Hindawi Publishing Corporation
BioMed Research International
Volume 2014, Article ID 194812, 11 pages
http://dx.doi.org/10.1155/2014/194812

Research Article

Tools and Databases of the KOMICS Web Portal for Preprocessing, Mining, and Dissemination of Metabolomics Data

Nozomu Sakurai,[1,2] Takeshi Ara,[1,2] Mitsuo Enomoto,[1,2] Takeshi Motegi,[1] Yoshihiko Morishita,[1] Atsushi Kurabayashi,[1] Yoko Iijima,[1,3] Yoshiyuki Ogata,[1,4] Daisuke Nakajima,[1] Hideyuki Suzuki,[1] and Daisuke Shibata[1]

[1] *Kazusa DNA Research Institute, 2-6-7 Kazusa-kamatari, Kisarazu, Chiba 292-0818, Japan*
[2] *JST, National Bioscience Database Center (NBDC), 5-3 Yonbancho, Chiyoda-ku, Tokyo 102-0081, Japan*
[3] *Department of Nutrition and Life Science, Kanagawa Institute of Technology, 1030 Shimo-ogino, Atsugi, Kanagawa 243-0292, Japan*
[4] *Graduate School of Life and Environmental Sciences, Osaka Prefecture University, Sakai, Osaka 599-8531, Japan*

Correspondence should be addressed to Nozomu Sakurai; sakurai@kazusa.or.jp

Received 5 December 2013; Revised 7 February 2014; Accepted 24 February 2014; Published 9 April 2014

Academic Editor: Shigehiko Kanaya

A metabolome—the collection of comprehensive quantitative data on metabolites in an organism—has been increasingly utilized for applications such as data-intensive systems biology, disease diagnostics, biomarker discovery, and assessment of food quality. A considerable number of tools and databases have been developed to date for the analysis of data generated by various combinations of chromatography and mass spectrometry. We report here a web portal named KOMICS (The Kazusa Metabolomics Portal), where the tools and databases that we developed are available for free to academic users. KOMICS includes the tools and databases for preprocessing, mining, visualization, and publication of metabolomics data. Improvements in the annotation of unknown metabolites and dissemination of comprehensive metabolomic data are the primary aims behind the development of this portal. For this purpose, PowerGet and FragmentAlign include a manual curation function for the results of metabolite feature alignments. A metadata-specific wiki-based database, Metabolonote, functions as a hub of web resources related to the submitters' work. This feature is expected to increase citation of the submitters' work, thereby promoting data publication. As an example of the practical use of KOMICS, a workflow for a study on *Jatropha curcas* is presented. The tools and databases available at KOMICS should contribute to enhanced production, interpretation, and utilization of metabolomic Big Data.

1. Introduction

A metabolome, which comprises comprehensive data on quantification of metabolites in an organism calculated using metabolomic technologies [9, 10], has been increasingly used for the analysis and practical applications of biological and environmental systems. Within the data-intensive systems biology discipline, metabolomics is particularly important compared to other "omics" (genome, transcriptome, and proteome) disciplines since metabolomes are more closely related to phenotype and regulate gene and protein expression networks [11–13]. Mass spectrometry (MS) and nuclear magnetic resonance spectroscopy (NMR) are complementary techniques often used for the detection and identification of metabolites. MS technology has integrated separation techniques and is used in most cases because of its sensitivity, selectivity, speed, and broad applicability [14–16]. Owing to the wide range of chemical diversity, there is no ideal apparatus that is capable of analyzing all possible metabolites. Combinations of separation techniques with MS, such as liquid chromatography- (LC-) MS, gas chromatography- (GC-) MS, and capillary electrophoresis-

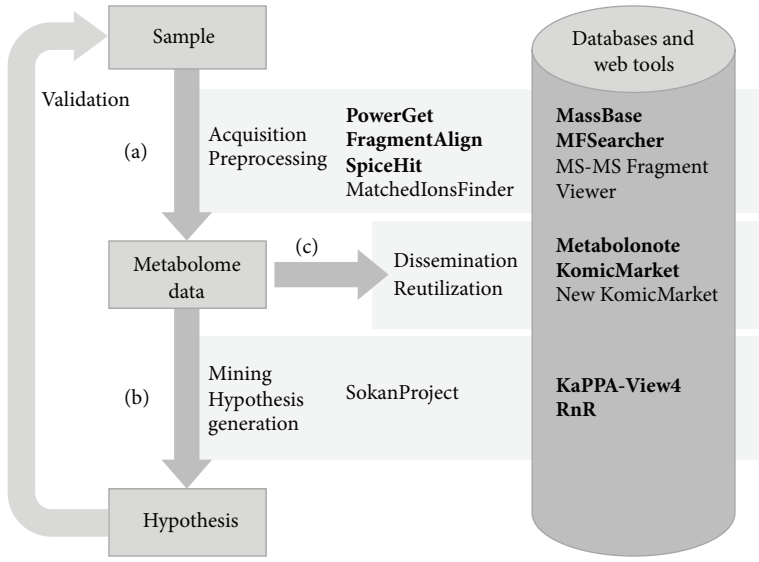


Figure 1: A typical workflow of a metabolomics study and KOMICS-relevant tools and databases. The process of data acquisition and preprocessing (a) is required for generating the metabolome data. A working hypothesis is generated by interpreting the metabolome data (b), and the cycle is completed after validating the hypotheses by further analyses (the arrow on the left side). The metabolome data are published in the databases (c) and utilized for preprocessing and data mining. The tools and databases introduced in the main text of this paper are shown in bold face.

(CE-) MS, are chosen according to a study's purpose [17–19]. Metabolomics technology, including instrumental analysis, detection and identification of metabolites, statistical interpretation, and generation of hypotheses with computational support, is used for a variety of studies, such as functional analysis of biological systems [20–22], biomarker discovery [23, 24], medical diagnostics [14, 25], quality assessment of foods [26, 27], evaluation of genetically modified crops [28, 29], and assessment of environmental pollution [30, 31].

A considerable number of software tools and databases have been developed for processing the complicated and multidimensional metabolome datasets generated by various types of MS-based instruments [32–35]. A typical workflow of metabolomic data analysis includes the following processes: (a) preprocessing of raw data for extraction of metabolite features, annotation of the metabolites, and finally generation of metabolome data; (b) mining and visualization of metabolome data for statistical interpretation of its nature and hypothesis generation; (c) storing and disseminating the data for further utilization and comparison (Figure 1). XCMS2 [36], MzMine2 [37], MathDAMP [38], MetAlign [39], and MET-IDEA [40] are typical tools for preprocessing including detection, alignment, and annotation of metabolite features. Some of these tools also provide statistical analysis functions for data interpretation. MassBank [41], METLIN [42], PRIMe [43], and HMDB [44] are available as references of mass spectra for metabolite annotation. The metabolite data are interpreted by means of the genome information from compound databases such as KNApSAcK [45], PubChem [46], and Chemspider (http://www.chemspider.com/) and by means of metabolic pathway databases including KEGG [47], BioCyc [48], and Reactome [49], which enable data visualization on pathway maps. The raw and processed data are stored publicly in databases such as PlantMetabolomics.org [50], GMD@CSB.DB [51], SetupX (currently not available), MetabolomeExpress [52], MetaboLights [53], and Metabolomics Workbench (http://www.metabolomicsworkbench.org/).

We report here a portal website named *KOMICS* (The Kazusa Metabolomics Portal, http://www.kazusa.or.jp/komics/), which hosts tools and databases that we developed for metabolomics. Although an increasing number of tools and databases have become available, two major issues remain to be resolved, that is, comprehensiveness of metabolites [54, 55] and data dissemination [53, 56, 57]. Our primary aim in developing data preprocessing tools is to help researchers with the manual annotation process that remains essential for nontarget metabolomics [54]. PowerGet for LC-high-resolution-MS and FragmentAlign for GC-MS are tools that enable curation of peak alignment results. SpiceHit is a high-throughput metabolite identification tool for CE-MS analysis using the selected ion monitoring (SIM) method. We have also developed data mining and visualization tools for the generation of working hypotheses (KaPPA-View and RnR). Real data is indispensable for comparative analysis and for the development and improvement of preprocessing tools [53, 58]. MassBase is one of the largest raw data repositories, and KomicMarket is a database of metabolic profiling data. We developed a metadata-specific database, Metabolonote, to promote data publication by researchers. These resources for a wide range of metabolome data processing are expected to contribute to improved production and utilization of metabolomic data.

2. Materials and Methods

The standalone tools for metabolome data production, PowerGet, FragmentAlign, and SpiceHit, were developed in Java (Oracle Corporation). The web-based tools and databases were developed and are run in Apache, PHP, Perl, MySQL, Java, and Tomcat on Linux servers. The KOMICS website was constructed using the content management system "Joomla!" running on a Linux server with Apache, PHP, and MySQL. The details of the development and license information are described in the individual introduction pages of KOMICS, in manuals, or in other relevant help resources. The tools and databases are freely available to academic users.

Details of the analytical methods for the evaluation of preprocessing tools are described in the Supplementary Material (see Supplementary Material available online at http://dx.doi.org/10.1155/2014/194812).

3. Results and Discussion

The tools and databases we have developed and provided at the KOMICS web portal are classified into three categories according to the typical workflow of metabolomic data analysis, namely, (a) preprocessing tools, (b) data mining tools, and (c) databases for data dissemination (Figure 1). Here we describe several representative examples. All the currently available tools and databases are listed in Table 1. The number of records in each metabolomics-related database is shown in Table 2. The formats of input and output files and the availability of sample datasets are summarized in Table 3.

3.1. Data Preprocessing Tools

3.1.1. PowerGet. PowerGet is a standalone Java software package for detection, alignment, and annotation of metabolite features from data obtained using LC-high-resolution-MS (HRMS). Accurate mass values measured by HRMS, such as Fourier transform ion cyclotron resonance MS and Orbitrap MS (Thermo Fisher), allow users to predict the elemental composition of a metabolite. The intensity ratio of ^{13}C to ^{12}C isotopic ion peaks is useful for estimating the number of carbon atoms in a molecule. Estimation of ion adducts attached to the metabolites by coeluted ions is helpful for calculating elemental composition and for search of compound databases by mass values of nonionized molecules. The PowerFT module in PowerGet attaches these data automatically to all metabolite features in the LC-HRMS data. In the PowerMatch module, the metabolite features are aligned among the samples taking into account the similarity of MS/MS fragmentation patterns. A tool for refining the alignment results, MatchedIonsFinder [1], is also available via KOMICS.

To evaluate the accuracy of mass values of the peaks detected using PowerGet, the mass differences between a theoretical mass and a detected mass were compared to those of the peaks detected using the commercial software, Xcalibur (see Supplementary Method S1). PowerGet exhibited greater accuracy (0.579 ± 0.481 ppm (mean ± SD)) than Xcalibur (0.783 ± 0.563 ppm) in the evaluation of 143 standard compounds (Supplementary Table S1).

One of the unique functions of PowerGet is that the alignment results are manually editable: a user can promptly check metabolite's characteristics, such as mass chromatogram shape, existence of adjacent features, and MS/MS fragmentation patterns, by means of a graphical user interface (GUI), as shown in Figure 2. Alignment is essential for preparing matrices of samples to metabolite intensity data for further comparison and statistical analysis. Alignment is highly valuable when users need to annotate the metabolites, especially for unknown features. By comparing the features from several replicate samples, (1) the estimation error of the ion adducts is verified, (2) accuracy of mass measurement can be improved, and (3) reproducibly detected features are prioritized for further annotation. Therefore, alignment errors should be assessed and corrected during detailed annotation of unknown metabolites. PowerGet is utilized in preparing data for KomicMarket and Bio-MassBank (http://bio.massbank.jp/).

3.1.2. FragmentAlign. This is a standalone Java tool designed for GC-MS data analysis with functions for alignment and annotation of metabolite features. A GUI for editing the alignment results is also implemented in this software (Figure 3). The similarity of fragment ion patterns generated by electron ionization (EI) is taken into account in the alignment of metabolite features. The metabolite features can also be annotated based on EI fragment patterns, by comparing to patterns from standard compounds. The fragment pattern data of standard compounds can be imported and utilized when the data is written in the format defined by the National Institute of Standards and Technology (NIST), USA.

To evaluate the applicability of data matrices generated by FragmentAlign for further statistical analyses, a principal component analysis (PCA) was conducted using the GC-MS data obtained from 3 biological sources: *Arabidopsis* leaves, *Lotus japonicus* leaves, and *Arabidopsis* cultured cells. Five replicates of each source were mapped to similar positions, whereas the 3 sources were mapped separately from one another on the score plot (Supplementary Figure S1). High-correlation coefficients for peak intensity within the replicates were observed (Supplementary Figure S2). These results suggest that appropriate feature extraction and generation of data matrices can be performed successfully using FragmentAlign.

3.1.3. SpiceHit. The standalone Java tool SpiceHit is intended for high-throughput identification of metabolite features detected using the selected ion monitoring (SIM) method in CE-MS. The metabolite features are identified based on migration times relative to internal standard compounds and are compared to those of the standard compound library prepared in-house. The tool is designed for processing a large number of data files in a high-throughput manner; it requires checking and correcting the assignment errors manually.

To ascertain whether SpiceHit is applicable to practical data analysis, the accuracy of peak quantification was

Table 1: The tools and databases available at KOMICS (as of November 2013).

Name	Description	URL	Reference
	Standalone tools		
PowerGet	Metabolite detection, alignment, and annotation tool for LC-high-resolution-MS.	http://www.kazusa.or.jp/komics/software/PowerGet	
MatchedIonsFinder	Revising tool for metabolite alignment results from LC-MS analyses.	http://www.kazusa.or.jp/komics/software/MatchedIonsFinder	[1]
FragmentAlign	Metabolite alignment and annotation tool for GC-MS.	http://www.kazusa.or.jp/komics/software/FragmentAlign	
SpiceHit	High-throughput metabolite detection and annotation tool for SIM analysis in CE-MS.	http://www.kazusa.or.jp/komics/software/SpiceHit	
KAGIANA	Microsoft Excel-based tool for exploring the function of *Arabidopsis* genes.	http://webs2.kazusa.or.jp/kagiana/	[2]
SokanProject	A tool for calculating Pearson's correlation coefficients.	http://www.kazusa.or.jp/komics/software/SokanProject	
	Web tools		
MFSearcher	Web service for rapid prediction of elemental composition and database searching by accurate mass values.	http://webs2.kazusa.or.jp/mfsearcher/	[3]
DAGViz	Visualization tool for similarities of gene ontology annotations.	http://www.pgb.kazusa.or.jp/dagviz/	[4]
	Databases		
MassBase	Largest repository of metabolomics raw data.	http://webs2.kazusa.or.jp/massbase/	
KomicMarket	Sample-centric database for metabolomic profile data.	http://webs2.kazusa.or.jp/komicmarket/	
New KomicMarket temporary website	Developmental version of KomicMarket.	http://webs2.kazusa.or.jp/new_km_tmp/	
KaPPA-View4	Pathway database for visualizing metabolome and transcriptome data.	http://kpv.kazusa.or.jp/	[5]
Metabolonote	Metadata-specific Semantic MediaWiki-based database.	http://metabolonote.kazusa.or.jp/	
MS-MS Fragment Viewer	Database for MS/MS fragmentation data of 115 flavonoids.	http://webs2.kazusa.or.jp/msmsfragmentviewer/	
RnR	Database providing metabolite-to-gene relationships calculated from ~200 transgenic *Arabidopsis* cells.	http://webs2.kazusa.or.jp/kagiana/rnr	
CoP	Gene-to-gene coexpression database for 8 plant species calculated using the Confeito algorithm.	http://webs2.kazusa.or.jp/kagiana/cop0911	[6]
KATANA	Cross-search system for *Arabidopsis* genes.	http://www.kazusa.or.jp/katana/	[7]
ARTRA	Database of probe information of *Arabidopsis* DNA microarray data (developed by Takara Bio Inc.).	http://artra.kazusa.or.jp/artra/ARI3_101/	[8]
FuLoja	Database of *Lotus japonicus* full-length cDNA obtained in the NEDO project.	http://webs2.kazusa.or.jp/IntegrationDBRS/FuLoja/	
PMPj-Blast	Database of ESTs, cDNAs, and oligo DNA microarray probes for *Lotus japonicus* and some other plants.	http://webs2.kazusa.or.jp/IntegrationDBRS/pmpj-blast/	

Table 2: The number of records in metabolomics-related databases at KOMICS (as of November 2013).

Database name	Number	Description
MassBase	43959	Binary raw datasets
KomicMarket	85	Biological samples, including 251 instrumental analysis datasets
	215	Chemical samples, including 488 instrumental analysis datasets
New KomicMarket temporary website	16	Number of studies, including 166 analyzed datasets
Metabolonote	34	Number of studies, including metadata for 375 instrumental analysis datasets and 765 computational analysis datasets
MS-MS Fragment Viewer	115	Analyzed flavonoids
RnR	194	Metabolite features

Figure 2: The alignment-editing function of the PowerMatch module of PowerGet. (a) The Alignment Table shows the alignment results of the peaks detected in each sample. The intensity of peaks is summarized in another window (f). The details of the peak information (b), MS/MS fragments (c), appearance of peripheral peaks (d), and peak shape (e) are shown for the user-selected peaks. A misaligned peak (the blue colored cell in panel a) can be merged to an appropriate row using the Edit Alignment function (g), by immediately checking the detailed information for the peak (b–e).

compared to that acquired using the commercial software ChemStation (Agilent Technologies, Palo Alto, CA). In the detection of amino acids, the results from SpiceHit were strongly correlated with those from ChemStation, as well as with theoretical concentrations (Supplementary Table S4). Similar relative standard deviation (RSD) values for each amino acid in triplicate analyses were observed for SpiceHit and ChemStation (Supplementary Figure S3). Good linearity of peak areas common to SpiceHit and ChemStation was observed in the amino acid peaks detected in the biological samples (Supplementary Figure S4). These results suggest that the accuracy and the sensitivity of peak quantification by SpiceHit are similar to those of ChemStation and that SpiceHit is suitable for practical use.

3.1.4. MFSearcher. This is a web service that allows for rapid prediction of elemental composition from accurate mass values and for rapid searching of compound databases [3]. A GUI tool for MFSearcher queries is also provided as a module in PowerGet. PowerGet has a batch search function

Table 3: A summary of input and output file formats and availability of sample data for preprocessing tools. The precise formats are described in the instruction manuals for each tool.

Tool name	Input	Output	Availability of sample data
PowerGet	(1) PowerGet format (text file): MSGet tool is available for generating the text files from Xcalibur raw files (2) MassBase SMS format (text file) (3) mzXML file generated from the Xcalibur raw files using the ReAdW tool[*a]	(1) PowerGet format (text file): users can select the items and formats of the output file (2) TogoMD format (text file)	KOMICS website
FragmentAlign	*Deconvoluted peak data* (1) FragmentAlign format (text file, one of the NIST formats) (2) MassBase SMT format (text file) (3) GMD[*b] format (text file in NIST[*c] MSP format) (4).ELU file of AMDIS[*d] software *Chromatogram data for deconvolution* CSV file exported by Pegasus III (text file)	FragmentAlign format (text file)	KOMICS website A sample file of a standard compound library is included in the tutorial data
SpiceHit	*Electropherogram data* (1) CSV file (text file) (2) ChemStation .MS file (binary) *Standard compound library* Excel file (binary)	(1) Tab-delimited text file (2) Excel file	(1) KOMICS website (2) Included in the tool A sample of a standard compound library is included

[*a]ReAdW tool: available at http://tools.proteomecenter.org/wiki/index.php?title=Software:ReAdW.
[*b]GMD: Golm Metabolome Database, http://gmd.mpimp-golm.mpg.de.
[*c]NIST: National Institute of Standards and Technology, http://www.nist.gov/.
[*d]AMDIS: available at http://chemdata.nist.gov/mass-spc/amdis/.

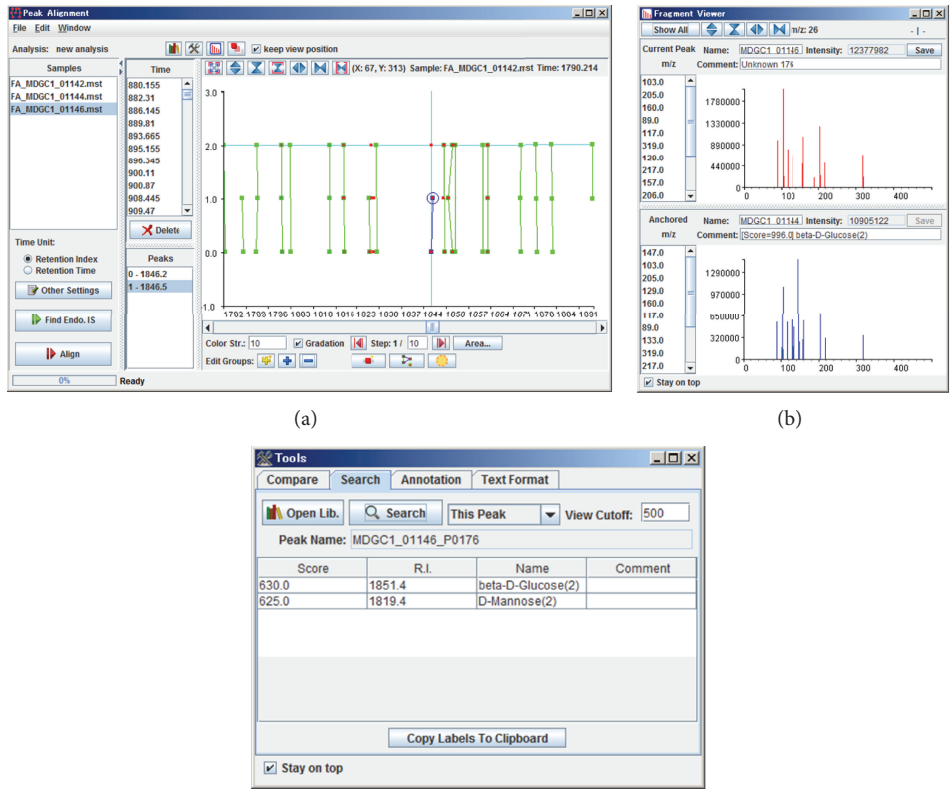


FIGURE 3: Screenshots of FragmentAlign. An alignment result from 3 samples is depicted (a). The electron ionization (EI) mode of a fragmentation pattern of the metabolite peak is presented in the Fragment Viewer panel (b). The metabolites are annotated by comparing the similarity of the fragment patterns to those obtained from standard compounds (c).

for querying thousands of detected metabolite features via MFSearcher.

Because MFSearcher is a RESTful web service, the query parameters for MFSearcher should be included in the description of a URL. Numerous sample queries are available as URL links on the MFSearcher website.

3.2. Data Mining Tools

3.2.1. KaPPA-View. This is a web-based tool for the visualization of metabolomic data on metabolic pathway maps from the Kyoto Encyclopedia of Genes and Genomes (KEGG) [5]. The degree of change in metabolite abundance between several samples is expressed as hue of the compound symbols drawn on the KEGG pathway maps, based on the compound IDs assigned. Alterations in transcriptome data can be simultaneously depicted on the maps. This tool can be used for the integrated analysis of metabolomic, transcriptomic, and possibly proteomic data.

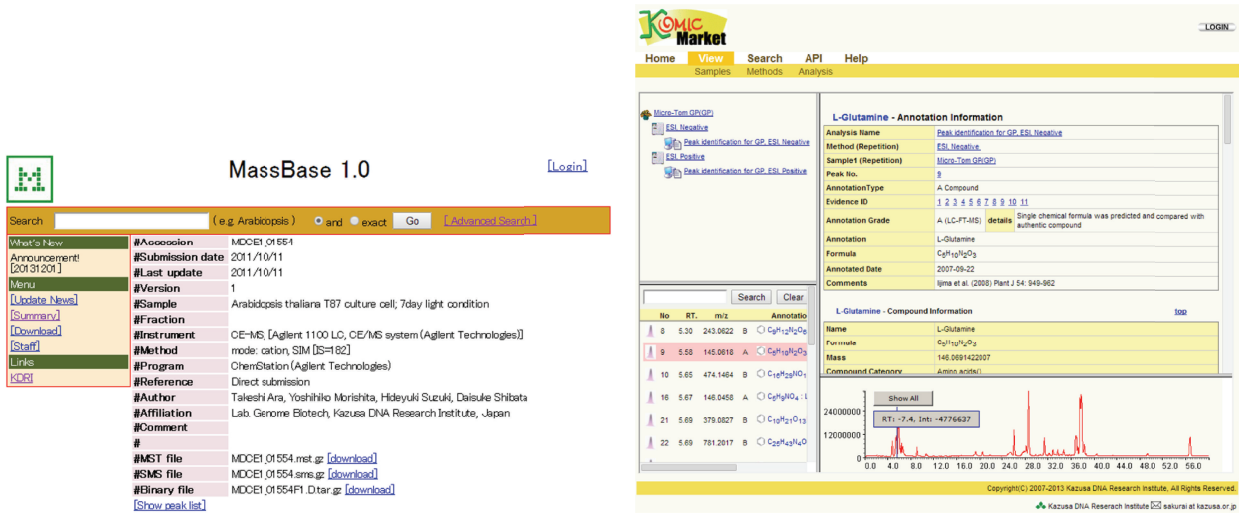
Sample data for testing the color representations on the pathway maps are available on the "Analysis" page of KaPPA-View. Users can select the items according to the directions presented on the page. Sample files for input data are available on the "Download" page.

3.2.2. RnR. This database contains data on the relationship between metabolites and genes; these data were generated via simultaneous measurement of the metabolome and transcriptome of approximately 200 transgenic cultured cell lines of the model plant *Arabidopsis*. The gene expression patterns and metabolite changes resulting from specific transgenes are compiled in the database. Users can search, for example, genes that can affect the abundance of the queried metabolites and vice versa. The database should contribute to knowledge discovery related to gene-to-metabolite regulatory networks in *Arabidopsis* cells.

To view an example dataset, a clickable pie chart of metabolites is presented on the main page of the RnR website. Clicking on a section of the pie chart will show a list of metabolites. After choosing a metabolite name, users will be able to view candidate genes that are related to the metabolite.

3.3. Databases for Data Dissemination

3.3.1. MassBase. The primary purpose of MassBase is the distribution of raw data generated by analytical instruments. Dissemination of raw data would enable the development and improvement of data analysis tools by bioinformatics researchers [53]. Binary raw data and near-raw text data exported from raw binary results are provided.



(a) (b)

FIGURE 4: Screenshots of the web interface of MassBase (a) and KomicMarket (b).

Users can search records by sample name, sample description, instrument type, and ionization mode on the "Advanced Search" page (Figure 4(a)). A summary of the records classified by species and instrument type is available on the "Summary" page.

3.3.2. KomicMarket. KomicMarket is a sample-centric database aimed at the distribution of metabolic profiles with and without metabolite annotations (Figure 4(b)). Previous results on the detection and annotation of metabolite features in certain samples can serve as good references for future metabolite annotations [56].

The records can be queried by keywords in the sample descriptions, including peak characteristics such as mass values, and in annotations of the peaks via the GUI on the KomicMarket website. The system provides application programming interfaces (APIs) for performing software-based querying of the data. Using the APIs, we employed the MFSearcher module of PowerGet to search metabolites in KomicMarket.

3.3.3. Metabolonote. This Semantic MediaWiki-based database is intended for managing metadata, which is the detailed information on experimental procedures accompanying the generation of data. Metabolonote is expected to accelerate publication of metabolomics data. The raw data obtained from the experiments or the processed data derived from them are not the target of Metabolonote, and the "actual data" are normally stored in other databases specifically built for a given purpose. Separation of the management of complicated metadata of metabolomics from the management of actual data makes it possible to share the same metadata among multiple actual databases such as raw data repositories, metabolic profile databases, reference libraries of MS/MS, and research papers. One-stop-shop management of complicated metadata of metabolomics eliminates the redundant management of metadata in several databases and reduces labor for data submitters. We defined a simple data format named *Togo metabolomics data format* (TogoMD) for easier description of metadata. Specifications of the TogoMD format are documented on the Metabolonote website (http://metabolonote.kazusa.or.jp/TogoMetabolomeDataFormat). Metabolonote provides application programming interfaces (APIs) for semantic searching of the records and retrieval of metadata. Because Metabolonote is a wiki system, it allows the submitters to attach free additional information about the metadata, such as images of the samples, video recordings of tricky analytical procedures, and links to a journal's website where the results are published. Therefore, metadata written by the submitter function as a hub of the web data resources related to the submitter's work. The increased presence of the submitters' published work on the web should increase the citations by others [58]. Therefore, the wiki system is expected to facilitate the dissemination of data to the public. Metabolonote is already linked to seven actual databases, including MassBase, KomicMarket, Bio-MassBank, and Riken PRIMe.

The metadata deposited in Metabolonote are listed on the "Public Pages." The registered metadata are semantically searchable by various items (and combinations thereof) on the "Metadata Search" page.

3.4. Practical Use. Here we present a workflow for a metabolomics study of *Jatropha curcas* L. [59, 60], a biofuel plant, to illustrate an example of the practical use of the KOMICS resources (Figure 5). LC-Orbitrap-MS analyses were conducted using 4 developmental stages of *J. curcas* fruit samples. The acquired data were primarily recorded as a binary raw file with commercial software (Xcalibur, Thermo Fisher). To analyze the data with PowerGet, the chromatogram data were exported to text files using the MSGet tool, which is also available on the KOMICS website. The raw files and extracted text files were published on MassBase. The text data were then processed using the PowerGet tool

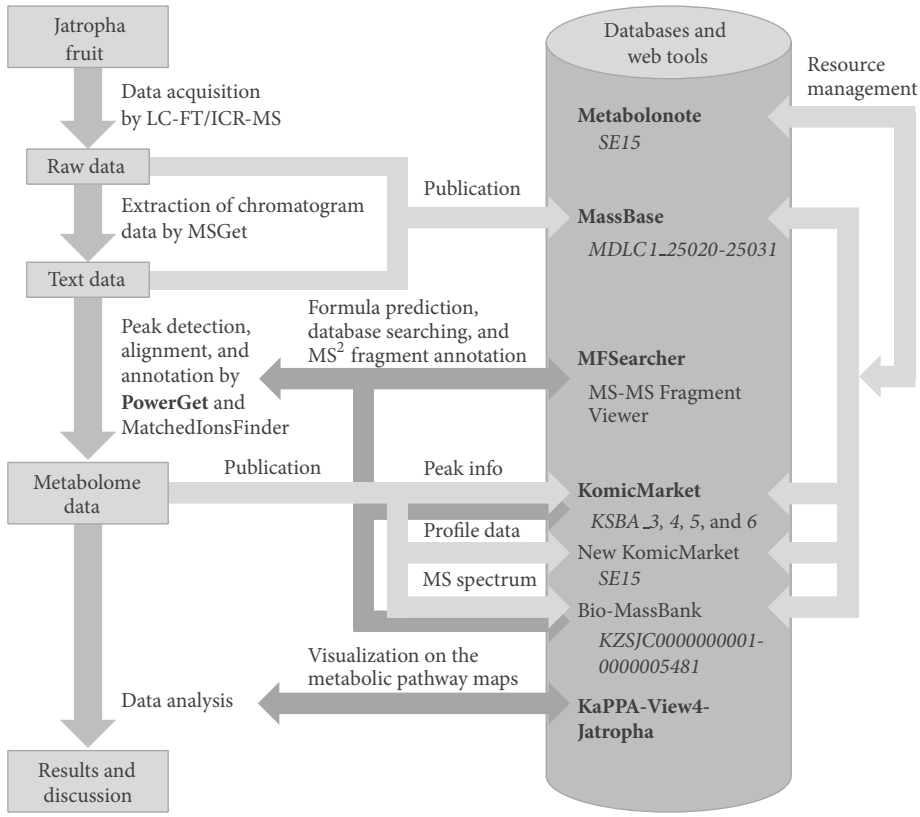


Figure 5: A schematic representation of the workflow for analysis of metabolomic changes in the developing fruit of *Jatropha curcas* L. The tools and databases introduced in the main text are shown in bold face. The accession IDs of the data in each database are shown in italics.

to generate the metabolomic data. MatchedIonsFinder was used to refine the alignment results. MFSearcher was used for high-throughput search of elemental composition and compound databases. MS-MS Fragment Viewer was used for interpreting MS/MS fragments in the metabolite annotations. The peak information, profile data (in the TogoMD format), and MS spectrum data were stored on KomicMarket (on the New KomicMarket temporary website) and on Bio-MassBank, respectively. These data were recursively used for metabolite annotations during the preprocessing step. Subsequently, the nature of the metabolomic data was interpreted by visualization on pathway maps using KaPPA-View4 and other statistical analyses. Consequently, a drastic change in metabolites during the maturation of *J. curcas* fruit was detected, and these data should contribute to further analysis of oil production by *J. curcas*. The record in Metabolonote (metadata ID: SE5) is a good guidepost for finding data resources deposited in various databases on the web.

4. Conclusions

We have developed various tools and databases for a wide range of processes related to metabolomic studies: preprocessing, data mining, and publication. To our knowledge, PowerGet and FragmentAlign are the first tools to allow users to curate alignment results via GUI. The unique concept of a metadata-specific database should accelerate data publication and dissemination. This infrastructure is expected to assist researchers to attain superior utilization of metabolomics' Big Data. Nonetheless, annotation of novel metabolites (the so-called unknown unknowns) remains a serious bottleneck in building comprehensive metabolomic datasets [16, 54]. Continuous efforts are needed to improve and automate annotation tasks. In addition, a systematic collection of annotation skills from experts will be necessary in the near future, as will the analysis and transfer of these skills to the public domain for education of fledgling annotators.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was partly supported by a grant from the New Energy and Industrial Technology Development Organization (NEDO, Japan) as part of the project "Development of Fundamental Technologies for Controlling the Material Production Process of Plants," a grant from the National Bioscience Database Center (NBDC) of Japan Science and Technology Agency (JST) as part of the project titled "The Life-Science Database Integration Project," and a grant from the Kazusa DNA Research Institute.

References

[1] N. Yamamoto, T. Suzuki, T. Ara et al., "MatchedIonsFinder: a software tool for revising alignment matrices of spectrograms from liquid chromatography-mass spectrometry," *Plant Biotechnology*, vol. 29, no. 1, pp. 109–113, 2012.

[2] Y. Ogata, N. Sakurai, K. Aoki et al., "KAGIANA: an excel-based tool for retrieving summary information on *Arabidopsis* genes," *Plant and Cell Physiology*, vol. 50, no. 1, pp. 173–177, 2009.

[3] N. Sakurai, T. Ara, S. Kanaya et al., "An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values," *Bioinformatics*, vol. 29, no. 2, pp. 290–291, 2013.

[4] K. Yano, K. Aoki, H. Suzuki, and D. Shibata, "DAGViz: a directed acyclic graph browser that supports analysis of gene ontology annotation," *Plant Biotechnology*, vol. 26, no. 1, pp. 9–13, 2009.

[5] N. Sakurai, T. Ara, Y. Ogata et al., "KaPPA-View4: a metabolic pathway database for representation and analysis of correlation networks of gene co-expression and metabolite co-accumulation and omics data," *Nucleic Acids Research*, vol. 39, no. 1, pp. D677–D684, 2011.

[6] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata, "CoP: a database for characterizing co-expressed gene modules with biological information in plants," *Bioinformatics*, vol. 26, no. 9, pp. 1267–1268, 2010.

[7] K. Yano, T. Dansako, N. Sakurai, H. Suzuki, and D. Shibata, "KATANA: a web-based guide to public databases for *Arabidopsis* genomic information," *Plant Biotechnology*, vol. 22, no. 3, pp. 225–229, 2005.

[8] T. Ohba, K. Suzuki, T. Oura et al., "ARTRA: a new database of the *Arabidopsis* transcriptome and gene-specific sequences for microarray probes and RNAi triggers," *Plant Biotechnology*, vol. 26, no. 1, pp. 161–165, 2009.

[9] S. G. Oliver, M. K. Winson, D. B. Kell, and F. Baganz, "Systematic functional analysis of the yeast genome," *Trends in Biotechnology*, vol. 16, no. 9, pp. 373–378, 1998.

[10] W. B. Dunn, "Current trends and future requirements for the mass spectrometric investigation of microbial, mammalian and plant metabolomes," *Physical Biology*, vol. 5, no. 1, Article ID 011001, 2008.

[11] K. Dettmer, P. A. Aronov, and B. D. Hammock, "Mass spectrometry-based metabolomics," *Mass Spectrometry Reviews*, vol. 26, no. 1, pp. 51–78, 2007.

[12] K. Hollywood, D. R. Brison, and R. Goodacre, "Metabolomics: current technologies and future trends," *Proteomics*, vol. 6, no. 17, pp. 4716–4723, 2006.

[13] J. Nielsen and S. Oliver, "The next wave in metabolome analysis," *Trends in Biotechnology*, vol. 23, no. 11, pp. 544–546, 2005.

[14] W. B. Dunn and T. Hankemeier, "Mass spectrometry and metabolomics: past, present and future," *Metabolomics*, vol. 9, no. 1, supplement, pp. 1–3, 2013.

[15] A. Zhang, H. Sun, P. Wang, Y. Han, and X. Wang, "Modern analytical techniques in metabolomics analysis," *Analyst*, vol. 137, no. 2, pp. 293–300, 2012.

[16] D. S. Wishart, "Computational strategies for metabolite identification in metabolomics," *Bioanalysis*, vol. 1, no. 9, pp. 1579–1596, 2009.

[17] Z. Lei, D. V. Huhman, and L. W. Sumner, "Mass spectrometry strategies in metabolomics," *The Journal of Biological Chemistry*, vol. 286, no. 29, pp. 25435–25442, 2011.

[18] R. D. Hall, "Plant metabolomics: from holistic hope, to hype, to hot topic," *New Phytologist*, vol. 169, no. 3, pp. 453–468, 2006.

[19] K. Saito and F. Matsuda, "Metabolomics for functional genomics, systems biology, and biotechnology," *Annual Review of Plant Biology*, vol. 61, pp. 463–489, 2010.

[20] S. H. Khoo and M. Al-Rubeai, "Metabolomics as a complementary tool in cell culture," *Biotechnology and Applied Biochemistry*, vol. 47, part 2, pp. 71–84, 2007.

[21] M. R. Mashego, K. Rumbold, M. de Mey, E. Vandamme, W. Soetaert, and J. J. Heijnen, "Microbial metabolomics: past, present and future methodologies," *Biotechnology Letters*, vol. 29, no. 1, pp. 1–16, 2007.

[22] J. Kopka, A. Fernie, W. Weckwerth, Y. Gibon, and M. Stitt, "Metabolite profiling in plant biology: platforms and destinations," *Genome Biology*, vol. 5, no. 6, article 109, 2004.

[23] A. Koulman, G. A. Lane, S. J. Harrison, and D. A. Volmer, "From differentiating metabolites to biomarkers," *Analytical and Bioanalytical Chemistry*, vol. 394, no. 3, pp. 663–670, 2009.

[24] J. Jansson, B. Willing, M. Lucio et al., "Metabolomics reveals metabolic biomarkers of Crohn's disease," *PLoS ONE*, vol. 4, no. 7, Article ID e6386, 2009.

[25] R. Madsen, T. Lundstedt, and J. Trygg, "Chemometrics in metabolomics—a review in human disease diagnosis," *Analytica Chimica Acta*, vol. 659, no. 1-2, pp. 23–33, 2010.

[26] M. A. Fitzgerald, S. R. McCouch, and R. D. Hall, "Not just a grain of rice: the quest for quality," *Trends in Plant Science*, vol. 14, no. 3, pp. 133–139, 2009.

[27] W. Pongsuwan, T. Bamba, K. Harada, T. Yonetani, A. Kobayashi, and E. Fukusaki, "High-throughput technique for comprehensive analysis of Japanese green tea quality assessment using ultra-performance liquid chromatography with time-of-flight mass spectrometry (UPLC/TOF MS)," *Journal of Agricultural and Food Chemistry*, vol. 56, no. 22, pp. 10705–10708, 2008.

[28] A. E. Ricroch, J. B. Bergé, and M. Kuntz, "Evaluation of genetically engineered crops using transcriptomic, proteomic, and metabolomic profiling techniques," *Plant Physiology*, vol. 155, no. 4, pp. 1752–1761, 2011.

[29] M. Kusano, H. Redestig, T. Hirai et al., "Covering chemical diversity of genetically-modified tomatoes using metabolomics for objective substantial equivalence assessment," *PLoS ONE*, vol. 6, no. 2, Article ID e16989, 2011.

[30] M. Krauss, H. Singer, and J. Hollender, "LC-high resolution MS in environmental analysis: from target screening to the identification of unknowns," *Analytical and Bioanalytical Chemistry*, vol. 397, no. 3, pp. 943–951, 2010.

[31] C. Y. Lin, M. R. Viant, and R. S. Tjeerdema, "Metabolomics: methodologies and applications in the environmental sciences," *Journal of Pesticide Science*, vol. 31, no. 3, pp. 245–251, 2006.

[32] A. Fukushima and M. Kusano, "Recent progress in the development of metabolome databases for plant systems biology," *Frontiers in Plant Science*, vol. 4, article 73, 2013.

[33] T. Tohge and A. R. Fernie, "Web-based resources for mass-spectrometry-based metabolomics: a user's guide," *Phytochemistry*, vol. 70, no. 4, pp. 450–456, 2009.

[34] G. Blekherman, R. Laubenbacher, D. F. Cortes et al., "Bioinformatics tools for cancer metabolomics," *Metabolomics*, vol. 7, no. 3, pp. 329–343, 2011.

[35] M. Hur, A. A. Campbell, M. Almeida-de-Macedo et al., "A global approach to analysis and interpretation of metabolic data for plant natural product discovery," *Natural Product Reports*, vol. 30, no. 4, pp. 565–583, 2013.

[36] H. P. Benton, D. M. Wong, S. A. Trauger, and G. Siuzdak, "XCMS2: processing tandem mass spectrometry data for metabolite identification and structural characterization," *Analytical Chemistry*, vol. 80, no. 16, pp. 6382–6389, 2008.

[37] T. Pluskal, S. Castillo, A. Villar-Briones, and M. Orešič, "MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data," *BMC Bioinformatics*, vol. 11, article 395, 2010.

[38] R. Baran, H. Kochi, N. Saito et al., "MathDAMP: a package for differential analysis of metabolite profiles," *BMC Bioinformatics*, vol. 7, article 530, 2006.

[39] A. Lommen, "Metalign: interface-driven, versatile metabolomics tool for hyphenated full-scan mass spectrometry data preprocessing," *Analytical Chemistry*, vol. 81, no. 8, pp. 3079–3086, 2009.

[40] Z. Lei, H. Li, J. Chang, P. X. Zhao, and L. W. Sumner, "MET-IDEA version 2.06; improved efficiency and additional functions for mass spectrometry-based metabolomics data processing," *Metabolomics*, vol. 8, pp. 105–110, 2012.

[41] H. Horai, M. Arita, S. Kanaya et al., "MassBank: a public repository for sharing mass spectral data for life sciences," *Journal of Mass Spectrometry*, vol. 45, no. 7, pp. 703–714, 2010.

[42] C. A. Smith, G. O'Maille, E. J. Want et al., "METLIN: a metabolite mass spectral database," *Therapeutic Drug Monitoring*, vol. 27, no. 6, pp. 747–751, 2005.

[43] T. Sakurai, Y. Yamada, Y. Sawada et al., "PRIMe update: innovative content for plant metabolomics and integration of gene expression and metabolite accumulation," *Plant and Cell Physiology*, vol. 54, no. 2, article e5, 2013.

[44] D. S. Wishart, T. Jewison, A. C. Guo et al., "HMDB 3.0—the human metabolome database in 2013," *Nucleic Acids Research*, vol. 41, pp. D801–D807, 2013.

[45] F. M. Afendi, T. Okada, M. Yamazaki et al., "KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research," *Plant and Cell Physiology*, vol. 53, no. 2, article e1, 2012.

[46] Y. Wang, J. Xiao, T. O. Suzek, J. Zhang, J. Wang, and S. H. Bryant, "PubChem: a public information system for analyzing bioactivities of small molecules," *Nucleic Acids Research*, vol. 37, no. 2, pp. W623–W633, 2009.

[47] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, "KEGG for integration and interpretation of large-scale molecular data sets," *Nucleic Acids Research*, vol. 40, pp. D109–D114, 2012.

[48] R. Caspi, T. Altman, K. Dreher et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Research*, vol. 38, no. 1, pp. D742–D753, 2009.

[49] D. Croft, G. O'Kelly, G. Wu et al., "Reactome: a database of reactions, pathways and biological processes," *Nucleic Acids Research*, vol. 39, no. 1, pp. D691–D697, 2011.

[50] P. Bais, S. M. Moon, K. He et al., "Plantmetabolomics.org: a web portal for plant metabolomics experiments," *Plant Physiology*, vol. 152, no. 4, pp. 1807–1816, 2010.

[51] J. Kopka, N. Schauer, S. Krueger et al., "GMD@CSB.DB: the Golm metabolome database," *Bioinformatics*, vol. 21, no. 8, pp. 1635–1638, 2005.

[52] A. J. Carroll, M. R. Badger, and A. H. Millar, "The MetaboloneExpress project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets," *BMC Bioinformatics*, vol. 11, article 376, 2010.

[53] K. Haug, R. M. Salek, P. Conesa et al., "MetaboLights-an open-access general-purpose repository for metabolomics studies and associated meta-data," *Nucleic Acids Research*, vol. 41, pp. D781–D786, 2013.

[54] B. P. Bowen and T. R. Northen, "Dealing with the unknown: metabolomics and metabolite atlases," *Journal of the American Society for Mass Spectrometry*, vol. 21, no. 9, pp. 1471–1476, 2010.

[55] S. Neumann and S. Böcker, "Computational mass spectrometry for metabolomics: identification of metabolites and small molecules," *Analytical and Bioanalytical Chemistry*, vol. 398, no. 7-8, pp. 2779–2788, 2010.

[56] R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan, and D. B. Kell, "Metabolomics by numbers: acquiring and understanding global metabolite data," *Trends in Biotechnology*, vol. 22, no. 5, pp. 245–252, 2004.

[57] T. Kind, M. Scholz, and O. Fiehn, "How large is the metabolome? A critical analysis of data exchange practices in chemistry," *PLoS ONE*, vol. 4, no. 5, Article ID e5440, 2009.

[58] H. A. Piwowar, R. S. Day, and D. B. Fridsma, "Sharing detailed research data is associated with increased citation rate," *PLoS ONE*, vol. 2, no. 3, article e308, 2007.

[59] N. Sakurai, Y. Ogata, T. Ara et al., "Development of KaPPA-View4 for omics studies on Jatropha and a database system KaPPA-Loader for construction of local omics databases," *Plant Biotechnology*, vol. 29, no. 2, pp. 131–135, 2012.

[60] R. Sano, T. Ara, N. Akimoto et al., "Dynamic metabolic changes during fruit maturation in *Jatropha curcas* L," *Plant Biotechnology*, vol. 29, no. 2, pp. 175–178, 2012.