analytical chemistry

LETTER

pubs.acs.org/ac

metaXCMS: Second-Order Analysis of Untargeted Metabolomics Data

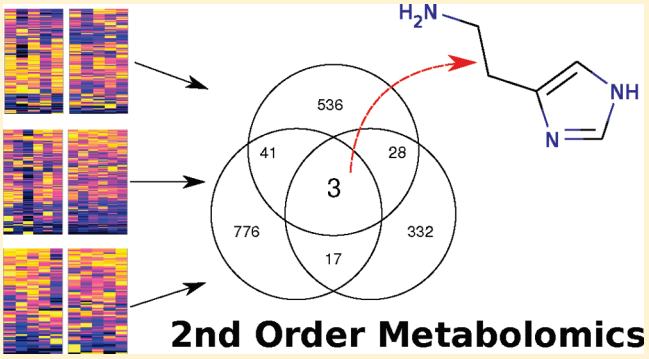
Ralf Tautenhahn,*[,†] Gary J. Patti,[†] Ewa Kalisiak,[†] Takashi Miyamoto,[‡] Manuela Schmidt,[‡] Fang Yin Lo,[§] Joshua McBee,[§] Nitin S. Baliga,[§] and Gary Siuzdak*[,†]

[†]Department of Chemistry and Molecular Biology, Center for Metabolomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[‡]Department of Cell Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, United States

[§]Institute for Systems Biology, Seattle, Washington 98103-8904, United States

ABSTRACT: Mass spectrometry-based untargeted metabolomics often results in the observation of hundreds to thousands of features that are differentially regulated between sample classes. A major challenge in interpreting the data is distinguishing metabolites that are causally associated with the phenotype of interest from those that are unrelated but altered in downstream pathways as an effect. To facilitate this distinction, here we describe new software called metaXCMS for performing second-order ("meta") analysis of untargeted metabolomics data from multiple sample groups representing different models of the same phenotype. While the original version of XCMS was designed for the direct comparison of two sample groups, metaXCMS enables meta-analysis of an unlimited number of sample classes to facilitate prioritization of the data and increase the probability of identifying metabolites causally related to the phenotype of interest. metaXCMS is used to import XCMS results that are subsequently filtered, realigned, and ultimately compared to identify shared metabolites that are up- or down-regulated across all sample groups. We demonstrate the software's utility by identifying histamine as a metabolite that is commonly altered in three different models of pain. metaXCMS is freely available at http://metlin.scripps.edu/metaxcms/.



Metabolites are small molecules within biological systems that serve as the substrates and products of cellular reactions. There is enormous structural diversity among metabolites ranging from polar compounds to lipids and drug derivatives. Untargeted metabolomics describes the process by which these molecules are globally profiled without bias. With the use of modern mass spectrometers interfaced with either gas or liquid chromatography, tens of thousands of metabolic features can typically be detected from cells, biofluids, and tissues.[1,2] A metabolomics feature represents a peak in the chromatogram and is defined as a molecular entity with a unique mass and retention time. Unbiased metabolomics is performed by first comprehensively identifying every feature within a sample group, and then comparing the relative intensity of each of them among different sample classes (e.g., healthy versus disease) for statistically significant changes.

Over the course of the past 5 years, several software tools for differential analysis of mass spectrometry-based metabolomics data have been developed (e.g., XCMS,[3,4] MZmine,[5,6] and MathDAMP[7]). These programs identify features whose relative intensity varies between sample groups and are therefore useful in screening for biomarkers of disease. In addition, however, the identification of dysregulated metabolites has been useful in making advances to our understanding of fundamental biochemistry. For example, untargeted metabolomics programs have successfully been applied to reveal new insights related to inborn errors of human metabolism,[8] extremophile bacteria,[9] viral pathogenesis,[10] the gut microbiome,[11,12] and stem cell differentiation.[13] A major challenge in interrogating complex biological phenomena at the metabolite level, however, is in distinguishing dysregulated pathways that are causally associated with the phenotype of interest from those that are unrelated but altered as a downstream effect. Knockout model organisms provide exciting opportunities to study disease, but metabolomics data sets comparing these organisms to wildtype controls are complicated by the potentially large number of altered features causally unrelated to the pathology. Examining more animal models of the same phenotype increases the likelihood of identifying features associated with underlying disease pathology, but previous metabolomics software limits this type of analysis in that only two sample groups can be compared.

It is important to emphasize that metabolomics programs such as XCMS identify dysregulated features, not metabolites. The process of identifying a feature as a metabolite requires searching databases on the basis of accurate mass and comparing the retention time and tandem mass spectrometry (MS/MS) data to that of a model compound for structural confirmation. Growing metabolite databases with advanced functionality have

Received: November 11, 2010
Accepted: December 13, 2010
Published: December 21, 2010

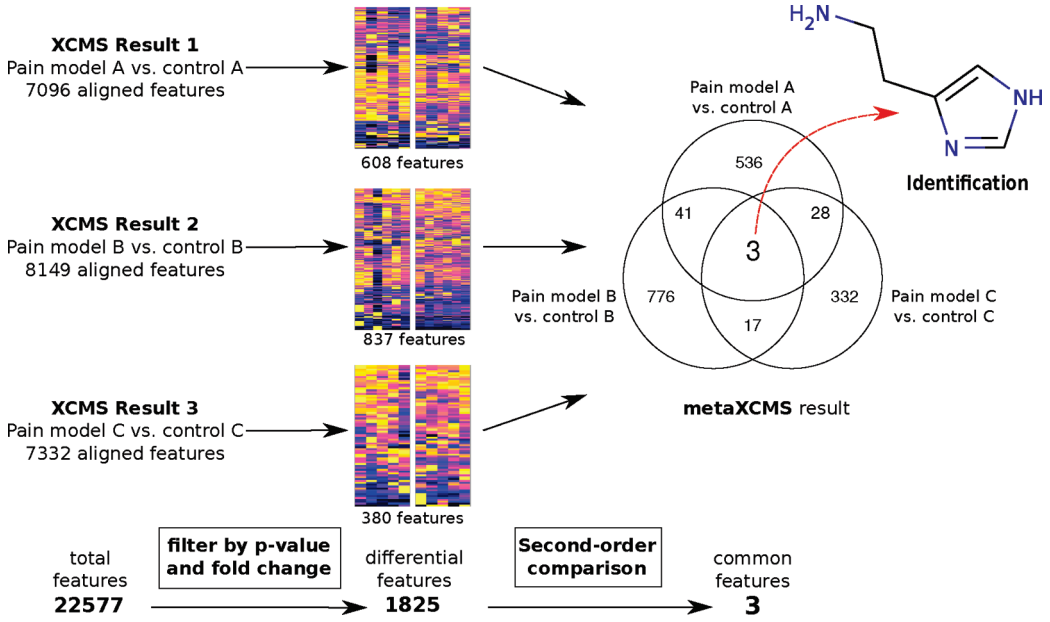


Figure 1. Data reduction with metaXCMS. The workflow is demonstrated using the pain data set, where pain model A is CFA-treated animals, pain model B is heat-treated animals, and pain model C is KRN-treated animals. The applied fold changes and *p*-values for the first filtering step were ≥1.5 and ≤0.05, respectively.

facilitated the procedure of metabolite identification,[14−16] but it is still a time-consuming and labor-intensive step of the metabolomics workflow. Thus, data reduction is essential to maximizing the physiological relevancy obtained from metabolomics experiments. The challenge is implementing an intelligent methodology to accomplish data reduction at the feature level prior to metabolite identification.

An effective data reduction strategy used in other fields has been performing second-order comparisons to identify shared disturbances among shared phenotypes. Such second-order analyses require the input of multiple sample groups, which previously has not been feasible with existing metabolomics programs. Here we describe new untargeted metabolomics software that can be used in conjunction with XCMS to perform second-order ("meta") analysis. Pairs of sample groups are first traditionally analyzed with XCMS, and the output files from any number of pair comparisons are then subsequently input into metaXCMS where they are realigned, statistically evaluated, and compared for shared differences. This offers an important metabolomics data reduction tool that has the potential to significantly decrease the number of interesting features selected for subsequent metabolite identification (Figure 1). metaXCMS is freely available as an open-source R-package that includes a graphical user interface. It can be downloaded from http://metlin.scripps.edu/metaxcms/.

WORKFLOW

The data processing workflow using metaXCMS can be summarized in the following steps. First, metaXCMS is used to import the data from multiple metabolomics experiments as TSV (tab separated) files or Excel sheets (.xlsx), both of which are standard formats exported by XCMS that can be directly imported into metaXCMS without any further preprocessing. After loading of the experimental data, sample class assignments are verified and the control group for each experiment is defined. During the second processing step, feature lists are filtered by a fold change (e.g., ≥2), *p*-value (e.g., ≤0.01), or predefined patterns of up- and down-regulation (e.g., features that are up-regulated in a first experiment, wildtype 1 vs knockout 1, but down-regulated in a second experiment, wildtype 2 vs knockout 2). In addition, feature lists can be designed to be subtracted from the final result so that metabolic changes in a control experiment (such as a reverse knockout) can be disregarded. The next step is the automated alignment of the feature lists from the different experiments on the basis of both m/z and retention time. The alignment method "group.nearest" that is implemented in XCMS is employed to align the data within user-defined m/z and retention time windows. The best results are achieved if the same liquid chromatography/mass spectrometry (LC/MS) conditions are used for all samples that are to be aligned.

While the number of data sets that can be compared with metaXCMS is generally not limited, a direct visualization of the result as a Venn diagram is only possible for instances in which the number of sample groups is five or less. The Venn diagram shows the number of common features to all sample groups as well as the number of features contained within other possible intersections. It should be noted that in some experiments, features that are *not* shared among sample groups may be the most biologically interesting, depending on the model systems being investigated and the question being asked. All metaXCMS results are displayed in tables that can be exported as Excel sheets.

For a detailed visual verification of the results, the retention time correction for all raw data files is recalculated using OBI-Warp[17] and extracted ion chromatograms (EIC) are generated for all selected features. Furthermore, boxplots are generated to visualize the distribution of feature intensities across the experiments. All graphic results can be exported as PNG or PDF files. The visualization displays are shown in Figure 2 for an example data set.

EXPERIMENTAL SECTION

As a demonstration of the utility of the software, here we briefly describe the experimental application of metaXCMS to

dx.doi.org/10.1021/ac102980g |*Anal. Chem.* 2011, 83, 696–700

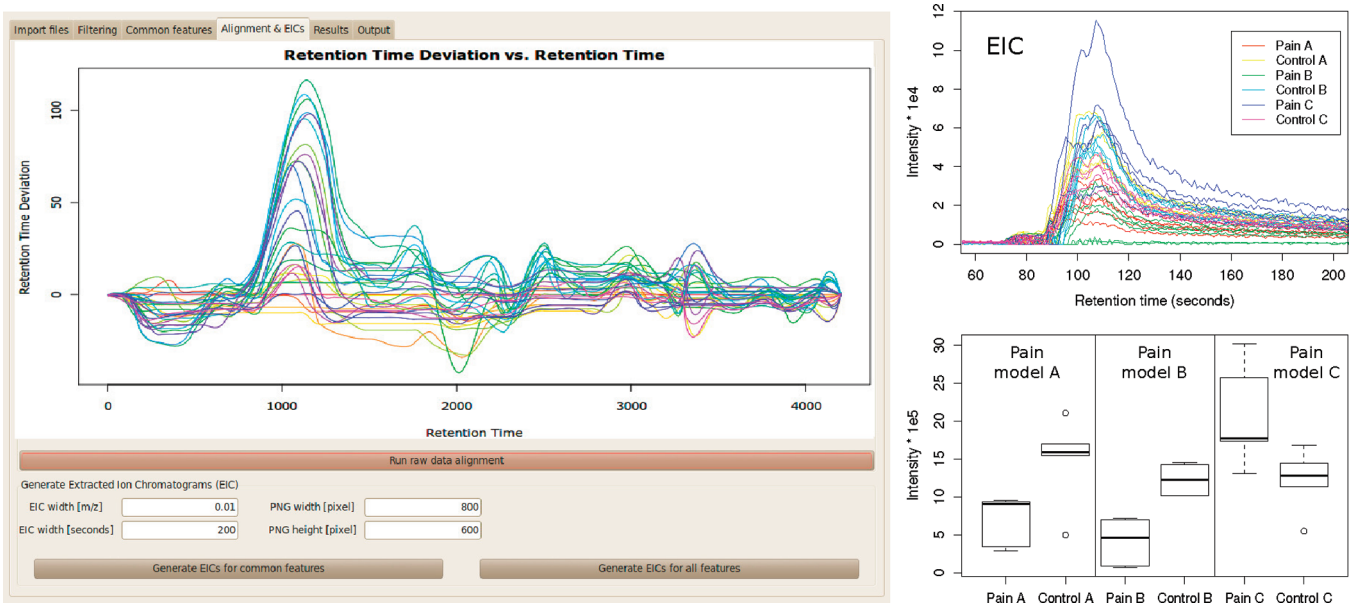


Figure 2. metaXCMS screenshot showing retention time correction curves and parameter settings for extracted ion chromatogram (EIC) generation (left), EIC overlay showing the ion intensity for m/z 112.09 ± 0.01 for all samples from the pain data set (upper right), and boxplots showing the distribution of integrated intensities of the feature m/z 112.09 for all pain samples (lower right). Pain models A, B, and C are defined in the legend of Figure 1.

the investigation of three mouse models of pain as well as five *Halobacterium salinarum* knockout organisms. For the pain study, metabolites were extracted from 10 mg pieces of skin isolated from the hind paw of animals. The following animal groups were compared: (A) animals plantar injected with Complete Freund's Adjuvant (CFA) and control animals, (B) animals to which noxious heat was acutely applied to the hind paw and room temperature-treated control animals; and (C) animals intraperitoneally injected with serum from K/BxN mice (i.e., KRN-treated mice) and vehicle-treated controls. These animals represent an inflammatory model,[18] an acute heat model,[19] and a spontaneous arthritis model of pain,[20] respectively. All experiments were conducted in accordance with the National Institutes of Health and the Scripps Research Institute animal care and use guidelines. Five biological replicates were used for each pain model and control group.

Halobacterium salinarum cultures of four knockout strains (Δ*VNG1816G*, Δ*VNG2094G*, Δ*VNG1179C*, and Δ*VNG0314G*) were grown to logarithmic-growth phase and compared to their parent control strain (Δ*ura3*). Cultures were centrifuged, rinsed with phosphate buffer solution, and lyophilized. Metabolites were extracted from 5 mg of frozen cell pellets from each culture. The genes VNG1816G, VNG2094G, and VNG1179C encode leucine-responsive regulatory protein (Lrp) family transcription factors.[21,22] The proteins encoded by VNG1816G and VNG2094G share binding sites that are upstream to genes involved in glutamic acid metabolism. The Cu-responsive transcription factor VNG1179C represses VNG2094G, suggesting that glutamic acid metabolism is transcriptionally influenced by Cu-trafficking.[22] VNG0314G encodes an enzyme in the shikimate biosynthesis pathway. VNG0314G served as a negative control for a metabolic perturbation that does not affect glutamic acid biosynthesis.

From skin tissue and lyophilized cell cultures, metabolites were extracted using cold methanol and acetone as described before.[13] Liquid chromatography was performed using a reverse-phase C18 column (Zorbax C18, Agilent, 5 mM, 150 mm × 0.5 mm diameter column) with a flow rate of 20 μL/min. Samples were analyzed by using electrospray ionization time-of-flight mass spectrometry (Agilent 6510 TOF) with water and acetonitrile for mobile phases A and B, each containing 0.1% formic acid. The chromatography started at 90% mobile phase A with a 45 min linear gradient to 98% mobile phase B.

RESULTS AND DISCUSSION

Although each of the pain models used in this study involves different pathogenic etiologies and mechanisms, we hypothesized that there may be common metabolites involved in triggering the transduction of nociceptive signals. We first compared each of the pain models to its respective control by using XCMS. XCMS performs feature detection as well as nonlinear retention time alignment and calculates statistics (Welch t test) for each feature. The XCMS result is a table that contains the m/z and retention time coordinates, p-value and fold change for each feature, and the integrated feature intensities from all aligned samples. Each one of the three pain model comparisons resulted in more than 7 000 features, with the total number of summed features from all comparisons being 22 577. The three TSV files that were generated by XCMS were imported into metaXCMS and filtered by fold change (≥1.5) and p-value (≤0.05). No restrictions were made on up- or down-regulation.

The filter step yielded 380, 837, and 608 differentially regulated features for each one of the pairwise comparisons, resulting in a total number of 1825 dysregulated features (Figure 1, heatmaps). The second-order comparison was applied using a tolerance of 0.01 m/z and 60 s retention time. Three features were found to be differentially regulated in all three pain models (Table 1).

Retention time correction was then applied to the raw data for the generation of extracted ion chromatograms for each shared feature. In addition, boxplots were created on the basis of

dx.doi.org/10.1021/ac102980g |*Anal. Chem.* 2011, 83, 696–700

Table 1. Features That Were Found to Be Differentially Regulated in the Pain Models A, B, and C As Defined in Figure 1[a]

		pain model A vs control A		pain model B vs control B		pain model C vs Control C	
m/z	retention time (s)	fold change	p-value	fold change	p-value	fold change	p-value
112.09	107	2.2	0.04	3.0	4×10^{-4}	1.7	0.05
233.10	85	1.8	0.04	1.8	0.02	1.7	0.05
112.18	107	2.2	0.04	4.1	4×10^{-5}	2.0	0.05

[a] The fold change and p-values are shown as calculated by XCMS for the pairwise comparison.

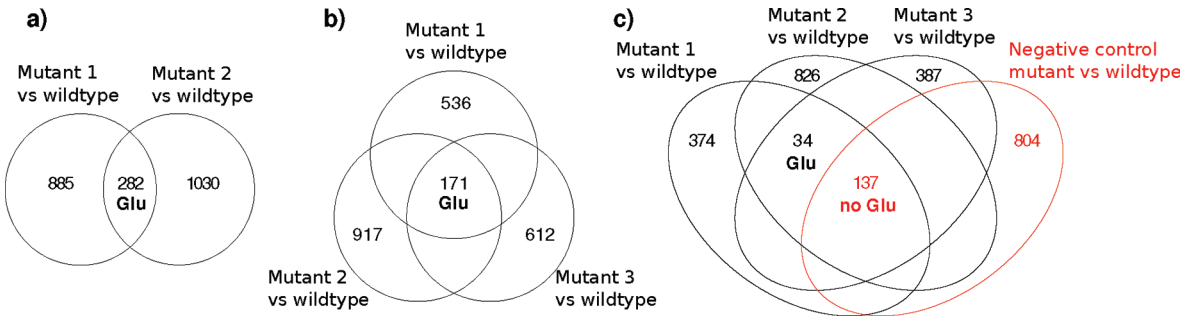


Figure 3. Venn diagrams showing the results of the second-order comparisons of four different *Halobacterium salinarum* knockout strains using metaXCMS. Mutants 1, 2, and 3 represent the strains Δ*VNG2094G*, Δ*VNG1816G*, and Δ*VNG1179C*, all of which are characterized by perturbations in glutamic acid metabolism. Δ*VNG0314G* does not affect glutamic acid metabolism, so Δ*VNG0314G* served as a negative control.

integrated intensity as exported from XCMS. Retention time correction curves for all pain samples is shown in Figure 2, along with overlaid extracted ion chromatograms and boxplots for the feature m/z 112.09. The compound with an observed m/z value of 112.09 was identified as histamine using accurate mass, retention time, and MS/MS fragmentation data as compared with a model compound. As validation of the approach, our result is consistent with the literature in that histamine has been well characterized as a mediator of pain by several mechanisms.[23−25] It has been shown that irritation of the skin by mechanical, electrical, or chemical stimuli causes release of histamine by mast cells resulting in sensory nerve ending depolarization.[26] In the skin, histamine receptors are found on Aβ fibers, on keratinocytes, on Merkel cells, and on deep dermal Aδ fibers terminating on dermal blood vessels.[27] The precise physiological effects of histamine are complex due to its immunomodulatory and neurotransmitter properties,[28,29] but our observation that it is commonly dysregulated is consistent with that which has been reported previously.

For the *Halobacterium salinarum* study, second-order analysis was performed on Δ*VNG2094G* and Δ*VNG1816G* with respect to their parent strains. For these mutants, 282 shared differences were detected. Among those, glutamic acid was found to be similarly dysregulated as expected (Figure 3a). The identity of glutamic acid was confirmed using accurate mass and MS/MS data as compared with a model compound. A higher-order analysis was also performed in which the difference profile from Δ*VNG1179C* with respect to its wildtype was introduced into the comparison. The number of shared differences decreased from 282 to 171 (Figure 3b). Importantly, glutamic acid was similarly dysregulated in all three mutants, supporting the previously suggested physiological link between Cu-trafficking and glutamic acid metabolism (Δ*VNG2094G*, fold change 1.6, p-value 0.01; Δ*VNG1816G*, fold change 2.2, p-value 0.03; Δ*VNG1179C*, fold change 1.9, p-value 0.01). Finally, as a negative control, the comparison of Δ*VNG0314G* to its parent strain was introduced into the analysis. VNG0314G encodes an enzyme involved in shikimate biosynthesis that is unrelated to glutamic acid metabolism and therefore Δ*VNG0314G* served as a negative control. Glutamic acid was not detected as a differentially regulated metabolite among all four mutants (Figure 3c). The decrease in shared features among all samples with the addition of Δ*VNG0314G* demonstrates the utility of eliminating features not specifically related to the phenotype of interest with metaXCMS by using a negative control.

CONCLUSIONS

In summary, metaXCMS provides software for second-order analysis of metabolomics data facilitating meta-comparisons similar to those already used in genomics and transcriptomics.[30−33] The introduction of such software in metabolomics is of significant value as it not only provides an analytical tool for distinguishing metabolites fundamentally associated with the underlying origin of a particular phenotype, but it also allows for data reduction at the feature level. Structural characterization of features is a rate-limiting step in the metabolomics workflow, and therefore metaXCMS offers a method to efficiently identify features with a higher likelihood to be biologically relevant prior to the time commitment of compound identification. In addition, metaXCMS provides a tool to analyze large cohorts of clinical samples from different groups or with complex subgroup variability.

AUTHOR INFORMATION

Corresponding Author

*E-mail: rtautenh@scripps.edu (R.T.); siuzdak@scripps.edu (G.S.).

ACKNOWLEDGMENT

We thank Matt Petrus at the Genomics Institute of the Novartis Research Foundation for help with preparation of the animal models used in the study. This work was supported by the California Institute of Regenerative Medicine

(Grant TR1-01219), the National Institutes of Health (Grants R24 EY017540-04, P30 MH062261- 10, and P01 DA026146-02), and NIH/NIA Grant L30 AG0 038036 (G.J.P.). Financial support was also received from the Department of Energy (Grants FG02-07ER64325 and DE-AC0205CH11231).

REFERENCES

(1) Dunn, W. B. *Phys. Biol.* **2008,** *5*, 011001.
(2) Want, E. J.; Nordström, A.; Morita, H.; Siuzdak, G. *J. Proteome Res.* **2007,** *6*, 459–468.
(3) Smith, C.; Want, E.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006,** *78*, 779–787.
(4) Tautenhahn, R.; Böttcher, C.; Neumann, S. *BMC Bioinf.* **2008,** *9*, 504.
(5) Katajamaa, M.; Oresic, M. *J. Chromatogr., A* **2007,** *1158*, 318–328.
(6) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010,** *11*, 395.
(7) Baran, R.; Kochi, H.; Saito, N.; Suematsu, M.; Soga, T.; Nishioka, T.; Robert, M.; Tomita, M. *BMC Bioinf.* **2006,** *7*, 530.
(8) Wikoff, W. R.; Gangoiti, J. A.; Barshop, B. A.; Siuzdak, G. *Clin. Chem.* **2007,** *53*, 2169–2176.
(9) Kalisiak, J.; Trauger, S. A.; Kalisiak, E.; Morita, H.; Fokin, V. V.; Adams, M. W. W.; Sharpless, K. B.; Siuzdak, G. *J. Am. Chem. Soc.* **2009,** *131*, 378–386.
(10) Wikoff, W. R.; Kalisak, E.; Trauger, S.; Manchester, M.; Siuzdak, G. *J. Proteome Res.* **2009,** *8*, 3578–3587.
(11) Jia, W.; Li, H.; Zhao, L.; Nicholson, J. K. *Nat. Rev. Drug Discovery* **2008,** *7*, 123–129.
(12) Wikoff, W. R.; Anfora, A. T.; Liu, J.; Schultz, P. G.; Lesley, S. A.; Peters, E. C.; Siuzdak, G. *Proc. Natl. Acad. Sci. U.S.A.* **2009,** *106*, 3698–3703.
(13) Yanes, O.; Clark, J.; Wong, D. M.; Patti, G. J.; Sánchez-Ruiz, A.; Benton, H. P.; Trauger, S. A.; Desponts, C.; Ding, S.; Siuzdak, G. *Nat. Chem. Biol.* **2010,** *6*, 411–417.
(14) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, M. Y.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010,** *45*, 703–714.
(15) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; Fung, C.; Nikolai, L.; Lewis, M.; Coutouly, M. A.; Forsythe, I.; Tang, P.; Shrivastava, S.; Jeroncic, K.; Stothard, P.; Amegbey, G.; Block, D.; Hau, D. D.; Wagner, J.; Miniaci, J.; Clements, M.; Gebremedhin, M.; Guo, N.; Zhang, Y.; Duggan, G. E.; Macinnis, G. D.; Weljie, A. M.; Dowlatabadi, R.; Bamforth, F.; Clive, D.; Greiner, R.; Li, L.; Marrie, T.; Sykes, B. D.; Vogel, H. J.; Querengesser, L. *Nucleic Acids Res.* **2007,** *35*, D521–526.
(16) Smith, C. A.; Maille, G. O.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio,D. E.; Abagyan, R.; Siuzdak, G. METLIN: A Metabolite Mass Spectral Database. *Proceedings of the 9th International Congress of Therapeutic Drug Monitoring and Clinical Toxicology*, Louisville, KY, 2005; pp 747−751.
(17) Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006,** *78*, 6140–6152.
(18) Chu, Y.-C.; Guan, Y.; Skinner, J.; Raja, S. N.; Johns, R. A.; Tao, Y.-X. *Pain* **2005,** *119*, 113–123.
(19) Bölcskei, K.; Petho, G.; Szolcsányi, J. *Methods Mol. Biol.* **2010,** *617*, 57–66.
(20) Kyburz, D.; Corr, M. *Springer Sem. Immunopathol.* **2003,** *25*, 79–90.
(21) Goo, Y. A.; Yi, E. C.; Baliga, N. S.; Tao, W. A.; Pan, M.; Aebersold, R.; Goodlett, D. R.; Hood, L.; Ng, W. V. *Mol. Cell. Proteomics* **2003,** *2*, 506–524.
(22) Kaur, A.; Pan, M.; Meislin, M.; Facciotti, M. T.; El-Gewely, R.; Baliga, N. S. *Genome Res.* **2006,** *16*, 841–854.
(23) Kajihara, Y.; Murakami, M.; Imagawa, T.; Otsuguro, K.; Ito, S.; Ohta, T. *Neuroscience* **2010,** *166*, 292–304.
(24) Yoshida, A.; Mobarakeh, J. I.; Sakurai, E.; Sakurada, S.; Orito, T.; Kuramasu, A.; Kato, M.; Yanai, K. *Eur. J. Pharmacol.* **2005,** *522*, 55–62.
(25) Rosenthal, S. R.; Minard, D. *J. Exp. Med.* **1939,** *70*, 415–425.
(26) Rosenthal, S. R.; Sonnenschein, R. R. *Am. J. Physiol.* **1948,** *155*, 186–190.
(27) Hough, L.; Rice, F. L. *J. Pharmacol. Exp. Ther.* **2011,** *336*, 30–37.
(28) Aarmstrong, D.; Dry, R. M. L.; Keele, C. A.; Markham, J. W. *J. Physiol.* **1953,** *120*, 326–351.
(29) Fjallbrant, N.; Iggo, A. *J. Physiol.* **1961,** *156*, 578–590.
(30) Cantor, R. M.; Lange, K.; Sinsheimer, J. S. *Am. J. Hum. Genet.* **2010,** *86*, 6–22.
(31) Borges, F.; Gomes, G.; Gardner, R.; Moreno, N.; McCormick, S.; Feijó, J. A.; Becker, J. D. *Plant Physiol.* **2008,** *148*, 1168–81.
(32) Güimil, S.; Chang, H.-S.; Zhu, T.; Sesma, A.; Osbourn, A.; Roux, C.; Ioannidis, V.; Oakeley, E. J.; Docquier, M.; Descombes, P.; Briggs, S. P.; Paszkowski, U. *Proc. Natl. Acad. Sci. U.S.A.* **2005,** *102*, 8066–8070.
(33) Higdon, R.; Haynes, W.; Kolker, E. *OMICS* **2010,** *14*, 309–314.