Note

MassBase: A large-scaled depository of mass spectrometry datasets for metabolome analysis

Takeshi Ara^{1,*}, Nozomu Sakurai^{1,b}, Hideyuki Suzuki^{1,c}, Koh Aoki^{1,2},
Kazuki Saito^{1,3,d}, Daisuke Shibata^{1,4}

¹Kazusa DNA Research Institute, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan; ²Graduate School of Life and Environmental Sciences, Osaka Prefecture University, 1-1 Gakuencho, Naka, Sakai, Osaka 599-8531, Japan; ³Graduate School of Pharmaceutical Sciences, Chiba University, 1-8-1 Inohana, Chuo, Chiba, Chiba 260-8675, Japan; ⁴Department of Biological Sciences, Kisarazu Campus of Tohoku University, 2-6-7 Kazusa-Kamatari, Kisarazu, Chiba 292-0818, Japan

*E-mail: ara@rishi.kyoto-u.ac.jp Tel: +81-774-38-3653 Fax: +81-774-38-3682

Received August 6, 2020; accepted September 11, 2020 (Edited by M. Mizutani)

Abstract Depository of low-molecular-weight compounds or metabolites detected in various organisms in a non-targeted manner is indispensable for metabolomics research. Due to the diverse chemical compounds, various mass spectrometry (MS) setups with state-of-the-art technologies have been used. Over the past two decades, we have analyzed various biological samples by using gas chromatography-mass spectrometry, liquid chromatography-mass spectrometry, or capillary electrophoresis-mass spectrometry, and archived the datasets in the depository MassBase (<http://webs2.kazusa.or.jp/massbase/>). As the format of MS datasets depends on the MS setup used, we converted each raw binary dataset of the mass chromatogram to text file format, and thereafter, information of the chromatograph peak was extracted in the text file from the converted file. In total, the depository comprises 46,493 datasets, of which 38,750 belong to the plant species and 7,743 are authentic or mixed chemicals as well as other sources (microorganisms, animals, and foods), as on August 1, 2020. All files in the depository can be downloaded in bulk from the website. Mass chromatograms of 90 plant species obtained by LC-Fourier transform ion cyclotron resonance MS or Orbitrap MS, which detect the ionized molecules with high accuracy allowing speculation of chemical compositions, were converted to text files by the software PowerGet, and the chemical annotation of each peak was added. The processed datasets were deposited in the annotation database KomicMarket2 (<http://webs2.kazusa.or.jp/km2/>). The archives provide fundamental resources for comparative metabolomics and functional genomics, which may result in deeper understanding of living organisms.

Key words: database, mass spectrometry, metabolome.

As the total number of unique metabolites in the plant, animal, and microbial kingdoms is estimated to be over one million (Afendi et al. 2012), archiving metabolite information in databases is challenging for metabolomics research. The differential analysis of whole metabolites of an organism (metabolome) under environmental conditions or of metabolomes between distinct species is essential to understand the gene function and regulatory mechanisms in the metabolic pathways (Bino et al. 2004; Fiehn et al. 2000; Quanbeck et al. 2012). The minimum information about a metabolomics experiment (MIAMET) has been proposed as the standardized format for metadata in metabolomics by the Metabolomics Standards Initiative (MSI); however,

standardization of the format is still in progress due to the continuous advancement of metabolomics technologies (Fiehn et al. 2007; Haug et al. 2017; Spicer et al. 2017). As public databases, MetaboLights (Steinbeck et al. 2012) and Metabolomics Workbench (Sud et al. 2016) have been used as data depositories for the metabolite information before publishing the articles that include metabolite analysis, the dataset amount is limited when compared to the genome and transcriptome datasets. In total, 2,087 datasets were publicly available from MetaboLights, Metabolomics Workbench, Metabolonote (Ara et al. 2015), and Metabolomic Repository Bordeaux on July 20, 2020, as reported by the international data aggregation and notification service

Abbreviations: MS, Mass spectrometry; GC-MS, gas chromatography-mass spectrometry; LC-MS, liquid chromatography-mass spectrometry; CE-MS, capillary electrophoresis-mass spectrometry; SMS, scanned mass spectra; GC-TOF-MS, gas chromatography/time-of-flight mass spectrometry; MST, mass spectral tag.

^a Present address: Research Institute for Sustainable Humansphere, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan

^b Present address: Bioinformation and DDBJ Center, National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

^c Present address: Department of Research and Development, Hirata Corporation, 111 Hitotsugi, Ueki, Kita, Kumamoto, Kumamoto 861-0198, Japan

^d Present address: RIKEN Center for Sustainable Resource Science, 1-7-22 Suehirocho, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

This article can be found at <http://www.jspcmb.jp/>

Published online March 11, 2021

for metabolomics MetabolomeXchange (<http://www.metabolomexchange.org/>).

Over two decades, we developed several tools and constructed databases for metabolomics, which are available on the Kazusa Metabolomics Portal KOMICS (<http://www.kazusa.or.jp/komics/>) at the Kazusa DNA Research Institute (Sakurai et al. 2014). Here we report a large-scaled depository of mass spectrometry datasets for metabolome analysis, MassBase (<http://webs2.kazusa.or.jp/massbase/>), which is accessible from the KOMICS site.

The MassBase was designed to archive information of low-molecular-weight compounds or metabolites found in the biological samples obtained via gas chromatography-mass spectrometry (GC-MS), liquid chromatography-mass spectrometry (LC-MS), and/or capillary electrophoresis-mass spectrometry (CE-MS) measurements in a non-targeted manner. The binary mass chromatogram datasets from these MS machines with distinct data formats are converted to text files in the unified format. The information of chromatogram peaks, chemical annotation, and experimental metadata (sample name, method name) is described in the other text file. All datasets in the MassBase can be freely accessed and downloaded in bulk. The details of the contents and the analyzed datasets deposited in the MassBase are described below.

We used two types of data formats to describe the information of MS analyses. Although various file formats for MS data (e.g., netCDF, mzXML) are available, we used simple text formats through which the metabolite datasets can be applied for tools and databases listed in the KOMICS (Figure 1). Raw mass chromatogram signals are converted into the text format, namely, scanned mass spectra (SMS), in a unified manner, respectively, from the distinct formats of vender's MS machines. SMS format file contains mass signal intensities in all scans ordered along with the retention time of the scan points. The header line in the SMS format file includes the experimental metadata such as instrument and software used for the measurement. From SMS files, information of chromatogram peaks is extracted and is then described in a text format, namely mass spectral tag (MST) format. MST format file comprises all peak models from a corresponding mass chromatogram. Each peak model is represented by three parts, i) a header line that contains information of peak properties such as peak identifier, retention time, base peak, retention index, peak intensity, and peak identifier of parent ion in case of MS/MS spectra, ii) a fragmentation pattern represented by continuous list of mass-to-charge ratio (m/z) and intensity pairs, and iii) a delimiter line. The metadata of each mass chromatogram (sample name, method name) are simply described. The details of the metadata format are summarized in

Figure 1. Workflow for metabolome analysis coupled with genome and transcriptome datasets. MassBase comprises three types of mass spectrometry data (binary raw data (RAW), text-formatted mass chromatogram data (SMS), and text-formatted peak data (MST)) as metabolomics datasets. Several commercial and free tools are used to analyze each type of metabolome dataset. Some tools support automated and/or manual metabolite annotation using public metabolite databases after peak detection process. Annotated metabolites are presumably linked to related enzymes and/or genes described in the genome and transcriptome datasets by pathway databases such as KEGG. KaPPA-View can integrate these metabolome datasets with genome and transcriptome datasets.

Supplementary Table S1.

The SMS and MST files are generated as described below. Each raw data file from an MS setup is renamed by using a unique identifier with a sequential number (e.g., MDGC1_00001), that is, MassBase accession number. The SMS format files are converted from the corresponding text files exported by a commercial software (ChromaTOF (LECO Co., USA) for GC-MS data and MassLynx (Waters Co., MA, USA) for UPLC-MS data, respectively) or in-house programs (MSGGet for other LC-MS data and ChemStationParser for CE-MS data). Moreover, the MST format files are converted from the corresponding text files exported by a commercial software (ChromaTOF (LECO Co., USA) for GC-MS data) or in-house programs (PowerFT for LC-MS data, SpiceHIT for CE-MS data; available from KOMICS

(<http://www.kazusa.or.jp/komics/>). To browse each mass spectrum graphically on the website, PDF formatted files are created using in-house Perl scripts. All peak data and metadata are registered into a MySQL (<http://www.mysql.com/>) database system. The MassBase Web interface is runs on an Apache web server (<http://www.apache.org/>) with PHP 5.3.3 (<http://www.php.net/>) on the Linux system (Red Hat release 4.8).

The datasets in the MassBase comprise 39,033 biological samples from 165 species on August 1 of 2020 (Table 1), analyzed by eight MS setups; 8,331 gas chromatography/time-of-flight mass spectrometry (GC-TOF-MS) data obtained by the Pegasus III system (LECO Co., USA); 29,442 CE-MS data obtained by the LC/MSD SL (Agilent Technologies, USA); and 8,720 LC-MS data obtained by multiple instruments. The LC-MS instruments include LCQ DECA XP plus or LTQ or LTQ-FT or LTQ Orbitrap XL (Thermo Fisher Scientific Inc., San Jose, CA, USA) connected to the LC system of Agilent 1100/1200 system (Agilent Technologies, USA), and Q-TOF Premier or Micromass LCT Premier (Waters Co., MA, USA) connected to the LC system of Acquity UPLC (Waters Co., MA, USA). Of the 39,033 samples, 38,750 samples (99.2%) were from 116 plant species that include the model plant *Arabidopsis thaliana*, wild plants, crops, vegetables, fruits, and algae. The depository also contains 7,460 authentic or mixed chemicals used for chemical identification and 283 datasets from microorganisms, animals, and various foods (Table 1). Among the plant datasets, 20 species were reported in other metabolomics data repositories (MetaboLights, Metabolomics Workbench, and/or GNPS); however, 96 species are unique in the MassBase (Supplementary Table S2).

The main page of the MassBase contains an interface to search and browse the mass spectrometry experimental datasets (Figure 2). The accession number (e.g., MDGC1_00001), sample name (species or compound name; e.g., *Arabidopsis* or Glucose) and sample condition name (any treatment; e.g., wounding) can be used as search terms. All hit records are listed in the search result page. Each MST information page is

displayed by clicking the accession number in the search result page. The MST information page is available for each mass spectrometry measurement. It contains 1) the experimental metadata (see details in Supplementary Table S1), 2) download files (MST, SMS, and binary raw data files) of each measurement, and 3) the predicted peak models represented in the MST format and PDF files that contain mass spectrum of each peak as a graphical image. The list of whole peak models appears by clicking “Show peak list”. Each peak has a unique identifier (e.g., MDGC1_00001_P000001) with the peak profile such as retention time and peak intensity. The identified compound name of the peak model is described at the right side of the peak intensity. Clicking “Show NIST” displays the mass spectrum of each peak in NIST-like text format; whereas, clicking “PDF” presents the PDF file of the mass spectrum in a graphical manner. This view helps in comparing the mass spectra of multiple peaks in manual annotation process.

The left side menu in each page of MassBase contains

Figure 2. Snapshot of the MST information page. All information of MST was summarized in the MST information page. The page was constructed from three sections (metadata section, download file section, and peak model section). Properties of each peak in the MST were presented as a list and text or graphical format of the distribution pattern of molecule ions. All datasets can be downloaded from the Download page. MST information page contains simple search function, and more detailed search function is provided in the advanced search page. All hit records of simple search in the top page and advanced search are listed in the search result page.

Table 1. Number of records in the MassBase.

Sample type	Number of records			
	LC-MS	GC-MS	CE-MS	Total
Biological samples	5,760	6,352	26,921	39,033
Plant and algae (116 species)	5,487	6,342	26,921	38,750
<i>Arabidopsis thaliana</i>	1,405	4,417	25,638	31,460
<i>Lotus japonicus</i>	361	851	370	1,582
<i>Solanum lycopersicum</i>	1,481	436	0	1,917
Other plant species (113 species)	2,240	638	913	3,791
Others (microorganisms, animals, foods)	273	10	0	283
Authentic chemicals and others	2,960	1,979	2,521	7,460

Details are listed on the website (http://webs2.kazusa.or.jp/massbase/index.php?action=Massbase_ShowSummary).

links to the advanced search page, download page and summary page. In the advanced search page, multiple search terms of distinct metadata categories are available for searching the datasets. For example, by selecting the type of mass spectrometry, comparable datasets with the same resolution and accuracy, which depend on the machine type, are retrieved from the database. In the summary page, links to the search result for selected terms such as species and instrument names are listed with the number of the records. All datasets in four formats (binary raw data format, SMS, and MST text format, and PDF format) can be downloaded in bulk from the download page. In total, 1.1 TB downloadable datasets are archived in gzip format. MassBase has a function of batch download for the datasets in the download page, while such batch download function is not available in MetaboLights and Metabolomics Workbench. Furthermore, the size of each download file is designed to be less than 1GB, considering the users who have only a weak internet connection.

The complete set of metadata, list of chemically identified peaks, and compound names used for annotation are provided as tab-delimited text format files. All sample names and experimental conditions used as search terms in MassBase are described in these files. Moreover, the “pre” version datasets that comprise the binary raw data files and metadata description files are provided for quick reference to users who can process the datasets using their own bioinformatics tools.

The metabolomics portal site KOMCS contains three distinct metabolome databases MassBase for raw datasets of mass spectrometers, KomicMarket2 for annotated datasets that were generated by annotating some of MassBase datasets, and Metabolonote for referring the metadata of KomicMarket2 (Figure 1). Metabolonote also refers to the MassBase datasets directly, and other databases such as the PRIME (<http://prime.psc.riken.jp/>).

Currently, part of the MassBase datasets was annotated for KomicMarket2 and referred in Metabolonote. As single or multiple chemical formulas can be calculated from the accurate mass value of ion peak, we processed 2,290 SMS files obtained from LC-FTICR or Orbitrap MS, which detect the m/z values within ~ 5 ppm. We used PowerGet software to extract ion peaks, and thereafter, annotated compounds using public metabolite databases, KEGG, and KNApSACk (Afendi et al. 2012; Kanehisa et al. 2002; Sakurai et al. 2014) (Figure 1). The datasets comprise metabolite annotations such as amino acids, sugars, flavonoids, and the results of database search and MS/MS spectra, if present. The information of the annotated peaks was deposited in the database KomicMarket2 (<http://webs2.kazusa.or.jp/km2/>) as a resource of differential metabolome analysis, from which all datasets can be freely downloaded. Details of the analytical methods of these datasets are described in

the corresponding pages in the metadata management system Metabolonote (<http://metabolonote.kazusa.or.jp/>); these complement the metadata in MassBase for better understanding of the datasets in the comparative metabolome analysis.

The datasets of *Jatropha curcas* fruit development were analyzed in detail and were then registered in an omics browser KaPPA-View4-Jatropha (<http://kpv.kazusa.or.jp/kpv4-jat/>), in which the gene expression and metabolite changes are represented on the metabolic maps to understand the regulatory mechanisms of metabolite changes during fruit ripening (Sakurai et al. 2012; Sano et al. 2012).

To identify the metabolites, we recognized 273 mass spectra of authentic chemicals measured via GC-TOF-MS and deposited them in the MassBase. Moreover, these datasets were also registered as reference mass spectra in the public mass spectral database MassBank (<http://www.massbank.jp/>) with accession numbers KZ000001–KZ000273 (Horai et al. 2010).

As part of the project “Platform of metabolome informatics considering the material cycle”, supported by the National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST), the datasets in the MassBase will be integrated into the MSI-compliant public data repository MetaboBank (<https://biosciencedbc.jp/en/funding/project/18062862.html>). Currently, through the database Metabolonote, part of MassBase datasets was aggregated in MetabolomeXchange, which is the site for data integration for MetaboLights, Metabolomics Workbench, and Metabolomic Repository Bordeaux. The datasets of MassBase will be incorporated into other databases such as MetaboLights and Metabolomics Workbench in the future.

Acknowledgements

We thank our colleagues, Takeshi Motegi, Mitsuo Enomoto, Yoko Iijima, Yoshihiko Morishita, Ryosuke Sasaki, Takashi Matsuura, Daisuke Nakajima, Kunihiro Suda, Atsushi Kurabayashi, Yoshiaki Nagashima, Makoto Hirose, Yoshie Tange, Yuuki Asami, and Tsurue Aoyama for their technical assistance. This work was supported by a grant from the New Energy and Industrial Technology Development Organization (NEDO) of Japan as part of the project “Development of Fundamental Technologies for Controlling the Material Production Process of Plants” and partly supported by National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST).

References

- Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, et al. (2012) KNApSACk family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* 53: e1
- Ara T, Enomoto M, Arita M, Ikeda C, Kera K, Yamada M, Nishioka

- T, Ikeda T, Nihei Y, Shibata D, et al. (2015) Metabolonote: A wiki-based database for managing hierarchical metadata of metabolome analyses. *Front Bioeng Biotechnol* 3: 38
- Bino RJ, Hall RD, Fiehn O, Kopka J, Saito K, Draper J, Nikolau BJ, Mendes P, Roessner-Tunali U, Beale MH, et al. (2004) Potential of metabolomics as a functional genomics tool. *Trends Plant Sci* 9: 418–425
- Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. *Nat Biotechnol* 18: 1157–1161
- Fiehn O, Sumner LW, Rhee SY, Ward J, Dickerson J, Lange BM, Lane G, Roessner U, Last R, Nikolau B (2007) Minimum reporting standards for plant biology context information in metabolomic studies. *Metabolomics* 3: 195–201
- Haug K, Salek RM, Steinbeck C (2017) Global open data management in metabolomics. *Curr Opin Chem Biol* 36: 58–63
- Horai H, Arita M, Kanaya S, Nihei Y, Ikeda T, Suwa K, Ojima Y, Tanaka K, Tanaka S, Aoshima K, et al. (2010) MassBank: A public repository for sharing mass spectral data for life sciences. *J Mass Spectrom* 45: 703–714
- Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res* 30: 42–46
- Quanbeck SM, Brachova L, Campbell AA, Guan X, Perera A, He K, Rhee SY, Bais P, Dickerson JA, Dixon P, et al. (2012) Metabolomics as a hypothesis-generating functional genomics tool for the annotation of *Arabidopsis thaliana* genes of “unknown function”. *Front Plant Sci* 3: 15
- Sakurai N, Ara T, Enomoto M, Motegi T, Morishita Y, Kurabayashi A, Iijima Y, Ogata Y, Nakajima D, Suzuki H, et al. (2014) Tools and databases of the KOMICS web portal for preprocessing, mining, and dissemination of metabolomics data. *BioMed Res Int* 2014: 1–11
- Sakurai N, Ogata Y, Ara T, Sano R, Akimoto N, Suzuki H, Kajikawa M, Widyastuti U, Suharsono S, Yokota A, et al. (2012) Development of KaPPA-View4 for omics studies on *Jatropha* and a database system KaPPA-Loader for construction of local omics databases. *Plant Biotechnol* 29: 131–135
- Sano R, Ara T, Akimoto N, Sakurai N, Suzuki H, Fukuzawa Y, Kawamitsu Y, Ueno M, Shibata D (2012) Dynamic metabolic changes during fruit maturation in *Jatropha curcas* L. *Plant Biotechnol* 29: 175–178
- Spicer RA, Salek R, Steinbeck C (2017) Compliance with minimum information guidelines in public metabolomics repositories. *Sci Data* 4: 170137
- Steinbeck C, Conesa P, Haug K, Mahendraker T, Williams M, Maguire E, Rocca-Serra P, Sansone SA, Salek RM, Griffin JL (2012) MetaboLights: Towards a new COSMOS of metabolomics data management. *Metabolomics* 8: 757–760
- Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, et al. (2016) Metabolomics Workbench: An international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res* 44(D1): D463–D470