Untargeted Metabolomics:
Data Analysis Platform for
All-Ion Fragmentation Mass Spectrometry

Tada, Ipputa

Doctor of Philosophy

Department of Genetics
School of Life Science
The Graduate University for Advanced Studies, SOKENDAI
September 2020

Abstract

New technologies make a leap forward in research fields by inspiring ideas and requiring new methods. As genetics has been advanced with DNA sequencers, metabolomics has gradually matured with mass spectrometry. Liquid chromatography-tandem mass spectrometry (LC-MS2) is a common technology for metabolomics studies. Major metabolomics experiments are roughly classified into two groups, targeted- and untargeted metabolomics. While targeted metabolomics aims to quantify pre-defined compounds with high accuracy, untargeted metabolomics tries to detect and identify as many compounds as possible for discovery studies. For reliable compound identification and estimation, MS2 spectrum is utilized because the fragmentation pattern of each molecule is consistent under the almost same experimental settings. In classical acquisition methods, a precursor ion is selected to acquire MS2 spectrum from co-eluting compounds. In contrast, All Ion Fragmentation (AIF) can generate all MS2 spectra by setting quite large *m/z* range (e.g. 40–1200 Da). Although AIF-MS is unbiased and reproducible, the acquired MS2 spectra are highly complex and difficult to interpret.

To solve the complex AIF MS2 spectra, I have developed a new Correlation-based Deconvolution (CorrDec) method. The CorrDec method utilizes intensity correlation between precursor ion and its fragment ions among samples. As a demonstration of CorrDec, it was applied to two datasets: dilution series of chemical standards and a 224-sample urinary metabolomics cohort. The serial dilution study showed that the peak intensities of fragment ions were highly correlated with their precursor ions. In the urine cohort study, 105 compounds were identified and CorrDec could generate clean MS2 spectra for 85 compounds out of them (>80% MS2 match with reference). CorrDec can separate completely co-eluting compounds and work well in even low concentration compounds. Consequently, CorrDec

enables more reliable compound annotations and identifications in multi-sample studies for untargeted metabolomics.

In order to confidently annotate and identify compounds in LC-MS2 data, reliable chemical standard libraries including three orthogonal properties—accurate mass (AM), retention time (RT), and MS2 spectrum—are required. In AIF projects, an MS2 spectrum is generated from a mixture of several adduct types and isoforms; therefore, MS2 spectral library measured by AIF mode can improve the quality of compound identification. However, AIF MS2 spectrum contains many noise peaks even in the measurement of a standard. I describe a workflow to confidently obtain AM, RT, and MS2 for a given compound using the AIF method and provide practical recommendations for library development. So far, 814 deconvoluted spectra by CorrDec and MS2Dec of 140 compounds were generated with manual curation as a chemical library. I illustrated how the library increases the confidence of compound identification in complex AIF data. The construction of high-quality, open-access libraries makes compound identifications more transparent, reliable, and transferable to the broader community.

I have proposed the AIF platform consisting of three metabolomics tools—MS-DIAL, MS-FINDER, and MS-LIMA. MS-DIAL and MS-FINDER were improved for AIF data, and MS-LIMA was newly developed. CorrDec was implemented into MS-DIAL, which is universal metabolomics software that supports various instruments. I have also improved MS-DIAL to adopt measurements of multiple collision energies and be fast and stable for large-scale study. MS-DIAL has grown as a modern user-friendly tool by my contributions. Second, MS-FINDER supports compound estimation by characterizing MS2 spectra, which is a key process in untargeted metabolomics. MS-FINDER can annotate MS2 peaks as molecular formulas, chemical substructures, and types of adduct/isotopic ions. Lastly, MS-LIMA helps to properly manage MS2 spectra acquired from both biological samples and chemical standards. MS-LIMA is open-source software to curate, search, compare, and visualize MS2 spectra for stable and reliable management. The freely available AIF platform

supports reliable data analysis, biological and technical insights, and reanalysis using public metabolomics raw data.

For further development in untargeted metabolomics, I believe that reusable data acquisition (such as AIF), reliable compound identification, and a universal and integrated data analysis platform are important. I hope that the AIF platform including CorrDec and the reliable library can improve the quality of compound identification, increase the number of annotated compounds, and help to exploit large-scale untargeted studies.

Acknowledgments

Foremost, I would like to express the deepest appreciation to my supervisor, Professor Masanori Arita, for his kind support, patient guidance, and expert advice. I want to thank him for encouraging me to go abroad (many international conferences and meetings, workshop at UC Davis, research at Karolinska Institute). I am indebted to Project Associate Professor Nozomu Sakurai, Assistant Professor Takeshi Kawashima, Dr. Wataru Tanaka, and other members of the Biological Networks laboratory, for supporting me and having general discussions to get a wide view.

I am deeply grateful to my progress committee members: Professor Yasukazu Nakamura, Professor Ituro Inoue, Professor Ken Kurokawa, and Associate Professor Shoko Kawamoto, for their encouragements and insightful comments, although my research topic is out of their main fields. I would like to show my appreciation to Professor Eiichiro Fukusaki (Osaka University) for constructive discussion.

I want to thank Dr. Hiroshi Tsugawa (RIKEN CSRS and IMS), who is developing several metabolomics tools such as MS-DIAL, for willingly allowing me to use and develop his tools and nonpublic source codes, and teaching me about fundamental computational metabolomics. The metabolomics hackathon with him was very interesting and inspired me to develop user-friendly GUI tools.

I special thank to Associate Professor Craig Wheelock (Karolinska Institute and Gunma University), for providing me with the standardized urine datasets measured by AIF, and offering me the research opportunity at Karolinska Institute for two months. My deepest appreciation goes his laboratory member, Assistant Professor Romanas Chaleckis (Gunma University), for giving me various experimental and computational experiences. The discussions with him had a very positive impact on me, and I learned a lot about

metabolomics from him. I also thank to the rest of Wheelock laboratory members, Dr. Isabel Meister and Dr. Pei Zhang, for their useful suggestions and discussions.

I owe a debt to Japan Society for the Promotion of Science (JSPS), for funding my studies by Grant-in-Aid for JSPS Fellows (Grant Number: JP18J23133).

Contents

List of Figures

List of Tables

Abbreviations

AIF: all-ion fragmentation

DDA: data-dependent acquisition

DIA: data-independent acquisition

MS: mass spectrometry

MS1: mass spectrometry without fragmentation (comparing with MS2)

MS2: tandem mass spectrometry

LC: liquid chromatography

GC: gas chromatography

CE: collision energy

Q: quadrupole

TOF: time of flight

IM: ion mobility

AM: accurate mass

RT: retention time

QC: quality control

IS: internal standard

tIS: technical internal standard

Chapter 1

Introduction

In the omics era, metabolomics is a next challenging research field following genomics, transcriptomics, and proteomics. The human metabolome includes >100,000 metabolites as the intermediate- and end products of metabolism affected by genetic- and environmental factors (**Figure 1-1**). To elucidate new findings, many metabolomics studies have been widely performed in not only human, but also microbes, plants, and other animals. While genomics can reveal many biological insights from genotype, metabolomics can directly observe metabolites related to phenotype.

In this chapter, the fundamental knowledge of metabolomics, especially for all-ion fragmentation-based metabolomics will be introduced for easily understanding other chapters.

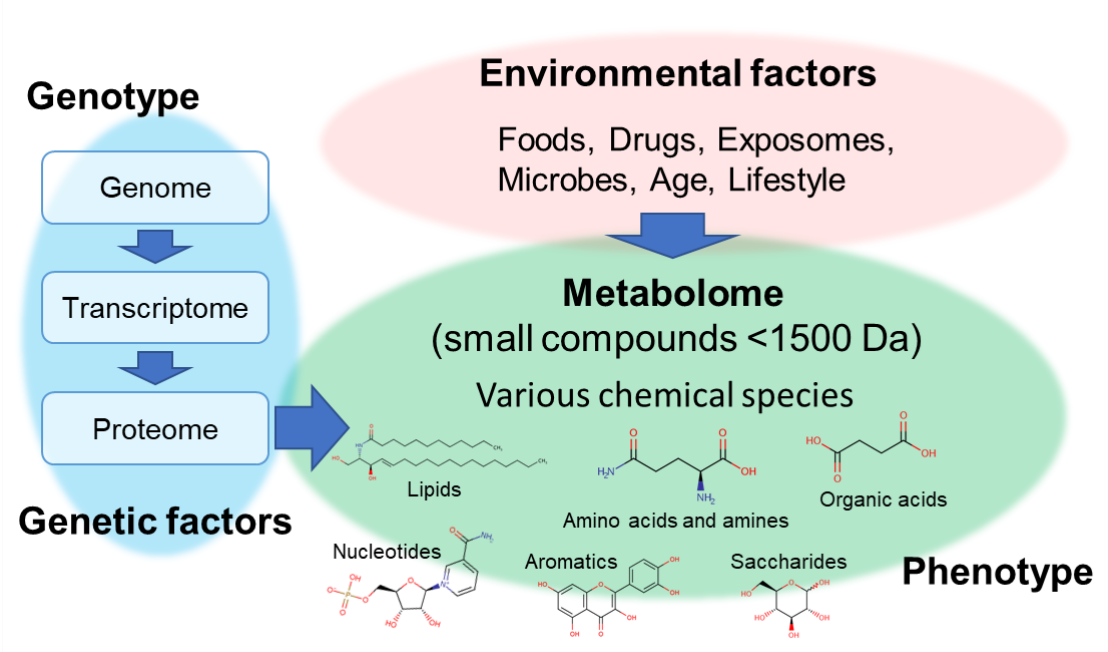


Figure 1-1. Overview of metabolome.

1-1 History of metabolomics

The suffix *-ome* and *-omics* are buzz words in life science [1,2]. The suffix *-ome* is well known as referring to wholeness/completion from the Greek origin [2]. The first *-ome* term *genome*, however, was coined as a blend word of "GENe" and "chromosOME" by Hans Winkler in 1920 [2–4]. The term *genomics* was proposed by Thomas H. Roderick during discussion about a new journal name with Frank Ruddle and Victor McKusick in 1986 [2,5]. 78 years after *genome*, the term *metabolome* was introduced in the scientific literature by Stephen G. Oliver et al. in 1998 [6,7]. In 2002, Oliver Fiehn defined the *metabolomics* as "a comprehensive analysis in which all the metabolites of a biological system are identified and quantified" and described the difference between metabolomics and metabolite profiling (or metabolic profiling; the measurement of pre-defined metabolites related to particular metabolic pathways in a biological sample) [8]. Although the definition of *metabolite* was written in several literatures with a slight difference [8–10], the term can be broadly defined as "small molecule (<1500 Da) found in a biological sample". Because true *metabolomics* is difficult, currently, the term *metabolomics* is typically used as "a large-scale study of small molecules (<1500 Da) in a biological sample", and classified into two groups, targeted- and untargeted metabolomics. Untargeted metabolomics tries to detect and identify as many compounds as possible for discovery studies, while targeted metabolomics aims to measure and quantify pre-defined metabolites with high accuracy.

Since the late 1960s, before proposing the terms *metabolome* and *metabolomics*, many analysis methods have been developed for metabolomics with the improvement of separation and detection technologies [8,11–13]. In 1971, 250 and 280 substances were detected by gas-liquid partition chromatography in a sample of breath and urine vapor, respectively [11]. It was an initial separation study; however, all detected substances were unknown compounds, and they could include many noise peaks. Three years later, amino acid profiles were measured for protein hydrolysates, plant tissue extracts, urines, and sera by gas chromatography (GC) [14]. For comparative measurements, retention index (normalized retention time) was utilized instead of absolute retention time in 1980 [15]. At almost the

same era, mass spectrometry (MS) was advanced to profile many metabolites (de Jongh et al. 1969 [16], Jellum et al. 1988 [17], and Kimura et al. 1999 [18]; see next section for detailed explanation of MS). In 2000, 326 distinct compounds were quantitatively detected using GC-MS from leaf extracts of *Arabidopsis thaliana*, and a chemical structure was assigned to about half of them [19]. Although liquid chromatography (LC) methods typically have lower chromatographic resolution than GC, they have developed and utilized well because they can measure a broad coverage of compounds without derivatization (Wolfender et al. 1998 [20] and Tolstikov et al. 2002 [21]) [22]. For metabolic profiling and metabolomics studies, eventually, GC and LC became major measurement technologies coupled with a detector, such as fluorescence, UV, and MS. In the early 2000s, GC- and LC-MS could separate and detect thousands of compounds (not only metabolites, also many contaminants) from a sample [23]. With developments of technologies, metabolomics has been advanced for measuring, identifying, and quantifying as many metabolites as possible, i.e., the number of known metabolites was increased. Indeed, the Human Metabolome Database (HMDB) originally contained 2,180 endogenous metabolites in 2007 [10], and expanded the number to 114,100 in 2018 [24], although the latter number was counted among not only known but also expected and predicted metabolites.

Besides MS, several other technologies have been utilized for metabolomics and are briefly introduced here. One of the major technologies is nuclear magnetic resonance (NMR). NMR was discovered in 1940s but began to be used in metabolic study in the early 1970s [25,26]. The advantages of NMR over MS are reproducibility and *de novo* identification of chemical structure, although it is known as a less metabolite coverage and low sensitivity [27].

1-2 Mass spectrometry-based metabolomics

Mass spectrometry (MS) is a well-known technology that measures the mass-to-charge ratio (*m/z*) of ionized molecules as mass spectrum, which is a plot of ion intensity versus *m/z*. The *m/z* value indicates the mass of an ion divided by the number of charges, and the unit of mass is Da (or u) in MS. Because MS separates isotopes, monoisotopic mass (calculated by the

mass of major isotope of each element) is used for interpretation. While the experimentally determined mass with high accuracy is called “accurate mass” (AM), the calculated mass of elemental formula is “exact mass” [28].

The performance of MS is mainly determined by mass resolution and scan speed of mass spectrum. Many vendors (AB SCIEX, Agilent, Bruker, Shimadzu, Thermo Scientific, Waters Corporation, and so on; in alphabetical order) have improved the performance and release multiple instruments with different concepts, such as Q-TOF (Quadrupole-Time Of Flight), FTICR (Fourier Transform Ion Cyclotron Resonance), and Orbitrap [29,30]. From a bit different aspect, imaging MS has been established by MALDI (Matrix Assisted Laser Desorption/Ionization) and related technologies [31–33]. Lastly, the recent breakthrough is the combination technology with ion mobility (IM) spectrometry for improving the performance of ion separation by a new separation dimension [34]. For simplification, this chapter focuses the modern common knowledge of LC-MS.

Liquid chromatography-mass spectrometry (LC-MS)

Mass spectrometer roughly comprises three major parts, ion source, mass analyzer, and detector. Analyte molecules are ionized in the ion source, separated by *m/z* in the mass analyzer, and detected as a mass spectrum in the detector part (**Figure 1-2A**). In the ion source, a molecule (M) is ionized with adducts (H, Na, K, and so on from matrix) by several methods (electrospray ionization (ESI) is the most famous in LC-MS). In general, the major adduct ion is $[M+H]^+$ and $[M-H]^-$ in positive- and negative mode, respectively. Therefore, the *m/z* value indicates the mass of an adduct ion such as $[M+H]^+$ instead of the molecular mass of M. For example, in the case of leucine ($C_6H_{13}NO_2$), the molecular weight is 131.173 but the monoisotopic mass is 131.0946, and the exact mass of the adduct ion $[M+H]^+$ is 132.1014 calculated as following: 131.094635 + 1.00728 − 0.000549 (M + H − e). In addition, the exact mass of the adduct ion of an isotope (^{13}C-leucine) $[M+H]^+$ is 133.1046, the mass of double molecule adduct ion [2M+H] is 263.1960, and the mass of double charged adduct ion $[M+2H]^{2+}$ is 66.5540 (**Table 1-1**). The mass and relative abundance of possible isotopes can be calculated based on natural abundance, and the isotopic pattern is determined by

consisting elements. However, the isotopic ions are typically classified as M+1, M+2 and so on due to the limitation of mass accuracy. By investigating the exact masses of several adducts and the isotope pattern, mass spectrum is utilized for molecular formula estimation.

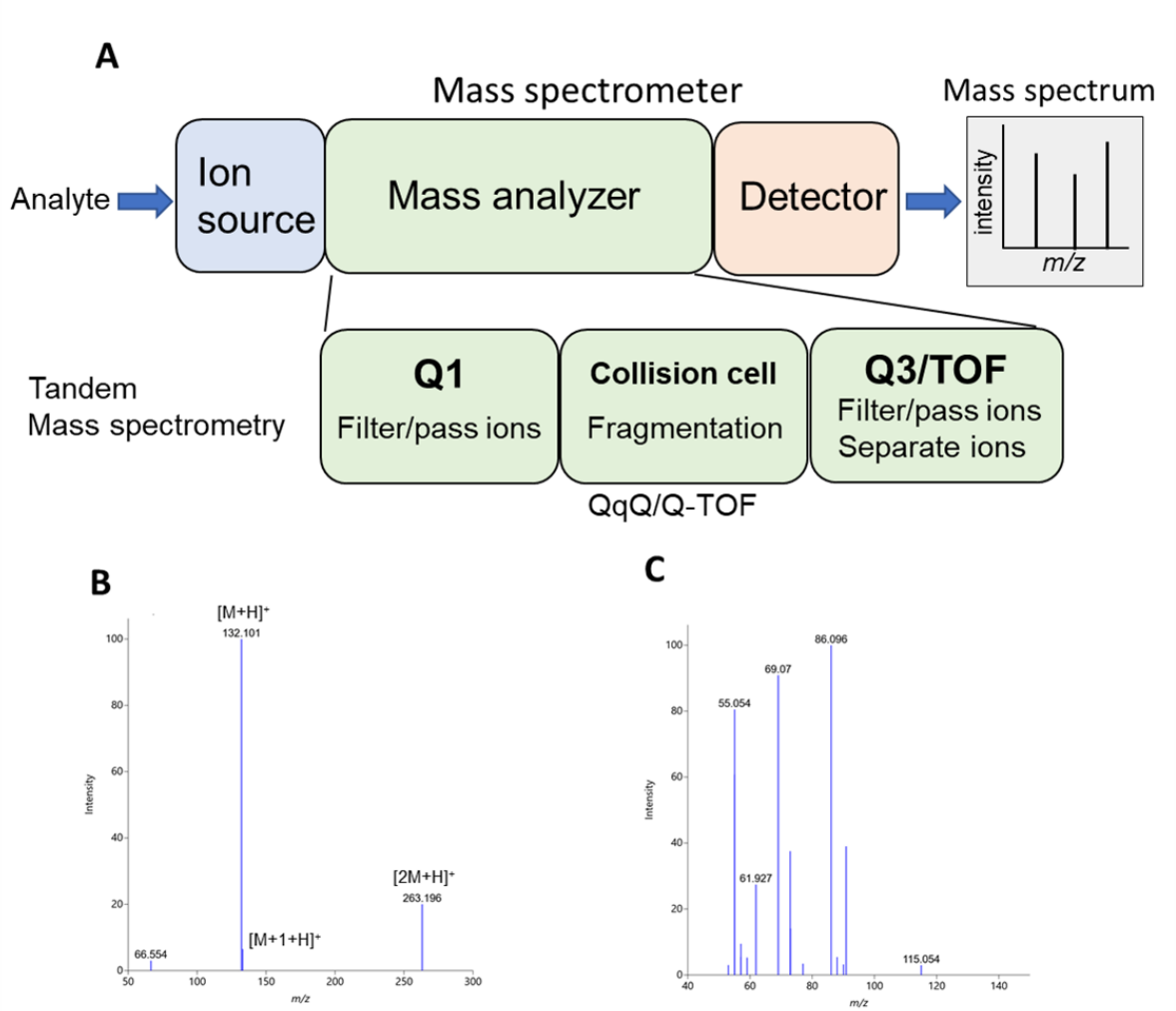


Figure 1-2. Overview of mass spectrometry (MS). **A**. Flowchart of MS and tandem mass spectrometry (MS2). Mass spectrum of leucine in MS1 (**B**) and MS2 (**C**).

Table 1-1. Mass list of $C_6H_{13}NO_2$.

Name	Value	Formula
Molecular weight	131.1732	6 * 12.01074 + 13 * 1.007941 + 14.00670 + 2 * 15.99940
Monoisotopic mass	131.0946	6 * 12.000000 + 13 * 1.007825 + 14.00307 + 2 * 15.99491
$[M+H]^+$	132.1014	131.094635 + 1.00728 - 0.000549
$[M(^{13}C)+H]^+$	133.1046	132.097850 + 1.00728 - 0.000549
$[2M+H]^+$	263.1960	2 * 131.094635 + 1.00728 - 0.000549
$[M+2H]^{2+}$	66.5540	(131.094635 + 2 * 1.00728 - 2 * 0.000549) / 2

With chromatographic technologies, mass spectra are also utilized for chromatographic peak detection by generating extracted ion chromatogram (EIC). LC-MS provides three-dimensional data consisting of *m/z*, retention time (RT), and ion intensity. In **Figure 1-3A**, thousands of ions can be separated in two-dimensions (RT and *m/z*) and detected as features from human urine data measured by my collaborators in public repository (the EMBL-EBI MetaboLights repository [35] with identifier MTBLS816). In this field, the feature consists of *m/z*, RT, ion intensity (peak height), peak area, and so on. The additional annotations of each detected feature will be assigned, such as adduct type, molecular formula, and chemical structure. As readers know, the molecular formula of leucine is the same as isoleucine and their structure is very similar; therefore, it is difficult to discriminate them. Using established methods, LC-MS can separate them in the EIC of 132.1014 *m/z* with the mass tolerance 0.001 from the same human urine data (**Figure 1-3B**). Comparing with the chemical standards measured under the same environment can reveal that left chromatographic peak is leucine and right peak is isoleucine.

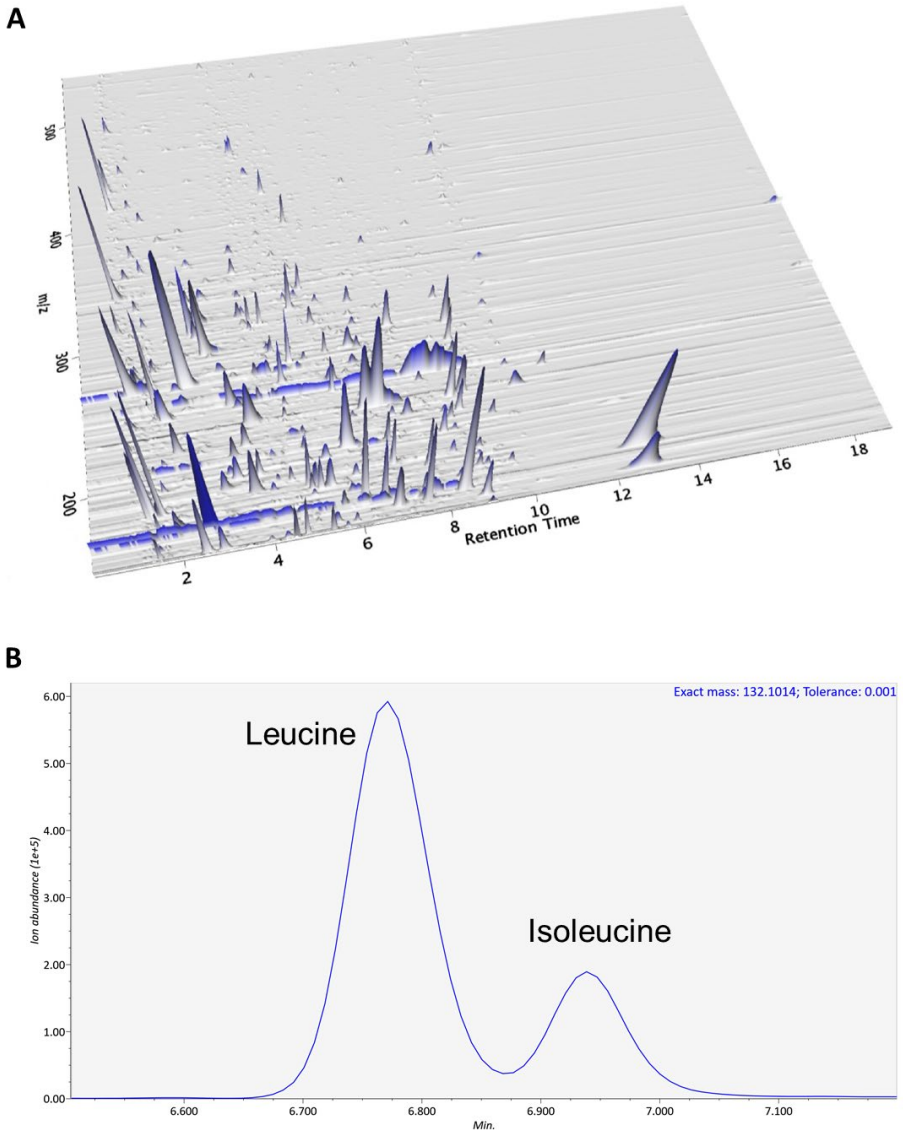


Figure 1-3. LC-MS provides three-dimensional data. A. Three-dimensional data of human urine consisting of *m/z*, retention time (RT), and intensity generated by MZmine2 [36], B. extracted ion chromatogram (EIC) of 132.1014 *m/z* with mass tolerance 0.01 Da generated by MS-DIAL [37]. Leucine and isoleucine are the left- and right peak, respectively.

Tandem mass spectrometry

Tandem mass spectrometry (MS2 or MS/MS) is a remarkable technology for chemical structure elucidation. The fragmentation pattern of each molecule is consistent under the almost same experimental settings; therefore, MS2 spectrum is utilized for compound identification and annotation in LC-MS2 measurement. A triple quadrupole MS (QqQ) consists of a collision cell sandwiched by two quadrupole mass analyzers (**Figure 1-2A**). Quadrupole mass analyzer (Q) filters ions by pre-defined value or pass all ions (called full scan). In the collision cell, analyte molecules are fragmented by collision-induced

dissociation with pre-defined collision energy typically from 0 eV (full scan) to 70 eV. A selected ion (precursor ion) in the first Q is dissociated in the collision cell, passed through the second Q, and measured as an MS2 spectrum (product ion spectrum, precisely). For more accurate analyses, Q-Time of Flight (Q-TOF) system is also widely used in general. In LC-MS2-based metabolomics, full scan without fragmentation is basic (called as MS1 scan) to detect features, also its spectrum called as MS1 spectrum. MS2 scans are additionally measured after MS1 scans with user-defined frequency due to the limited scan speed. In targeted metabolomics, pre-defined precursors are selected to measure MS2 spectra. As an example, mass spectra of leucine in MS1 and MS2 are shown in **Figure 1-2**. In contrast, in the case of untargeted metabolomics, high abundant ions are typically selected as precursor ions and measured in MS2 to give as many MS2 spectra as possible for compound identification.

Compound identification and estimation

Compound identification is a key process in metabolomics study. There are several ways and levels to identify compounds as described above, so the process should be scandalized with minimum metadata. In 2007, the minimum reporting standards and four identification levels were proposed by the Metabolomics Standard Initiative (MSI) [38]. To obtain the level-1 identification (most reliable), at least two orthogonal experimental properties of compound should match with those of an authentic standard. In LC–MS metabolomics, this criterion is often interpreted as an exact match of the peak feature in the measured sample and to the chemical standard by accurate mass (AM) and RT. However, these two properties may not be enough to reliably identify compounds due to co- or closely eluting compounds and RT fluctuations of certain chromatography techniques (e.g., HILIC). To further increase the reliability of metabolite identification, MS2 spectra are used in addition to AMRT. Recently, other identification levels have been reported [39–41], but they are not widespread yet.

For compound estimation, molecular formula prediction and chemical structure elucidation are general approaches using precursor *m/z*, isotopic pattern, and product ion spectrum (MS2 spectrum). Three properties of candidate compounds, exact mass, predicted

RT, and *in silico* MS2 spectrum, are also useful for the estimation. Exact mass can be easily calculated from molecular formula. RT can be predicted from chemical properties by machine learning methods, such as PredRet [42]. Lastly, there are several tools for *in silico* fragmentation, such as CSI:FingerID [43], MetFrag [44], and MS-FINDER [45] by machine learning or fragmentation rules. These powerful tools can support compound estimation; however, manual interpretation and confirmation are still important to avoid mis-annotations.

Difficulties of data reuse

Data reuse analysis is typically difficult using public metabolomics data by bioinformatics. In genomics, raw sequences, public repositories/databases, tools, and technical knowledge have been accumulated and matured due to the high reusability and open data culture; therefore, bioinformaticians can easily utilize them for new findings [46–49]. In fact, I reported a new visualization tool for bacterial complete genomes to compare public genomes with their consensus and highlight genome rearrangements [50,51]. Moreover, I have worked for comparative genomics of *Lactobacillus* and revealed a new species named *Lactobacillus paragasseri* from database [52,53].

While genomics is assisted with bioinformatics, metabolomics has three major unsuitable points to compare with different studies and reuse them as follows: (1) large chemical diversity, (2) low reproducible measurement, (3) insufficient standardization. First, metabolome consists of large number of compounds with different chemical property. Thus, the perfect measurement methods cannot exist. Many measurement methods have been established based on study purposes and designs, and reported compounds are different by studies. Second, LC-MS and LC-MS2 are very sensitive but less reproducible. Even if the same measurement methods, the result will be slightly differed by laboratory, experimenter, and batch, although there are many normalization and calibration methods. Lastly, the characteristic of raw measurement data is machine-dependent; moreover, the file formats of raw data are also different by venders. Because metabolomics is also a field of analytical chemistry, measurement methods tend to be specialized and sophisticated with improved instruments in each laboratory, instead they are standardized and widespread worldwide.

Data processing is also difficult to be standardized. In LC-MS, detected chromatographic peaks depends on the parameters and manual curation by researchers. Compound identification also requires manual curation by expert as mentioned in the previous subsection. These issues are also major challenges in untargeted metabolomics study.

In the era of open data, data sharing and further analysis are necessary in metabolomics [54,55]. To support and enhance metabolomics studies, public and commercial databases have been developed and maintained. PubChem [56] and ChemSpider [57] are comprehensive chemical databases, HMDB [24] and KNApSAcK [58] are curated metabolite databases, MassBank [59], METLIN [60], GNPS [61], and NIST [62] are well known MS2 spectral databases. Moreover, several raw data repositories/databases are also developed. MetabolomeXchange [63] and Metabolonote [64] are measurement metadata databases, MetaboLights [35], Metabolomics Workbench [65], and Food Metabolome Repository [66,67] are raw measurement data repositories with standardized metadata. Reused study of public raw data in above databases has been anticipated.

1-3 Data-Independent Acquisition (DIA)

MS2 spectrum is very useful for compound identification. In the previous section, MS2 acquisition was roughly explained as “a selected ion (precursor ion) in the first Q is dissociated in the collision cell, passed through the second Q, and measured as an MS2 spectrum”. More precisely, the narrow *m/z* range (typically 1 Da) is used in Q1 for filtering, and passed ions are fragmented to acquire the corresponding MS2 spectrum. In general, MS2 spectra are sequentially acquired from highly abundant MS1 ions, because the number of MS2 spectra are limited by the scan speed. Therefore, this classic measurement method is called as data-dependent acquisition (DDA) suffering from the low MS2 spectral coverage. For more comprehensive measurements, data-independent acquisition (DIA) has been developed to acquire all MS2 spectra from all co-eluting compounds.

In DDA method, MS2 spectra are relatively clean and easily usable for compound identification because the link between precursor ion and its product ions remain, while the

limitation of the MS2 spectral coverage restricts further analysis for untargeted metabolomics. In contrast, DIA method can acquire all MS2 spectra from all co-eluting compounds; however, the link between precursor and product ions is missing, i.e. its MS2 spectra are very complicated and necessary to be deconvoluted for compound identification. The difference of DDA and DIA was reviewed by Xiaochun Zhu et al. in 2013 [68] and Ruohong Wang in 2019 [69].

Two major DIA methods, AIF (All-Ion Fragmentation) and SWATH (Sequential Window Acquisition of all THeoretical fragmentation ion spectra), are briefly introduced. AIF is the simplest DIA method for acquiring full MS2 scan after full MS1 scan without precursor selection. Therefore, AIF MS2 spectra are highly complicated and consists of all co-eluting compounds and measurement noise. In SWATH, precursor ions are selected to acquire MS2 spectra by the sequential *m/z* window (typically 25 Da) after full MS1 scan. Thus, SWATH MS2 spectra from each 25 Da window are cleaner than AIF spectra from all *m/z* range (Figure 1-4). Even SWATH MS2 spectra require MS2 deconvolution for compound identification. Various deconvolution tools have been developed for SWATH, but high-performance tools for AIF spectra are still anticipated.

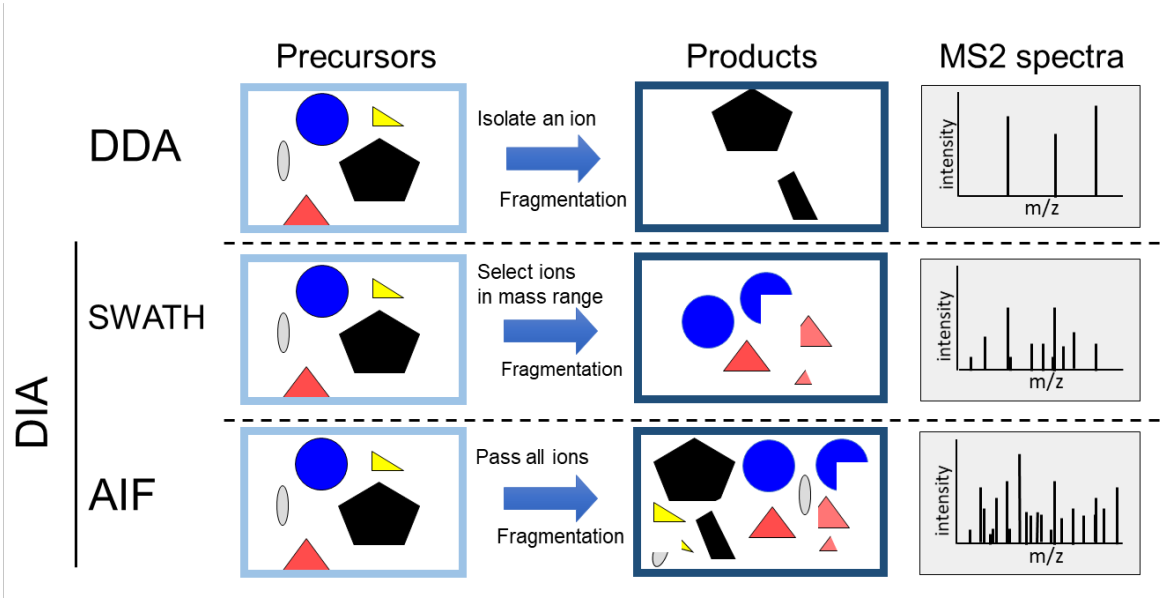


Figure 1-4. The difference of DDA and DIA (SWATH and AIF).

1-4 Thesis Outline

In this doctoral thesis, I will report my works related to AIF-MS for reusable untargeted metabolomics. For further development in metabolomics, I think that reusable data acquisition (such as AIF), reliable compound identification, and universal and integrated data analysis platform are required. The major disadvantages of AIF are the complex MS2 spectra and lack of good analysis platform. To generate clean MS2 spectra from AIF, I developed a new MS2 deconvolution method, named CorrDec (Correlation-based Deconvolution). For improving the accuracy of compound identification, reliable chemical standard library was created using AIF-MS and CorrDec. Lastly, a data analysis platform for AIF was developed and released in public.

In Chapter 2, Correlation-based Deconvolution (CorrDec) method will be demonstrated using dilution series and human urine dataset. CorrDec is a new method for MS2 deconvolution to generate clean MS2 spectra from complex AIF spectra. The advantage and limitation are discussed by comparing with the previous method (MS2Dec). By random resampling analysis, the minimum number of samples is roughly estimated for CorrDec. I also describe the usability of both MS2Dec and CorrDec based on different concepts.

In Chapter 3, creating reliable chemical library consisting of AM, RT, and MS2 spectra will be reported. For accurate compound identification, a reliable chemical library is required. I provide practical recommendations for library development and mention the benefit and necessity of open data in metabolomics.

In Chapter 4, the AIF platform consisting of three metabolomics tools, MS-DIAL, MS-FINDER, and MS-LIMA, which are improved and developed by the author will be introduced. To analyze complex AIF-based metabolomics data, various functions are implemented in MS-DIAL and MS-FINDER. MS-LIMA has been developed by the author to manage MS2 spectra. The usability of the AIF platform is also summarized.

In Chapter 5, I will summarize the new data analysis platform including CorrDec, reliable library, and useful software, for AIF-based untargeted metabolomics. Future perspective and works are also described.

Note

In this chapter, I briefly summarized the history of metabolomics and fundamental knowledge related to AIF-MS. Through this thesis, I tried to avoid explaining the principles of measurement technologies and techniques to clarify the scope and be easy for understanding the context. Readers who are interested in more details about the history, measurement technologies, and analysis methods are referred reviews and textbooks [7,9,23,68,70–74].

Chapter 2

Correlation-Based Deconvolution

For complex AIF MS2 spectra, a new Correlation-based Deconvolution method, namely CorrDec, was developed. CorrDec utilizes intensity variations among samples and calculates intensity correlations between precursor peak and its MS2 peaks (**Figure 2-1**).

The contents of this chapter are also described in Tada et al. *Analytical Chemistry* 2020 [75].

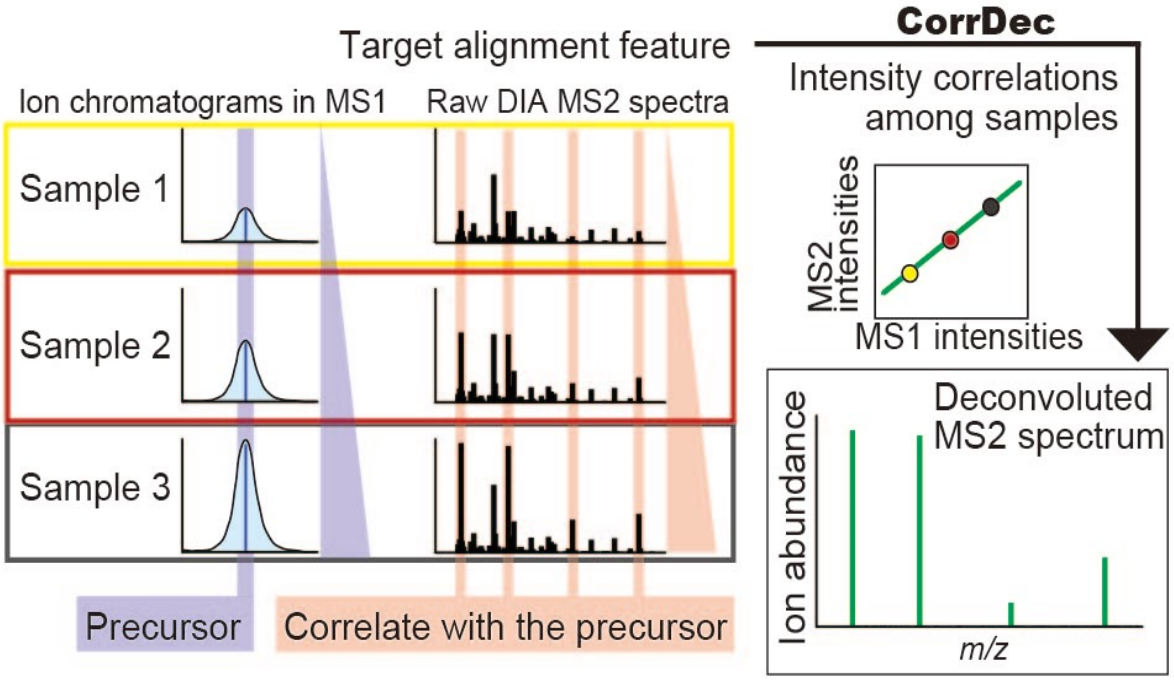


Figure 2-1. Graphical abstract of CorrDec

2-1 Introduction

High-resolution tandem mass spectra (MS2) in combination with public mass spectra and the associated computational tools are considered indispensable for compound identification. A number of resources are now available including MassBank [59], GNPS[61], CSI:FingerID [43] and MS-FINDER [76]. In the classical data-dependent acquisition mass spectrometry (DDA-MS), ions are isolated in a narrow window (typically 1 Da) of the target *m/z* value. In contrast, for the data-independent acquisition mass spectrometry (DIA-MS), wider *m/z* windows of 10 to 1000 Da are used to obtain highly complex mixture spectra that require computational approaches to interpret.

To overcome the trade-off between comprehensiveness and cleanness of spectra, various deconvolution tools have been proposed for DIA data, such as OpenSWATH [77], Specter [78], MetDIA [79], decoMetDIA [80], and MS-DIAL [37]. OpenSWATH, Specter, and MetDIA were designed for the targeted analyses utilizing predefined spectral libraries to deconvolute spectra. MS-DIAL (MS2Dec) and decoMetDIA can deconvolute MS2 spectra *de novo* by fitting MS2 chromatograms to their precursor chromatogram. These powerful methods are suitable for the SWATH (Sequential Window Acquisition of all THeoretical fragment ion spectra) type of DIA data [68]. However, MS2 spectra become highly complex when precursor ions of all *m/z* are fragmented together (e.g. all ion fragmentation (AIF), MS^{All}, or MS^{E}). Especially busy chromatographic regions with multiple co-eluting compounds pose a challenge. In the case of MS2Dec, at least two data-point difference between the liquid chromatographic peak tops is required for deconvolution. Therefore, the MS2Dec and decoMetDIA are not suitable for untangling complex MS2 spectra from the AIF acquisition and its equivalent.

Correlation has been widely used in mass spectrometry-based metabolomics [81,82]. For example, the Pearson correlation is used in CAMERA to estimate the similarity of different mass chromatograms to extract compound spectra, and to annotate adduct ions and isotopic peaks [83]. For DIA data correlation-based approaches such as RAMClust can be used to assign precursor-product relationships based on detected features in MS1 and MS2

[84,85]. Herein, I present a new MS2 deconvolution method based on the correlation of ion abundances between precursor- and product ions among biological samples, named CorrDec (Correlation-based Deconvolution). The CorrDec method is based on three assumptions: (1) metabolite concentrations differ across study samples in multi-sample studies, (2) the MS2 fragmentation pattern is identical under identical experimental conditions, and (3) intensities of fragment ions correlate with those of their precursors. CorrDec (implemented in MS-DIAL version 3.22 and later) is designed to generate MS2 spectra using untargeted multi-sample AIF metabolomics data and does not require a pre-defined spectral library.

In this study, I utilized the idea for MS2 deconvolution to purify the DIA spectra. In contrast to the previous approaches, CorrDec does not perform the feature detection procedure for MS2 chromatograms to retrieve as many characteristic product ions as possible from the DIA-MS2 spectra; instead, the noisy spectra are effectively excluded by integrating multi-samples profile data. I demonstrate the concept and utility of CorrDec in dilution series of chemical standards in urine and a case study from a urinary metabolomics cohort.

2-2 Results and Discussion

CorrDec workflow

CorrDec starts with the aligned peak list from multiple samples. The peak list consists of 'aligned features', which include the averages of RT, *m/z*, peak height and width obtained from the detected peaks in the samples, their ion abundances, and corresponding DIA MS2 spectra. The peak height is used for the quantification of MS1 and MS2 peaks. The MS2 deconvolution is performed as follows.

(Step 1) For each aligned feature Ft1, Pearson correlations are calculated between all product ions and their precursors. The MS2 spectra of Ft1 for all samples are retrieved to create a "MS2Mat" data matrix, consisting of the ion abundances of each product ion (P) binned by an *m/z* threshold (0.01 in this study) in all samples. The precursor ion abundances of all

samples are retrieved to create a "MS1Vec" data vector, and Pearson correlations are calculated for all pairs of the features in MS1Vec and product ions in MS2Mat (**Figure 2-2A**). For each product ion, its existence ratio within the samples (the number of samples having the product ion above the threshold value (1000 in this study) divided by the number of all samples) is also recorded.

(Step 2) All correlation values in all features are joined into a matrix based on the *m/z* of the product ion using the same *m/z* threshold (0.01 in this study) as MS2Mat (**Figure 2-2B**).

(Step 3) Each product ion is assessed using the correlation value $Corr_{MS1vsMS2}$ for its inclusion to the deconvoluted spectrum of Ft1. Three criteria are applied (**Figure 2-2C**):

(Criterion 1) $Corr_{MS1vsMS2}$ > minimum threshold,

(Criterion 2) $Corr_{MS1vsMS2} > MaxCorr_{Ft} - margin1$, and

(Criterion 3) $Corr_{MS1vsMS2} > MaxCorr_{P} - margin2$.

Criterion 1 is a simple cut-off threshold to suppress noise signals; correlations between the ion abundances of a MS1 precursor ion and the ion abundances of a MS2 product ion should be higher than a predefined minimum correlation threshold. In this study, I use 0.7 as a threshold (recommended range is between 0.3 and 0.7, the lower the threshold the higher the possibility to introduce noise peaks into the spectra). Criterion 2 is a threshold to judge whether each product ion is eligible as a fragment derived from the feature Ft1 (e.g. for removing lower correlating peaks due to ionization enhancement and/or biological correlation between compounds). The maximum of all correlations for each MS1 feature (Ft1), $MaxCorr_{Ft}$, is used for calculating the cut-off. Product ions with correlations smaller than $MaxCorr_{Ft} - margin1$ are excluded. In this study I use 0.2 as *margin1* (recommended range is between 0.1 and 0.3, the larger the *margin* the higher the possibility of including noise peaks into the spectra). For example, in **Figure 2-2B**, the MS2 peak P1 (0.73) is removed because $MaxCorr_{Ft}$ for the Ft1 feature is 0.98. Criterion 3 is used to avoid false-positive assignments by Criterion 2 when the same product ion shows high correlation values for multiple precursor ions. For each product ion P*x* a maximum correlation $MaxCorr_{P}$ with its

neighboring features (eluting within ± 0.5 * peak width of Ft1) is determined. When the correlation value between the Ft1 and P*x* is less than $MaxCorr_P$ − *margin2* (0.1 in this study, recommended range is between 0.1 and 0.3, the larger the *margin2* the higher the possibility of including noise peaks into the spectra), P*x* is excluded from the deconvoluted spectrum of Ft1. For example, the product ion P2 is excluded from the Ft1 deconvoluted spectrum because the value of 0.81 is less than $MaxCorr_P$ (0.93) − 0.1 (**Figure 2-2B**). These threshold values are strict; they may require tuning when applied to different datasets. The *m/z* value and the intensity in a deconvoluted spectrum are represented by their respective median value of *m/z* and intensities in biological samples, where the intensities are normalized by the abundance of the precursor ion in each sample.

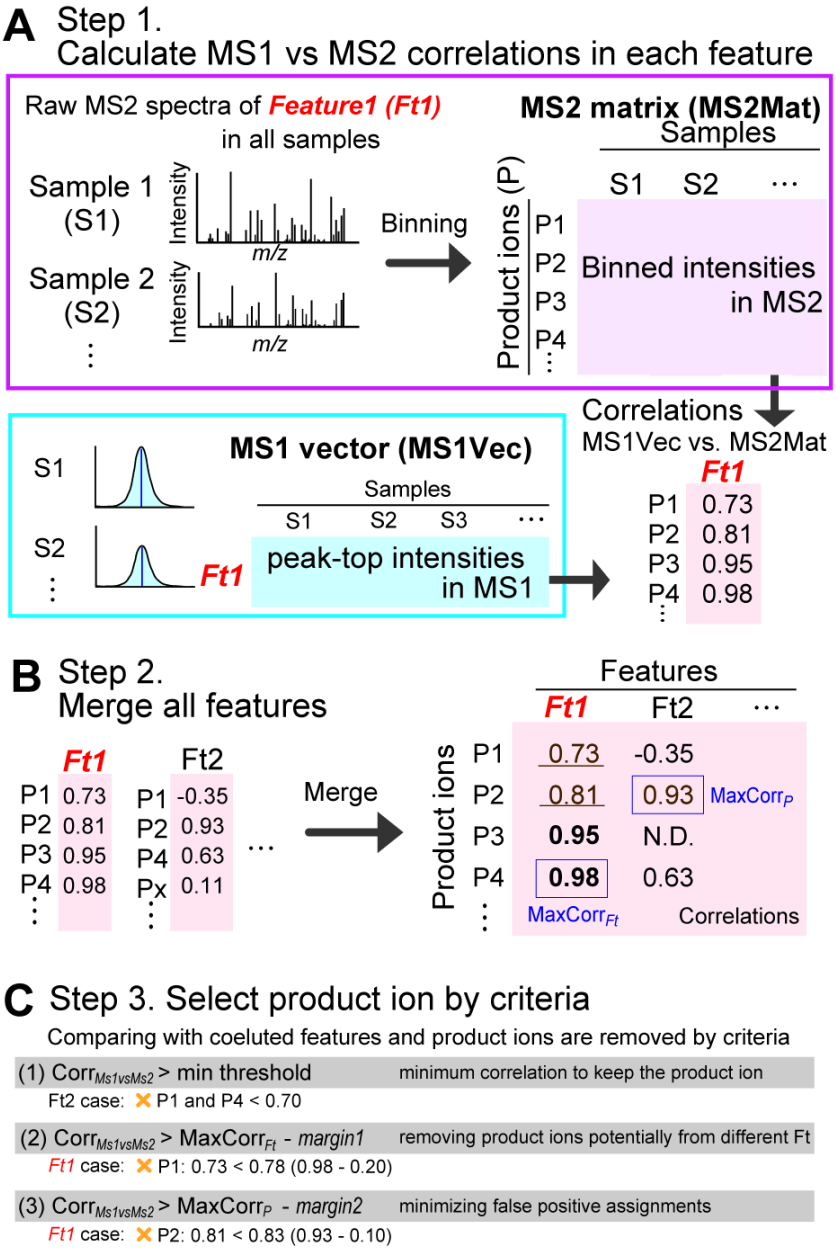


Figure 2-2. Flowchart of the CorrDec method for a target feature Ft1. **A.** For each feature, the Pearson correlations are calculated for all pairs of precursor (MS1 vector) and product ion (MS2 matrix). **B.** All correlation values of all features are merged into a single matrix. **C.** Product ions satisfying the three criteria (see the main text for details) are selected to produce the deconvoluted MS2 spectrum of Ft1.

Serial dilution study

Using a dilution series of chemical standards, I tested whether the intensities of MS2 fragments were highly correlated with those of their precursors. The 11-point dilution series of 8 chemical standards were measured by AIF mode (see Methods) with diluted urine as the matrix. In such a setup, only the concentrations of the spiked compound vary while concentrations of compounds in the urine matrix remain stable. In **Figure 2-3**, I show the result of the tyrosine standard. The tyrosine dilution series was partially masked by the endogenous tyrosine present in the matrix (diluted urine). In AIF mode, the MS2 spectra of tyrosine contained 193 and 280 peaks in the 11-point dilution series for 10 and 30 eV, respectively. The similarity scores (simple dot product) of all raw MS2 spectra with the reference spectra were less than 30%. When processed by the CorrDec, 12 and 13 peaks, respectively, showed >0.9 correlations with their precursors, clearly deviating from the normal distribution formed by the correlation values of the other peaks (**Figure 2-3B** bottom). These highly correlated peaks exhibited intensities proportional to the dilution (**Figure 2-3B** top in the log scale), and the MS2 similarity scores with the reference spectra were 90.5% and 86.5% for 10 and 30 eV, respectively.

Similar results were reproduced for the other 7 compounds; MS2 spectra were successfully generated with high MS2 matches (1 compound >80%, other 6 compounds > 90% at least one collision energy) by the CorrDec (**Table 2-1**). In addition to the MS2 spectra at 10 and 30 eV, deconvoluted spectra were obtained for 0 eV resulting from in-source fragmentation which can be used for metabolite identification [84–87]. The degree of in-source fragmentation depends on the ionization source settings and in this study in-source fragmentation is facilitated by rather high fragmentor voltage (380 V). In this setup, the MS2 similarity of the 0 eV spectra was comparable to the 10 and 30 eV spectra, corroborating the usability of in-source fragmentation data.

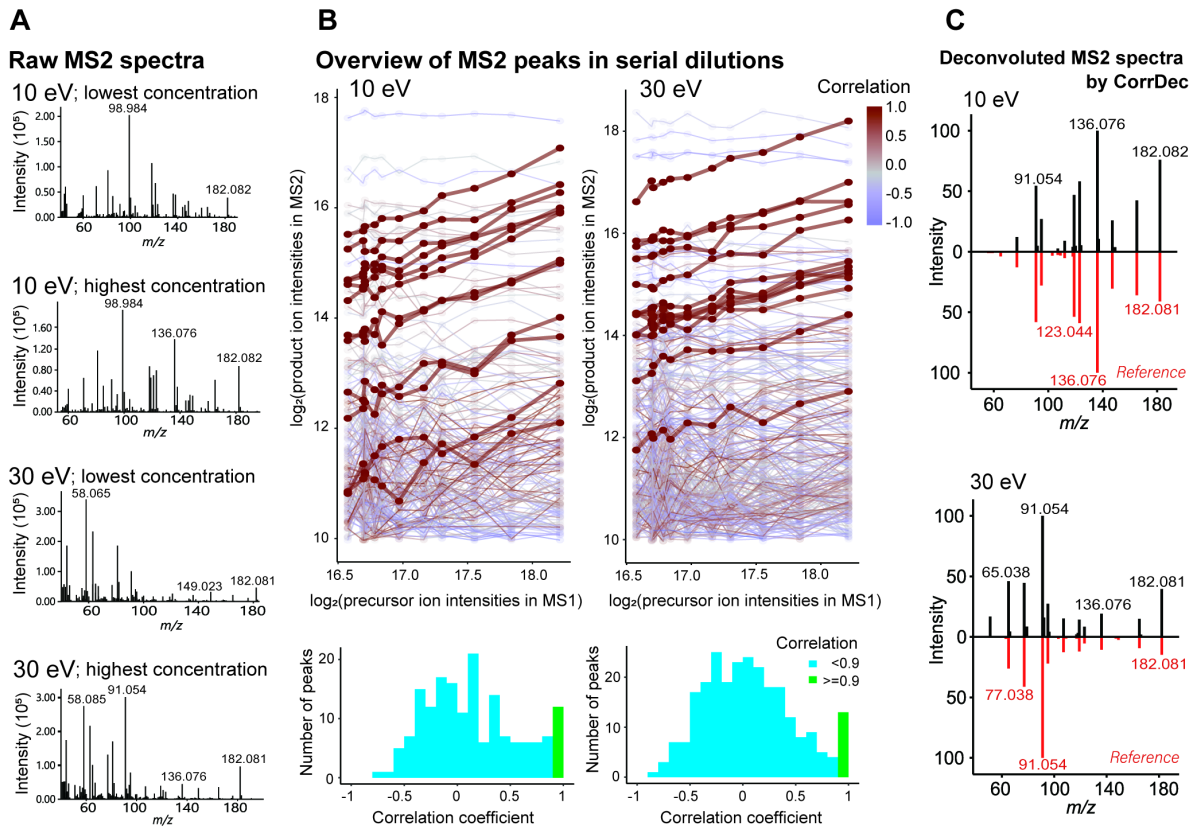


Figure 2-3. Demonstration of the CorrDec method using tyrosine dilution series spiked into diluted urine as background matrix. **A.** Raw MS2 spectra of tyrosine $[M+H]^+$ (*m/z*: 182.082) at the lowest (69 nM) and the highest (4 μM) spiked concentrations in dilution series. Raw MS2 spectra contain over one hundred peaks masking the ions derived from tyrosine, especially at low spiked-in concentrations. **B.** Linked scatter plots visualizing the intensity correlations between the MS1 *m/z* 182.082 and MS2 peaks in 11 dilution series samples. Only 12 out of 193 (10 eV) and 13 out of 280 peaks (30 eV) correlated >0.9 (highlighted lines). **C.** Deconvoluted MS2 spectra (above, in black) matched well with the library reference spectra (below, in red). The MS2 similarities of deconvoluted spectra were 90.5% (10 eV) and 86.5% (30 eV), while the MS2 similarities of raw spectra at 0, 10, and 30 eV were less than 30% in the all samples.

Table 2-1. MS2 similarity scores for the CorrDec deconvoluted spectra of chemical standards

					Similarity score (%)		
ID	RT	*m/z*	Adduct	Metabolite name	0 eV	10 eV	30 eV
1	7.37	144.1053	$[M+H]^+$	Proline betaine	92.9	92.9	82.9
2	7.47	138.0582	$[M+H]^+$	Trigonelline	94.5	96.8	88.8
3	7.81	104.0710	$[M+H]^+$	Dimethylglycine	95.1	99.2	69.5
4	7.01	76.0762	$[M+H]^+$	Trimethylamine N-oxide	93.3	98.3	91.8
5	7.62	182.0815	$[M+H]^+$	Tyrosine	89.6	90.5	86.5
6	7.56	118.0881	$[M+H]^+$	Betaine	94.5	95.5	94.1
7	7.74	116.0710	$[M+H]^+$	Proline	93.8	92.2	84.3
8	7.43	225.0872	$[M+H]^+$	3-Hydroxykynurenine	86.7	80.9	81.5

To further confirm the usability of CorrDec for metabolomics studies where the concentration of the compounds varies only little between the samples measured 1.07-fold dilution series of tyrosine in diluted urine matrix. CorrDec successfully generated MS2 spectra showing >80% MS2 match at 10 eV using only 4 samples (3.05-4.00 μM). Therefore, only small concentration changes between the samples (<25%) can be enough for the correlation-based methods which is consistent with the previous report [84]. Based on the results of chemical standard dilution series, the CorrDec was applied to a metabolomics study.

Urine cohort study

As proof of concept for the performance of CorrDec, I analyzed a LC-MS (HILIC chromatography) metabolomics dataset consisting of 224 unique urine samples, 58 pooled QCs and 4 blanks acquired in positive ionization AIF mode. Data was processed by MS-DIAL version 4.12. A total of 4159 features were aligned; the alignment of 64 features was manually curated. In the CorrDec deconvolution process, I discarded product ions that appear in <50% of all samples for computational efficiency. This threshold of 50% is arbitrary and should be set for each study considering the sample number and the desired level of reliability. After manual curation, 105 compounds were confidently identified at the MSI level 1 [38] by matching AM, RT and MS2 spectra to the reference library.

For all of the 105 compounds, MS2Dec and CorrDec were able to generate MS2 spectra. However, while for MS2Dec only 34 compounds showed >80% match to the reference spectra, 85 of the 105 compounds CorrDec spectra could achieve >80% match (**Figure 2-4A**). Furthermore, the distribution of MS2 similarity scores (**Figure 2-4B**) for the two deconvolution methods shows that MS2Dec spectra for 50 compounds had <60% match. Median similarity values were 59.1% and 91.3% for MS2Dec and CorrDec, respectively. The reason for the disparity is that CorrDec is especially effective in obtaining cleaner spectra for compounds of low abundance or smaller peak intensity (**Figure 2-4C**), which are often the majority in the metabolomics datasets and generally challenging to identify [72].

In addition to the 105 compounds identified at the AMRT and MS2 match level, I could identify six metabolites that highly matched (>80%) to the applied MS2 library using CorrDec spectra, but not with MS2Dec spectra, and have been previously reported to be detected in human urine (imidazole acetic acid [88], homocitrulline [89], aminohippuric acid [90], isobutyryl (C4) carnitine [90], liquiritigenin [91], and AICA-riboside [92]). Among the 111 identified compounds over half (61) are amino acids and their metabolites (standard amino acids (13), methylated (9), acetylated (6), other amino acid metabolites (22), conjugates (11)). The other major compound groups include products of nucleic acid metabolism (13), food and drug metabolites (8). I summarized identified compounds and their mass spectral similarities in **Table 2-2**.

In addition to the MS2 library matching, CorrDec can provide more reliable MS spectra than MS2Dec for annotation tools such as MS-FINDER [93]. For example, I could annotate two features based on their CorrDec MS2 spectra as acetaminophen-sulfate and valerylcarnitine, two compounds not present in the used MS2 spectral library, but likely to present in urine [90]. At the early stages of this study, from the currently 85 AMRT+MS2 confirmed compounds (**Figure 2-4A**), 25 compounds were annotated by MS2 match only and were purchased for confirmation. MS2 spectra of compounds with variable levels in the samples, for example metabolites of drugs and dietary components, are particularly suitable for deconvolution by CorrDec and subsequent structural interpretation. Therefore, CorrDec spectra are valuable for the identification using spectral libraries and as well as *in silico* annotations.

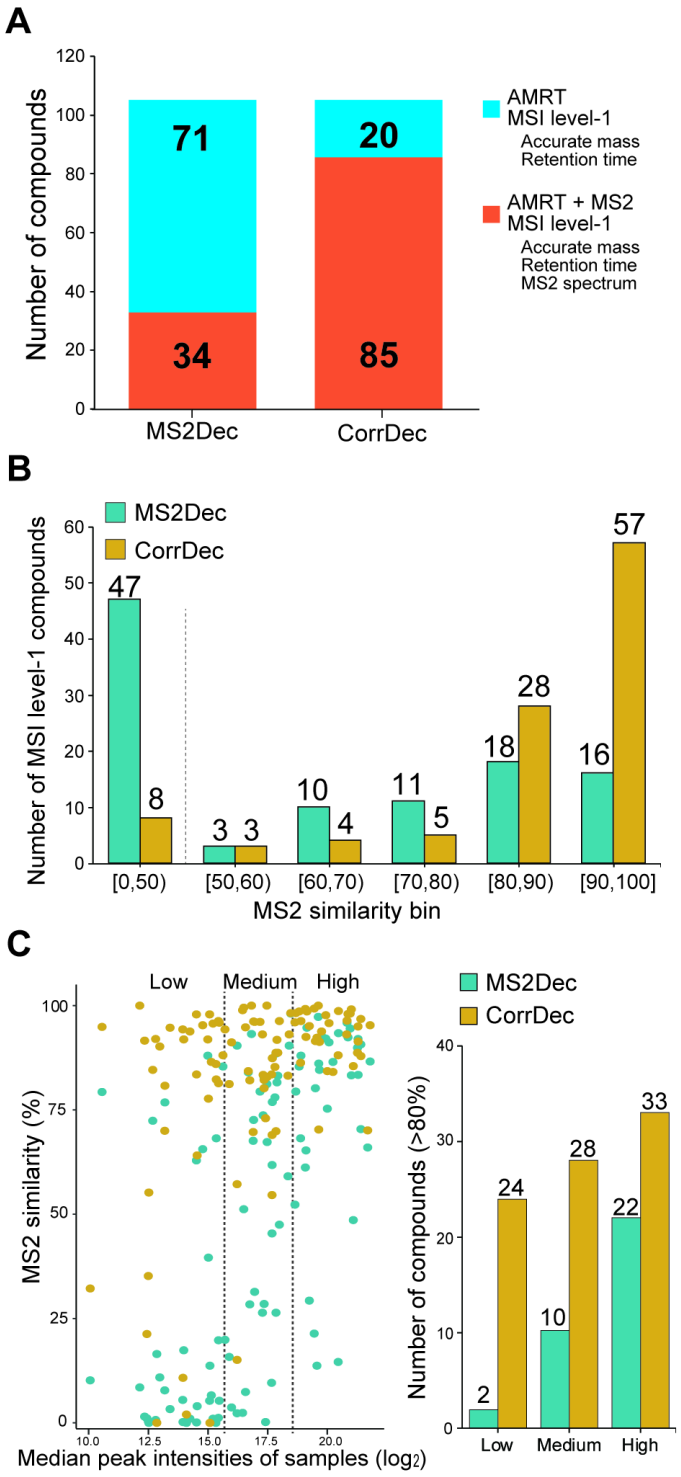


Figure 2-4. CorrDec MS2 spectra provide more confidence in compound identification than those obtained by MS2Dec in the urinary metabolomics DIA dataset. **A.** Number of compounds in each identification category identified using MS2Dec and CorrDec. **B.** Distribution of the MS2 similarity scores for the MSI level-1 compounds spectra deconvoluted by the CorrDec and MS2Dec. **C.** MS2 similarity scores from CorrDec were higher than MS2Dec, especially for low-intensity peaks.

Table 2-2. 111 identified and annotated compounds

Average RT	Average *m/z*	Adduct	Metabolite name	Identification rank	MS2 similarity score	
					MS2Dec	CorrDec
7.43	76.080	$[M+H]^+$	Trimethylamine N-oxide	AMRT+MS2	92.4	98.0
2.02	100.079	$[M+H]^+$	2-Piperidone	AMRT+MS2	13.7	92.3
7.81	104.072	$[M+H]^+$	Dimethylglycine	AMRT+MS2	97.3	100.0
8.50	104.109	$[M+H]^+$	Choline	AMRT+MS2	86.1	70.3
4.82	112.042	$[M+H]^+$	2,3-Dihydroxypyridine	AMRT+MS2	76.9	87.4
8.74	112.050	$[M+H]^+$	Cytosine	AMRT+MS2	85.4	88.1
2.52	113.035	$[M+H]^+$	Uracil	AMRT+MS2	88.0	91.9
8.62	114.069	$[M+H]^+$	Creatinine	AMRT+MS2	86.6	95.3
7.74	116.072	$[M+H]^+$	Proline	AMRT+MS2	93.2	100.0
9.05	118.066	$[M+H]^+$	Glycocyamine	AMRT+MS2	83.4	91.4
7.60	118.094	$[M+H]^+$	Betaine	AMRT+MS2	48.6	85.5
8.65	120.067	$[M+H]^+$	Threonine	AMRT+MS2	81.6	91.8
7.35	121.065	$[M+H-NH3]^+$	Tyramine	AMRT+MS2	51.2	99.5
7.72	126.024	$[M+H]^+$	Taurine	AMRT+MS2	75.3	84.3
2.50	130.051	$[M+H]^+$	Pyroglutamic acid	AMRT+MS2	80.2	91.5
7.47	130.088	$[M+H]^+$	Pipecolic acid	AMRT+MS2	81.2	100.0
9.12	131.118	$[M+H]^+$	N-Acetylputrescine	AMRT+MS2	82.8	80.2
8.28	132.085	$[M+H]^+$	Creatine	AMRT+MS2	90.7	96.8
6.72	132.102	$[M+H]^+$	Isoleucine	AMRT+MS2	79.4	96.1
6.72	132.103	$[M+H]^+$	Leucine	AMRT+MS2	98.1	98.2
9.09	133.061	$[M+H]^+$	Asparagine	AMRT+MS2	28.4	82.1
4.97	137.052	$[M+H]^+$	Hypoxanthine	AMRT+MS2	14.6	88.1
8.75	137.060	$[M+H]^+$	Dopamine	AMRT+MS2	5.5	94.3
7.51	138.066	$[M+H]^+$	Trigonelline	AMRT+MS2	70.4	87.9
7.64	143.084	$[M+H]^+$	Ectoine	AMRT+MS2	84.0	84.3
7.95	146.095	$[M+H]^+$	4-Guanidinobutanoic acid	AMRT+MS2	68.2	86.3
8.93	147.080	$[M+H]^+$	Glutamine	AMRT+MS2	86.5	84.1
13.03	147.113	$[M+H]^+$	Lysine	AMRT+MS2	79.3	94.9
7.61	152.057	$[M+H]^+$	Guanine	AMRT+MS2	1.2	96.3
1.52	152.074	$[M+H]^+$	Acetaminophen	AMRT+MS2	68.2	82.3
4.75	153.043	$[M+H]^+$	Xanthine	AMRT+MS2	61.2	99.2
1.72	154.050	$[M+H]^+$	3-Hydroxyanthranilic acid	AMRT+MS2	5.3	97.9
12.56	156.082	$[M+H]^+$	Histidine	AMRT+MS2	92.1	97.7
6.28	159.052	$[M+H]^+$	Allantoin	AMRT+MS2	6.6	86.5
8.31	162.120	$[M+H]^+$	Carnitine	AMRT+MS2	92.0	95.0
7.51	166.074	$[M+H]^+$	7-Methylguanine	AMRT+MS2	52.3	96.0
6.55	166.091	$[M+H]^+$	Phenylalanine	AMRT+MS2	91.2	95.8
6.16	169.042	$[M+H]^+$	Uric acid	AMRT+MS2	94.5	98.1
1.56	170.048	$[M+H]^+$	2-Furoylglycine	AMRT+MS2	67.3	82.9
12.60	170.105	$[M+H]^+$	3-Methylhistidine	AMRT+MS2	95.9	96.4
4.20	176.056	$[M+H]^+$	N-Acetylaspartic acid	AMRT+MS2	3.7	91.2
9.34	176.104	$[M+H]^+$	Citrulline	AMRT+MS2	62.9	97.9
1.44	180.070	$[M+H]^+$	Hippuric acid	AMRT+MS2	93.0	93.1

1.62	181.082	$[M+H]^+$	Dimethylxanthine	AMRT+MS2	83.3	99.1
8.06	182.049	$[M+H]^+$	Methionine sulfone	AMRT+MS2	19.8	81.4
7.73	182.084	$[M+H]^+$	Tyrosine	AMRT+MS2	84.6	91.3
4.95	183.055	$[M+H]^+$	1-Methyluric acid	AMRT+MS2	92.7	93.8
4.13	184.062	$[M+H]^+$	4-Pyridoxic acid	AMRT+MS2	73.7	83.4
8.36	189.124	$[M+H]^+$	N6-Acetyllysine	AMRT+MS2	47.5	96.3
8.96	189.124	$[M+H]^+$	N2-Acetyllysine	AMRT+MS2	28.5	82.2
3.07	190.051	$[M+H]^+$	Kynurenic acid	AMRT+MS2	90.4	93.1
3.88	190.073	$[M+H]^+$	N-Acetylglutamate	AMRT+MS2	15.8	81.2
1.24	194.082	$[M+H]^+$	Methylhippuric acid	AMRT+MS2	1.0	93.8
1.31	195.096	$[M+H]^+$	Caffeine	AMRT+MS2	79.4	98.2
1.41	196.061	$[M+H]^+$	o-Hydroxyhippuric acid	AMRT+MS2	17.4	91.9
6.33	196.065	$[M+H]^+$	MES	AMRT+MS2	59.1	83.2
3.89	197.070	$[M+H]^+$	1,7-Dimethylurate	AMRT+MS2	85.4	98.6
8.66	198.088	$[M+H]^+$	N,N-Acetylhistidine	AMRT+MS2	7.4	94.8
5.44	199.084	$[M+H]^+$	5-Acetylamino-6-amino-3-methyluracil	AMRT+MS2	83.2	88.7
7.10	201.984	$[M+H]^+$	Cysteine-S-sulfate	AMRT+MS2	0.1	83.5
12.41	203.151	$[M+H]^+$	Dimethylarginine	AMRT+MS2	90.4	15.1
5.77	204.128	$[M+H]^+$	Acetylcarnitine	AMRT+MS2	93.4	98.7
6.94	205.100	$[M+H]^+$	Tryptophan	AMRT+MS2	88.0	96.3
4.24	206.046	$[M+H]^+$	Xanthurenic acid	AMRT+MS2	26.4	93.1
1.36	206.082	$[M+H]^+$	N-Cinnamoylglycine	AMRT+MS2	19.9	94.3
4.98	208.101	$[M+H]^+$	CHES	AMRT+MS2	26.4	88.7
6.71	209.094	$[M+H]^+$	Kynurenine	AMRT+MS2	0.0	86.0
1.70	211.084	$[M+H]^+$	1,3,7-Trimethyluric acid	AMRT+MS2	1.0	95.7
8.73	217.132	$[M+H]^+$	N-Acetylarginine	AMRT+MS2	65.3	92.5
5.02	218.115	$[M+H]^+$	Acetylcitrulline	AMRT+MS2	72.6	96.2
4.69	218.140	$[M+H]^+$	Propionyl-carnitine	AMRT+MS2	2.4	98.9
2.13	220.120	$[M+H]^+$	Pantothenic acid	AMRT+MS2	94.7	98.6
6.22	222.099	$[M+H]^+$	N-Acetyl-D-glucosamine	AMRT+MS2	72.4	84.6
7.57	225.087	$[M+H]^+$	3-Hydroxykynurenine	AMRT+MS2	0.7	92.0
8.88	229.123	$[M+H]^+$	Proline-hydroxyproline	AMRT+MS2	21.4	99.3
6.53	238.095	$[M+H]^+$	Biopterin	AMRT+MS2	5.3	95.9
11.05	239.107	$[M+H]^+$	HEPES	AMRT+MS2	88.5	69.9
10.71	241.032	$[M+H]^+$	Cystine	AMRT+MS2	76.8	80.8
5.81	247.145	$[M+H]^+$	Trimethyl-tryptophan	AMRT+MS2	10.9	90.2
2.46	265.126	$[M+H]^+$	Phenylacetylglutamine	AMRT+MS2	89.9	88.7
6.95	268.105	$[M+H]^+$	Adenosine	AMRT+MS2	31.4	96.1
5.58	269.088	$[M+H]^+$	Inosine	AMRT+MS2	8.5	100.0
8.92	269.127	$[M+H]^+$	N-acetylcarnosine	AMRT+MS2	29.3	95.1
6.59	284.099	$[M+H]^+$	Guanosine	AMRT+MS2	1.5	91.6
7.51	290.135	$[M+H]^+$	Ophthalmic acid	AMRT+MS2	3.3	94.7
5.57	298.097	$[M+H]^+$	Methylthioadenosine	AMRT+MS2	65.6	95.3
8.86	303.068	$[M+H]^+$	PIPES	AMRT+MS2	9.6	83.5
6.69	310.114	$[M+H]^+$	N-Acetylneuraminic acid	AMRT+MS2	78.0	85.3
8.37	104.072	$[M+H]^+$	Aminobutyric acid	AMRT	0.2	73.0
9.10	106.050	$[M+H]^+$	Serine	AMRT	61.8	69.0
6.73	110.062	$[M+H]^+$	2-Aminophenol	AMRT	45.4	54.6
7.60	118.086	$[M+H]^+$	5-Aminovaleric acid	AMRT	2.3	57.2

3.45	123.055	$[M+H]^+$	Nicotinamide	AMRT	13.7	0.0
1.87	127.051	$[M+H]^+$	Thymine	AMRT	16.5	0.0
7.47	144.107	$[M+H]^+$	Proline betaine	AMRT	66.0	70.1
8.53	148.061	$[M+H]^+$	Glutamate	AMRT	7.8	70.0
8.82	166.054	[M+H]+	Methionine sulfoxide	AMRT	0.2	10.8
7.20	175.108	$[M+H]^+$	Theanine	AMRT	0.0	55.2
6.88	183.087	$[M+H]^+$	Mannitol	AMRT	4.0	64.1
4.42	187.072	$[M+H]^+$	Pyroglutamylglycine	AMRT	0.0	2.0
5.20	193.035	$[M+H]^+$	Citric acid	AMRT	10.2	32.2
9.52	244.093	$[M+H]^+$	Cytidine	AMRT	1.1	21.3
0.99	267.174	$[M+H]^+$	Tri-N-butyl phosphate	AMRT	67.6	69.7
5.65	285.084	$[M+H]^+$	Xanthosine	AMRT	39.6	77.7
1.16	361.200	$[M+H]^+$	Cortisone	AMRT	0.4	35.2
8.16	127.050	$[M+H]^+$	Imidazoleacetic acid	AM+MS2	56.6	89.2
9.07	190.119	$[M+H]^+$	Homocitrulline	AM+MS2	2.6	84.6
1.70	195.078	$[M+H]^+$	Aminohippuric acid	AM+MS2	36.8	88.3
3.76	232.157	$[M+H]^+$	Isobutyryl-carnitine	AM+MS2	34.9	97.9
4.56	257.081	$[M+H]^+$	Liquiritigenin	AM+MS2	5.9	90.2
6.57	259.104	$[M+H]^+$	AICA-riboside	AM+MS2	6.0	85.5

Glutamine and N-acetylcarnosine

Untargeted LC methods often contain regions with multiple co-eluting compounds. In this analytical method, the distribution of the 4159 features ranged from a few to over 250 peaks per 20 s (approximate average peak width at base) across the 0.8-15 min of gradient elution. Such coeluting peaks pose a challenge to deconvolution methods relying on mass chromatograms, but the CorrDec could deconvolute even completely coeluting compounds, such as abundant glutamine and little *N*-acetylcarnosine (**Figure 2-5A**). The peak intensities of the two compounds fit well with the reported average concentrations in the literature: 18-72 and 1-2 µM/mmol creatinine for glutamine [90] and *N*-acetylcarnosine (see supplemental material), respectively. Using MS2Dec, the deconvoluted spectrum of *N*-acetylcarnosine contained all fragment peaks of glutamine, reducing the MS2 match with the reference to only 29.3%. The deconvoluted spectrum of glutamine kept the MS2 match of >80% (**Figure 2-5C**). With the same dataset, CorrDec could deconvolute the MS2 spectrum of *N*-acetylcarnosine with >80% match (**Figure 2-5D**). Low abundance metabolites such as *N*-acetylcarnosine arguably constitute the larger part of most metabolomics datasets [72]. The high-quality MS2 spectra deconvoluted by the CorrDec enabled us to untangle the complex AIF dataset, by improving the identifications and annotations of smaller peaks in chromatographically dense sections.

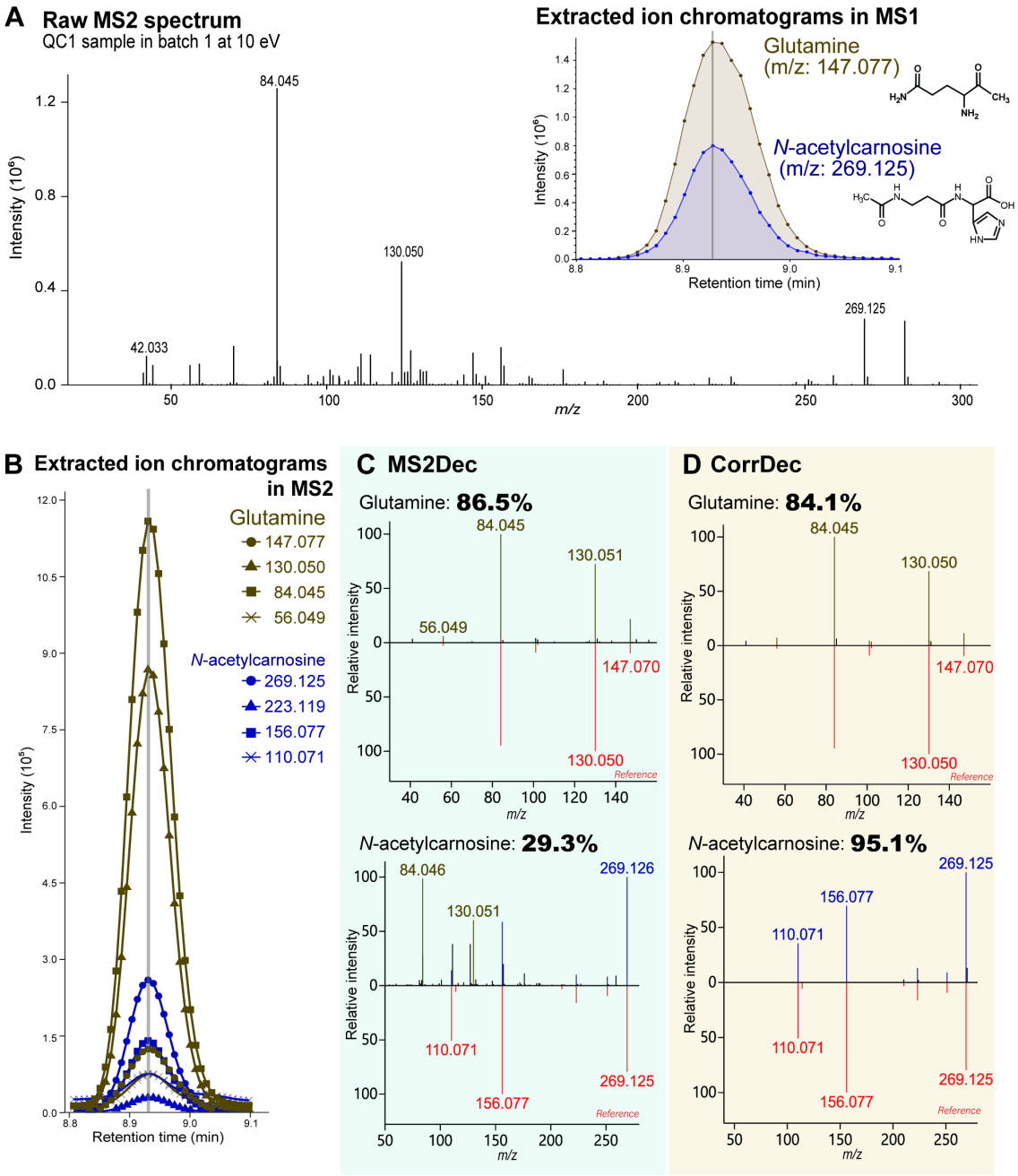


Figure 2-5. CorrDec can successfully deconvolute the MS2 spectra of completely coeluting compounds, glutamine and *N*-acetylcarnosine. **A.** The raw MS2 spectrum and extracted ion chromatograms in MS1 (0 eV) of completely coeluting glutamine and *N*-acetylcarnosine as well as **B.** their fragments in MS2 (10 eV) from the urine data (QC1 sample in batch 1). **C.** MS2 spectra of glutamine and *N*-acetylcarnosine deconvoluted by the MS2Dec. **D.** MS2 spectra of glutamine and *N*-acetylcarnosine deconvoluted by the CorrDec.

Random resampling validation

Estimating the number of samples required for CorrDec is difficult, because it depends on multiple factors (study design, sample matrix, metabolite, etc). Here for a rough estimation, I used 85 compounds confidently annotated (AMRT and MS2 match) in the urine study to perform random resampling analysis. Based on the median MS2 similarity from 100 iterations for each resampling, I plotted the number of compounds (total 85) having more than the particular MS2 similarity scores for each sampling number (**Figure 2-6**). Already with 10 samples, 47% (40 of 85) of the compounds showed over 80% MS2 similarity, when using 30 samples this number rose to 85% (72 of 85) of the compounds. Therefore, even smaller studies with tens of samples can benefit from the CorrDec method. Keeping in mind that urine is more variable compared to homeostatic fluids such as blood, I speculate that a larger number of samples might be required for successful application of CorrDec in studies with less metabolite variations between samples. The quality of MS2 spectra are largely dependent on compound classes and study designs; defining the best parameters or the minimum sample number required for all studies is therefore difficult.

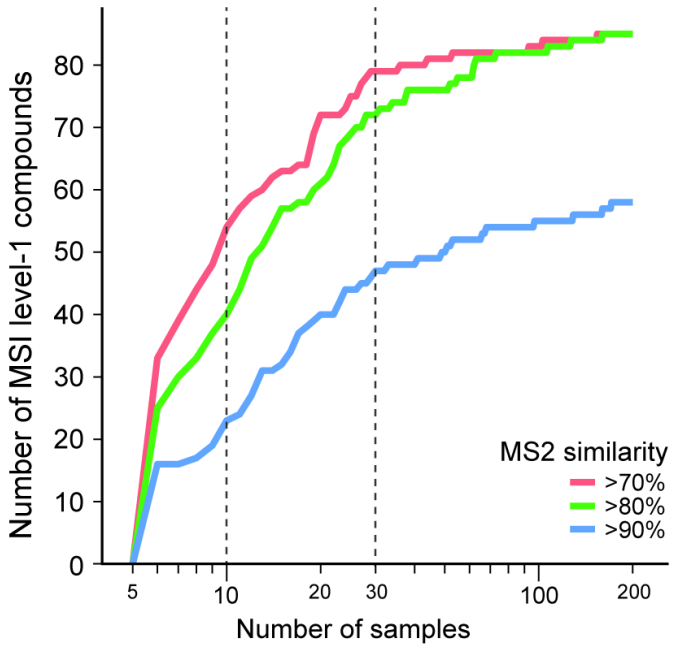


Figure 2-6. Summary of the randomized resampling analysis for the 85 CorrDec AMRT+MS2 compounds (Figure 2-4) to assess the relationship between the number of samples (urinary metabolomics dataset) used for the CorrDec and quality of the deconvoluted MS2 spectra compared library MS2 spectrum.

The benefits of CorrDec are summarized as: 1) cleaner MS2 spectra, and 2) statistical annotations (frequency and correlation) for MS2 peaks. CorrDec can generate clear spectra without noise signals from the matrix, mobile phase or mass spectrometer artifacts, enabling better matching to spectral databases and improving library search results. In the deconvolution process, each MS2 peak is assigned with a correlation value and frequency among samples. Advanced users can manually interpret deconvoluted MS2 spectra of unknown or marginally matching metabolites with reference spectra using these statistical annotations.

On the other hand, CorrDec has two disadvantages: 1) requiring multiple samples with varying compound concentrations, 2) possibly missing shared fragments with coeluting compounds. First, in principle, CorrDec cannot be performed on a single sample and at least three samples are required to calculate the correlation coefficient. While I observed that four spiked samples can be sufficient to obtain >80% similarity match, I investigated further to estimate the required sample number in an actual study using random resampling of the urine metabolomics dataset. Second, if coeluting compounds produce same *m/z* product ions, their intensity correlations might be small and be filtered out from deconvoluted spectra depending on the CorrDec parameters. MS2Dec spectra might be useful to complement missing peaks. Moreover, if advanced users try to retrieve missing peaks, they can carefully interpret the spectral statistical annotations and MS2 chromatograms.

In MS-DIAL, CorrDec is not intended to replace MS2Dec, as both deconvolution methods are based on different concepts and have different usage scenarios. The CorrDec method provides a reasonably clean deconvoluted MS2 spectrum per feature and sample set, therefore it is suitable for annotating and identifying a feature at the level of the whole sample set. MS2Dec can deconvolute MS2 spectra for each feature in a single sample; therefore, while noisier, the MS2Dec can be utilized to evaluate the feature identification for each sample in the dataset. In DIA metabolomics, MS2 spectra are obtained from only a small number of MS scans. For such complex and noisy data, traditional deconvolution methods such as multivariate curve resolution (MCR) is difficult to apply because the multivariate

method requires proper constraints to deconvolute spectra. When the number of coeluting compounds and peak shapes are interfered with noise, error-minimization is not a good algorithmic choice. Here, MS2Dec and CorrDec methods can function in complement to clean MS2 spectra from a relatively large dataset. Lastly, regardless of how clean the MS2 spectra or how good the MS2 library similarity matches are, it is still necessary to manually confirm compound annotations with chemical standards.

2-3 Conclusion

To obtain MS2 spectra for as many compounds as possible, AIF approach is useful in untargeted metabolomics. However, the complex AIF MS2 spectra require computational approaches for interpretation. To overcome the trade-off between comprehensiveness and cleanness of spectra, I have developed CorrDec—a new MS2 spectra deconvolution method for AIF data based on the correlations of the peak intensities across samples.

The serial dilution study of chemical standards showed that the peak intensities of fragment ions were highly correlated with those of their precursor ions across samples. The performance of CorrDec was demonstrated in the urine cohort study; the improved quality of the MS2 spectra and the ability to deconvolute completely coeluting compounds are the main advantages over retention-time based deconvolution methods. Additionally, CorrDec is useful for compound estimations by i*n silico* fragmentation tools such as MS-FINDER, because CorrDec spectra were generated without any reference MS2 libraries. Although CorrDec requires multiple samples to calculate intensity correlations across samples, it is applicable for almost all untargeted studies because multiple samples are usually measured in untargeted metabolomics for comparison. Of course, it should be noted that the quality and reliability of CorrDec spectra tend to be low in small-scale studies. In any case, manual confirmation is still needed because computational approaches might lead to misinterpretation.

CorrDec is available in MS-DIAL version 3.22 or later, and is already utilized by not only my collaborators, but also a few other MS-DIAL users. CorrDec can help to utilize

complex AIF MS2 spectra for reliable compound annotation and identification in various untargeted metabolomics studies.

2-4 Methods

Sample information and data acquisition

LC-MS measurements in AIF mode were performed as described previously [40,94], they are measured by my collaborators in Craig Wheelock laboratory. Metabolites were separated on a 15 minute gradient using HILIC chromatography with acidified water and acetonitrile. Data were acquired in positive ionization mode on an Agilent 6550 Q-TOF-MS system with a mass range of 40−1200 *m/z* in AIF mode with three alternating collision energies (full scan, 10 and 30 eV). The data acquisition rate was 6 scans/s.

Dilution series of eight chemical standards were prepared using urine as a matrix (proline betaine, trigonelline, dimethylglycine, trimethylamine *N*-oxide, tyrosine, glycine betaine, proline, 3-hydroxy-kynurenine). The original concentration of 4 μM in urine was diluted 1.5-fold with an equal amount of urine 10 times, resulting in an 11-point series to the final concentration of 69 nM. In addition, for tyrosine, a small step (1.07-fold) serial dilutions were also acquired.

Urine samples (n = 224) were used as the proof of concept for assessing the CorrDec performance. A detailed description of the full study is found in the original publication [95]. Samples were measured in four analytical batches, with pooled quality control (QC) sample injections every 5 samples and a water blank at the end of the batch sequence.

Chemical standard library

An in-house MS2 spectral library containing 13597 compounds was used for identification. The retention times (RT) for 280 compounds were obtained from purchased chemical standards [40,96].

Data processing and analysis

The CorrDec method was implemented into MS-DIAL [37]. Data were processed in MS-DIAL version 4.12 (peak detection, alignment, and deconvolution). Important parameters were: minimum peak height MS1: 3000, noise level of MS2: 1000, total identified score

cutoff: 80%, detected in 20% of all samples, not in blank (maximum sample intensity/average blank intensity > 5). As the used library contained records of both DDA and DIA spectra, during the identification processes I used the deconvoluted spectra with and without the ions above the precursor. The higher scored match was kept. Detailed data processing settings of MS-DIAL are shown in **Table 2-3** and **Table 2-4**. After the alignment of the features, MS2 spectra were deconvoluted using the CorrDec and the MS2Dec method independently.

Table 2-3. Experimental file of MS-DIAL for multiple collision energy mode

ID	MS Type	Start *m/z*	End *m/z*	Name	Collision energy	Deconvolution target (0: No, 1:Yes)
0	ALL	40	1200	10eV	10	1
1	ALL	40	1200	30eV	30	1
2	SCAN	40	1200	0eV	0	1

Table 2-4. MS-DIAL project settings

Start up a project	
Ionization type	Soft ionization
Method type	All-ions with multiple CEs (Table 2-2 experiment file)
Data type (MS1)	Centroid
Data type (MS/MS)	Centroid
Ion mode	Positive ion mode
Target omics	Metabolomics
New project window (file selection) Compound characterization	
Sample type	Sample
Set Class ID	Sample
New project window (file selection) E-TYPE	
Sample type	Sample, QC, and Blank as in Table S7
Set Class ID	same as sample type
Data collection	
MS1 tolerance	0.01
MS2 tolerance	0.01
Retention time begin	0
Retention time end	16
Mass range begin	40
Mass range end	1200

Maximum charged number		2
Consider Cl and Br elements	Unchecked	
Number of threads		20
Execute retention time corrections	Unchecked	
Peak detection		
Minimum peak height		3000
Mass slice width		0.1
Smoothing method	Linear weighted moving average	
Smoothing level		3
Minimum peak width		5
Exclusion mass list (tolerance: 0.01Da)	121.051, 922.0098, and 923.0129	
MS2Dec		
Sigma window value		0.5
MS2Dec amplitude cut off		1000
Exclude after precursor	Unchecked	
Keep isotope until		0.5
Keep the isotopic ion w/o MS2Dec	Unchecked	
Identification		
Retention time tolerance		1
Accurate mass tolerance (MS1)		0.01
Accurate mass tolerance (MS2)		0.01
Identification score cut off		80
Using retention time for scoring	Checked	
Postidentification	Not used	
Adduct		
Molecular species	$[M+H]^+$, $[M+Na]^+$, $[M+H-H2O]^+$, and $[2M+H]^+$	
Alignment		
Retention time tolerance		0.5
MS1 tolerance		0.01
Retention time factor		0.5
MS1 factor		0.5
Peak count filter		20
N% detected in at least one group		20
QC at least filter	Unchecked	
Remove feature based on blank information	Checked	
Sample max / blank average		5
Keep identified and annotated metabolites	Unchecked	
Keep removable features and assign the tag for checking	Unchecked	
Tracking of isotopic labels	Not used	

For the urine data, I manually confirmed and curated the alignment results to correct missed or doubtful peak picking, feature alignment, and compound identification. I also annotated all features using three criteria: (i) accurate mass (AM) match (tolerance: 0.01 Da), (ii) retention time (RT) match (tolerance: 1 min), and (iii) MS2 spectrum match (similarity >80%). The MS2 similarity was scored by the simple dot product without any weighting [97]:

$$MS2\ similarity\ (\%) = 100 * \frac{(\sum Am\ Ar)^2}{\sum Am^2 \sum Ar^2}$$

where *Am* and *Ar* are the arrays of *m/z* intensities in a measured- and reference mass spectrum, respectively. To avoid erroneous high similarity match resulting from only a few peaks, I adopted the following additional criteria: 1) match of at least two MS2 peaks with the reference spectra when the RT also matches, and 2) match of at least three MS2 peaks without the RT match. The MS2 similarities with reference spectra were compared between the CorrDec and the MS2Dec using three different collision energies (0, 10, and 30 eV).

Random sampling analysis

I evaluated the performance of CorrDec for different sample sizes by randomized resampling analysis of the urine metabolomics dataset. After chromatographic alignment was performed using all samples, I re-selected the study- and QC samples for deconvolution by the CorrDec. The number of samples varied from four to the number of detected samples (depending upon the chosen compound) with 100 iterations. For each iteration, I calculated the MS2 similarity between the deconvoluted spectrum from the resampling and the reference spectrum. The MS2 similarity of resampling was the average of 100 iterations.

Data availability

The datasets have been deposited to the EMBL-EBI MetaboLights repository with the identifiers MTBLS787 (chemical standards) and MTBLS816 (urine metabolomics).

Chapter 3

Chemical Library

Reliable chemical standard libraries also required consisting of accurate mass (AM), retention time (RT), and MS2 spectrum. In this chapter, a workflow to obtain AM, RT, and MS2 for a given compound using the AIF method will be proposed. I also provide practical recommendations for library development.

The contents of this chapter are also described in Tada et al. *Metabolites* 2019 [96].

3-1 Introduction

Interest in the analysis of the metabolome has increased significantly due to its utility for understanding biological processes and for biomarker discovery [98]. Liquid chromatography coupled to mass spectrometry (LC-MS) is a widespread metabolomics method owing to its sensitivity, and its measurement strategies are broadly classified into targeted and untargeted approaches [99]. Targeted approaches using LC-MS2 offer increased selectivity and quantification [100]; however, they are by nature limited to the measurement of preselected compounds. Untargeted metabolomics enables the discovery of unknown compounds; however, metabolite identification is a major bottleneck in data interpretation [72]. The criteria for metabolite identification was proposed [38], it is not enough for current untargeted metabolomics as detailed in Chapter 1.

To further increase the reliability of metabolite identification, MS2 spectra are used in addition to accurate mass and retention time (AMRT), MS2 spectra can be obtained from either data dependent acquisition (DDA) or data independent acquisition (DIA) [68]. In DDA, a narrow window of a few daltons or less is isolated around the precursor ion, and relatively clean MS2 spectra with a clear connection to their precursors are obtained [101]. However,

MS2 information is obtained only for a fraction of all detected ions in a measured sample. In DIA, on the other hand, all ions are sent to the collision cell to obtain their cumulative MS2 spectra (**Figure 3-1A**); this means that MS2 information is collected for virtually all ions in the sample (provided that they are of sufficient abundance). DIA-based data such as AIF (all ion fragmentation), MS^E, or SWATH (sequential windowed acquisition of all theoretical fragment ion mass spectra) [69] are therefore rich in content, but require spectral deconvolution. Towards this end, multiple software programs such as MS2Dec [37], MetDIA [79], and CorrDec (See Chapter 2) have been developed for interpretation of DIA-based data. In this process, there is little consensus on the treatment of spectra originating from identical compounds such as in-source fragmentation and different adducts [87]. In addition, peak intensities of MS2 spectra also depend on individual LC-MS instruments and measurement conditions [102]. Data analysis in DIA metabolomics is currently limited to the use of libraries constructed using DDA MS2 spectra without information on in-source fragmentation or multiple adduct types [40,103,104], or libraries with RT that are not suitable for the available measurement settings.

To address these difficulties and to provide a useful workflow for library construction, I demonstrate the creation of a reliable AMRT+MS2 library for LC-MS AIF metabolomics of hydrophilic compounds on a zic-HILIC column (**Figure 3-1B**). RT shifts were rigorously assessed using technical internal standards (tIS), and spectral deconvolution was fully exploited to obtain high-quality mass spectra for accurate metabolite annotation. A dedicated software tool was developed for comparing and sharing spectra in the NIST MSP format, named Mass Spectral LIbrary MAnager (MS-LIMA; see Chapter 4) [105]. Step-by-step tutorials can be downloadable for constructing (Tutorial 1) and application (Tutorial 2) of the AMRT+MS2 library on an AIF metabolomics dataset [106]. While for simplicity the application in this work is limited to zic-HILIC chromatography, this approach is generally applicable to any chromatographic system.

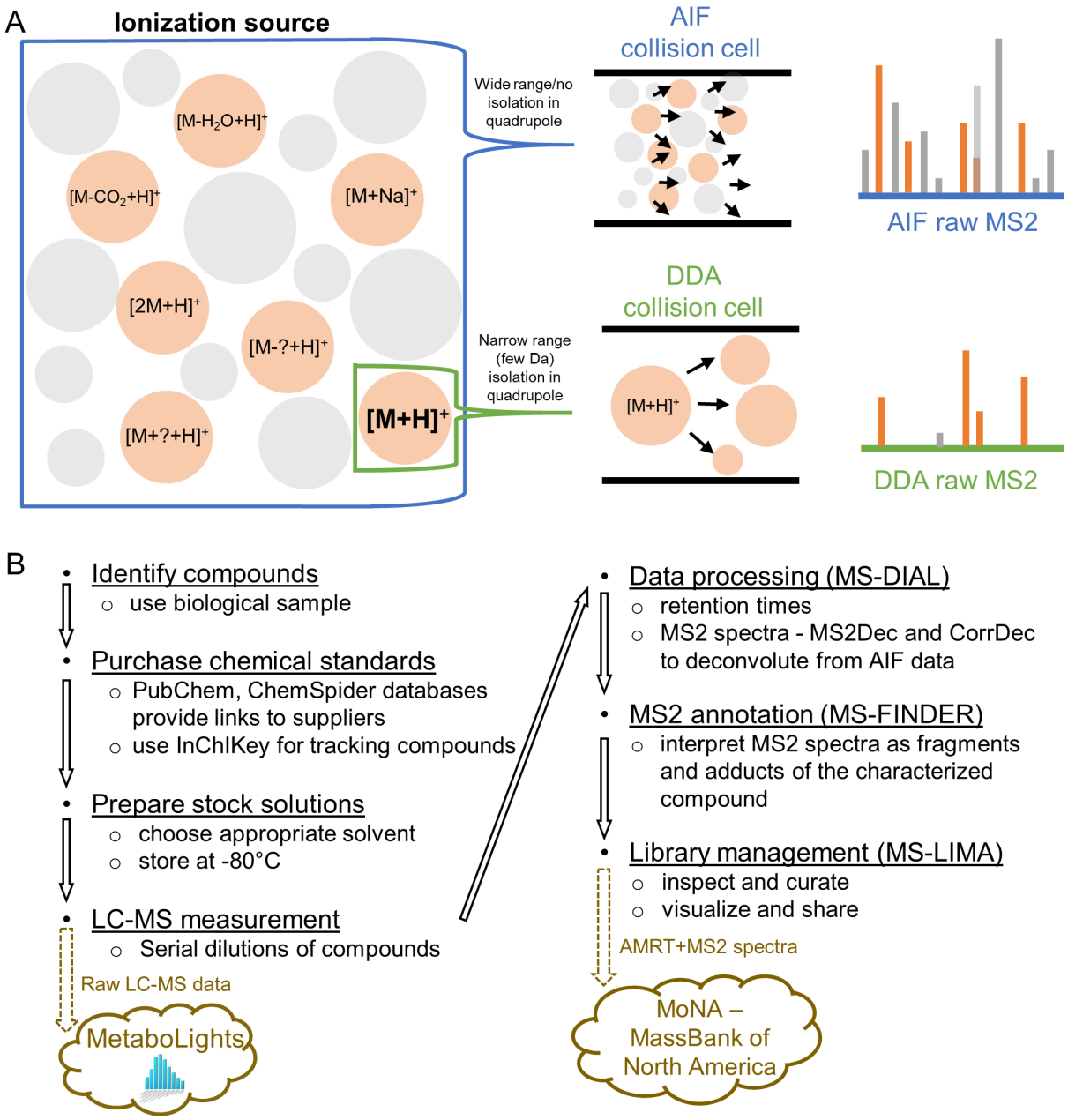


Figure 3-1. A. Comparison of AIF (all ion fragmentation) and DDA (data dependent acquisition) MS2 spectra acquisition, **B**. MS2 library construction workflow used in the current study

3-2 Results and Discussion

Chemical standard selection

Due to the need to perform multiple injections per compound, compound selection for inclusion in the library should be performed based upon likelihood of detection in authentic samples. I recommend establishing a list of compounds based upon feature annotation in the target sample matrix (e.g., pooled quality control samples, pilot study samples) [86,107]. Compounds can have multiple common names: for example, 5-pyrrolidone-2-carboxylic acid, pidolic acid, and pyroglutamic acid all designate the same chemical compound. In addition, identifiers from chemical databases such as HMDB [24], ChEBI [108], PubChem [56], ChemSpider [57], do not necessarily contain all synonyms for a given compound. InChIKey is a universal and unique compound identifier developed under the auspices of IUPAC (International Union of Pure and Applied Chemistry) [109], which can be used to search for other identifiers automatically (for example, with the R webchem package [110] or Chemical Translation Service [111]). PubChem and ChemSpider provide comprehensive information on the compounds, including a list of vendors when available. Commercial compounds are often available as salts (e.g., trigonelline chloride), with varying degrees of purity. While composition and purity of the chemical standard is crucial for direct infusion, it is not critical when LC separation is used (**Figure 3-2**)

Many plant- and food-based compounds are difficult to obtain commercially, as well as phase II metabolized forms (e.g., sulfates or glucuronides) of compounds other than drugs. While custom synthesis is an option, it is time-consuming, costly, and requires specific expertise [112]. When chemical standards are not available, the spectra of putatively annotated compounds in the samples can be used as an MSI level-2 or 3 compound library in order to reproduce consistent putative annotations across several studies.

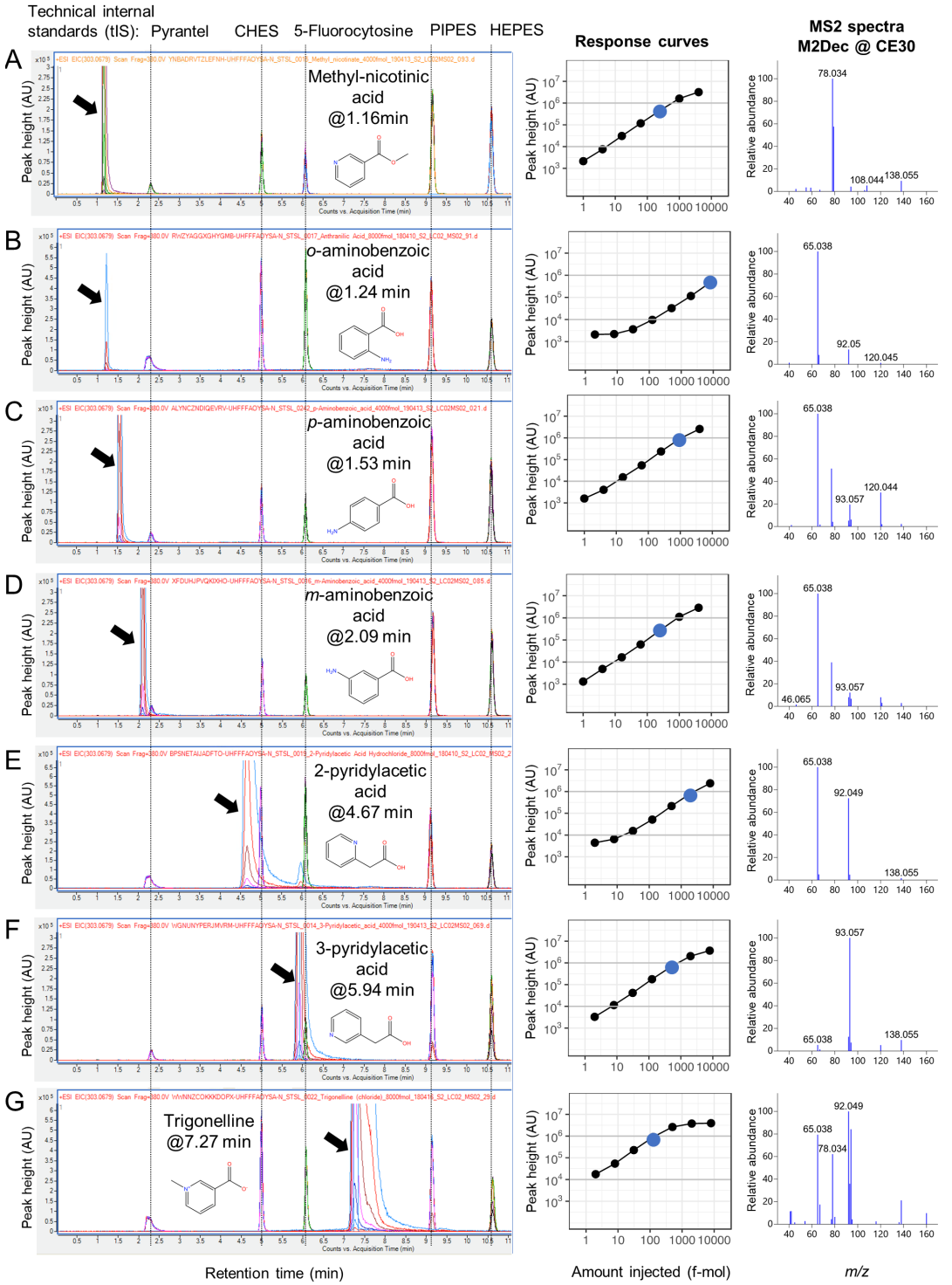


Figure 3-2. Retention time (RT) and response curve characterization of seven compounds with $C_7H_7NO_2$ formula in positive ionization mode on zic-HILIC chromatography. Peaks of the characterized compounds are indicated by black arrows. The elution order of the methyl-nicotinic acid and aminobenzoates (**A-D**) was confirmed by the constant RTs of the tIS. The analytical standard of 2-pyridylacetic acid (**E**) shows two peaks at 4.6 and 5.9 min, the later having the same RT as 3-pyridylacetic acid (**F**). Trigonelline (**G**) is detected at lower amounts than other compounds with the same formula. The shown MS2 spectra were deconvoluted using MS2Dec from the injection indicated by a blue dot in the response curve.

LC-MS acquisition of the chemical standard

When high-quality spectra are available, AIF data can be used to distinguish isobaric co- or closely eluting compounds [40,94,104]. However, compounds have different ionization efficiencies and response curves [113,114]. To produce a clean MS2 spectrum using MS2Dec [37], an appropriate amount for each compound should be injected into the LC-MS system. CorrDec requires multiple samples, with varying levels of the target compound (see Chapter 2). Therefore, multiple injections at different dilutions are necessary. Multiple injections also enable estimation of the detection and saturation limits for each compound. In positive ionization mode, as used in the current study, compounds with positively charged nitrogen atoms (e.g., trigonelline or trimethylamino groups in betaines and carnitines) ionize very well (**Figure 3-2**). The detection limits for such compounds can be an order of magnitude lower (around 0.1 fmol) compared with the standard amino acids and nucleosides (1–10 fmol). On the other hand, compounds containing only carbon, oxygen and hydrogen (e.g., carboxylic acids) are often poorly detected in positive ionization, and negative ionization mode should therefore be used [115]. In addition, depending upon the compound, the molecular ion might not always be the major species [86]. For example, in this study the main ions of chenodeoxycholic and cholic acids in positive ionization mode are $[M+H-2H_2O]^+$ and $[M+H-3H_2O]^+$.

Retention time normalization

RT characterization initially appears to be straightforward, simply requiring notation of the elution time of the injected chemical standard on the LC-MS system. However, RT can fluctuate depending on many factors, including the LC-MS system setup, solvents, column batches, etc. [116]. For example, some HILIC columns are prone to fluctuations in RT even within the same system and sorbent batch, which can complicate method transfer across laboratories and decrease long-term consistency. The challenge of RT shifts can be illustrated using two isobaric compounds, valine and betaine. In Naz et al. [40], who employed the same zic-HILIC method and instrumentation as this study, valine and betaine eluted at 6.79 and 7.10 min, respectively, while in the current work, they eluted at 7.21 and 7.41 min,

respectively. It is difficult to confidently identify these two compounds based solely on AMRT. The addition of MS2 spectra does not easily resolve this RT complication because low-molecular-mass metabolites with different structures may exhibit similar MS2 spectra as shown in **Figure 3-2** for compounds with the formula $C_7H_7NO_2$. RT characterization is necessary for reliable identification. Chemical standards may also contain impurities; for example, the peak of 2-pyridylacetic acid standard is separated by RT from 3-pyridylacetic acid (**Figure 3-2E** and **F**). To address this issue, I include multiple tIS in each injection to check (1) the performance of the instrumentation (e.g., peak shape, intensity); and (2) RT shifts. In the GC-MS field, the Kovats retention indices have been used for decades to adjust the RT shifts. However, in the LC-MS field, there is no single set of widely adopted retention index standards [117–119]. RT standards were only recently proposed for HILIC chromatography [120]. A practical solution for selection of tIS is a mix of common metabolites or exogenous compounds as in this study, with RT spread across the elution profile. To adjust the RT, first, the reference RTs of the tIS is obtained from an authentic representative analysis. Second, when processing each chemical standard data, their RTs are adjusted using the RTs of the tIS, based on a linear correction between each tIS. This is a relatively coarse correction, and other sophisticated approaches are available for larger deviations [121]. Information on the fluctuations of the tIS RTs from the library construction can be used when setting RT tolerance for compound identification in a dataset. For the five tIS used in this study, I observed that RT deviations <0.55 min from average and coefficient of variation (CV) across the seven injections of the 140 compounds in most cases <10%. I also observed ion suppression when a tIS coeluted with a characterized compound (e.g., fluorocytosine coeluted with norvaline betaine, resulting in ion suppression at 6.10 and 6.16 min, respectively). Currently the AMRT libraries can only be used for MSI level-1 annotation if generated in the same laboratory under identical experimental conditions. I demonstrate here that, in reality, experimental conditions fluctuate over time, even in the same laboratory on the same instrument (e.g., solvent, column production batches), greatly affecting the RT precision. Therefore, in current practice, untargeted metabolomics studies should only report MSI level-2 annotations, unless all standard compounds are simultaneously analyzed within

the same analytical batch/study. However, the use of measurable parameters such as RT deviations of the tIS should enable researchers to assess whether the library is suitable for the AMRT MSI level-1 annotations of a dataset.

MS2 spectra characterization

A high-quality library requires annotation of reliable product ions in MS2 spectra of the chemical standards. Comparison of the annotated compound MS2 spectra enables the search for compound-specific fragment ions. In the case of complex AIF data from biological samples, such compound-specific ions enable quantification of coeluting compounds such as threonine/homoserine [40] methylxanthines [94], or leucine/isoleucine [104]. In principle, DDA MS2 spectra can be used to identify such compound-specific ions, however, for example, DDA MS2 spectra obtained by direct infusion do not account for the in-source fragmentation as well as may contain peaks from isobaric impurities. Therefore, I recommend using annotated AIF MS2 spectra obtained from the characterization of chemical standard dilution series.

I used two deconvolution methods based on different concepts. MS2Dec [37] applies a least square regression method to consider the difference of liquid chromatographic peak tops, while CorrDec calculates the Pearson's correlation among multiple samples to identify correlated MS2 peaks with the precursor. In other words, MS2Dec and CorrDec consider different information: ion intensity over RT in MS2Dec, and ion intensity across samples in CorrDec.

From the dilution series, a representative sample (at unsaturated ion intensity corresponding to 10^4–10^6 AU, with the instrumentation and settings used in this study) was selected for each chemical standard. For all 140 compounds, raw MS2 spectra were obtained at 0, 10, and 30 eV collision energies. The median number of peaks in raw MS2 spectra were 52 (0 eV), 91 (10 eV), and 128 (30 eV) after removing small peaks with <1% relative ion intensities. Spectra were then deconvoluted using both MS2Dec and CorrDec. CorrDec was able to generate deconvoluted MS2 spectra for 132 of the 140 compounds, with eight

compounds not fulfilling the CorrDec criteria (at least four spectra of each compound have to be above the noise level). The two deconvolution methods produced similar spectra (the median dot product similarity: 81.3%), although their concepts and calculation methods are fundamentally different. The median number of peaks in MS2Dec spectra were 8, 15, 19, and in CorrDec spectra, 10, 19, 22 at 0, 10, 30 eV, respectively.

After deconvolution, MS2 peaks in each spectrum were annotated using the fragment annotation method implemented in MS-FINDER [93]. The MS-FINDER version 3.22 or later can estimate not only formula and substructure, but also isotopic ions and different adduct types of MS2 peaks from AIF data (AIF MS2 spectra may include different adduct types due to multiple precursors as explained in the Introduction). Nonannotated peaks were removed from the spectra, and the median number of removed peaks was four in both MS2Dec and CorrDec.

I detail the approach using the example of trigonelline, a betaine-type compound, made by plants and often detected in human biofluids [112]. Trigonelline ionizes well, and a relatively low amount of 125 fmol was sufficient to obtain a high (ion intensity: 907588), but unsaturated signal (**Figure 3-2G**). In the raw MS2 spectra at 30 eV (**Figure 3-3A** left column), the difference in the fragment patterns among the dilution series was observed. There was a common peak (149.022 *m/z*) detected in even the lowest concentration, which was most likely chemical noise (possible formula: $C_8H_5O_3$, corresponding to the common contaminant phthalic acid $[M+H-H2O]^+$ ion [122,123]). The MS2Dec spectra (**Figure 3-3A**, right column) were similar (the median similarity of all MS2Dec pairwise comparisons: 90.8%) over the dilution series. The only exception was the 31 fmol sample, whose base peak was 65.038 *m/z* (the median similarity between MS2Dec 31 fmol spectra and the other MS2Dec spectra: 49.0%); however, this peak was a fragment of trigonelline in combination with noise. A comparison of trigonelline's raw spectrum (**Figure 3-3A**, left column) to MS2Dec spectra (**Figure 3-3A**, right column) shows that deconvolution is indeed effective. The CorrDec spectra were generated using seven raw MS2 spectra and compared to representative MS2Dec spectrum, showing a good match (**Figure 3-3B**). In both spectra, the

primary adduct type observed was $[M+H]^+$ (138.055 *m/z*). Additionally, $[M+Na]^+$ (160.038 m/z) and $[M+K]^+$ (176.012 *m/z*) were also detected. The sodium and potassium adducts probably originate from the chemical standard, purchased as trigonelline chloride. To confirm the reliability of trigonelline's MS2Dec and CorrDec deconvoluted spectra, they were compared with the DDA MS2 spectra measured in house (**Figure 3-3C**). Although raw AIF MS2 spectra are noisy, the deconvoluted and curated MS2 spectra were well matched with the DDA MS2 spectrum. MS2 spectra deconvoluted from AIF data offer advantages relative to DDA MS2 spectra, including good coverage of isotopic patterns and inclusion of the adducts relevant to the LC method used in the acquisition (**Figure 3-3C**).

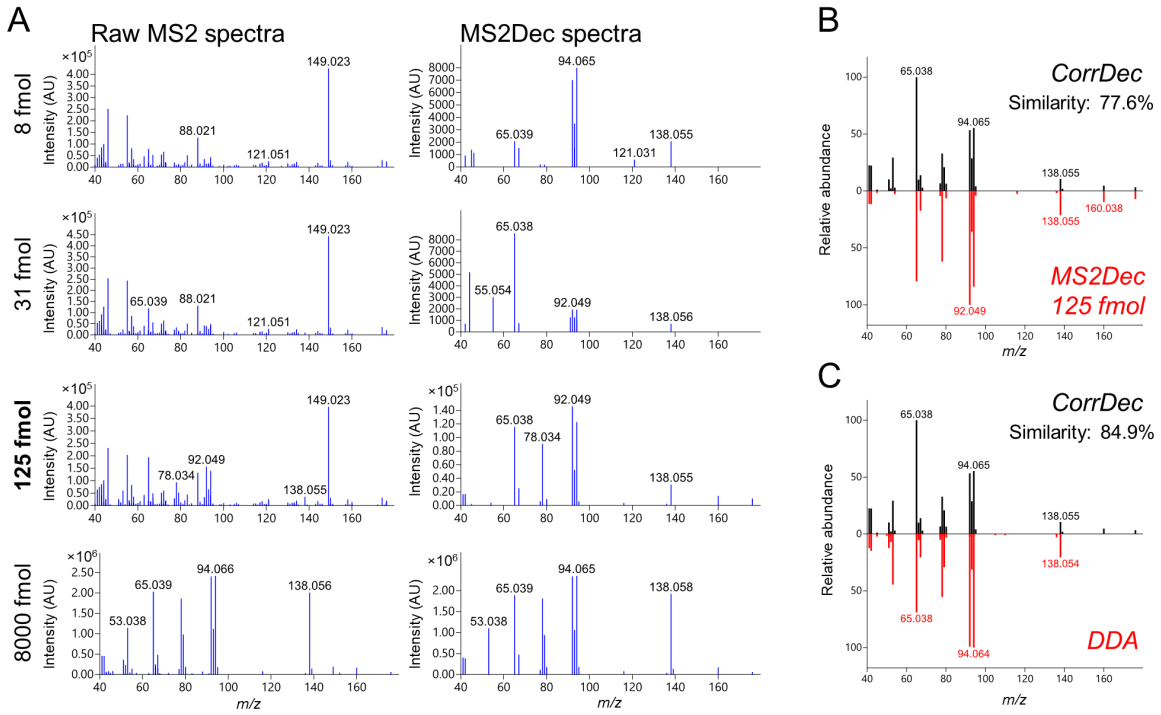


Figure 3-3. Deconvolution of trigonelline ($C_7H_7NO_2$, monoisotopic mass 137.0477) MS2 spectra from AIF data at 30 eV. **A.** Raw trigonelline AIF spectra contain multiple noise peaks (left column), compared to MS2 spectra deconvoluted by MS2Dec (right column), especially when lower amounts were injected. **B.** MS2Dec and CorrDec yield similar MS2 spectra. **C.** Comparison between CorrDec and DDA MS2 spectra acquired in house at 30 eV (MoNa ID: MoNA011431) confirms the solid MS2 deconvolution from the AIF data. Similarity reported as the dot product.

Confirmation and curation of MS2 spectra using MS-LIMA

With the MS-LIMA version 1.52, I examined 814 MS2 spectra (140 compounds) exported from MS-FINDER: compared the precursor *m/z* difference with theoretical *m/z*, confirmed adduct type and collision energies, and removed nonannotated MS2 peaks. The experimental precursor *m/z* was replaced with the theoretical precursor *m/z*, because the characterized compounds were known and theoretical precursor *m/z* values should be used in the mass spectral search to calculate the mass accuracy. The original experimental *m/z* values were stored, because it is also important to know the mass accuracy of spectral records. For example, the information of mass accuracy is necessary for structure elucidation tools such as MS-FINDER [93] and CSI:FingerID [43]. Although the MS1 mass accuracy cannot directly be transferred to the MS2 mass accuracy, the experimental precursor *m/z* value is a criterion to access accuracy in MS1 and MS2 spectra. Finally, I modified and added metadata, including SMILES, InChI, spectrum type, instrument, instrument type, chromatography, author, and license. As described in the methods section, raw data has been deposited to the EMBL-EBI MetaboLights repository [124] with the identifier MTBLS816, the MS2 spectral library was submitted to MoNA [125], and the RTs of compounds were also deposited at PredRet database [42], with the benefit of predicting RTs for uncharacterized compounds by mapping between multiple chromatographic systems. Raw data and MS spectra can also be deposited in other repositories (e.g., Metabolomics Workbench [65] and GNPS [61]). In this study, I used MS-DIAL and MS-FINDER to obtain the MS spectra from the AIF data; however, alternative workflows can be created using other available tools including MZmine [36], XCMS [126], CAMERA [83], RAMClust [85], MetFrag [44], and CSI:FingerID [43]. In the era of open science, sharing and obtaining feedback on the MS2 libraries is necessary for improving the quality as well as for developing the metabolomics community.

Library application for human urine study

A 224-sample urinary metabolomics study measured by AIF was used for library assessment. The dataset has been deposited to the EMBL-EBI MetaboLights repository with the identifier MTBLS816. To highlight the benefits of the library, I focused on the particular *m/z* window, 138.055 ± 0.01, which could correspond to $[C_7H_7NO_2+H]^+$; the details and additional examples are provided in the supplemental compound identification in the LC-MS AIF data tutorial (Tutorial 2) [106]. Based upon AMRT match only, which qualifies for MSI level-1, three features had plausible matches in the library (**Figure 3-4A**). With respect to MS2, two features at 4.99 min and 6.58 min did not match to any spectra in spite of relatively high ion abundance (**Figure 3-4B** and **C**). In contrast, a peak at 7.46 min could be identified as trigonelline, based on not only the AMRT, but also the MS2 match (**Figure 3-4D**). Therefore, I consider the two peaks at 4.99 and 6.58 min as adduct ions, in-source fragments, or unknown compounds. Due to RT fluctuations in HILIC chromatography, relatively large tolerances are used at the cost of reliable identification, and it is essential to use MS2 matching whenever possible to ensure accurate annotation.

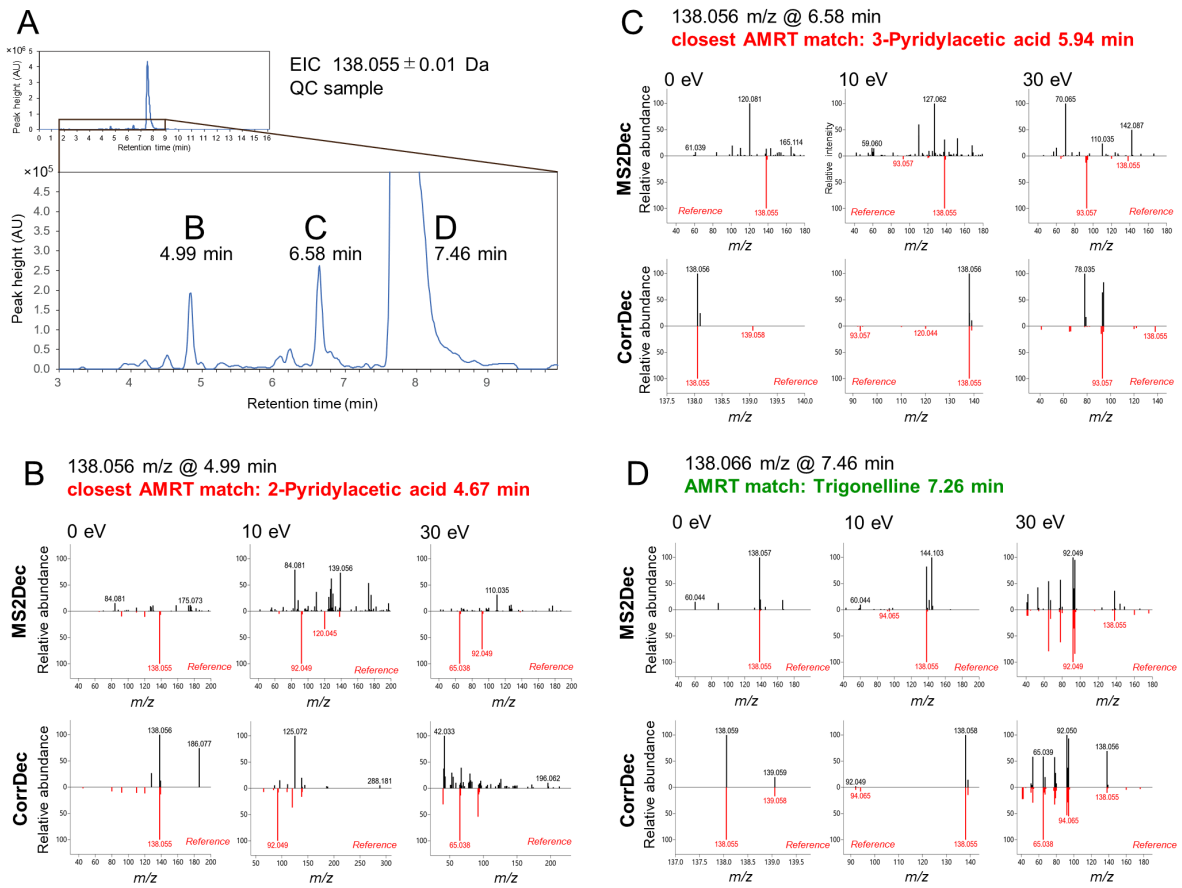


Figure 3-4. Application of the AMRT+MS2 library to urine metabolomics data acquired in positive ionization mode on a zic-HILIC column. **A**. Extracted ion chromatogram of *m/z* 138.055±0.01 Da (corresponding to $[C_7H_7NO_2+H]^+$) from a quality control (QC) sample. Two peaks at (**B**) 4.99 min and (**C**) 6.58 min have AMRT matches within 0.7 min, but poor MS2 match despite relative high abundance. A peak at 7.46 min (**D**) despite the mass shift due to high abundance could unequivocally be identified as trigonelline based on the AMRT+MS2 match (trigonelline was not spiked into the sample or known *a priori* to be present in the samples).

3-3 Conclusion

Reliable AMRT+MS2 libraries are necessary to confidently annotate and identify compounds in untargeted metabolomics. I describe workflow to obtain AM, RT, and MS2 for a given compound using the AIF data acquisition method and provide practical recommendations for library development. AIF spectra are useful as a library because they contain several adduct types. The main features of the library are normalized RT and annotated MS2 spectrum including several adduct types. The serial dilution measurements of standards can improve the confidence of measurements, confirm the ionization efficacy and suitable concentration, and be deconvoluted by CorrDec. To facilitate library curation and visualization, I developed the mass spectral manager MS-LIMA. The workflow can be easily reproduced by the AIF platform explained in the Chapter 4.

Although I highlighted the advantages of the created library, there are limitations. The library spectra were obtained from the LC-MS platform (Agilent Technologies, Santa Clara, CA, USA), and the spectra will most likely differ on platforms from other MS vendors with different ionization configurations. The set of tIS was chosen for the zic-HILIC method using positive ionization mode, and a different set may offer improved performance for a different combination of chromatography system, sample type, and ionization mode. For example, positive ionization mode is suitable for the urine study due to its efficient ionization of nitrogen-containing metabolites. However, negative ionization mode will require a different set of tIS, while reversed phase would yet again require a unique set of tIS. In this sense, it is difficult to assess the efficiency of the library only from a single study. However, the methodology introduced here is clearly transferrable, and there is a need to standardize this process within the metabolomics community. The construction of high-quality, open-access libraries makes compound annotations more transparent, reliable, and transferable to the broader community.

3-4 Materials and Methods

Sample information and data acquisition

Water, acetonitrile, methanol, and isopropanol used for the LC-MS analysis and sample preparation were of LC-MS grade and purchased from Wako (Osaka, Japan).

A stock solution (1–10 mM) for each chemical standard was prepared in water, methanol, acetonitrile, or other suitable solvent and stored at −80 ◦C. For the LC-MS characterization, seven 4-fold serial dilutions from 4.0–0.001 µM were prepared for each compound in acetonitrile containing tIS. An Agilent Bravo liquid handling system (Agilent Technologies, Santa Clara, CA, USA) with 96-well 0.2 mL PCR plates (PCR-96-MJ, BMBio, Tokyo, Japan) was employed to automate the serial dilutions. Pierceable seals 4Ti-0531 (4titude, Wotton, UK) were used to seal the plates for 4 s at 185 ◦C, using a PX1 heat sealer (Bio-Rad, Hercules, CA, USA). The plates were stored at 4 ◦C until measurement by LC-MS. See also tutorial chemical standard characterization using LC-MS AIF data (section "Handling of chemical standards and LC-MS measurements").

LC-MS measurements in AIF mode were performed as described previously [40,94]. My collaborators in Craig Wheelock laboratory measured all data with their instruments. In short, metabolites were separated on a 15 min gradient using a zic-HILIC column (100 × 2.1 mm, 3.5 µm particle size; Merck, Darmstadt, Germany) with acidified water and acetonitrile. Data were acquired in positive ionization mode on an Agilent 6550 Q-TOF-MS system (Agilent Technologies, Santa Clara, CA, USA), with a mass range of 40−1200 *m/z* in AIF mode, with three alternating collision energies (full scan, 10, and 30 eV). The data acquisition rate was 6 scans/s. One or two microliters of the solution were injected into the LC-MS system, corresponding to 1–8000 fmol. Solutions were injected from the lowest to the highest concentration, with a blank sample between each compound. The LC system was conditioned with several injections before each LC-MS sequence, and in each injection, a 7 min re-equilibration step was implemented after the gradient to maintain stable RTs.

Data processing and analysis

Data files were converted to mzML format using ProteoWizard version 3.0 [60] and processed in MS-DIAL [37] version 3.66 to obtain RT and MS2 spectra using MS2Dec and CorrDec deconvolution algorithms. The CorrDec function is implemented in the MS-DIAL (version 3.32 or later), which is freely available. Next, peaks in each MS2 spectra were annotated in MS-FINDER [93] version 3.22 and exported in NIST MSP format. Detailed settings of MS-DIAL and MS-FINDER are showed in **Table 3-1, 3-2,** and **3-3**. See also tutorial chemical standard characterization using LC-MS AIF data (Tutorial 1, sections "Deconvolution MS2 spectra in MS-DIAL" and "Annotation of MS fragments in MS-FINDER") [106].

Table 3-1. Internal standard settings for retention time normalization

Compound name	RT	RT tolerance	*m/z*	*m/z* tolerance	Minimum intensity	Include
Pyrantel STD $[M+H]^+$	2.3	0.5	207.09505	0.01	10000	TRUE
CHES STD $[M+H]^+$	5	0.5	208.10019	0.01	10000	TRUE
5-Fluorocytosine STD $[M+H]^+$	6.1	0.5	130.0441	0.01	10000	TRUE
PIPES STD $[M+H]^+$	9.1	0.5	303.0679	0.01	10000	TRUE
HEPES STD $[M+H]^+$	10.7	0.5	239.106	0.01	10000	TRUE

Table 3-2. MS-DIAL console project settings

#Data type
MS1 data type: Centroid
MS2 data type: Centroid
Ion mode: Positive
DIA file: File path to DIA file setting (Table S5)

#Data collection parameters
Retention time begin: 0.5
Retention time end: 15
Mass range begin: 40
Mass range end: 1200

#Centroid parameters
MS1 tolerance for centroid: 0.01
MS2 tolerance for centroid: 0.01

#Retention time correction
iSTD file: File path to internal standard file (**Table 2-3**)

Excute RT correction: TRUE
RT correction with smoothing for RT diff: TRUE
User setting intercept: 0
RT diff calc method: SampleMinusReference
Interpolation Method: Linear
Extrapolation method (begin): UserSetting
Extrapolation method (end): LastPoint

#Peak detection parameters
Smoothing method: LinearWeightedMovingAverage
Smoothing level: 3
Minimum peak width: 5
Minimum peak height: 1000
Mass slice width: 0.02

#Deconvolution parameters
Sigma window value: 0.5
Amplitude cut off: 1000
Exclude after precursor: FALSE

#Adduct list
Adduct list: [M+H]+,[M+Na]+,[2M+H]+,[M+H-H2O]+,[M+H-2H2O]+,[M+K]+

#MSP file and MS/MS identification setting
MSP file: File path to msp file
Retention time tolerance for identification: 1
Accurate ms1 tolerance for identification: 0.01
Accurate ms2 tolerance for identification: 0.01
Identification score cut off: 60

#Text file and post identification (retention time and accurate mass based) setting
Text file: File path to text format library
Retention time tolerance for post identification: 0.5
Accurate ms1 tolerance for post identification: 0.01
Post identification score cut off: 85

#Alignment parameters setting
Retention time tolerance for alignment: 0.1
MS1 tolerance for alignment: 0.015
Retention time factor for alignment: 0.5
MS1 factor for alignment: 0.5
Peak count filter: 20
QC at least filter: FALSE

#CorrDec setting
CorrDec excute: TRUE
CorrDec MS2 tolerance: 0.01
CorrDec minimum MS2 peak height: 500
CorrDec minimum number of detected samples: 4
CorrDec exclude highly correlated spots: 0.9

CorrDec minimum correlation coefficient (MS2): 0.9
CorrDec margin 1 (target precursor): 0.1
CorrDec margin 2 (coeluted precursor): 0.1
CorrDec minimum detected rate: 0.7
CorrDec minimum MS2 relative intensity: 1
CorrDec remove peaks larger than precursor: FALSE

Table 3-3. MS-FINDER settings for fragment annotation

Mass spectrum	
Mass tolerance (MS2)	0.01 Da
Relative abundance cut off	1%
Mass range max	1200 Da
Mass range min	40 Da
Advanced settings for AIF	
MS2 positive adduct list	$[M+H]^+$,$[M+Na]^+$,$[M+K]^+$,$[2M+H]^+$
Structure finder	
Tree depth	3

In order to curate and maintain the mass spectral libraries, I developed MS-LIMA software (open source, available on GitHub MS-LIMA project [105], see the next chapter). The library presented here was curated using MS-LIMA version 1.52 in the following manner: I replaced the experimental precursor *m/z* with the theoretical values (because the identity of the compound being characterized was known in each case) and kept only the peaks with the MS-FINDER formula annotation (isotopes, fragments, adducts) in the mass spectra. See also tutorial chemical standard characterization using LC-MS AIF data (Tutorial 1, section "Library assembly and curation in MS-LIMA") [106].

Data availability

The dataset has been deposited to the EMBL-EBI MetaboLights repository with the identifier MTBLS1040. MS2 spectra were submitted to RIKEN PRIMe website and MoNA (MassBank of North America [125]) with the tags: "zicHILIC_POS_KI-GIAR", "Agilent_6550_Q-TOF_AIF". RT were submitted to PredRet [42] and assigned to the chromatography named "KI_GIAR_zic_HILIC_pH2_7".

Chapter 4
AIF platform

In this chapter, the recent developments of three tools for AIF technology—MS-DIAL, MS-FINDER, and MS-LIMA—are introduced as the integrated platform.

4-1 Introduction

Metabolomics tools have been developed by vendors and researchers for feature picking, adduct ion estimation, compound identification, feature alignment, and statistical analysis. Integrated software for whole data processing is the recent mainstream. Comparing with genomics, many software supports graphic user interface (GUI) instead of console user interface (CUI), because the results (detected chromatographic peak region, adduct type, annotated compound, and alignment) should be manually confirmed and curated. Especially, chromatographic peaks and aligned features should be visually confirmed.

Currently, three integrated tools are widely used in untargeted metabolomics except for vendor software, XCMS [126], MZmine2 [127], and MS-DIAL [37]. XCMS is an R based software package, and its interactive online service also exists with METLIN database. MZmine2 is Java based software consisting of modular frameworks. MS-DIAL is C# based software for both DDA and DIA data. They have been utilized as reference in tool articles and compared with each other [79,128–130]. Each of them has advantages and disadvantages, and fairness evaluation is difficult. The overview of MS-DIAL is described in the next section.

Several functions have been anticipated for AIF data analysis. Multiple collision energies (CEs) can be sequentially measured such as 0, 10, and 30 eV in AIF. For each CE, the deconvolution and visualization are needed. Although AIF MS2 spectra consists of

several adduct types, current *in silico* fragmenters (CSI:FingerID [43], MetFrag [44], and MS-FINDER [45]) consider only an adduct ion in an MS2 spectrum. Lastly, user-friendly tools for AIF does not exist.

I have proposed the AIF platform consisting of MS-DIAL, MS-FINDER, and MS-LIMA (**Figure 4-1**). MS-DIAL and MS-FINDER were improved for AIF data. MS-LIMA was developed for mass spectral library management. Finally, the connections between MS-DIAL, MS-NDER, and MS-LIMA were enhanced to export/import in bulk and keep comments and sample information.

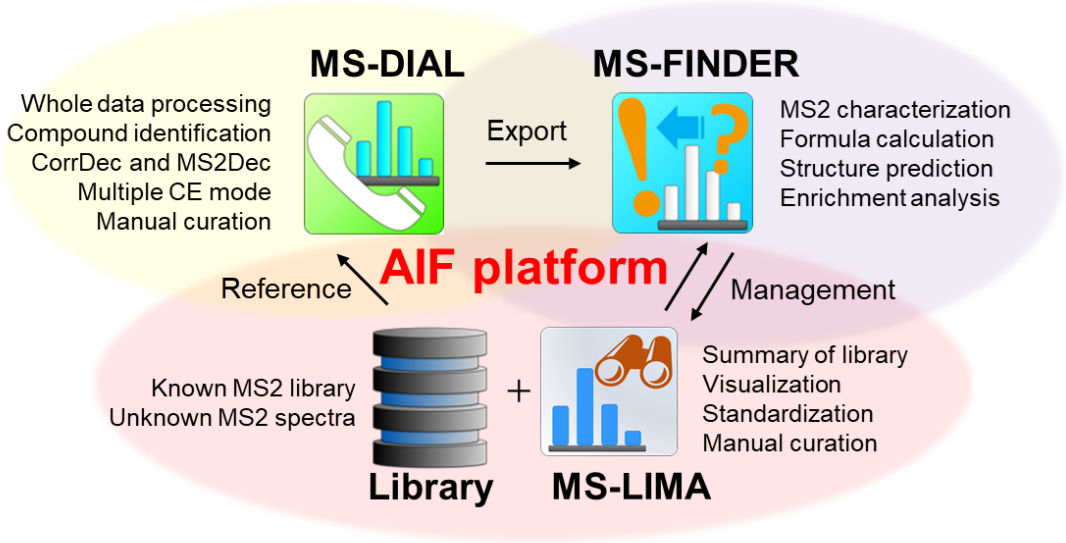


Figure 4-1. Overview of the AIF platform.

4-2 MS-DIAL

MS-DIAL is universal software for metabolomics and lipidomics developed by Tsugawa Hiroshi in 2015 [37]. I began to contribute to the MS-DIAL development in 2017 and became one of the main developers from 2019.

MS-DIAL can perform feature detection, adduct estimation, deconvolution, compound identification, feature alignment, data normalization, and statistical analysis. Moreover, it supports multiple instrument data (GC-MS, LC-MS, LC-MS2, LC-IM-MS2) with several common data formats (abf, mzML, and netCDS) into which all major vendor

formats can be converted. MS-DIAL is also known for the rapid development and frequent updates. Indeed, MS-DIAL is the first vendor-free software that supports all data processing for LC-IM-MS2 [131]. Nowadays, MS-DIAL is the most active category in the Metabolomics Society Forum [132].

In Chapter 2 and 3, several results and settings of MS-DIAL are documented. In this chapter, the functions mainly developed and implemented by the author are introduced in the following subsections using public AIF data of yeast strains as a demonstration [133].

Multiple collision energy mode

I have developed new methods and GUIs to adopt multiple collision energies (CEs; **Figure 4-2**). The main functions of them are following; (1) accept metadata of multiple CEs (**Table 2-3**), (2) identify compounds using multiple MS2 spectra with different CEs, (3) export all MS2 spectra, (4) simultaneously visualize multiple MS2 spectra in a feature. The main window and its usability are almost same as DDA and SWATH mode. As a difference, AIF controller window appears in multiple CE mode, and the additional windows are launched as in **Figure 4-2.**

The multiple CE mode in MS-DIAL is first integrated program for large-scale AIF data with user-friendliness. According to discussion with AIF users, certain tools take a few minutes or more to just visualize a graph for manual curation, but MS-DIAL can visualize the same graph at once. In addition to the speed, the simultaneous visualization of all deconvoluted MS2 spectra of multiple CEs helps to understand the characteristic of product ions. For encouraging AIF study and data analysis, the multiple CE mode is essential. The demonstration and details are also described in Chapter 8 of the MS-DIAL online tutorial [134].

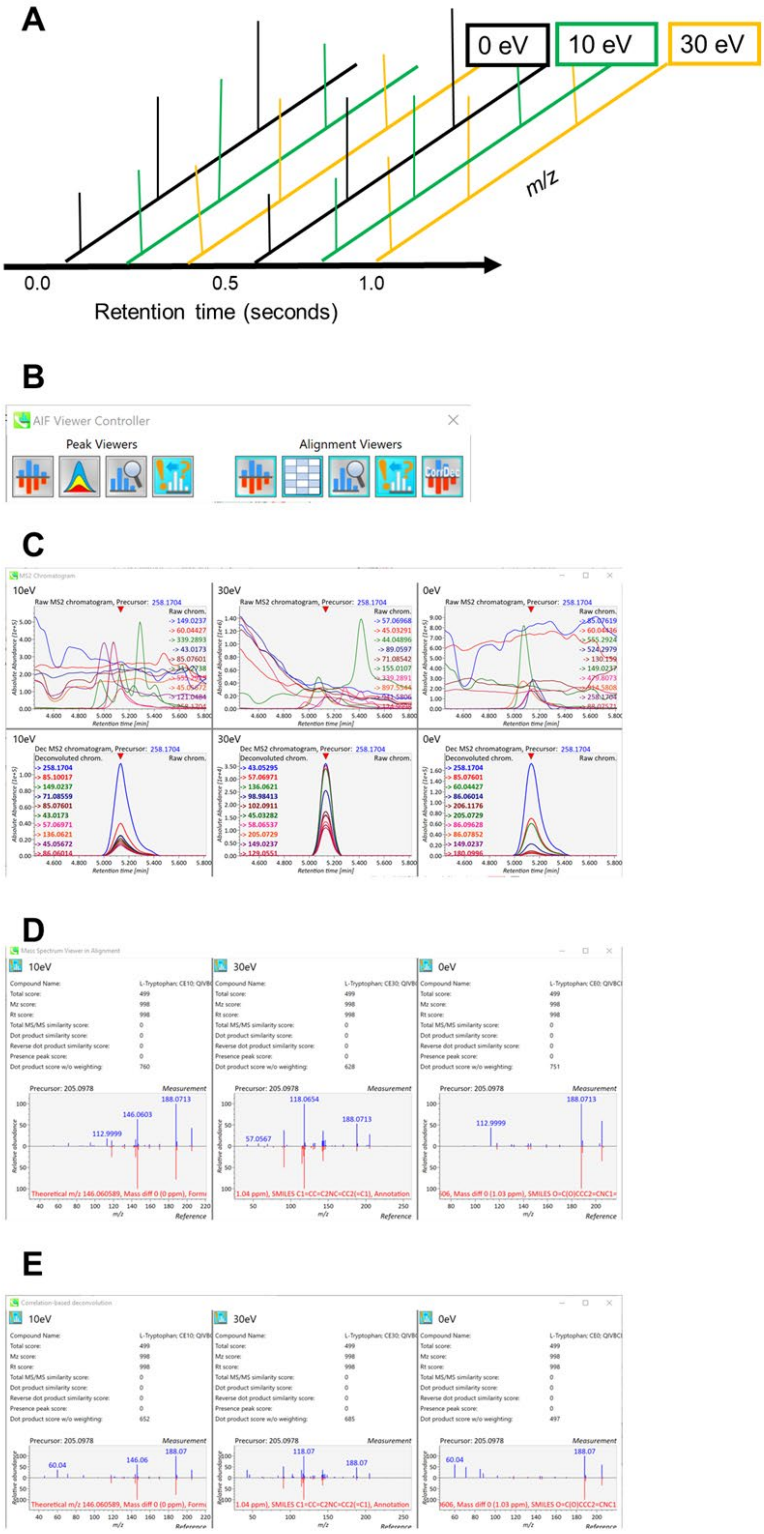


Figure 4-2. Multiple collision energy (CE) mode. **A.** An example of multiple CEs (0, 10, and 30 eV). **B.** AIF viewer controller to launch additional windows, **B**. raw and deconvoluted MS2 chromatograms of tryptophan at 0, 10, and 30 eV. Deconvoluted MS2 spectra of tryptophan at 0, 10, 30 eV by MS2Dec (**C**) and CorrDec (**D**).

Correlation-based deconvolution (CorrDec)

For complex AIF MS2 spectra, CorrDec was developed as detailed in Chapter 2, and implemented into MS-DIAL. According to its workflow (**Figure 2-2**), CorrDec program can run after feature alignment in multi-sample projects (>6 samples without blank samples).

MS-DIAL supports two different concept deconvolution methods, CorrDec and MS2Dec, and can visualize both deconvoluted spectra at all CEs of a feature. In the case of demonstration data, 6 patterns MS2 deconvoluted spectra of tryptophan are simultaneously visualized and users can utilize them for interpretation (**Figure 4-2D** and **E**).

Aligned extracted ion chromatogram

To confirm the accuracy of feature alignment, I developed a new "graphical interface" for aligned extracted ion chromatograms (aligned EIC; **Figure 4-3A**). In aligned EICs, EICs from all samples in an alignment are overlaid to easily understand the difference among samples and detect miss-alignment. At the last part of data processing, all aligned EICs are calculated and saved as an additional file. Therefore, the visualization is very fast and stable; even >1000 samples project can visualize it in a few seconds.

Chromatographic peak modification

To curate miss-alignment manually, Tsugawa and I added chromatographic peak modification. Tsugawa implemented a function to curate a chromatographic peak, and I developed a function to simultaneously modify all peaks in an alignment to quickly curate even in large-scale projects (**Figure 4-3B**). First, when user select a region on the top panel in **Figure 4-3B**, EICs in the middle graph is aligned to adjust peak top in the selected region. Then, user can select chromatographic peak from RT aligned chromatogram. This function helps to save time for concentrating biological things.

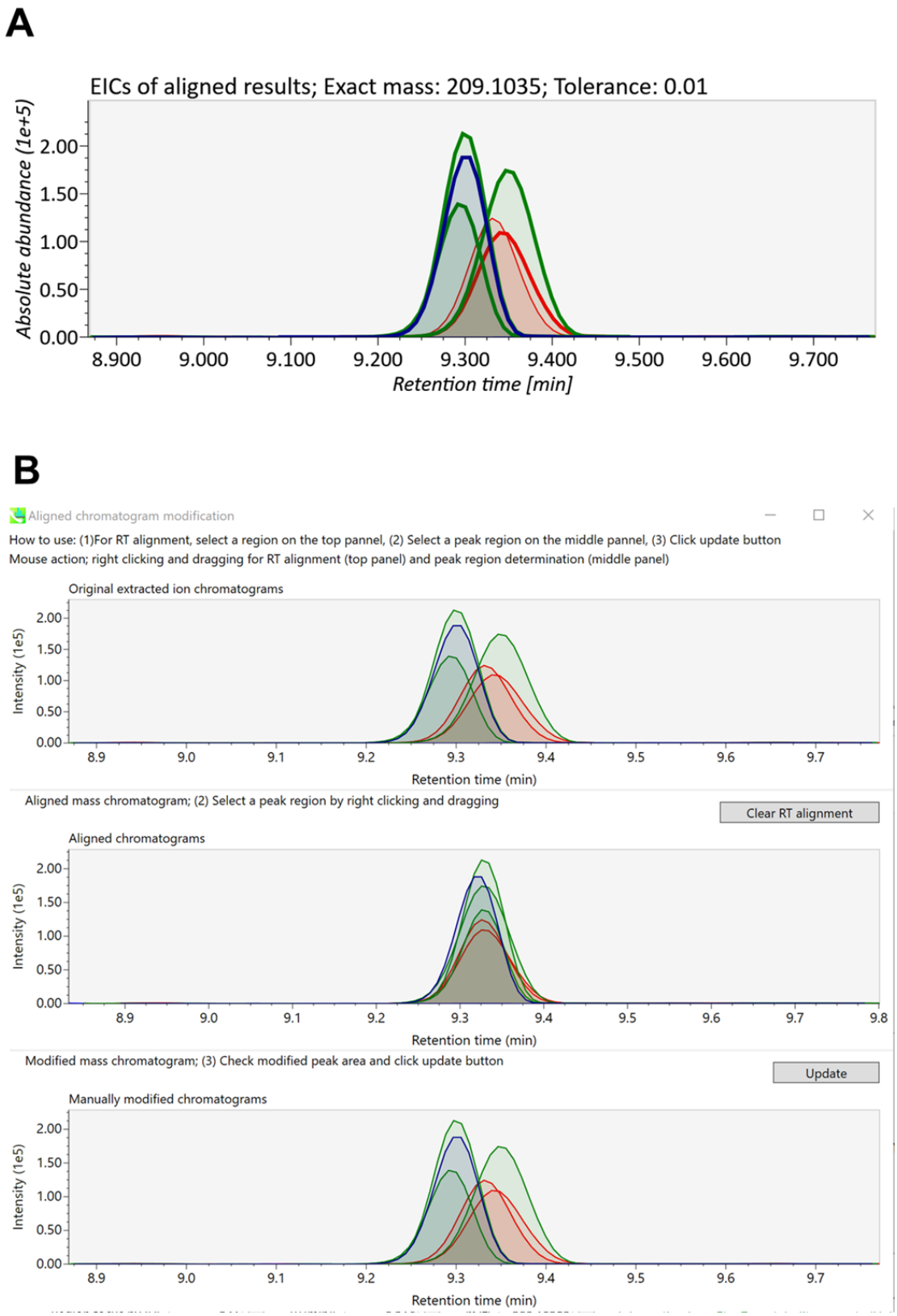


Figure 4-3. Aligned extracted ion chromatograms (EICs). **A.** The aligned EICs graph in the MS-DIAL main window, **B.** the multiple peak modification window.

Retention time correction

To reduce miss-alignments, I implemented a simple method of retention time (RT) correction based on linear correction by several internal standards. For reliable RT correction, the correction result should be manually confirmed. It provides several graphs to compare their EICs before and after correction (**Figure 4-4**). As a demonstration, EICs of an alignment are shown in **Figure 4-5** before/after RT correction.

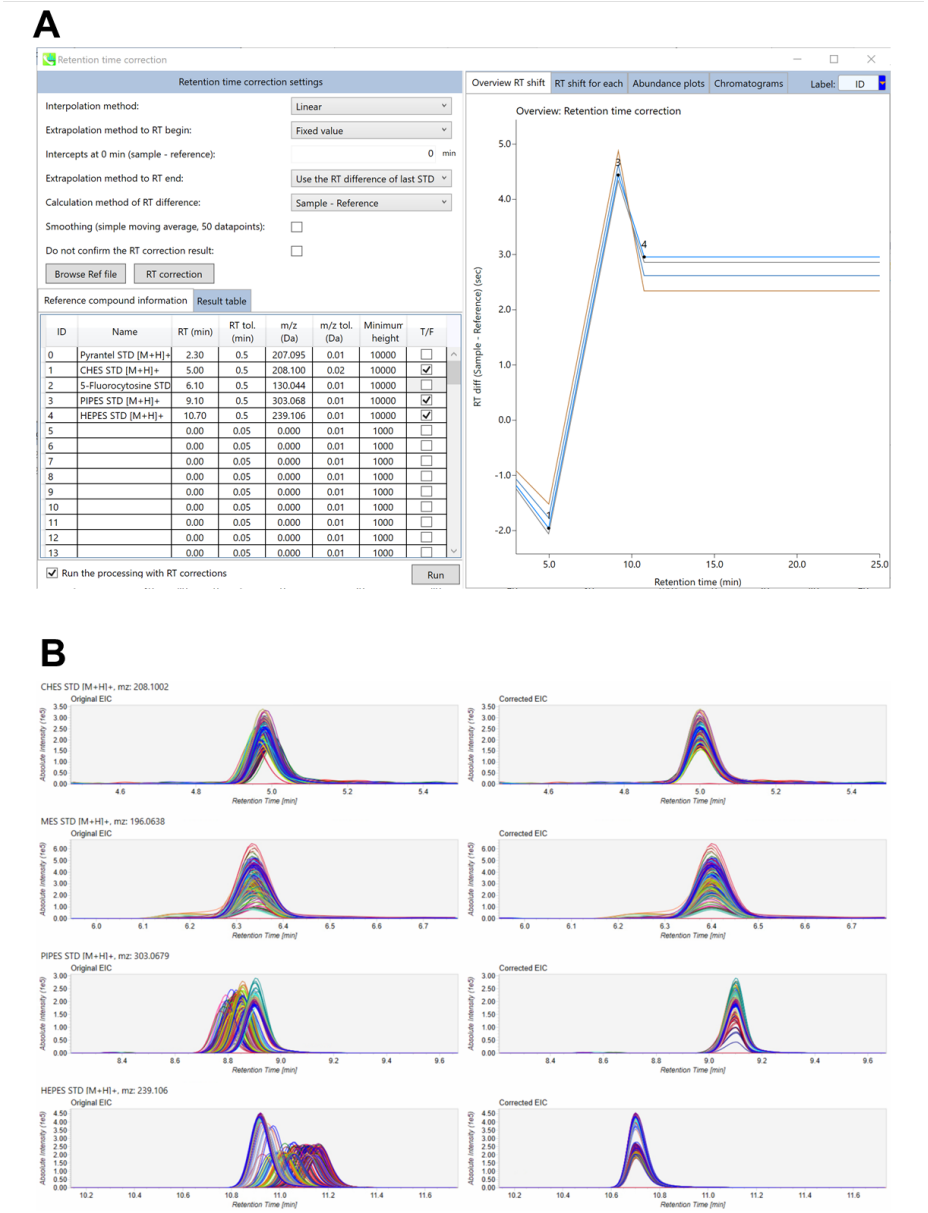


Figure 4-4. RT correction windows **A.** Screenshot of overview of the window, **B.** EICs of selected standards before/after RT correction.

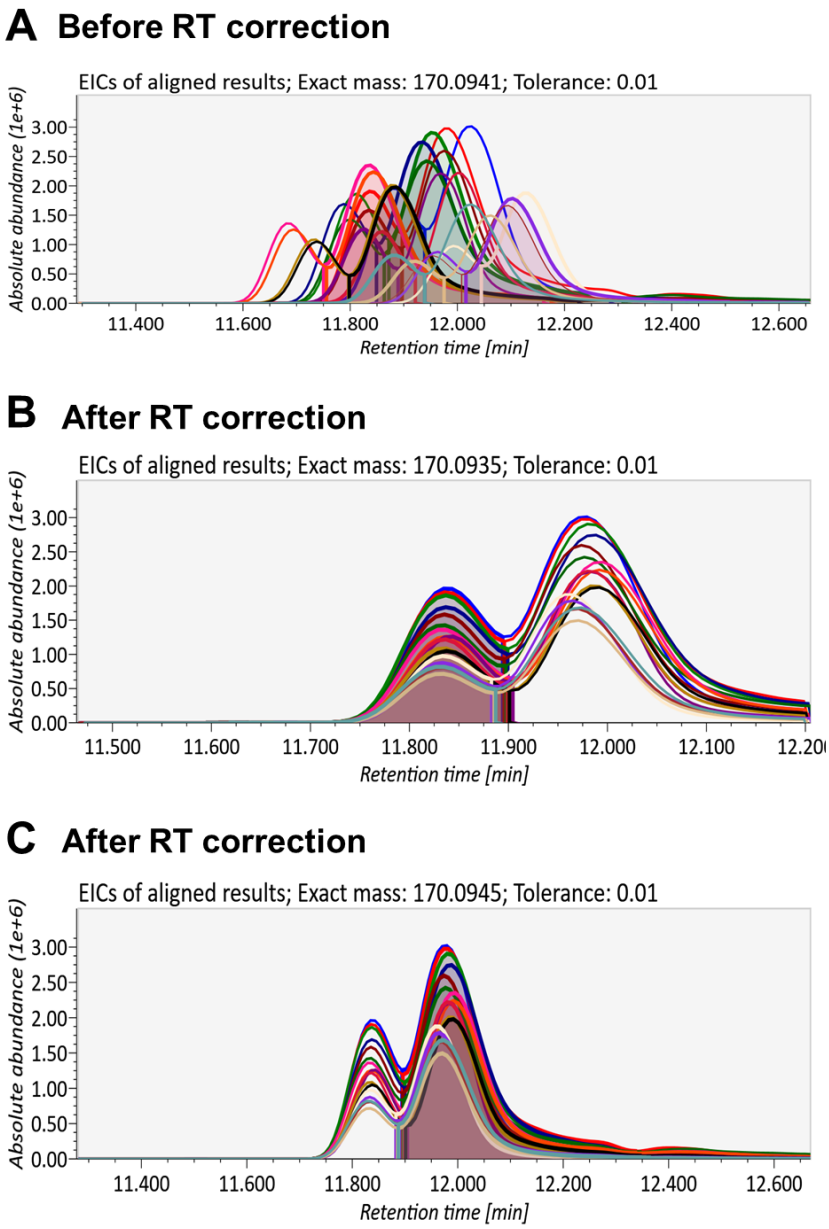


Figure 4-5. EICs of 170.094 *m/z* with mass tolerance 0.01 before (**A**) and after (**B, C**) RT correction.

Data format and multi-threading

Recently, the large number of samples are measured in metabolomics such as hundreds and thousands. I remodeled result file formats to quickly save, load, and select data even in >1000 sample projects. Moreover, I changed data processing from single thread to multi thread. Even in large-scale study, MS-DIAL requires relatively small RAM (16GB for hundreds of samples, 64 GB for thousands of samples; of course, it depends on data size) while the process is very fast.

Table viewers

Table viewers are just additional windows to visualize results as table with filtering function. However, it helps many users to save their time and find interesting metabolites (in personal communications).

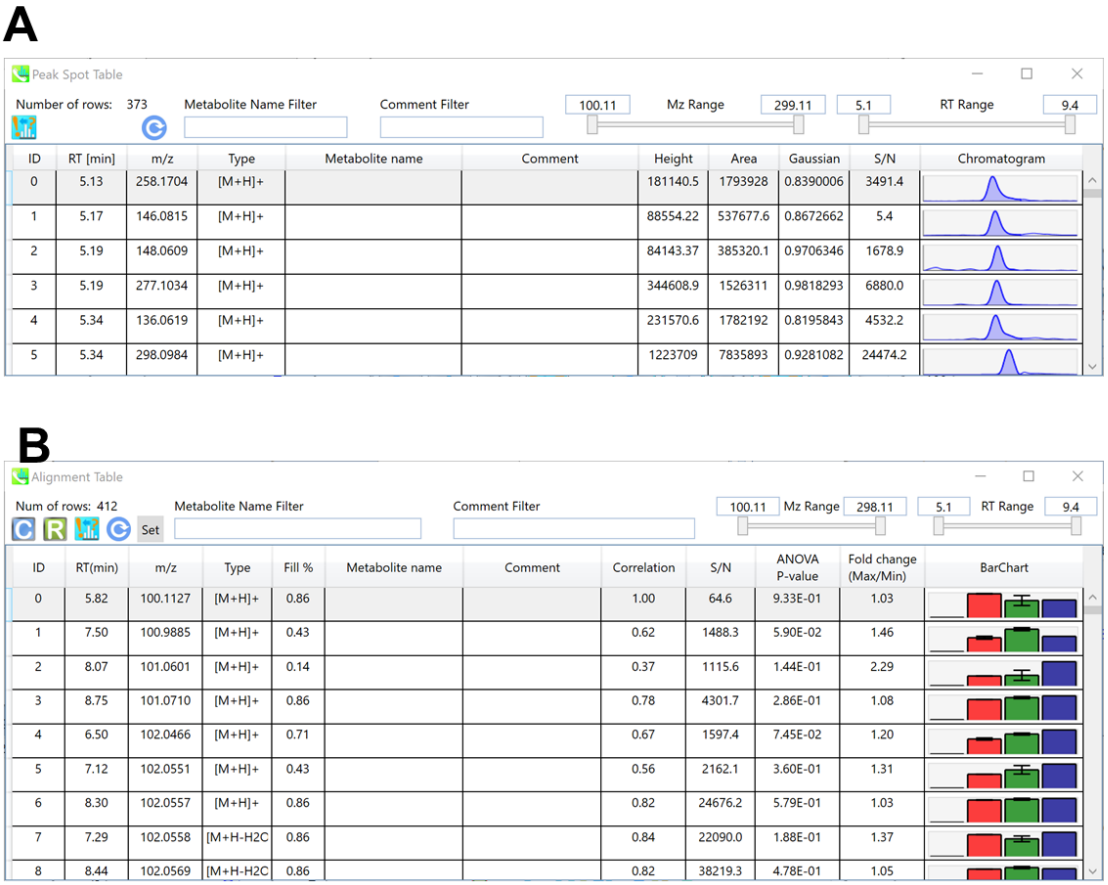


Figure 4-6. Screenshots of sample table viewer (**A**) and alignment table viewer (**B**).

MS-DIAL console for Windows, Linux, and MacOS

Lastly, I am trying to develop a console program worked in all computational resources. MS-DIAL can perform in only Windows because it is written by C#. C# is a programing language to provide very nice GUIs, but it supports mainly Windows OS. Recently, a C# framework for Unix/Linux has been provided by Microsoft, therefore, universal MS-DIAL console program can be released for Windows, Linux, and MacOS.

4-3 MS-FINDER

For untargeted metabolomics, chemical structure elucidation from MS2 spectrum is a key process. MS-FINDER is also originally developed by Hiroshi Tsugawa to elucidate chemical structure by *in silico* fragmentation. MS-FINDER can also assign molecular formula and substructure to MS2 peaks of known compound. As a result of my contribution, MS-FINDER can predict different adduct ions in MS2 spectrum, and export detailed annotations as comments of MS2 peaks. These functions were utilized in the library creation study described in Chapter 3.

Adduct ion annotation for MS2 spectrum

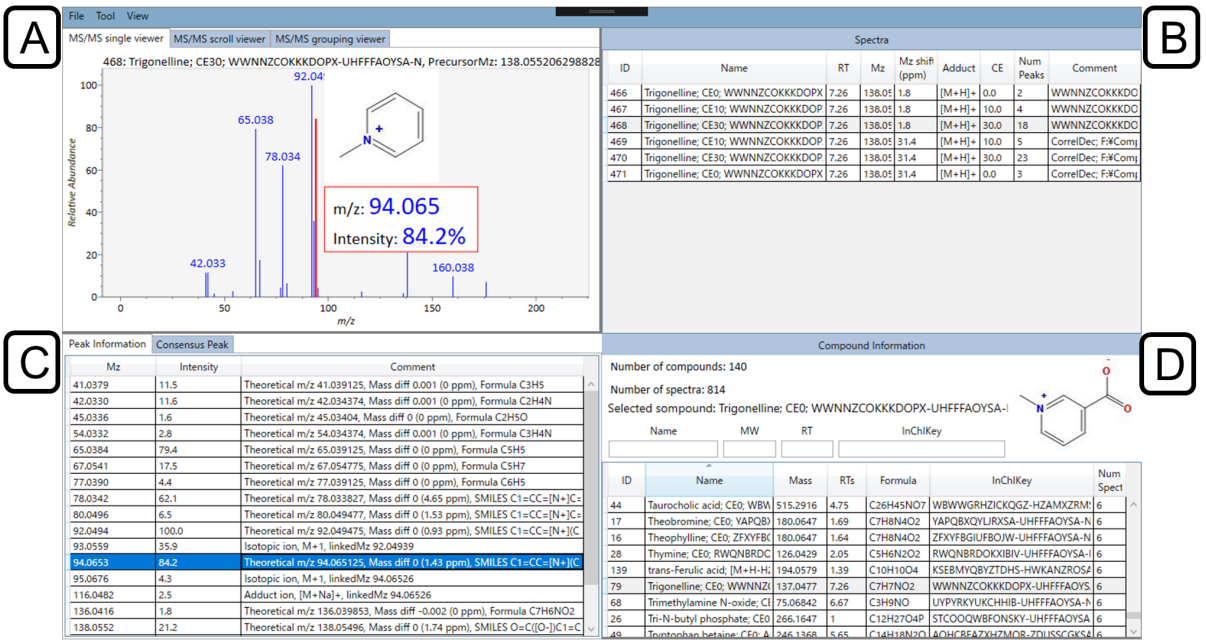
AIF MS2 spectrum includes various adduct ions even from chemical standard for library creation. I have implemented adduct ion annotations for MS2 spectrum into MS-FINDER. In the setting window, target adduct types can be selected. After molecular formula assignment in MS2, unknown MS2 peaks are candidates of adduct ion estimation. Based on pre-defined adduct types, the candidate peaks are annotated as a different adduct ion.

4-4 MS-LIMA development

A chemical standard library should be confirmed and maintained by researchers who use the library. The main advantage of NIST MSP format (also MGF and MassBank) is that text editors can open and edit MSP files - MSP is a plain text format. Even though it can be opened by text editors, visualizing mass spectra is still important to intuitively understand their features. This is possible within NIST MS Search Program [62], which is primarily

designed for searching NIST Tandem Mass Spectral Libraries. Although the software is useful for mass spectral library searches, it is not convenient for the curation and management of AMRT+MS2 libraries containing MS2 peak annotations from MS-FINDER.

I have developed a new open-source software, MS-LIMA (Mass Spectral Library MAnager), to visualize, manage, and curate mass spectral libraries. The main window of MS-LIMA is shown in **Figure 4-8** after opening the library described in Chapter 3 and selecting peak at 94.065 *m/z* originating from trigonelline spectrum at 30 eV. MS-LIMA supports MassBank, MGF, and many subtypes of MSP formats from different institute and databases, such as RIKEN, MoNA (MassBank of North America), and NIST.



E

```
NAME: Trigonelline
RETENTIONTIME: 7.26
PRECURSORMZ: 138.0552
FORMULA: C7H7NO2
PRECURSORTYPE: [M+H]+
INCHIKEY: WWNNZCOKKKDOPX-UHFFFAOYSA-N
COLLISIONENERGY: 30
SMILES: C[N+]1=CC=CC(=C1)C(=O)[O-]
Num Peaks: 18
41.03791    11.47238    "Theoretical m/z 41.039125, Mass diff 0.001 (0 ppm), Formula C3H5"
42.03302    11.61852    "Theoretical m/z 42.034374, Mass diff 0.001 (0 ppm), Formula C2H4N"
…
```

Figure 4-7. MS library organization and editing with MS-LIMA. **A.** Visualization of MS spectrum with **B.** editable annotations from MS-FINDER for each peak. **C.** Available MS spectra for **D.** a selected compound in loaded AMRT+MS2 library. **E.** For MS-LIMA libraries, I recommend including the following lines for each record with trigonelline as an example.

After opening library files, MS-LIMA groups compound spectra based on InChIKey or short InChIKey (the first 14 characters corresponding to the bonding pattern are termed the short InChIKey and the default setting uses the short InChIKey) to easily compare and assess MS2 spectra originating from the same compound. In the grouping process, MS-LIMA checks all MS2 records from the same compound whether they share an identical formula and similar retention time (<1 min difference as default) limiting the possibility that the given MSP files contains retention times from different LC methods. MS-LIMA also supports commenting to MS2 annotated peaks by MS-FINDER version 3.22 or later, and visualization of the substructure for the selected peak. To curate spectra, users can check precursor *m/z* differences and modify all information in the library. Also, MS-LIMA has various functions to manage and curate the library, such as MS2 spectra comparison between two libraries, making a consensus spectrum of a compound, calculating the frequency of product ions among library, automatically saving, exporting spectrum as several formats, replacing metadata based on InChIKey (see GitHub repository for details [105]). Moreover, since it is open-source, anyone can contribute to MS-LIMA development to support additional formats or add new functions.

In MSP format, inconveniently, there are no strict rules for describing meta information. For example, “retention_time” and “RETENTIONTIME” can be used to indicate retention time in MS-LIMA. I recommend adding metadata for each record in AMRT+MS2 library as shown in **Figure 4-7E** using trigonelline record as an example. Additional information lines (e.g. full InChI) can be included in MSP without affecting MS-LIMA functionality. I expect that standards and guidelines for AMRT+MS2 libraries will gradually evolve, as in recent MS spectra annotations.

4-5 AIF platform

The AIF platform consisting of MS-DIAL, MS-FINDER, and MS-LIMA is useful not only for compound identification with reference libraries, but also for unknown compound characterization. Unknown compounds detected by MS-DIAL can be exported to MS-FINDER, then they can be annotated molecular formula and chemical structures.

MS-FINDER can export them to MS-LIMA as MSP format with MS2 peak annotations. In MS-LIMA, all characterized unknown compounds from known biological samples can be accumulated as a detected unknown compound library. The accumulated unknown compound library can be utilized to elucidate new biological findings, and MS-DIAL can utilize the library for comparison with other projects.

The created AIF platform can be utilized in three cases: (1) data analysis using own metabolomics data (Chapter 2), (2) creating reliable chemical library (Chapter 3), and (3) reanalysis of public metabolomics data. I described the usability and examples of the AIF platform for the first and second cases in Chapter 2 and 3. Lastly, reanalysis of public metabolomics data can be easily conducted by the AIF platform, although it is hard task to elucidate new findings from public raw metabolomics data.

4-6 Conclusion

In this chapter, I introduced three metabolomics tools developed for AIF data analysis. MS-DIAL provides whole data processing for AIF data with multiple collision energies. In MS-DIAL, AIF MS2 spectra can be deconvoluted by both CorrDec and MS2Dec, visualized, and utilized for compound identification. MS-FINDER performs compound estimation from MS2 spectra by *in silico* fragmentation. For AIF spectra, several adduct ions can be annotated to MS2 peaks in a spectrum by MS-FINDER. MS-LIMA supports management of mass spectra as the MSP format with annotations, compare them with each other. The AIF platform enables user-friendly and reliable data analysis for AIF-based untargeted metabolomics.

The AIF platform and raw AIF data can be downloaded from websites [105,135] and online databases, such as MetaboLights [35] and Metabolome Workbench [65]. Therefore, the AIF platform can help not only AIF users but also many metabolomics researchers to reuse public AIF data.

Chapter 5
Conclusion

This thesis has presented a new data analysis platform for untargeted metabolomics based on All Ion Fragmentation (AIF). Compound identification is a major bottleneck in untargeted metabolomics. MS2 spectra are required for reliable compound identification; moreover, MS2 spectra are necessary to annotate unknown compounds, which are most detected features in samples. Although the performance of mass spectrometers is dramatically improved, the number of MS2 scans are restricted by the scan speed. AIF is a very powerful approach to get comprehensive but complex MS2 spectra from limited MS2 scans by setting quite large *m/z* range (e.g. 40–1200 Da). Therefore, AIF is suitable for untargeted metabolomics to avoid precursor selection bias and annotate as many compounds as possible. However, AIF spectra require computational approaches for interpretation. In other words, AIF spectra consisting of several fragment ions from some precursor ions with different adduct types should be separated and correctly re-assigned to their precursor. Good data analysis methods and tools for AIF have been anticipated. In this study, I have developed a data analysis platform for AIF including a new deconvolution method, reliable chemical library, and user-friendly software.

To overcome the trade-off between comprehensiveness and cleanness of spectra, I have developed CorrDec—a new MS2 spectra deconvolution method for AIF data based on the correlations of the peak intensities across samples. I confirmed that the peak intensities of fragment ions were highly correlated with those of their precursor ions across samples. The high quality of MS2 spectra and the ability to deconvolute completely coeluting compounds are the main advantages over retention-time (RT) based deconvolution methods. Additionally,

CorrDec is useful for compound estimations by i*n silico* fragmentation tools such as MS-FINDER, because CorrDec spectra are generated without any reference MS2 libraries. Although CorrDec requires multiple samples to calculate intensity correlations across samples, it is applicable for almost all untargeted studies because multiple samples are usually measured in untargeted metabolomics for comparison. Of course, it should be noted that the quality and reliability of CorrDec spectra tend to be low in small-scale studies. In any case, manual confirmation is still needed because computational approaches might lead to misinterpretation.

Reliable AMRT+MS2 libraries are necessary to confidently annotate and identify compounds in untargeted metabolomics. I described workflow to obtain accurate mass (AM), RT, and MS2 for a standard using the AIF data acquisition method and provided practical recommendations for library development. AIF spectra are useful as a library because they contain several adduct types. The main advantages of the library are normalized RT and annotated MS2 spectrum including several adduct types. The serial dilution measurements of standards can improve the confidence of measurements, confirm the ionization efficacy and suitable concentration, and be deconvoluted by CorrDec. The construction of high-quality, open-access libraries makes compound annotations more transparent, reliable, and transferable to the broader community.

In Chapter 4, I introduced three metabolomics tools developed for AIF data analysis. MS-DIAL provides whole data processing for AIF data with multiple collision energies. In MS-DIAL, AIF spectra can be deconvoluted by both CorrDec and MS2Dec, visualized, and utilized for compound identification. MS-FINDER performs compound estimation from MS2 spectra by *in silico* fragmentation. For AIF spectra, adduct ions can be annotated to MS2 peaks in a spectrum. MS-LIMA supports management of mass spectra as the MSP format with annotations, compare with each other. The AIF platform enables user-friendly and reliable data analysis for AIF-based untargeted metabolomics.

Reuse of public metabolomics data is still a challenge; however, it is necessary for further development in metabolomics. For accumulating open data in public repositories, data

analysis platform should be prepared and improved. Because AIF data contains all MS2 spectra, it is suitable for further analysis. The AIF platform and raw AIF data can be freely downloaded from websites and online databases/repositories. Therefore, the AIF platform can help not only AIF users but also many metabolomics researchers to reuse public AIF data.

CorrDec is available in MS-DIAL and utilized by some users for DIA data analysis, not only AIF but also SWATH. CorrDec was originally developed for AIF data, so the workflow and several parameters should be modified to adopt for SWATH data if SWATH user requested. MS-DIAL has two different deconvolutions, CorrDec and MS2Dec, which are based on different concepts, and produces each deconvolution result. For easily understanding, the summary and comparison of deconvoluted spectra between CorrDec and MS2Dec will be useful.

The current AIF platform can easily analyze public metabolomics data. However, it is difficult to compare with other data measured by different methods/instruments. To enhance reuse of public data, a comparison workflow and tool should be developed. If researchers can easily compare own data with public metabolomics data, many metabolomics studies will be more advanced.

Bibliography

1. Lederberg, J., and McCray, A. T. 'Ome Sweet 'Omics-- a genealogical treasury of words. *The Scientist* **2001**, *15* (7), 8.
2. Yadav, S. P. The wholeness in suffix -omics, -omes, and the word om. *J. Biomol. Tech.* **2007**, *18* (5), 277.
3. Karahalil, B. Overview of systems biology and omics technologies. *Curr. Med. Chem.* **2016**, *23* (37), 4221–4230. doi:10.2174/0929867323666160926150617.
4. Hans Winkler. *Verbreitung Und Ursache Der Parthenogenesis Im Pflanzen - Und Tierreiche*, Gustav Fischer, Jena, **1920**.
5. Kuska, B. Beer, bethesda, and biology: how "genomics" came into being. *J. Natl. Cancer Inst.* **1998**, *90* (2), 93. doi:10.1093/jnci/90.2.93.
6. Oliver, S. G., Winson, M. K., Kell, D. B., and Baganz, F. Systematic functional analysis of the yeast genome. *Trends Biotechnol.* **1998**, *16* (9), 373–378. doi:10.1016/S0167-7799(98)01214-1.
7. Kell, D. B., and Oliver, S. G. The metabolome 18 years on: a concept comes of age. *Metabolomics* **2016**, *12* (9), 148. doi:10.1007/s11306-016-1108-4.
8. Fiehn, O. Metabolomics - The link between genotypes and phenotypes. *Plant Mol. Biol.* **2002**, *48* (1–2), 155–171. doi:10.1023/A:1013713905833.
9. Patti, G. J., Yanes, O., and Siuzdak, G. Metabolomics: The apogee of the omics trilogy. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263–269. doi:10.1038/nrm3314.
10. Wishart, D. S., Tzur, D., Knox, C., Eisner, R., Guo, A. C., Young, N., Cheng, D., Jewell, K., Arndt, D., Sawhney, S., et al. HMDB: The human metabolome database. *Nucleic Acids Res.* **2007**, *35*, 521–526. doi:10.1093/nar/gkl923.
11. Pauling, L., Robinson, A. B., Teranishi, R., and Cary, P. Quantitative analysis of urine vapor and breath by gas-liquid partition chromatography. *Proc. Natl. Acad. Sci. U. S. A.* **1971**, *68* (10), 2374–2376. doi:10.1073/pnas.68.10.2374.
12. Robinson, A. B., and Robinson, N. E. Origins of metabolic profiling. *Methods Mol. Biol.* **2011**, *708*, 1–23. doi:10.1007/978-1-61737-985-7_1.
13. Saurina, J., and Sentellas, S. Strategies for metabolite profiling based on liquid chromatography. *J. Chromatogr. B* **2017**, *1044–1045*, 103–111. doi:10.1016/j.jchromb.2017.01.011.
14. Adams, R. F. Determination of amino acid profiles in biological samples by gas chromatography. *J. Chromatogr. A* **1974**, *95* (2), 189–212. doi:10.1016/S0021-9673(00)84078-9.
15. Tanaka, K., Hine, D. G., West-Dull, A., and Lynn, T. B. Gas-chromatographic method of analysis for urinary organic acids. I. Retention indices of 155 metabolically important compounds. *Clin. Chem.* **1980**, *26* (13), 1839–1846. doi:10.1093/clinchem/26.13.1839.
16. De Jongh, D. C., Radford, T., Hribar, J. D., Hanessian, S., Bieber, M., Dawson, G., and Sweeley, C. C. Analysis of trimethylsilyl derivatives of carbohydrates by gas chromatography and mass spectrometry. *J. Am. Chem. Soc.* **1969**, *91* (7), 1728–1740. doi:10.1021/ja01035a022.

17. Jellum, E., Kvittingen, E. A., and Stokke, O. Mass spectrometry in diagnosis of metabolic disorders. *Biol. Mass Spectrom.* **1988**, *16* (1–12), 57–62. doi:10.1002/bms.1200160111.
18. Kimura, M., Yamamoto, T., and Yamaguchi, S. Automated metabolic profiling and interpretation of GC/MS data for organic acidemia screening: A personal computer-based system. *Tohoku J. Exp. Med.* **1999**, *188* (4), 317–334. doi:10.1620/tjem.188.317.
19. Fiehn, O., Kopka, J., Dörmann, P., Altmann, T., Trethewey, R. N., and Willmitzer, L. Metabolite profiling for plant functional genomics. *Nat. Biotechnol.* **2000**, *18* (11), 1157–1161. doi:10.1038/81137.
20. Wolfender, J. L., Rodriguez, S., and Hostettmann, K. Liquid chromatography coupled to mass spectrometry and nuclear magnetic resonance spectroscopy for the screening of plant constituents. *J. Chromatogr. A* **1998**, *794* (1–2), 299–316. doi:10.1016/S0021-9673(97)00939-4.
21. Tolstikov, V. V., and Fiehn, O. Analysis of highly polar compounds of plant origin: combination of hydrophilic interaction chromatography and electrospray ion trap mass spectrometry. *Anal. Biochem.* **2002**, *301* (2), 298–307. doi:10.1006/abio.2001.5513.
22. Liu, X., and Locasale, J. W. Metabolomics: A Primer. *Trends Biochem. Sci.* **2017**, *42* (4), 274–284. doi:10.1016/j.tibs.2017.01.004.
23. German, J. B., Hammock, B. D., and Watkins, S. M. Metabolomics: building on a century of biochemistry to guide human health. *Metabolomics* **2005**, *1* (1), 3–9. doi:10.1007/s11306-005-1102-8.
24. Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608–D617. doi:10.1093/nar/gkx1089.
25. Wilson, D. M., and Burlingame, A. L. Deuterium and carbon-13 tracer studies of ethanol metabolism in the rat by 2H, 1H-decoupled 13C nuclear magnetic resonance. *Biochem. Biophys. Res. Commun.* **1974**, *56* (3), 828–835. doi:10.1016/0006-291x(74)90680-9.
26. Wishart, D. S. NMR metabolomics: A look ahead. *J. Magn. Reson.* **2019**, *306*, 155–161. doi:10.1016/j.jmr.2019.07.013.
27. Markley, J. L., Brüschweiler, R., Edison, A. S., Eghbalnia, H. R., Powers, R., Raftery, D., and Wishart, D. S. The future of NMR-based metabolomics. *Curr. Opin. Biotechnol.* **2017**, *43*, 34–40. doi:10.1016/j.copbio.2016.08.001.
28. Brenton, A. G., and Godfrey, A. R. Accurate mass measurement: Terminology and treatment of data. *J. Am. Soc. Mass Spectrom.* **2010**, *21* (11), 1821–1835. doi:10.1016/j.jasms.2010.06.006.
29. Ghaste, M., Mistrik, R., and Shulaev, V. Applications of fourier transform ion cyclotron resonance (FT-ICR) and orbitrap based high resolution mass spectrometry in metabolomics and lipidomics. *Int. J. Mol. Sci.* **2016**, *17* (6). doi:10.3390/ijms17060816.
30. Schollée, J. E., Schymanski, E. L., Stravs, M. A., Gulde, R., Thomaidis, N. S., and Hollender, J. Similarity of high-resolution tandem mass spectrometry spectra of structurally related micropollutants and transformation products. *J. Am. Soc. Mass Spectrom.* **2017**, *28* (12), 2692–2704. doi:10.1007/s13361-017-1797-6.
31. Li, Y., Shrestha, B., and Vertes, A. Atmospheric pressure infrared MALDI imaging mass spectrometry for plant metabolomics. *Appl. Environ. Microbiol.* **2003**, *2* (5), 407. doi:10.1021/ac701703f.
32. Feng, X., Liu, X., Luo, Q., and Liu, B. F. Mass spectrometry in systems biology: An overview. *Mass Spectrom. Rev.* **2008**, *27* (6), 635–660. doi:10.1002/mas.20182.

33. Fujimura, Y., and Miura, D. MALDI mass spectrometry imaging for visualizing in situ metabolism of endogenous metabolites and dietary phytochemicals. *Metabolites* **2014**, *4* (2), 319–346. doi:10.3390/metabo4020319.
34. D'Atri, V., Causon, T., Hernandez-Alba, O., Mutabazi, A., Veuthey, J.-L., Cianferani, S., and Guillarme, D. Adding a new separation dimension to MS and LC-MS: What is the utility of ion mobility spectrometry? *J. Sep. Sci.* **2018**, *41* (1), 20–67. doi:10.1002/jssc.201700919.
35. Kale, N. S., Haug, K., Conesa, P., Jayseelan, K., Moreno, P., Rocca-Serra, P., Nainala, V. C., Spicer, R. A., Williams, M., Li, X., et al. MetaboLights: An open-access database repository for metabolomics data. *Curr. Protoc. Bioinforma.* **2016**, *2016*, 14.13.1-14.13.18. doi:10.1002/0471250953.bi1413s53.
36. Pluskal, T., Castillo, S., Villar-Briones, A., and Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11* (1), 395. doi:10.1186/1471-2105-11-395.
37. Tsugawa, H., Cajka, T., Kind, T., Ma, Y., Higgins, B., Ikeda, K., Kanazawa, M., VanderGheynst, J., Fiehn, O., and Arita, M. MS-DIAL: Data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **2015**, *12* (6), 523–526. doi:10.1038/nmeth.3393.
38. Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., Fan, T. W.-M., Fiehn, O., Goodacre, R., Griffin, J. L., et al. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Inititative (MSI). *Metabolomics* **2007**, *3* (3), 211–221. doi:10.1007/s11306-007-0082-2.Proposed.
39. Griss, J., Jones, A. R., Sachsenberg, T., Walzer, M., Gatto, L., Hartler, J., Thallinger, G. G., Salek, R. M., Steinbeck, C., Neuhauser, N., et al. The mzTab data exchange format: Communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell. Proteomics* **2014**, *13* (10), 2765–2775. doi:10.1074/mcp.O113.036681.
40. Naz, S., Gallart-Ayala, H., Reinke, S. N., Mathon, C., Blankley, R., Chaleckis, R., and Wheelock, C. E. Development of a liquid chromatography-high resolution mass spectrometry metabolomics method with high specificity for metabolite identification using all ion fragmentation acquisition. *Anal. Chem.* **2017**, *89* (15), 7933–7942. doi:10.1021/acs.analchem.7b00925.
41. The MS-DIAL annotation code. Available online: http://prime.psc.riken.jp/compms/msdial/annotationcode.html (accessed in March 2020).
42. Stanstrup, J., Neumann, S., and Vrhovšek, U. PredRet: Prediction of retention time by direct mapping between multiple chromatographic systems. *Anal. Chem.* **2015**, *87* (18), 9421–9428. doi:10.1021/acs.analchem.5b02287.
43. Dührkop, K., Shen, H., Meusel, M., Rousu, J., and Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proc. Natl. Acad. Sci. U. S. A.* **2015**, *112* (41), 12580–12585. doi:10.1073/pnas.1509788112.
44. Ruttkies, C., Schymanski, E. L., Wolf, S., Hollender, J., and Neumann, S. MetFrag relaunched: Incorporating strategies beyond in silico fragmentation. *J. Cheminform.* **2016**, *8* (1), 3. doi:10.1186/s13321-016-0115-9.
45. Lai, Z., Tsugawa, H., Wohlgemuth, G., Mehta, S., Mueller, M., Zheng, Y., Ogiwara, A., Meissen, J., Showalter, M., Takeuchi, K., et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **2018**. doi:10.1038/nmeth.4512.

46. Rung, J., and Brazma, A. Reuse of public genome-wide gene expression data. *Nat. Rev. Genet.* **2013**, *14* (2), 1–11. doi:10.1038/nrg3394.
47. Goodwin, S., McPherson, J. D., and McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17* (6), 333–351. doi:10.1038/nrg.2016.49.
48. Haft, D. H. Using comparative genomics to drive new discoveries in microbiology. *Curr. Opin. Microbiol.* **2015**, *23*, 189–196. doi:10.1016/j.mib.2014.11.017.
49. Yu, J., Blom, J., Glaeser, S. P., Jaenicke, S., Juhre, T., Rupp, O., Schwengers, O., Spänig, S., and Goesmann, A. A review of bioinformatics platforms for comparative genomics. Recent developments of the EDGAR 2.0 platform and its utility for taxonomic and phylogenetic studies. *J. Biotechnol.* **2017**, *261*, 2–9. doi:10.1016/j.jbiotec.2017.07.010.
50. Tada, I., Tanizawa, Y., and Arita, M. Visualization of consensus genome structure without using a reference genome. *BMC Genomics* **2017**, *18*. doi:10.1186/s12864-017-3499-7.
51. Noureen, M., Tada, I., Kawashima, T., and Arita, M. Rearrangement analysis of multiple bacterial genomes. *BMC Bioinformatics* **2019**, *20* (Suppl 23). doi:10.1186/s12859-019-3293-4.
52. Tada, I., Tanizawa, Y., Endo, A., Tohno, M., and Arita, M. Revealing the genomic differences between two subgroups in Lactobacillus gasseri. *Biosci. Microbiota, Food Heal.* **2017**, *36* (4), 155–159. doi:10.12938/bmfh.17-006.
53. Tanizawa, Y., Tada, I., Kobayashi, H., Endo, A., Maeno, S., Toyoda, A., Arita, M., Nakamura, Y., Sakamoto, M., Ohkuma, M., et al. Lactobacillus paragasseri sp. nov., a sister taxon of Lactobacillus gasseri, based on whole-genome sequence analyses. *Int. J. Syst. Evol. Microbiol.* **2018**, *68* (11), 3512–3517. doi:10.1099/ijsem.0.003020.
54. Rocca-Serra, P., Salek, R. M., Arita, M., Correa, E., Dayalan, S., Gonzalez-Beltran, A., Ebbels, T., Goodacre, R., Hastings, J., Haug, K., et al. Data standards can boost metabolomics research, and if there is a will, there is a way. *Metabolomics* **2016**, *12* (1), 1–13. doi:10.1007/s11306-015-0879-3.
55. Haug, K., Salek, R. M., and Steinbeck, C. Global open data management in metabolomics. *Curr. Opin. Chem. Biol.* **2017**, *36*, 58–63. doi:10.1016/j.cbpa.2016.12.024.
56. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., et al. PubChem substance and compound databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213. doi:10.1093/nar/gkv951.
57. Pence, H. E., and Williams, A. Chemspider: An online chemical information resource. *J. Chem. Educ.* **2010**, *87* (11), 1123–1124. doi:10.1021/ed100697w.
58. Nakamura, Y., Afendi, F. M., Parvin, A. K., Ono, N., Tanaka, K., Hirai Morita, A., Sato, T., Sugiura, T., Altaf-Ul-Amin, M., and Kanaya, S. KNApSAcK Metabolite Activity Database for retrieving the relationships between metabolites and biological activities. *Plant Cell Physiol.* **2014**, *55* (1), e7. doi:10.1093/pcp/pct176.
59. Horai, H., Arita, M., Kanaya, S., Nihei, Y., Ikeda, T., Suwa, K., Ojima, Y., Tanaka, K., Tanaka, S., Aoshima, K., et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **2010**, *45* (7), 703–714. doi:10.1002/jms.1777.
60. Guijas, C., Montenegro-Burke, J. R., Domingo-Almenara, X., Palermo, A., Warth, B., Hermann, G., Koellensperger, G., Huan, T., Uritboonthai, W., Aisporna, A. E., et al. METLIN: A technology platform for identifying knowns and unknowns. *Anal. Chem.* **2018**, *90* (5), 3156–3164. doi:10.1021/acs.analchem.7b04424.

61. Wang, M., Carver, J. J., Phelan, V. V, Sanchez, L. M., Garg, N., Peng, Y., Nguyen, D. D., Watrous, J., Kapono, C. A., Luzzatto-Knaan, T., et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **2016**, *34* (8), 828–837. doi:10.1038/nbt.3597.
62. NIST Standard Reference Database 1A v17. Available online: https://www.nist.gov/srd/nist-standard-reference-database-1a-v17 (accessed in March 2020).
63. MetabolomeXchange. Available online: http://www.metabolomexchange.org/site/ (accessed in March 2020).
64. Ara, T., Enomoto, M., Arita, M., Ikeda, C., Kera, K., Yamada, M., Nishioka, T., Ikeda, T., Nihei, Y., Shibata, D., et al. Metabolonote: A wiki-based database for managing hierarchical metadata of metabolome analyses. *Front. Bioeng. Biotechnol.* **2015**, *3* (APR), 38. doi:10.3389/fbioe.2015.00038.
65. Sud, M., Fahy, E., Cotter, D., Azam, K., Vadivelu, I., Burant, C., Edison, A., Fiehn, O., Higashi, R., Sreekumaran Nair, K., et al. Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **2015**, *44* (8). doi:10.1093/nar/gkv1042.
66. Sakurai, N., and Shibata, D. Tools and databases for an integrated metabolite annotation environment for liquid chromatography–mass spectrometry-based untargeted Metabolomics. *Carotenoid Sci.* **2017**, *22*, 16–22.
67. Food Metabolome Repository. Available online: http://metabolites.in/foods/ (accessed in March 2020).
68. Zhu, X., Chen, Y., and Subramanian, R. Comparison of information-dependent acquisition, SWATH, and MS All techniques in metabolite identification study employing ultrahigh-performance liquid chromatography-quadrupole time-of-flight mass spectrometry. *Anal. Chem.* **2014**, *86* (2), 1202–1209. doi:10.1021/ac403385y.
69. Wang, R., Yin, Y., and Zhu, Z. J. Advancing untargeted metabolomics using data-independent acquisition mass spectrometry technology. *Anal. Bioanal. Chem.* **2019**, *411* (19), 4349–4357. doi:10.1007/s00216-019-01709-1.
70. Villas-boas, S. G., Rossener, U., Hansen, M. A. E., Smedsgaard, J., and Nielsen, J. *Metabolome Analysis: An Introduction*, Wiley, **2007**.
71. Vaidyanathan, S., Harrigan, G. G., and Goodacre, R. *Metabolome Analyses: Strategies for Systems Biology*, Springer Nature, Boston, **2005**. doi:10.1007/0-387-25240-1.
72. Chaleckis, R., Meister, I., Zhang, P., and Wheelock, C. E. Challenges, progress and promises of metabolite annotation for LC–MS-based metabolomics. *Curr. Opin. Biotechnol.* **2019**, *55*, 44–50. doi:10.1016/J.COPBIO.2018.07.010.
73. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Mol. Case Stud.* **2015**, *1* (1), a000588. doi:10.1101/mcs.a000588.
74. Kell, D. B. Metabolomics and systems biology: Making sense of the soup. *Curr. Opin. Microbiol.* **2004**, *7* (3), 296–307. doi:10.1016/j.mib.2004.04.012.
75. Tada, I., Chaleckis, R., Tsugawa, H., Meister, I., Zhang, P., Lazarinis, N., Dahlén, B., Wheelock, C. E., and Arita, M. Correlation-Based Deconvolution (CorrDec) To Generate High-Quality MS2 Spectra from Data-Independent Acquisition in Multisample Studies. *Anal. Chem.* **2020**, *92* (16), 11310–11317. doi:10.1021/acs.analchem.0c01980.
76. Tsugawa, H., Kind, T., Nakabayashi, R., Yukihira, D., Tanaka, W., Cajka, T., Saito, K., Fiehn, O., and Arita, M. Hydrogen rearrangement rules: Computational MS/MS fragmentation and structure elucidation using MS-FINDER software. *Anal. Chem.* **2016**, *88* (16), 7946–7958. doi:10.1021/acs.analchem.6b00770.

77. Röst, H. L., Rosenberger, G., Navarro, P., Gillet, L., Miladinović, S. M., Schubert, O. T., Wolski, W., Collins, B. C., Malmström, J., Malmström, L., et al. OpenSWATH enables automated, targeted analysis of data-independent acquisition MS data. *Nat. Biotechnol.* **2014**, *32*, 219–223. doi:10.1038/nbt.2840.
78. Peckner, R., Myers, S. A., Jacome, A. S. V., Egertson, J. D., Abelin, J. G., MacCoss, M. J., Carr, S. A., and Jaffe, J. D. Specter: linear deconvolution for targeted analysis of data-independent acquisition mass spectrometry proteomics. *Nat. Methods* **2018**, *15* (5), 371–378. doi:10.1038/nmeth.4643.
79. Li, H., Cai, Y., Guo, Y., Chen, F., and Zhu, Z. J. MetDIA: Targeted metabolite extraction of multiplexed MS/MS spectra generated by data-independent acquisition. *Anal. Chem.* **2016**, *88* (17), 8757–8764. doi:10.1021/acs.analchem.6b02122.
80. Yin, Y., Wang, R., Cai, Y., Wang, Z., and Zhu, Z.-J. DecoMetDIA: Deconvolution of multiplexed MS/MS spectra for metabolite identification in SWATH-MS-based untargeted metabolomics. *Anal. Chem.* **2019**, *91* (18), 11897–11904. doi:10.1021/acs.analchem.9b02655.
81. Brown, M., Wedge, D. C., Goodacre, R., Kell, D. B., Baker, P. N., Kenny, L. C., Mamas, M. A., Neyses, L., and Dunn, W. B. Automated workflows for accurate mass-based putative metabolite identification in LC/MS-derived metabolomic datasets. *Bioinformatics* **2011**. doi:10.1093/bioinformatics/btr079.
82. Alonso, A., Marsal, S., and Julià, A. Analytical methods in untargeted metabolomics: State of the art in 2015. *Front. Bioeng. Biotechnol.* **2015**, *3* (MAR). doi:10.3389/fbioe.2015.00023.
83. Kuhl, C., Tautenhahn, R., Bö, C., Larson, T. R., and Neumann, S. CAMERA: An integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal. Chem.* **2011**. doi:10.1021/ac202450g.
84. Broeckling, C. D., Heuberger, A. L., Prince, J. A., Ingelsson, E., and Prenni, J. E. Assigning precursor–product ion relationships in indiscriminant MS/MS data from non-targeted metabolite profiling studies. *Metabolomics* **2013**, *9* (1), 33–43. doi:10.1007/s11306-012-0426-4.
85. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A., and Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral- Matching-Based Annotation for Metabolomics Data. *Anal. Chem.* **2014**, *86*, 6812–6817. doi:10.1021/ac501530d.
86. Domingo-Almenara, X., Montenegro-Burke, J. R., Benton, H. P., and Siuzdak, G. Annotation: A Computational solution for streamlining metabolomics analysis. *Anal. Chem.* **2017**, acs.analchem.7b03929. doi:10.1021/acs.analchem.7b03929.
87. Domingo-Almenara, X., Montenegro-Burke, J. R., Guijas, C., Majumder, E. L. W., Benton, H. P., and Siuzdak, G. Autonomous METLIN-guided in-source fragment annotation for untargeted metabolomics. *Anal. Chem.* **2019**, *91* (5), 3246–3253. doi:10.1021/acs.analchem.8b03126.
88. Tsuruta, Y., Tomida, H., Kohashi, K., and Ohkura, Y. Simultaneous determination of imidazoleacetic acid and N tau- and N pi-methylimidazoleacetic acids in human urine by high-performance liquid chromatography with fluorescence detection. *J. Chromatogr.* **1987**, *416* (1), 63–69. doi:10.1016/0378-4347(87)80485-1.
89. Pohjanpelto, P., Niemi, K., and Sarmela, T. Anterior chamber haemorrhage in the newborn after spontaneous delivery. A case report. *Acta Ophthalmol.* **1979**, *57* (3), 443–446. doi:10.1111/j.1755-3768.1979.tb01827.x.

90. Bouatra, S., Aziat, F., Mandal, R., Guo, A. C., Wilson, M. R., Knox, C., Bjorndahl, T. C., Krishnamurthy, R., Saleem, F., Liu, P., et al. The human urine metabolome. *PLoS One* **2013**, *8* (9), e73076. doi:10.1371/journal.pone.0073076.
91. Li, C., Homma, M., and Oka, K. Characteristics of delayed excretion of flavonoids in human urine after administration of Shosaiko-to, a herbal medicine. *Biol. Pharm. Bull.* **1998**, *21* (12), 1251–1257. doi:10.1248/bpb.21.1251.
92. Hornik, P., Vyskocilová, P., Friedecký, D., and Adam, T. Diagnosing AICA-ribosiduria by capillary electrophoresis. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* **2006**, *843* (1), 15–19. doi:10.1016/j.jchromb.2006.05.020.
93. Tsugawa, H., Nakabayashi, R., Mori, T., Yamada, Y., Takahashi, M., Rai, A., Sugiyama, R., Yamamoto, H., Nakaya, T., Yamazaki, M., et al. A cheminformatics approach to characterize metabolomes in stable-isotope-labeled organisms. *Nat. Methods* **2019**, *16* (4), 295–298. doi:10.1038/s41592-019-0358-2.
94. Chaleckis, R., Naz, S., Meister, I., and Wheelock, C. E. LC-MS-based metabolomics of biofluids using all-ion fragmentation (AIF) acquisition. In *Clinical Metabolomics*; Humana Press, New York, NY, **2018**; pp 45–58. doi:10.1007/978-1-4939-7592-1_3.
95. Lazarinis, N., Bood, J., Gomez, C., Kolmert, J., Lantz, A.-S., Gyllfors, A., Davis, A., Wheelock, C. E., Dahl en, S.-E., Dahl en, B., et al. Leukotriene E 4 induces airflow obstruction and mast cell activation through the cysteinyl leukotriene type 1 receptor. *J. Allergy Clin. Immunol.* **2018**. doi:10.1016/j.jaci.2018.02.024.
96. Tada, I., Tsugawa, H., Meister, I., Zhang, P., Shu, R., Katsumi, R., Wheelock, C. E., Arita, M., and Chaleckis, R. Creating a reliable mass spectral-retention time library for all ion fragmentation-based metabolomics. *Metabolites* **2019**, *9* (11). doi:10.3390/metabo9110251.
97. Moorthy, A. S., Wallace, W. E., Kearsley, A. J., Tchekhovskoi, D. V., and Stein, S. E. Combining fragment-ion and neutral-loss matching during mass spectral library searching: A mew general purpose algorithm applicable to illicit drug identification. *Anal. Chem.* **2017**, *89* (24), 13261–13268. doi:10.1021/acs.analchem.7b03320.
98. Rinschen, M. M., Ivanisevic, J., Giera, M., and Siuzdak, G. Identification of bioactive metabolites using activity metabolomics. *Nat. Rev. Mol. Cell Biol.* **2019**, *20* (6), 353–367. doi:10.1038/s41580-019-0108-4.
99. Cajka, T., and Fiehn, O. Toward Merging Untargeted and Targeted Methods in Mass Spectrometry-Based Metabolomics and Lipidomics. doi:10.1021/acs.analchem.5b04491.
100. Chamkasem, N., Ollis, L. W., Harmon, T., Lee, S., and Mercer, G. Analysis of 136 pesticides in avocado using a modified QuEChERS method with LC-MS/MS and GC-MS/MS. *J. Agric. Food Chem.* **2013**, *61* (10), 2315–2329. doi:10.1021/jf304191c.
101. Nikolskiy, I., Mahieu, N. G., Chen, Y. J., Tautenhahn, R., and Patti, G. J. An untargeted metabolomic workflow to improve structural characterization of metabolites. *Anal. Chem.* **2013**, *85* (16), 7713–7719. doi:10.1021/ac400751j.
102. Vinaixa, M., Schymanski, E. L., Neumann, S., Navarro, M., Salek, R. M., and Yanes, O. Mass spectral databases for LC/MS- and GC/MS-based metabolomics: State of the field and future prospects. *TrAC Trends Anal. Chem.* **2016**, *78*, 23–35. doi:10.1016/j.trac.2015.09.005.
103. Bruderer, T., Varesio, E., Hidasi, A. O., Duchoslav, E., Burton, L., Bonner, R., and Hopfgartner, G. Metabolomic spectral libraries for data-independent SWATH liquid chromatography mass spectrometry acquisition. *Anal. Bioanal. Chem.* **2018**, *410* (7), 1873–1884. doi:10.1007/s00216-018-0860-x.
104. Sentandreu, E., Peris-Díaz, M. D., Sweeney, S. R., Chiou, J., Muñoz, N., and Tiziani, S. A survey of orbitrap all ion fragmentation analysis assessed by an R MetaboList

package to study small-molecule metabolites. *Chromatographia* **2018**, *81* (7), 981–994. doi:10.1007/s10337-018-3536-y.
105. GitHub Repository of MS-LIMA. Available online: https://github.com/tipputa/MS-LIMA-Standard (accessed in February 2020).
106. Supplemental Tutorial Material in Chapter 3. Available online: http://prime.psc.riken.jp/compms/msdial/aiftutorial.zip (accessed in March 2020).
107. Dunn, W. B., Erban, A., Weber, R. J. M., Creek, D. J., Brown, M., Breitling, R., Hankemeier, T., Goodacre, R., Neumann, S., Kopka, J., et al. Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics* **2013**, *9* (SUPPL.1), 44–66. doi:10.1007/s11306-012-0434-4.
108. Hastings, J., Owen, G., Dekker, A., Ennis, M., Kale, N., Muthukrishnan, V., Turner, S., Swainston, N., Mendes, P., and Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**, *44* (D1), D1214-9. doi:10.1093/nar/gkv1031.
109. Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. InChI, the IUPAC international chemical identifier. *J. Cheminform.* **2015**, *7* (1), 23. doi:10.1186/s13321-015-0068-4.
110. Szöcs, E., Münch, D., and Ranke, J. {webchem}: retrieve chemical information from the web. *Zenodo* **2015**. doi:10.5281/zenodo.33823.
111. Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T., and Fiehn, O. The chemical translation service--a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **2010**, *26* (20), 2647–2648. doi:10.1093/bioinformatics/btq476.
112. Kärkkäinen, O., Lankinen, M. A., Vitale, M., Jokkala, J., Leppänen, J., Koistinen, V., Lehtonen, M., Giacco, R., Rosa-Sibakov, N., Micard, V., et al. Diets rich in whole grains increase betainized compounds associated with glucose metabolism. *Am. J. Clin. Nutr.* **2018**, *108* (5), 971–979. doi:10.1093/ajcn/nqy169.
113. Pozo, O. J., Van Eenoo, P., Deventer, K., and Delbeke, F. T. Ionization of anabolic steroids by adduct formation in liquid chromatography electrospray mass spectrometry. *J. Mass Spectrom.* **2007**, *42* (4), 497–516. doi:10.1002/jms.1182.
114. Pluskal, T., Nakamura, T., Villar-Briones, A., and Yanagida, M. Metabolic profiling of the fission yeast S. pombe: Quantification of compounds under different temperatures and genetic perturbation. *Mol. Biosyst.* **2010**, *6* (1), 182–198. doi:10.1039/b908784b.
115. Liigand, P., Kaupmees, K., Haav, K., Liigand, J., Leito, I., Girod, M., Antoine, R., and Kruve, A. Think negative: Finding the best electrospray ionization/MS mode for your analyte. *Anal. Chem.* **2017**, *89* (11), 5665–5668. doi:10.1021/acs.analchem.7b00096.
116. Abate-Pella, D., Freund, D. M., Ma, Y., Simón-Manso, Y., Hollender, J., Broeckling, C. D., Huhman, D. V, Krokhin, O. V, Stoll, D. R., Hegeman, A. D., et al. Retention projection enables accurate calculation of liquid chromatographic retention times across labs and methods. *J. Chromatogr. A* **2015**, *1412*, 43–51. doi:10.1016/j.chroma.2015.07.108.
117. Baker, J. K., and Ma, C. Y. Retention index scale for liquid-liquid chromatography. *J. Chromatogr. A* **1979**, *169* (C), 107–115. doi:10.1016/0021-9673(75)85036-9.
118. Bogusz, M., and Aderjan, R. Corrected retention indices in HPLC: their use for the identification of acidic and neutral drugs. *J. Anal. Toxicol.* **1988**, *12* (2), 67–72. doi:10.1093/jat/12.2.67.
119. Zhang, T., Creek, D. J., Barrett, M. P., Blackburn, G., and Watson, D. G. Evaluation of coupling reversed phase, aqueous normal phase, and hydrophilic interaction liquid chromatography with Orbitrap mass spectrometry for metabolomic studies of human urine. *Anal. Chem.* **2012**, *84* (4), 1994–2001. doi:10.1021/ac2030738.

120. Zhu, Q. F., Zhang, T. Y., Qin, L. L., Li, X. M., Zheng, S. J., and Feng, Y. Q. Method to calculate the retention index in hydrophilic interaction liquid chromatography using normal fatty acid derivatives as calibrants. *Anal. Chem.* **2019**, *91* (9), 6057–6063. doi:10.1021/acs.analchem.9b00598.
121. Smith, R., Ventura, D., and Prince, J. T. LC-MS alignment in theory and practice: a comprehensive algorithmic review. *Brief. Bioinform.* **2015**, *16* (1), 104–117. doi:10.1093/bib/bbt080.
122. Guo, X., Bruins, A. P., and Covey, T. R. Characterization of typical chemical background interferences in atmospheric pressure ionization liquid chromatography-mass spectrometry. *Rapid Commun. Mass Spectrom.* **2006**, *20* (20), 3145–3150. doi:10.1002/rcm.2715.
123. Keller, B. O., Sui, J., Young, A. B., and Whittal, R. M. Interferences and contaminants encountered in modern mass spectrometry. *Anal. Chim. Acta* **2008**, *627* (1), 71–81. doi:10.1016/j.aca.2008.04.043.
124. Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., Mahendraker, T., Williams, M., Neumann, S., Rocca-Serra, P., et al. MetaboLights - An open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **2013**, *41* (D1), D781. doi:10.1093/nar/gks1004.
125. MassBank of North America. Available online: https://mona.fiehnlab.ucdavis.edu/ (accessed in March 2020).
126. Gowda, H., Ivanisevic, J., Johnson, C. H., Kurczy, M. E., Benton, H. P., Rinehart, D., Nguyen, T., Ray, J., Kuehl, J., Arevalo, B., et al. Interactive XCMS online: Simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal. Chem.* **2014**, *86* (14), 6931–6939. doi:10.1021/ac500734c.
127. Andreev, V. P., Rejtar, T., Chen, H.-S., Moskovets, E. V., Ivanov, A. R., Karger, B. L., Schütte, C., Scheel, D., and Clemens, S. A universal denoising and peak picking algorithm for LC−MS based on matched filtration in the chromatographic time domain. *Anal. Chem.* **2003**, *75* (22), 6314–6326. doi:10.1021/ac0301806.
128. Ahmed, Z., Mayr, M., Zeeshan, S., Dandekar, T., Mueller, M. J., Fekete, A., Beecher, C. W. W., Garrett, T. J., and Yost, R. A. Lipid-Pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms. *Bioinformatics* **2015**, *31* (7), 1150–1153. doi:10.1093/bioinformatics/btu796.
129. Li, Z., Lu, Y., Guo, Y., Cao, H., Wang, Q., and Shui, W. Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection. *Anal. Chim. Acta* **2018**, *1029*, 50–57. doi:10.1016/j.aca.2018.05.001.
130. Yu, Y.-J., Zheng, Q.-X., Zhang, Y.-M., Zhang, Q., Zhang, Y.-Y., Liu, P.-P., Lu, P., Fan, M.-J., Chen, Q.-S., Bai, C.-C., et al. Automatic data analysis workflow for ultra-high performance liquid chromatography-high resolution mass spectrometry-based metabolomics. *J. Chromatogr. A* **2019**, *1585*, 172–181. doi:10.1016/J.CHROMA.2018.11.070.
131. Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., Okahashi, N., Yamada, Y., Tada, I., Bonini, P., et al. A lipidome atlas in MS-DIAL 4. *Nat. Biotechnol.* **2020**, 1–5. doi:10.1038/s41587-020-0531-2.
132. Metabolomics Society Forum. Available online: http://www.metabolomics-forum.com/index.php (accessed in March 2020).
133. Ohashi, K., Chaleckis, R., Takaine, M., Wheelock, C. E., and Yoshida, S. Kynurenine aminotransferase activity of Aro8/Aro9 engage tryptophan degradation by producing

kynurenic acid in Saccharomyces cerevisiae. *Sci. Rep.* **2017**, *7* (1), 1–8. doi:10.1038/s41598-017-12392-6.
134. MS-DIAL online tutrial. Available online: https://mtbinfo-team.github.io/mtbinfo.github.io/MS-DIAL/tutorial (accessed in March 2020).
135. Computational mass spectrometry (CompMS). Available online: http://prime.psc.riken.jp/compms/index.html (accessed in June 2020).

Publication Records by the Author

1. **Tada, I.**, Tanizawa, Y., and Arita, M. Visualization of Consensus Genome Structure without Using a Reference Genome. *BMC Genomics* **2017**, *18*. doi:10.1186/s12864-017-3499-7.

2. **Tada, I.**, Tanizawa, Y., Endo, A., Tohno, M., and Arita, M. Revealing the Genomic Differences between Two Subgroups in Lactobacillus Gasseri. *Bioscience of Microbiota, Food and Health*. BMFH Press 2017, pp 155–159. doi:10.12938/bmfh.17-006.

3. Tanizawa, Y., **Tada, I.**, Kobayashi, H., Endo, A., Maeno, S., Toyoda, A., Arita, M., Nakamura, Y., Sakamoto, M., Ohkuma, M., et al. Lactobacillus Paragasseri Sp. Nov., a Sister Taxon of Lactobacillus Gasseri, Based on Whole-Genome Sequence Analyses. *Int. J. Syst. Evol. Microbiol.* **2018**, *68* (11), 3512–3517. doi:10.1099/ijsem.0.003020.

4. **Tada, I.**, Tsugawa, H., Meister, I., Zhang, P., Shu, R., Katsumi, R., Wheelock, C. E., Arita, M., and Chaleckis, R. Creating a Reliable Mass Spectral–Retention Time Library for All Ion Fragmentation-Based Metabolomics. *Metabolites* **2019**, *9* (11). doi:10.3390/metabo9110251.

5. Noureen, M., **Tada, I.**, Kawashima, T., and Arita, M. Rearrangement Analysis of Multiple Bacterial Genomes. *BMC Bioinformatics* **2019**, *20* (Suppl 23). doi:10.1186/s12859-019-3293-4.

6. Tsugawa, H., Ikeda, K., Takahashi, M., Satoh, A., Mori, Y., Uchino, H., Okahashi, N., Yamada, Y., **Tada, I.**, Bonini, P., Higashi, Y., Okazaki, Y., Zhou, Z., Zhu Z. J., Koelmel, J., Cajka, T., Fiehn, O., Saito, K., Arita, Masanori, and Arita, Makoto. A lipidome atlas in MS-DIAL 4, *Nature Biotech.* **2020**. doi:10.1038/s41587-020-0531-2.

7. **Tada, I.**, Chaleckis, R., Tsugawa, H., Meister, I., Zhang, P., Lazarinis, N., Dahlén, B., Wheelock, C. E., and Arita, M. Correlation-based Deconvolution (CorrDec) to generate high-quality MS2 spectra from data-independent acquisition in multi-sample studies. *Anal. Chem.* **2020**, *92* (16), 11310-11317. doi: 10.1021/acs.analchem.0c01980.