Plant Biotechnology 26, 445–449 (2009)

Bioinformatics Note

Special Issue

Improvement of the quantitative differential metabolome pipeline for gas chromatography-mass spectrometry data by automated reliable peak selection

Takeshi Ara[1], Nozomu Sakurai[1], Yoshie Tange[1], Yoshihiko Morishita[1], Hideyuki Suzuki[1], Koh Aoki[1], Kazuki Saito[1,2], Daisuke Shibata[1,*]

[1] Kazusa DNA Research Institute, Kisarazu, Chiba 292-0818, Japan; [2] Graduate School of Pharmaceutical Sciences, Chiba University, Chiba 263-8522, Japan
* E-mail: shibata@kazusa.or.jp Tel: +81-438-52-3947 Fax: +81-438-52-3948

Received August 27, 2009; accepted October 8, 2009 (Edited by A. Fukushima)

Abstract Recent advances in metabolomics technology have enabled large-scale comprehensive analyses of metabolites, but the throughput of data processing of non-targeted, quantitative differential analyses is very low. It is crucial to solve this problem to generate biological hypotheses from a large-scale dataset. To improve the analysis of metabolite data, we focused on the processing of quantitative differential analysis after multiple peak alignment. We have developed a program named FAQuant that automatically selects reliable peaks from each chromatogram, quantifies the mean of peak intensity to compare between sample groups, and selects the peaks with differences in accumulation of metabolites. This program was incorporated into a quantitative differential metabolome pipeline as a module to improve the throughput of gas chromatography-mass spectrometry dataset analysis. As a result, the module incorporation largely reduced the total processing time. Furthermore, differential analysis of metabolites in soybean (*Glycine max*) cultivars was demonstrated by use of the system. This system might facilitate biological hypothesis generation from large-scale comparative metabolome analysis.

Key words: Differential analysis, GC-TOF-MS, mass spectrometry, metabolomics, quantification.

Recently, metabolomics has emerged as a new omics field that aims to measure all metabolites in living organisms, and metabolome analyses have been applied in various research fields (Hall 2006). However, there are various technical problems in comprehensive analysis of whole metabolite profiles of living organisms. The accumulation of metabolomics data is therefore limited to a smaller scale than other omics fields such as genomics and transcriptomics (Kind et al. 2009; Tohge and Fernie 2009). Many tools and systems for metabolome analysis have been developed to improve various analytical processes for gas chromatography-mass spectrometry (GC-MS; Duran et al. 2003; Jonsson et al. 2005; Tikunov et al. 2005; Broeckling et al. 2006; Bunk et al. 2006; Luedemann et al. 2008; Neuweger et al. 2008; Hiller et al. 2009; Oishi et al. 2009), liquid chromatography-mass spectrometry (LC-MS; Katajamaa et al. 2006; Smith et al. 2006; Sturm et al. 2008) and capillary electrophoresis-mass spectrometry (CE-MS; Baran et al. 2006; Morohashi et al. 2007). However, throughput of comparative analysis of metabolome data, especially for quantitative differential analysis, is very low since there are many time-consuming processes. For example, analytical processes such as noise filtering, peak deconvolution, multiple alignments, annotation of metabolite names, and selection of peaks for statistical analysis require many manual correction steps with the help of experts to produce reliable biological hypotheses. Moreover, the number of parameters to optimize these processes is enormous, and there are also computational limitations such as data transaction speed and treatable data size, since the size of metabolomics datasets is extremely large (~100 megabytes/run). A few cases have been reported that have addressed these issues, including large-scale metabolome data analysis, comprehensive and large-scale parameter estimation, and improvement of biological hypothesis generation efficiency (van den Berg et al. 2006; Lu et al. 2008; Peters et al. 2009). It is essential to solve these problems for high-throughput generation of biological hypotheses from large-scale metabolome datasets.

In the current study, we developed a program to automate a reliable peak selection process (such as removal of outlier and estimation of missing peak values)

Abbreviations: GC-TOF-MS, gas chromatography-time-of-flight-mass spectrometry; MST, mass spectral tag.
This article can be found at http://www.jspcmb.jp/

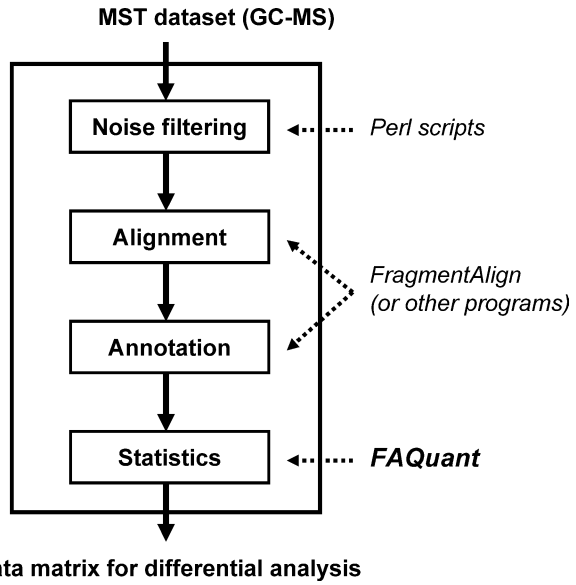

Figure 1. Quantitative differential metabolome pipeline for GC-TOF-MS datasets. Deconvoluted MST datasets in a public metabolome database are applied to the pipeline as input. The pipeline contains steps for noise filtering, multiple peak alignment (FragmentAlign or other programs), compound name annotation, and statistical processing of alignment (FAQuant) to output data matrix for differential analysis.

for quantitative differential analysis after peak alignment (Figure 1). The incorporation of this program into a quantitative differential metabolome pipeline as a module largely reduced the total processing time and enabled batch processing of the peak selection process. We applied the pipeline to public gas chromatography-time-of-flight-mass spectrometry (GC-TOF-MS) datasets in MassBase (http://webs2.kazusa.or.jp/massbase/) and the resulting biological hypotheses are described here.

As a test dataset, mass spectral tags (MSTs; Schauer et al. 2005) from 19 analytical datasets which contained three to four repetitive GC-TOF-MS measurements of metabolites in wild soybean (*Glycine max* subsp. soja) and three soybean cultivars (*G. max* Misuzu, Masshoku, and Koitozairai) seeds were obtained from the public metabolome database MassBase (Accession numbers: MDGC1_02413, 02414, 02417 to 02419, 02422, 02423, 02426, 02427, 02430, 02431, 02433, 02434, 02439, 02440, 02442, 02443, 02445, 02446). MSTs of the same cultivar were grouped into a sample group. Metadata was acquired from the database or an additional survey was performed (e.g., sample condition: dry seed; sample preparation method: methanol extraction after crush by a mortar; sample fraction: polar; internal standard: ribitol; retention index method: using n-alkane (C12–36); deconvolution program: ChromaTOF, etc).

For each MST, noise peaks were removed by peak property assignment. The definition of a noise peak is described in the README file from the MassBase download page (http://webs2.kazusa.or.jp/massbase/index.php?action=Massbase_ShowDownload). Multiple alignment and compound name annotation were performed by the in-house program FragmentAlign ver 1.12 that can align peaks by similarity of fragmentation patterns and can correct alignments manually (Sakurai et al. in preparation). The parameters of this program were set at the default values except for the retention index permission width (15) and correlation coefficients to calculate the similarity between mass spectra (−1, 0.8, −1 for Pearson correlation, cosine correlation, Spearman correlation, respectively) for alignment. After the automatic alignment process, manual corrections of the resulting alignment were performed. Compound names of aligned peaks were identified using an in-house mass spectral library of known compounds. Several ambiguous annotations were assigned with a term "putative" in these annotations. Additionally, aligned peaks with retention indices below 900 and above 3500 were too complicated to correct manually. These peaks were removed from later analysis. Finally a tab-delimited table of aligned peaks with intensity values was produced (Supplemental Table 1). The average mass spectrum of each aligned peak was output to a file in NIST format (Linstrom and Mallard 2005).

For differential analysis of metabolites, we have developed a program named FAQuant that aims to evaluate the reliability of each peak by appearance frequency and distribution of peak intensity in each sample group for each aligned peak, and to select high confidence peaks to perform quantification and differential analysis among sample groups from a multiple alignment. Input file format is based on the output files of FragmentAlign but it is possible to adjust to text format output file of other peak alignment programs. The algorithm and parameters of the peak selection process were designed based on conventional manual procedures and were implemented using the script language Perl (De Souza et al. 2006; Steinfath et al. 2008). This program consists of the four following steps: 1) selection of the available peaks, 2) evaluation of the reliability of peaks in sample groups and calculation of the mean of peak intensity, 3) comparison of the mean peak intensity between sample groups for all aligned peaks, and 4) selection of aligned peaks that are reliable and show intensity differences among the sample groups (Figure 2). Details of the algorithm are described below.

At first, each peak intensity is corrected by the intensity of the internal standard (IS) incorporated in each sample so that IS intensities are the same value, the mean IS intensity of all measurements. After this correction, peaks with higher intensities than the threshold (e.g., 10000) are selected as available peaks. When more than two available peaks are observed in a sample group of an aligned peak, this is defined as the aligned peak being detected in the sample group. These peaks are represented as peak group $P_{i,j}$ [i-th aligned

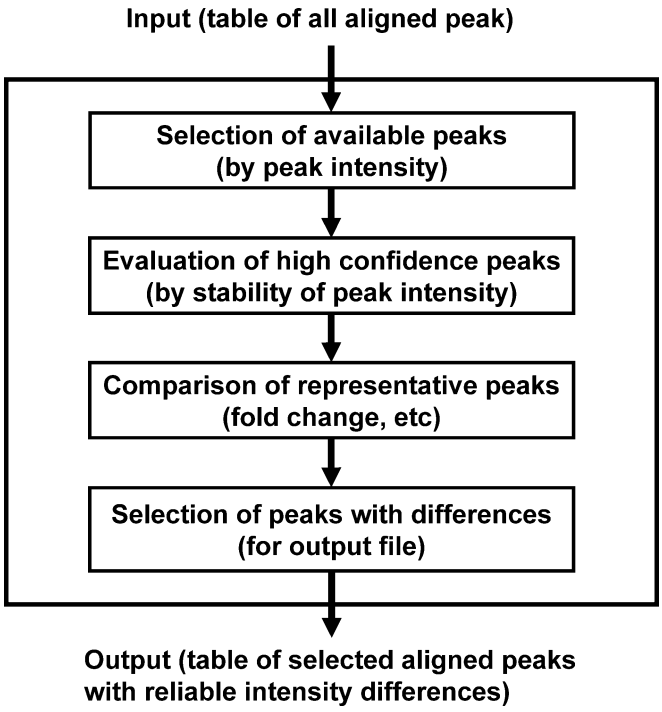

Figure 2. Processing steps of FAQuant. FAQuant converts the input file (table of all aligned peaks) to an output file (table of selected aligned peaks with reliable intensity differences) through 1) selection of available peaks by peak intnsity, 2) evaluation of high confidence peaks by stability of peak intensity among repetitive experiments, 3) comparison of representative peaks (fold change, etc), and 4) selection of peaks that show intensity differences. Details of the algorism are descrived in the text.

peak, *j*-th sample group]. To estimate the magnitude of mean peak intensity of $P_{i,j}$, the integer part of a logarithm to base 10 of peak intensity (C_{peak}) is calculated. This approximation is effective to detect outliers in $P_{i,j}$ easier prior to calculation of the mean peak intensity for differential analysis. If there are peaks having the same value in each $P_{i,j}$, these peaks are grouped as a "high-confidence peak set". After this calculation, the level of reliability of each $P_{i,j}$ is defined as follows. In the case of only one high-confidence peak set existing in a $P_{i,j}$, and the number of peaks in the peak set is >50% of the number of samples for the sample group, and >50% of the number of the whole available peaks are in the $P_{i,j}$, the level of reliability is set to "1". If there is a single high-confidence peak set but the criteria of the number of peaks in the peak set described above is not satisfied, the level of reliability is set to "2". If there are multiple high-confidence peak sets, the level of reliability is set to "3", and in the case that no high-confidence peak set is observed, the level of reliability is set to "4".

The mean peak intensity and its width (the difference of maximum C_{peak} and minimum C_{peak}; C_{max}–C_{min}) are calculated from peaks in the peak set. In the case that the level of reliability of $P_{i,j}$ is 3 or 4, all available peaks are used for the calculation of the mean peak intensity and its width. Accordingly, all peaks in $P_{i,j}$ are represented by a peak model that has this mean peak intensity value as a property of peak intensity. This peak model is defined as a representative peak of $P_{i,j}$. In cases where the difference in the mean peak intensity of a representative peak between a target sample group (M_{target}) and the blank sample group (M_{blank}) in an aligned peak is over the threshold (e.g., 10000) and the ratio of the mean peak intensity of the target sample group and the blank sample group (M_{target}/M_{blank}) is over the threshold (e.g., 1.1), the difference M_{target}–M_{blank} is used to calculate the ratio of the mean peak intensity between the target sample and control sample groups. If the difference M_{target}–M_{blank} is below zero, the value is replaced with zero and the representative peak is classified as "not detected". If the ratio of the mean peak intensity between the target sample and control sample groups is more than (e.g., 2) or less than (e.g., 0.5) the threshold, the representative peak is classified into the "up" or "down" category, respectively. If the level of reliability of one or both peaks is "2", then the term "putative" is added to the category. If the level of reliability of either or both representative peaks is 3, the peak is classified into "increase" or "decrease" when the following criteria are satisfied:

Criteria: X and Y<threshold (e.g. 3), Z>X, Z>Y

where $X = C_{max(target)} - C_{min(target)}$

$Y = C_{max(control)} - C_{min(control)}$

$Z = |C_{mean(target)} - C_{mean(control)}|$

When a representative peak is detected with only one sample group to compare, the difference is classified into the "new" or "lost" category. Furthermore, if the level of

reliability of the representative peak is “2” or “3”, the term “putative” is added to the category. Other cases are categorized into “no difference”. Any aligned peak is removed from the original alignment if no representative peak exists except for blank samples or if its normalized fragmentation pattern is assigned to a noise peak. After these processes, a summarized result of aligned peaks is generated as a tab-delimited text format file (Supplemental Table 2). Any threshold described above is easily changeable for recalculations. Accuracy of the calculation by FAQuant was manually evaluated by using randomly selected peaks in the input/output files.

A differential analysis of accumulated metabolites among cultivars of soybean was performed to demonstrate the usage of this system. FAQuant greatly improved the throughput of the reliable peak selection process for hundreds of peaks in an alignment from a more than one week manual process to a less than 1 min automated process. This performance can adjust parameters faster and achieve high throughput batch processing for this part of the pipeline. After the peak selection process, about 30% of representative peaks of the dataset were classified as “1” and 20% were classified as “2” or “3” as the level of reliability. Out of 701 aligned peaks, 490 were subjected to a differential analysis as the selected reliable dataset. Compound name annotation was given in about 20% of the 490 aligned peaks. Finally, biological hypotheses were generated from these results with some manual operations. 1) A total of about 200 reliable (~level of reliability=1) representative peaks were detected in each soybean species. Among these peaks, 32 peaks were observed in all cultivers and 49 cultivar-specific accumulated peaks were observed (Table 1). 2) A total of 319 reliable peaks showed differences in accumulation against peaks observed in wild soybean. Among these 319 reliable representative peaks, 47, 52, 96, and 124 were classified into “up”, “down”, “new”, and “lost” categories of difference, respectively (Figure 3). 3) After this statistical survey, individual peaks with known metabolite names were investigated in detail. Compared with wild soybean, the S-adenosyl-methionine and arginine contents tended to be decreased in other soybean species. In addition, uracil accumulation was observed only in wild soybean. Further investigation is necessary to interpret these generated results since there are imperfect annotations and metadata in the dataset. However, improving the throughput of the peak selection process for generating biological hypotheses under various parameter settings greatly advances the quantitative differential metabolome pipeline. The pipeline will greatly enhance comparative metabolomics research and these results will be registered in the metabolome database Komics (http://webs2.kazusa.or.jp/komics/) in the future.

Table 1 Difference of reliable peaks among cultivars of *Glycine max*

Sample group name	#sample	#peak (reliable peak)	#specific peak
G. max subsp. soja	4	353 (174)	18
G. max Misuzu	4	337 (186)	9
G. max Masshoku	4	291 (150)	4
G. max Koitozairai	3	242 (192)	18

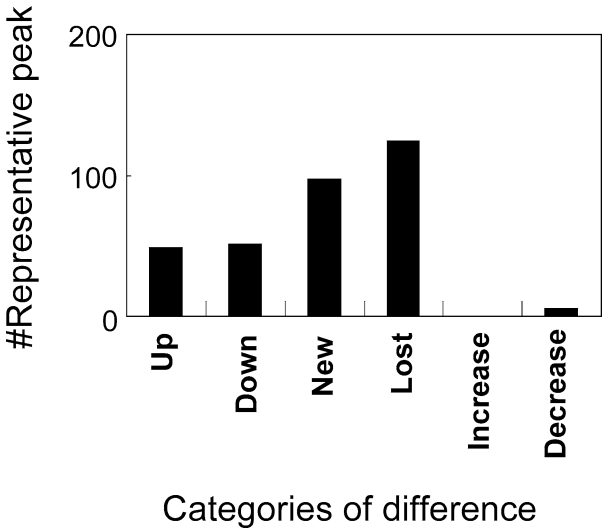

Figure 3. Distribution of a number of reliable peaks that are automatically classified into categories of difference. The number of reliable representative peaks classified into categories (Up, Down, New, Lost, Increase, and Decrease) selected by FAQuant process with default parameters are shown.

In conclusion, a program to automate the peak selection process in differential analysis of metabolites was developed and incorporated into a quantitative differential metabolome pipeline resulting in a great improvement of the throughput for generating a reliable data matrix, thereby producing biological hypotheses from GC-TOF-MS datasets. In addition, a parameter search by large-scale batch processing is now possible for the differential analysis. This system might enhance the accumulation of a large-scale comparative metabolomics results available to the public. This program is available on request.

Acknowledgements

We thank Dr. K. Harada (Chiba University) for his gift of *G. max* seeds. This work was supported by a grant from the New Energy and Industrial Technology Development Organization (NEDO) of Japan as part of the “Development of Fundamental Technologies for Controlling the Material Production Process of Plants”.

References

Baran R, Kochi H, Saito N, Suematsu M, Soga T, Nishioka T, Robert M, Tomita M (2006) MathDAMP: a package for differential analysis of metabolite profiles. *BMC*

Bioinformatics 7: 530
Broeckling CD, Reddy IR, Duran AL, Zhao X, Sumner LW (2006) MET-IDEA: data extraction tool for mass spectrometry-based metabolomics. *Anal Chem* 78: 4334–4341
Bunk B, Kucklick M, Jonas R, Münch R, Schobert M, Jahn D, Hiller K (2006) MetaQuant: a tool for the automatic quantification of GC/MS-based metabolome data. *Bioinformatics* 22: 2962–2965
De Souza DP, Saunders EC, McConville MJ, Likić VA (2006) Progressive peak clustering in GC-MS metabolomic experiments applied to *Leishmania parasites*. *Bioinformatics* 22: 1391–1396
Duran AL, Yang J, Wang L, Sumner LW (2003) Metabolomics spectral formatting, alignment and conversion tools (MSFACTs). *Bioinformatics* 19: 2283–2293
Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. *New Phytol* 169: 453–468
Hiller K, Hangebrauk J, Jäger C, Spura J, Schreiber K, Schomburg D (2009) MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal Chem* 81: 3429–3439
Jonsson P, Johansson AI, Gullberg J, Trygg J, A J, Grung B, Marklund S, Sjöström M, Antti H, Moritz T (2005) High-throughput data analysis for detecting and identifying differences between samples in GC/MS-based metabolomic analyses. *Anal Chem* 77: 5635–5642
Katajamaa M, Miettinen J, Orešič M (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* 22: 634–636
Kind T, Scholz M, Fiehn O (2009) How large is the metabolome? A critical analysis of data exchange practices in chemistry. *PLoS One* 4: e5440
Linstrom PJ, Mallard WG (2005) NIST Chemistry WebBook, NIST Standard Reference Databases Number 69, National Institute of Standards and Technology, Gaithersburg, MD, http://webbook.nist.gov.
Lu H, Dunn WB, Shen H, Kell DB, Liang Y (2008) Comparative evaluation of software for deconvolution of metabolomics data based on GC-TOF-MS. *Trends Anal Chem* 27: 215–227
Luedemann A, Strassburg K, Erban A, Kopka J (2008) TagFinder for the quantitative analysis of gas chromatography-mass spectrometry (GC-MS)-based metabolite profiling experiments. *Bioinformatics* 24: 732–737
Morohashi M, Shimizu K, Ohashi Y, Abe J, Mori H, Tomita M, Soga T (2007) P-BOSS: a new filtering method for treasure hunting in metabolomics. *J Chromatogr A* 1159: 142–148
Neuweger H, Albaum SP, Dondrup M, Persicke M, Watt T, Niehaus K, Stoye J, Goesmann A (2008) MeltDB: a software platform for the analysis and integration of metabolomics experiment data. *Bioinformatics* 24: 2726–2732
Oishi T, Tanaka K, Hashimoto T, Shinbo Y, Jumtee K, et al. (2009) An approach to peak detection in GC-MS chromatograms and application of KNApSAcK database in prediction of candidate metabolites. *Plant Biotechnol* 26: 167–174
Peters S, van Velzen E, Janssen HG (2009) Parameter selection for peak alignment in chromatographic sample profiling: objective quality indicators and use of control samples. *Anal Bioanal Chem* 394: 1273–1281
Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, et al. (2005) GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Lett* 579: 1332–1337
Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G (2006) XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem* 78: 779–787
Steinfath M, Groth D, Lisec J, Selbig J (2008) Metabolite profile analysis: from raw data to regression and classification. *Physiol Plant* 132: 150–161
Sturm M, Bertsch A, Gröpl C, Hildebrandt A, Hussong R, et al. (2008) OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinformatics* 9: 163
Tikunov Y, Lommen A, de Vos CH, Verhoeven HA, Bino RJ, Hall RD, Bovy AG (2005) A novel approach for nontargeted data analysis for metabolomics. Large-scale profiling of tomato fruit volatiles. *Plant Physiol* 139: 1125–1137
Tohge T, Fernie AR (2009) Web-based resources for mass-spectrometry-based metabolomics: A user's guide. *Phytochemistry* 70: 450–456
van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ (2006) Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics* 7: 142
Xia J, Psychogios N, Young N, Wishart DS (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. *Nucl Acids Res* 37: W652–660