based on LASSO (least absolute shrinkage and selection operator). However, the deconvolution is not suitable if any convolved component(s) are missing. In addition, given the complexity of MS/MS spectrum, LASSO regression model parameter needs to be manually tuned to effectively clean the contamination. Finally, integration of MS1 and MS2 results for further downstream functional analysis remains a key bottleneck in LC-MS based metabolomics.

Here we introduce MetaboAnalystR 4.0 as a unified framework for processing raw LC-MS and MS/MS data and integrating the results for deep functional insights. By leveraging and consolidating the comprehensive statistical and functional analysis functions, as well as LC-MS data processing workflow established in the previous version (102, 105, 122), version 4.0 aims to address the key demands from our user community to support MS/MS spectra processing and integrative analysis. It contains four key new features:

1) An auto optimized DDA data deconvolution workflow to remove contamination signals in chimeric spectra.

2) A highly efficient SWATH-DIA data deconvolution pipeline to process the SWATH-DIA data.
3) Comprehensive MS/MS databases curated from all public database that can support diverse application purposes.
4) Accurate functional activity prediction by integrating LC-MS and MS/MS results.

4.3 Results

4.3.1 General workflow

The LC-MS/MS data processing workflow in MetaboAnalystR involves several steps, including raw spectral data import, MS data processing (auto-optimized peak picking, alignment, gap filling and annotation) (102), DDA/SWATH-DIA data deconvolution, spectrum consensus from replicates, MS/MS reference library searching, results export, and integration into functional prediction. The workflow is depicted in Figure 4.1a.

For DDA spectral data, MetaboAnalystR assigns all MS/MS spectra of an individual spectral data into different feature groups based on precursors' information (*m/z*, RT), and chimeric status is evaluated based on the nearest MS scan. The MS/MS spectra of all ions (including main precursor and other contaminating ions within the isolation window and above the intensity threshold) are extracted from reference libraries as candidate spectra (Methods, Figure 4.1b). If any reference spectrum is missing, a predicted spectrum will be generated using a similarity-network model (84) (Methods, Figure 4.1c). All candidates are then used to obtain the deconvolved spectrum based on elastic-net regression with extra penalties to the predicted spectra (Methods). The SWATH-DIA MS/MS data processing module of MetaboAnalystR has been developed based on the DecoMetDIA (76). The entire deconvolution workflow was written using Rcpp/C++ framework

and further optimized to enable high-throughput processing. In addition, MetaboAnalystR supports multi-threaded data processing to further speed up the analysis through parallel computing.

After the deconvolution step, all deconvolved spectra for a specific MS feature from replicates (if any) are then subjected to consensus - a step aiming to produce single spectrum from technical/biological replicates to reduce potential errors and noise. This process can optionally be performed in a database-assisted manner to prevent over-fitting from a hard-coded cut-off value (Methods, Figure S4.1).

The consensus spectral results are submitted to database searching for compound annotation (Methods). Users can choose different databases based on MS instrument type, collision energy, and other database options. The dot-product or spectral entropy similarity method is used to evaluate MS/MS matching similarity (194). The candidates for a particular feature are scored by considering *m/z*, RT, isotope, and MS/MS similarity together, based on the rule from MS-DIAL (40). The matching score ranges between 0 and 100, where 0 indicates negative matching and 100 indicates perfect matching (Methods). The top N chemical candidates, defined by the user, can be exported as the database searching results. If the matching score is below 10, MetaboAnalystR optionally performs neutral loss scan to further improve compound annotation (195).

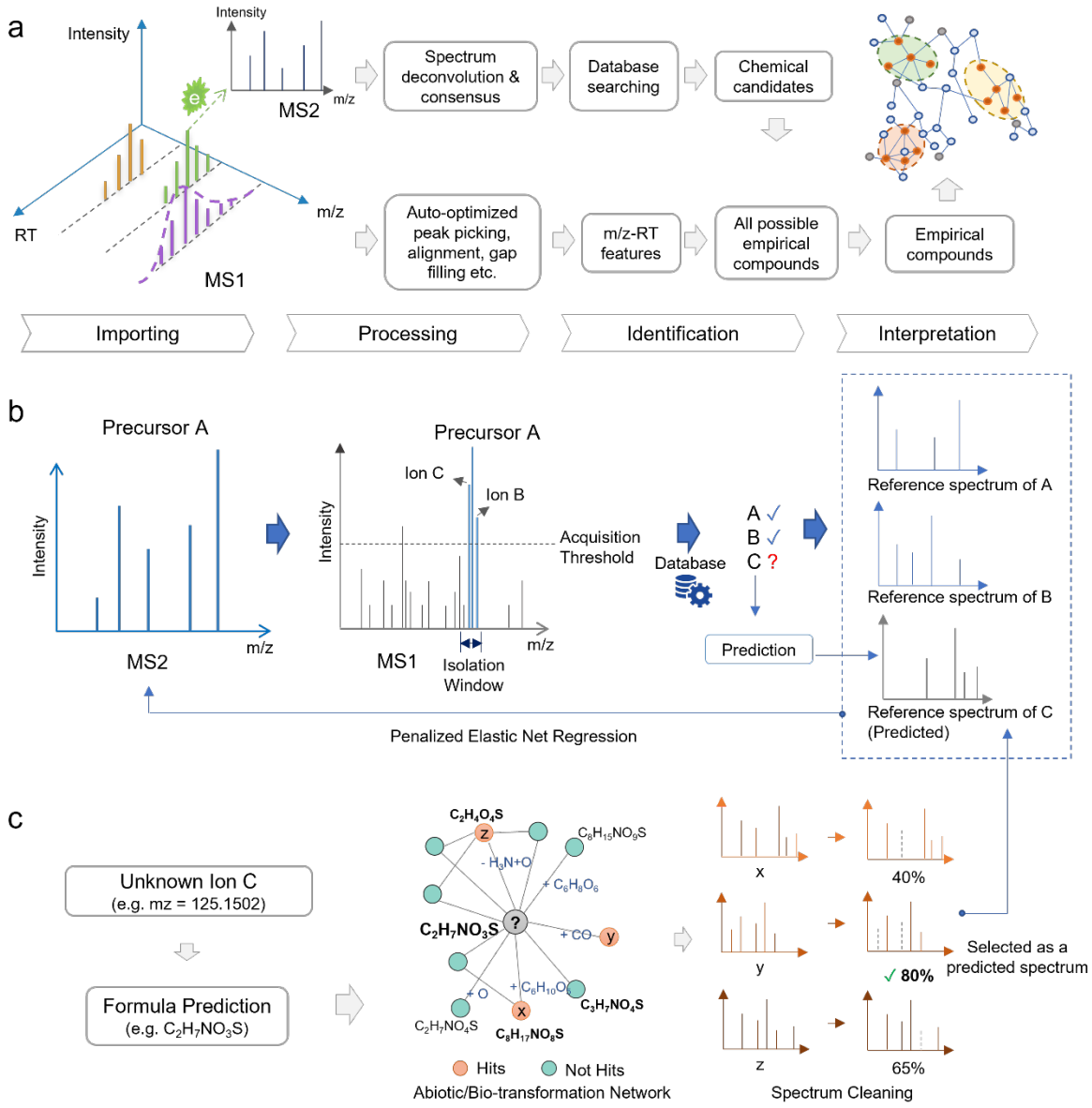


Figure 4.1. Implementation of MetaboAnalystR for LC-MS/MS data processing and biological interpretation. a. Raw spectra data processing workflow. MetaboAnalystR accepts common open-source formats. Centroid format is highly recommended for both MS and MS/MS. Mass spectral signals are processed separately from MS and MS/MS levels. All detected MS features are mapped as empirical compounds, and filtered based on the chemical candidates from MS/MS. The resulting clean empirical compounds list would be permuted to predict biological functions. b. Workflow

of DDA spectral deconvolution. All spectra acquired by DDA are evaluated as “clean” or “chimeric” based on the MS signals of the nearest MS scan. The reference spectra of all ions (A, B and C) acquired within the isolation windows and above the intensity threshold are extracted from reference MS/MS library for regression analysis. c. Diagram of reference spectrum prediction strategy. MetaboAnalystR predicts one or more candidates missing in the MS/MS library. Formula(s) of the ion (e.g., ion C) is predicted at first. An abiotic/bio-transformation network is constructed for the formula (e.g., C2H7NO3S), and all neighbors with reference spectra of the formula are extracted as a list. Each fragment of a single spectrum in the list is predicted into formula. If the formula (e.g., C3H9) of the fragment includes more chemical elements (number or type) than the original formula (C2H7NO3S), it is excluded from the spectrum. The clean spectrum is returned into the list. The similarities to original MS/MS spectrum (the most left one in b.) of all spectra in the list are evaluated, and the one with the highest similarity score is selected as the predicted spectrum for the ion (e.g., ion C).

MetaboAnalystR 4.0 offers comprehensive database options to facilitate high-throughput MS/MS spectra processing and compound annotation. A total of five databases are provided, including pathway compound database, biological compound database, lipids database, exposomics database and the complete database. All these databases are curated from public MS and MS/MS data, including HMDB (196), MoNA Series (87), LipidBlast (91), MassBank (87), GNPS (89), LipidBank (94), MINEs (90), LipidMAPs (184), KEGG (92, 179) (Methods). The summary of these five databases is provided in Table S4.1.

Table 4.1. Summary of identified compounds by different tools (DDA, ESI^+)

Tools	Detected (MS1)	Annotated (MS2)	Percentage (%)	CPU Time (min)
MS-DIAL/MS-FINDER	271	121	26.4	36
MZmine/SIRIUS	317	124	27.0	84
MetaboAnalystR (deco)	**336**	**194**	**42.3**	**28**
MetaboAnalystR (non-deco)	336	185	40.3	15

Obtaining functional insights underlying the observed phenotypic differences is among the main objectives of most metabolomics studies. However, conventional approaches generally require manual annotation of a significant portion of spectral features, which is a very time-consuming process. The bottleneck has been addressed by the *mummichog* algorithm (101). We have recently shown that the algorithm can significantly improve the accuracy and specificity in pathway activity prediction by leveraging HRMS and MS/MS (192). This algorithm has been enhanced in MetaboAnalystR 4.0 by introducing MS/MS-based compound identification results lists to filter out impractical chemical possibilities. Briefly, after MS and MS/MS spectral data processing, MetaboAnalystR can automatically perform statistical analysis from the peak intensity table and format the database searching results for functional enrichment analysis (see Figure 4.1a and Methods). The functional analysis is based on the known biological functional databases curated from KEGG (179), BioCyc (197), *etc*, supporting >120 species.

4.3.2 Benchmarking and validation

Using a total of seven datasets including three standard mixtures (190, 198, 199), one serial dilution data, one whole blood exposomics data, and two COVID-19 plasma metabolomics datasets (30, 200), we benchmarked the performance of MetaboAnalystR together with other widely used tools. For DDA workflow, we included MS-DIAL/MS-FINDER (39, 40) and MZmine (38)/SIRIUS (81), while for the SWATH-DIA workflow, we included MS-DIAL/MS-FINDER and XCMS

(35)/SIRIUS pipelines because MZmine does not support deconvolution on LC-SWATH-DIA data at the moment.

4.3.3 Characterizing performance of compound annotation with standard mixtures

We firstly used three standard mixture (SM) datasets with different complexities from Mass Spectrometry Metabolite Library (MSMLS, IROA Technologies). The 1st SM data include 15 DDA samples (198). Each contains 10-15 non-isobaric compounds (simple mixture). The 2nd SM data contain a mixture of 526 standards (complex mixture), including one DDA spectrum and one SWATH-DIA spectrum (190). The 3rd SM data contain 91 compounds (199). Both DDA and SWATH-DIA (ESI^+ and ESI^-) modes are included, with three replicates for each mode.

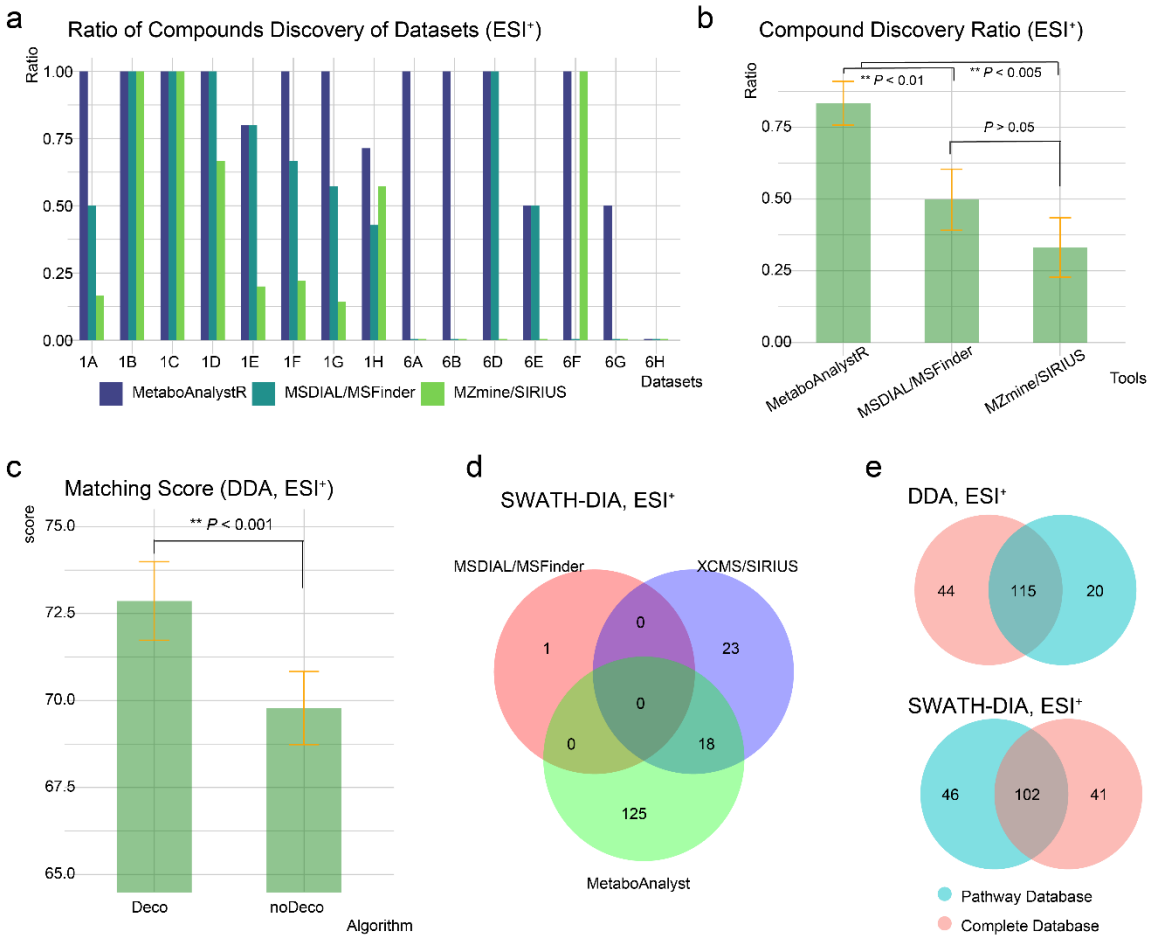


Figure 4.2. Validation of MetaboAnalystR with standard mixtures. a. Compound discovery ratio of simple standards mixture samples in three workflows (ESI^+ mode). For all samples, MetaboAnalystR could detect the highest ratio of compounds as the top first candidate. b. Statistical analysis of the compound discovery results. Compared to other two workflows, MetaboAnalystR reported significantly higher compound discovery ratio ($P < 0.01$). c. Comparison of matching scores of DDA (w/o deconvolution, ESI^+) in complex standard mixture sample. In contrast to the non-deconvolved spectra, the deconvolution algorithm in MetaboAnalystR could significantly improve the matching score of chemical candidates (paired t-test, $P < 0.001$). d. Venn Diagram of compounds identified from the complex standard mixture by different tools (SWATH-DIA, ESI^+). e. Performance evaluation of compound discovery with different reference libraries by MetaboAnalystR. The majority of the compounds identified with different database are shared for both DDA and SWATH-DIA (ESI^+).

Using the 1st SM data (198), MetaboAnalystR always detected most correct compounds as the top first candidate (ESI^+, Figure 4.2a), with significantly higher compound discovery rate than MS-DIAL/MS-FINDER and MZmine/SIRIUS (ESI^+, Figure 4.2b). Similar results were also observed in ESI^- mode (Figure S4.2). Using the DDA samples in 2nd SM data (190), MetaboAnalystR could find the highest number of MS features (RT deviation < 20 sec, *m/z* error < 10 ppm) and compounds in an efficient way, compared to MS-DIAL/MS-FINDER and MZmine/SIRIUS pipeline (Table 4.1). Deconvolution increased the number of compounds correctly identified as the top candidate (Table 4.1, Table S4.2), and the matching scores reported by deconvolution pipeline were significantly improved in comparison to the non-deconvolution pipeline (Figure 4.2c).

For SWATH-DIA samples in the 2nd SM data, MetaboAnalystR correctly detected most compounds (ESI^+, Figure 4.2d, Table S4.3). In this study, MS-DIAL/MS-FINDER did not find sufficient MS features with default parameters as only one compound was discovered (ESI^+, Figure 4.2d, Table S4.3). However, XCMS/SIRIUS workflow reported 51 compounds correctly, with 23 of them not detected correctly by MetaboAnalystR. Similar results were also observed in ESI^- mode (Figure S4.3, Tables S2 and S4).

4.3.4 Effects of reference spectral databases

Using the 2nd SM data (190), we compared the results obtained using the pathway reference library or the complete reference library. Only 75% compounds in the standards mixtures are included in the pathway reference library, while all of them are contained in the complete reference library. As shown in Figure 4.2e, most compounds identified correctly are shared in both libraries. For DDA dataset (ESI^+), MetaboAnalystR could identify two times more compounds using the complete library compared to using the pathway library. For SWATH-DIA dataset (ESI^+), the numbers of unique compounds identified by both libraries are similar (41 vs. 46). Similar results are found in ESI^- mode (Figure S4.3).

4.3.5 Evaluating false discovery rate

Using the 3rd SM data (199), we first tested the number of compounds that could be identified correctly from both acquisition and ion modes. For DDA dataset, MetaboAnalystR could correctly identify most compounds by using either the complete database or the pathway database compared to other tools (for both ESI^+ and ESI^-, Figure 4.3a). For the data from ESI^- mode, MZmine/SIRIUS only identified one compound correctly. However, for SWATH-DIA dataset, MetaboAnalystR

identify most compounds from ESI^+ mode, but not from ESI^- mode (4 compounds fewer than MS-DIAL/MS-FINDER, even with the complete library, Figure 4.3b).

To evaluate the false discovery rate, we generated a series of decoy spectral data by randomly increasing *m/z* error and replacing the original MS/MS spectra with synthetic spectra from isobaric compounds (Methods, Figure 4.3c). Both DDA and SWATH-DIA workflows were tested with decoy spectra datasets (Methods, Figure 4.3c, Figure S4.4). The results showed that MetaboAnalystR did not produce significantly higher number of false positives in comparison to other tools (Figure 4.3d-e). For SWATH-DIA dataset, MS-DIAL/MS-FINDER detected more false positives. Similar results were observed in ESI^- modes (Figure S4.5). This study indicates that MetaboAnalystR could significantly improve chemical identifications without increasing false identifications, independent of the reference library.

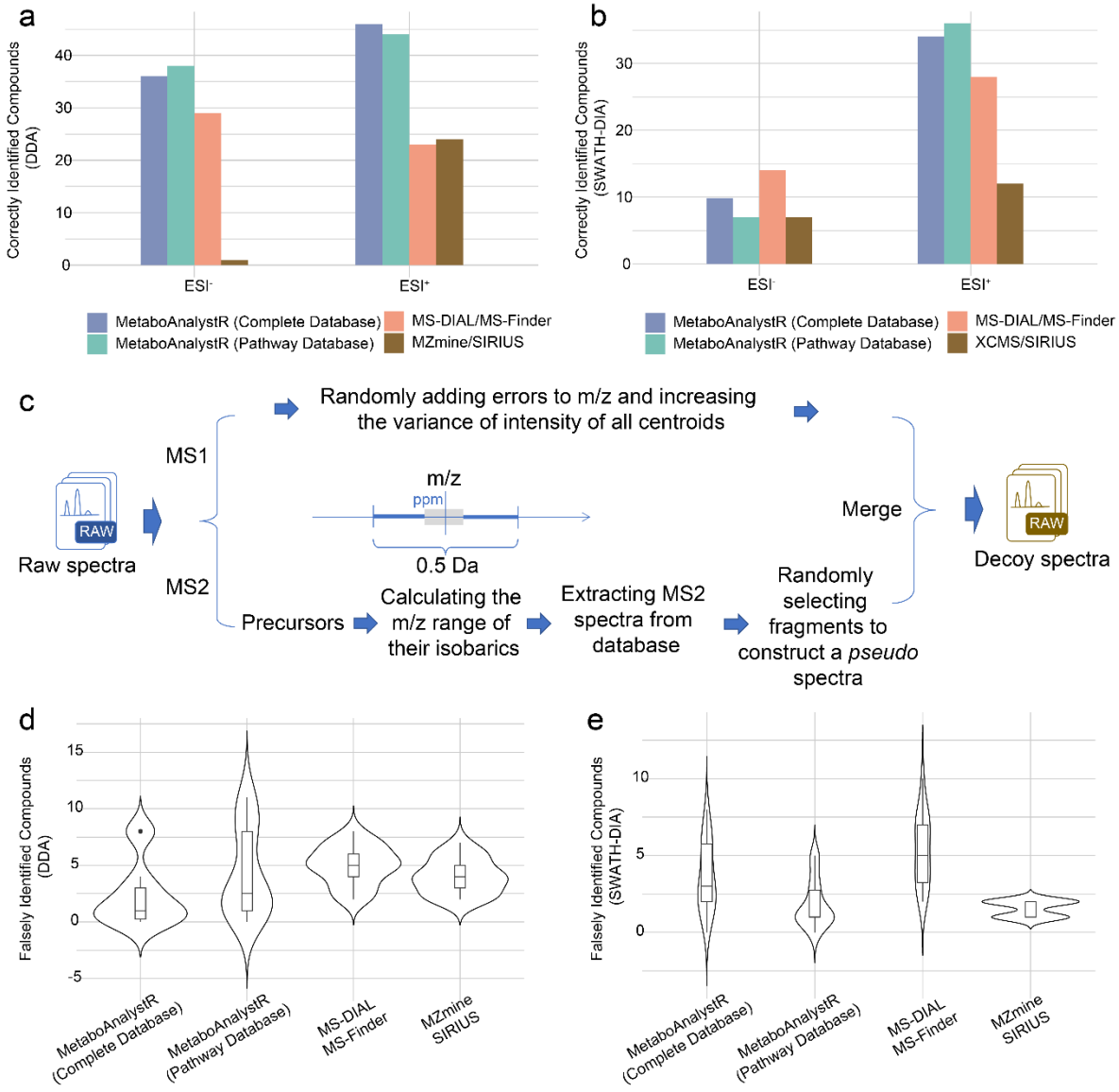


Figure 4.3. Validation of MetaboAnalystR with standard mixtures and false discoveries. a. Statistics of correctly identified compounds from DDA dataset. b. Statistics of correctly identified compounds from SWATH-DIA dataset. c. Workflow to generate decoy spectra data (DDA). d. Falsely identified compounds from decoy spectra data (DDA, ESI^+).

4.3.6 Characterizing unique metabolome of different types of blood samples

To evaluate the performance of compounds annotation in real biological samples, we conducted LC-MS/MS based metabolomics including both DDA and SWATH-DIA on paired blood samples of different types (plasma, serum and whole blood) and compared their chemical differences. Understanding the unique chemical compositions of different types of blood samples are important. For instance, compared to serum or plasma, whole blood can provide extra insights for many critical illnesses such as sepsis (201) and COV-SARS-2 infection (202) that are closely related to blood cellular components.

The general design of this study is shown in Figure 4.4a. Iterative targeted DDA was optimized with HERMES (78) to improve the coverage on the metabolome (Methods). All MS features were detected with MetaboAnalystR and other workflows. Overall, the metabolomes of the three different blood samples show significant intrinsic chemical difference (PCAs, Figure S4.6). In this study, we focused on elucidating the distinct features among different blood types (Figure 4.4b, Figure S4.7-4.10). All these "unique features" are used as target lists for MS/MS-based compound identification. For both DDA and SWATH-DIA, MetaboAnalystR identified highest number of compounds compared to other tools, especially from SWATH-DIA datasets (Figure 4.4c), indicating SWATH-DIA spectral data processing workflow in MetaboAnalystR could improve the chemical identification rate (level 2a (203)).

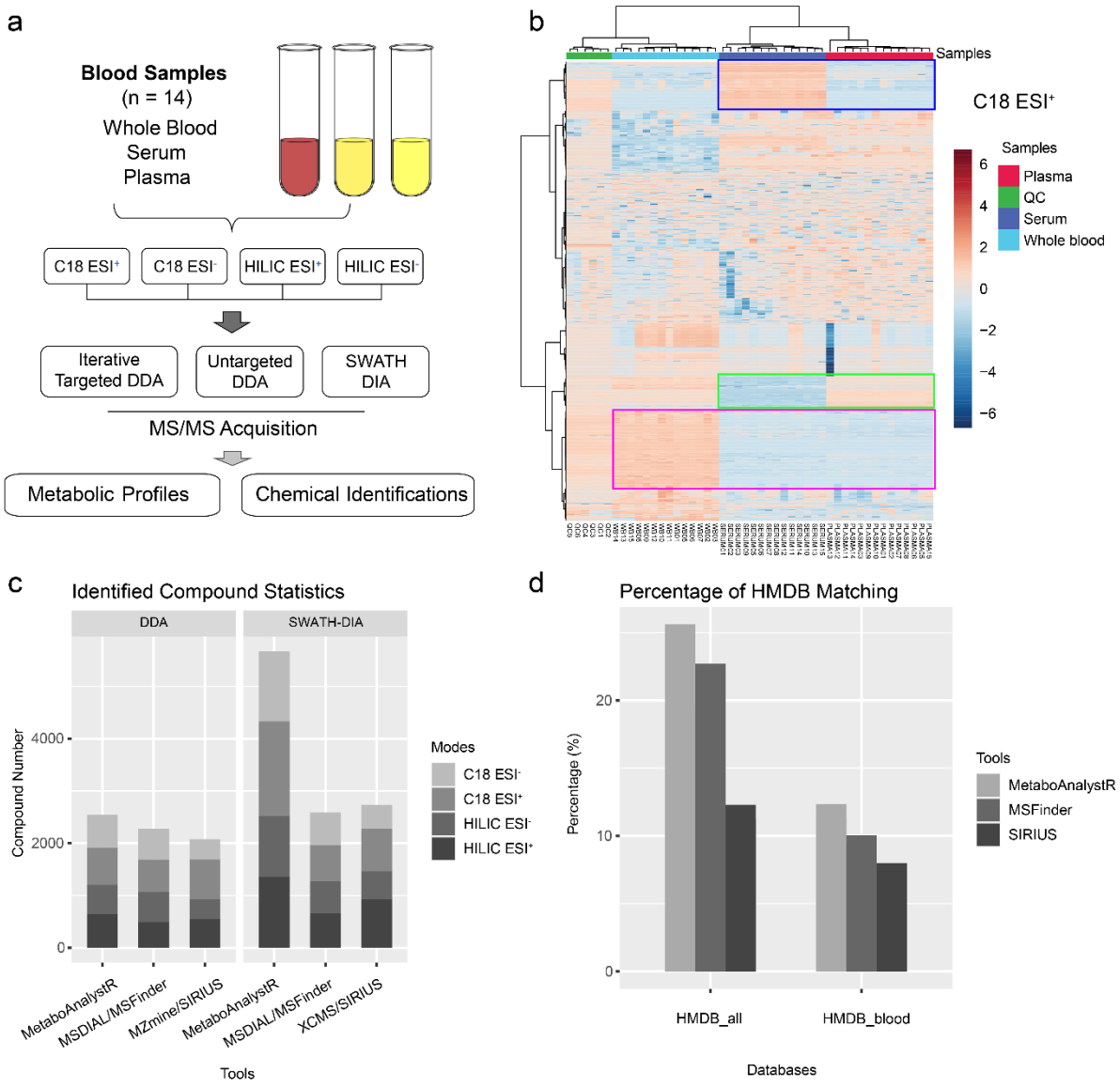


Figure 4.4. Comparison of chemical identification from different blood samples. a. Experiment design of metabolomics study. b. Heatmap of complete metabolic profiles (MS level, C18, ESI^+). Unique MS features for a specific blood type were highlighted with rectangles. Blue, unique features for serum compared to plasma; Green, unique features for plasma compared to serum. Ruby, unique features for whole blood compared to plasma and serum. c. Summary of all

compounds identified by different tools from all modes (including both DDA and SWATH-DIA). d. Percentage of HMDB matching of compounds identified by different tools.

To evaluate the validity of all chemical identifications, all identified compounds results from different tools were matched to HMDB database and HMDB blood database (Methods). As shown in Figure 4.4d, compounds identified by MetaboAnalystR had the highest percentage (and absolute number) of compounds matched into databases compared to other tools. The chemical composition analysis showed that whole blood sample contains more lipids, organic acids and organic heterocyclic components in contrast to serum and plasma (Methods, Figure S4.11). The main chemical difference between serum and plasma are lipids (Figure S4.11), which is expected based on the blood coagulation of serum and consistent to the previous study (204).

4.3.7 Evaluation on quantitative performance with serial dilutions

To benchmark the quantitative performance of MetaboAnalystR, a dilution series was prepared by mixing serum and urine in a cross-gradient manner (205) (Figure 4.5a, Methods). The unique features detected from MS1 level (Figure 4.5c) in serum or urine were extracted as the target features for further analysis. Other features shared by urine and serum are considered as generic features; therefore, they are excluded to evaluate the quantitation performance.

Correlation analysis of MS features' intensities with dilution ratios was performed. Compared to XCMS, MS-DIAL and MZmine, unique features detected by MetaboAnalystR showed the highest average correlation coefficient (C18, ESI^+, Figure 4.5b). Similar results were also found from other modes (C18 ESI^-, HILIC ESI^+ and HILIC ESI^-, Figure S4.12), indicating that quantification by

MetaboAnalystR performs better than others. The serial dilution patterns could be observed clearly from heatmaps for all modes (Figure 4.5c and Figure S4.13).

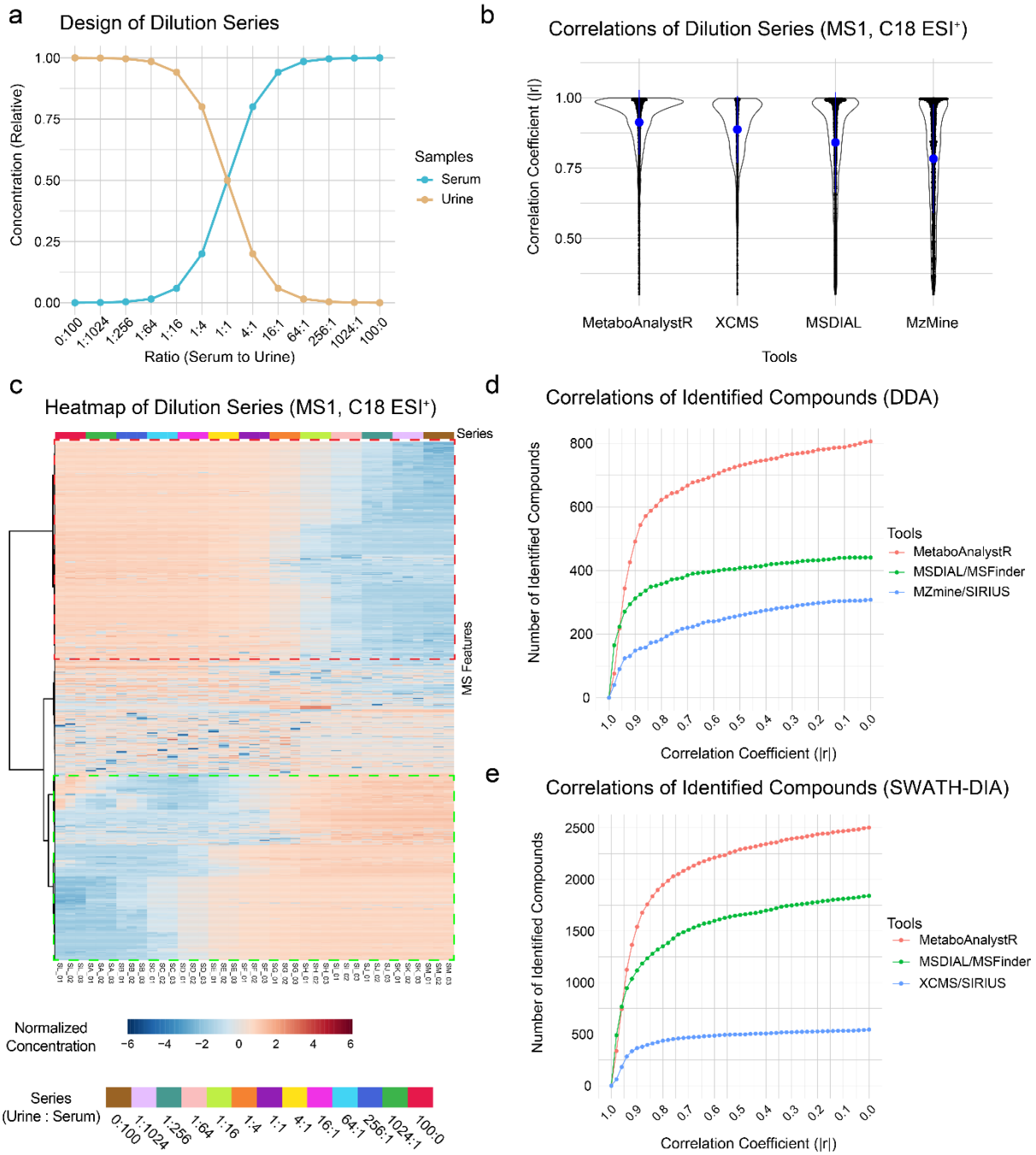


Figure 4.5. Evaluation of quantitative and qualitative performance based on serial dilutions. a. Design of serial dilutions. Urine and serum are mixed according the ratio labelled at x-axis. b. Correlation analysis of MS features from serial dilutions (detected by different tools under C18 ESI^+ mode). MetaboAnalystR reported highest average correlation coefficients compared to other tools. c. Heatmap of all MS features (MetaboAnalystR, C18 ESI^+). All features are normalized and clustered. Samples are sorted based on the dilution series. Pure/undiluted urine and serum samples are also included. Unique features for urine and serum are highlighted with red and green dashed rectangles respectively. d. Association between the number of identified compounds from DDA mode and correlation coefficient cut-offs. e. Association between the number of identified compounds from SWATH-DIA mode and correlation coefficient cut-offs.

To further assess the annotation performance from MS/MS-based chemical identification, all unique features were extracted as the targets. As the correlation coefficient threshold relaxing, the number of compounds identified by all tools was increasing. In comparison, for both DDA and SWATH-DIA datasets, MetaboAnalystR chemically identified the highest number of compounds (at level 2a) which following the dilution gradients. MS-FINDER workflow identified more compounds than SIRIUS from both DDA and SWATH-DIA datasets. Compared to DDA, the SWATH-DIA based dataset showed a higher coverage (of the chemical identification) on the gradient features. In brief, MetaboAnalystR could more effectively quantify and annotate the MS and MS/MS features.

4.3.8 Biological interpretation of COVID-19 metabolomics data

We used two COVID-19 metabolomics datasets to showcase the complete workflow of MetaboAnalystR - from raw LC-MS and MS/MS spectra to biological interpretation. The 1st dataset (30) includes a total of 160 samples both polar metabolites and non-polar lipids datasets (ESI^+ and ESI^-, DDA, Methods) categorized into COVID-19 with multiple severities. While the 2nd dataset (200) includes 30 samples (ESI^+ and ESI^-, SWATH-DIA, Methods) categorized into COVID-19 and healthy control.

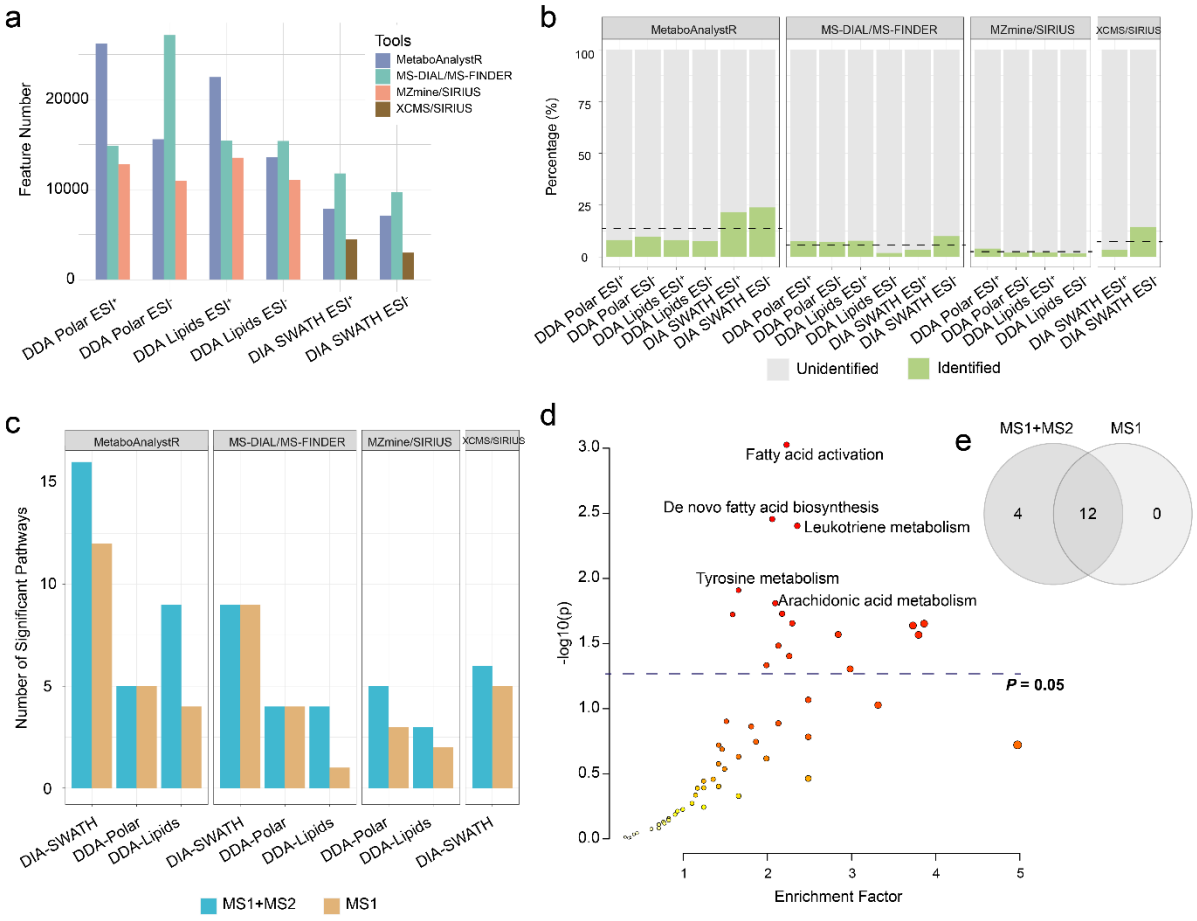


Figure 4.6. Interpretation of biological insights of COVID-19. a. Summary of MS features detected by different tools from DDA and SWATH-DIA datasets. b. Percentage of MS features which has been identified with MS/MS spectra. c. Summary of pathway prediction results from

different tools by integrating MS/MS identification results (MS+MS/MS) or not (MS_only). d. Scatter plot of pathway enrichment analysis results of SWATH-DIA dataset (processed by MetaboAnalystR). e. Venn diagram of pathway analysis results from SWATH-DIA dataset (processed by MetaboAnalystR). A total of 12 pathways are reported as significant by both methods. However, integrating MS/MS results into pathway prediction added another 4 significant pathways.

Table 4.2. Comparison of computational performance of different tools. The values are presented as relative values to the MS-DIAL/MS-FINDER workflow.

Comparisons	Tools	MS			MS/MS			Mean
		DDA Polar*	DDA Lipids	SWATH DIA	DDA Polar	DDA Lipids	SWATH DIA	
Clock Time Elapse	MS-DIAL/MS-FINDER	1	1	1	1	1	1	1
	MetaboAnalystR	**0.76**	**0.86**	**0.61**	**0.33**	**0.55**	**0.43**	**0.59**
	MZmine/SIRIUS	0.53	0.57	-	0.91	1.39	-	0.85
	XCMS/SIRIUS	-	-	0.33	-	-	0.54	0.44
RAM Usage	MS-DIAL/MS-FINDER	1	1	1	1	1	1	1
	MetaboAnalystR	**1.04**	**1.45**	**1.34**	**1.17**	**1.22**	**0.56**	**1.13**
	MZmine/SIRIUS	2.38	3.57	-	2.57	2.02	-	2.64
	XCMS/SIRIUS	-	-	0.91	-	-	2.13	1.52

* Polar, polar compounds

MS features were detected at first by different workflows. According to the summary of MS features (Figure 4.6a), different tools showed significantly distinct sensitivities. For metabolites (ESI^+) and lipids (ESI^+), MetaboAnalystR detected the highest number of MS features; while for the other four datasets, MS-DIAL detected the highest number. In contrast, MZmine and XCMS did not detect as many MS features as MetaboAnalystR and MS-DIAL. We acknowledge that

manual parameter tunning by expert users may change the sensitivities of these algorithms to some extent, while MetaboAnalystR performs automated parameter tuning to avoid such needs.

Chemical identification was performed by different tools with the corresponding MS features detected above as targets (Figure 4.6a). MetaboAnalystR could identify the highest percentage of compounds from MS features (level 2a, Figure 4.6b). However, for DDA metabolite ESI^- dataset, MS-DIAL/MS-FINDER identified most compounds based on the absolute number (Figure S4.14), indicating MS-DIAL/MS-FINDER can be used to complement MetaboAnalystR in some situations.

Next, we applied the enhanced *mummichog* algorithm (Methods) to the results generated by the three tools. In comparison to the previous version, the enhanced algorithm leverages MS/MS-based chemical identifications to filter out impractical compound assignments to improve pathway activity prediction. We chose to compare biological differences between Mild and Fatal COVID-19 cases for the 1st data (DDA), and between COVID-19 case and healthy controls for the 2nd data (SWATH-DIA). There are two sub datasets (polar and non-polar lipids) in 1st data and one dataset in 2nd data. Therefore, a total of nine comparisons were performed for different workflows. As shown by COVID-19 datasets (Figure 4.6c). Six out of nine comparisons reported more pathways if both MS and MS/MS data were used. In addition, the enhanced *mummichog* algorithm predicted more pathways using spectral processing results from MetaboAnalystR compared to other tools. For example, 16 pathways were predicted by MetaboAnalystR from DIA-SWATH dataset if both MS and MS/MS are utilized together (Figure 4.6d). Compared to the prediction based only on MS features, four more pathways were reported as significantly perturbed (Figure 4.6e). These pathways are related to phosphatidylinositol phosphate, vitamin D, vitamin C metabolism and arachidonic cascades (Table S4.5). They have been reported to be related to the

pathogenesis of COVID-19 by previous studies (206-209). All pathway predicting results are provided in Figure S4.15-4.18 and Table S4.5-4.7.

4.3.9 Computational performance assessment

To assess the computation performance, the same COVID-19 datasets were used to compare the difference of computational performance of all tools included in this study. This assessment was performed in a standard workstation (Dell OptiPlex 7070, 64GB RAM, Intel-i7-9700 CPU, Ubuntu 20.04.2). To ensure the fairness, this comparison was performed and controlled with Simple Linux Utility for Resource Management (SLURM44, Methods).

In comparison to other tools, MetaboAnalystR could finish MS features detection at a similar speed to MZmine and MS-DIAL. MetaboAnalystR is slower than XCMS for MS feature detection on SWATH-DIA dataset due to the extra parameter optimization step (Table 4.2 and Figure S4.19-4.20). The RAM usage of MetaboAnalystR is approximative to MS-DIAL, but obviously lower than MZmine/SIRIUS or XCMS/SIRIUS. As for the MS/MS data processing, MS-FINDER and SIRIUS is significantly slower than MetaboAnalystR. According to the execution logs, MS/MS spectra searching in SIRIUS is based on API access to remote web services, which is highly depending on the network traffic and responsiveness of remote server. MS-FINDER also partially uses remote access to predict formulas. Different from SIRIUS and MS-FINDER, raw spectra processing and annotation in MetaboAnalystR is all based on local databases to allow users to detect MS features and annotate MS/MS features in a highly efficient way.

4.4 Methods and materials

4.4.1 Chemicals

Standard human serum, ammonium acetate (NH_4AC) was purchased from Sigma-Aldrich (Sigma, St. Louis, MO, USA). Acetonitrile (ACN) and methanol (MeOH), 0.1% formic acid (FA) in Water, 0.1% FA in ACN and pure water were purchased from Fisher Chemical (Morris Plains, NJ, U.S.A.).

4.4.2 Sample preparation of bloods

Healthy volunteers were recruited from McGill University as previously discussed (210). About five milliliters of venous whole blood were drawn from each volunteer into a BD K2-EDTA Trace Element free Vacutainer. A sub-sample of this whole blood was used to obtain plasma (i.e., whole blood centrifuged for 15 min at 4 ℃ at 2,700 rpm). From each individual, another sample of blood was collected into a BD Vacutainer tube not containing any anticoagulant, which was allowed to sit for ~30-60 minutes for clots to form following which serum was obtained by centrifugation (15 min at 4 ℃ at 2,700 rpm). Blood samples from 14 individuals were collected and included for this study. All blood samples are paired with three different types (whole blood, serum and plasma). All samples were immediately frozen at -80 ℃ until analysis. The demographic information of all subjects is summarized in Table S4.8. This study was approved by Research Ethics Office of McGill University (Study ID: A05-M26-16B).

The different blood sample types were prepared based on the previously published protocols (211, 212). The three blood sample types (WB, serum and plasma) were thawed on ice for 1 hour, and then vortexed for 30 seconds to ensure homogeneity. 100 µL of each sample type was transferred to a 1.5 mL Eppendorf microcentrifuge tube, to which 400 µL of -20°C 1:1 ACN:MeOH (v/v) was added. Samples were next vortexed for 60 seconds and stored at -20°C for 1 hour. Samples were then centrifuged at 16,100 × g for 10 minutes at 4°C. The supernatants were collected (250 µL) and filtered by centrifugation using 0.2 µm Nanosep centrifugal filters (PALL Life Sciences) at 14,000 × g for 15 minutes at 4°C. Filtered samples (120 µL) were then transferred to LC-MS vials

equipped with 250 µL glass inserts and run in the LC-MS. A Quality Control (QC) sample was made by pooling equal volumes from each filtered all samples supernatant into one 1.5 mL Eppendorf microcentrifuge tube.

4.4.3 Sample preparation of serial dilutions

A urine sample was collected from a donor of McGill University (210). A total of 100mL urine was sub-sampled and frozen at -80 ℃. A human standard serum sample (Sigma-Aldrich, Sigma, St. Louis, MO, USA) and the urine were thawed in ice for 1 hour, and then vortexed for 30 seconds to ensure homogeneity. A total of 13 Eppendorf microcentrifuge tubes were prepared and labelled from A to M. For tubes A to E, 150 µL urine was transferred into each of them. For tube G to K, 150 µl standard serum was transferred into each of them. 250 µL pure urine were transferred into tube L, and 250 µl pure serum were transferred into tube M. Then, 50 µL pure serum were extracted and mixed into tube E. Then, 50 µl liquids were extracted and mixed into tube D, and so on until tube A. Same operations were repeated for pure urine tube L, and tube G-K. Finally, 75 µl pure urine and serum were extracted respectively and mixed into tube F. At a result, a total of 11 dilution mixtures and 2 pure samples were generated. The whole preparation workflow is shown in Figure S4.21. After the preparation of the serial dilutions, all samples were processed similarly as the samples of blood. But the ratio of organic reagents to samples is 1:2.5 instead of 1:4 above for blood samples. All processed samples (120 µL) were then transferred to LC-MS vials equipped with 250 µl glass inserts and run in the LC-MS. No QC samples prepared in this case study. Three replicates were prepared for each of the serial dilutions. This study has been approved by Research Ethics Office of McGill University as described above.

4.4.4 LC-MS/MS analysis

Metabolic profiling at the MS1 level was performed on an UHPLC system (Thermo Scientific™ UltiMate™ 3000 System). A hydrophobic column (Hypersil GOLD™ aQ C18 Polar Endcapped HPLC Column, 100mm × 2.1mm, 1.9µm) and a hydrophilic (Accucore™ 150 Amide HILIC HPLC Column, 100mm × 2.1mm, 2.5µm) column were used for reverse phase (C18 column) and hydrophilic interaction liquid chromatography (HILIC column) separation, respectively. The chromatogram system was coupled to a Thermo Scientific Q-Exactive Orbitrap mass spectrometer.

The chromatographic conditions for the C18 and HILIC columns were optimized as follows. For both columns, the flow rate was fixed as 0.4 mL/min. For C18 columns, the composition of the mobile phases A and B were 0.1% FA in water and 0.1% FA in ACN, respectively. For HILIC chromatography, the composition of the mobile phases A and B were 50% ACN in water with 5 mmol/L NH_4AC and 95% ACN in Water with 5mmol/L NH_4AC, respectively. The gradient procedures and other instrumental parameters are provided in Table S4.9.

The Q-Exactive Orbitrap MS was configured as follows. For the C18 column, an electrospray ion (ESI) source with a spray voltage of 4 keV in positive mode and 3.5 keV in negative mode were used, and for HILIC a voltage of 4 keV in positive mode and 3.8 keV in negative mode were used. Additional MS parameters were set for the C18 and HILIC columns, which are summarized in Table S4.10. Both positive (ESI^+, pos) and negative (ESI^-, neg) ion modes were adopted for ion acquisition.

LC-MS/MS was performed immediately after the LC-MS experiment with the corresponding mode (C18-ESI^+, C18-ESI^-, HILIC-ESI^+ or HILIC-ESI^-). The chromatographic conditions were the same as the ones detailed in the LC-MS section, while the mass spectrometer was specifically configured for untargeted DDA, DIA and iterative targeted DDA, respectively. DIA was

performed with sequential windowed acquisition of all theoretical fragment ion mass spectra (SWATH) strategy. All parameters for the MS/MS acquisitions are summarized in Table S4.10.

Untargeted DDA was performed to detect the top ten ions with highest intensity of each full MS scan. Immediately after the acquisition of MS1 and untargeted DDA, the SWATH was performed. Each cycle of SWATH consisted of a full MS scan and 10 MS/MS windows with different window size. The *m/z* size of MS/MS window was determined based on the general distribution of metabolic features at the MS1 level. The adjacent windows were sequentially used for DIA MS/MS detection. The windows were overlapped with their neighbors at 1.0 *m/z*. The design of SWATH windows is summarized in Table S4.11. Approximately 0.9s elapsed in total for each SWATH cycle.

The sample type-specific metabolic features were extracted based on the results of MS1 and used as the inclusion list. Iterative targeted DDA was executed with the updated inclusion list (optimized using HERMES (78)) to exhaust the targets as much as possible48. In detail, the sample-specific ions were input as the acquisition targets for DDA. Once the 1st round DDA was finished, the detected ions were excluded from the target inclusion list. Then, the 2nd round of targeted DDA was performed with the updated inclusion list until the targets or samples were exhausted.

4.4.5 MS/MS spectra reference library curation

A total of 9 public MS/MS database were collected and curated. The schema of reference library in MetaboAnalystR is displayed in Figure 4.21. HMDB database (196) were downloaded directly HMDB websites (https://hmdb.ca/downloads) as xml file. The database was parsed with XML package into R and curated into the SQLite format. Four tables (HMDB_experimental_NegDB,

HMDB_experimental_PosDB, HMDB_predicted_NegDB, HMDB_predicted_NegDB) were generated from HMDB database. MoNA series and LipidBlast database were downloaded from MassBank of North Americ (https://mona.fiehnlab.ucdavis.edu/downloads) as msp format. Nine tables (MoNA_PosDB, MoNA_NegDB, ReSpect_PosDB, ReSpect_NegDB, VaniyaNP_PosDB, Vaniya_NegDB, BMDMS_PosDB, LipidBlast_PosDB and LipidBlast_NegDB) were generated. MassBank database was downloaded from MassBank websites (https://github.com/MassBank/MassBank-data/releases/latest) as msp files. Four tables were generated (RIKEN_PosDB, RIKEN_NegDB, MassBank_PosDB, MassBank_NegDB). GNPS database was downloaded from GNPS website (https://gnps-external.ucsd.edu/gnpslibrary) as msp format. Two tables (GNPS_PosDB, GNPS_NegDB) were generated. MINEs database was downloaded from MINEs website (https://minedatabase.mcs.anl.gov/#/download) as msp format. Two tables were curated (MINEs_PosDB, MINEs_NegDB) from it. The these downloaded msp files were curated with tailored in-house R scripts into SQLite format. All spectra data tables are collected and formatted based on the same schema into a database, which is named as "Complete library". Starting from this "Complete library", we curated another 3 specific MS/MS reference libraries (Pathway library, Biology library and Lipid library) to improve the accuracy and avoid false positives.

Pathway library was mainly curated according to the KEGG pathway information. KEGG pathways of 120 species (which are common model species or pathogenic microorganisms) were downloaded with KEGGREST (213). All compounds from the metabolic pathways of all species were extracted as a compound list. A total of 3,456 compounds were included (Figure S4.22). All MS/MS records in the "Complete library" matching to these compounds were extracted as a "Pathway library".

Similarly, Biology library was curated based on the compound information from KEGG (92) compound database and HMDB (196). In details, all compounds and glycans from KEGG are summarized as compound list 1. All compounds in HMDB labelled as “Serum”, “Urine”, “Sweat”, “Saliva”, “Feces” and “Cerebrospinal Fluid” were summarized as compound list 2. Compound list 1 and 2 were merged as a target list. All MS/MS records in the “Complete library” matching to these compounds were extracted as a “Biology library”.

Lipid library was curated based on the compound information from LIPIDMAPS (183), LipidBank (94), LipidBlast (91) and compounds in HMDB database (which were classified (214) as “Lipids and lipid-like compounds”). All lipids from these databases were summarized as a lipid list. All MS/MS records in the “Complete library” matching to these compounds were extracted as a “Lipid library”. Besides, all lipids in this library are classified into super classes, main classes and sub-classes based on RefMet (182).

Exposomics library was curated from KEGG Drug database (215), Microbial Metabolites Database (MiMeDB) (216), Toxin-Toxin-Target Database (T3DB) (217), FooDB (www.foodb.ca), Phenol-Explorer (218), Exposome-Explorer (219), and NORMAN Suspect List Exchange database (220). All compounds from these databases were extracted as exposomics compound list. All MS/MS records in the “Complete library” matching to these compounds were separately extracted as a “Exposomics library”.

In addition, four neutral loss spectra databases were pre-calculated, corresponding to the five options above. The curation of these neutral loss databases was based on the algorithm implemented by METLIN neutral loss database (195). In brief, neutral loss spectra was calculated by deducting the *m/z* from precursor ion as the neutral loss ion. The intensity values were directly mirrored. Schema of all MS/MS reference libraries is shown in Figure S4.23.

4.4.6 DDA data deconvolution algorithm

The first step of DDA spectral data deconvolution is to assign all MS/MS spectra into all individual target features based on the information of precursors. If users provide a targeted feature list, MetaboAnalystR can automatically perform MS/MS data processing. Otherwise, MetaboAnalystR uses the complete MS features detected for MS/MS data processing. By default, MS1 features list generated by MetaboAnalystR includes minimum and maximum values for *m/z* and RT. If *m/z* and RT are not provided as in ranges, the *m/z* and RT ranges are going to be calculated automatically based on tolerance values defined by users (ppm for *m/z*, and rt_tol for RT). If there are multiple MS/MS spectra assigned to an individual target MS feature, the spectra would be merged in a weighted manner developed from MZmine (38). The median *m/z* (mz_med) and median RT (rt_med) is extracted or calculated based on the MS feature's information. Then the nearest MS1 is extracted for following analysis. If there are multiple different centroids within the (mz_med centered) isolation window, and any of their intensity values are over the acquisition threshold (user defined), the spectrum is considered as "Chimeric", otherwise the spectrum is categorized as "Clean". All Chimeric spectrum are organized to be deconvolved in the next steps. The centroid ion corresponding to the MS feature is considered as "main ion", others are considered as "contamination ions".

The purpose of deconvolution is to remove the fragments produced from "contamination ions" in the chimeric spectrum ("Spectrum 0") and generate a clean deconvoluted spectrum for "main ion". Technically, the "main ion" is the ion of target feature, while the "contamination ion" may come from multiple sources, such as isobarics, orphan isotopologues (190), and other known or unknown ions with their *m/z*s falling into the isolation windows, etc. At the second step, MetaboAnalystR goes through all contamination ions and determine if they are orphan isotopologues. If any of them

is identified as orphan isotopologues based on MS1 scan, the spectrum of this orphan isotopologue is predicted with the method from DecoID (190). We name the predicted spectrum of the orphan isotopologue as "Candidate Spectrum I". Then, if there is any ion has been detected and identified as a clean spectrum in one or two nearest MS/MS scans inside the data itself, the clean spectrum is also extracted for deconvolution, named as "Candidate Spectrum II". Next, MetaboAnalystR extracts potential spectra from MS/MS library as the reference spectrum for the ions, which is neither orphan isotopologues nor the ones with clean spectrum included by the spectra data itself. All spectra from reference are extracted, and the one showing highest similarity to the original chimeric spectrum ("Spectrum 0") is retained as "Candidate Spectrum III". The spectral similarity is evaluated with dot-product (190) or spectral entropy similarity (194) methods based on user's preference. If all ions in this isolation window have been assigned with a reference spectrum, the deconvolution can be executed directly.

But in many cases, some ions can be identified as neither orphan isotopologues nor the one with a clean spectrum contained in the data itself nor the one with a reference spectrum from library. These ions are named as "Unknown ions" here. MetaboAnalystR predicts the spectrum of these "Unknown ions" based on a hypothesis that the ions with abiotic/bio-transformation relationships share highly similar MS/MS spectra pattern (83, 84). In this case, the most accurate formula is firstly predicted for the "Unknown ion". Then an abiotic/bio-transformation network is constructed around the "Unknown ion" (Figure 4.1c) based on the rules from NetID (84). Different from the network in NetID, this prediction network model is not propagatable to avoid potential redundancy. Once the network is constructed successfully, all neighbors of the "Unknown ion" are searched against the library to get their spectra data. All fragments of the spectra data are predicted as the most accurate formula. If the chemical elements' composition of the formula is against the formula

of the “Unknown ion”, this fragment is considered as an unreasonable fragment, and then removed from the spectrum. Once this cleaning step is completed, the similarity of all spectra to the original chimeric spectrum (“Spectrum 0”) are evaluated. The one with the highest similarity score is returned as the predicted spectrum for the “Unknown ion”. It is named as “Candidate Spectrum IV”.

Next, Candidate Spectra I-IV are returned as the components to deconvolve the original chimeric spectrum (“Spectrum 0”). Given that the Candidate Spectrum IV are neither from a real data nor reference library, a penalty (0-10; 0, no penalty for perfect match; 10, 10 times penalty for negative match) is given based on the similarity to the “Spectrum 0”. Deconvolution on the “Spectrum 0” is performed with a penalized elastic-net regression model (221, 222). The purpose of this deconvolution model is to minimize the residue. The deconvolution method is shown as the formula below:

$$Residue = \min\left(\sum_{i=0}^{n}(\mathrm{y} - \mathrm{x}_i\beta)^2 + \lambda P_\alpha(\beta)\right)$$

where $P_\alpha(\beta) = \frac{1}{2}(1-\alpha)||\beta||_2^2 + \alpha||\beta||_1$ is the elastic net penalty(222). y is the response vector (Spectrum 0), x is the candidate components (Candidate Spectra I-IV). In this model, α and λ are two critical parameters. If α=1, the model is a LASSO regression model, similar to DecoID2. Instead of using an arbitrary value for α and λ from DecoID, MetaboAnalystR permutates a matrix of α and λ combination. In detail, 11 α values (starting from 0, and end with 1, step by 0.1) and 10 λ values, estimated by the correlations between response vector and component vectors (222). Therefore, 110 α and λ combinations are prepared to automatically optimize the elastic-net model. Then, deconvolution based on the penalized elastic model is executed and results in 110 solutions for “Spectrum 0”. All residues of 110 solutions are iterated and the one with minimal residue is

returned (Solution 0). Different from DecoID, MetaboAnalystR optimizes α and λ for every individual peak, instead of imply a hard value for all peaks.

Finally, all β values for contamination ions (β_{contms}) in Solution 0 is used to remove fragments in "Spectrum 0". The remaining fragments are normalized and exported as "deconvoluted" spectrum for "main ion". If there is no fragment left after the cleaning or the β value for "main ion" is 0, the deconvolution failed. The original "Spectrum 0" will be retained and exported directly for MS/MS reference library searching. The deconvolution of DDA data can be achieved with the function, *PerformDDADeconvolution*.

4.4.7 SWATH-DIA data deconvolution algorithm

SWATH-DIA data deconvolution algorithm follows the steps described by DecoMetDIA (76). In brief, for a specific MS feature, all extracted ion chromatograms (EICs) of MS/MS peaks from the corresponding SWATH window are detected and clustered based on peak similarity and RT information. One model peak is selected from each cluster, and all model peaks are organized to decompose all EICs. Each decomposed component from different EICs is used to reconstruct the composition of the MS/MS cluster. The cluster containing the original MS feature is exported as a *pseudo*-MS/MS spectrum. Unlike DecoMetDIA, the entire data deconvolution workflow is implemented in Rcpp/C++. The deconvolution of SWATH-DIA data can be achieved with the function, *PerformDIADeconvolution*.

4.4.8 Spectra consensus of replicates algorithm

MS/MS data acquisition with multiple replicates is common. In such cases, all deconvolved M/MS spectra corresponding to the same MS1 peak must be processed to generate a single consensus spectrum. If there are no replicates, this step is skipped. All MS/MS fragments across different

replicates are initially summarized by count. If the frequency of an individual fragment is above a user-defined threshold (e.g., 50%), it is kept; otherwise, the fragment is removed. Optionally, a database-assisted spectrum consensus in MetaboAnalystR can be used to avoid potential over-deletion (see Figure S4.1). If database-assisted option is enabled by users, all spectra of the precursor are extracted as referent list (L) from the reference library. All fragments not meeting the (user-defined) frequency threshold are then searched against L. If this fragment can be found from L and the frequency of the fragment across the replicates is over 2, the fragment is kept; otherwise, it is discarded. All kept fragments are normalized and merged to generate a consensus spectrum for database searching in the next step. The spectrum consensus can be achieved with the function, *PerformSpectrumConsenus*.

4.4.9 Reference library searching and scoring algorithms

Reference library searching is based on the *m/z* (and optionally, the information of RT) of precursors. All matches are extracted from database for formula prediction. MetaboAnalystR uses the same scoring rule as MS-DIAL14. The matching score is calculated using the following formula:

$$Matching\ Score = \frac{MSMS\ Similarity + MS1\ Similarity + RT\ Similarity + 0.5 \times Isotope\ similarity}{3.5} \times 100$$

where the MS/MS similarity (ranging from 0 to 1) is calculated using the popular dot product similarity (40, 190) or spectral entropy similarity (194) algorithms. In this study, dot-product similarity was used to evaluate the MS/MS matching results. MS1 similarity and RT similarity (both ranging from 0 to 1) are calculated with exponential distribution method based on deviation. Isotope similarity is also calculated using a similar method as implemented in MS-DIAL. However, the calculation of isotope distribution similarity only considers [M+n] (n<3), as the intensity of

isotopes [M+n] (n≥3) is very low and highly variant to be considered. Briefly, the isotope distribution similarity evaluation is performed based on the experimental isotope distribution and the theoretical distribution of formulas extracted from the reference library. Isotope elements considered here include carbon (C13), hydrogen (H2), nitrogen (N15), oxygen (O17, O18) and sulfur (S33, S34). Other elements are not considered due to their extremely low abundance in nature. RT is optionally used based on users' request. If RT is disabled (by default), the RT similarity is not calculated, and denominator is modified to 2.5. The database searching is performed based on SQLite query, and can be achieved with the function, *PerformDBSearchingBatch*.

4.4.10 Neutral loss searching

If the matching score is below 10 (out of a maximum of 100), the option neutral loss searching can be performed to find the potential chemical identification. In such cases, the target precursor (P0) is extracted and incorporated into the abiotic/bio-transformation network model (84) (described in "DDA data deconvolution algorithm") to find all potential neighbors. These neighbors are used as targets for extracting spectra from neutral loss reference library. Neutral loss of P0 is calculated directly (195) and is matched against the neutral loss spectra extracted from reference library. All potential matches are scored and exported as reference of chemical identification. The results are labelled as "Neutral loss matching". The database searching with neutral loss can be achieved by enabled the "enableNL" parameter in the *PerformDBSearchingBatch* function.

4.4.11 Result export

All compound identification results can be exported as a data frame and saved as a .csv or .txt file. The exported information includes compound names, chemical formula, InChIKeys, and matching

scores. If the reference library is lipid library, the exported information also includes lipid classifications (super class, main class and sub-class). The database searching results can be exported using *PerformResultsExport* function and formatted as a data frame table using *FormatMSnAnnotation* function.

4.4.12 Decoy spectra generation and null evaluation

To generate decoy spectra, a standard mixture of 91 compounds (199) was used as the raw spectra data in the mzML format, which was imported using the mzR package (144). Spectral scans were split into MS data and MS/MS data based on MS level (as shown in Figure 4.3c). MS data was processed similarly for both DDA and SWATH-DIA methods. Specifically, the *m/z* values of mass centroids of MS data were randomly adjusted by adding mass errors ranging from 10 to 30 ppm, while the intensity values were randomly distorted by multiplying with a coefficient (ranging from 0.01 to 50.0). The RT dimension was kept unchanged. For DDA spectra, the MS/MS spectrum pattern was replaced with a synthetic MS/MS spectrum randomly simulated from isobaric compounds. In contrast, for SWATH-DIA spectra, the MS/MS spectrum pattern was processed in the same way as MS spectra, while the SWATH window and RT were kept unchanged. A total of 18 spectra decoy spectra data were generated for each replicate in the both DDA and SWATH-DIA datasets. These decoy spectra were processed in the same way as the original real dataset using different tools, to perform null evaluations.

4.4.13 MetaboAnalystR usage

The MS spectral data used in this study were converted into mzML centroid mode using Proteowizard (144) for both MS and MS/MS levels. The auto-optimized workflow was first applied to process the MS spectra, including peak picking, peak alignment, gap filling, and peak

annotation, to generate complete MS feature tables. These tables were used as the target list for MS/MS spectra processing in both DDA and SWATH-DIA. The chemical classification analysis of blood samples was performed using ClassyFire database (214). Pearson correlation analysis from the R stats package was employed to perform the correlation analysis of serial dilutions. The heatmap analysis was directly performed using MetaboAnalystR. Identification and matching of compounds to the standards list were performed based on InChIKeys.

4.4.14 MS-DIAL/MS-FINDER usage

MS-DIAL (v4.9.22, Linux version) and MS-FINDER (v3.52, Linux version) were used. A mass accuracy parameter of 0.005 Da was set for "MS1 tolerance" and 0.01 Da for "MS2 tolerance". The minimum percentage of peaks within one group was set to 50%, while other parameters were left as default. Following DDA or SWATH-DIA data processing, peak area alignment results were exported as the results of MS level. All features with MS/MS information were exported for MS-FINDER analysis performed in batch mode. PubChem access was only allowed when there was no candidate from other databases. All MS/MS spectra libraries were selected for searching, while other parameters were left as default values. The identified formula and structures were exported automatically by MS-FINDER. Compound identification and matching to the standards list were based on InChIKeys.

4.4.15 MZmine usage

We used MZmine (v3.2.8, Linux version) to process the raw spectral data. Firstly, we imported the raw data and performed mass detection at MS level 1. Next, we executed the ADAP chromatogram builder (161) with a parameter scan-to-scan accuracy set at 0.005 Da or 10 ppm, while keeping other parameters as default values. We then performed smoothing and joint

alignment, which resulted in an aligned feature list that we exported as the MS feature table. Subsequently, we executed the MSn feature list builder, followed by mass detection at MS level 2. Finally, we exported the feature list in the SIRIUS/CSI-FingerID format with all feature lists with MS2 features selected. We enabled the merge MS/MS option, leaving the other parameters as default values in this step. To evaluate computational performance, we processed the COVID-19 dataset in batch mode from the command line. For other datasets, we used the MZmine UI for analysis.

4.4.16 XCMS usage

XCMS (v3.20.0, an R package) was used to process SWATH-DIA data. All parameters from the XCMS online platform (https://xcmsonline.scripps.edu/) were extracted to process MS data. For SWATH-DIA data processing, we utilized function, reconstructChromPeakSpectra, to deconvolve SWATH-DIA data. We exported all deconvoluted spectra into a msp file using an in-house R script.

4.4.17 SIRIUS usage

We used SIRIUS (v5.6.3, Linux headless version) to search MS/MS spectra. For formula prediction, we set the program to use the entire database, while enabling ZODIAC, CSI:FingerID, and CANOPUS. For CSI:FingerID, we selected all available databases. Other parameters were left at their default values. We exported the MS/MS searching results with the "write summarize" option enabled. To identify chemicals and match them to the standards list, we used the InChIKeys generated by InChIs from SIRIUS.

4.4.18 Computational performance assessment

To evaluate the computational performance of the tools, we used SLURM (v22.05.6) to execute and record the usage of the computational resources for each job. All tools had a command-line interface to be executed. We allocated two CPU cores and all RAM resources for each tool for comparison. We recorded the clock time between the starting and ending of the job, as well as the maximum usage of RAM.

4.4.19 Integration of MS/MS results into mummichog algorithm

MetaboAnalystR can process MS and MS/MS directly and convert the results into formatted lists for pathway enrichment prediction with enhanced *mummichog*. It also accepts MS peaks list/table individually or in combination with MS/MS-based compounds identification. The *mummichog* algorithm was improved by incorporating MS/MS-based chemical identification results. Initially, the algorithm matches all features based on their *m/z* and/or RT to generate empirical compounds (102). One MS feature may be mapped to multiple empirical compounds. In such cases, MS/MS-based chemical identifications are utilized to filter out those empirical compounds that are not feasible based on the MS/MS spectrum. This process results in a shorter but more accurate list of empirical compounds. The permutation test is then performed based on the filtered empirical compound list. The underlying pathway libraries have been updated with additional compound IDs to be more compatible with the results generated from MS/MS identification. In MetaboAnalystR 4.0, all compounds in various pathway databases have been converted into different types of chemical IDs, including InChIKeys, KEGG IDs, HMDB IDs, PubChem SIDs, PubChem CIDs, and SMILES.

4.4.20 Interfacing with other tools

MetaboAnalystR is capable of processing metabolomics data from raw spectra and providing biological insights directly. However, it can also accept results from other raw spectra data processing tools, such as MS-DIAL (in mat format), MZmine (in msp format), and XCMS (as an R object) for MS/MS identification with a comprehensive/specific database. In addition, MetaboAnalystR can automatically convert MS/MS identifications from MS-FINDER (structure result table) and SIRIUS (compound identification table) into a compound list for *mummichog*-based pathway enrichment analysis. All MS/MS reference libraries are curated as open-source SQLite files. Users can easily convert their in-house reference library into SQLite format to incorporate into the workflow of MetaboAnalystR.

4.5 Discussion

MetaboAnalystR is a comprehensive toolkit for LC-MS metabolomics data analysis, including raw spectra processing, statistical analysis, and functional analysis (102, 105, 122). In version 4.0, MetaboAnalystR has been enhanced by introducing a series of functions to perform an end-to-end LC-MS/MS raw data processing and biological interpretation.

By deconvolving DDA spectra with a penalized elastic-net regression model and spectra prediction network, together with comprehensive MS/MS reference library options, MetaboAnalystR enables high throughput MS/MS-based compound identification. By introducing the SWATH-DIA data processing workflow and adapting it into MetaboAnalystR raw data processing workflow, the computing efficiency and compound annotation coverage has been significantly improved. MetaboAnalystR accepts all common open-source raw spectra data formats. MS/MS data acquired by DDA or SWATH-DIA strategies can be processed directly. Finally, MetaboAnalystR enables accurate pathway enrichment analysis directly with the results from raw spectra processing workflow.

Within the realm of metabolomics, two tools have emerged for processing DDA data deconvolution: DecoID (190) and MS2Purifier (77). In our present study, we introduced a network-based spectral prediction model and an auto-optimized elastic net regression to address the limitations of DecoID. MS2Purifier, on the other hand, is grounded in the principle of establishing similarity between the elution profiles of MS and MS/MS features. By integrating a machine learning model, contaminations are discriminated and eliminated. Given the algorithmic strategies at play, MS2Purifier potentially complements MetaboAnalystR. This prompts the need for a comprehensive comparison and benchmark study to further enhance MetaboAnalystR's performance in the future.

To be more compatible and interoperable with other popular tools, MetaboAnalystR supports the MS/MS database searching workflow with MS-DIAL and MZmine software. The pre-processed results from other tools could also been easily formatted into msp format for searching by MetaboAnalystR. However, compared to MS-DIAL and MZmine, the current version does not support ion mobility spectrometry MS data. The functionalities to process ion mobility spectral data, direct injection and flow-injection spectral data will be achieved in next version.

4.6 Conclusion

MetaboAnalystR is a comprehensive workflow for LC-MS metabolomics data analysis, including raw spectra processing, statistical analysis, and functional analysis. The version 4.0 has introduced a series of important functions to enable streamlined LC-MS/MS raw data processing, annotation, and biological interpretation. Through careful design and efficient implementation, it allows unified spectral deconvolution, consensus and annotation for both DDA and DIA data. The results can be directly integrated for more accurate pathway enrichment analysis. MetaboAnalystR

accepts all common open-source raw spectra data formats. MS/MS data acquired by DDA or SWATH-DIA strategies can be processed directly. To be more compatible and interoperable with other popular tools, MetaboAnalystR supports the MS/MS database searching workflow with MS-DIAL and MZmine software. The pre-processed results from other tools could also been easily formatted into MSP format for searching by MetaboAnalystR. Compared to MS-DIAL and MZmine, the current version does not support ion mobility spectrometry MS data. The functionalities to process ion mobility spectral data, direct injection and flow-injection spectral data will be achieved in the future release.

4.7 Supplementary materials

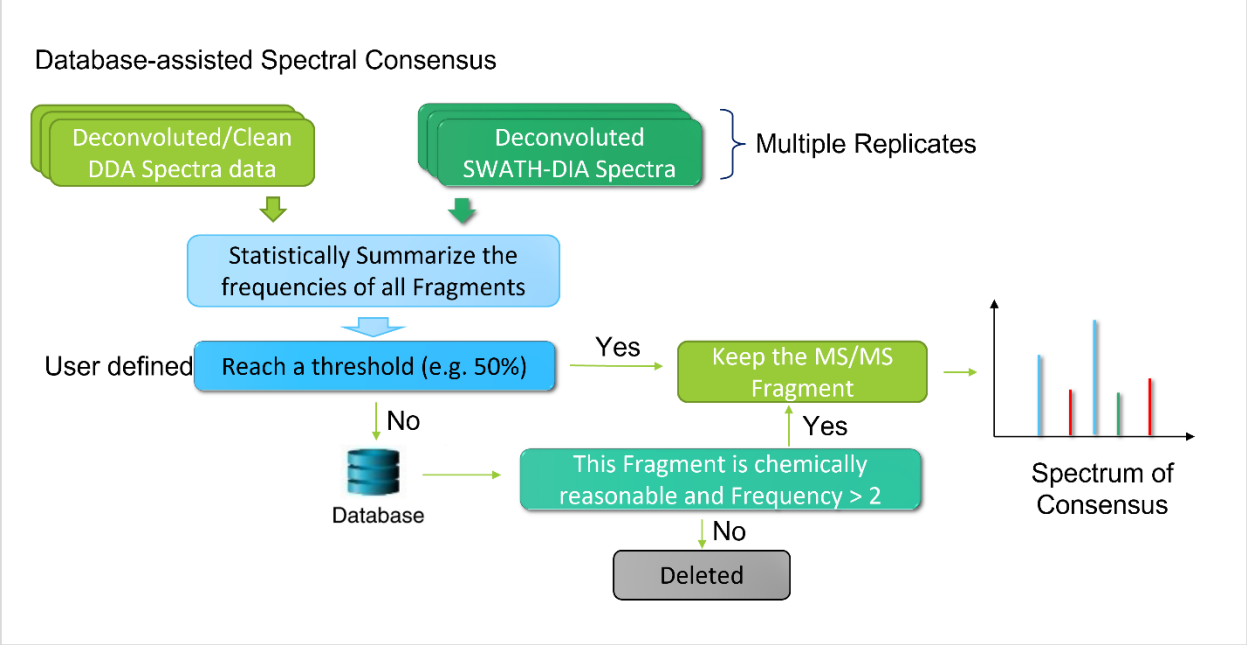


Figure S4.1. Workflow of spectral consensus of replicates. The deconvoluted/clean spectra obtained from DDA or SWATH-DIA are summarized by counting the frequency of each fragment. Fragments that meet a user-defined threshold (e.g., 50%) are retained. If the databases-assisted consensus is enabled, fragments that do not meet the threshold are searched against the reference

library. All spectra are extracted based on the precursors' information, and if the fragment can be found in any of the extracted spectra, it is considered chemically reasonable. Fragments with a frequency of over 2 are retained, while those with lower frequency are considered noise and deleted.

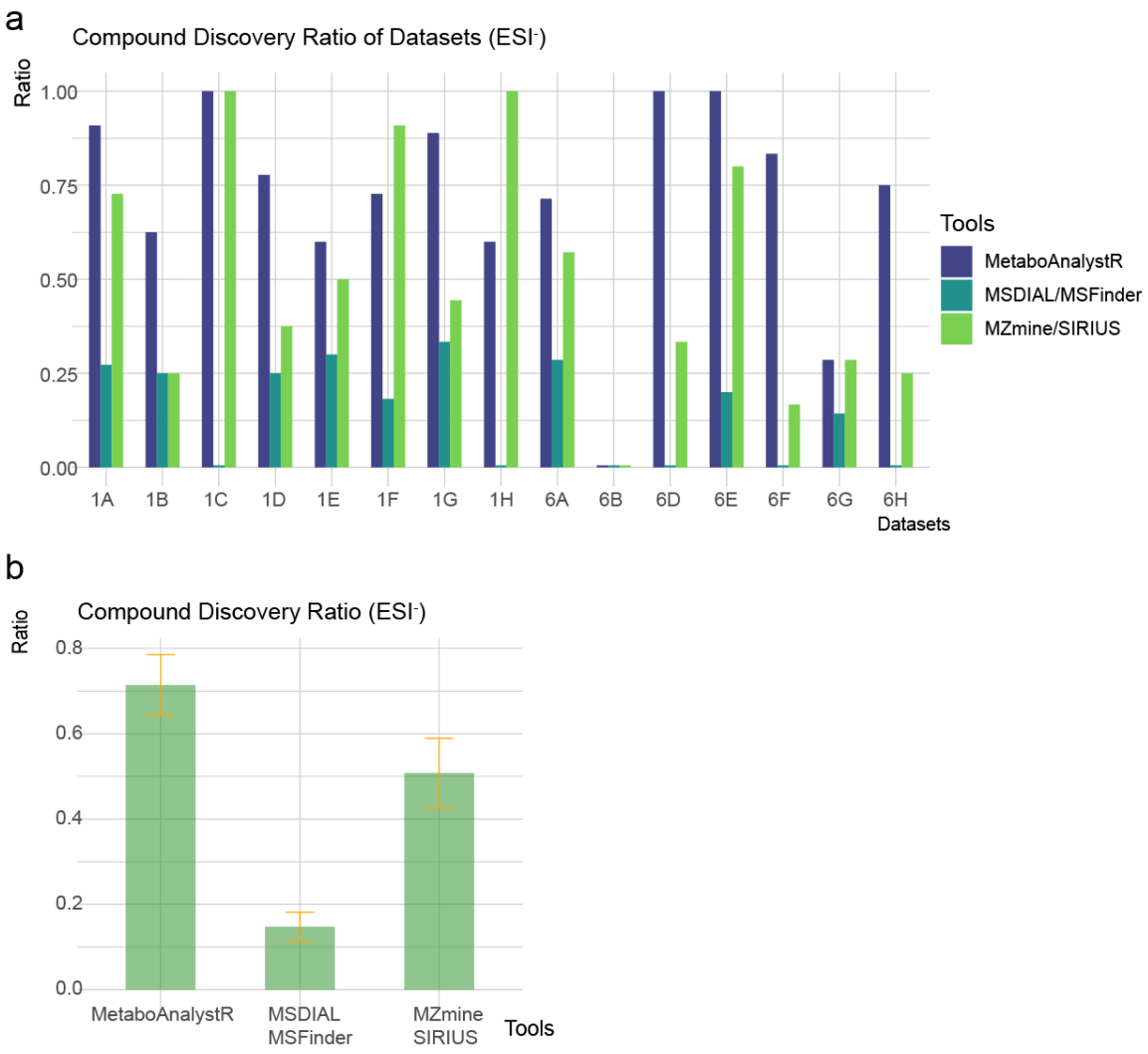


Figure S4.2. Validation results of MetaboAnalystR with simple standard mixtures in ESI- mode. a. Compound discovery ratio of simple standards mixture samples in three workflows. For all samples, MetaboAnalystR detected the highest ratio of compounds as the top first candidate. b. Statistical analysis of the compound discovery results. Compared to other two workflows, MetaboAnalystR reported significantly higher compound discovery ratio ($P < 0.01$ vs. MSDIAL/MSFINDER and vs. MZmine/SIRIUS; $P < 0.01$, MSDIAL/FINDER vs. MZmine/SIRIUS).

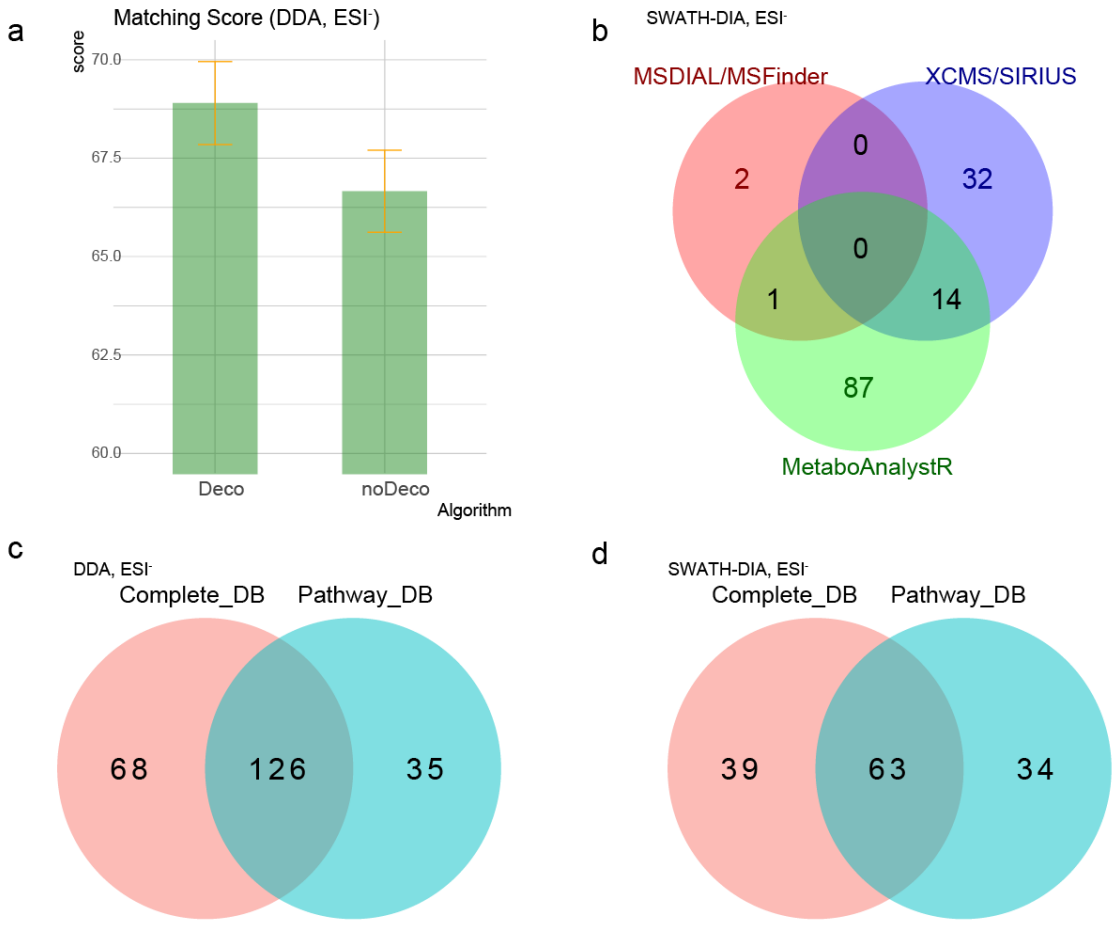


Figure S4.3. Analysis results of complicated standards mixture. a. Comparison of matching scores of DDA (w/o deconvolution, ESI$^-$) in complex standard mixture sample. The deconvolution algorithm in MetaboAnalystR significantly improved the matching score of chemical candidates, as compared to the non-deconvolved spectra (paired t-test, $P < 0.01$). b. Venn Diagram of compounds identified from the complex standard mixture by different tools (SWATH-DIA, ESI$^-$). MetaboAnalystR identified the highest number of correct compounds compared to other tools. MSDIAL/MSFinder and XCMS/SIRIUS workflow could find compounds that were not identified by MetaboAnalystR. e. Performance evaluation of compound discovery with different reference libraries by MetaboAnalystR. The majority of the compounds identified with different databases are shared for both DDA and SWATH-DIA (ESI$^-$). However, compared to the complete reference library, the pathway library could also be used to find some unique compounds that were masked by false positives from the complete reference library.

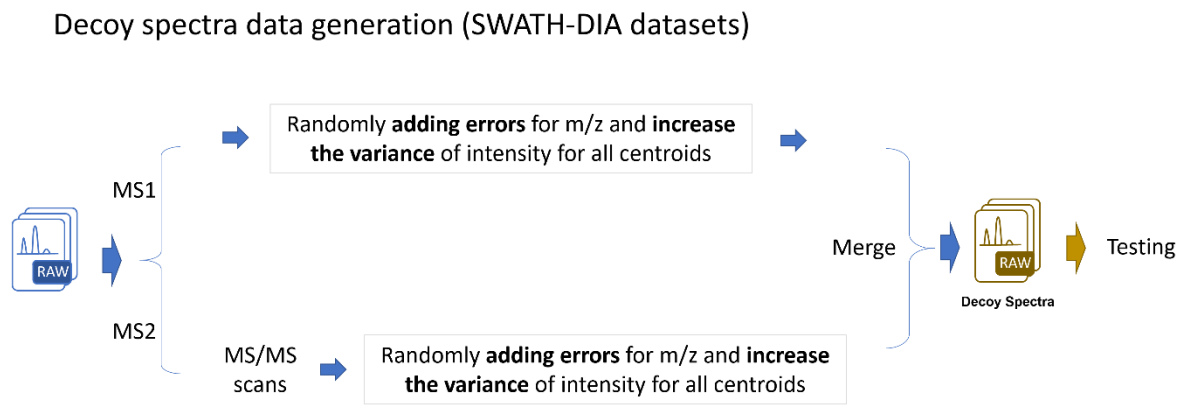


Figure S4.4. Workflow to generate decoy spectra data (SWATH-DIA). Raw spectra data is initially split into MS and MS/MS. For signals from the MS level, mass errors (10~30 ppm) for *m/z* values and variance for intensities were randomly added, while RT information is retained in its original status. For MS/MS data, the original SWATH windows and cycles are retained, while MS/MS spectra were modified by adding mass errors and variance for all MS/MS centroids. Finally, all modified MS and MS/MS scans were merged into a decoy spectra data. A total of 18 decoy spectra datasets were generated for both ESI^+ and ESI^-.

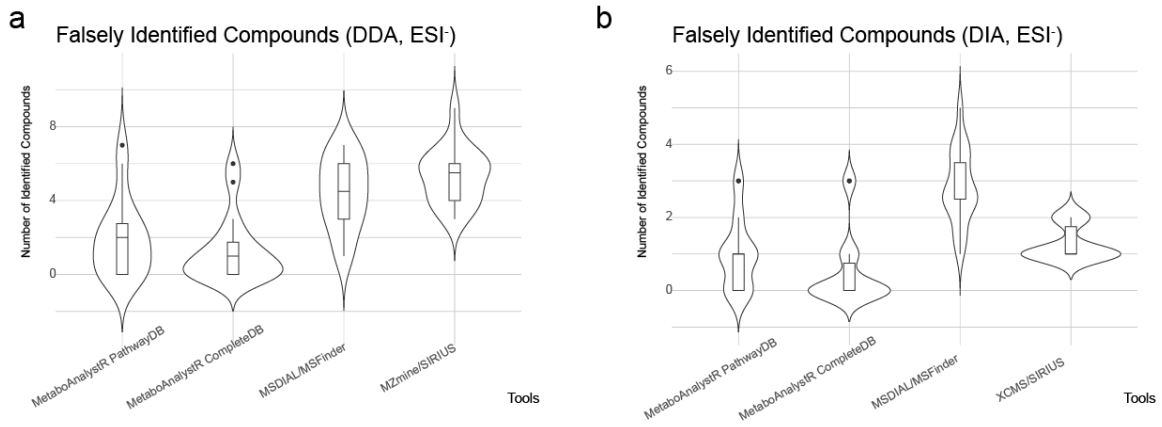


Figure S4.5. Summary of falsely identified compounds from null evaluations in ESI^- mode. a. falsely identified compounds from decoy spectra data (DDA) are compared between different tools. MetaboAnalystR did not significantly increase the false identification (ANOVA, $P > 0.05$) with any reference library, outperforming other tools. b. falsely identified compounds from decoy spectra data (SWATH-DIA) are compared between MetaboAnalystR, XCMS/SIRIUS, and MSDIAL/MSFinder workflows. MetaboAnalystR did not significantly increase false

identification compared to XCMS/SIRIUS ($P > 0.05$). However, MetaboAnalystR had a significantly lower false identification number compared to MSDIAL/MSFinder workflow ($P \approx 0.01$).

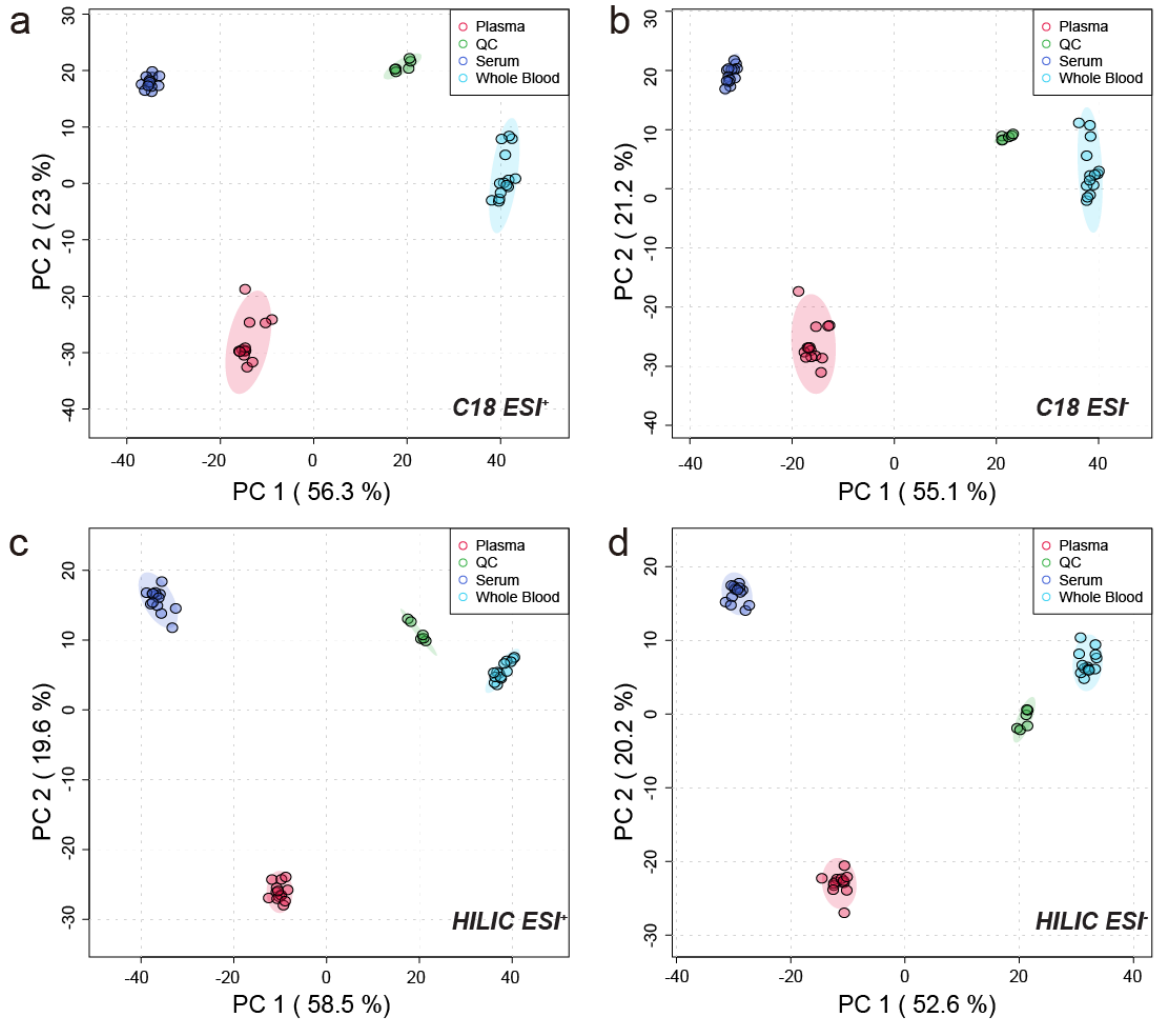


Figure S4.6. PCA results of metabolomic profiles of four different modes analyzed by MetaboAnalystR. a. C18-ESI$^+$; b. C18-ESI$^-$, c. HILIC-ESI$^+$ and d. HILIC-ESI$^-$.

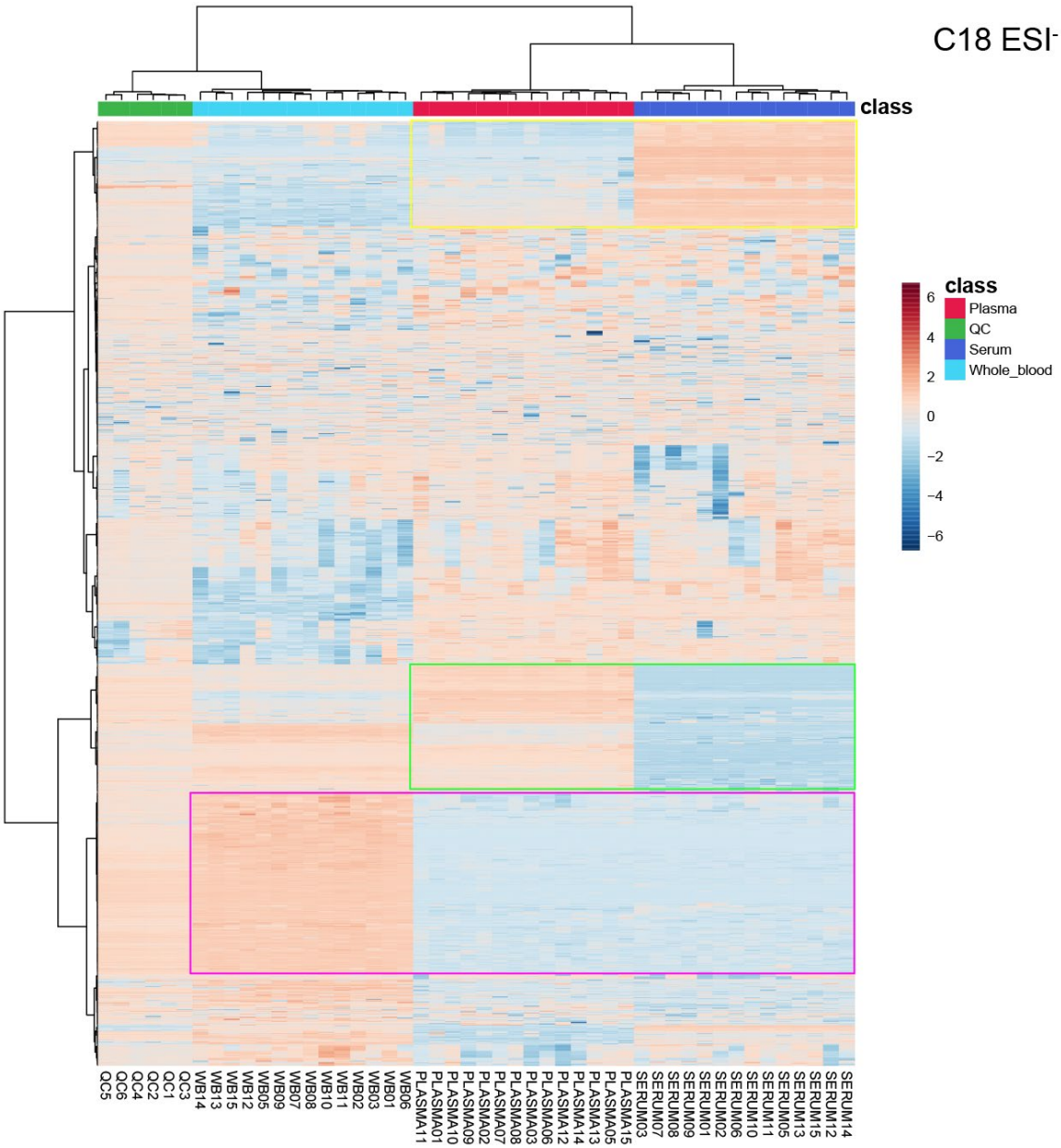


Figure S4.7. Heatmap of blood samples from C18 ESI- mode. MetaboAnalystR was used to detect MS features, and the heatmap was generated based on the detected features. The ruby rectangle highlights unique MS features detected in whole blood compared to serum and/or plasma. The green rectangle indicates unique MS features detected in plasma compared to serum, while the yellow rectangle highlights unique MS features detected in serum compared to plasma.

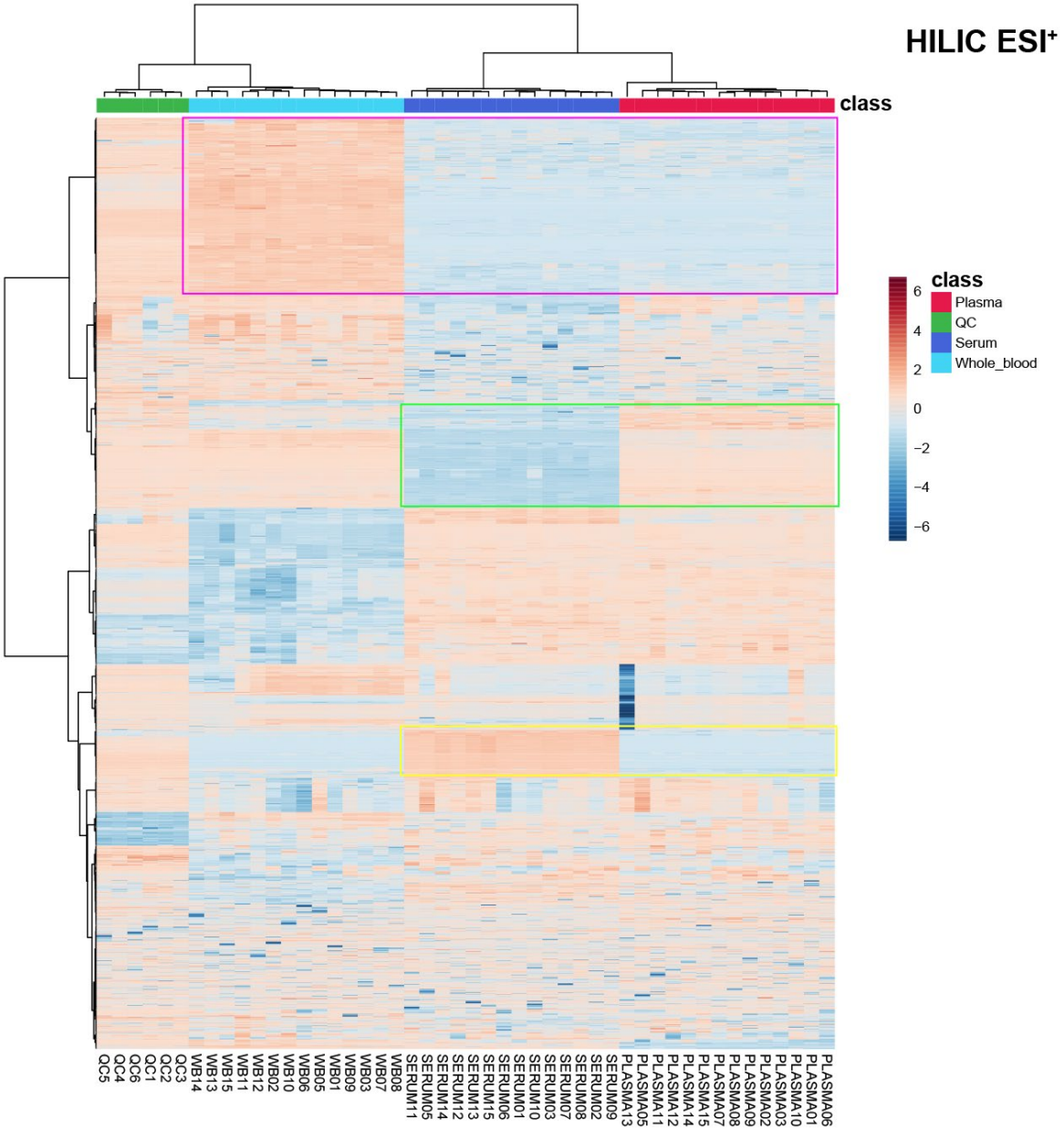


Figure S4.8. Heatmap of blood samples from HILIC ESI$^+$ mode. MetaboAnalystR was used to detect MS features, and the heatmap was generated based on the detected features. The ruby rectangle highlights unique MS features detected in whole blood compared to serum and/or plasma. The green rectangle indicates unique MS features detected in plasma compared to serum, while the yellow rectangle highlights unique MS features detected in serum compared to plasma.

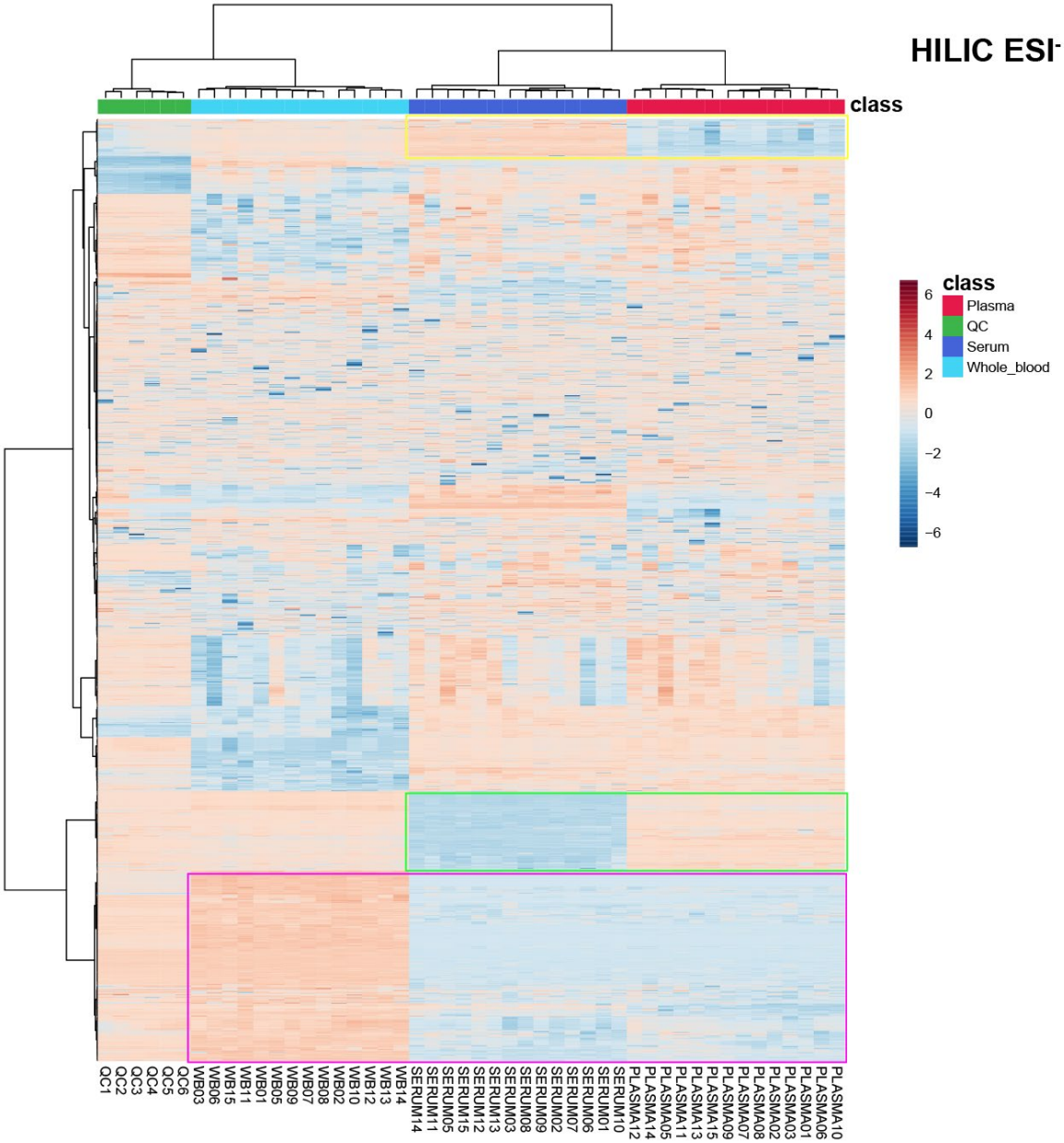


Figure S4.9. Heatmap of blood samples from HILIC ESI$^-$ mode. MetaboAnalystR was used to detect MS features, and the heatmap was generated based on the detected features. The ruby rectangle highlights unique MS features detected in whole blood compared to serum and/or plasma. The green rectangle indicates unique MS features detected in plasma compared to serum, while the yellow rectangle highlights unique MS features detected in serum compared to plasma.

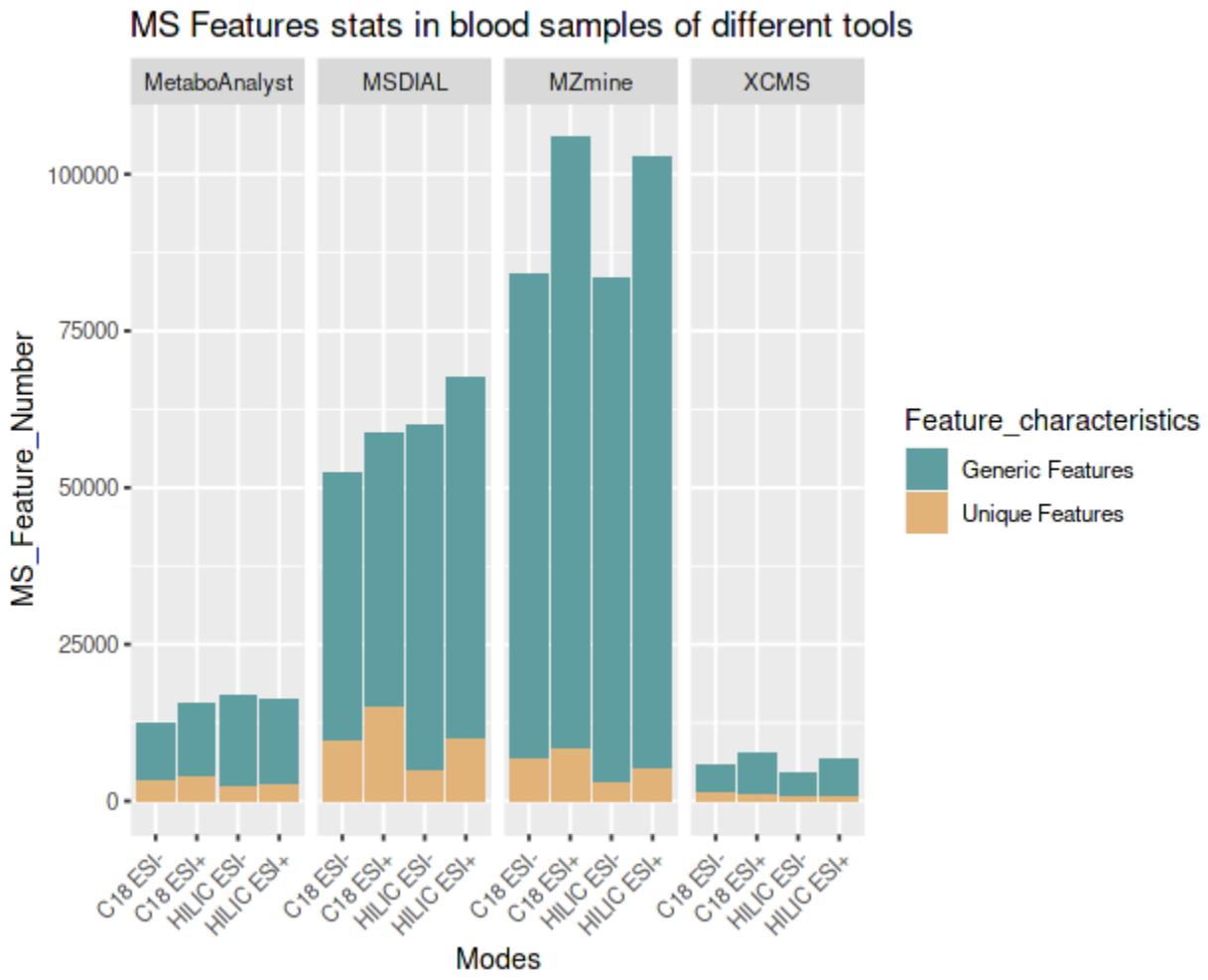


Figure S4.10. Statistics of MS features detected by different tools. Features are classified as "Generic Features" and "Unique Features". Generic features can be detected from all blood samples, while unique features are the features detected from a certain blood sample type specifically. The results show that MZmine and MSDIAL are more sensitive compared to MetaboAnalystR and XCMS, indicating that these tools may be more suitable for detecting features at a trace level.

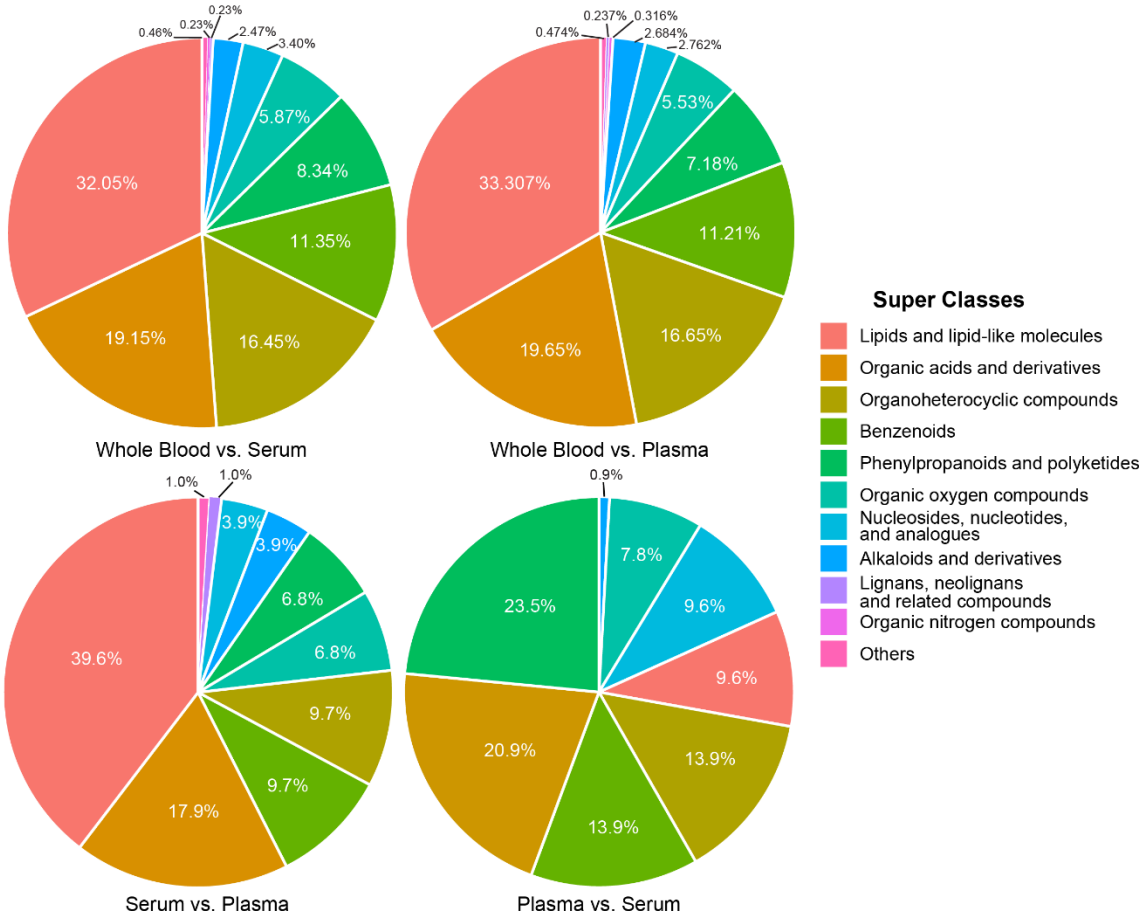


Figure S4.11. Summary of chemical classification of identified compounds from MetaboAnalystR. All unique features from Supplementary Fig. 10 were targeted for compound identification. More lipids, organic acids, and organic heterocyclic compounds were identified in whole blood compared to serum and plasma. More lipids, organic acids, and benzenoids were identified in serum compared to plasma. Conversely, more phenylpropanoids, organic heterocyclic and benzenoids were identified in plasma samples compared to serum.

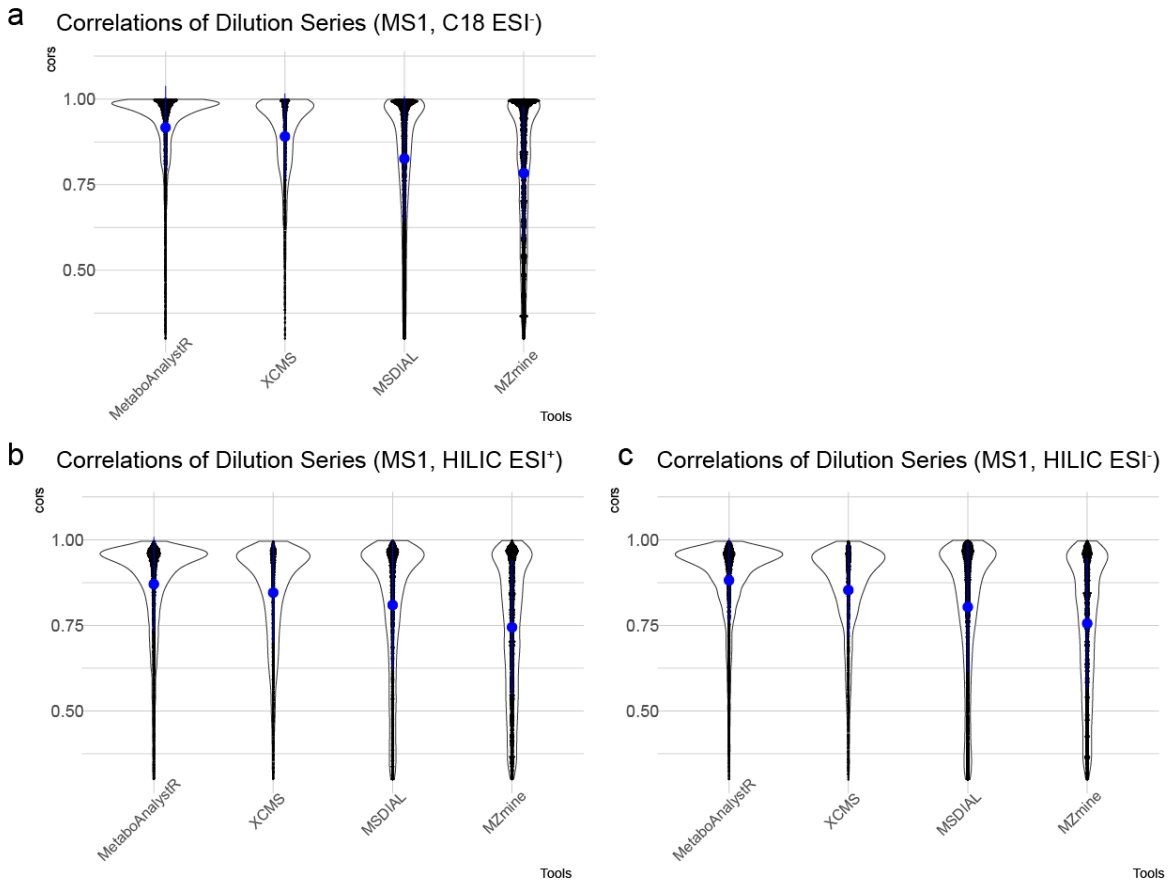


Figure S4.12. Evaluation of quantitative performance based on serial dilutions**.** The correlation analysis of MS features from serial dilutions was detected by different tools under different modes. Under the C18 ESI^- mode in **a**, MetaboAnalystR reported the highest average correlation coefficients compared to other tools. The same trend was observed for HILIC ESI^+ mode in **b** and HILIC ESI^- mode in **c**. This indicates that MetaboAnalystR is better at accurately quantifying the identified compounds in serial dilutions compared to other tools.

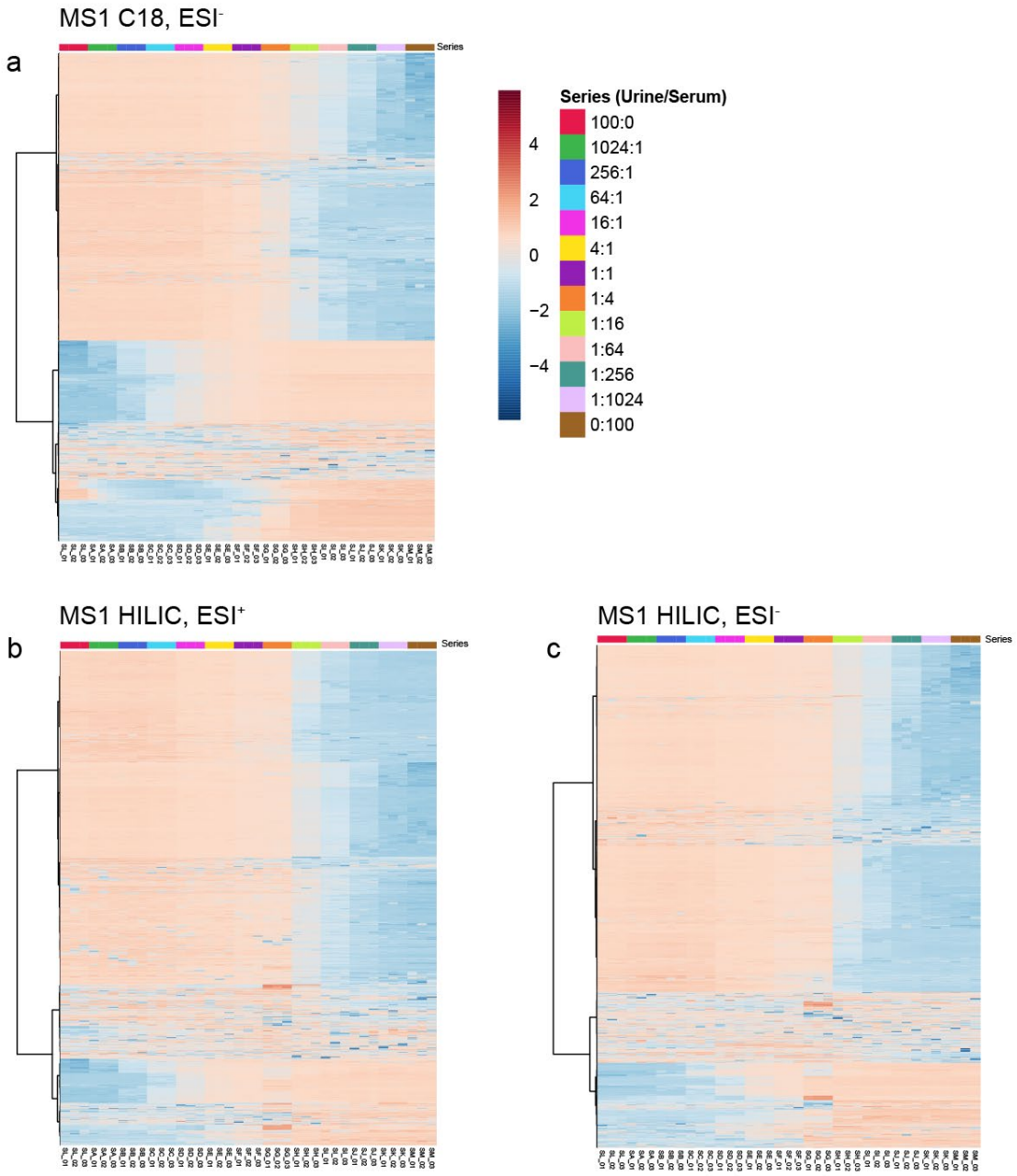


Figure S4.13. Serial dilution heatmaps. Heatmaps of all MS features detected by MetaboAnalystR under different modes (a. C18 ESI^-; b. HILIC ESI^+; c. HILIC ESI^-). All features have been normalized and clustered, and the samples are not sorted based on the dilution series. Pure urine (100:0) and serum (0:100) samples are also included. Clear serial dilution patterns are evident for all modes. Only the features that are not shared by urine and serum are used for compound identification.

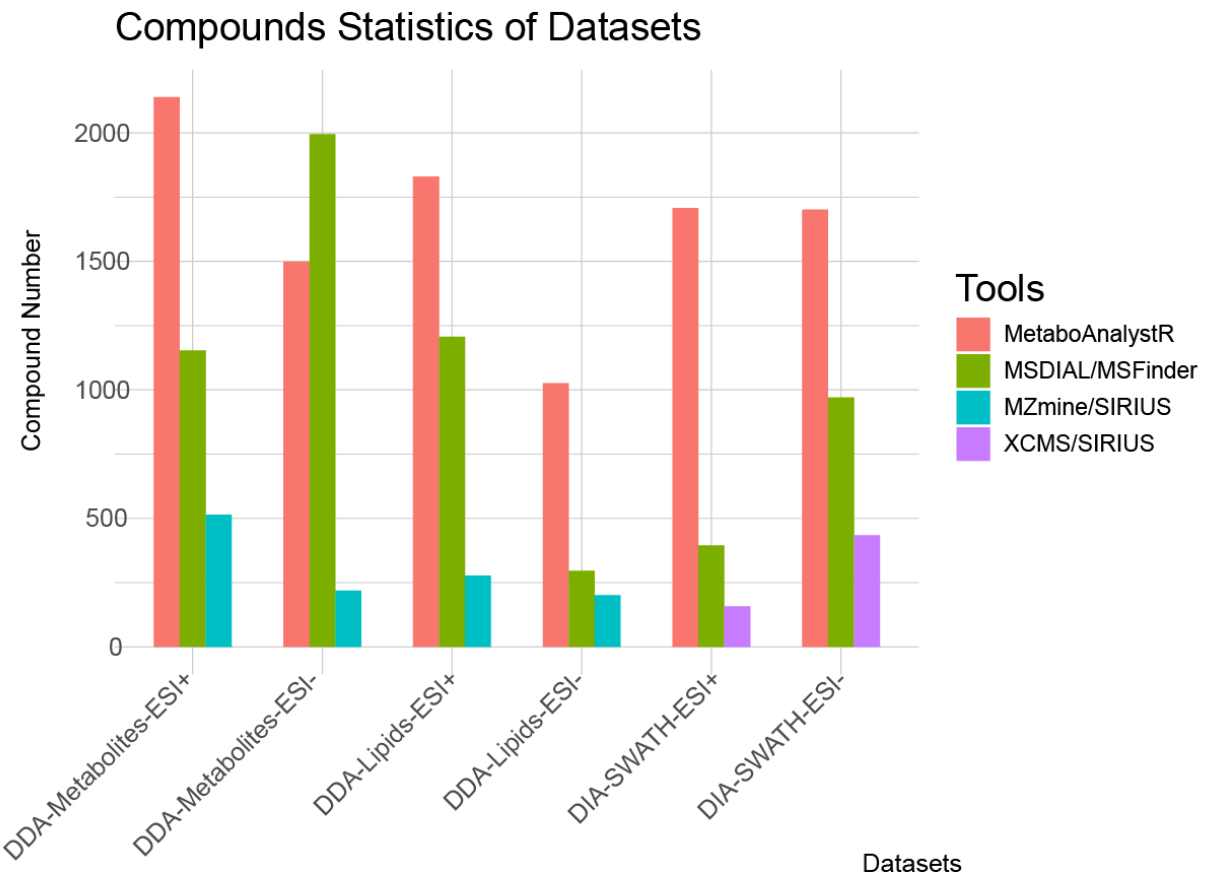


Figure S4.14. Statics of compound identification by different tools. The results show that in most cases, MetaboAnalystR identified the highest number of compounds from the datasets, except for the DDA metabolomics dataset in ESI^- mode. Overall, MetaboAnalystR showed superior performance in compound identification when compared to the other tools.

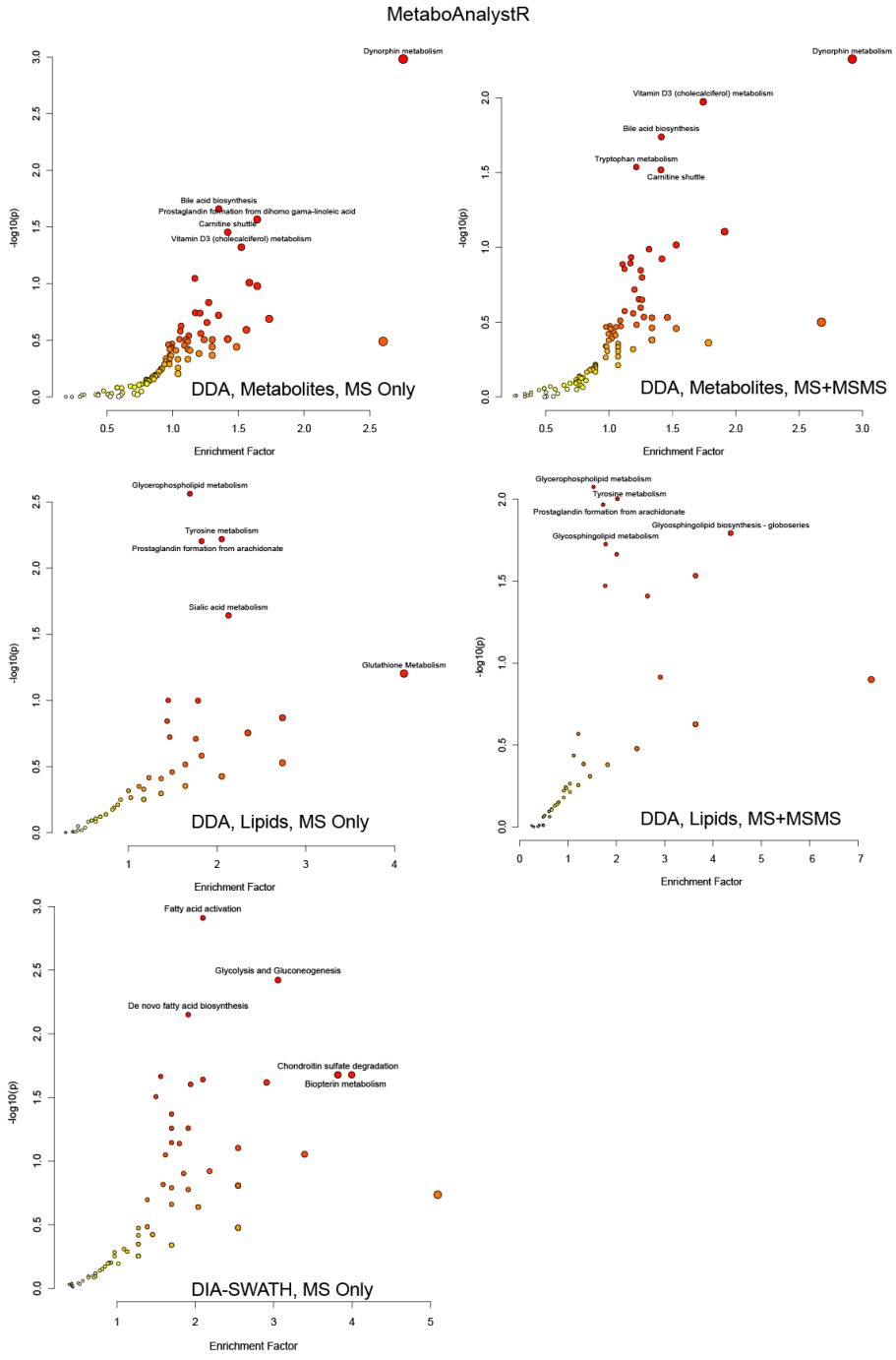


Figure S4.15. Scatter plot of pathway enrichment of datasets analyzed by MetaboAnalystR. The MS features and MS/MS compounds used in this figure were processed by MetaboAnalystR. In most cases, integrating MS/MS results into the pipeline increases the discovery of pathways or improves the statistical significance compared to using "MS only".

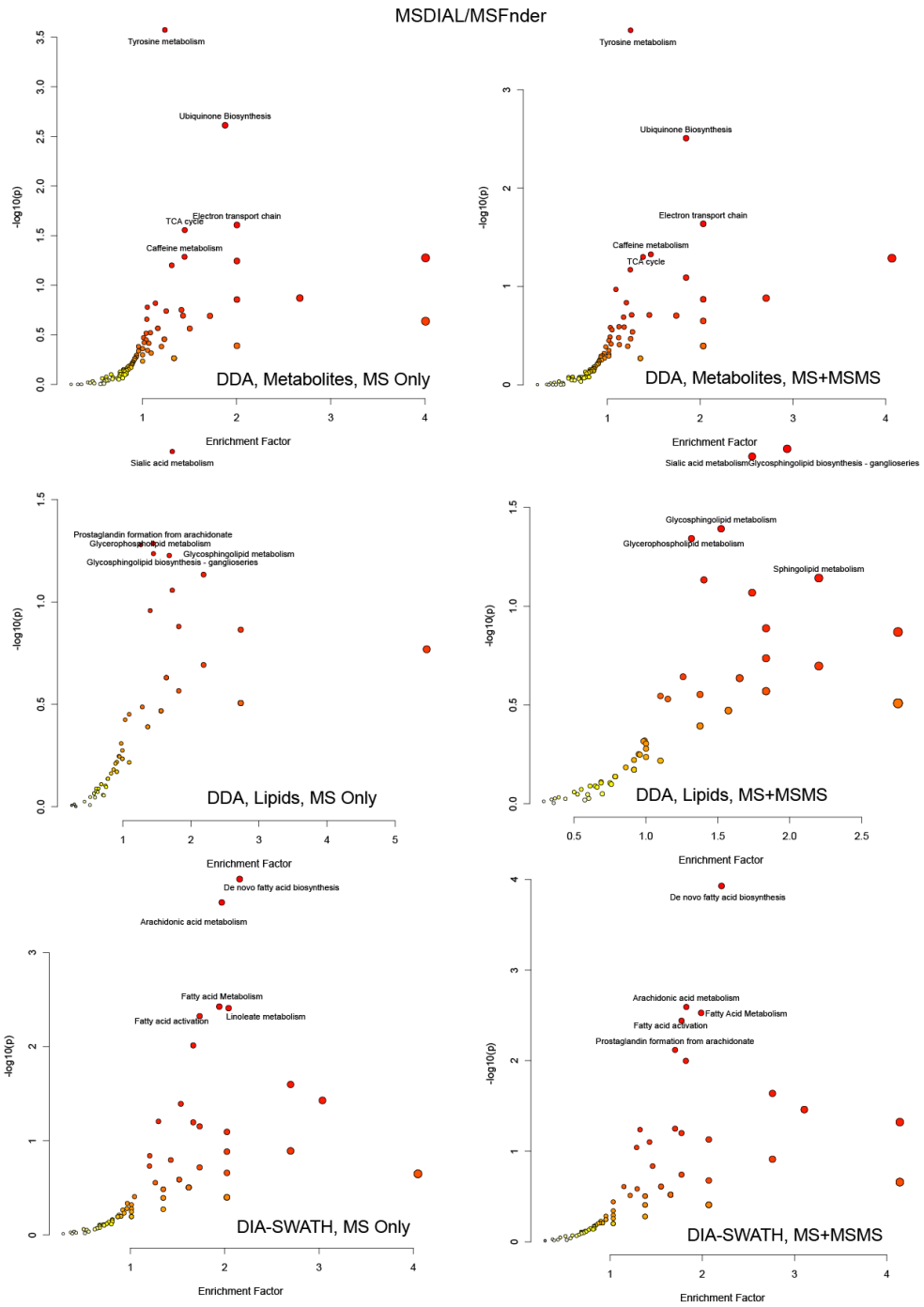


Figure S4.16. Scatter plot of pathway enrichment of datasets generated by MSDIAL / MSFinder. The MS features and MS/MS compounds used in this figure were processed by MSDIAL/MSFinder. Pathway analysis was performed by MetaboAnalystR. In most cases, integrating MS/MS results into the pipeline increases the discovery of pathways or improves the statistical significance compared to using "MS only".

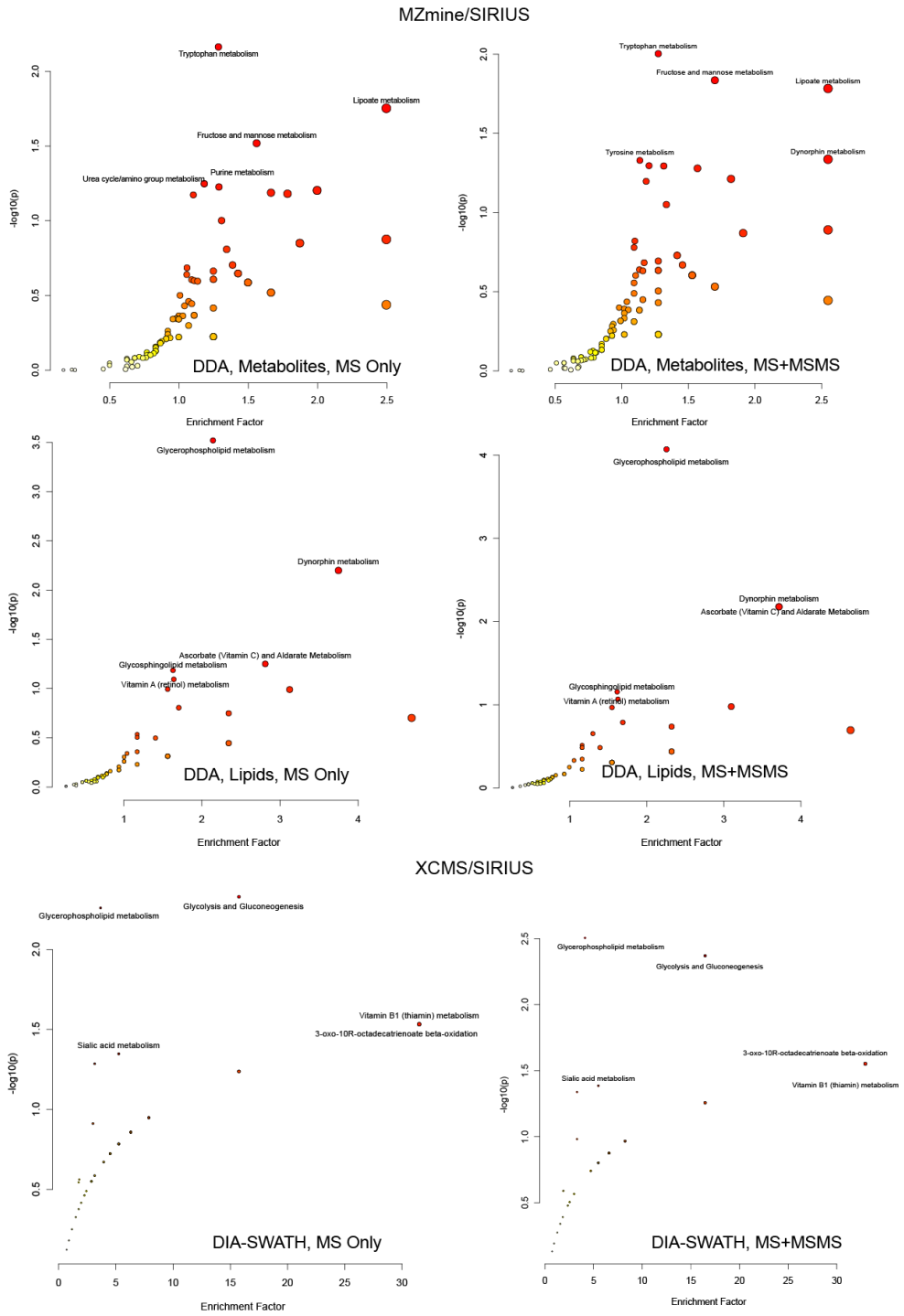


Figure S4.17. Scatter plot of pathway enrichment of datasets generated by MZmine / SIRIUS or XCMS / SIRIUS. The MS features and MS/MS compounds used in this figure were processed by MZmine/SIRIUS (for DDA) or XCMS/SIRIUS (for SWATH-DIA). Pathway analysis was performed by MetaboAnalystR. In most cases, integrating MS/MS results into the pipeline

increases the discovery of pathways or improves the statistical significance compared to using "MS only".

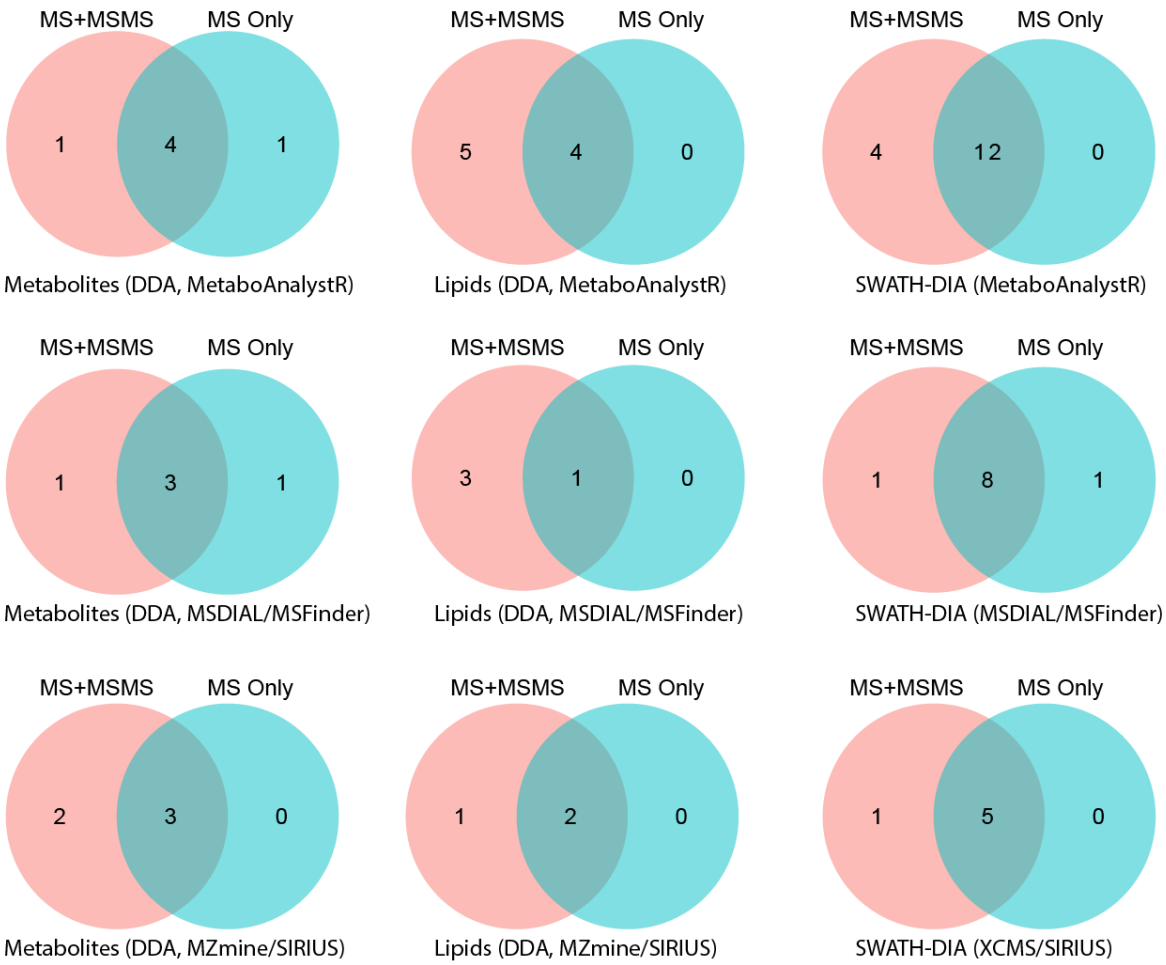


Figure S4.18. Venn diagram of pathway analysis results from different datasets. These diagrams summarize the intersections of pathways identified by MetaboAnalystR *mummochog* function on the MS features or MS/MS features from different tools. In general, integrating MS/MS results into *mummichog* functions increases the discovery of pathways or improves the statistical significance compared to using "MS only" for most cases.

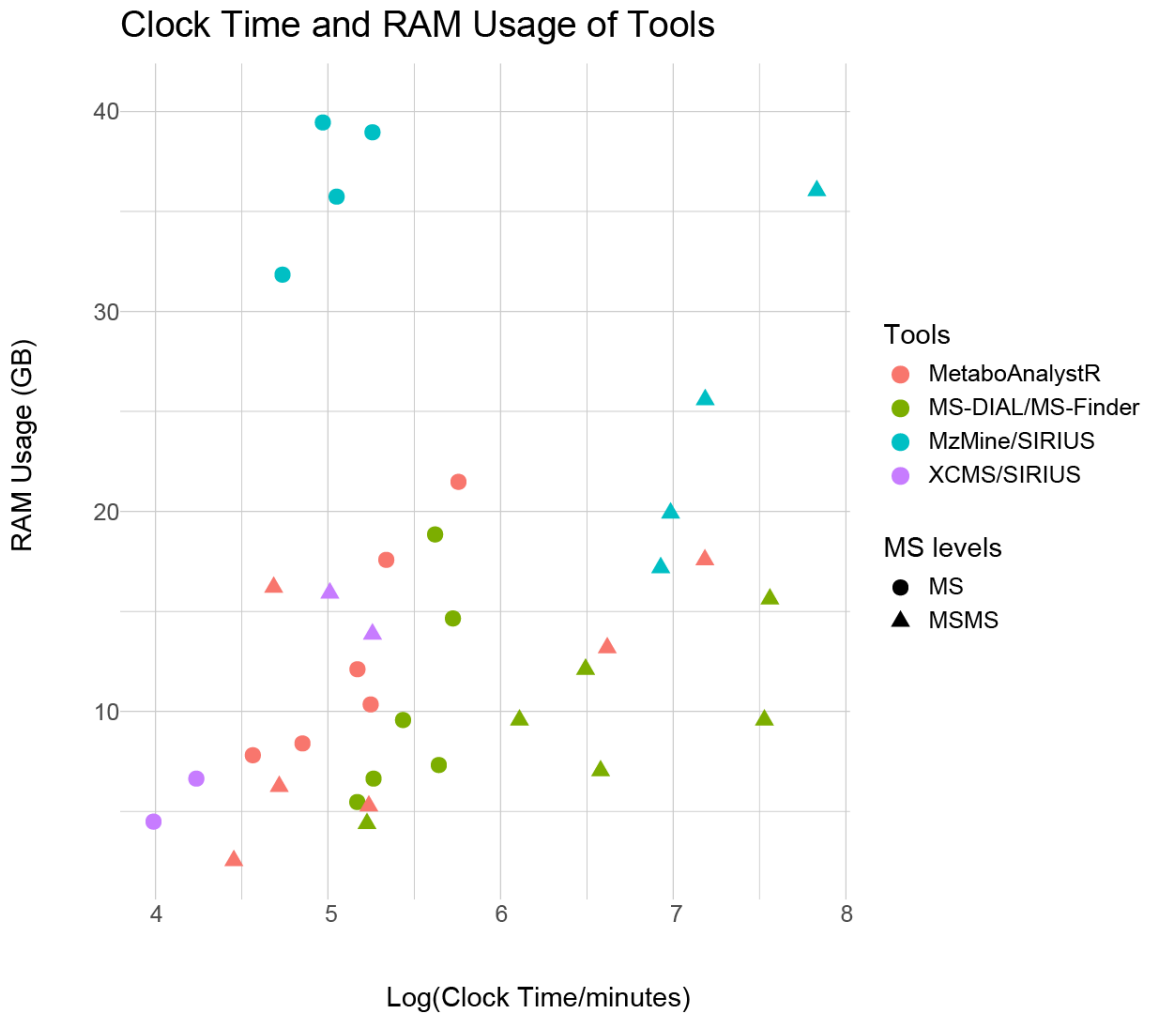


Figure S4.19. Comparison of computational efficiency of different tools. This figure accompanies Table 4.2 in the manuscript, and displays efficiency scores of different tools in different colors based on tools. Compared to other tools, MetaboAnalystR took less RAM and finished both MS and MS/MS detection in an efficient way. MSDIAL/MSFinder usually consumed less RAM, but more time. MZmine could finish the MS features detection by taking less time, but the RAM usage may be high. SIRIUS is very slow based on the testing results because of the remote API access is easily affected by network connection. In addition, the RAM usage may be quite high when there is a big number of features to be annotated.

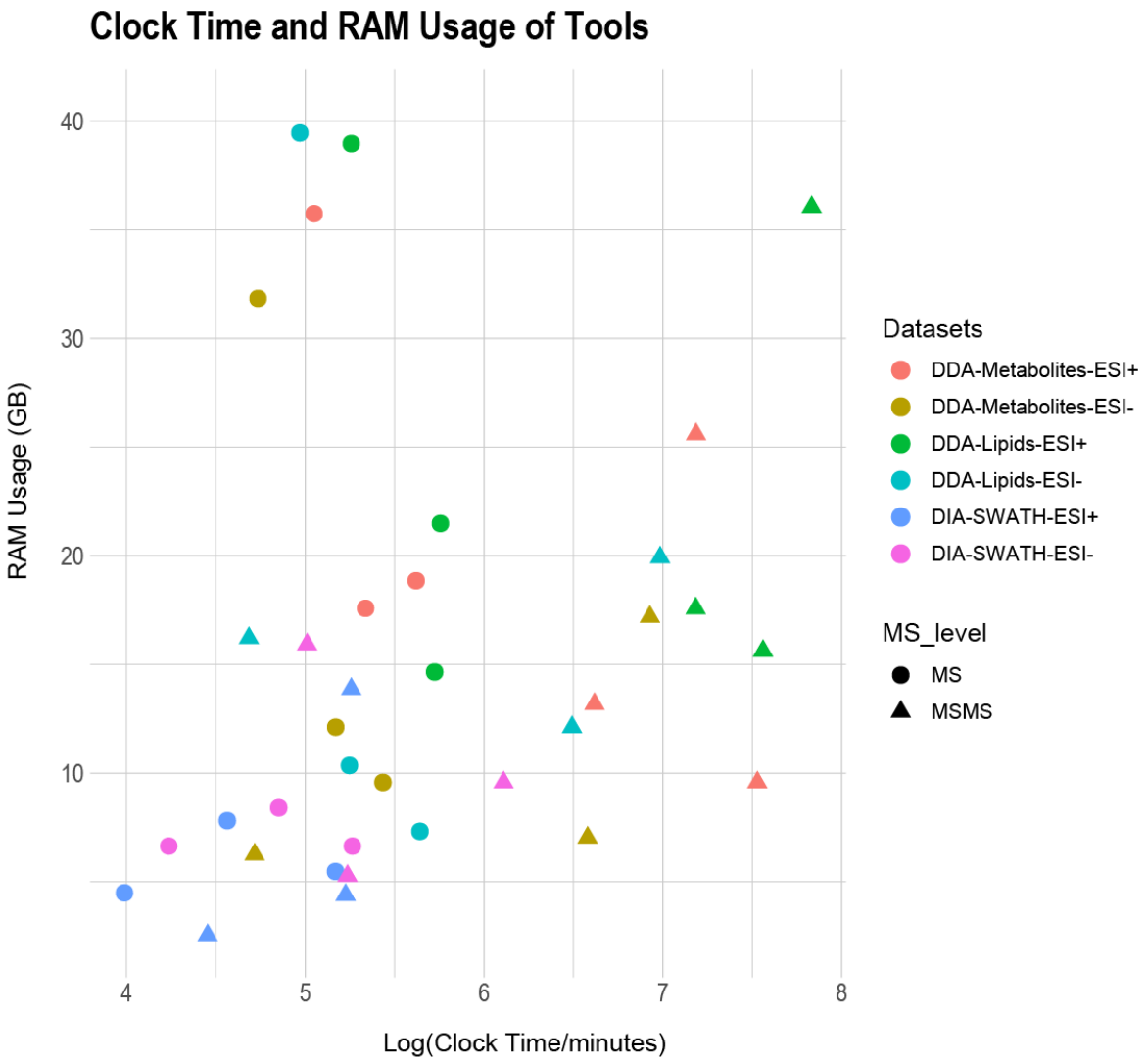


Figure S4.20. Comparison of computational efficiency of different datasets. This figure accompanies Figure 4.7 in the manuscript, and displays efficiency scores of different tools in different colors based on datasets. Processing of SWATH-DIA datasets by various tools is notably faster than other datasets, which is attributed to the smaller size of SWATH-DIA (n=30) compared to others (n=160). The computational performance shows no clear difference among DDA datasets, suggesting that the computational efficiency is primarily determined by the tools used, rather than by the datasets.

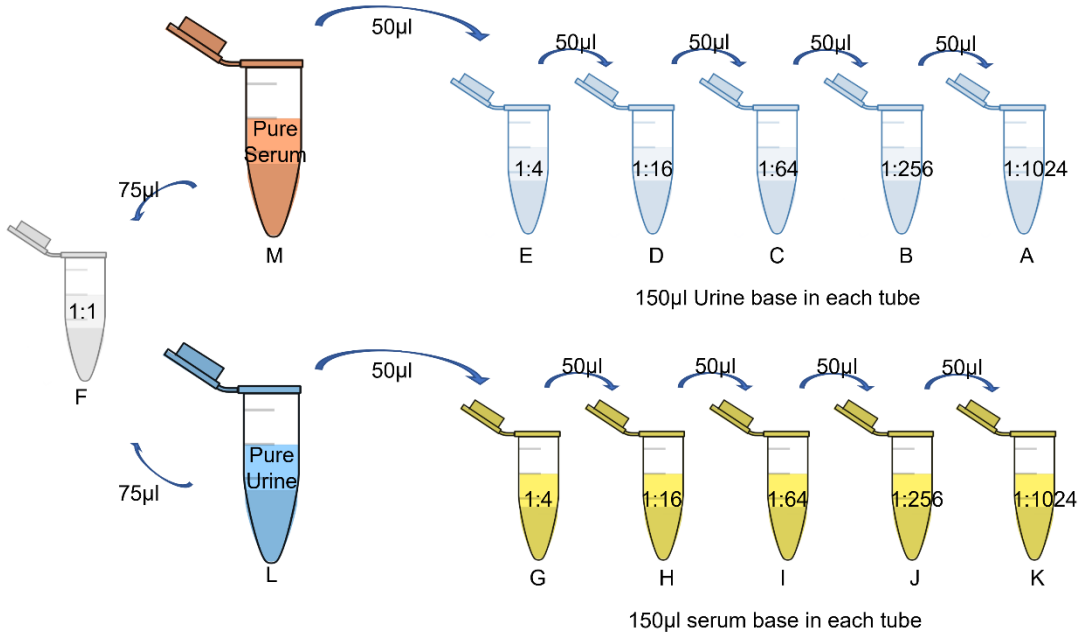


Figure S4.21. Steps involved in preparing a serial dilution. To begin, tubes A to E are filled with 150 µl of base urine, while tubes G to K are filled with 150 µl of base serum. The quadruple dilution process is carried out by adding urine to serum or serum to urine in a quadruple dilution manner. Tube F contains an equal volume of serum and urine.

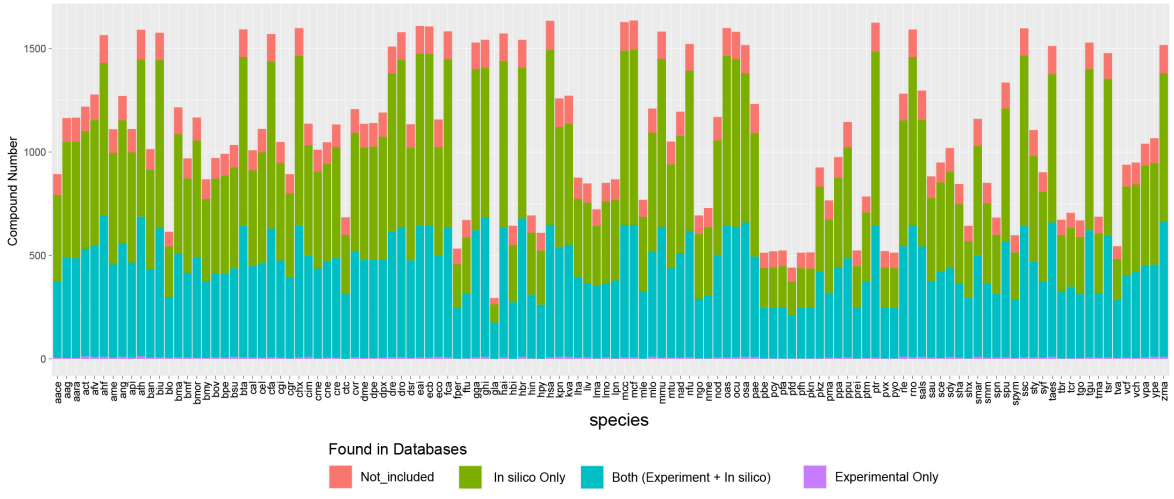


Figure S4.22. Statistics of compounds in pathway reference library. The summary encompasses all 3,456 compounds in the database, with over 90% of the compounds having MS/MS records

included. Moreover, over half of the compounds in the library feature both experimental and in-silico MS/MS spectra.

ID	CompoundName	DBID	PrecursorMZ	PrecursorType	Formula	Smiles	InchiKey	InstrumentType	CollisionEnergy	RetentionTime	Ontology	NumberOfPeak	MS2Peaks
Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter	Filter
1	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	7	68.952 2...
2	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	4	110.975 69...
3	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	18	68.452 21...
4	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	4	110.975 64...
5	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	21	67.951 7...
6	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	5	110.975 76...
7	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	20	68.452 14...
8	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	7	83.049 10...
9	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	19	68.452 77...
10	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	8	83.049 15...
11	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	19	68.452 28...
12	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	6	110.975 94...
13	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	21	67.951 12...
14	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	4	110.975 74...
15	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	20	67.951 10...
16	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	5	110.975 31...
17	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	18	67.951 6...
18	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	7	110.975 29...
19	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	20	68.452 25...
20	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	6	110.975 65...
21	Pyruvic acid	BMDM...	89.02	[M+H]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	22	68.452 48...
22	Pyruvic acid	BMDM...	111.01	[M+Na]+	C3H4O3	O=C(O)C...	LCTONWC...	Orbitrap	10.0	2.2	Alpha-keto...	3	110.975 88...

Figure S4.23. Schema of MS/MS reference library for MetaboAnalystR. The reference library is based on an SQLite database, which supports multiple tables. Each table in the database should contain the following mandatory columns: ID, CompoundName, Precursor MZ, Precursor Type, Formula, InChIKeys, and MS2Peaks. Other columns can be included as required, but they are not mandatory.

Table S4.1. Summary of compounds in different database options

MS/MS Reference Library	Records	Unique Compounds	Size
Pathway Library	172,370	3,456	138.2 MB
Biological Library	864,386	49,055	744.0 MB
Lipids Library	3,221,409	878,220	1.9 GB
Complete Library	10,420,215	1,551,012	7.2 GB

Table S4.2. Summary of identified compounds by different tools (DDA, ESI^-)

Tools	Number of detected standards (MS1)	Compounds correctly annotated (MS2)	Percentage	Time elapsed (1 CPU core)
MS-DIAL + MS-FINDER	271	121	26.4%	36 min
MZmine + SIRIUS	317	124	27.0%	84 min
MetaboAnalyst	336	194	42.3%	28 min
MetaboAnalyst (nonDeco)	336	185	40.3%	15 min

Table S4.3. Summary of identified compounds by different tools (SWATH-DIA ESI^+)

Tools	Number of detected standards (MS1)	Compounds correctly annotated (MS2)	Percentage	Time elapsed (1 CPU core)
MS-DIAL + MS-FINDER	5	1	0.25%	3 min
XCMS + SIRIUS	108	42	10.3%	~ 12 h
MetaboAnalyst	324	143	35.22%	14 min
MetaboAnalyst (PathwayDB)	324	148	36.45%	5 min

Table S4.4. Summary of identified compounds by different tools (SWATH-DIA ESI^-)

Tools	Number of detected standards (MS1)	Compounds correctly annotated (MS2)	Percentage	Time elapsed (1 CPU core)
MS-DIAL + MS-FINDER	6	3	0.65%	12 min
XCMS + SIRIUS	107	46	10.02%	~ 13 h
MetaboAnalyst	241	102	22.22%	24 min
MetaboAnalyst (PathwayDB)	241	97	21.13%	3 min

Table S4.5. Pathway enrichment results of polar metabolites datasets from three tools

Pathways	Pathway total	Hits.total	Hits.sig	Expected	P value	Results	Tools
Fatty acid activation	74	35	14	6.6766	0.0012	MS_only	MetaboAnalystR
Glycolysis and Gluconeogenesis	49	16	13	1.9637	0.0037	MS_only	
De novo fatty acid biosynthesis	106	30	10	6.2838	0.0070	MS_only	
Chondroitin sulfate degradation	37	4	4	0.78548	0.0209	MS_only	
Biopterin metabolism	22	3	3	0.78548	0.0209	MS_only	
Heparan sulfate degradation	34	4	4	0.78548	0.0209	MS_only	
Glycerophospholipid metabolism	156	35	14	9.6221	0.0216	MS_only	
Butanoate metabolism	34	14	8	3.3383	0.0228	MS_only	
TCA cycle	31	8	4	1.3746	0.0241	MS_only	
Arachidonic acid metabolism	95	30	23	4.1238	0.0249	MS_only	
Tyrosine metabolism	160	51	18	10.015	0.0311	MS_only	
Leukotriene metabolism	92	28	13	5.302	0.0427	MS_only	
Fatty acid activation	74	27	12	5.8327	0.0009	MS+MS/MS	
De novo fatty acid biosynthesis	106	26	10	5.8327	0.0035	MS+MS/MS	
Leukotriene metabolism	92	21	12	3.8214	0.0039	MS+MS/MS	
Tyrosine metabolism	160	44	18	9.0508	0.0123	MS+MS/MS	
Arachidonic acid metabolism	95	18	13	3.8214	0.0154	MS+MS/MS	
Prostaglandin formation from arachidonate	78	19	8	3.218	0.0186	MS+MS/MS	
Glycerophospholipid metabolism	156	32	13	9.453	0.0189	MS+MS/MS	
Phosphatidylinositol phosphate metabolism	59	16	9	2.6147	0.0221	MS+MS/MS	
Chondroitin sulfate degradation	37	3	3	0.80451	0.0229	MS+MS/MS	
Biopterin metabolism	22	3	3	0.80451	0.0229	MS+MS/MS	
Heparan sulfate degradation	34	3	3	0.80451	0.0229	MS+MS/MS	
Glycolysis and Gluconeogenesis	49	13	11	1.4079	0.0269	MS+MS/MS	
Butanoate metabolism	34	10	6	2.8158	0.0328	MS+MS/MS	
Vitamin D3 (cholecalciferol) metabolism	16	10	5	2.2124	0.0394	MS+MS/MS	
Ascorbate (Vitamin C) and Aldarate Metabolism	29	13	7	3.0169	0.0463	MS+MS/MS	
TCA cycle	31	6	3	1.0056	0.0495	MS+MS/MS	
De novo fatty acid biosynthesis	106	32	15	7.4104	0.0001	MS_only	MSDIAL MSFINDER
Arachidonic acid metabolism	95	54	45	9.1394	0.0002	MS_only	
Linoleate metabolism	46	23	13	6.1753	0.0037	MS_only	
Fatty Acid Metabolism	63	18	9	6.1753	0.0037	MS_only	
Fatty acid activation	74	36	14	8.6454	0.0047	MS_only	
Prostaglandin formation from arachidonate	78	40	18	8.3984	0.0096	MS_only	
Heparan sulfate degradation	34	6	5	1.4821	0.025	MS_only	

Chondroitin sulfate degradation	37	4	4	0.98805	0.0372	MS_only	
Glycosphingolipid metabolism	67	23	15	7.1633	0.0404	MS_only	
De novo fatty acid biosynthesis	106	32	15	7.2442	0.0001	MS+MS/MS	
Arachidonic acid metabolism	95	53	42	8.2101	0.0025	MS+MS/MS	
Fatty Acid Metabolism	63	18	9	6.0368	0.0029	MS+MS/MS	
Fatty acid activation	74	36	14	8.4516	0.0036	MS+MS/MS	
Prostaglandin formation from arachidonate	78	39	17	8.2101	0.0076	MS+MS/MS	
Linoleate metabolism	46	23	13	6.0368	0.0100	MS+MS/MS	
Heparan sulfate degradation	34	6	5	1.4488	0.0229	MS+MS/MS	
Chondroitin sulfate degradation	37	4	4	0.96589	0.0346	MS+MS/MS	
Prostaglandin formation from dihomo gama-linoleic acid	11	1	1	0.48295	0.0476	MS+MS/MS	
Glycolysis and Gluconeogenesis	49	5	4	0.12698	0.0046	MS_only	XCMS SIRIUS
Glycerophospholipid metabolism	156	30	6	1.3651	0.0055	MS_only	
3-oxo-10R-octadecatrienoate beta-oxidation	27	4	4	0.031746	0.0293	MS_only	
Vitamin B1 (thiamin) metabolism	20	1	1	0.031746	0.0293	MS_only	
Sialic acid metabolism	107	14	6	0.38095	0.0449	MS_only	
Glycerophospholipid metabolism	156	28	6	1.2146	0.0031	MS+MS/MS	
Glycolysis and Gluconeogenesis	49	5	4	0.12146	0.0042	MS+MS/MS	
3-oxo-10R-octadecatrienoate beta-oxidation	27	4	4	0.030364	0.0280	MS+MS/MS	
Vitamin B1 (thiamin) metabolism	20	1	1	0.030364	0.0280	MS+MS/MS	
Sialic acid metabolism	107	13	6	0.36437	0.0412	MS+MS/MS	
Glycosphingolipid metabolism	67	21	7	0.91093	0.0458	MS+MS/MS	

Table S4.6. Pathway enrichment results of polar metabolites datasets from three tools

Pathways	Pathway total	Hits.total	Hits.sig	Expected	P value	Methods	Tools
Dynorphin metabolism	8	5	4	3.2683	0.0010	MS_only	MetaboAnalystR
Bile acid biosynthesis	82	56	45	19.994	0.0219	MS_only	
Prostaglandin formation from dihomo gama-linoleic acid	11	8	6	7.3056	0.0271	MS_only	
Carnitine shuttle	72	30	18	12.689	0.0352	MS_only	
Vitamin D3 (cholecalciferol) metabolism	16	14	12	7.8823	0.0477	MS_only	
Dynorphin metabolism	8	5	3	2.0546	0.00552	MS+MS/MS	
Vitamin D3 (cholecalciferol) metabolism	16	14	13	8.0316	0.0106	MS+MS/MS	
Bile acid biosynthesis	82	55	39	16.997	0.01829	MS+MS/MS	
Tryptophan metabolism	94	65	36	36.236	0.0290	MS+MS/MS	
Carnitine shuttle	72	32	21	14.195	0.0303	MS+MS/MS	
Tyrosine metabolism	160	100	70	82.367	0.0002	MS_only	MSDIAL MSFINDER
Ubiquinone Biosynthesis	10	8	6	7.9871	0.00244	MS_only	
Electron transport chain	7	2	2	3.4944	0.0246	MS_only	
TCA cycle	31	17	11	11.731	0.0277	MS_only	
Tyrosine metabolism	160	99	68	79.193	0.0002	MS+MS/MS	
Ubiquinone Biosynthesis	10	8	6	8.1161	0.0031	MS+MS/MS	
Electron transport chain	7	2	2	3.4432	0.0230	MS+MS/MS	
Caffeine metabolism	11	10	8	8.8539	0.0471	MS+MS/MS	
Tryptophan metabolism	94	69	41	27.25	0.0068	MS_only	MZmine SIRIUS
Lipoate metabolism	8	4	4	1.6029	0.0176	MS_only	
Fructose and mannose metabolism	33	27	22	6.4118	0.0303	MS_only	
Tryptophan metabolism	94	69	40	26.683	0.0099	MS+MS/MS	
Fructose and mannose metabolism	33	25	21	5.8859	0.0146	MS+MS/MS	
Lipoate metabolism	8	4	4	1.5696	0.0164	MS+MS/MS	
Dynorphin metabolism	8	3	3	1.1772	0.04609	MS+MS/MS	
Tyrosine metabolism	160	103	54	36.1	0.046879	MS+MS/MS	

Table S4.7. Pathway enrichment results of lipids datasets from three tools

Pathways	Pathway total	Hits.total	Hits.sig	Expected	P value	Methods	Tools
Glycerophospholipid metabolism	156	31	13	11.809	0.00274	MS_only	MetaboAnalystR
Tyrosine metabolism	160	35	10	5.3565	0.00603	MS_only	
Prostaglandin formation from arachidonate	78	57	11	7.6696	0.00625	MS_only	
Sialic acid metabolism	107	18	4	3.287	0.02279	MS_only	
Glycerophospholipid metabolism	156	31	14	13.746	0.00841	MS+MS/MS	
Tyrosine metabolism	160	25	10	4.9485	0.00993	MS+MS/MS	
Prostaglandin formation from arachidonate	78	53	13	8.11	0.01077	MS+MS/MS	
Glycosphingolipid biosynthesis - globoseries	16	4	2	0.68729	0.01609	MS+MS/MS	
Glycosphingolipid metabolism	67	23	7	6.1856	0.01877	MS+MS/MS	
Glycosphingolipid biosynthesis - ganglioseries	62	10	6	3.9863	0.02166	MS+MS/MS	
Fructose and mannose metabolism	33	12	4	0.82474	0.02925	MS+MS/MS	
Sialic acid metabolism	107	18	5	5.0859	0.03374	MS+MS/MS	
Sphingolipid metabolism	3	3	2	1.512	0.03896	MS+MS/MS	
Sialic acid metabolism	107	26	8	6.9532	0.01843	MS_only	MSDIAL MSFINDER
Glycosphingolipid biosynthesis - ganglioseries	62	13	6	4.5404	0.01600	MS+MS/MS	
Sialic acid metabolism	107	26	8	6.9014	0.01749	MS+MS/MS	
Glycosphingolipid metabolism	67	24	11	8.536	0.04060	MS+MS/MS	
Glycerophospholipid metabolism	156	33	16	16.709	0.04551	MS+MS/MS	
Glycerophospholipid metabolism	156	32	15	7.4704	0.00030	MS_only	MZmine SIRIUS
Dynorphin metabolism	8	5	4	1.0672	0.00632	MS_only	
Glycerophospholipid metabolism	156	32	16	7.5352	8.54E-05	MS+MS/MS	
Ascorbate (Vitamin C) and Aldarate Metabolism	29	5	4	1.0765	0.00664	MS+MS/MS	
Dynorphin metabolism	8	5	4	1.0765	0.00664	MS+MS/MS	

Table S4.8. Demographics of all subjects involved in current study

Items	Values
Age	28.4 ± 5.6
Gender (F/M)	12 (5/7)
BMI	22.7 ± 4.1
Ethnicities	Caucasian/South Asian/East Asian/Latino/Mixed (7/1/2/1/1)
Diabetes	None
Pregnant	None
Breastfeeding	None

Table S4.9. Chromatographic conditions, gradient procedure instrumental settings

Columns	Time (min)	Flow rate (mL/min)	A (%)	B (%)
C18	0	0.4	95	5
	1	0.4	95	5
	3	0.4	50	50
	15	0.4	20	80
	15.5	0.4	0	100
	19.5	0.4	0	100
	20	0.4	95	5
	Column Temperature	40℃		
	Autosampler Temperature	4℃		
	Injection Volume	5µL		
HILIC	Time (min)	Flow rate (mL/min)	A (%)	B (%)
	0	0.4	1	99
	3	0.4	1	99
	20	0.4	50	50
	21	0.4	95	5
	24	0.4	95	5
	25	0.4	1	99
	35	STOP		
	Column Temperature	35℃		
	Autosampler Temperature	4℃		
	Injection Volume	1µL		

Table S4.10. Parameters of mass spectrometers for both MS1 and MS/MS

MS levels	Parameters		Values
MS1	MS Scan Range		70~1000 m/z
	MS Resolution		70,000
	AGC target		1×10^6
	Maximum IT		200 ms
	Capillary temperature		350℃
	Sheath Gas flow		55 arb
	Aux Gas flow		10 arb
MS/MS	MS level(s)	Parameters	Values
	Full MS	Resolution	70,000 (DDA) 35,000 (DIA)
		AGC target	3×10^6 (DDA) 1×10^6 (DIA)
		Maximum IT	200 ms (DDA) 100 ms (DIA)
		Scan Range	70~1000 *m/z*
	MS2-Settings	Resolution	17,500
		AGC target	2×10^5
		Maximum IT	50 ms (DDA) auto (DIA)
		Loop Count	10 (DDA) 7+3 (DIA of HILIC) 8+2 (DIA of C18)
		TopN	10 (DDA)
		Isolation Window	1.0 (DDA)
		Scan range	200~2,000
		(N)CE	15, 30, 45
		Minimum AGC target	1×10^2 (Targeted DDA of HILIC) 8×10^3 (Untargeted DDA) 1×10^2 (Targeted DDA of C18 Negative) 2×10^2 (Targeted DDA of C18 Positive)
		Intensity threshold	2×10^3 (Targeted DDA of HILIC) 1.6×10^5 (Untargeted DDA) 2×10^3 (Targeted DDA of C18 Negative) 4×10^3 (Targeted DDA of C18 Positive)

Table S4.11. Design of SWATH-DIA for different modes for blood Samples

Mode	MS levels	MZ Starting	MZ Ending	Scan Time/ms	Cycle Duration/ms
C18 ESI^+	Full MS1	69.5	1000.5	140	~900
	SWATH	69.5	140.5	75	
	SWATH	139.5	210.5		
	SWATH	209.5	280.5		
	SWATH	279.5	350.5		
	SWATH	349.5	420.5		
	SWATH	419.5	490.5		
	SWATH	489.5	560.5		
	SWATH	559.5	630.5		
	SWATH	629.5	700.5		
	SWATH	699.5	1000.5		
C18 ESI^-	Full MS1	69.5	1000.5	140	~900
	SWATH	69.5	140.5	75	
	SWATH	139.5	210.5		
	SWATH	209.5	280.5		
	SWATH	279.5	350.5		
	SWATH	349.5	420.5		
	SWATH	419.5	490.5		
	SWATH	489.5	560.5		
	SWATH	559.5	630.5		
	SWATH	629.5	750.5		
	SWATH	749.5	1000.5		
HILIC ESI^+	Full MS1	69.5	1000.5	140	~900
	SWATH	69.5	130.5	75	
	SWATH	129.5	190.5		
	SWATH	189.5	250.5		
	SWATH	249.5	310.5		
	SWATH	309.5	410.5		
	SWATH	409.5	510.5		
	SWATH	509.5	610.5		
	SWATH	609.5	710.5		
	SWATH	709.5	810.5		
	SWATH	809.5	1000.5		
HILIC ESI^-	Full MS1	69.5	1000.5	140	~900
	SWATH	69.5	135.5	75	
	SWATH	134.5	175.5		
	SWATH	174.5	215.5		
	SWATH	214.5	255.5		
	SWATH	254.5	295.5		
	SWATH	294.5	335.5		
	SWATH	334.5	375.5		
	SWATH	374.5	500.5		
	SWATH	499.5	750.5		
	SWATH	749.5	1000.5		

Acknowledgments: The authors truly appreciate the support from all members of the Xia lab.

Conflicts of Interest: The authors declare no conflicts of interest.

Data availability: MS and MS/MS data of blood samples and serial dilutions used in the evaluation of MetaboAnalystR 4.0 have been uploaded to MetabolomicsWorkbench repository as studies ST002796 and ST002798 (https://www.metabolomicsworkbench.org/). The complex standards mixture data was obtained from MetaboLights repository (ID: MTBLS2207). The simple standard mixtures series data was from Curatr (https://curatr.mcf.embl.de/). The standard mixtures used to test false discovery was also from MetaboLights repository (ID: MTBLS1311). Polar and non-polar DDA metabolomics datasets of COVID-19 was downloaded from MetaboLights repository MetaboLights repository (ID: MTBLS2542). The COVID-19 SWATH-DIA metabolomics dataset was obtained from MassIVE repository (ID: MSV000089568).

Code availability: Source code of MetaboAnalystR 4.0 is available from Github (https://github.com/xia-lab/MetaboAnalystR). The codes used for benchmark studies is available also from Github (https://github.com/Zhiqiang-PANG/MetabR4_scripts). A step-by-step tutorial of MetaboAnalystR 4.0 has been provided in MetaboAnalyst website (https://www.metaboanalyst.ca/docs/RTutorial.xhtml).

Preface to Chapter 5

Chapter 5 showcases a meta-analysis of COVID-19 global metabolomics datasets. This chapter mainly aims to achieve the Objective 4. In this chapter, we performed a comprehensive meta-analysis of seven metabolomics datasets obtained from three countries. In this study, the performance of raw spectral processing pipeline and updated functional analysis workflow was validated. This meta-analysis has confirmed the efficacy of MetaboAnalystR and MetaboAnalyst website when solving biological questions from the real world. This study also suggests that extensive dysregulations of multiple metabolic pathways and bioactive metabolites are the metabolic characteristics underlying the progression of COVID-19. Overall, this chapter validated and confirmed the previous chapters by using seven COVID-19 metabolomics datasets.

Chapter 5: Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets

Zhiqiang Pang [1], Guangyan Zhou [1], Jasmine Chong [1] and Jianguo Xia [1,2,*]

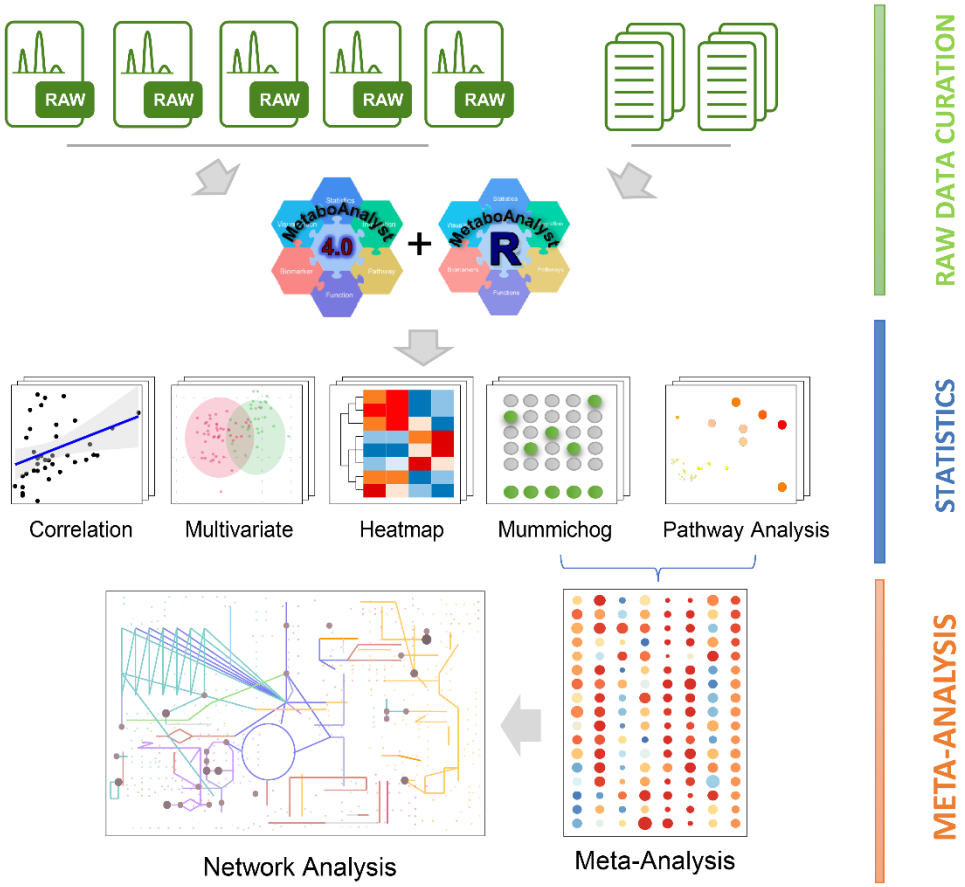
[1] Institute of Parasitology, McGill University, 21111 Lakeshore Road, Ste Anne de Bellevue, QC H9X 3V9, Canada; zhiqiang.pang@mail.mcgill.ca (Z.P.); guangyan.zhou@mail.mcgill.ca (G.Z.); jasmine.chong@mail.mcgill.ca (J.C.)

[2] Department of Animal Science, McGill University, 21111 Lakeshore Road, Ste Anne de Bellevue, QC H9X 3V9, Canada

* Correspondence: jeff.xia@mcgill.ca; Tel.: +1-514-398-8668

This chapter has been published in Metabolites (Metabolites 2021, 11(1), 44)

5.1 Abstract



The novel coronavirus SARS-CoV-2 has spread across the world since 2019, causing a global pandemic. The pathogenesis of the viral infection and the associated clinical presentations depend primarily on host factors such as age and immunity, rather than the viral load or its genetic variations. A growing number of omics studies have been conducted to characterize the host immune and metabolic responses underlying the disease progression. Meta-analyses of these datasets have great potential to identify robust molecular signatures to inform clinical care and to facilitate therapeutics development. In this study, we performed a comprehensive meta-analysis of publicly available global metabolomics datasets obtained from three countries (United States,

China and Brazil). To overcome high heterogeneity inherent in these datasets, we have (a) implemented a computational pipeline to perform consistent raw spectra processing; (b) conducted meta-analyses at pathway levels instead of individual feature levels; and (c) performed visual data mining on consistent patterns of change between disease severities for individual studies. Our analyses have yielded several key metabolic signatures characterizing disease progression and clinical outcomes. Their biological interpretations were discussed within the context of the current literature. To the best of our knowledge, this is the first comprehensive meta-analysis of global metabolomics datasets of COVID-19.

5.2 Introduction

COVID-19 is an unprecedented health emergency driven by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (223). This disease had led to over 1.2 million deaths globally by 5 November 2020 according to the WHO (224). A broad spectrum of clinical presentations has been observed, ranging from asymptomatic, mild, moderate, or severe symptoms, to fatal illness. Such diverse trajectories are believed to be the result of the differences in individual immune responses to COVID-19 (223, 225, 226). A comprehensive understanding of the molecular events underlying different clinical courses is urgently needed to help improve patient management and to accelerate the development of therapeutic strategies.

Metabolism fuels all biological processes in the human body, including immune responses. Blood metabolites are the end products of many systematic processes and are informative indictors of biochemical activities or diseases' phenotypes (11, 227). Powered by the growing applications of high-resolution mass spectrometry (MS), metabolomics has become a key member of the omics toolkit in biomedical research. Multiple metabolomics studies have been recently conducted across the world to study COVID-19, revealing key metabolic dysregulations during the disease's

progression (228-237). For instance, several amino acids have been observed to be positively correlated with the severity of COVID-19 as key indicators of clinical prognosis of the disease (229, 230, 233-235). Perturbations in energy metabolisms such as glycolysis and pentose phosphate pathway, TCA and urea cycle have also been reported (228, 230, 233). The changes in lipid metabolites such as fatty acid, arachidonic acid, glycerophospholipid and sphingolipids are now considered important hallmarks in the pathogenesis of COVID-19 (238, 239). To help to accelerate diagnostics, prognostics, and treatment of the disease, the COVID-19 MS Coalition has been recently launched as a collective community effort to combat the pandemic (240).

Meta-analysis of the available datasets is a promising approach to gain a comprehensive understanding of the pathogenesis of the disease (241, 242), as well as to help to identify robust biomarkers to inform better clinical care and to facilitate therapeutics development. Indeed, meta-analyses of the COVID-19 transcriptomics datasets are quickly emerging and have produced important insights into common and unique gene expression patterns of the disease (243-245). However, to the best of our knowledge, meta-analyses of COVID-19 metabolomics datasets have not been conducted so far. This could be due to a much smaller number of metabolomics studies reported so far or even more likely, due to the practical challenges in dealing with the high levels of heterogeneity inherent in global metabolomics datasets. Unlike transcriptomics in which genes or transcripts can be reliably identified and quantified directly from sequencing data, the features reported by liquid chromatography (LC)-MS-based global metabolomics are peaks characterized by their RTs and *m/z* values, which are insufficient for metabolite identification in general. Moreover, spectral peaks are not usually comparable across different studies due to differences in chromatographic and/or MS conditions.

To address this research gap and to gain a better understanding of the metabolic changes underlying the disease, we systematically collected the COVID-19 global metabolomics datasets that were publicly available as of 5 November 2020 and implemented a computational pipeline for spectra processing, visual exploration and meta-analysis. In this manuscript, we report our findings and discuss their implications within the context of the current understanding of the disease.

5.3 Results

5.3.1. Summary of different datasets and their clinical characteristics

A total of 175 COVID-19 papers were identified in our initial search. After filtering these studies based on our inclusion/exclusion criteria, six studies from the USA, China and Brazil were finally included in this meta-analysis (Figure 5.1). One study from the USA generated two datasets using two different metabolomic platforms. As a result, seven datasets were finally included in this meta-analysis. Among them, five datasets were obtained as raw spectra, including two from MetaboLights (246), one from MassIVE (https://massive.ucsd.edu/) and two directly from the authors. The remaining two datasets were annotated metabolite intensity tables obtained from the Supplementary Materials of the original publications. In total, 438 samples from 337 subjects were included. Table 5.1 summarizes the key information about these datasets. More details on the patient classification criteria, technical information on experimental conditions and the demographic characteristics of all subjects are provided in Tables S5.1–S5.3, respectively.

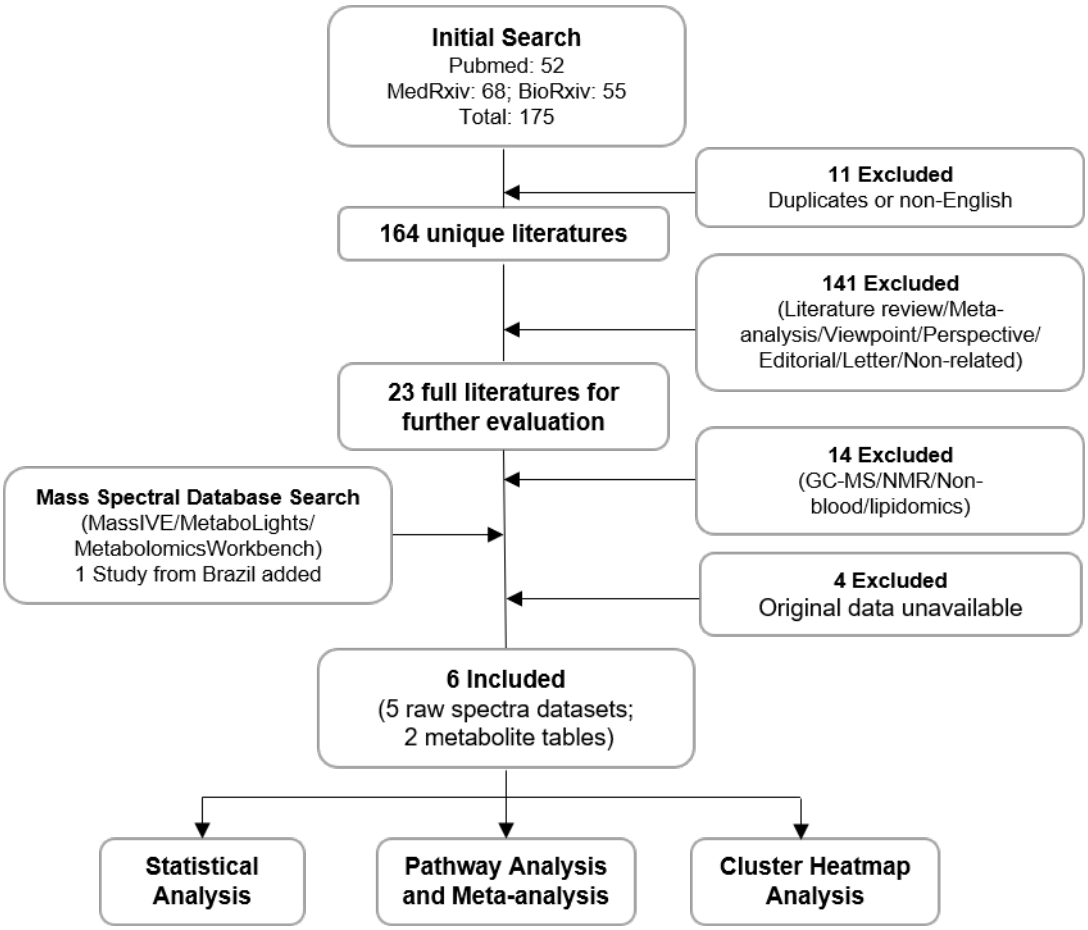


Figure 5.1. The workflow diagram of our data curation process and analysis strategy. The six studies contain seven datasets, five as raw spectra and two as putatively annotated peak tables.

5.3.2. Processing and overview of individual datasets

The five raw spectra datasets were processed using our MetaboAnalystR 3.0 pipeline for optimized peak detection, quantification and alignment (with peak numbers ranging from 2553 to 11,665). The final optimized parameters are provided in Table S5.4. The resulting peak intensity tables from both positive and negative ion modes were combined, median normalized and log transformed for an initial data quality check and visual inspection. The two annotated peak tables were directly used to perform multivariate analysis. Figure 5.2 shows the results from Principal Component Analysis (PCA) of samples between COVID-19 and healthy controls (HCs). No clear

batch effects were observed in the normalized datasets. The first two PCs showed the clear patterns of separation for all datasets (except the C1). The relative low variances explained by the top two PCs could be due to the very high dimensionality of the global metabolomics data, similar to PCA of transcriptomics data. We further analyzed C1 using Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA), which showed a significant separation. The model was evaluated with cross validations (Q2 0.964 and R2 0.803) and permutation tests (*p-value* < 0.001). Overall, these results indicated overall significant metabolic perturbations in COVID-19 patients across all study populations. For two randomly selected datasets, we also compared the results from our spectra processing pipeline against those from two other public spectra processing tools (34, 42) and observed that the PCA from our pipeline produced better separation patterns (data not shown).

Table 5.1. Summary of the seven datasets and the corresponding COVID-19 patient classifications.

Datasets	Chromatogram	MS	Patient Classification					Country
			Total	HC	MM	Severe	Fatal	
A1 (233)	UPLC-C18	Q/E	49	16	27	6	0	USA
A2 * (234)	UPLC-HILIC	Q/TOF	59	20	39	0	0	USA
A3 * (234)	UPLC-C18							
B1 (247)	HPLC- C18	micrOTOF	28	13	6	3	6	Brazil
C1 (229)	UPLC-C18	Triple TOF	76	26	37	11	2	China
C2 ** (230)	UPLC-C18	QE-HF	71	25	37	28	0	China
C3 ** (228)	UPLC- C30	Q/TRAP	96	10	14	11	9	China

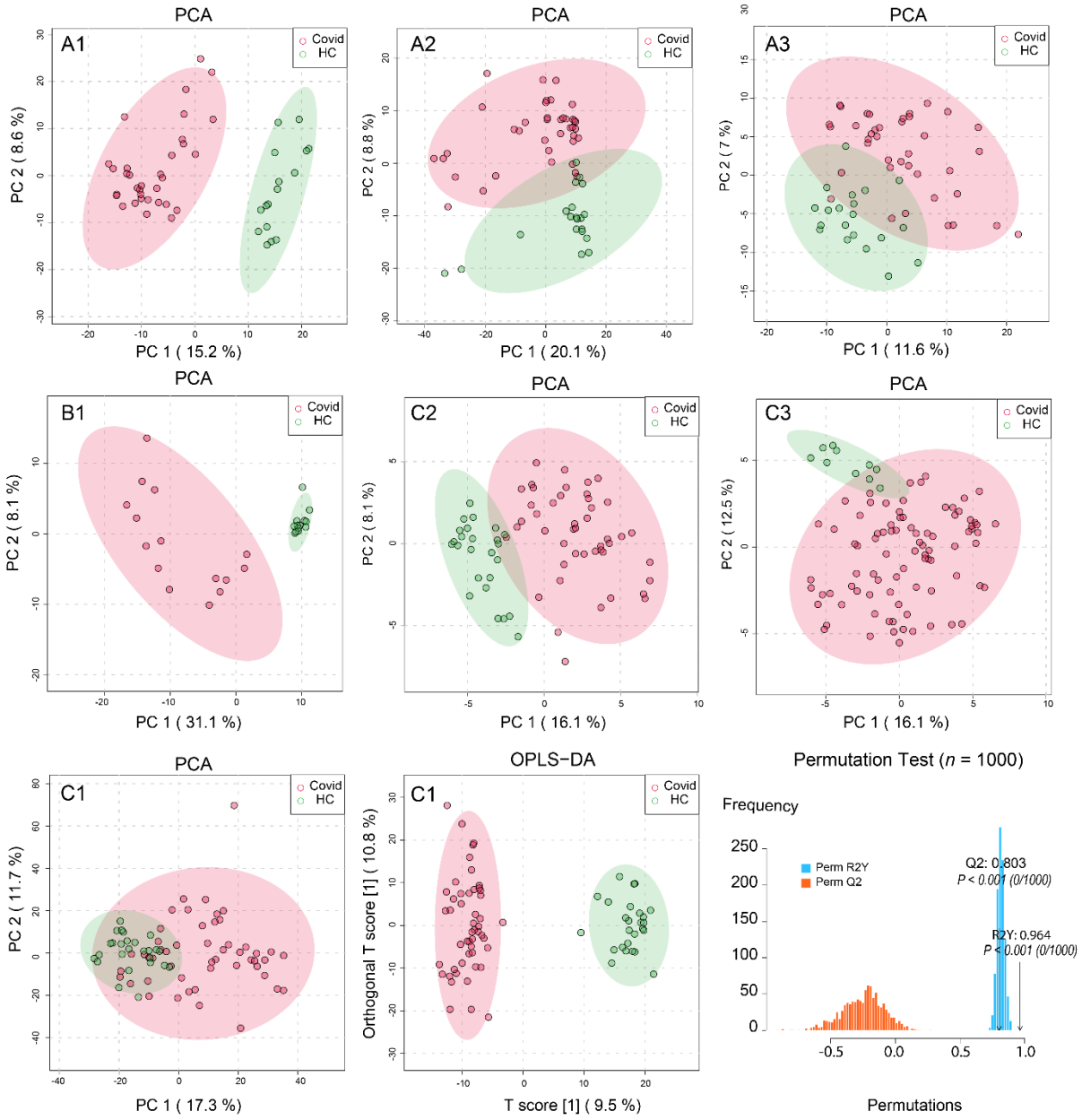


Figure 5.2. Overview of the separation patterns between COVID-19 and healthy controls (HCs) across the seven datasets. The 1st and 2nd rows are the principal component analysis (PCA) results of datasets A1, A2, A3, B1, C2 and C3, respectively. For C1 (3rd row), we performed PCA, followed by Orthogonal Projections to Latent Structures Discriminant Analysis (OPLS-DA) and its validation by permutations (n = 1000).

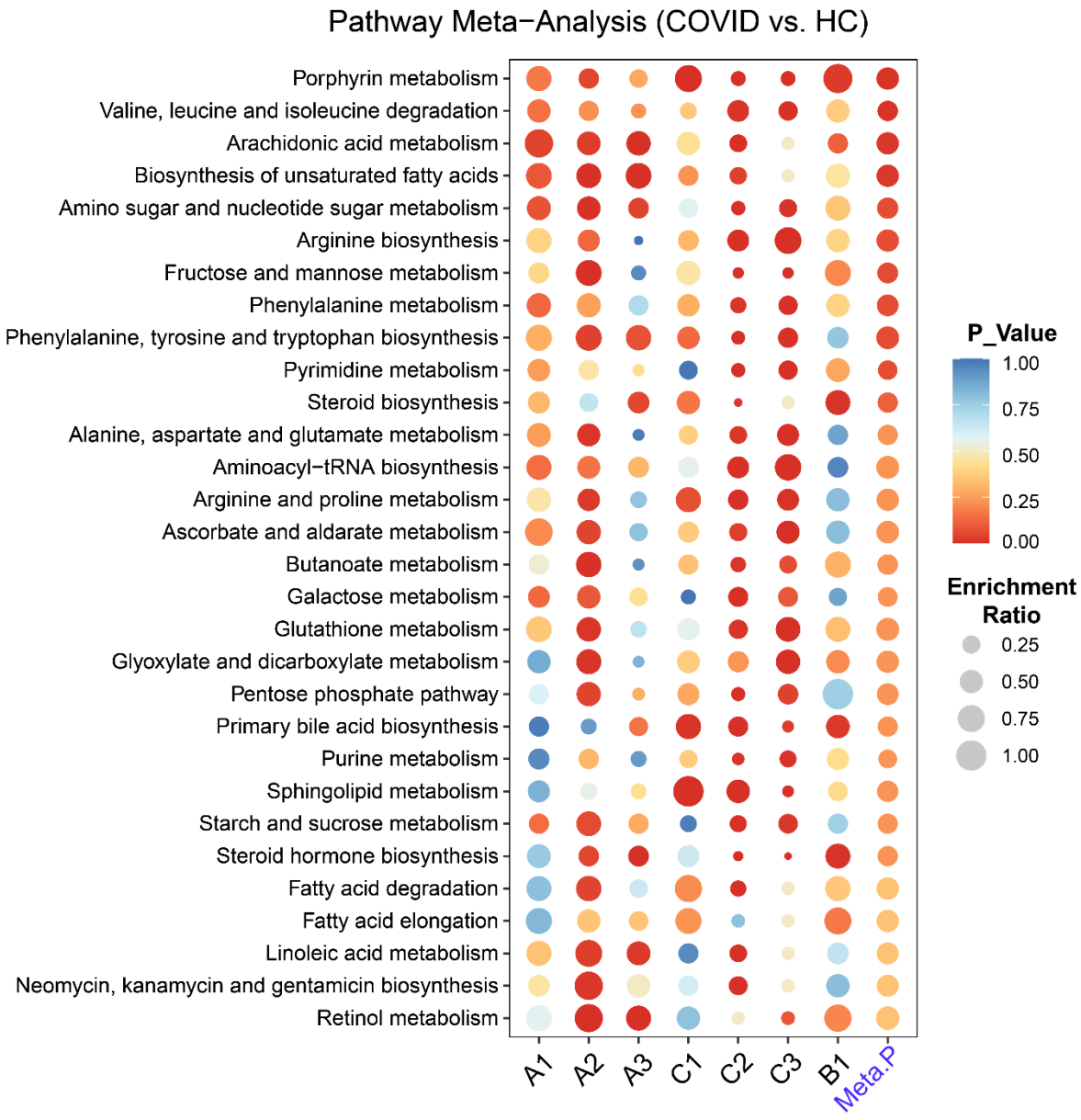


Figure 5.3. Pathway analysis and meta-analysis between COVID-19 and healthy controls (HC) across the seven datasets. Each row represents a pathway and each column represents a dataset. The rightmost column shows the result from the meta-analysis.

5.3.3. Metabolic pathways changes in COVID-19 patient

For each raw spectral dataset, we performed metabolic pathway activity predictions using the *mummichog* approach (101) in MetaboAnalystR 3.0. For the two annotated peak tables, we

performed pathway analysis using the quantitative enrichment method based on their annotations. The human Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database was used in both cases. The pathway-level p-values were further integrated to produce a final ranked list of perturbed pathways (Figure 5.3). Four common pathways were significantly changed between COVID-19 patients and HCs (*p-value* < 0.05). Despite the ambiguities in individual compound assignments, we also attempted to extract the peaks underlying these four perturbed pathways from individual studies. The correlations between these peaks with the symptom onset days were then statistically evaluated. Nine peaks were significant (*p-value* < 0.05), with one negative and eight positive associations (Figure S5.1).

5.3.4. Identification of metabolic hot spots in COVID-19

In order to gain a high-level overview of the changes in metabolic activities caused by COVID-19, we mapped all significant metabolites (based on putative peak annotations) onto the KEGG global metabolic network (Figure 5.4). Network visualizations could reveal coordinated metabolic activities as clusters of metabolites distributed both within and across pathway boundaries. A total of 65 compounds have been reported by at least two datasets within these pathways. The five colored areas indicate the top five pathways identified in Figure 5.3. Other metabolic pathways also contain many metabolites that have received hits from multiple datasets. For instance, cholesterol, d-Mannose, Tyrosine, L-phenylalanine and Bilirubin are the top five most common compounds identified in our meta-analysis, which indicates their potentials as metabolic biomarkers. To complement the meta-analysis, we performed cluster heatmap analysis at feature levels for each dataset. We were able to identify clusters with consistently upregulated or downregulated metabolic patterns between the two conditions in six out of the seven datasets

(Figures S5.2–S5.7). The pathway analyses based on these patterns reported similar results to those in Figure 5.3.

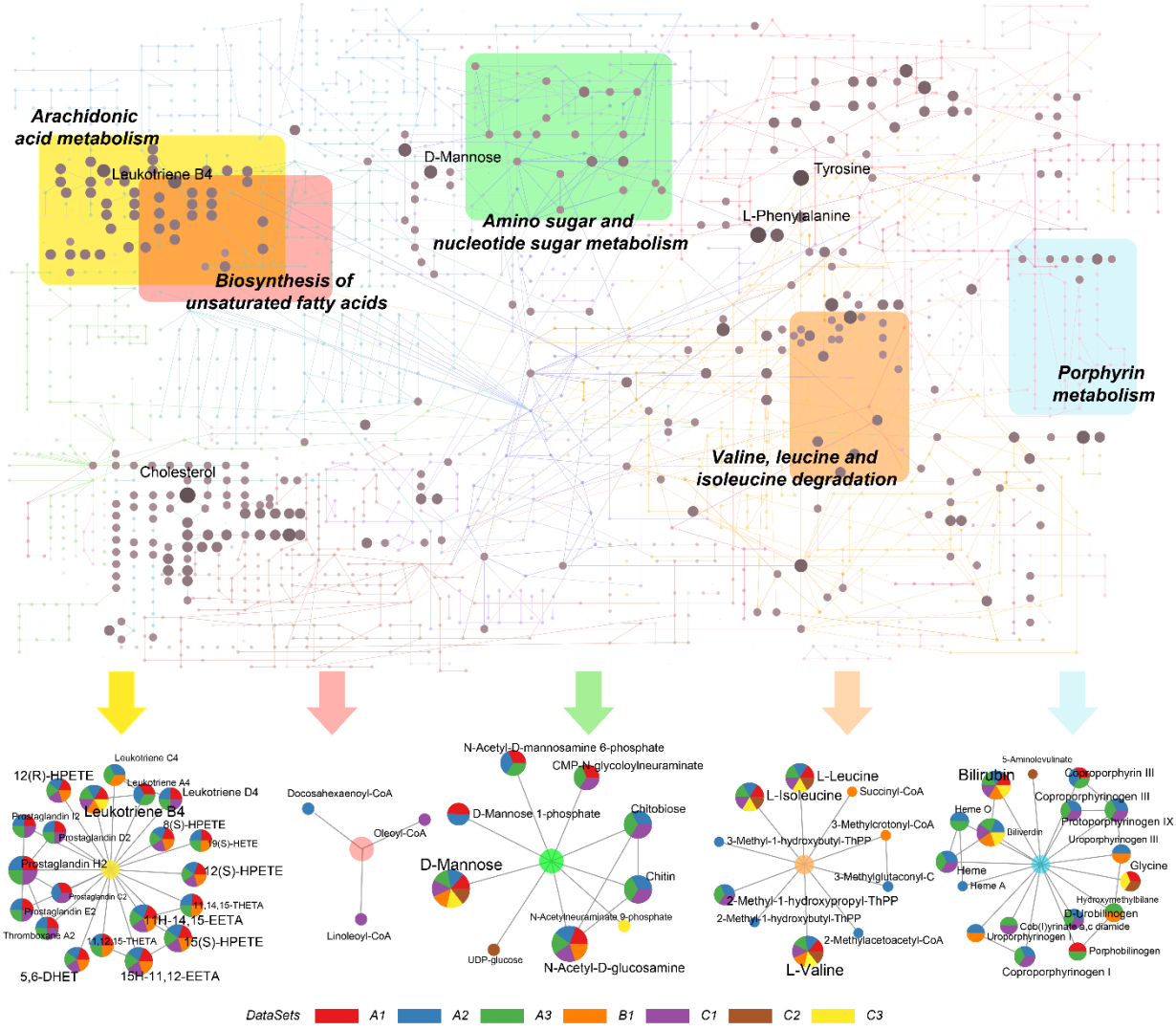


Figure 5.4. Overview of potentially perturbed metabolites and extracted metabolic pathways based on the seven datasets. The top five pathways ranked by their integrated p-values are shown here. The top part is the KEGG global metabolic map, with nodes in brown showing the matched metabolites whose sizes are based on the total number of hits from different datasets. Different colored areas represent different pathways. At the bottom are the five extracted pathways

corresponding to the five colored regions in the map, with nodes shown as pie charts whose sizes and components correspond to the hits from different datasets.

5.3.5. Metabolic changes between mild-to-moderate and severe COVID-19

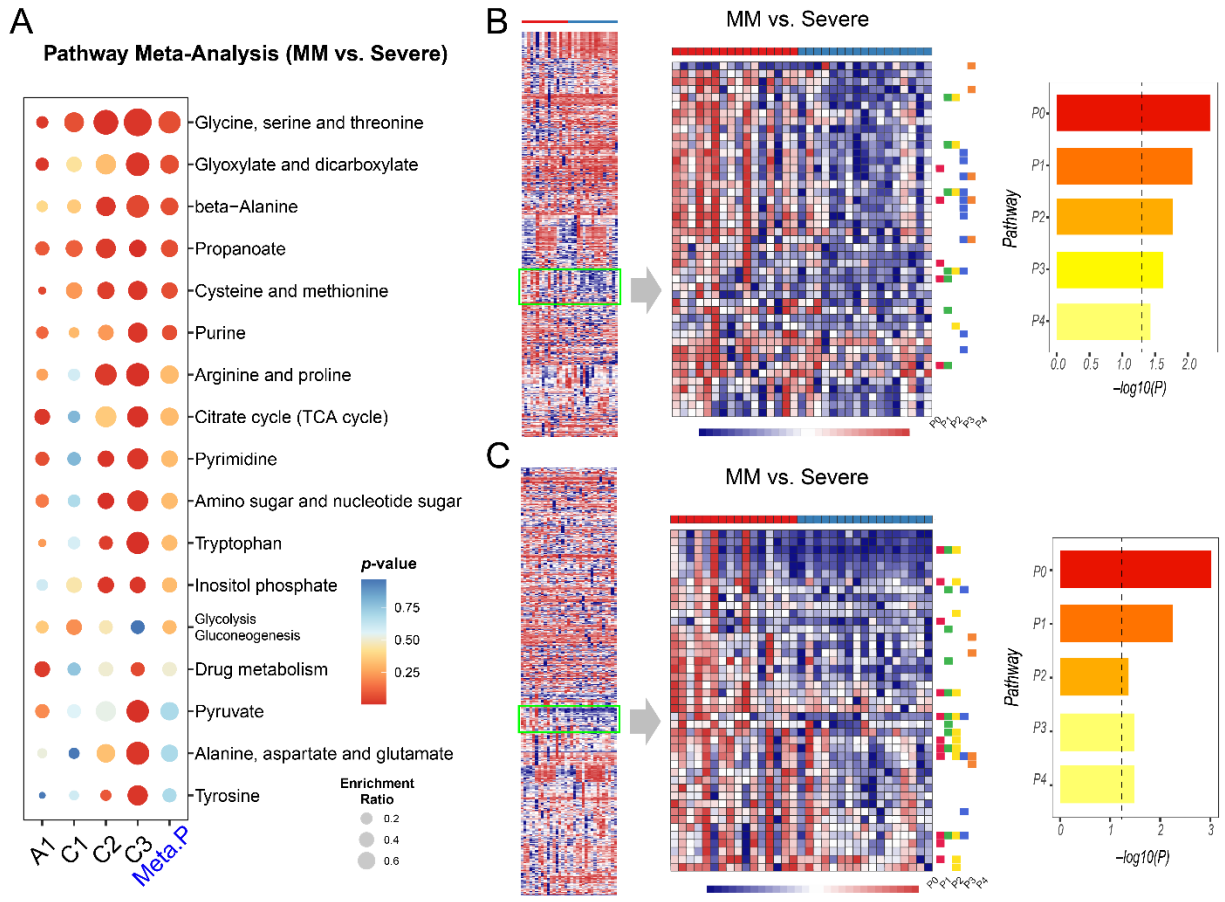


Figure 5.5. Metabolic pathway analysis and cluster heatmap analysis between mild-to-moderate (MM) and severe groups. (A) Summary of pathway analysis and meta-analysis result. (B) Enrichment analysis on a pattern of interest identified in dataset A1 (negative ion mode). P0: Caffeine metabolism; P1: Glyoxylate and dicarboxylate metabolism; P2: Citrate cycle (TCA cycle); P3: Purine metabolism; P4: Lysine degradation. The vertical dashed line in the bar plot is the threshold of $p = 0.05$. (C) Enrichment analysis on a pattern of interest in A1 (positive ion mode).

P0: Glycine, serine and threonine metabolism; P1: Glyoxylate and dicarboxylate metabolism; P2: Cysteine and methionine metabolism; P3: Citrate cycle (TCA cycle); P4: Selenocompound metabolism.

Four datasets contain samples from patients classified as MM and severe COVID-19. The patients with fatal outcomes were also included in the severe group for this comparison based on their clinical status. We first aimed to identify commonly perturbed metabolic pathways across the four datasets. As summarized in Figure 5.5A, six pathways were ranked as the top changed metabolic pathways between MM and severe groups. Similarly, we also mapped the significant metabolites onto the KEGG global metabolic map and noticed that only a few metabolites (L-Alanine, Uridine and Uracil) were shared across the four datasets (Figure S5.8). We then performed cluster heatmap analysis on individual datasets and visually examined the cluster patterns to identify consistent changes between MM and, severe COVID-19. As shown in Figure 5.5B,C, there are some regions that show a general decrease in abundance in the Severe group of A1. A total of eight metabolic pathways were significantly downregulated in this group. Similarly, a consistent metabolic pattern was also found in dataset C3 (Figure S5.9), but not in C1 and C2 (Figure S5.10).

5.3.6 Exploration of metabolic perturbations in fatal COVID-19

Three datasets (C1, C3 and B1) contained COVID-19 patients with mortality information. The C1 dataset was excluded because it contained only two cases to perform meaningful statistical analysis. Metabolic differences between the severe and fatal patients were evaluated with the remaining two datasets (Figure 5.6). Several common metabolic pathway changes were identified from these two datasets (Figure 5.6A). Six metabolites were found as the common hits after mapping to the KEGG global metabolic map (Figure S5.11). From the cluster heatmap of the B1 dataset (positive ion mode), we identified a consistent pattern of change showing five enriched metabolic pathways

(Figure 5.6B). From the cluster heatmap of C3, we combined two regions of interest and identified three enriched metabolic pathways (Figure 5.6C).

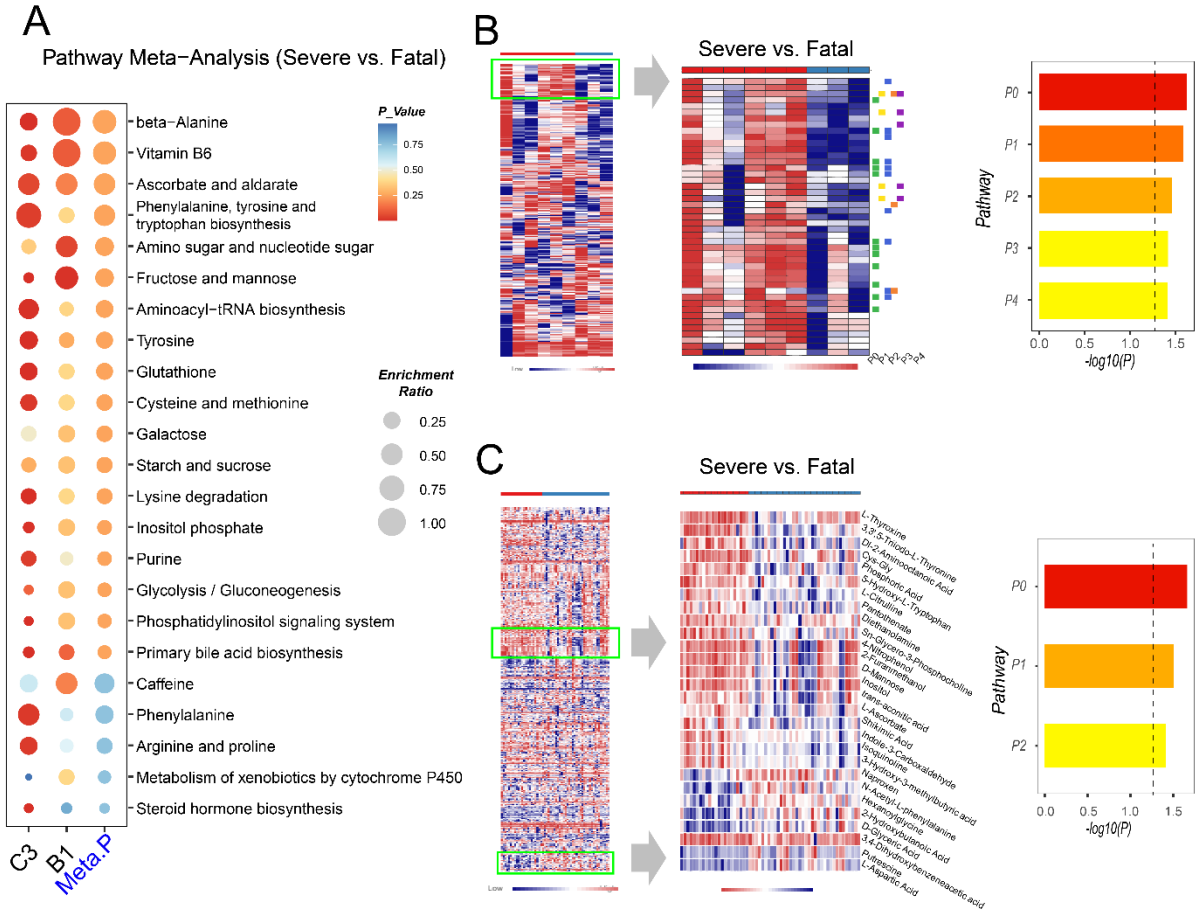


Figure 5.6. Pathway analysis and cluster heatmap analysis between severe and fatal groups. (A) Summary of pathway analysis and meta-analysis result. (B) Enrichment analysis on a metabolic pattern of interest in dataset B1 (positive ion mode). P0: Primary bile acid biosynthesis; P1: D-Glutamine and D-glutamate metabolism; P2: Steroid biosynthesis; P3: Ubiquinone and other terpenoid-quinone biosynthesis; P4: Alanine, aspartate and glutamate metabolism. The vertical dashed line in the bar plot is the threshold of p = 0.05. (C) Enrichment analysis on the combined metabolic patterns of interest in dataset C3. P0: Arginine biosynthesis; P1: Tyrosine metabolism; and P2: Pantothenate and Coenzyme A (CoA) biosynthesis.

5.4 Discussion

Patients with SARS-CoV-2 infection manifest a classical respiratory virus-like clinical course with activated innate and adaptive immune responses (225, 248). Multiple metabolic pathways, such as amino acid metabolism, energy metabolism and lipid metabolism, are involved in the initiation and maintenance of the immune responses in COVID-19. Our meta-analysis has not only confirmed the dysregulations of these pathways as reported by original studies, but also observed novel patterns of metabolic changes underlying the pathogenesis of COVID-19.

Several common metabolic pathways were identified by comparing COVID-19 patients with healthy subjects. The most significantly perturbed pathway is Porphyrin metabolism or Heme biosynthesis, which is consistent with previous reports (249, 250). The SARS-CoV-2 virus could capture hemoglobin, displace iron and decrease the ability of carrying oxygen, thus causing respiratory distress and coagulation reactions, damaging multi-organs (251). The hijacking of the cellular amino acid metabolism to fuel viral proliferation might be a critical mechanism underlying the COVID-19 pathogenesis (252). Arachidonic acid is an endogenous bioactive antiviral lipid, and this metabolic pathway has been suggested to play an important role in susceptibility to COVID-19 (253, 254). The elevated levels of free poly-unsaturated fatty acids are characteristics of COVID-19 patients (231, 255). However, their roles are still controversial (256-258) and warrant further studies.

The heterogeneity of COVID-19 patients shows a wide spectrum of symptoms as well as disease severity. The risk categorization of COVID-19 is difficult because of the complexity of the pathophysiological status of the patients. Therefore, understanding the molecular underpinnings of the disease severities is important to help to reduce the mortality.

Patients with mild-to-moderate (MM) cases of COVID-19 typically have an optimistic prognosis and can recover very quickly. Pathway analysis between MM and severe COVID-19 showed six common perturbed pathways. Most of them were amino acids pathways. Our analysis identified propanoate as a novel pathway in the progression of COVID-19. Propanoate metabolism usually starts with the gut microbiota and enters into immune cells such as macrophages, thereby modulating the biological process (259). The glyoxylate and dicarboxylate metabolism pathway has been reported to be decreased after infection (260). We observed that this pathway was downregulated in severe compared to MM. Downregulation of the TCA might be related to the high energy consumption of SARS-CoV-2 (228). Decreased TCA metabolism would cause an imbalance of anti-oxidization and inflammatory damage (261, 262). Finally, selenocompound is an ex vivo compound originating mainly from gut microbiota (263), and the biological effect of its decrease needs further investigation. Both propanoate metabolism and selenocompound metabolism suggest potential roles played by gut microbiota in the progression of COVID-19, a topic which has gained increasing attention recently (259, 264).

SARS-CoV-2 infection can not only cause pathogenic changes in the respiratory system but can also lead to systematic multi-organ damages and death (265). Preventing fatal COVID-19 is the most important objective in current clinical care. In addition to the observation of extensive dysregulations in amino acid metabolism, our analysis also detected other energy-related pathways such as mannose metabolism as reported previously (230). The change in glutathione metabolism was observed in fatal COVID-19, providing direct evidence for a recent clinical hypothesis that glutathione deficiency could lead to serious manifestation and death in COVID-19 (266). This metabolic pattern also reveals other interesting metabolic signatures. For instance, biosynthesis of bile acid might be a key clinical manifestation of liver damage by SARS-CoV-2 infection (228,

267). The inhibition of its synthesis might accelerate the deterioration of COVID-19 to death (268). Endogenous steroid biosynthesis was found to be decreased, although it could have been caused by medical treatments. Ubiquinone has been reported to alleviate the cytokine storm and restore exhausted T cells in COVID-19 (269). The suppression of its biosynthesis could worsen the disease condition. The role of vitamin B5 biosynthesis on the deterioration of COVID-19 remains unclear, but vitamin B6 has been proposed to ameliorate the severity of COVID-19 (270).

The high level of heterogeneity inherent in global metabolomics datasets poses tremendous challenges to conduct metabolomics meta-analysis at the feature (MS peaks) level. In this study, we utilized the well-established *mummichog* method to first compute pathway activities from MS peaks and performed meta-analysis at the pathway level. There are however, several limitations to this analysis method. The potential bias caused by differences in the extraction procedures and analytical platforms at pathway level remains an open question. Due to the nature of putative annotations, the significant metabolites reported in this study need to be further validated using more targeted approaches. Although the potential confounding factors (diet, ethnicity, medical treatment, etc.) were controlled within each study, they were not considered in the current meta-analysis because most meta-data are incomplete or missing from the original studies. We intend to address this issue by expanding this analysis to include multiple-cohorts-based metabolomics studies when more datasets become available in the coming year. In addition, many signatures are likely to reflect general immune and inflammatory responses. We plan to include studies on other viral infections (such as SARS-CoV and influenza) to identify unique metabolic signatures of this disease as illustrated in a recent meta-analysis based on transcriptomics (271).

5.5 Methods and materials

5.5.1 Data curation

This meta-analysis was strictly conducted based on the PRISMA guidelines (272). All studies were searched for on PubMed, medRxiv (www.medrxiv.org/), and bioRxiv (www.biorxiv.org/) using the search term "(COVID-19) AND (Metabolomics)" before 5 November 2020. The inclusion criteria for further processing were as below: (1) The study should have had a matched healthy control for COVID-19 samples; (2) all raw spectra data or original/annotated peak tables should have been available publicly or upon request; (3) to ensure comparability, only LC-MS-based global metabolomics datasets were included; other metabolomics datasets generated by gas chromatography (GC)-MS or nuclear magnetic resonance (NMR) were excluded. The PRISMA 2009 Flow Diagram is provided in Figure S5.12.

5.5.2 Patient classification

All COVID-19 patients were diagnosed separately at their original hospitals or testing centers. Their disease severities were classified according to a combined standard based on the Guideline of Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (8th) published by the National Health Commission of China (273), WHO R&D Blueprint novel Coronavirus COVID-19 Therapeutic Trial Synopsis (274), and an inflammation correlated cytokine, IL-6 as used in the original studies (233).

5.5.3 Raw spectra processing

Raw LC–MS spectra were first converted and centroided from vendor format to mzML using ProteoWizard (144). All centroided spectra were processed with an automated pipeline with built-in parameter optimization procedures as described in MetaboAnalystR 3.0 (102). For annotated

peak tables, the names were standardized with the ID conversion tool in MetaboAnalyst (275). The remaining ambiguous compounds/peaks were manually corrected based on HMDB (196).

5.5.4 Statistical analysis

Chemometrics analysis (PCA and OLS-DA) was performed based on the normalized peak tables using the corresponding functions in MetaboAnalystR 3.0. Spearman correlations between the onset days of symptoms and metabolic features were calculated using base R package (v4.0.2). The confidence interval of the significant correlation was set to 0.95.

5.5.5 Metabolic pathway analysis and meta-analysis

The pathways analysis on the datasets from raw spectra in this present study was performed independently for every dataset using *mummichog* (101) from the MetaboAnalystR 3.0 workflow (102). The pathways analysis on the two annotated peak tables was completed with the Pathway Analysis module based on the default quantitative enrichment analysis method and the human KEGG database (215). The meta-analysis was performed at the pathway levels. The combined p-values were computed based on the vote counting method in the metap package in R (v4.0.2) by counting the p-value from two directions and outputting an integrated p-value based on the counting results. The enrichment ratio describes the relative percentage of the empirical compound hits to the whole empirical pathway. The enrichment ratio of the compounds from the annotated peak tables was calculated with the average of the other empirical pathway size as the denominator.

5.5.6 Global metabolic network visualization

The MS peaks from different studies were putatively annotated based on the *mummichog* algorithm and mapped to the KEGG global metabolic network using the Peaks to Pathway module in MetaboAnalyst (275). The sizes of the matched nodes (compounds) corresponded to the number

of hits received from different studies. For those highlighted pathways, the corresponding compounds were extracted, with edges between compounds representing direct interactions based on the KEGG global metabolic reaction network.

5.5.7 Cluster heatmap analysis

The peak intensity tables from the individual datasets were uploaded to the Peaks to Pathway module in MetaboAnalyst 4.0 (275). After normalization, the peak tables were displayed as an interactive heatmap with different clustering options. From the Overview on the left panel, we manually selected patterns of interest to be displayed on the Focus view on the central panel. Pathway activity predictions were performed based on *mummichog* using the peaks in the current Focus view as significant peaks.

5.6 Conclusion

There are significant knowledge gaps in the systems biology of COVID-19. The ongoing multi-omics investigations will continue to yield valuable insights to fill this gap in the coming year. Global metabolomics can provide rich data that complement other omics layers to inform the development of diagnostics, prognostics, and treatment of COVID-19. In this study, we have systematically curated public metabolomics datasets and performed comprehensive data processing, analysis and meta-analysis to identify common as well as unique metabolic signatures underlying different clinical courses of COVID-19. Our results suggest that extensive dysregulations of amino acids metabolism, damage to the oxygen transport in red blood cells, exhaustion of endogenous immune bioactive metabolites and the suppression of multiple physiological processes are the metabolic characteristics underlying the progression of COVID-

19. We will continue to improve the computational workflow and expand the scale and scope of the current meta-analysis when more metabolomics datasets become available in the coming year.

5.7 Supplementary materials

Supplementary Materials: The following are available online at https://www.mdpi.com/2218-1989/11/1/44/s1. Table S5.1: Classification Standards for Different Severities of COVID-19; Table S5.2: Technical information of all datasets included in this study; Table S5.3: Clinical demographics characteristics of all subjects; Table S5.4: Optimized parameters of all datasets for raw spectral processing. Figure S5.1: The Spearman correlation analysis on the onset time (days) with the metabolites in the significantly perturbed pathways; Figure S5.2: Cluster heatmap analysis between COVID-19 and HC groups of Dataset A1; Figure S5.3: Cluster heatmap analysis between COVID-19 and HC groups of Dataset A2; Figure S5.4: Cluster heatmap analysis between COVID-19 and HC groups of Dataset A3; Figure S5.5: Cluster heatmap analysis between COVID-19 and HC groups of Dataset C1; Figure S5.6: Cluster heatmap analysis between COVID-19 and HC groups of Dataset C2; Figure S5.7: Cluster heatmap analysis between COVID-19 and HC groups of Dataset B1; Figure S5.8: Overview of perturbed pathways in COVID-19 across datasets for comparison between mild-to-moderate (MM) and severe COVID-19; Figure S5.9: The metabolic pattern between MM and severe of dataset C3; Figure S5.10: The metabolic pattern between MM and severe of dataset C1 and C2; Figure S5.11: Overview of perturbed pathways in COVID-19 across datasets for comparison between severe and fatal COVID-19; Figure S5.12: PRISMA 2009 Flow Diagram.

Table S5.1. Classification Standards for Different Severities of COVID-19

Severities	Classification Criteria
Mild to Moderate	1). Hospitalized but without oxygen or with low-flow oxygen. 2). Imagological changes found or not, but not meet the standard of Severe. 3). Serum IL-6 concentration less than 90 pg/mL. Meet one of the above.
Severe	1). Intubation and ventilation or high-flow oxygen, or additional organ support. 2). Shortness of breath, RR≥30 beats/min. 3). SO2 less than 93% at rest. 4). PaO2/FiO2 less than 300mmHg. 5). Acute progressing with imagological lesion grew more than 50% in past 24-48 hours. 6). Shocked. 7). Serum IL-6 concentration more than 90 pg/mL Meet one of the above.
Fatal	Reported death as the clinical end.
Non-Covid	Shown similar clinical characteristics including fever and/or cough as COVID-19 patients but tested as negative with nucleic acid.

Table S5.2. Technical Information of all datasets included in this study

	Data	**Platform**	**Column**	**Main Extraction***	**Ion Mode**	**Sample**
A1	Raw	UPLC-QE	C18	Folch -> Choroform/methanol	Neg + Pos	Serum
A2	Raw	UPLC-QTOF	HILIC	Acetone + methanol ->acetonitrile + H2O	Neg + Pos	Serum
A3			C18			
C1	Raw	UPLC-TTOF	C18	Chloroform/methanol -> water -> aqueous/methanol	Neg + Pos	Plasma
C2	Table	UPLC-QEHF	C18	Ethanol ->methanol	Neg + Pos	Serum
C3	Table	UPLC-QTRAP	C30	Methanol + MTBE	Neg + Pos	Plasma
B1	Raw	HPLC-micrOTOF	C18	MeOH:MeCN -> Water	Neg + Pos	Plasma

* Extraction method only show the main steps of all studies briefly.

Table S5.3. Clinical Demographics Characteristics of All Samples

	CONTROL	**COVID-19**		
Characteristics/Severity	Healthy	Mild to Moderate	Severe	Fatal
Age/Years/Sample#				
<20	0	4	0	0
20~60	69	96	36	0
> 60	3	40	23	4
Not Clear *	56	28	22	36
Sex (Sample#)				
Male	50	72	39	7
Female	33	57	19	1
Not Clear *	26	28	22	36
Onset time/Days	-	8.26	9.6	8.12
Treatment				
Hydroxychloroquine	-	28	-	-
Remdesivir	-	3	-	-

* Not Clear: refers to that the related information is not available from the original manuscript. The number refers to the samples number rather than the subjects' number.

Table S5.4. Optimized parameters of all datasets for raw spectral processing

NO	ION	Algorithm	ppm	Peak width	mzdiff	s/n	noise	prefilter	bw
A1	Negative	*centWave*	1.78	3.625, 35.375	0.004	10.65	0	2, 1172.26	2
A1	Positive	*centWave*	1.78	3.625, 42.125	0.006	12.05	0	2, 1172.26	2
A2	Negative	*centWave*	11.13	6.25, 14.75	0.013	15	0	2, 10	2
A2	Positive	*centWave*	17.99	10.125, 40.75	0.016	9.5	0	2, 10	2
A3	Negative	*centWave*	13.8	8.25, 51.5	0.012	16.05	0	2, 10	2
A3	Positive	*centWave*	13.26	4.375, 48.875	0.016	14.5	0	2, 10	2
C1	Negative	*centWave*	10.58	4.5, 19	0.013	13.75	0	2, 10	2
C1	Positive	*centWave*	35.96	7, 15.5	-0.02	10	0	2, 10	2

NO	ION	Algorithm	Critical Value	Consec Missed Limit	Unions	Check back
B1	Negative	*Massifquant*	1.125	2	1	0
B1	Positive	*Massifquant*	1.125	2	1	0

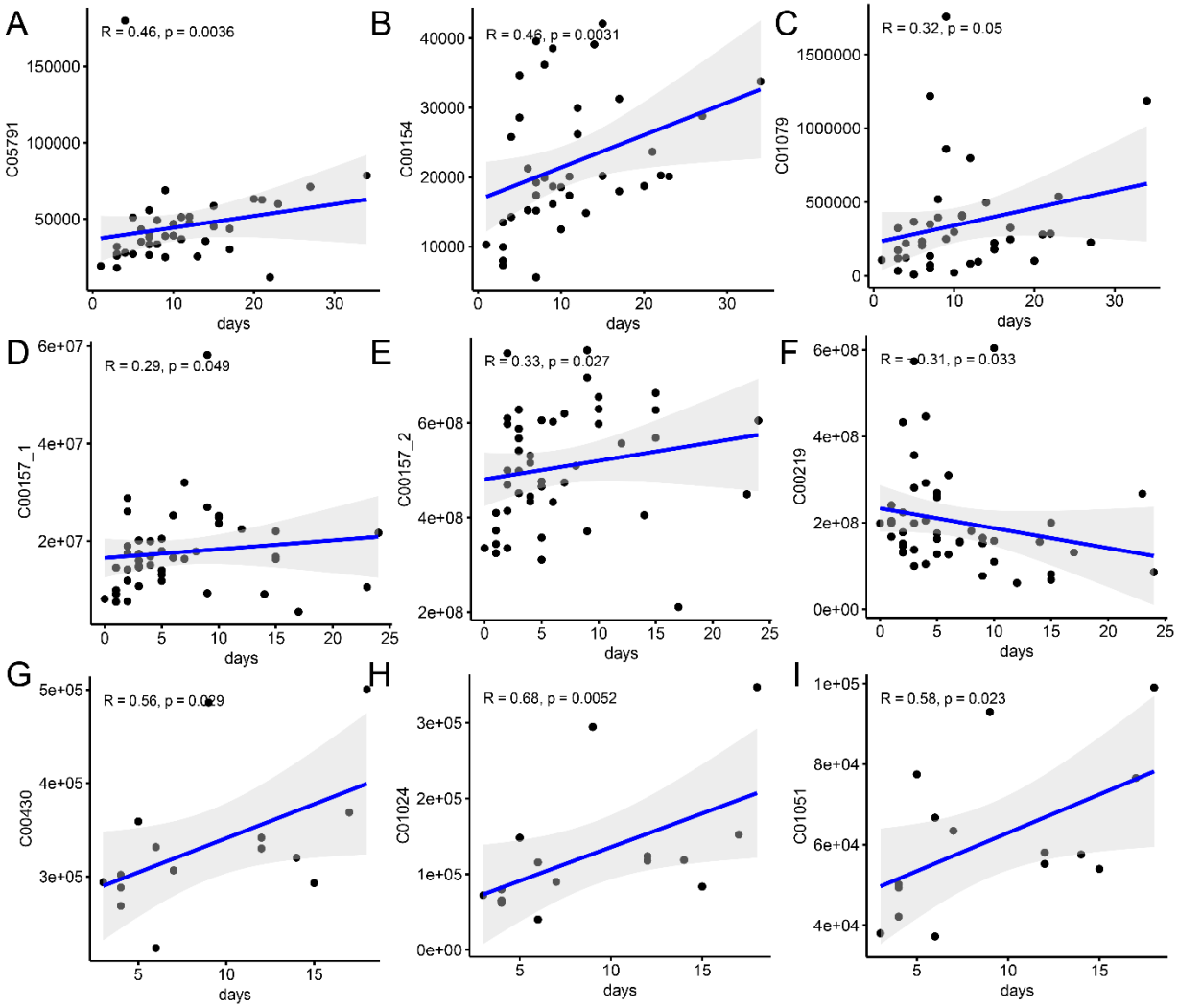


Figure S5.1. The spearman correlation analysis on the onset time (days) with the metabolites in the significantly perturbed pathways. (A). C05791, D-Urobilinogen. (B). C00154, Palmitoyl-CoA. (C). C01079, Protoporphyrinogen IX. (D). C00157, PC (16:0/16:1(9Z)). (E). C00157, PC (16:0/18:0). (F). C00219, Arachidonic acid. (G). C00430, 5-Aminolevulinate. (H). C01024, Hydroxymethylbilane. (I). C01051, Uroporphyrinogen III.

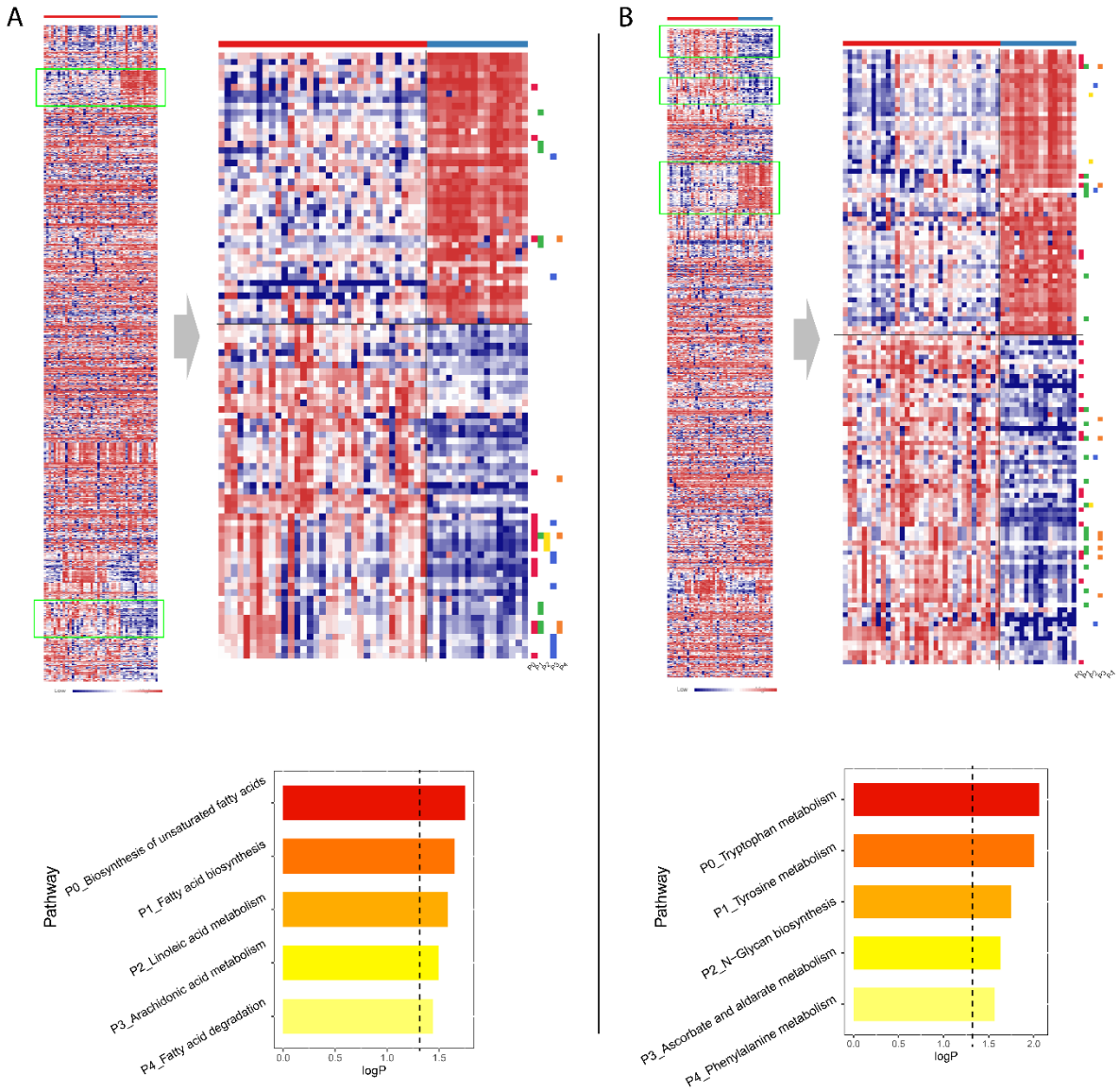


Figure S5.2. Cluster heatmap analysis between Covid and HC groups of Dataset A1 in negative (**A**) and positive mode (**B**).

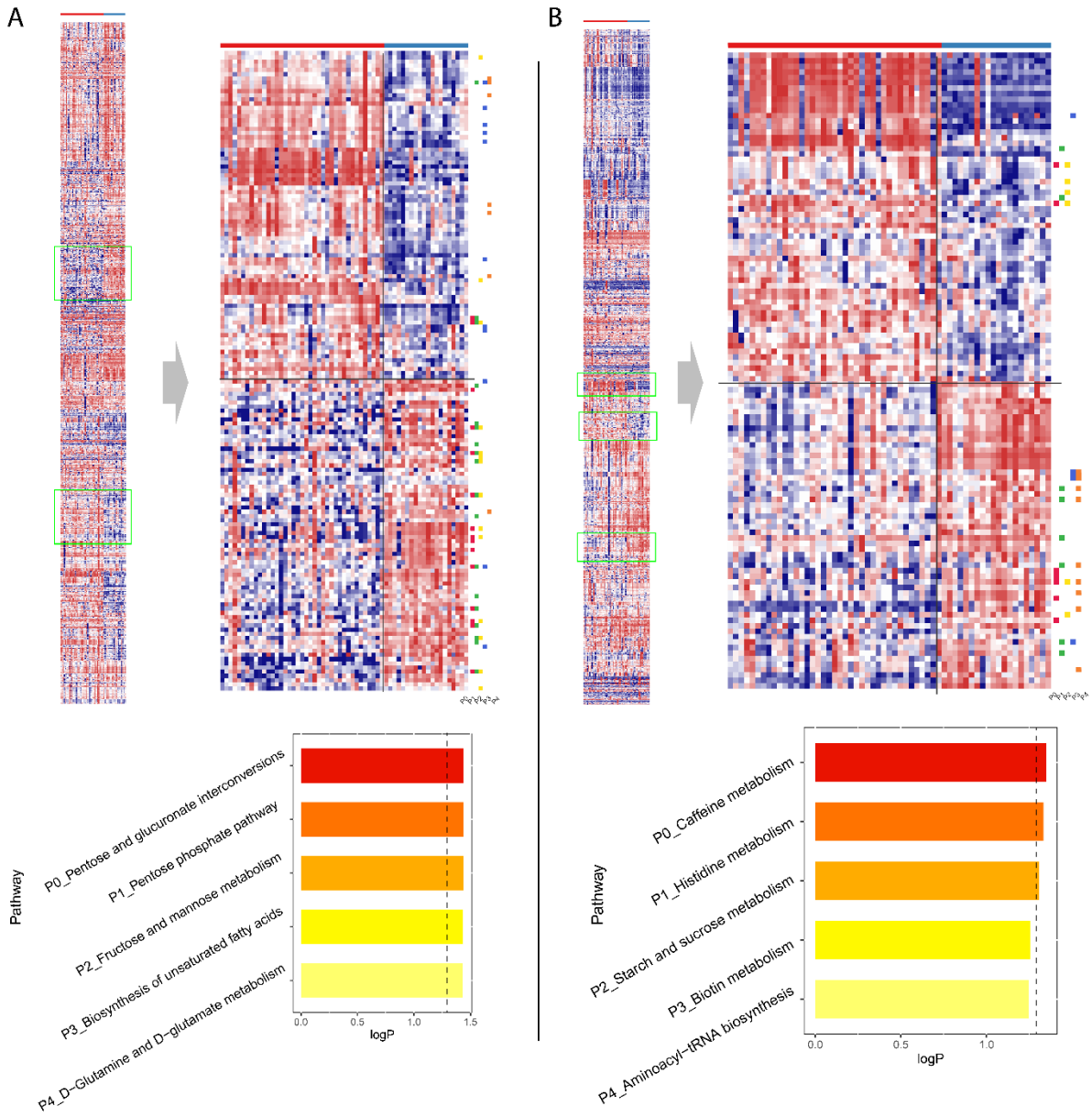


Figure S5.3. Cluster heatmap analysis between Covid and HC groups of Dataset A2 in negative (**A**) and positive mode (**B**).

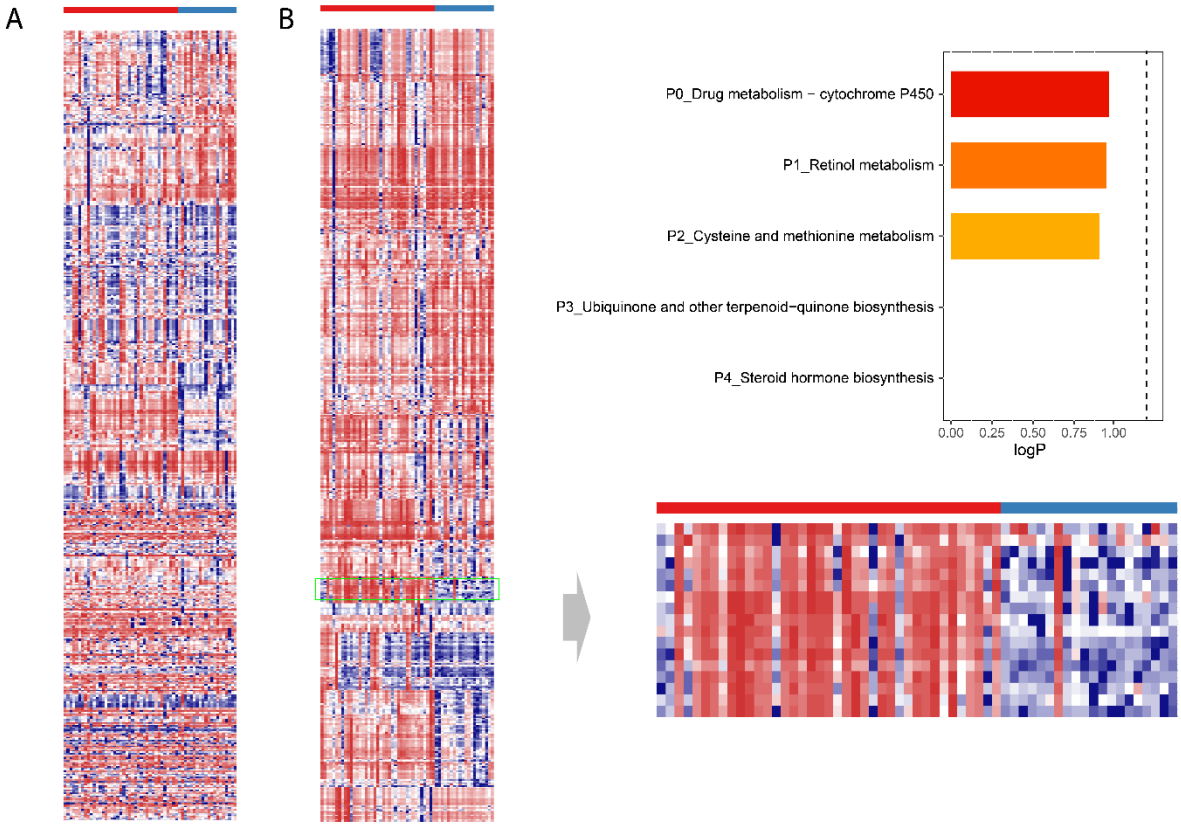


Figure S5.4. Cluster heatmap analysis between Covid and HC groups of Dataset A3 in negative (**A**) and positive mode (**B**).

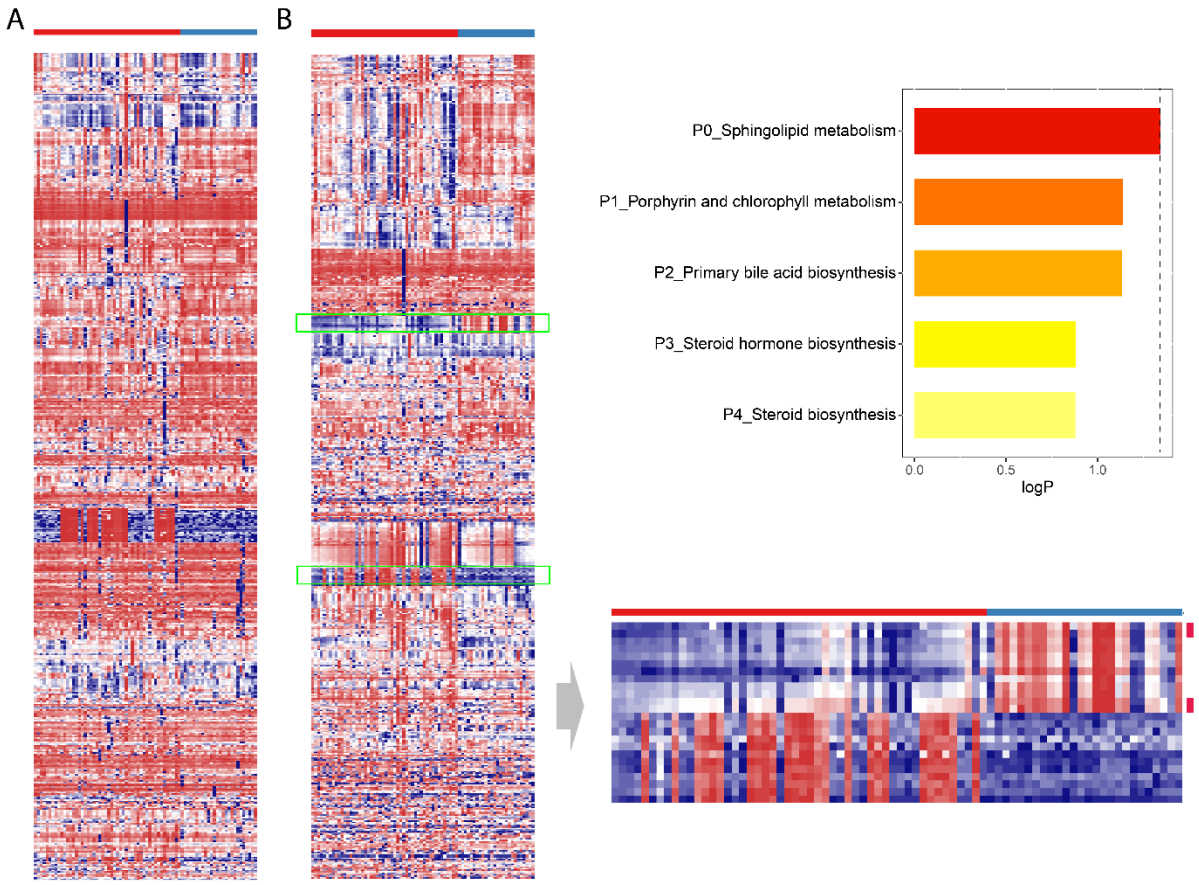


Figure S5.5. Cluster heatmap analysis between Covid and HC groups of Dataset C1 in negative (**A**) and positive mode (**B**).

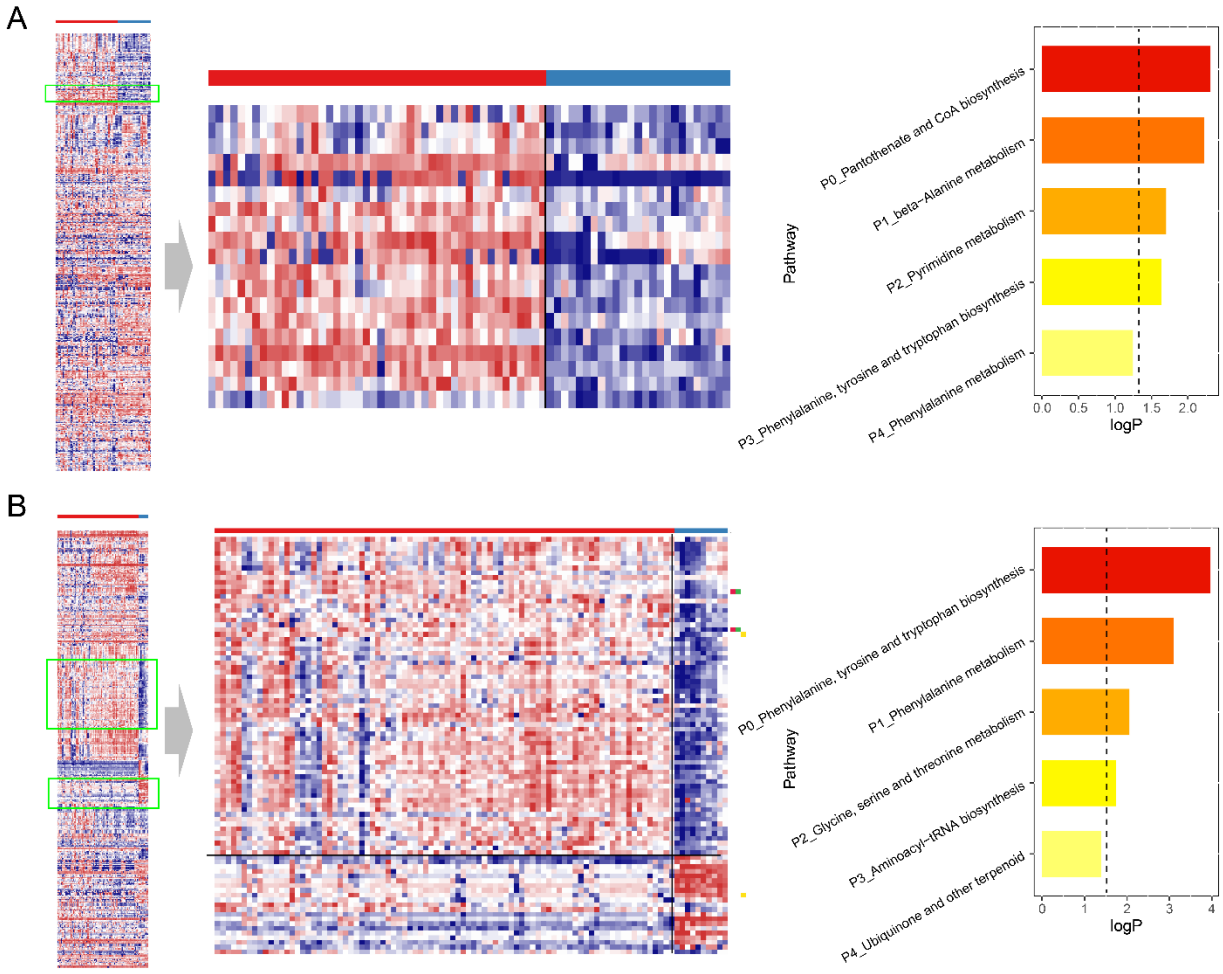


Figure S5.6. Cluster heatmap analysis between Covid and HC groups of Dataset C2 (**A**) and C3 (**B**).

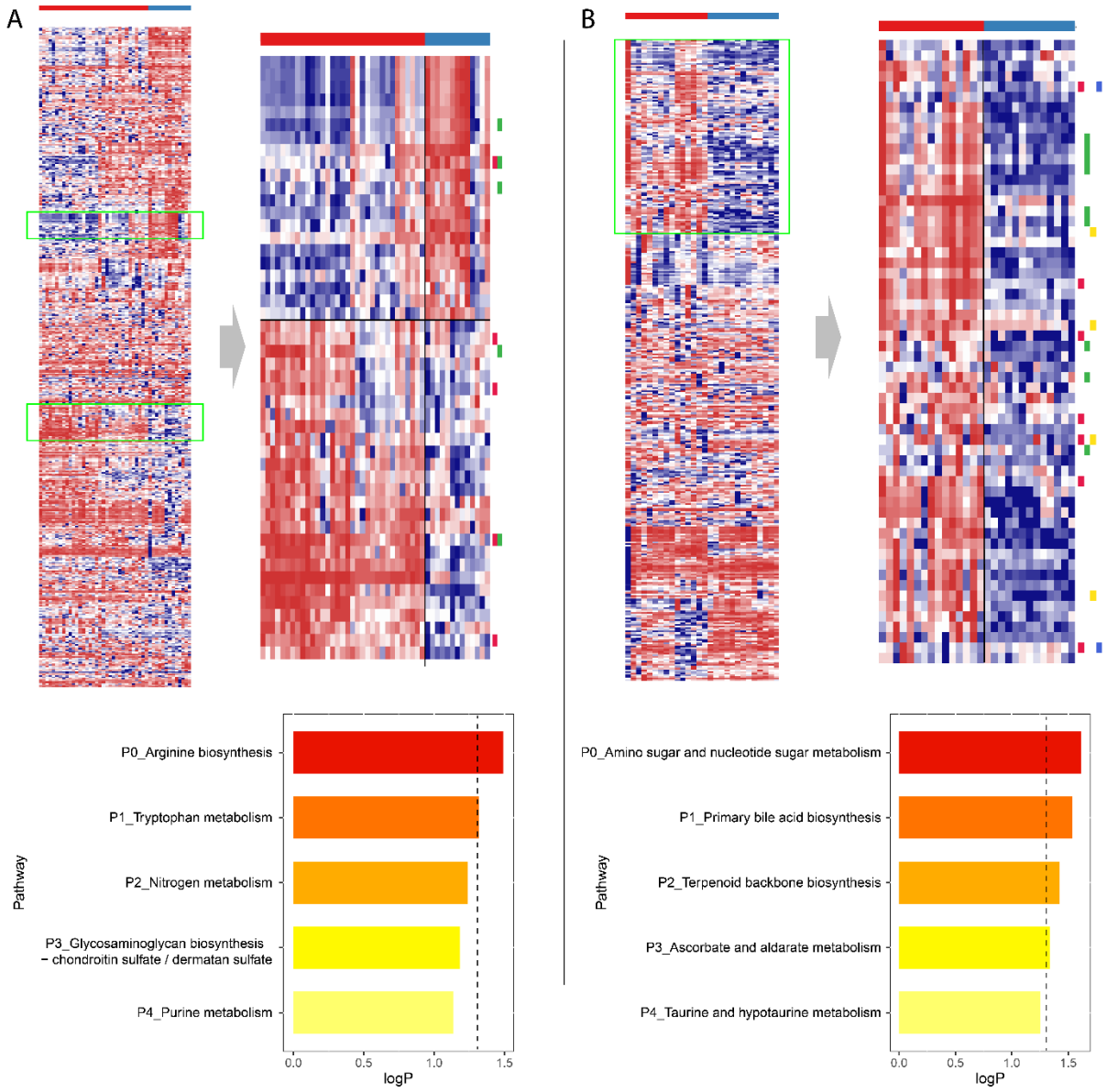


Figure S5.7. Cluster heatmap analysis between Covid and HC groups of Dataset B1 in negative (**A**) and positive mode (**B**).

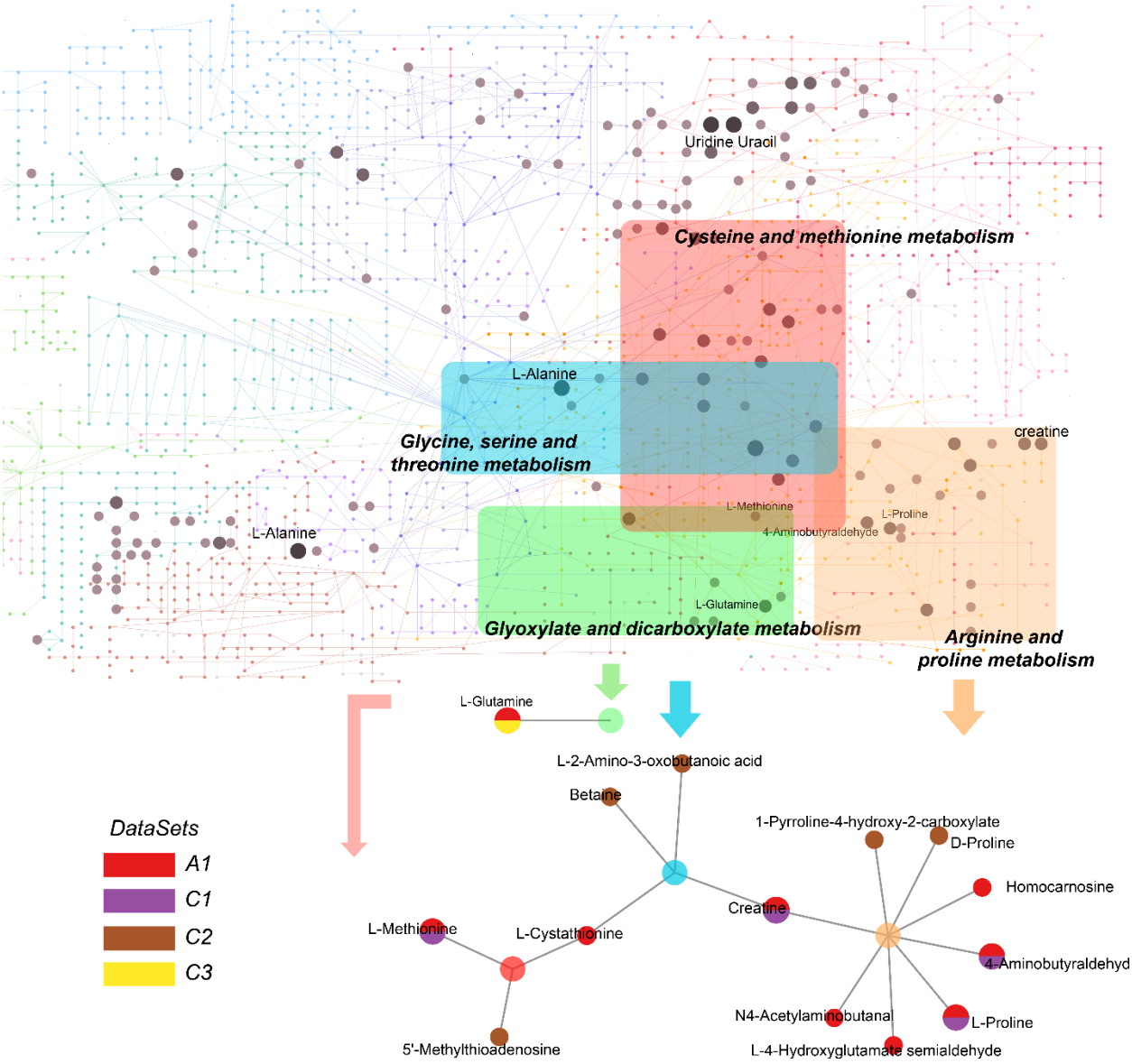


Figure S5.8. Overview of perturbed pathways in COVID-19 across datasets for comparison between MM (Mild-to-moderate) and Severe COVID-19. Every dataset is marked as a specific color. Multiple hits from different datasets on the same metabolite is shown a pie chart with each part colored with the same color code.

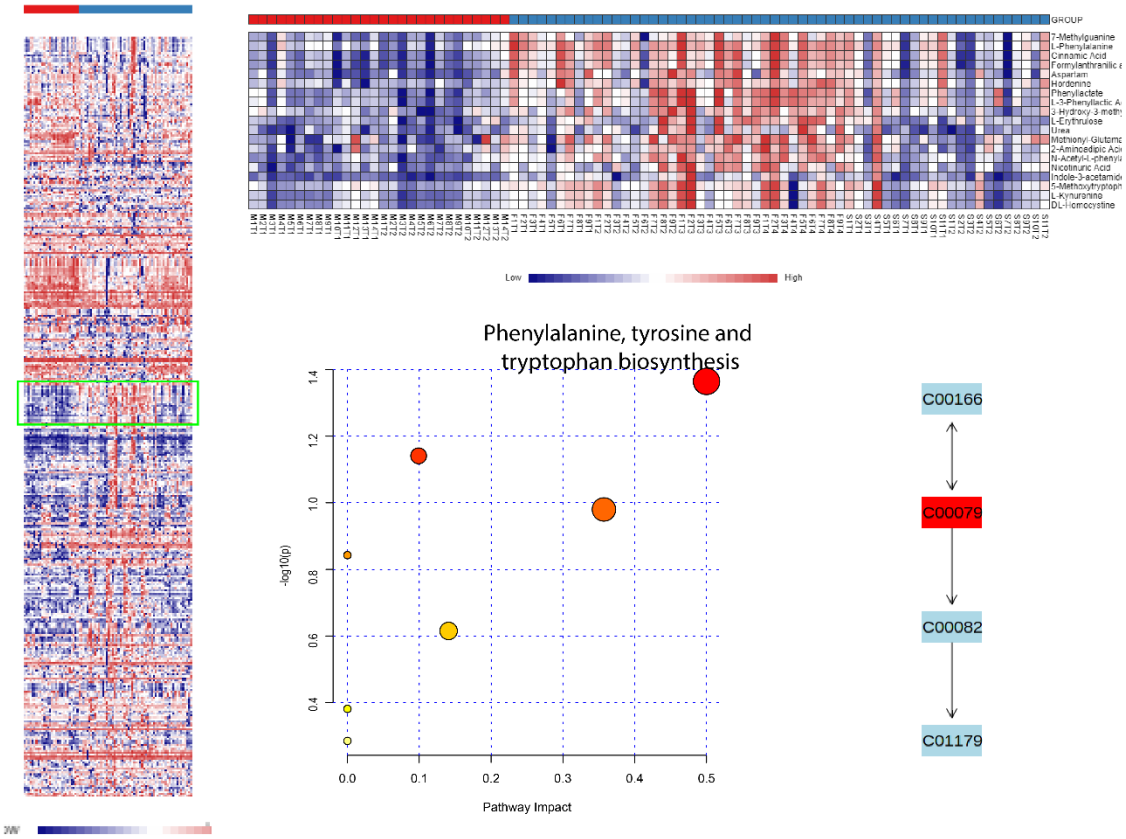


Figure S5.9. The metabolic pattern between MM and Severe of dataset C3. A consistent decreased metabolic pattern was observed and enriched as only one significantly perturbed pathway (Phenylalanine, tyrosine and tryptophan biosynthesis, containing 1 changed metabolite).

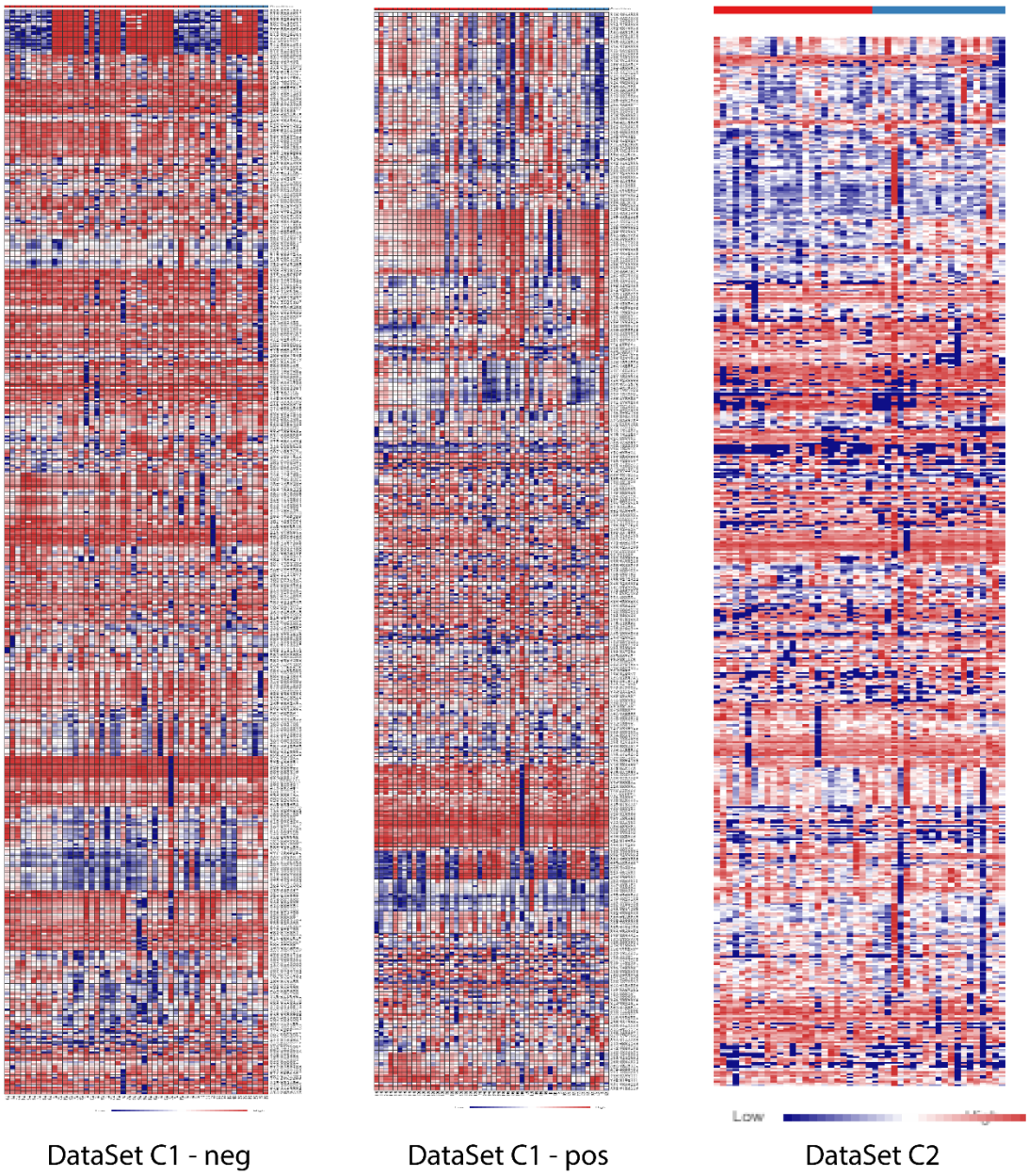


Figure S5.10. The metabolic pattern between MM and Severe of dataset C1 and C2. There is no consistent metabolic pattern cluster found in these datasets.

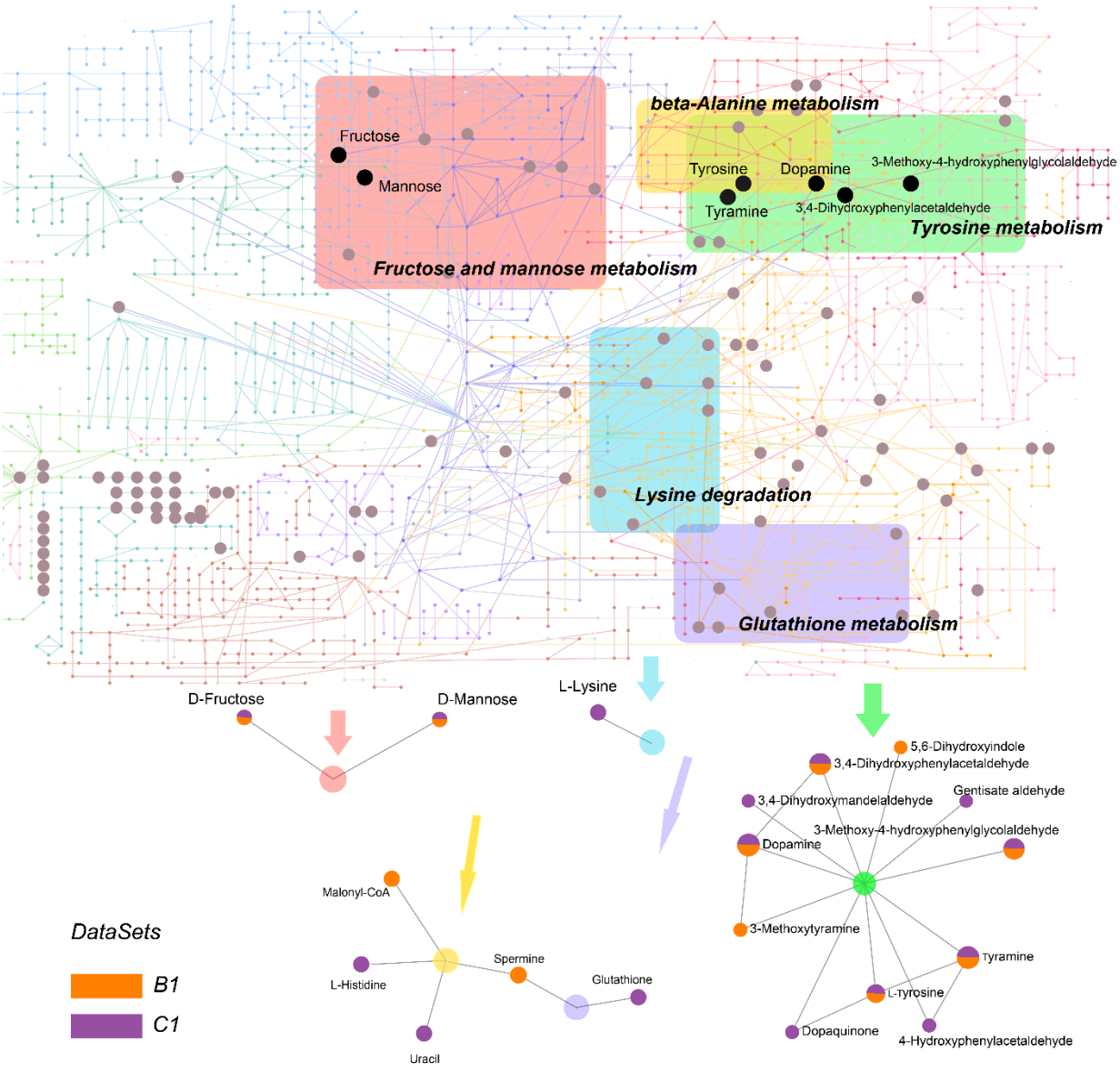


Figure S5.11. Overview of perturbed pathways in COVID-19 across datasets for comparison between Severe and Fatal COVID-19. Each dataset is marked as a specific color. Multiple hits from different datasets on the same metabolite is shown a pie chart with each part colored with the same color code.

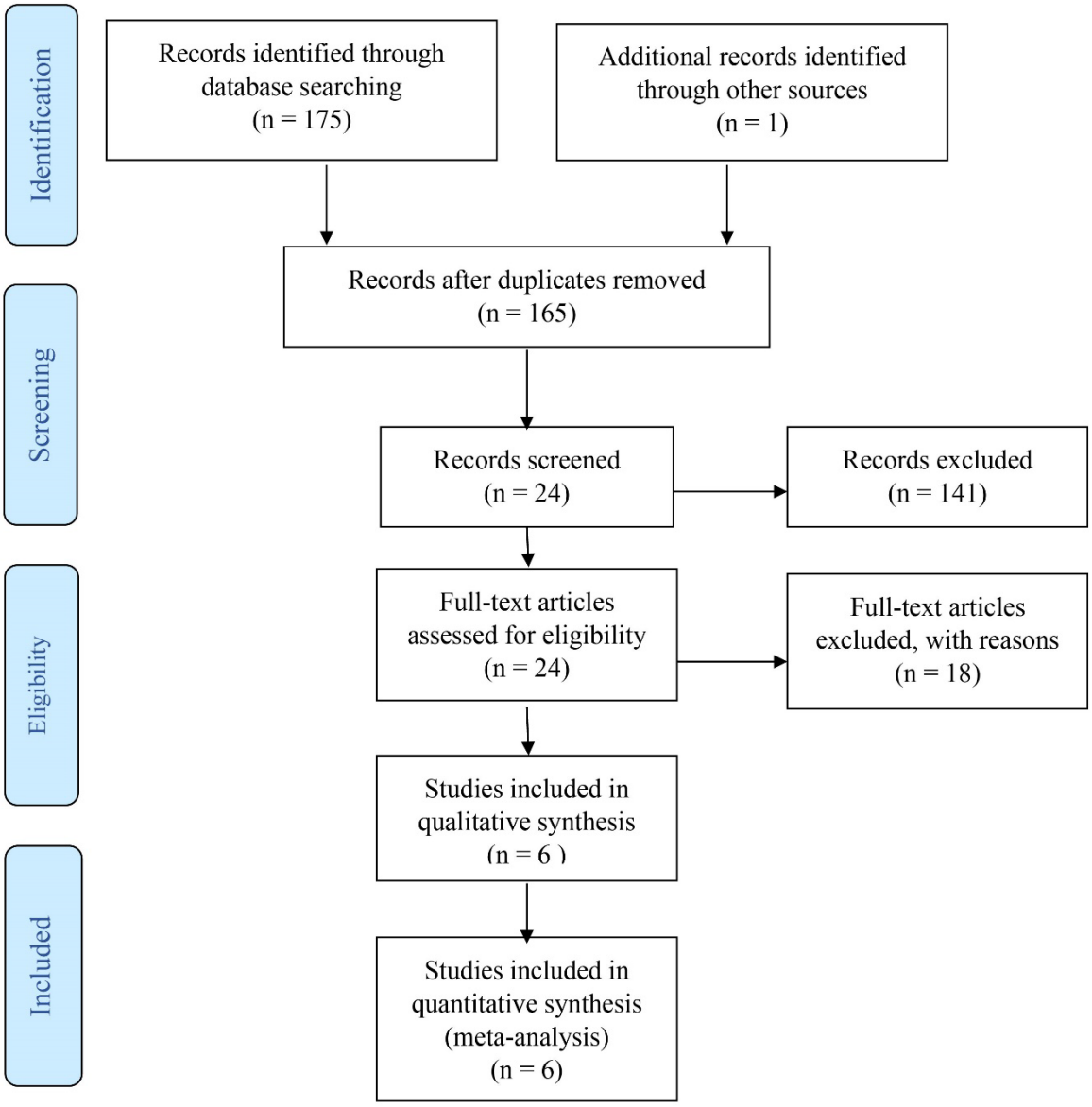


Figure S5.12. The PRISMA Flow Diagram

Author Contributions: Conceptualization, J.X.; data curation, Z.P.; formal analysis, Z.P., G.Z. and J.C.; funding acquisition, J.X.; methodology, Z.P., G.Z., J.C. and J.X.; supervision, J.X.; writing, original draft, Z.P.; review and editing, J.X. and J.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Genome Canada, Génome Québec, US National Institutes of Health (U01 CA235493), Natural Sciences and Engineering Research Council of Canada (NSERC) and Canada Research Chairs (CRC) Program.

Data Availability Statement: The data presented in this study are openly available from this link: https://drive.google.com/drive/folders/1R_I_gu5D3SkD_9q_J93HOA9GuKxZiGNG.

Acknowledgments: The authors truly appreciate the support from original authors Angelo D'Alessandro and Guanghou Shui for providing the raw spectra datasets.

Conflicts of Interest: The authors declare no conflict of interest.

Chapter 6: General Discussion

6.1 Brief summary of this thesis

The overarching theme woven throughout this thesis and extending beyond its chapters is the pursuit of a highly efficient computational approach to empower LC-MS based metabolomics. The primary objective of this thesis has been to bridge the gap between raw LC-MS spectral data preprocessing and the extraction of functional insights. To realize this goal, four chapters (Chapter 2 to 4) have been presented.

Briefly speaking, Chapter 2 introduced the development of the MetaboAnalystR 3.0 package (102). This package established a streamlined workflow for the automated optimization of raw LC-MS spectral data processing. It incorporated empirical compounds characterized by both *m/z* and RT, enhancing the functional analysis based on *Mummichog*. Building upon Chapter 2, Chapter 3 upgraded the MetaboAnalyst website to version 5.0 (276). This version offers a user-friendly interface, enabling users without programming expertise to process their raw LC-MS spectral data using automated optimization. Furthermore, it enhanced functional analysis by integrating results from multiple datasets at the pathway level, addressing the challenge of heterogeneity in global metabolomics functional meta-analysis. Additionally, a heatmap-based functional analysis feature was introduced to explore specific metabolic patterns resulting from perturbations (276, 277).

In Chapter 4, we advanced to MetaboAnalystR 4.0, which introduced an automated workflow for LC-MS/MS data deconvolution. This version also incorporated a comprehensive collection of spectral reference libraries, significantly enhancing MS/MS-based compound identification capabilities. Furthermore, MS/MS-based compound identifications were seamlessly integrated into the functional analysis workflow, thereby augmenting the accuracy of pathway predictions.

Lastly, Chapter 5 saw the comprehensive meta-analysis of multiple COVID-19 global metabolomics datasets, serving as a practical demonstration of the functional analysis workflows we developed (29).

In summary, this thesis culminates in the creation of the MetaboAnalystR package and the enhancement of the MetaboAnalyst website. Subsequent sections will delve into the conceptual underpinnings of Chapters 2 to 5, the strengths and limitations of this thesis, as well as ongoing work and future directions.

6.2 More discussions on chapters 2-5

Global metabolomics has emerged as the primary tool for comprehensively exploring overall changes in the metabolic profile. Over the past decades, numerous algorithms have been developed to process raw LC-MS metabolomics spectral data. In chapter 2, *centWave* was employed as the fundamental algorithm within this thesis to create an auto-optimized workflow, chosen for its established performance and robustness, as evidenced by numerous studies. IPO (46) is a widely used tool for optimizing parameters in *centWave*, although it does suffer from performance defects. To tackle this, MetaboAnalystR 3.0 addresses the performance issue by extracting a subset of spectra data containing the most abundant MS signals. This approach has greatly expedited the parameter optimization step, leading to optimal outcomes. This forms the nucleus of the MetaboAnalystR 3.0 package. However, it's important to note that only *centWave* was considered in the current phase. For future progress, it's necessary to also consider other popular algorithms and optimization methods. More discussions will be included in the limitations and future works section below. Additionally, in this auto-optimized workflow, the regions of MS spectra used for parameter optimization are automatically extracted. It is also possible to extend this capability to allow users to pre-define and extract specific regions of MS spectra based on their personalized

interests and needs. The functionality to achieve this goal has been established, but further validations and case studies are required.

In Chapter 3, the development of the MetaboAnalyst v5.0 website aimed to furnish a user-friendly interface that streamlines the workflow for processing raw spectra. The primary objective of this chapter was to implement all the functionalities introduced in Chapter 2, with a special emphasis on the auto-optimized raw spectral processing workflow. This chapter introduced the capability for users to upload their raw spectral data, initiate processing by simply clicking website buttons, and subsequently obtain their raw spectral processing results without the need for manual adjustment of parameters for the *centWave* algorithm (an example results page is depicted in Figure 6.1). Similar to the MetaboAnalystR 3.0 package, the MetaboAnalyst website exclusively employs an auto-optimization pipeline for the *centWave* algorithm.

When compared to another widely used online web-based tool for raw spectral processing, XCMS Online, MetaboAnalyst has an advantage in terms of parameter optimization functionalities. However, it's important to note that the version of MetaboAnalyst in this context lacks MS/MS-based compound identification due to the absence of a connecting algorithm and reference libraries within MetaboAnalystR 3.0.

Hence, in Chapter 4, our focus shifted towards refining the workflow for processing LC-MS/MS raw spectral data, resulting in the development of the latest iteration, MetaboAnalystR 4.0. Compound identification in global metabolomics poses challenges when contrasted with targeted metabolomics, which relies on established standards (278). MS/MS-based metabolite identification stands out as the predominant approach, primarily due to its independence from standards. However, MS/MS-based compound identification is characterized by reduced accuracy and increased complexity due to the heterogeneous nature of publicly available MS/MS reference

libraries. Over the past decades, several algorithms have been devised to process MS/MS spectral data. Regrettably, none of these algorithms have exhibited precise compound identification. Consequently, a series of additional compound identification steps must be undertaken prior to leveraging the compound information for subsequent functional interpretation.

In more detail, the issue of contamination is widespread in DDA spectral data and necessitates thorough cleansing for increased result accuracy. Until recently, two algorithms have been developed specifically for purging chimeric DDA spectra: DecoID (190) and MS2Purifier (77). DecoID employs a linear regression of spectral references to eliminate contamination fragments. However, this approach can result in failed deconvolutions due to the absence of spectral candidates for certain contaminants. MS2Purifier, on the other hand, distinguishes between contaminated and true fragments by assessing elution profile similarities. The incorporation of a machine learning model enables the recognition and elimination of contaminated fragments.

In Chapter 4, MetaboAnalystR 4.0 integrates a network-based spectrum prediction model to tackle the issue of chimeric spectra. Unlike MS2Purifier, MetaboAnalystR 4.0 primarily relies on an auto-optimized linear regression model and the spectrum prediction model to eliminate contamination. In terms of algorithmic nature, MetaboAnalystR 4.0 can be complementary with MS2Purifier, which may need more demonstration and comparison in the future.

Raw LC-MS spectral processing stands as the foremost critical aspect addressed within this thesis. Nevertheless, functional analysis takes precedence after raw spectral processing in order to glean valuable biological insights. Pathway enrichment analysis has proven precise for targeted metabolomics due to its accurate quantification and compound identification. In global metabolomics, the functional analysis is facilitated through the utilization of the *Mummichog*

algorithm. In Chapter 2, we enhanced the original *Mummichog* algorithm by incorporating *m/z* and RT information, thereby augmenting the accuracy of pathway perturbation prediction.

Moving into Chapter 3 and building upon the functionalities present in MetaboAnalystR 3.0, we have further refined the functional analysis process. This was achieved by introducing metabolic pattern-based functional analysis and conducting functional meta-analysis across multiple datasets. Unlike targeted metabolomics (95, 100), functional analysis in the context of global metabolomics capitalizes on the clustering effect exhibited by perturbed metabolites within their respective pathways. All information of the global metabolomics is included to evaluate the functional perturbation. In this vein, Chapter 3 not only expands the scope of functional analysis of global metabolomics data but also introduces a user-friendly interface meticulously designed to facilitate comprehensive analysis. This is a novel contribution to global metabolomics field.

To further enhance biological exploration and maximize the utilization of information within MS data, Chapter 4 introduces the integration of MS/MS into the functional analysis workflow. The incorporation of MS/MS-based compound identifications serves to filter out impractical empirical compounds used for pathway prediction. This refinement results in a significant enhancement of the discovery of biological insights. At its current stage, MetaboAnalystR 4.0 also accommodates results from other raw spectral processing tools such as MS-DIAL/MS-Finder and SIRIUS, enabling a comprehensive approach to functional analysis. As the platform is still under review, additional functionalities will be incorporated to facilitate connections with other widely used tools.

Since the emergence of the COVID-19 outbreaks in late 2019, understanding the pathogenesis of this severe respiratory disease has emerged as a paramount concern for both clinicians and researchers. Simultaneously, numerous metabolomics studies have been conducted to identify perturbations in metabolic profiles. However, the heterogeneity of metabolomics platforms and

sample processing has led to a considerable degree of variability and inconsistency in results for certain cases. Against this backdrop, Chapter 5 undertakes a comprehensive meta-analysis. Its primary objectives are to assess the performance of functional analysis algorithms and, concurrently, to offer a comprehensive overview based on evidence gleaned from multiple metabolomics datasets. Consequently, the metabolic pathways we elucidated through functional meta-analysis have demonstrated consistency with other clinical researches, as discussed in Chapter 5. This study not only validates effectiveness but also strengthens the comprehension of the pathogenesis from a metabolic perspective.

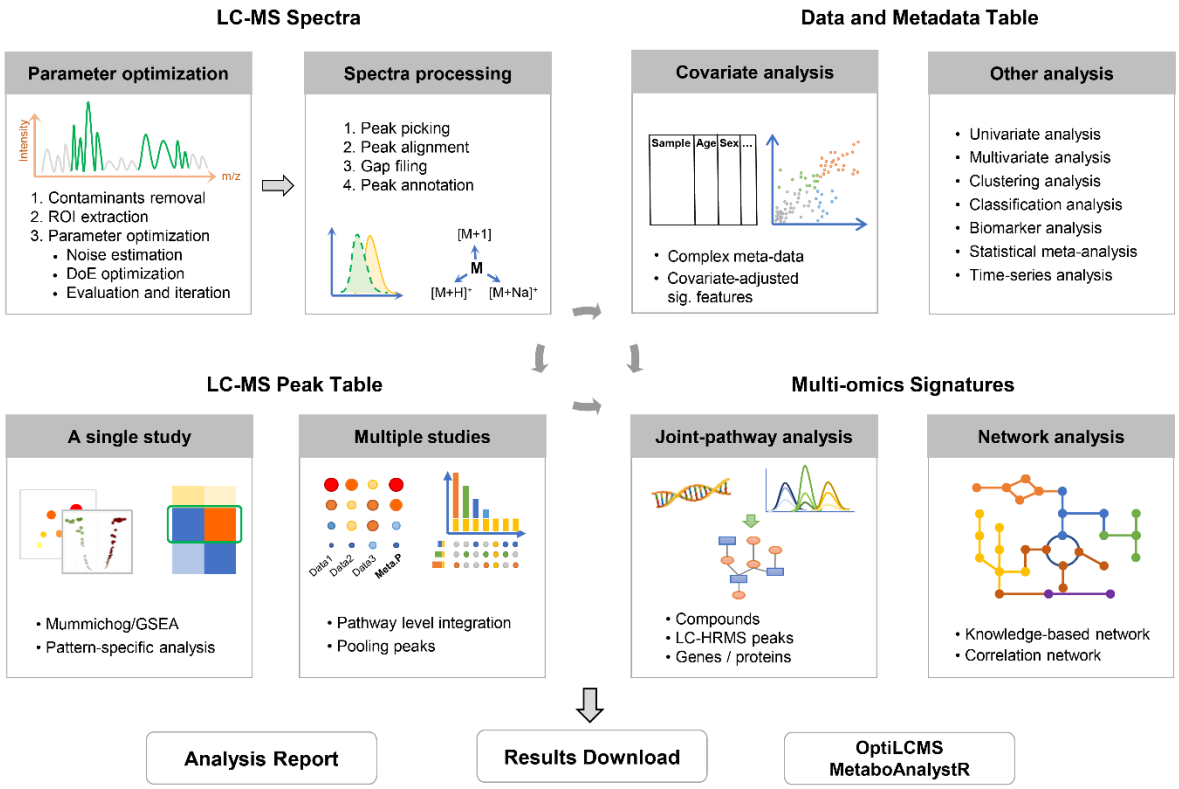


Figure 6.1. Overview of MetaboAnalyst and MetaboAnalystR. MetaboAnalyst and MetaboAnalystR focuses on comprehensive support for LC–MS-based global metabolomics

including spectral processing, functional interpretation, statistical analysis with complex metadata, and multi-omics integration. ROI, regions of interest; DoE, design of experiments.

Finally, both MetaboAnalyst and MetaboAnalystR stand as comprehensive toolkits equipped with functionalities tailored for both targeted and untargeted metabolomics. In its earlier versions (v1.0-v3.0) (154, 155, 157), the MetaboAnalyst website primarily focused on statistical and functional analyses of targeted metabolomics. In the initial iterations of MetaboAnalystR (v.1.0-v.2.0) (105, 122), its primary role was to reproduce website results. However, with the release of MetaboAnalyst version 4.0 and MetaboAnalystR version 3.0, a growing array of features dedicated to processing untargeted metabolomics data has been introduced. These functionalities constitute the core elements of this thesis.

Encompassing statistical analysis, integration of multiple omics data, raw LC-MS spectral processing, and functional analysis utilities, both MetaboAnalyst and MetaboAnalystR have evolved into comprehensive toolkits offering all-encompassing functionalities for the analysis of global metabolomics data, as shown in Figure 6.1 (277).

6.3 Strengths

MetaboAnalyst and MetaboAnalystR stand out as the most comprehensive toolkits, as described earlier. In addition to these attributes, there are several key features that contribute to the superior performance of MetaboAnalyst and MetaboAnalystR.

Firstly, MetaboAnalystR has established an auto-optimized workflow for both MS and MS/MS data processing. This aspect is vital for achieving optimal outcomes without the need for manual

parameter adjustments. Nonetheless, advanced users also have the option to customize parameters manually.

Secondly, the parameters' optimization process is both automatic and exceptionally rapid. Optimizing parameters for MS1 data processing typically takes minutes to a maximum of two hours, even on a standard laptop. In contrast, other automatic parameter optimization tools, like IPO, might take days or even weeks for parameter refinement. The linear regression model employed for DDA MS/MS data deconvolution is also swift and automatic. The algorithm incorporates a penalized elastic net model that achieves real-time optimization of critical parameters within milliseconds, ensuring optimal deconvolution outcomes without manual intervention.

Thirdly, diverse MS/MS reference libraries have been curated to cater to various studies or sample types, such as biological, lipidomics, and exposomics studies. MetaboAnalystR supports searches within these MS/MS reference libraries. These libraries have been assembled from public repositories to ensure comprehensive metabolome coverage. Importantly, the entries within these libraries are meticulously categorized based on MS instrumental types and adduct information, further enhancing accuracy and coverage.

Fourthly, functional analysis of global metabolomics data has undergone significant enhancement through the integration of RT and MS/MS candidates with *m/z* values, resulting in improved accuracy. Standards-based and labor-intensive compound identification procedures can now be circumvented directly. Moreover, functionalities established in this thesis has empowered users to engage in meta-analysis of metabolomics datasets at the pathway level, bolstering result confidence through multiple similar or complementary metabolomics studies. This represents a substantial advancement. Furthermore, pattern-based metabolic pathway perturbation prediction

empowers users to conduct precision exploration of distinct feature clustering patterns within the entire metabolome.

Lastly, a series of user-friendly interfaces (UIs) have been developed to facilitate these functionalities, encompassing auto-optimized MS processing and comprehensive functional analysis of metabolomics datasets, among others. These UIs enable non-programming users to conveniently upload and process their data. Crucially, these interfaces empower users to intuitively and interactively explore their data for biological insights. Reproducing results from the MetaboAnalyst website using the local R package is also seamless.

Besides, as a comprehensive toolkit, MetaboAnalyst and MetaboAnalystR have been regularly updated and maintained in response to user feedback. When compared to the previous versions and other popular tools in the field, the latest iterations of both MetaboAnalyst and MetaboAnalystR boast the most stable and comprehensive functionalities for processing global metabolomics data.

6.4 Limitations

While MetaboAnalyst and MetaboAnalystR possess distinctive features, certain limitations persist. For the MetaboAnalyst website, a restriction emerges due to proprietary formats and the substantial file sizes produced by LC-MS instruments. Presently, MetaboAnalyst does not support the upload of raw spectra in vendor-specific formats. Consequently, raw data from various MS instruments necessitates preliminary conversion into an open data format using either a vendor-provided conversion tool or a free alternative such as ProteoWizard.

Furthermore, the public iteration of MetaboAnalyst currently imposes a cap on the processing of raw spectra, limiting it to a maximum of 200 samples per job. This limitation, in our experience,

typically accommodates common metabolomic studies efficiently. For expansive projects, users are encouraged to conduct spectra processing locally utilizing MetaboAnalystR. Neither MetaboAnalyst nor MetaboAnalystR currently facilitate the processing of spectra from GC-MS or ion mobility spectrometry, which are commonly employed within global metabolomics.

Moreover, flow-injection and direct injection mass spectrometry have gained traction in profiling the metabolic or ionic composition of samples, particularly for large-scale sample assays (27, 279). By bypassing chromatography-based separation and directly injecting samples into the MS instrument, data acquisition is expedited and batch effects are minimized. However, neither MetaboAnalyst nor MetaboAnalystR support the processing of flow/direct injection data, constituting an additional limitation.

As of now, only the MetaboAnalystR package has an implemented MS/MS data processing algorithm. A UI-based platform is not currently available, necessitating the use of the R/RStudio console for MS/MS data analysis. The MS/MS data processing pipeline within MetaboAnalystR 4.0 has exclusively undergone benchmarking against a selection of popular data processing tools, each with their comprehensive raw data processing workflows. Other contamination removal solutions, such as MS2Purifier, require further evaluation and comparison.

Functional analysis and integration in MetaboAnalyst mainly focus on biological samples, while environmental and industrial samples are not well supported owing to lack of well-established conceptual frameworks and knowledgebases required for these types of analysis.

6.5 Future directions

To address the aforementioned limitations, our forthcoming efforts will center on the development of additional algorithms to facilitate data processing for flow/direct injection and ion mobility MS

data. This could further enhance the raw data processing ability of MetaboAnalystR package. Furthermore, plans are in place to seamlessly integrate MS/MS data deconvolution and reference library search functionalities into the MetaboAnalyst website with user-friendly UIs, ensuring their availability in the subsequent release.

In parallel with the expansion of functionalities, an essential aspect involves conducting further case studies and benchmark evaluations. This will enable comprehensive comparisons and assessments of the performance exhibited by newly introduced MS data processing pipelines, such as *asari* (205). Additionally, for MS/MS DDA data deconvolution, conducting extensive comparisons with other robust tools like MS2Purifier will be crucial. Through such comparisons, we can further refine and enhance performance by harnessing the potential of machine-learning-based models.

Furthermore, an imperative aspect involves the undertaking of additional biological studies to further assess the efficacy of functional analysis. Within this thesis, we conducted a meta-analysis on seven COVID-19 metabolomics datasets. Since the emergence of COVID-19, the landscape boasts a proliferation of over a hundred metabolomics publications. This surge in data prompts the necessity for a more expansive meta-analysis across these datasets. Such an undertaking holds a two-fold objective: firstly, to comprehensively evaluate and refine the algorithm's performance; and secondly, to enhance the understanding of COVID-19's pathogenesis from a metabolic perspective.

Chapter 7: Conclusions and Future works

MetaboAnalystR 3.0 in Chapter 2 has created an auto-optimized LC-MS raw data processing workflow, which could process raw LC-MS data to obtain optimal results automatically in a highly-efficient way. The pipeline has been implemented into MetaboAnalyst website (v5.0, Chapter 3) to offer an intuitive and easy-to-use interface for users. Data deconvolution functionalities created in MetaboAnalystR 4.0 (in Chapter 4) has also enabled the processing of LC-MS/MS raw spectral data (for DDA and SWATH-DIA) in a high-precision and ultra-fast approach. Multiple comprehensive reference MS/MS library options further allow users mining the chemical candidates based on MS/MS.

Function analysis of global metabolomics dataset has also been updated by integrating RT and MS/MS identification information in MetaboAnalystR 3.0 and 4.0, respectively. A user-friendly interface has enabled users to perform function analysis either based on the complete metabolomics dataset or a pattern of metabolomics features in MetaboAnalyst website. Besides, the implementation of functional meta-analysis in the website has further empowered the biological interpretation from multiple metabolomics datasets. The efficacy of functional analysis and functional meta-analysis has been demonstrated by a COVID-19 study (in Chapter 5).

Other utilities for metabolomics data processing, including batch effect correction and multi-omics integrative analysis has also been offered, together with the main features above.

In summary, MetaboAnalystR and MetaboAnalyst offers a comprehensive and powerful toolkit to bridge LC-MS and LC-MS/MS raw spectral processing to accurate functional analysis.

Moving forward, both MetaboAnalystR and MetaboAnalyst does not support processing ion mobility spectral data, direct injection and flow-injection spectral data at current stage. The related

functionalities will be achieved in the future release. Besides, LC-MS/MS raw spectral processing is only available from MetaboAnalystR package. A user-friendly interface for LC-MS/MS would be released in the next version MetaboAnalyst website (v6.0).

References:

1. Karahalil B. Overview of Systems Biology and Omics Technologies. Curr Med Chem. 2016;23(37):4221-30.
2. Tyers M, Mann M. From genomics to proteomics. Nature. 2003;422(6928):193-7.
3. Weckwerth W. Metabolomics in systems biology. Annu Rev Plant Biol. 2003;54:669-89.
4. Hanash S. Disease proteomics. Nature. 2003;422(6928):226-32.
5. Al-Amrani S, Al-Jabri Z, Al-Zaabi A, Alshekaili J, Al-Khabori M. Proteomics: Concepts and applications in human medicine. World J Biol Chem. 2021;12(5):57-69.
6. Garza DR, van Verk MC, Huynen MA, Dutilh BE. Towards predicting the environmental metabolome from metagenomics with a mechanistic model. Nature Microbiology. 2018;3(4):456-60.
7. Wishart DS, Tzur D, Knox C, Eisner R, Guo AC, Young N, et al. HMDB: the Human Metabolome Database. Nucleic Acids Res. 2007;35(Database issue):D521-6.
8. Krug S, Kastenmüller G, Stückler F, Rist MJ, Skurk T, Sailer M, et al. The dynamic range of the human metabolome revealed by challenges. The FASEB Journal. 2012;26(6):2607-19.
9. Wishart DS, Mandal R, Stanislaus A, Ramirez-Gaona M. Cancer Metabolomics and the Human Metabolome Database. Metabolites. 2016;6(1):10.
10. Saito K, Matsuda F. Metabolomics for Functional Genomics, Systems Biology, and Biotechnology. Annual Review of Plant Biology. 2010;61(1):463-89.
11. Pang Z, Wang G, Wang C, Zhang W, Liu J, Wang F. Serum Metabolomics Analysis of Asthma in Different Inflammatory Phenotypes: A Cross-Sectional Study in Northeast China. BioMed Research International. 2018;2018:2860521.
12. García-Sevillano MÁ, García-Barrera T, Gómez-Ariza JL. Environmental metabolomics: Biological markers for metal toxicity. ELECTROPHORESIS. 2015;36(18):2348-65.
13. Bundy JG, Davey MP, Viant MR. Environmental metabolomics: a critical review and future perspectives. Metabolomics. 2009;5(1):3-21.
14. Klassen A, Faccio AT, Canuto GAB, da Cruz PLR, Ribeiro HC, Tavares MFM, et al. Metabolomics: Definitions and Significance in Systems Biology. In: Sussulini A, editor. Metabolomics: From Fundamentals to Clinical Applications. Cham: Springer International Publishing; 2017. p. 3-17.
15. Vinayavekhin N, Saghatelian A. Untargeted metabolomics. Curr Protoc Mol Biol. 2010;Chapter 30:Unit 30 1 1-24.
16. Roberts LD, Souza AL, Gerszten RE, Clish CB. Targeted metabolomics. Curr Protoc Mol Biol. 2012;Chapter 30:Unit 30 2 1-24.
17. Pang Z, Wang G, Ran N, Lin H, Wang Z, Guan X, et al. Inhibitory Effect of Methotrexate on Rheumatoid Arthritis Inflammation and Comprehensive Metabolomics Analysis Using Ultra-Performance Liquid Chromatography-Quadrupole Time of Flight-Mass Spectrometry (UPLC-Q/TOF-MS). Int J Mol Sci. 2018;19(10).
18. Shan L, Yang J, Meng S, Ruan H, Zhou L, Ye F, et al. Urine Metabolomics Profiling of Lumbar Disc Herniation and its Traditional Chinese Medicine Subtypes in Patients Through Gas Chromatography Coupled With Mass Spectrometry. Front Mol Biosci. 2021;8:648823.

19. Zhang ZM, Chen MJ, Zou JF, Jiang S, Shang EX, Qian DW, et al. UPLC-Q-TOF/MS based fecal metabolomics reveals the potential anti-diabetic effect of Xiexin Decoction on T2DM rats. J Chromatogr B Analyt Technol Biomed Life Sci. 2021;1173:122683.
20. Cameron SJ, Lewis KE, Beckmann M, Allison GG, Ghosal R, Lewis PD, et al. The metabolomic detection of lung cancer biomarkers in sputum. Lung Cancer. 2016;94:88-95.
21. Zukunft S, Prehn C, Rohring C, Moller G, Hrabe de Angelis M, Adamski J, et al. High-throughput extraction and quantification method for targeted metabolomics in murine tissues. Metabolomics. 2018;14(1):18.
22. Diaz-Cruz MS, Lopez de Alda MJ, Lopez R, Barcelo D. Determination of estrogens and progestogens by mass spectrometric techniques (GC/MS, LC/MS and LC/MS/MS). J Mass Spectrom. 2003;38(9):917-23.
23. Nagana Gowda GA, Raftery D. Analysis of Plasma, Serum, and Whole Blood Metabolites Using (1)H NMR Spectroscopy. Methods Mol Biol. 2019;2037:17-34.
24. Tsizin S, Bokka R, Keshet U, Alon T, Fialkov AB, Tal N, et al. Comparison of electrospray LC–MS, LC–MS with Cold EI and GC–MS with Cold EI for sample identification. International Journal of Mass Spectrometry. 2017;422:119-25.
25. Zhou B, Xiao JF, Tuli L, Ressom HW. LC-MS-based metabolomics. Mol Biosyst. 2012;8(2):470-81.
26. Becker S, Kortz L, Helmschrodt C, Thiery J, Ceglarek U. LC-MS-based metabolomics in the clinical laboratory. J Chromatogr B Analyt Technol Biomed Life Sci. 2012;883-884:68-75.
27. Sarvin B, Lagziel S, Sarvin N, Mukha D, Kumar P, Aizenshtein E, et al. Fast and sensitive flow-injection mass spectrometry metabolomics by analyzing sample-specific ion distributions. Nat Commun. 2020;11(1):3186.
28. Nassan FL, Kelly RS, Kosheleva A, Koutrakis P, Vokonas PS, Lasky-Su JA, et al. Metabolomic signatures of the long-term exposure to air pollution and temperature. Environ Health. 2021;20(1):3.
29. Pang Z, Zhou G, Chong J, Xia J. Comprehensive Meta-Analysis of COVID-19 Global Metabolomics Datasets. Metabolites. 2021;11(1):44.
30. Wu P, Chen D, Ding W, Wu P, Hou H, Bai Y, et al. The trans-omics landscape of COVID-19. Nat Commun. 2021;12(1):4543.
31. Romero R, Espinoza J, Gotsch F, Kusanovic J, Friel L, Erez O, et al. The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. BJOG: An International Journal of Obstetrics & Gynaecology. 2006;113(s3):118-35.
32. Blaženović I, Kind T, Ji J, Fiehn O. Software Tools and Approaches for Compound Identification of LC-MS/MS Data in Metabolomics. Metabolites. 2018;8(2):31.
33. Melamud E, Vastag L, Rabinowitz JD. Metabolomic analysis and visualization engine for LC-MS data. Anal Chem. 2010;82(23):9818-26.
34. Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G. XCMS Online: a web-based platform to process untargeted metabolomic data. Anal Chem. 2012;84(11):5035-9.
35. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. Anal Chem. 2006;78(3):779-87.

36. Pluskal T, Castillo S, Villar-Briones A, Oresic M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. BMC Bioinformatics. 2010;11:395.
37. Katajamaa M, Miettinen J, Oresic M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. Bioinformatics. 2006;22(5):634-6.
38. Schmid R, Heuckeroth S, Korf A, Smirnov A, Myers O, Dyrlund TS, et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. Nature Biotechnology. 2023.
39. Tsugawa H, Ikeda K, Takahashi M, Satoh A, Mori Y, Uchino H, et al. A lipidome atlas in MS-DIAL 4. Nature Biotechnology. 2020;38(10):1159-63.
40. Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. Nature Methods. 2015;12(6):523-6.
41. Rost HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016;13(9):741-8.
42. Yu T, Park Y, Johnson JM, Jones DP. apLCMS—adaptive processing of high-resolution LC/MS data. Bioinformatics. 2009;25(15):1930-6.
43. Deutsch EW. Mass Spectrometer Output File Format mzML. In: Hubbard SJ, Jones AR, editors. Proteome Bioinformatics. Totowa, NJ: Humana Press; 2010. p. 319-31.
44. Lai Z, Tsugawa H, Wohlgemuth G, Mehta S, Mueller M, Zheng Y, et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. Nat Methods. 2018;15(1):53-6.
45. Tautenhahn R, Böttcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. BMC Bioinformatics. 2008;9:504.
46. Libiseller G, Dvorzak M, Kleb U, Gander E, Eisenberg T, Madeo F, et al. IPO: a tool for automated optimization of XCMS parameters. BMC Bioinformatics. 2015;16:118.
47. Myers OD, Sumner SJ, Li S, Barnes S, Du X. Detailed Investigation and Comparison of the XCMS and MZmine 2 Chromatogram Construction and Chromatographic Peak Detection Methods for Preprocessing Mass Spectrometry Metabolomics Data. Anal Chem. 2017;89(17):8689-95.
48. Conley CJ, Smith R, Torgrip RJ, Taylor RM, Tautenhahn R, Prince JT. Massifquant: open-source Kalman filter-based XC-MS isotope trace feature detection. Bioinformatics. 2014;30(18):2636-43.
49. Aberg KM, Torgrip RJ, Kolmert J, Schuppe-Koistinen I, Lindberg J. Feature detection and alignment of hyphenated chromatographic-mass spectrometric data. Extraction of pure ion chromatograms using Kalman tracking. J Chromatogr A. 2008;1192(1):139-46.
50. Tengstrand E, Lindberg J, Aberg KM. TracMass 2--a modular suite of tools for processing chromatography-full scan mass spectrometry data. Anal Chem. 2014;86(7):3435-42.
51. Kenar E, Franken H, Forcisi S, Wörmann K, Häring HU, Lehmann R, et al. Automated label-free quantification of metabolites from liquid chromatography-mass spectrometry data. Mol Cell Proteomics. 2014;13(1):348-59.
52. Myers OD, Sumner SJ, Li S, Barnes S, Du X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry

Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. Anal Chem. 2017;89(17):8696-703.
53. Prince JT, Marcotte EM. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. Anal Chem. 2006;78(17):6140-52.
54. Kuhl C, Tautenhahn R, Böttcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. Anal Chem. 2012;84(1):283-9.
55. Senan O, Aguilar-Mogas A, Navarro M, Capellades J, Noon L, Burks D, et al. CliqueMS: a computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network. Bioinformatics. 2019;35(20):4089-97.
56. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, et al. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. Analyst. 2009;134(7):1322-32.
57. Jankevics A, Merlo ME, de Vries M, Vonk RJ, Takano E, Breitling R. Separating the wheat from the chaff: a prioritisation pipeline for the analysis of metabolomics datasets. Metabolomics. 2012;8(Suppl 1):29-36.
58. Mahieu NG, Patti GJ. Systems-Level Annotation of a Metabolomics Data Set Reduces 25000 Features to Fewer than 1000 Unique Metabolites. Anal Chem. 2017;89(19):10397-406.
59. DeFelice BC, Mehta SS, Samra S, Cajka T, Wancewicz B, Fahrmann JF, et al. Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-MS) Data Processing. Anal Chem. 2017;89(6):3250-5.
60. Tikunov YM, Laptenok S, Hall RD, Bovy A, de Vos RC. MSClust: a tool for unsupervised mass spectra extraction of chromatography-mass spectrometry ion-wise aligned data. Metabolomics. 2012;8(4):714-8.
61. Silva RR, Jourdan F, Salvanha DM, Letisse F, Jamin EL, Guidetti-Gonzalez S, et al. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. Bioinformatics. 2014;30(9):1336-7.
62. Uppal K, Walker DI, Jones DP. xMSannotator: An R Package for Network-Based Annotation of High-Resolution Metabolomics Data. Anal Chem. 2017;89(2):1063-7.
63. Defossez E, Bourquin J, von Reuss S, Rasmann S, Glauser G. Eight key rules for successful data-dependent acquisition in mass spectrometry-based metabolomics. Mass Spectrom Rev. 2023;42(1):131-43.
64. Fenaille F, Barbier Saint-Hilaire P, Rousseau K, Junot C. Data acquisition workflows in liquid chromatography coupled to high resolution mass spectrometry-based metabolomics: Where do we stand? J Chromatogr A. 2017;1526:1-12.
65. Ten-Doménech I, Martínez-Sena T, Moreno-Torres M, Sanjuan-Herráez JD, Castell JV, Parra-Llorca A, et al. Comparing Targeted vs. Untargeted MS(2) Data-Dependent Acquisition for Peak Annotation in LC-MS Metabolomics. Metabolites. 2020;10(4).
66. Sidoli S, Lin S, Xiong L, Bhanu NV, Karch KR, Johansen E, et al. Sequential Window Acquisition of all Theoretical Mass Spectra (SWATH) Analysis for Characterization and

Quantification of Histone Post-translational Modifications. Mol Cell Proteomics. 2015;14(9):2420-8.
67. Alka O, Shanthamoorthy P, Witting M, Kleigrewe K, Kohlbacher O, Röst HL. DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics. Nature Communications. 2022;13(1):1347.
68. Zhao T, Xing S, Yu H, Huan T. De Novo Cleaning of Chimeric MS/MS Spectra for LC-MS/MS-Based Metabolomics. Analytical Chemistry. 2023;95(35):13018-28.
69. Nikolskiy I, Mahieu NG, Chen Y, Jr., Tautenhahn R, Patti GJ. An Untargeted Metabolomic Workflow to Improve Structural Characterization of Metabolites. Analytical Chemistry. 2013;85(16):7713-9.
70. Lawson TN, Weber RJM, Jones MR, Chetwynd AJ, Rodríguez-Blanco G, Di Guida R, et al. msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. Analytical Chemistry. 2017;89(4):2432-9.
71. Yu M, Dolios G, Petrick L. Reproducible untargeted metabolomics workflow for exhaustive MS2 data acquisition of MS1 features. Journal of Cheminformatics. 2022;14(1):6.
72. Tada I, Chaleckis R, Tsugawa H, Meister I, Zhang P, Lazarinis N, et al. Correlation-Based Deconvolution (CorrDec) To Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies. Analytical Chemistry. 2020;92(16):11310-7.
73. Peckner R, Myers SA, Jacome ASV, Egertson JD, Abelin JG, MacCoss MJ, et al. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. Nat Methods. 2018;15(5):371-8.
74. Röst HL, Rosenberger G, Navarro P, Gillet L, Miladinović SM, Schubert OT, et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. Nature Biotechnology. 2014;32(3):219-23.
75. Li H, Cai Y, Guo Y, Chen F, Zhu Z-J. MetDIA: Targeted Metabolite Extraction of Multiplexed MS/MS Spectra Generated by Data-Independent Acquisition. Analytical Chemistry. 2016;88(17):8757-64.
76. Yin Y, Wang R, Cai Y, Wang Z, Zhu Z-J. DecoMetDIA: Deconvolution of Multiplexed MS/MS Spectra for Metabolite Identification in SWATH-MS-Based Untargeted Metabolomics. Analytical Chemistry. 2019;91(18):11897-904.
77. Xing S, Yu H, Liu M, Jia Q, Sun Z, Fang M, et al. Recognizing Contamination Fragment Ions in Liquid Chromatography–Tandem Mass Spectrometry Data. Journal of the American Society for Mass Spectrometry. 2021;32(9):2296-305.
78. Giné R, Capellades J, Badia JM, Vughs D, Schwaiger-Haber M, Alexandrov T, et al. HERMES: a molecular-formula-oriented method to target the metabolome. Nature Methods. 2021;18(11):1370-6.
79. Tsugawa H, Kind T, Nakabayashi R, Yukihira D, Tanaka W, Cajka T, et al. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. Analytical Chemistry. 2016;88(16):7946-58.
80. Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. Proceedings of the National Academy of Sciences. 2015;112(41):12580-5.

81. Dührkop K, Fleischauer M, Ludwig M, Aksenov AA, Melnik AV, Meusel M, et al. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. Nature Methods. 2019;16(4):299-302.
82. Schmid R, Petras D, Nothias L-F, Wang M, Aron AT, Jagels A, et al. Ion identity molecular networking for mass spectrometry-based metabolomics in the GNPS environment. Nature Communications. 2021;12(1):3832.
83. Shen X, Wang R, Xiong X, Yin Y, Cai Y, Ma Z, et al. Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. Nature Communications. 2019;10(1):1516.
84. Chen L, Lu W, Wang L, Xing X, Chen Z, Teng X, et al. Metabolite discovery through global annotation of untargeted metabolomics data. Nature Methods. 2021;18(11):1377-85.
85. Yi Z, Zhu ZJ. Overview of Tandem Mass Spectral and Metabolite Databases for Metabolite Identification in Metabolomics. Methods Mol Biol. 2020;2104:139-48.
86. Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. METLIN: a metabolite mass spectral database. Ther Drug Monit. 2005;27(6):747-51.
87. Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, et al. MassBank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom. 2010;45(7):703-14.
88. Wishart DS, Feunang YD, Marcu A, Guo AC, Liang K, Vazquez-Fresno R, et al. HMDB 4.0: the human metabolome database for 2018. Nucleic Acids Res. 2018;46(D1):D608-D17.
89. Wang M, Carver JJ, Phelan VV, Sanchez LM, Garg N, Peng Y, et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. Nat Biotechnol. 2016;34(8):828-37.
90. Jeffryes JG, Colastani RL, Elbadawi-Sidhu M, Kind T, Niehaus TD, Broadbelt LJ, et al. MINEs: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. Journal of Cheminformatics. 2015;7(1):44.
91. Kind T, Liu KH, Lee DY, DeFelice B, Meissen JK, Fiehn O. LipidBlast in silico tandem mass spectrometry database for lipid identification. Nat Methods. 2013;10(8):755-8.
92. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587-d92.
93. Fahy E, Subramaniam S, Murphy RC, Nishijima M, Raetz CR, Shimizu T, et al. Update of the LIPID MAPS comprehensive classification system for lipids. J Lipid Res. 2009;50 Suppl(Suppl):S9-14.
94. Watanabe K, Yasugi E, Oshima M. How to Search the Glycolipid data in “LIPIDBANK for Web”, the Newly Developed Lipid Database in Japan. Trends in Glycoscience and Glycotechnology. 2000;12:175-84.
95. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. Bioinformatics. 2010;26(18):2342-4.
96. Beger RD. A Review of Applications of Metabolomics in Cancer. Metabolites [Internet]. 2013; 3(3):[552-74 pp.].

97. Chagoyen M, Pazos F. Tools for the functional interpretation of metabolomic experiments. Briefings in Bioinformatics. 2013;14(6):737-44.
98. Wieder C, Frainay C, Poupin N, Rodríguez-Mier P, Vinson F, Cooke J, et al. Pathway analysis in metabolomics: Recommendations for the use of over-representation analysis. PLoS Comput Biol. 2021;17(9):e1009105.
99. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway Analysis: State of the Art. Front Physiol. 2015;6:383.
100. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. Nucleic Acids Res. 2010;38(Web Server issue):W71-7.
101. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, et al. Predicting Network Activity from High Throughput Metabolomics. PLOS Computational Biology. 2013;9(7):e1003123.
102. Pang Z, Chong J, Li S, Xia J. MetaboAnalystR 3.0: Toward an Optimized Workflow for Global Metabolomics. Metabolites. 2020;10(5):186.
103. Hartl J, Kiefer P, Kaczmarczyk A, Mittelviefhaus M, Meyer F, Vonderach T, et al. Untargeted metabolomics links glutathione to bacterial cell cycle progression. Nat Metab. 2020;2(2):153-66.
104. Uppal K, Walker DI, Liu K, Li S, Go YM, Jones DP. Computational Metabolomics: A Framework for the Million Metabolome. Chem Res Toxicol. 2016;29(12):1956-75.
105. Chong J, Yamamoto M, Xia J. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. Metabolites. 2019;9(3):57.
106. De Bruycker K, Welle A, Hirth S, Blanksby S, Barner-Kowollik C. Mass spectrometry as a tool to advance polymer science. Nature Reviews Chemistry. 2020.
107. Albóniga OE, González O, Alonso RM, Xu Y, Goodacre R. Optimization of XCMS parameters for LC-MS metabolomics: an assessment of automated versus manual tuning and its effect on the final results. Metabolomics. 2020;16(1):14.
108. Nash WJ, Dunn WB. From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data. TrAC Trends in Analytical Chemistry. 2019;120:115324.
109. Guo J, Shen S, Huan T. Paramounter: Direct Measurement of Universal Parameters To Process Metabolomics Data in a “White Box”. Analytical Chemistry. 2022;94(10):4260-8.
110. McLean C, Kujawinski EB. AutoTuner: High Fidelity and Robust Parameter Selection for Metabolomics Data Processing. Analytical Chemistry. 2020.
111. Zheng H, Clausen MR, Dalsgaard TK, Mortensen G, Bertram HC. Time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches. Anal Chem. 2013;85(15):7109-16.
112. Manier SK, Keller A, Meyer MR. Automated optimization of XCMS parameters for improved peak picking of liquid chromatography-mass spectrometry data using the coefficient of variation and parameter sweeping for untargeted metabolomics. Drug Test Anal. 2019;11(6):752-61.
113. Lloyd-Price J, Arze C, Ananthakrishnan AN, Schirmer M, Avila-Pacheco J, Poon TW, et al. Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases. Nature. 2019;569(7758):655-62.

114. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nat Protoc. 2011;6(7):1060-83.
115. Li B, Tang J, Yang Q, Li S, Cui X, Li Y, et al. NOREVA: normalization and evaluation of MS-based metabolomics data. Nucleic Acids Res. 2017;45(W1):W162-W70.
116. Pramann A, Noordmann J, Rienitz O. Investigation of mass-scale drift effects in the milli-mass range: Influence on high mass resolution mode multicollector-inductively coupled plasma mass spectrometer isotope ratio measurements. Journal of Mass Spectrometry. 2021;56(6):e4732.
117. Thonusin C, IglayReger HB, Soni T, Rothberg AE, Burant CF, Evans CR. Evaluation of intensity drift correction strategies using MetaboDrift, a normalization tool for multi-batch metabolomics data. J Chromatogr A. 2017;1523:265-74.
118. Deng K, Zhang F, Tan Q, Huang Y, Song W, Rong Z, et al. WaveICA: A novel algorithm to remove batch effects for large-scale untargeted metabolomics data based on wavelet analysis. Anal Chim Acta. 2019;1061:60-9.
119. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, Siuzdak G. Annotation: A Computational Solution for Streamlining Metabolomics Analysis. Anal Chem. 2018;90(1):480-9.
120. Chaleckis R, Meister I, Zhang P, Wheelock CE. Challenges, progress and promises of metabolite annotation for LC-MS-based metabolomics. Curr Opin Biotechnol. 2019;55:44-50.
121. Shuzhao L. 2019 [Available from: https://github.com/shuzhao-li/mummichog.
122. Chong J, Xia J. MetaboAnalystR: an R package for flexible and reproducible analysis of metabolomics data. Bioinformatics. 2018;34(24):4313-4.
123. Xia J. 2020 [Available from: https://github.com/xia-lab/MetaboAnalystR.
124. Li Z, Lu Y, Guo Y, Cao H, Wang Q, Shui W. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. Anal Chim Acta. 2018;1029:50-7.
125. Simon-Manso Y, Lowenthal MS, Kilpatrick LE, Sampson ML, Telu KH, Rudnick PA, et al. Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. Anal Chem. 2013;85(24):11725-31.
126. Eliasson M, Rannar S, Madsen R, Donten MA, Marsden-Edwards E, Moritz T, et al. Strategy for optimizing LC-MS data processing in metabolomics: a design of experiments approach. Anal Chem. 2012;84(15):6869-76.
127. Mena Bares LMf, Benitez Cantero JM, Iglesias Flores E, Gros Alcalde B, Moreno Ortega E, Maza Muret FR, et al. Bile acid malabsorption in patients with chronic diarrhea and Crohn's disease. Rev Esp Enferm Dig. 2019;111(1):40-5.
128. Uchiyama K, Kishi H, Komatsu W, Nagao M, Ohhira S, Kobashi G. Lipid and Bile Acid Dysmetabolism in Crohn's Disease. J Immunol Res. 2018;2018:7270486.
129. Kuroki F, Iida M, Tominaga M, Matsumoto T, Kanamoto K, Fujishima M. Is vitamin E depleted in Crohn's disease at initial diagnosis? Dig Dis. 1994;12(4):248-54.

130. Narula N, Cooray M, Anglin R, Muqtadir Z, Narula A, Marshall JK. Impact of High-Dose Vitamin D3 Supplementation in Patients with Crohn's Disease in Remission: A Pilot Randomized Double-Blind Controlled Study. Dig Dis Sci. 2017;62(2):448-55.
131. Dionne S, Calderon MR, White JH, Memari B, Elimrani I, Adelson B, et al. Differential effect of vitamin D on NOD2- and TLR-induced cytokines in Crohn's disease. Mucosal Immunol. 2014;7(6):1405-15.
132. Scoville EA, Allaman MM, Brown CT, Motley AK, Horst SN, Williams CS, et al. Alterations in Lipid, Amino Acid, and Energy Metabolism Distinguish Crohn's Disease from Ulcerative Colitis and Control Subjects by Serum Metabolomic Profiling. Metabolomics. 2018;14(1):17.
133. Kolacek M, Paduchova Z, Dvorakova M, Zitnanova I, Cierna I, Durackova Z, et al. Effect of natural polyphenols on thromboxane levels in children with Crohn's disease. Bratisl Lek Listy. 2019;120(12):924-8.
134. Petrey AC, de la Motte CA. Hyaluronan in inflammatory bowel disease: Cross-linking inflammation and coagulation. Matrix Biol. 2019;78-79:314-23.
135. Ramette A. Multivariate analyses in microbial ecology. FEMS Microbiol Ecol. 2007;62(2):142-60.
136. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.
137. Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM. Metabolomics data normalization with EigenMS. PLoS One. 2014;9(12):e116221.
138. Wehrens R, Hageman JA, van Eeuwijk F, Kooke R, Flood PJ, Wijnker E, et al. Improved batch correction in untargeted MS-based metabolomics. Metabolomics. 2016;12:88.
139. De Livera AM, Sysi-Aho M, Jacob L, Gagnon-Bartsch JA, Castillo S, Simpson JA, et al. Statistical methods for handling unwanted variation in metabolomics data. Anal Chem. 2015;87(7):3606-15.
140. De Livera AM, Dias DA, De Souza D, Rupasinghe T, Pyke J, Tull D, et al. Normalizing and integrating metabolomics data. Anal Chem. 2012;84(24):10768-76.
141. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896-902.
142. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M. Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics. 2007;8:93.
143. Redestig H, Fukushima A, Stenlund H, Moritz T, Arita M, Saito K, et al. Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. Anal Chem. 2009;81(19):7974-80.
144. Chambers MC, Maclean B, Burke R, Amodei D, Ruderman DL, Neumann S, et al. A cross-platform toolkit for mass spectrometry and proteomics. Nature Biotechnology. 2012;30(10):918-20.
145. Rinschen MM, Ivanisevic J, Giera M, Siuzdak G. Identification of bioactive metabolites using activity metabolomics. Nat Rev Mol Cell Biol. 2019;20(6):353-67.
146. Wishart DS. Metabolomics for Investigating Physiological and Pathophysiological Processes. Physiol Rev. 2019;99(4):1819-75.

147. Vermeulen R, Schymanski EL, Barabási AL, Miller GW. The exposome and health: Where chemistry meets biology. Science. 2020;367(6476):392-6.
148. Beger RD, Dunn W, Schmidt MA, Gross SS, Kirwan JA, Cascante M, et al. Metabolomics enables precision medicine: "A White Paper, Community Perspective". Metabolomics. 2016;12(10):149.
149. Panyard DJ, Kim KM, Darst BF, Deming YK, Zhong X, Wu Y, et al. Cerebrospinal fluid metabolomics identifies 19 brain-related phenotype associations. Commun Biol. 2021;4(1):63.
150. Aikaterini I, Emmanuel M, Ibrahim K, Freya E, Julian LG, Ioanna T, et al. Metabolic phenotyping and cardiovascular disease: an overview of evidence from epidemiological settings. Heart. 2021;107(14):1123.
151. Stanstrup J, Broeckling CD, Helmus R, Hoffmann N, Mathé E, Naake T, et al. The metaRbolomics Toolbox in Bioconductor and beyond. Metabolites. 2019;9(10):200.
152. Spicer R, Salek RM, Moreno P, Cañueto D, Steinbeck C. Navigating freely-available software tools for metabolomics analysis. Metabolomics. 2017;13(9):106.
153. Huan T, Forsberg EM, Rinehart D, Johnson CH, Ivanisevic J, Benton HP, et al. Systems biology guided by XCMS Online metabolomics. Nat Methods. 2017;14(5):461-2.
154. Xia J, Psychogios N, Young N, Wishart DS. MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res. 2009;37(Web Server issue):W652-60.
155. Xia J, Mandal R, Sinelnikov IV, Broadhurst D, Wishart DS. MetaboAnalyst 2.0--a comprehensive server for metabolomic data analysis. Nucleic Acids Res. 2012;40(Web Server issue):W127-33.
156. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013;9(2):280-99.
157. Xia J, Sinelnikov IV, Han B, Wishart DS. MetaboAnalyst 3.0--making metabolomics more meaningful. Nucleic Acids Res. 2015;43(W1):W251-7.
158. Chong J, Soufan O, Li C, Caraus I, Li S, Bourque G, et al. MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res. 2018;46(W1):W486-w94.
159. Yu B, Zanetti KA, Temprosa M, Albanes D, Appel N, Barrera CB, et al. The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies. Am J Epidemiol. 2019;188(6):991-1012.
160. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: beyond biomarkers and towards mechanisms. Nat Rev Mol Cell Biol. 2016;17(7):451-9.
161. Du X, Smirnov A, Pluskal T, Jia W, Sumner S. Metabolomics Data Preprocessing Using ADAP and MZmine 2. Methods Mol Biol. 2020;2104:25-48.
162. Chagoyen M, López-Ibáñez J, Pazos F. Functional Analysis of Metabolomics Data. Methods Mol Biol. 2016;1415:399-406.
163. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res. 2016;44(D1):D463-70.

164. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, et al. MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res. 2020;48(D1):D440-d4.
165. Basu S, Duren W, Evans CR, Burant CF, Michailidis G, Karnovsky A. Sparse network modeling and metscape-based visualization methods for the analysis of large-scale metabolomics data. Bioinformatics. 2017;33(10):1545-53.
166. Hoffmann N, Rein J, Sachsenberg T, Hartler J, Haug K, Mayer G, et al. mzTab-M: A Data Standard for Sharing Quantitative Results in Mass Spectrometry Metabolomics. Anal Chem. 2019;91(5):3302-10.
167. Forsberg EM, Huan T, Rinehart D, Benton HP, Warth B, Hilmers B, et al. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. Nat Protoc. 2018;13(4):633-51.
168. Huss M, Holme P. Currency and commodity metabolites: their identification and relation to the modularity of metabolic networks. IET Syst Biol. 2007;1(5):280-5.
169. Xia J, Lyle NH, Mayer ML, Pena OM, Hancock RE. INVEX--a web-based tool for integrative visualization of expression data. Bioinformatics. 2013;29(24):3232-4.
170. Xia J, Fjell CD, Mayer ML, Pena OM, Wishart DS, Hancock RE. INMEX--a web-based tool for integrative meta-analysis of expression data. Nucleic Acids Res. 2013;41(Web Server issue):W63-70.
171. Huan T, Palermo A, Ivanisevic J, Rinehart D, Edler D, Phommavongsay T, et al. Autonomous Multimodal Metabolomics Data Integration for Comprehensive Pathway Analysis and Systems Biology. Anal Chem. 2018;90(14):8396-403.
172. Srivastava V, Obudulu O, Bygdell J, Löfstedt T, Rydén P, Nilsson R, et al. OnPLS integration of transcriptomic, proteomic and metabolomic data shows multi-level oxidative stress responses in the cambium of transgenic hipI- superoxide dismutase Populus plants. BMC Genomics. 2013;14:893.
173. Zhou G, Li S, Xia J. Network-Based Approaches for Multi-omics Integration. Methods Mol Biol. 2020;2104:469-87.
174. Chong J, Xia J. Computational Approaches for Integrative Analysis of the Metabolome and Microbiome. Metabolites. 2017;7(4):62.
175. Krumsiek J, Bartel J, Theis FJ. Computational approaches for systems metabolomics. Curr Opin Biotechnol. 2016;39:198-206.
176. Cavill R, Jennen D, Kleinjans J, Briedé JJ. Transcriptomic and metabolomic data integration. Brief Bioinform. 2016;17(5):891-901.
177. Jana J, Sara van de G. Confidence intervals for high-dimensional inverse covariance estimation. Electronic Journal of Statistics. 2015;9(1):1205-29.
178. Wishart DS, Knox C, Guo AC, Eisner R, Young N, Gautam B, et al. HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res. 2009;37(Database issue):D603-10.
179. Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic Acids Res. 2021;49(D1):D545-d51.
180. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, et al. PubChem 2019 update: improved access to chemical data. Nucleic Acids Res. 2019;47(D1):D1102-d9.

181. de Matos P, Dekker A, Ennis M, Hastings J, Haug K, Turner S, et al. ChEBI: a chemistry ontology and database. Journal of Cheminformatics. 2010;2(1):P6.
182. Fahy E, Subramaniam S. RefMet: a reference nomenclature for metabolomics. Nat Methods. 2020;17(12):1173-4.
183. O'Donnell VB, Dennis EA, Wakelam MJO, Subramaniam S. LIPID MAPS: Serving the next generation of lipid researchers with tools, resources, data, and training. Science Signaling. 2019;12(563):eaaw2964.
184. Liebisch G, Fahy E, Aoki J, Dennis EA, Durand T, Ejsing CS, et al. Update on LIPID MAPS classification, nomenclature, and shorthand notation for MS-derived lipid structures. J Lipid Res. 2020;61(12):1539-55.
185. Reinhold D, Pielke-Lombardo H, Jacobson S, Ghosh D, Kechris K. Pre-analytic Considerations for Mass Spectrometry-Based Untargeted Metabolomics Data. Methods Mol Biol. 2019;1978:323-40.
186. Guijas C, Montenegro-Burke JR, Domingo-Almenara X, Palermo A, Warth B, Hermann G, et al. METLIN: A Technology Platform for Identifying Knowns and Unknowns. Anal Chem. 2018;90(5):3156-64.
187. Giacomoni F, Le Corguillé G, Monsoor M, Landi M, Pericard P, Pétéra M, et al. Workflow4Metabolomics: a collaborative research infrastructure for computational metabolomics. Bioinformatics. 2015;31(9):1493-5.
188. Kuo TC, Tian TF, Tseng YJ. 3Omics: a web-based systems biology tool for analysis, integration and visualization of human transcriptomic, proteomic and metabolomic data. BMC Syst Biol. 2013;7:64.
189. Yang Q, Wang Y, Zhang Y, Li F, Xia W, Zhou Y, et al. NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data. Nucleic Acids Res. 2020;48(W1):W436-w48.
190. Stancliffe E, Schwaiger-Haber M, Sindelar M, Patti GJ. DecoID improves identification rates in metabolomics through database-assisted MS/MS deconvolution. Nature Methods. 2021;18(7):779-87.
191. Raetz M, Bonner R, Hopfgartner G. SWATH-MS for metabolomics and lipidomics: critical aspects of qualitative and quantitative analysis. Metabolomics. 2020;16(6):71.
192. Lu Y, Pang Z, Xia J. Comprehensive investigation of pathway enrichment methods for functional interpretation of LC–MS global metabolomics data. Briefings in Bioinformatics. 2022;24(1).
193. Houel S, Abernathy R, Renganathan K, Meyer-Arendt K, Ahn NG, Old WM. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. J Proteome Res. 2010;9(8):4152-60.
194. Li Y, Kind T, Folz J, Vaniya A, Mehta SS, Fiehn O. Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification. Nature Methods. 2021;18(12):1524-31.
195. Aisporna A, Benton HP, Chen A, Derks RJE, Galano JM, Giera M, et al. Neutral Loss Mass Spectral Data Enhances Molecular Similarity Analysis in METLIN. J Am Soc Mass Spectrom. 2022;33(3):530-4.
196. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, et al. HMDB 5.0: the Human Metabolome Database for 2022. Nucleic Acids Res. 2022;50(D1):D622-d31.

197. Karp PD, Billington R, Caspi R, Fulcher CA, Latendresse M, Kothari A, et al. The BioCyc collection of microbial genomes and metabolic pathways. Brief Bioinform. 2019;20(4):1085-93.
198. Phapale P, Palmer A, Gathungu RM, Kale D, Brügger B, Alexandrov T. Public LC-Orbitrap Tandem Mass Spectral Library for Metabolite Identification. Journal of Proteome Research. 2021;20(4):2089-97.
199. Narayanaswamy P, Teo G, Ow JR, Lau A, Kaldis P, Tate S, et al. MetaboKit: a comprehensive data extraction tool for untargeted metabolomics. Molecular Omics. 2020;16(5):436-47.
200. Sameh M, Khalaf HM, Anwar AM, Osama A, Ahmed EA, Mahgoub S, et al. Integrated multiomics analysis to infer COVID-19 biological insights. Scientific Reports. 2023;13(1):1802.
201. Stringer KA, Younger JG, McHugh C, Yeomans L, Finkel MA, Puskarich MA, et al. Whole Blood Reveals More Metabolic Detail of the Human Metabolome than Serum as Measured by 1H-NMR Spectroscopy: Implications for Sepsis Metabolomics. Shock. 2015;44(3):200-8.
202. Thomas T, Stefanoni D, Dzieciatkowska M, Issaian A, Nemkov T, Hill RC, et al. Evidence of Structural Protein Damage and Membrane Lipid Remodeling in Red Blood Cells from COVID-19 Patients. Journal of Proteome Research. 2020;19(11):4455-69.
203. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, et al. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. Environmental Science & Technology. 2014;48(4):2097-8.
204. Ishikawa M, Maekawa K, Saito K, Senoo Y, Urata M, Murayama M, et al. Plasma and Serum Lipidomics of Healthy White Adults Shows Characteristic Profiles by Subjects' Gender and Age. PLOS ONE. 2014;9(3):e91806.
205. Li S, Siddiqa A, Thapa M, Chi Y, Zheng S. Trackable and scalable LC-MS metabolomics data processing using asari. Nature Communications. 2023;14(1):4113.
206. Ripon MAR, Bhowmik DR, Amin MT, Hossain MS. Role of arachidonic cascade in COVID-19 infection: A review. Prostaglandins Other Lipid Mediat. 2021;154:106539.
207. Bae JH, Choe HJ, Holick MF, Lim S. Association of vitamin D status with COVID-19 and its severity : Vitamin D and COVID-19: a narrative review. Rev Endocr Metab Disord. 2022;23(3):579-99.
208. Xu Y, Baylink DJ, Chen C-S, Reeves ME, Xiao J, Lacy C, et al. The importance of vitamin d metabolism as a potential prophylactic, immunoregulatory and neuroprotective treatment for COVID-19. Journal of Translational Medicine. 2020;18(1):322.
209. Theken KN, Tang SY, Sengupta S, FitzGerald GA. The roles of lipids in SARS-CoV-2 viral replication and the host immune response. Journal of Lipid Research. 2021;62:100129.
210. Santa-Rios A, Barst BD, Basu N. Mercury Speciation in Whole Blood and Dried Blood Spots from Capillary and Venous Sources. Analytical Chemistry. 2020;92(5):3605-12.
211. Johnson JM, Yu T, Strobel FH, Jones DP. A practical approach to detect unique metabolic patterns for personalized medicine. Analyst. 2010;135(11):2864-70.
212. Go YM, Walker DI, Liang Y, Uppal K, Soltow QA, Tran V, et al. Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. Toxicol Sci. 2015;148(2):531-43.

213. Tenenbaum D, Maintainer B. KEGGREST: Client-side REST access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version. 2021;1(0).
214. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, et al. ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. Journal of Cheminformatics. 2016;8(1):61.
215. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. 2017;45(D1):D353-d61.
216. Wishart DS, Oler E, Peters H, Guo A, Girod S, Han S, et al. MiMeDB: the Human Microbial Metabolome Database. Nucleic Acids Res. 2023;51(D1):D611-d20.
217. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, et al. T3DB: the toxic exposome database. Nucleic Acids Res. 2015;43(Database issue):D928-34.
218. Rothwell JA, Perez-Jimenez J, Neveu V, Medina-Remón A, M'Hiri N, García-Lobato P, et al. Phenol-Explorer 3.0: a major update of the Phenol-Explorer database to incorporate data on the effects of food processing on polyphenol content. Database (Oxford). 2013;2013:bat070.
219. Neveu V, Moussy A, Rouaix H, Wedekind R, Pon A, Knox C, et al. Exposome-Explorer: a manually-curated database on biomarkers of exposure to dietary and environmental factors. Nucleic Acids Research. 2016;45(D1):D979-D84.
220. Mohammed Taha H, Aalizadeh R, Alygizakis N, Antignac JP, Arp HPH, Bade R, et al. The NORMAN Suspect List Exchange (NORMAN-SLE): facilitating European and worldwide collaboration on suspect screening in high resolution mass spectrometry. Environ Sci Eur. 2022;34(1):104.
221. Cohen Freue GV, Kepplinger D, Salibián-Barrera M, Smucler E. Robust elastic net estimators for variable selection and identification of proteomic biomarkers. The Annals of Applied Statistics. 2019;13(4):2065-90, 26.
222. Kepplinger D. Robust estimation and variable selection in high-dimensional linear regression models [Text]2020.
223. Lucas C, Wong P, Klein J, Castro TB, Silva J, Sundaram M, et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. Nature. 2020;584(7821):463-9.
224. Organization WH. Coronavirus Disease (COVID-19) Pandemic. 2020. p. https://www.who.int/emergencies/diseases/novel-coronavirus-2019.
225. Azkur AK, Akdis M, Azkur D, Sokolowska M, van de Veen W, Brüggen MC, et al. Immune response to SARS-CoV-2 and mechanisms of immunopathological changes in COVID-19. Allergy. 2020;75(7):1564-81.
226. Grifoni A, Weiskopf D, Ramirez SI, Mateus J, Dan JM, Moderbacher CR, et al. Targets of T cell responses to SARS-CoV-2 coronavirus in humans with COVID-19 disease and unexposed individuals. Cell. 2020;181(7):1489-501. e15.
227. Psychogios N, Hau DD, Peng J, Guo AC, Mandal R, Bouatra S, et al. The human serum metabolome. PloS one. 2011;6(2):e16957.
228. Wu D, Shu T, Yang X, Song J-X, Zhang M, Yao C, et al. Plasma metabolomic and lipidomic alterations associated with COVID-19. National Science Review. 2020;7(7):1157-68.

229. Song J-W, Lam SM, Fan X, Cao W-J, Wang S-Y, Tian H, et al. Omics-driven systems interrogation of metabolic dysregulation in COVID-19 pathogenesis. Cell metabolism. 2020;32(2):188-202. e5.
230. Shen B, Yi X, Sun Y, Bi X, Du J, Zhang C, et al. Proteomic and metabolomic characterization of COVID-19 patient sera. Cell. 2020;182(1):59-72. e15.
231. Maras JS, Sharma S, Bhat A, Agarwal R, Gupta E, Sarin SK. Multi-Omics Analysis of Respiratory Specimen Characterizes Baseline Molecular Determinants Associated with COVID-19 Diagnosis and Outcome. Reshu and Gupta, Ekta and Sarin, Shiv Kumar, Multi-Omics Analysis of Respiratory Specimen Characterizes Baseline Molecular Determinants Associated with COVID-19 Diagnosis and Outcome. 2020.
232. Blasco H, Bessy C, Plantier L, Lefevre A, Piver E, Bernard L, et al. The specific metabolome profiling of patients infected by SARS-COV-2 supports the key role of tryptophan-nicotinamide pathway and cytosine metabolism. Scientific reports. 2020;10(1):16824.
233. Thomas T, Stefanoni D, Reisz JA, Nemkov T, Bertolone L, Francis RO, et al. COVID-19 infection alters kynurenine and fatty acid metabolism, correlating with IL-6 levels and renal status. JCI insight. 2020;5(14).
234. Cai Y, Kim DJ, Takahashi T, Broadhurst DI, Ma S, Rattray NJ, et al. Kynurenic acid underlies sex-specific immune responses to COVID-19. MedRxiv. 2020.
235. Su Y, Chen D, Lausted C, Yuan D, Choi J, Dai C, et al. Multiomic immunophenotyping of COVID-19 patients reveals early infection trajectories. BioRxiv. 2020:2020.07. 27.224063.
236. Zhao Y, Shang Y, Ren Y, Bie Y, Qiu Y, Yuan Y, et al. Omics study reveals abnormal alterations of breastmilk proteins and metabolites in puerperant women with COVID-19. Signal transduction and targeted therapy. 2020;5(1):247.
237. Delafiori J, Navarro LC, Siciliano RF, de Melo GC, Busanello ENB, Nicolau JC, et al. Covid-19 automated diagnosis and risk assessment through Metabolomics and Machine-Learning (preprint). 2020.
238. Hannun YA, Obeid LM. Principles of bioactive lipid signalling: lessons from sphingolipids. Nature reviews Molecular cell biology. 2008;9(2):139-50.
239. Chiurchiù V, Leuti A, Maccarrone M. Bioactive lipids and chronic inflammation: managing the fire within. Frontiers in immunology. 2018;9:38.
240. Struwe W, Emmott E, Bailey M, Sharon M, Sinz A, Corrales FJ, et al. The COVID-19 MS Coalition—accelerating diagnostics, prognostics, and treatment. The Lancet. 2020;395(10239):1761-2.
241. Zaas AK, Chen M, Varkey J, Veldman T, Hero AO, Lucas J, et al. Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell host & microbe. 2009;6(3):207-17.
242. Zhou G, Stevenson MM, Geary TG, Xia J. Comprehensive transcriptome meta-analysis to characterize host immune responses in helminth infections. PLoS neglected tropical diseases. 2016;10(4):e0004624.
243. Gardinassi LG, Souza CO, Sales-Campos H, Fonseca SG. Immune and metabolic signatures of COVID-19 revealed by transcriptomics data reuse. Frontiers in immunology. 2020;11:1636.

244. Cavalli E, Petralia MC, Basile MS, Bramanti A, Bramanti P, Nicoletti F, et al. Transcriptomic analysis of COVID-19 lungs and bronchoalveolar lavage fluid samples reveals predominant B cell activation responses to infection. International journal of molecular medicine. 2020;46(4):1266-73.
245. Pinto BG, Oliveira AE, Singh Y, Jimenez L, Gonçalves AN, Ogava RL, et al. ACE2 expression is increased in the lungs of patients with comorbidities associated with severe COVID-19. The Journal of infectious diseases. 2020;222(4):556-63.
246. Kale NS, Haug K, Conesa P, Jayseelan K, Moreno P, Rocca-Serra P, et al. MetaboLights: an open-access database repository for metabolomics data. Current protocols in bioinformatics. 2016;53(1):14.3. 1-.3. 8.
247. Gauglitz JM, Bittremieux W, Williams CL, Weldon KC, Panitchpakdi M, Di Ottavio F, et al. Reference data based insights expand understanding of human metabolomes. BioRxiv. 2020:2020.07. 08.194159.
248. García LF. Immune response, inflammation, and the clinical spectrum of COVID-19. Frontiers in immunology. 2020;11:1441.
249. San Juan I, Bruzzone C, Bizkarguenaga M, Bernardo-Seisdedos G, Laín A, Gil-Redondo R, et al. Abnormal concentration of porphyrins in serum from COVID-19 patients. British journal of haematology. 2020;190(5):e265.
250. Ponti G, Maccaferri M, Ruini C, Tomasi A, Ozben T. Biomarkers associated with COVID-19 disease progression. Critical reviews in clinical laboratory sciences. 2020;57(6):389-99.
251. Liu W, Li H. COVID-19: attacks the 1-beta chain of hemoglobin and captures the porphyrin to inhibit human heme metabolism. ChemRxiv. Preprint. 2020;10.
252. Thaker SK, Ch'ng J, Christofk HR. Viral hijacking of cellular metabolism. BMC biology. 2019;17:1-15.
253. Hoxha M. What about COVID-19 and arachidonic acid pathway? European journal of clinical pharmacology. 2020;76:1501-4.
254. Hammock BD, Wang W, Gilligan MM, Panigrahy D. Eicosanoids: the overlooked storm in coronavirus disease 2019 (COVID-19)? The American journal of pathology. 2020;190(9):1782-8.
255. Schwarz B, Sharma L, Roberts L, Peng X, Bermejo S, Leighton I, et al. Severe SARS-CoV-2 infection is defined by a shift in the serum lipidome resulting in dysregulation of eicosanoid lipid immune mediators. 2020.
256. Darwesh AM, Bassiouni W, Sosnowski DK, Seubert JM. Can N-3 polyunsaturated fatty acids be considered a potential adjuvant therapy for COVID-19-associated cardiovascular complications? Pharmacology & therapeutics. 2021;219:107703.
257. Margină D, Ungurianu A, Purdel C, Tsoukalas D, Sarandi E, Thanasoula M, et al. Chronic inflammation in the context of everyday life: dietary changes as mitigating factors. International journal of environmental research and public health. 2020;17(11):4135.
258. Iddir M, Brito A, Dingeo G, Fernandez Del Campo SS, Samouda H, La Frano MR, et al. Strengthening the immune system and reducing inflammation and oxidative stress through diet and nutrition: considerations during the COVID-19 crisis. Nutrients. 2020;12(6):1562.
259. Dhar D, Mohanty A. Gut microbiota and Covid-19-possible link and implications. Virus research. 2020;285:198018.

260. Gassen NC, Papies J, Bajaj T, Dethloff F, Emanuel J, Weckmann K, et al. Analysis of SARS-CoV-2-controlled autophagy reveals spermidine, MK-2206, and niclosamide as putative antiviral therapeutics. BioRxiv. 2020:2020.04. 15.997254.
261. Li Z, Liu G, Wang L, Liang Y, Zhou Q, Wu F, et al. From the insight of glucose metabolism disorder: oxygen therapy and blood glucose monitoring are crucial for quarantined COVID-19 patients. Ecotoxicology and Environmental Safety. 2020;197:110614.
262. Williams NC, O'Neill LA. A role for the Krebs cycle intermediate citrate in metabolic reprogramming in innate immunity and inflammation. Frontiers in immunology. 2018;9:141.
263. Takahashi K, Suzuki N, Ogra Y. Effect of gut microflora on nutritional availability of selenium. Food chemistry. 2020;319:126537.
264. Kumar P, Chander B. COVID 19 mortality: Probable role of microbiome to explain disparity. Medical Hypotheses. 2020;144:110209.
265. Bradley BT, Maioli H, Johnston R, Chaudhry I, Fink SL, Xu H, et al. Histopathology and ultrastructural findings of fatal COVID-19 infections in Washington State: a case series. The Lancet. 2020;396(10247):320-32.
266. Polonikov A. Endogenous deficiency of glutathione as the most likely cause of serious manifestations and death in COVID-19 patients. ACS infectious diseases. 2020;6(7):1558-62.
267. Wang D, Hu B, Hu C, Zhu F, Liu X, Zhang J, et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus–Infected Pneumonia in Wuhan, China. JAMA. 2020;323(11):1061-9.
268. Abdulrab S, Al-Maweri S, Halboub E. Ursodeoxycholic acid as a candidate therapeutic to alleviate and/or prevent COVID-19-associated cytokine storm. Medical Hypotheses. 2020;143:109897.
269. Ouyang L, Gong J. Mitochondrial-targeted ubiquinone: A potential treatment for COVID-19. Medical Hypotheses. 2020;144:110161.
270. Kumrungsee T, Zhang P, Chartkul M, Yanaka N, Kato N. Potential Role of Vitamin B6 in Ameliorating the Severity of COVID-19 and Its Complications. Frontiers in Nutrition. 2020;7.
271. Thair SA, He YD, Hasin-Brumshtein Y, Sakaram S, Pandya R, Toh J, et al. Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections. iScience. 2021;24(1):101947.
272. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009;6(7):e1000097.
273. China NHCotPsRo. Diagnosis and Treatment Protocol for Novel Coronavirus Pneumonia (8th). 2020. p. http://www.nhc.gov.cn/.
274. WHO. WHO R&D Blueprint novel Coronavirus COVID-19 Therapeutic Trial Synopsis. 2020. p. https://www.who.int/blueprint/priority-diseases/key-action/COVID-19_Treatment_Trial_Design_Master_Protocol_synopsis_Final_18022020.pdf.
275. Chong J, Wishart DS, Xia J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. Current Protocols in Bioinformatics. 2019;68(1):e86.

276. Pang Z, Chong J, Zhou G, de Lima Morais DA, Chang L, Barrette M, et al. MetaboAnalyst 5.0: narrowing the gap between raw spectra and functional insights. Nucleic Acids Research. 2021;49(W1):W388-W96.
277. Pang Z, Zhou G, Ewald J, Chang L, Hacariz O, Basu N, et al. Using MetaboAnalyst 5.0 for LC–HRMS spectra processing, multi-omics integration and covariate adjustment of global metabolomics data. Nature Protocols. 2022;17(8):1735-61.
278. Sindelar M, Patti GJ. Chemical Discovery in the Era of Metabolomics. Journal of the American Chemical Society. 2020;142(20):9097-105.
279. Haijes HA, Willemsen M, Van der Ham M, Gerrits J, Pras-Raves ML, Prinsen H, et al. Direct Infusion Based Metabolomics Identifies Metabolic Disease in Patients' Dried Blood Spots and Plasma. Metabolites. 2019;9(1).