analytical chemistry

Article

pubs.acs.org/ac

XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data

Ralf Tautenhahn,[†] Gary J. Patti,[‡] Duane Rinehart,[†] 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

[‡]Departments of Chemistry, Genetics, and Medicine, Washington University School of Medicine, 660 South Euclid Ave, Saint Louis, Missouri 63110, United States

ABSTRACT: Recently, interest in untargeted metabolomics has become prevalent in the general scientific community among an increasing number of investigators. The majority of these investigators, however, do not have the bioinformatic expertise that has been required to process metabolomic data by using command-line driven software programs. Here we introduce a novel platform to process untargeted metabolomic data that uses an intuitive graphical interface and does not require installation or technical expertise. This platform, called XCMS Online, is a web-based version of the widely used XCMS software that allows users to easily upload and process liquid chromatography/mass spectrometry data with only a few mouse clicks. XCMS Online provides a solution for the complete untargeted metabolomic workflow including feature detection, retention time correction, alignment, annotation, statistical analysis, and data visualization. Results can be browsed online in an interactive, customizable table showing statistics, chromatograms, and putative METLIN identities for each metabolite. Additionally, all results and images can be downloaded as zip files for offline analysis and publication. XCMS Online is available at https://xcmsonline.scripps.edu.

Untargeted metabolomics describes the global profiling of small molecules in a biological system without bias. Although several analytical technologies can be used to perform untargeted metabolomics, often liquid chromatography/mass spectrometry (LC/MS) is the technique of choice given the large number of metabolites that can be simultaneously measured in a single analysis. In a typical analysis, for example, tens of thousands of features can be measured by LC/MS in a metabolite extract (where a feature is defined as an ion with a unique m/z and retention time). Generally, the aim of untargeted metabolomics is to determine which of these features is dysregulated between two or more sample groups. Due to the complexity of the data sets, however, it is impractical to perform the comparison manually. Moreover, the retention time of a compound can vary from run to run as a result of experimental drifts such as column degradation, temperature fluctuations, pH changes of the mobile phase, etc. These nonlinear deviations can complicate interpretation of even a subset of untargeted metabolomic data and reinforce the need of bioinformatic software that performs automated data processing, such as retention time alignment.

Over the past decade, several software programs for automated processing of LC/MS-based metabolomic data have been introduced, including freely available software like MetAlign,[1] MZmine,[2] and XCMS[3] as well as commercial products like Mass Profiler Pro (Agilent) or Metabolic Profiler (Bruker). Although each software has some unique advantages depending on experimental design as reviewed by Castillo et al.,[4] all of the programs are similarly limited by the technical expertise required for installation and operation or by having proprietary restrictions that limit utility and applicability to different instrumentation. Accordingly, the limited accessibility of these programs to nonbioinformatic scientists has greatly hindered the growth of untargeted metabolomics, particularly in biological and clinical laboratories. Many of these same laboratories, however, have access to biological specimens that are well-suited for metabolomic analysis. While interest in untargeted metabolomics has surged in recent years, particularly among biologists and clinicians due to the high throughput and limited sample requirements, an overwhelming number of research endeavors have relied heavily on collaboration with laboratories having bioinformatic expertise in data processing.

In response to the growing interest from the general scientific community for a user-friendly program to process untargeted metabolomic data, we have created a web-based platform called XCMS Online. Unlike the web-based tools MetaboAnalyst[5] and metaP-Server[6] that have been recently introduced to perform statistical analysis of preprocessed data, XCMS Online is a solution for the entire untargeted metabolomic workflow ranging from the computationally expensive raw data processing and retention time correction calculations to statistical analysis and metabolite assignment. XCMS Online is based on software we developed in 2006 called XCMS, which has been widely accepted by the metabolomic community. In brief, XCMS identifies features whose relative intensity varies between sample groups and

Received: March 12, 2012
Accepted: April 25, 2012
Published: April 25, 2012

   dx.doi.org/10.1021/ac300698c | *Anal. Chem.* 2012, 84, 5035−5039

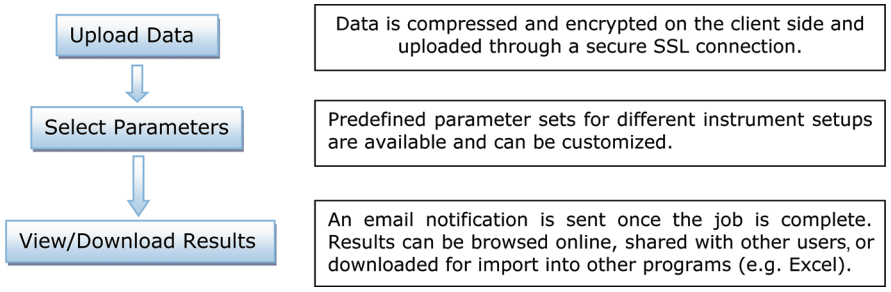

Figure 1. XCMS Online workflow: LC/MS-based metabolomic data is processed in three simple steps.

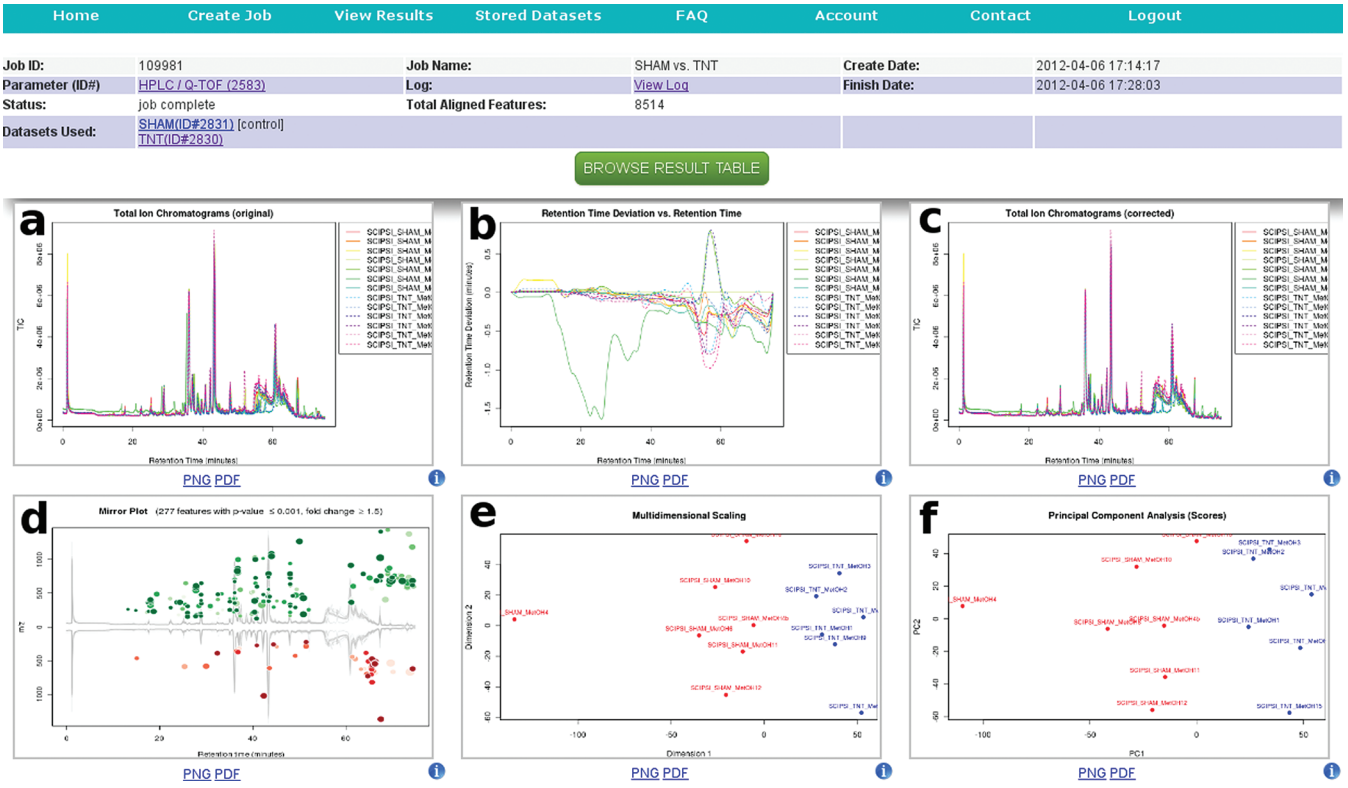

Figure 2. XCMS Online screen shot showing the overview of results from an untargeted analysis. The top display panel includes text on general job information such as data set names, parameter set names, and job date. The lower display panels are used to provide a global visualization of the experimental results (labeled d–f) in addition to plots that can be used for quality control (labeled a–c, e, and f). The display shows the overlay of total ion chromatograms before (a) and after (c) retention time correction, the retention time correction curve (b), mirror plot (d), multidimensional scaling (e), and principal component analysis (f). Color codings: samples are color coded in a–c; the same color mapping is used for the samples across these three panels. In the mirror plot (d) up-regulated features are shown in green and down-regulated features are in red. For the PCA and MDS plot (e and f) colors are used to highlight the different sample classes (here red for control group and blue for disease group).

calculates *p*-values as well as fold changes. The software has been used for a wide range of applications, including bud and fruit development in plants,[7−9] cancer research,[10−14] chronic pain,[15] pathogenesis of hepatitis B,[16] forensics,[17,18] long-term studies of human serum,[19] pregnancy-specific syndromes,[20] and stem cell differentiation.[21] It is not the purpose of this paper to provide a detailed description of the XCMS algorithms, which have already been described in detail elsewhere.[3,22] For more information related to the XCMS algorithms, please see http://metlin.scripps.edu/xcms/. XCMS is distributed as an R package and is operated through a command-line interface or customized scripts, making it flexible and particularly suitable for batch-processing. XCMS Online retains these same functionalities of the original XCMS software but does not require familiarity with a command-line interface or programming scripts, thereby making it accessible to a much greater scientific population.

WORKFLOW

The processing of metabolomic data by XCMS Online is organized in three simple steps: data upload, parameter selection, and result interpretation (Figure 1). The samples for each group that are to be compared are uploaded through a specific Java applet, which allows users to simply drag and drop their files into the upload area of the program. The currently accepted file formats are netCDF, mzXML, mzData, and Agilent .d folders. In future versions, additional vendor-specific formats will be supported. All files are automatically compressed and encrypted prior to being uploaded through a secure SSL connection to the XCMS Online server. Although

dx.doi.org/10.1021/ac300698c | *Anal. Chem.* 2012, 84, 5035−5039

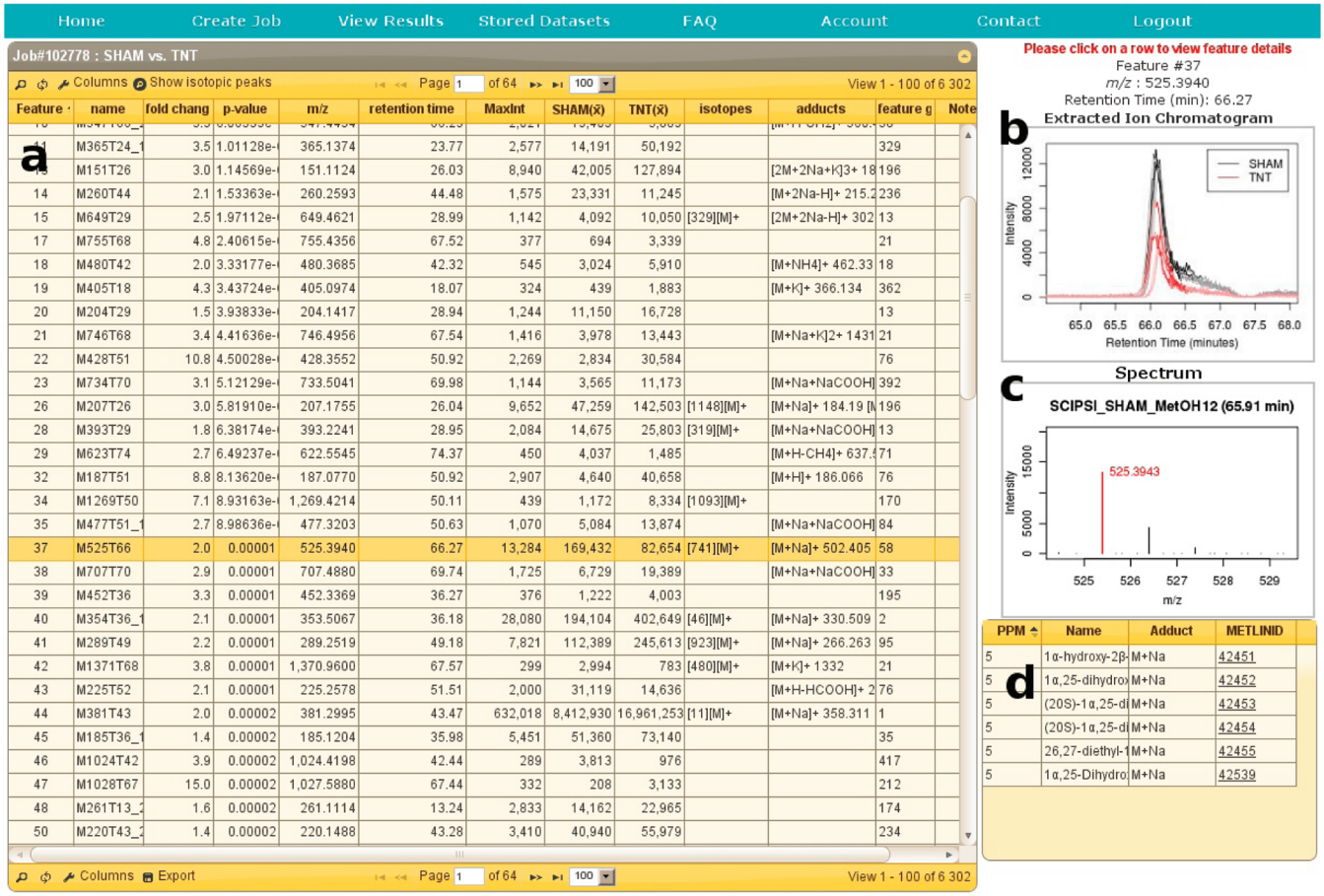

Figure 3. XCMS Online screen shot showing the feature browser that is used to display detailed information for each individual feature including statistics (a), extracted ion chromatograms (b), spectrum details (c), and putative METLIN identifications (d).

file upload can be time demanding (depending on file size, ranging from seconds up to several hours), it is not necessary for the user to wait for the upload to finish; the job can be submitted even before the upload is complete. File upload will continue in the background and the data processing will start automatically as soon as all files are successfully uploaded.

During file upload, users will be asked to select a parameter set that matches the instrument setup used to analyze the samples. Predefined parameter sets for different instrument setups are available (e.g., HPLC/Q-TOF, UPLC/Q-TOF, HPLC/Orbitrap, HPLC/single quad MS, GC/single quad MS). Customization of the parameter set is also possible, for example, to change the signal/noise threshold, to adjust the mass tolerance for the identification step, or to change the methods that are used for feature detection,[22] retention time correction,[23] alignment,[3] and annotation.[24] After the parameter set has been selected, the job will be submitted to the system. Depending on the data set size, data processing can take from several minutes up to several hours. Multiple jobs can be submitted simultaneously. The system also manages all data sets that have been uploaded previously by the user. These data sets can be utilized for additional comparisons, modified, or deleted.

RESULT INTERPRETATION

The user will receive an e-mail notification after a job has been completed and is ready for browsing or downloading. After selecting a finished job from the job list, XCMS Online displays several figures that provide an overview of the experimental results and also serve as a quality control mechanism (Figure 2). XCMS Online uses nonlinear methods[3,23] to compensate for retention time drifts between samples. As a visualization and for quality control of this correction procedure, an overlay of all total ion chromatograms (TICs) acquired is shown before (Figure 2a) and after retention time correction (Figure 2c), in addition to the retention time correction curves (Figure 2b). After retention time correction, all TICs should be in alignment. Potentially problematic samples with extreme deviations can be recognized from the correction curve plot and may be removed from the data set.

A plot showing dysregulated features, a so-called "mirror plot" (Figure 2d), is used to represent ions whose intensities are altered between sample groups according to statistical thresholds set by the user (e.g., p-value $\leq$0.001, fold change $\geq$2). Features that are down-regulated are represented as circles on the bottom of the plot, whereas features that are up-regulated are represented as circles on the top. The size of each circle corresponds to the (log) fold change of the feature (i.e., the average difference in relative intensity of the peak between sample groups). Larger circles correspond to peaks with greater fold changes. Up-regulated features are displayed in green, and down-regulated features are in red. The shade of color is used to represent p-value, with brighter circles having lower p-values. The retention time corrected TICs are also overlaid in gray in the background of the figure. The circles representing features with hits in the METLIN database are shown with a black

dx.doi.org/10.1021/ac300698c | Anal. Chem. 2012, 84, 5035−5039

outline. Additionally, an interactive version of the plot is available where feature statistics and putative identities are displayed in a pop-up window when users scroll their mouse over the circles in the plot.

Two additional plots for visualizing high-dimensional data sets are also included as part of the XCMS Online standard output: a multidimensional scaling plot (MDS, Figure 2e) and a principal component analysis plot (PCA, Figure 2f). PCA and MDS are performed on the centered and scaled data. Additional plots showing R2 and Q2 values for PCA model validation are available, as well as loadings plots.[25] PCA and MDS plots are based on the intensities of all aligned features and can depict similarities between the samples or help in identifying potential outliers. From the overview page, the complete results, including the aligned feature table and all graphics, can be downloaded as a zip file.

The feature table can be viewed online by clicking the "Browse Result Table" button. The feature table browser (Figure 3) displays detailed information for each individual feature including statistics, an extracted ion chromatogram, spectrum details, and putative METLIN assignments. An explanation of the individual columns that are shown (like m/z value, retention time, and fold change) is given in Table 1.

Table 1. Names and Descriptions of the Columns in the Result Browser of XCMS Online

column name	explanation
name	feature name (arbitrary), formed by nominal mass and retention time, e.g., M120T7
fold change	fold change (ratio of the mean intensities)
p-value	p-value (Welch t test, unequal variances)
q-value	q-value estimation for false discovery rate control[a]
m/z	m/z value (median value for the aligned features)
retention time	retention time (median value for the aligned features)
MaxInt	highest absolute intensity of this feature across all aligned samples; useful to decide if feature can be selected for tandem MS
[Sample class] ($\bar{x}$)	average feature intensity within sample class (actual sample class name is shown).
[Sample class] (sd)	standard deviation of the feature intensity within sample class (actual sample class name is shown).
isotopes	information about isotopic peaks (monoisotopic peaks are shown as, e.g., $[M]^+$ for singly charged metabolites, $[M + 1]^+$ for the first isotopic peak, etc.)
adducts	annotation of adducts and common fragments/neutral losses[b]
feature group	features that are likely related to the same compound have the same feature group number; grouping is based on retention time and peak intensity correlations[b]
METLIN MS/MS	shows if tandem MS data is available for this metabolite in METLIN
Notes	users can add notes in this field, e.g., to mark candidates for tandem MS

[a]Ref 26. [b]Ref 24.

The feature table (Figure 3a) can be browsed by clicking on a row or by using the cursor keys to flip through the features. In either case, the graphics on the right side (Figure3b–d) are updated automatically to show the data according to the feature that is selected in the table. Specifically, the top display panel (Figure 3b) shows an overlay of the extracted ion chromatograms from each sample for that particular feature. Users can click the image to see an enlarged version of higher resolution. In the middle display panel (Figure 3c), a detailed mass spectral view of the selected feature is displayed. The sample where the particular feature has the highest intensity is used to extract the data and to show the exact m/z value in the spectrum. Both the extracted ion chromatogram as well as the mass spectral view display panels can be used to quickly assess the quality of each feature and to identify possible problems during feature detection and alignment. The bottom display panel (Figure 3d) shows putative identifications based on a METLIN search of the accurate mass displayed, ordered by mass difference (ppm). Multiple adducts (adjustable) are used for the database search, including the adduct type that was automatically detected by the annotation algorithm.[24] All results are directly linked to the METLIN metabolite database such that metabolite details and MS/MS data (if available) are displayed when a database hit is selected.

To facilitate the interpretation and to prioritize features, multiple criteria are available to filter the results. Filtering can be based on p-value, fold change, intensity, m/z value, retention time ranges, etc. If desired, isotopic peaks can be removed. The filtered and annotated table can also be saved (TSV format) for import into Microsoft Excel or other programs.

CONCLUSIONS

In summary, the technical expertise that has historically been required to process LC/MS-based untargeted metabolomic data has limited the growth of the field and made the technology largely inaccessible to a substantial population of biological scientists. To facilitate the processing and analyzing of untargeted metabolomic data, we have created a web-based platform called XCMS Online. XCMS Online does not require the installation of any programs, is designed to be user-friendly, and accepts data files generated from most mass spectrometers currently used to perform untargeted metabolomics. Although the platform has adjustable settings for advanced users, it is designed to provide a complete metabolomic solution for investigators without expertise in the field. The processing output provided by XCMS Online can be downloaded and used directly in research proposals and publications.

AUTHOR INFORMATION

Notes

The authors declare no competing financial interest.

REFERENCES

(1) Lommen, A. *Anal. Chem.* **2009**, *81* (8), 3079–3086.
(2) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.
(3) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78* (3), 779–787.
(4) Castillo, S.; Gopalacharyulu, P.; Yetukuri, L.; Oresic, M. *Chemom. Intell. Lab. Syst.* **2011**, *108* (1), 23–32.
(5) Xia, J.; Psychogios, N.; Young, N.; Wishart, D. S. *Nucleic Acids Res.* **2009**, *37* (Web Server issue), W652–W660.
(6) Kastenmuller, G.; Romisch-Margl, W.; Wagele, B.; Altmaier, E.; Suhre, K. *J. Biomed. Biotechnol.* **2011**, *2011*, 839862.
(7) Ruttink, T.; Arend, M.; Morreel, K.; Storme, V.; Rombauts, S.; Fromm, J.; Bhalerao, R. P.; Boerjan, W.; Rohde, A. *Plant Cell* **2007**, *19* (8), 2370–2390.
(8) Mounet, F.; Moing, A.; Garcia, V.; Petit, J.; Maucourt, M.; Deborde, C.; Bernillon, S.; Le Gall, G.; Colquhoun, I.; Defernez, M.; Giraudel, J. L.; Rolin, D.; Rothan, C.; Lemaire-Chamley, M. *Plant Physiol.* **2009**, *149* (3), 1505–1528.

(9) Griesser, M.; Hoffmann, T.; Bellido, M. L.; Rosati, C.; Fink, B.; Kurtzer, R.; Aharoni, A.; Munoz-Blanco, J.; Schwab, W. *Plant Physiol.* **2008**, *146* (4), 1528−1539.
(10) Chen, J.; Wang, W. Z.; Lv, S.; Yin, P. Y.; Zhao, X. J.; Lu, X.; Zhang, F. X.; Xu, G. W. *Anal. Chim. Acta* **2009**, *650* (1), 3−9.
(11) Chiang, K. P.; Niessen, S.; Saghatelian, A.; Cravatt, B. F. *Chem. Biol.* **2006**, *13* (10), 1041−1050.
(12) Kind, T.; Tolstikov, V.; Fiehn, O.; Weiss, R. H. *Anal. Biochem.* **2007**, *363* (2), 185−195.
(13) Nomura, D. K.; Long, J. Z.; Niessen, S.; Hoover, H. S.; Ng, S. W.; Cravatt, B. F. *Cell* **2010**, *140* (1), 49−61.
(14) Qiu, Y. P.; Cai, G. X.; Su, M. M.; Chen, T. L.; Liu, Y. M.; Xu, Y.; Ni, Y.; Zhao, A. H.; Cai, S. J.; Xu, L. X.; Jia, W. *J. Proteome Res.* **2010**, *9* (3), 1627−1634.
(15) Patti, G. J.; Yanes, O.; Shriver, L. P.; Courade, J. P.; Tautenhahn, R.; Manchester, M.; Siuzdak, G. *Nat. Chem. Biol.* **2012**, *8* (3), 232−234.
(16) Yang, F.; Yan, S. K.; He, Y.; Wang, F.; Song, S. X.; Guo, Y. J.; Zhou, Q.; Wang, Y.; Lin, Z. Y.; Yang, Y.; Zhang, W. D.; Sun, S. H. *J. Hepatol.* **2008**, *48* (1), 12−19.
(17) Fraga, C. G.; Perez Acosta, G. A.; Crenshaw, M. D.; Wallace, K.; Mong, G. M.; Colburn, H. A. *Anal. Chem.* **2011**, *83* (24), 9564−9572.
(18) Tan, G.; Lou, Z.; Jing, J.; Li, W.; Zhu, Z.; Zhao, L.; Zhang, G.; Chai, Y. *Biomed. Chromatogr.: BMC* **2011**, *25* (12), 1343−1351.
(19) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B.; Consortium, H. *Anal. Chem.* **2009**, *81* (4), 1357−1364.
(20) Kenny, L. C.; Broadhurst, D. I.; Dunn, W.; Brown, M.; North, R. A.; McCowan, L.; Roberts, C.; Cooper, G. J. S.; Kell, D. B.; Baker, P. N.; Endpoints, S. P. *Hypertension* **2010**, *56* (4), 741−749.
(21) Yanes, O.; Clark, J.; Wong, D. M.; Patti, G. J.; Sanchez-Ruiz, A.; Benton, H. P.; Trauger, S. A.; Desponts, C.; Ding, S.; Siuzdak, G. *Nat. Chem. Biol.* **2010**, *6* (6), 411−417.
(22) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinf.* **2008**, *9*, 504.
(23) Prince, J. T.; Marcotte, E. M. *Anal. Chem.* **2006**, *78* (17), 6140−6152.
(24) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.
(25) Stacklies, W.; Redestig, H.; Scholz, M.; Walther, D.; Selbig, J. *Bioinformatics* **2007**, *23* (9), 1164−1167.
(26) Storey, J. D. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2002**, *64* (3), 479−498.

dx.doi.org/10.1021/ac300698c | *Anal. Chem.* 2012, 84, 5035−5039