Anal. Chem. **2006,** *78,* 779−787

XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification

Colin A. Smith, Elizabeth J. Want, Grace O'Maille, Ruben Abagyan, and Gary Siuzdak*

The Scripps Center for Mass Spectrometry and Department of Molecular Biology, The Scripps Research Institute, 10550 North Torrey Pines Road, BCC-007, La Jolla, California 92037

Metabolite profiling in biomarker discovery, enzyme substrate assignment, drug activity/specificity determination, and basic metabolic research requires new data preprocessing approaches to correlate specific metabolites to their biological origin. Here we introduce an LC/MS-based data analysis approach, XCMS, which incorporates novel nonlinear retention time alignment, matched filtration, peak detection, and peak matching. Without using internal standards, the method dynamically identifies hundreds of endogenous metabolites for use as standards, calculating a nonlinear retention time correction profile for each sample. Following retention time correction, the relative metabolite ion intensities are directly compared to identify changes in specific endogenous metabolites, such as potential biomarkers. The software is demonstrated using data sets from a previously reported enzyme knockout study and a large-scale study of plasma samples. XCMS is freely available under an open-source license at http://metlin.scripps.edu/download/.

Recent advances in analytical technology have enabled the high-throughput analysis of many of nature's biological building blocks. DNA microarrays can measure the transcription of the entire human genome using a single chip.[1] Liquid chromatography coupled to tandem mass spectrometry (LC−MS/MS) can be used to identify thousands of proteins from a complex mixture.[2] More recently, metabolite profiling has gained popularity using a number of techniques including nuclear magnetic resonance (NMR) or different combinations of liquid chromatography (LC), gas chromatography (GC), and mass spectrometry (MS).[3−5] One particularly popular platform for untargeted metabolite profiling is LC/MS using electrospray ionization (LC/ESI-MS). Unlike NMR, LC/ESI-MS resolves individual chemical components into separate peaks, where NMR provides only a chemical fingerprint. Unlike GC/MS, it additionally detects nonvolatile compounds, which make up a large proportion of metabolites. Finally, LC separation provides a means for resolving isobaric compounds and reducing signal suppression.[3,6]

The simultaneous separation and detection of metabolite analytes using both LC and MS produces complex data sets that require significant preprocessing before multiple samples can be analyzed statistically. In preprocessing spectral and separation data, two general strategies can be taken: (1) Divide the signal into bins, incorporating all data into a recognition profile for each sample. (2) Identify and individually quantify significant features, discarding data not deemed to be part of a feature. Variations and combinations of those strategies are possible, but here we will attempt to emphasize their differences. Depending on the expected variation in the data and the experimental question being asked, one technique may be more useful than the other. If discriminating features is nontrivial or if features are difficult to resolve from one another, then binning may be the most effective strategy. If correlated shifts in signal distribution are expected, either preprocessing method may be equally useful. Such variations are often analyzed with multivariate techniques such as principal components analysis.

However, if the variation between samples is largely random except for a small number of features, then an unsupervised multivariate technique may not be adequate to identify the significant differences. In those cases, applying a univariate statistical analysis procedure or supervised multivariate technique to each individual bin or feature may be the most effective method for identifying these differences. Variance or bias not attributable to true differences in analyte abundance generally reduces the power of such procedures. Thus, an important goal of any preprocessing method should be to reduce such variance. In that respect, binning has a number of drawbacks. If breaks between bins are chosen arbitrarily, the signal from a given analyte may get split between two adjacent bins. Additionally, multiple independent analytes may contribute to the signal of a single bin, thus decreasing the ability to discern significant differences in individual analytes. Feature detection helps avoid those problems by trying to ensure that each segment of the characterized signal corresponds to and entirely captures the signal from a single feature.

A preprocessing routine based on peak detection requires a robust method for reproducibly characterizing peaks in the three-

* To whom correspondence should be addressed. E-mail: siuzdak@scripps.edu.

(1) Kronick, M. N. *Expert Rev. Proteomics* **2004**, *1*, 19−28.
(2) Linscheid, M. W. *Anal. Bioanal. Chem.* **2005**, *381*, 64−66.
(3) Want, E. J.; Cravatt, B. F.; Siuzdak, G. *ChemBioChem* **2005**, *6*, 1941−1951.
(4) Saghatelian, A.; Trauger, S. A.; Want, E. J.; Hawkins, E. G.; Siuzdak, G.; Cravatt, B. F. *Biochemistry* **2004**, *43*, 14332−14339.
(5) Fiehn, O.; Kopka, J.; Dörmann, P.; Altmann, T.; Trethewey, R. N.; Willmitzer, L. *Nat. Biotechnol.* **2000**, *18*, 1157−1161.
(6) Siuzdak, G. *The Expanding Role of Mass Spectrometry in Biotechnology*; MCC Press: San Diego, 2003.

10.1021/ac051437y CCC: $33.50
Published on Web 01/07/2006

dimensional space (time, mass, intensity) defined by the LC/MS data. Such an algorithm should be able to detect peaks with a very low signal-to-noise (s/n) ratio while simultaneously filtering out random noise. An important property of LC/MS data is that it is highly anisotropic. That is, the number of data points sampling a peak in the chromatographic time domain is generally much greater than in the mass domain. Because of the greater amount of data per feature, processing data in the chromatographic domain is more likely to help discriminate between analyte peaks and noise peaks. Additionally, because the number of isobaric peaks is generally small, the complexity of the chromatographic domain is typically much lower.

Filtering noise in the chromatographic domain data requires some type of signal processing. One recently proposed method[7] was median filtering in a specified window size. A perhaps more optimal signal processing technique for increasing the s/n ratio is matched filtration. Matched filtration is based on applying a filter whose coefficients are equal to the expected shape of the signal. In chromatographic data, the Gaussian function is a useful simplified description of the peaks shape. A matched filter for a chromatographic peak would use coefficients equal to a Gaussian function whose width was the same as the expected peak width. The net effect of such a filter is the reduction of peaks whose widths are significantly less than the model peak shape. The application of matched filtration and the effect of model peak width on LC/MS data were recently reported.[8] Additionally, matched filtration was recently extended by using the noise characteristics in areas without signal to improve the filter in an algorithm known as MEND.[9]

Another important task in preprocessing data for comparative analysis is matching peaks representing the same analyte from different samples. This has been an area of recent development in the analysis of one-channel chromatographic data, such as gas chromatography with infrared detection. One recently developed method[10] accomplishes peak matching by first generating an average chromatogram whose peaks are used to define group centers. Individual sample peaks are then matched to the master list of group centers. As noted by the authors, their algorithm requires that the deviation in retention time from sample to sample be no greater than the time between two adjacent peaks. That observation is true of all current peak-matching algorithms as well as the one presented here. Another group developed a method[11] that combined and sorted peak lists from all samples, sequentially creating peak groups using a fixed retention time window. They use a number of techniques to resolve collisions in which two peaks from the same sample are assigned to the same group. A disadvantage of their method is that it is dependent on having an optimized retention time window, which may require a good deal of manual optimization. Importantly, these two methods were designed for data separated in only one dimension.

One way to overcome the limitations imposed by drifts in retention time from sample to sample is to use an alignment algorithm. One often-used technique involves spiking a small number of internal standards into all samples prior to data acquisition.[12] During data preprocessing, the peaks corresponding to those standards are identified and then used to linearly shift the retention time of each acquired sample. There are a number of major disadvantages to that method. One, it assumes that deviations in retention time are linear, which we will demonstrate is not the case. Two, it requires an additional step of sample preparation and the addition of chemicals that may mask the presence of other experimentally relevant analytes. To avoid those problems, a number of methods have been developed to align chromatographic profiles without internal standards and allowing nonlinearities. Correlation optimized warping[13] (COW) comprehensively searches possible sets of segmented warpings that can be used to align one chromatogram onto another, identifying the best set of warpings using a correlation metric. That pairwise method was recently extended[14] to use the mass spectral component of LC/MS data to further improve the alignment. Noting speed limitations of the comprehensive search performed by COW, another group developed a pairwise method[10] that used individual peak matching to calculate nonlinear deviations from a reference chromatogram. A similar strategy[15] was employed to align HPLC/UV chromatograms. As noted earlier, their peak-matching and alignment algorithm is dependent on the drift in retention time being less than the distance between adjacent peaks. As we will show, that limitation proves insignificant when the profile is also resolved by mass. Another method has been developed[16] that models the warping function itself, iteratively improving the coefficients of a quadratic warping function to minimize the difference between two chromatographic traces.

It is important to acknowledge other software specifically designed for metabolite profiling. MarkerLynx (Waters) incorporates peak detection and data set alignment using predefined internal standards. However, it is designed only to work with Waters MicroMass mass spectrometers, making import and analysis of data from other instruments difficult. XCMS, on the other hand, allows for data processing of various instrument origins including the Waters Q-Tof Micro, Finnigan LTQ, and Agilent 1100 LC/MSD. metAlign (Plant Research International B.V.) also includes peak selection and the option of nonlinear, iterative alignment, similar to XCMS but using a different algorithm. metAlign is not based on any published algorithms beyond the user manual. In addition, its code is proprietary and not open to outside modification or inspection.

Here we describe our approach to preprocessing LC/MS data for global, untargeted metabolite profiling. An implementation of the methodology illustrated here is freely available in an open-source package called XCMS (an acronym for various forms (X) of chromatography mass spectrometry). We demonstrate the

(7) Hastings, C. A.; Norton, S. M.; Roy, S. *Rapid Commun. Mass Spectrom.* **2002**, *16*, 462−467.
(8) Danielsson, R.; Bylund, D.; Markides, K. E. *Anal. Chim. Acta* **2002**, *454*, 167−184.
(9) Andreev, V. P.; Rejtar, T.; Chen, H. S.; Moskovets, E. V.; Ivanov, A. R.; Karger, B. L. *Anal. Chem.* **2003**, *75*, 6314−6326.
(10) Johnson, K. J.; Wright, B. W.; Jarman, K. H.; Synovec, R. E. *J. Chromatogr., A* **2003**, *996*, 141−155.
(11) Duran, A. L.; Yang, J.; Wang, L.; Sumner, L. W. *Bioinformatics* **2003**, *19*, 2283−2293.
(12) Frenzel, T.; Miller, A.; Engel, K.-H. *Eur. Food Res. Technol.* **2003**, *216*, 335−342.
(13) Nielsen, N.-P. V.; Carstensen, J. M.; Smedsgaard, J. *J. Chromatogr., A* **1998**, *805*, 17−35.
(14) Bylund, D.; Danielsson, R.; Malmquist, G.; Markides, K. E. *J. Chromatogr., A* **2002**, *961*, 237−244.
(15) Yang, J.; Xu, G.; Zheng, Y.; Kong, H.; Wang, C.; Zhao, X.; Pang, T. *J. Chromatogr., A* **2005**, *1084*, 214−221.
(16) Eilers, P. H. *Anal. Chem.* **2004**, *76*, 404−411.

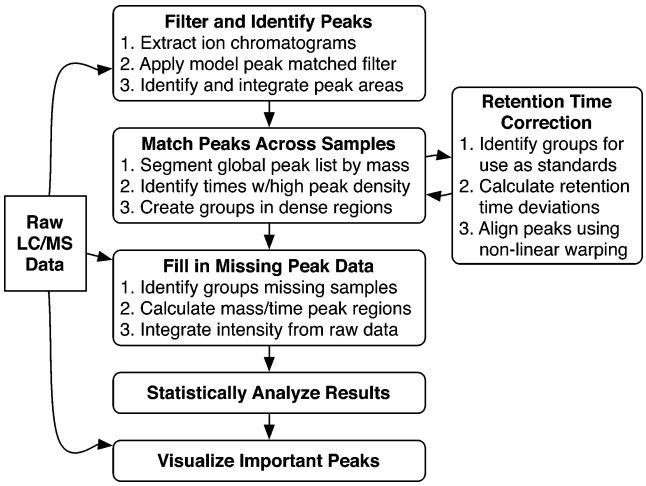

Figure 1. Flowchart showing the general strategy for preprocessing and analysis of LC/MS data for global untargeted analysis of metabolites and other analytes. Only one raw data file is read at a time, allowing the analysis of very large data sets. Multiple iterations of supervised retention time correction are possible, as necessary. The statistical analysis and visualization steps, not being the focus of this paper, are shown in less detail.

software's ability to align many chromatographic traces, handle large data sets, and discover novel differences between two sample groups.

THEORY

The preprocessing strategy described here builds on and combines previous ideas in chemometric analysis. In addition, it includes novel techniques for matching peaks across samples and aligning the retention times of many samples in a single step. A general overview of the preprocessing and analysis strategy is shown in Figure 1. Here, we will describe the first three steps in detail, peak detection, peak matching, and retention time alignment. The latter steps will be described in Results and Discussion. Importantly, the design of the method is modular and allows substitution or addition of supplementary processing steps. We have developed an implementation that enables others to easily extend and alter the methodology described here.

Peak Detection. The peak detection algorithm used for this work is based on cutting the LC/MS data into slices a fraction of a mass unit (0.1 m/z) wide and then operating on those individual slices in the chromatographic time domain. Within each slice, the signal is determined by taking the maximum intensity at each time point in the slice. We term this representation the extracted ion base-peak chromatogram (EIBPC). Representative examples of slices are shown in Figure 2. Before peak detection, each slice is filtered with matched filtration using a second-derivative Gaussian as the model peak shape. The standard deviation of the Gaussian was 13 s, equivalent to a full width at half-maximum (fwhm) of 30 s. As observed by Danielsson et al.,[8] a second-derivative Gaussian model peak generates a new chromatographic profile reflecting curvature rather than absolute intensity, accomplishing implicit background subtraction. They also showed that the second-derivative matched filter yielded consistent s/n improvement even if the model peak width was 1.5−4 times the signal peak width, indicating that performance of the matched filter is not overly sensitive to variations in peak width.

After filtration, peaks are selected using a signal-to-noise ratio cutoff. Because of the second-derivative transformation and the resulting negative component in the signal, determining a noise

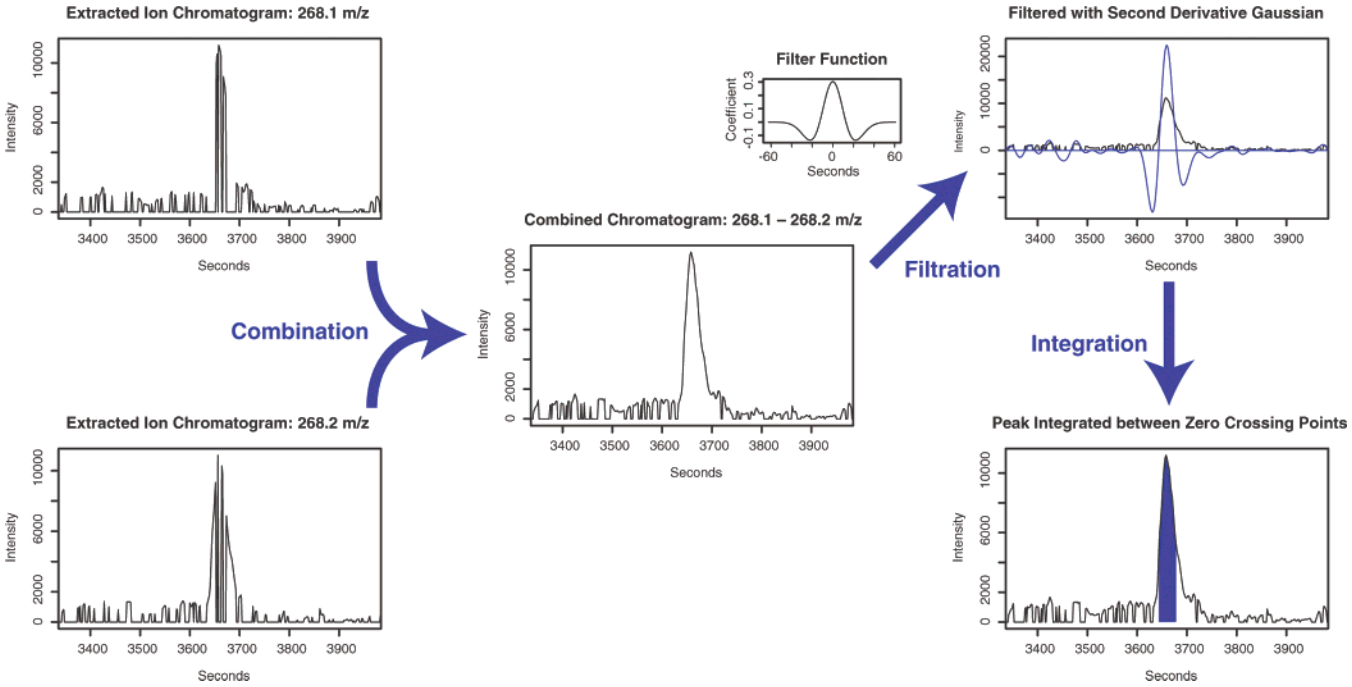

Figure 2. Illustration of the peak detection method using a single peak from the centroided FAAH knockout data. The peak shown here is contained in two adjacent chromatographic slices at 268.1 and 268.2 m/z. The algorithm creates overlapping combined chromatograms (i.e., m/z 268.0/268.1, 268.1/268.2, 268.2/268.3, etc.) with only the m/z 268.1/268.2 chromatogram showing a clean signal. The data are then processed with a matched filter whose coefficients are equal to a second-derivative Gaussian function. The second-derivative transformation causes the filtered chromatogram to cross the x-axis roughly at the peak inflection points. Those zero-crossing points define the area of peak integration. The peak tails, which typically contain only a small fraction of the signal and may overlap with other peaks, are not integrated.

value from the filtered data can be problematic. Through iterative visual inspection of the detected peaks, we determined that taking the mean of the unfiltered data was the most effective method for determining the noise. We also found that a s/n cutoff of 10 was the most effective. It is important to note that the actual s/n ratio of the peak in the unfiltered data could be much lower depending on the quality of the peak shape. Similar to MEND,[9] an additional limit of five peaks per slice was imposed.

As noted by Danielsson et al., the zero-crossing points of the filtered data provide a useful mechanism for determining peak width.[8] Peaks are characterized by integrating the unfiltered chromatogram between the zero-crossing points of the filtered chromatogram. It is worthy of note that no background subtraction is applied to the unfiltered spectrum prior to integration. Thus, any background intensity is included in the feature intensity. In our studies, background has generally been an insignificant fraction of peak area or remains largely constant from run to run, making its impact minimal. In the EIBPCs from LC/MS data, background subtraction may add more noise than it eliminates, although that question requires further study.

An important detail is the relationship between spectral peak width and slice width. If the peak width is larger than the slice width, then the signal from a single peak may bleed across multiple slices. Low-resolution mass spectrometers, such as those employing single-quadrupole mass detection, often produce peak widths greater than the default 0.1 m/z slice used by XCMS. The MEND peak detection algorithm uses a scoring function to assess whether a chromatographic peak is also at the maximum of a spectral peak, preemptively removing such bleed.[9] Instead of eliminating spurious extra peaks during detection, our algorithm uses a postprocessing step that descends through the peak list by intensity, eliminating any peaks in the vicinity (0.7 m/z) of higher intensity peaks.

Another possibility is that the peak width is significantly smaller than the slice width. High-resolution time-of-flight or Fourier transform mass spectrometers often exhibit such behavior. Another extreme example would be centroided mass spectral data, in which the centroid transformation causes peak width to be infinitely thin. In that case, depending on the scan-to-scan precision of the instrument, the signal from an analyte may oscillate between adjacent slices over chromatographic time, making an otherwise smooth peak shape appear jagged. (Figure 2) Based on operator knowledge of the mass spectrometer characteristics, we optionally combine the maximum signal intensity from adjacent slices into overlapping EIBPCs (i.e., 100.0/100.1, 100.1/100.2, etc.), That initial step produces both smooth and jagged chromatographic profiles, which are then used for filtration and peak detection (Figure 2). During the vicinity elimination postprocessing step, peaks detected from smooth profiles (integrated from the full signal) are selected over peaks detected from jagged profiles (integrated from an incomplete signal).

The generation of mass slices necessarily reduces precision. To overcome that loss, the mass of each peak is computed from the original, high-resolution spectra. First, the full-resolution mass is determined in each spectrum containing the peak. Second, the overall peak mass is calculated as a weighted mean of all the full-resolution masses, using intensities as the weights. The peak detection algorithm thus handles low-resolution, high-resolution, and centroided data in a flexible and robust manner.

Peak Matching. After identifying peaks in individual samples, those peaks must then be matched across samples to allow calculation of retention time deviations and relative ion intensity comparison. We developed a peak-matching algorithm that takes into account the two-dimensional, anisotropic nature of LC/MS data. Because the accuracy of mass spectrometers is often more understood and relatively better than corresponding retention time drifts, we make use of fixed-interval bins 0.25 m/z wide to match peaks in the mass domain. An example of the peaks contained in a bin is shown in Figure 3. To avoid splitting a group apart because of arbitrary bin borders, we use overlapping bins in which adjacent bins overlap by half (i.e., 100.0−100.25, 100.125−100.375, etc.). During binning, each peak is counted twice in two overlapping bins. Similar to peak picking, a postprocessing step is used to remove peak groups originating from overlapping bins.

After initially binning peaks by mass, we then resolve groups of peaks with different retention time in each bin. One could apply fixed interval matching as previously described;[10] however, that requires prior knowledge about the deviations in retention time, which may not be known at the outset of preprocessing. Instead, our algorithm calculates the overall distribution of peaks in chromatographic time and dynamically identifies boundaries of regions where many peaks have similar retention times. A robust method of calculating distributions is the kernel density estimator,[17] which can be thought of as a histogram smoothed by another function, in this case the Gaussian. We calculate the distribution using a fast Fourier transform implementation[18] of kernel density estimation. From that distribution, our method then identifies so-called "meta-peaks" which represent many peaks with similar retention times (Figure 3). Starting with the highest peak in the distribution, we descend down either side of the meta-peak until the distribution increases again. That process defines a fixed interval in which all peaks are placed into a group. That procedure is repeated for all meta-peaks in the distribution. Changing the width of the Gaussian smoothing function modulates the inclusiveness of the matching, although we have found that correct matching is not highly sensitive to the smoother width.

To prune out insignificant groups of peaks, the algorithm eliminates each group that contains peaks from fewer than half the samples. If samples are known to come from different conditions and can be divided into sets, such as wild type or knockout, then the method of elimination is slightly modified. In that case, groups are eliminated in which none of the sets has at least half its samples represented. To resolve conditions in which a sample has more than one peak in a group, the algorithm employs a number of different tie-breaking criteria depending on the application. For ion intensity comparison, the peak closest to the median retention time is used. For retention time alignment, described below, the peak with the highest intensity is used.

Retention Time Alignment. Unlike the previously cited retention time alignment algorithms, our method simultaneously corrects the retention times of all samples in a single step. The ability to do so depends on initially having a coarse matching of peaks into reasonable groups. That initial matching is possible because of the mass separation of the chromatographically

(17) Rosenblatt, M. *Ann. Math. Statistics* **1956**, *27*, 832−837.
(18) Silverman, B. W. *Appl. Statistics* **1982**, *31*, 93−99.

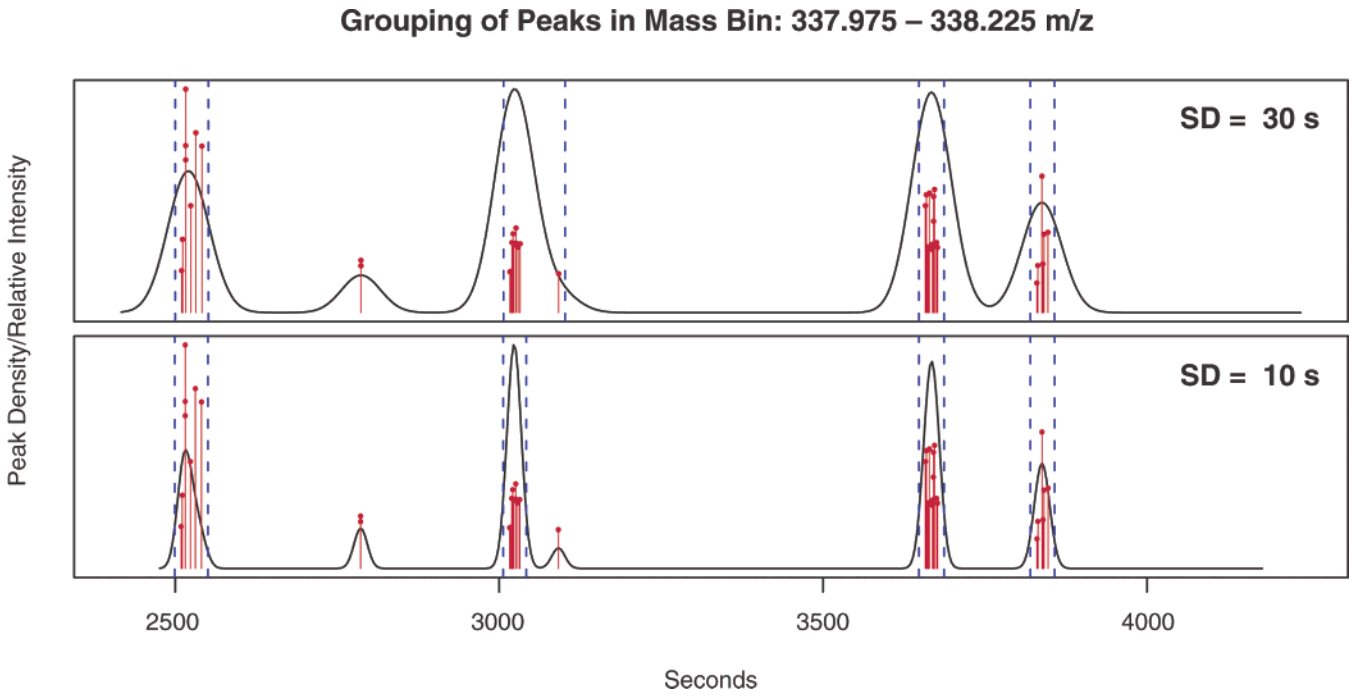

Figure 3. Example of cross-sample peak matching from the FAAH data involving 12 samples. Individual peaks are shown as sticks indicating relative intensity. The smoothed peak density profile is shown as a solid continuous line. Identified groups are flanked by dashed lines. The peak density profiles were smoothed with Gaussian functions of SD 30 and 10 s, respectively. Note how decreased smoothing eliminates a peak from the second group.

separated sample and the relatively low complexity of each individual EIBPC. From the initial grouping, the algorithm typically identifies hundreds of "well-behaved" peak groups in which very few samples have no peaks assigned and very few samples have more than one peak assigned. Such well-behaved groups have a high probability of being properly matched and can be used as temporary standards. For every group, the algorithm calculates the median retention time and the deviation from median for every sample in that group. Because the well-behaved peak groups are generally distributed evenly over the significant portions of the chromatographic profile, a detailed, nonlinear retention time deviation contour can be built for each sample.

After separation, it is possible and often the case that two peaks with different mass but similar retention times may show slightly different retention time deviations. Additionally, it is possible to have parts of the chromatogram where no "well-behaved" peak groups are present. To deal with that uncertainty, our method uses a mathematical function to approximate the differences between deviations and interpolate in sections where no peak groups are present. One recently developed alignment method[16] modeled the deviations using a quadratic function. However, the typical drifts we have observed are often not well approximated by a quadratic function. While one could use higher-order polynomials to fit the deviation data, we have elected to use a local regression fitting method, loess,[19] which uses segmented low-order polynomials to fit the data. Because of the segmented fitting, local perturbations in retention time can be accounted for and thus corrected. The loess fitting method provides the option of automatically removing residual outliers from the data, providing robustness against peaks that were initially mismatched.

(19) Cleveland, W. S.; Grosse, E.; Shyu, W. M. In *Statistical Models in S*; Chambers, J. M., Hastie, T. J., Eds.; Chapman & Hall/CRC: Pacific Grove, CA, 1991.

At the beginning and end of the LC/MS runs, where there are no well-behaved peak groups available for fitting, the deviation function is flattened to a constant value. The resulting deviation profiles are then used to correct the retention times of the original peak lists, after which they are matched into groups once more. The matching/alignment procedure can be repeated in an iterative fashion, successively discerning more and more well-behaved peak groups for increasingly precise alignment. An example of retention time deviation profiles determined from several iterative steps of retention time alignment can be seen in Figure 4.

There are a number of advantages to this single-step alignment procedure. It works quickly and does not require the selection or generation of a standard target sample for alignment. Additionally, straightforward visualization of the retention time deviation profiles and the data from which they were fitted allows the method to be manually supervised if desired. An additional strength of the alignment algorithm is that it works purely with peak data. Although this has been seen as a weakness during the development of other alignment algorithms, the preprocessing strategy described here has a specific interest in that peak data. Because peak identification is a prerequisite of this preprocessing approach, it is logical to leverage that information for retention time alignment. Furthermore, through separation, many overlapping, low-intensity peaks can be resolved in the same time window that a single peak would be seen in single-channel chromatography. The additional resolution provides sufficient peak density to perform quite granular retention time correction without incorporating peak shape. Depending on the application, it might be reasonable to consider employing other retention time alignment techniques, such as COW, as an alternative or complement to the peak-based technique described here. However, we have found the algorithm as described performs well in correcting for both global and local perturbations in retention time.

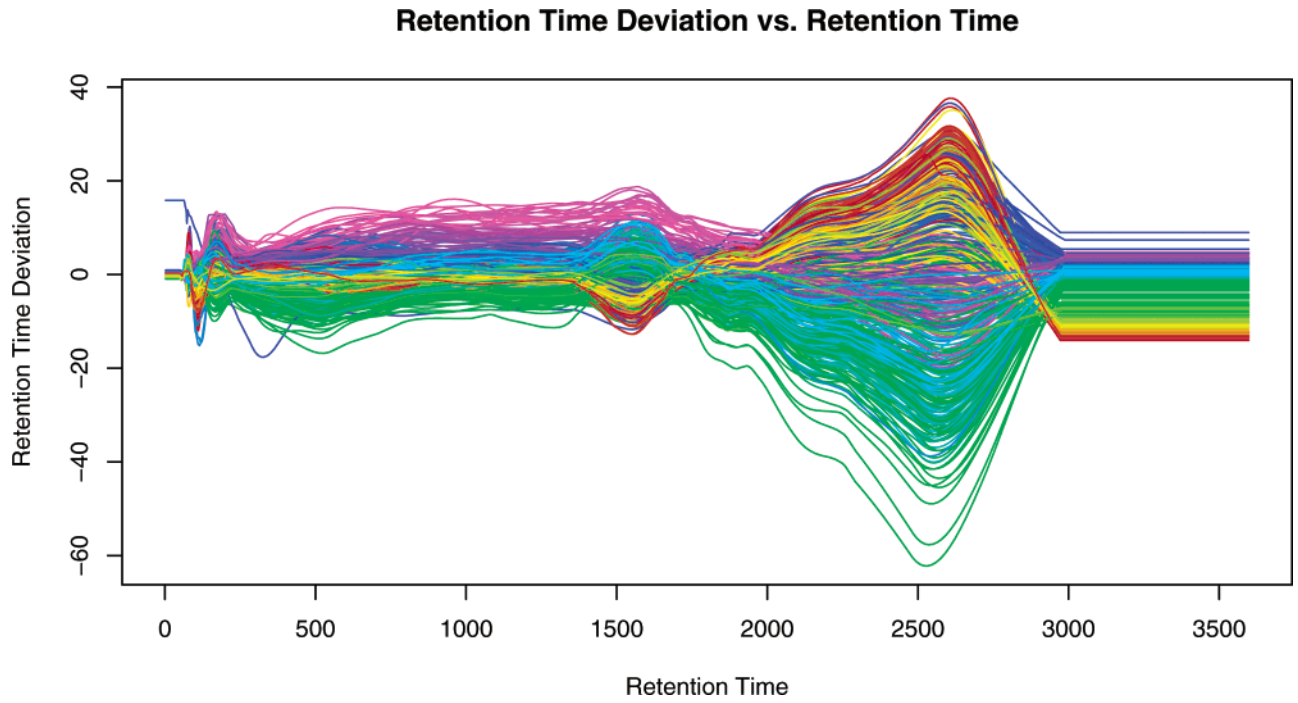

Figure 4. Retention time deviation profiles used for aligning 476 LC/MS analyses from the serum samples. All times are in seconds. The deviation profiles were created after two iterative steps of retention time alignment. A positive deviation indicates that the sample was eluting after the median retention time, and a negative deviation indicates that the sample was eluting before the median retention time. Sample profiles are colored in a rainbow by the order in which they were run, with red being the first samples and violet being the last samples run.

EXPERIMENTAL SECTION

Plasma Metabolite Extraction and LC/MS Analysis. A total of 238 plasma samples were included in this study. Ice-cold methanol (150 μL) was added to 50 μL of plasma, vortexed briefly, and incubated at 4 °C for 20 min. The supernatants were collected after centrifugation at 13200*g* for 10 min; these were dried and resuspended in 50 μL of 95/5 (v/v) water/acetonitrile. Reversed-phase chromatography and mass detection of the plasma metabolite extracts was performed on an Agilent 1100 LC/MSD SL system. Duplicate runs of each extraction were analyzed randomly, with a blank run between each sample to prevent carryover. For each run, 10 μL of serum metabolite extract was injected onto the same C18 column (Symmetry column, 2.1 × 100 mm, 3.5-μm particle size; Waters, Milford, MA) and eluted at a flow rate of 250 μL/min under gradient conditions of 5−90% B over 60 min. Mobile phase A consisted of water/acetonitrile/formic acid (95/5/0.1, v/v/v), and B consisted of acetonitrile/formic acid (100/0.1, v/v). Mass spectral data from 100 to 1000 m/z were collected in the positive ionization mode. Most samples were analyzed in duplicate. Several were analyzed in triplicate, and one was analyzed in singlet. The samples were analyzed in 31 days of continuous data acquisition.

Fatty Acid Amide Hydrolase (FAAH) Knockout Study. Data from a previously published study[4] examining the endogenous substrates of FAAH were reanalyzed. Briefly, spinal cord and brain tissue samples were taken from six wild type (WT) and six knockout (KO) mice and run in both positive and negative ion modes for a total of eight distinct sample types. Sample data were acquired in centroid mode on an Agilent 1100 LC/MSD system. Here, we only show data from the 12 positive-mode spinal cord KO/WT samples.

RESULTS AND DISCUSSION

To help illustrate the algorithmic details of XCMS, we reanalyzed a subset of the raw data from a previous FAAH knockout study[4] involving 12 samples. Additionally, this provides an example of XCMS discovering and visualizing novel differences between a set of LC/MS samples. To demonstrate the scalability, alignment of subtle shifts in retention time, and reproducibility, we analyzed data from over 200 plasma samples run in duplicate. All processing was done using XCMS version 1.0.0. Both analyses were initiated by compiling a list of all peaks in each sample using the peak detection algorithm. In both data sets, several thousand peaks were detected per sample. In our experience, the presence of isotopic peaks, adduct peaks, and multiply charged ions have not imposed significant limitations on data analysis. Therefore, a deisotoping routine has not yet been integrated into the peak detection algorithm, although there are plans to do so in the future. A summary of the analysis results and execution times is provided in Table 1.

After peak detection, the cross-sample peak-matching routine was used to determine initial groups of peaks thought to represent the same metabolite across samples. To allow for modest retention time differences between samples, the peak density smoother was initially set to a standard deviation (SD) of 30 s (Figure 3). Because cross-sample peak matching is based on the distribution of peaks across time, two peaks separated by several times the SD can be grouped together, provided there are other peaks distributed between them. Our experience has shown that the peak grouping of XCMS is much more dependent on the underlying peak distribution rather than the precise SD used. Unless there is an especially punctuated change in LC performance, most peaks representing the same analyte should be evenly distributed around

Table 1. Data Processing Results Summary[a]

	plasma samples	FAAH knockout
samples	476	12
data storage mode	continuum	centroid
mean file size/sample (MB)	175	28
sample classes	8	2
mean peaks detected/sample	3899	2564
well-behaved peak groups	167	316
reproducible peak groups	3071	1853
mean filled in peak groups/sample	898	741
algorithm running times (s)		
mean peak detection/sample	177	61
Cross-sample peak matching	74	32
Retention time alignment	209	3

[a] Well-behaved peak groups, used for retention time alignment, were those in which there was at most one sample with no peaks assigned and at most one sample with two peaks assigned. Reproducible peak groups, used for statistical analysis, were those in which one of the sample types had at least half its samples represented. For most of the reproducible peak groups, there were still samples that did not have a peak assigned to the group. Integrating samples in the area of those peak groups filled in the missing data points. Algorithm execution times were measured using an Intel 3.0-GHz Xeon processor with 2-GB memory.

a central retention time, as shown in the unaligned chromatograms of Figure 5. The peak-matching algorithm was designed with that conjecture in mind.

In most cases, the peak-matching algorithm can account for subtle shifts in retention time when there are few metabolites of similar mass. However, to more correctly match the isobaric peaks eluting near each other, the retention time alignment algorithm is used. The algorithm recognizes which groups have a high probability of being correctly matched (the well-behaved peak groups) and uses those groups to determine nonlinear profiles of retention time drift across the length of the LC run. An example of profiles from the plasma samples is shown in Figure 4. The algorithm then uses those profiles to warp all peaks into alignment, which can be thought of as transforming each profile into a straight line through 0 retention time deviation. For the plasma samples, the retention time varied by up to 100 s. For the FAAH knockout samples, the retention time varied by up to 50 s (data not shown). After retention time alignment, the smoother was reduced to a SD of 10 s and the peaks were regrouped.

One alignment strategy that can be employed for metabolite analysis is to use an internal standard to linearly shift the retention time across the entire run.[12] Such a strategy could be visualized by shifting each of the profiles in Figure 4 up or down so that they intersect at the time point of the internal standard. With such a strategy, it is impossible to correctly align the samples and may worsen the alignment depending at what time the internal standard elutes. For example, between 1500 and 2500 s, a large fraction of the samples eluting early switch to eluting late, and vice versa. If the internal standard had been chosen at one of those extremes, the alignment at the other extreme would be significantly worse. On the other hand, the alignment algorithm described here properly aligns peaks at both time points (Figure 5).

It is important to distinguish between peaks detected during peak picking and the features that are eventually used for relative ion intensity comparison. After peak picking, the peak-matching routine uses the combined peak information from all samples to determine where significant signal is located. A noise peak present in only a few samples is discarded as insignificant. Only peaks that appear reproducibly in a given fraction of samples are retained for relative ion intensity comparison. In that regard, when judging the analytical capacity of metabolite profiling techniques, it is perhaps more useful to consider the number of highly reproducible features rather than the average number of features detected per sample. Both data sets described here contained several thousand reproducible peak groups.

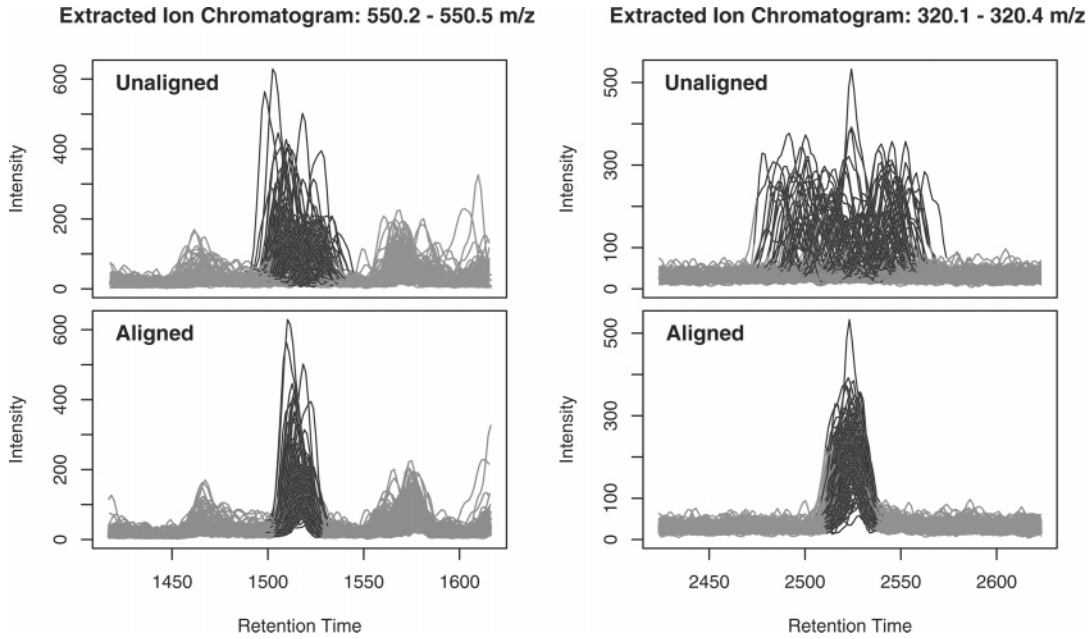

Figure 5. Overlaid extracted ion chromatograms from a subset of 126 random plasma samples. Retention time alignment is demonstrated for the extremes of the deviation profile at 1500 and 2500 s (Figure 4). A simple linear retention time shift could align peaks at only one of the time points, but not both. The nonlinear alignment algorithm in XCMS properly aligns peaks over all times. Darkened lines indicate where the peaks were integrated for relative ion intensity comparison.

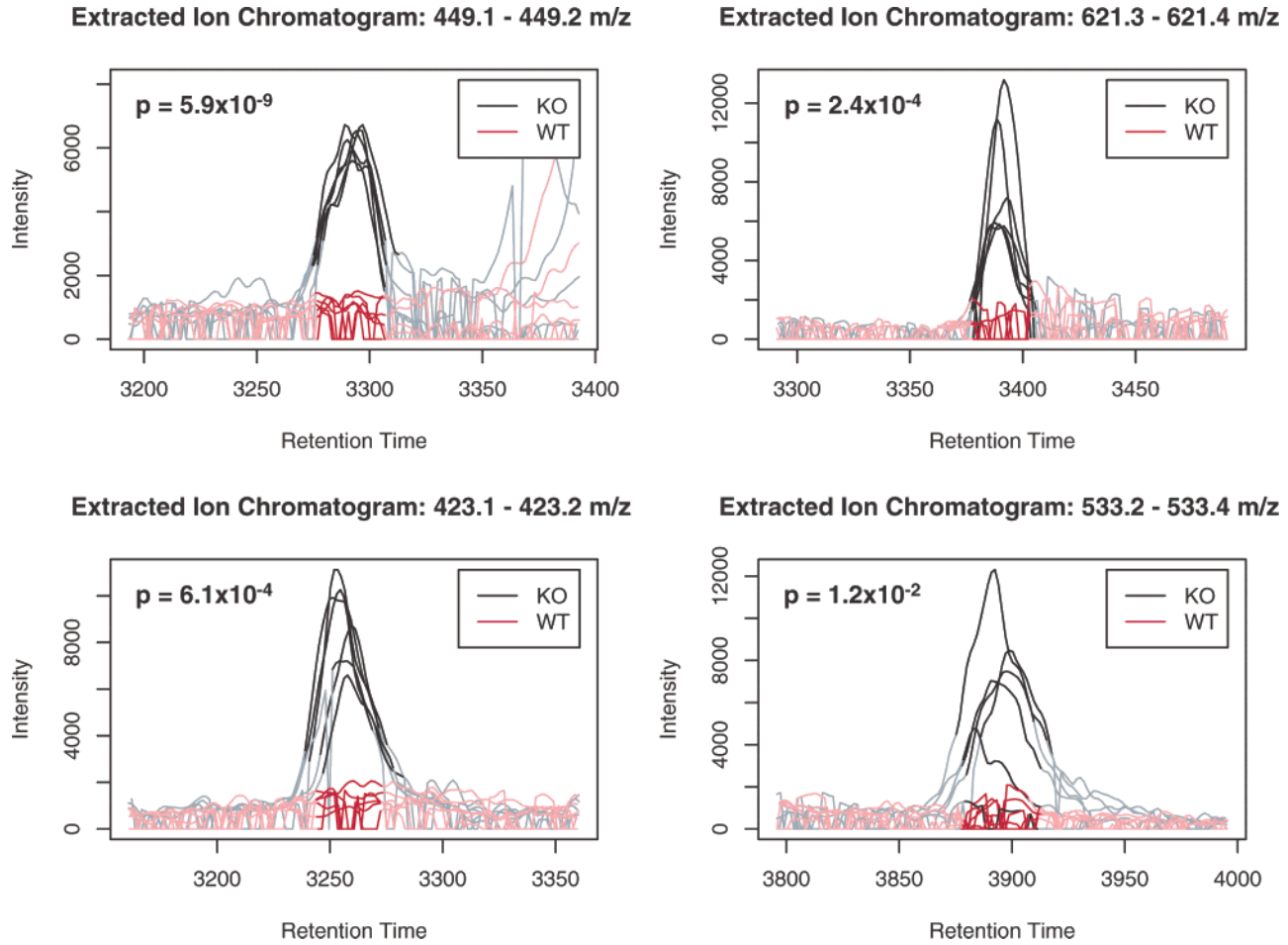

Figure 6. Using manual data analysis, Saghatelian et al.[4] characterized numerous *N*-acylethanolamines (NAE) and *N*-acyltaurines (NAT) as being metabolized by FAAH. On the same data set, the software described here identifies several additional metabolites also regulated by FAAH, but which are neither NAE/NAT nor previously characterized. Darkened lines indicate where the peaks were integrated for relative ion intensity comparison.

After determining a set of reproducible peak groups, most representing an individual analyte, XCMS determines which samples are missing from each peak group. Using statistics gathered during peak detection about where peaks begin and end, and aligned retention times for each sample, the raw LC/MS data are integrated to fill in intensity values for each of the missing data points (Figure 1). This provides significant advantages over using data from a vendor-supplied or third party peak detection program alone, as peaks that are missed during peak detection can be measured directly from the aligned raw data. As shown in Table 1, a significant number of potential peaks are missed during peak detection. That observation that is likely true of most, if not all, peak detection algorithms, given the inherent uncertainty close to the s/n cutoff. In the case that a peak shows up in one class of sample but not another, as in the FAAH knockout study, the step of filling in missing peak data is necessary for robust statistical analysis.

To assess the analytical reproducibility of the LC/MS acquisition and data processing, we calculated intensity differences between peaks from duplicate acquisitions of the same serum sample. Figure 7 shows median values of those differences for five fractions of the dynamic range. The reproducibility for high-intensity peaks was significantly better than that for low-intensity peaks. Because many of the lowest intensity data points were integrated purely from baseline noise, poor reproducibility is expected. The majority of duplicate data points showed less than a 20% difference in intensity. Given that degree of uncertainty, fold changes above 1.5 should be readily detectable with the current methodology. Importantly, no intensity normalization was used in the course of preprocessing. Normalization may further reduce variance and would be an important area of future study.

After preprocessing, any number of statistical analyses can be performed. For the task of identifying differentially regulated metabolites, we employed a simple univariate *t*-test to identify metabolites whose intensities are significantly different from sample class to sample class and rank the metabolites by *p*-value. While there are many other options for identifying differences, the focus of this work is on preprocessing prior to that analysis, not the analysis itself.

The final stage in the analysis strategy enabled by XCMS is to visualize the raw data for metabolites of interest. Most current software is not designed to visualize chromatograms from more than a few samples at once. However, when large data sets are analyzed, such verification is crucial and can be a limiting factor in purely numerical studies. To address that problem, XCMS automatically produces superimposed, aligned EICs for peaks of interest, allowing quick visual scanning of hundreds of metabolites at once. Examples of those EICs are shown in Figures 5 and 6.

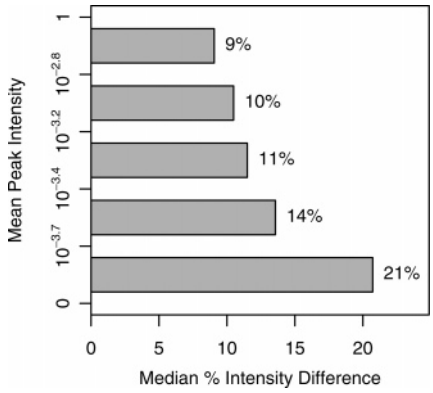

Figure 7. Sample percent intensity differences across the dynamic range (from duplicates). Random variations in both data acquisition and processing contributed to differences in peak intensity from otherwise identical samples. The highest peak intensity fraction showed the least error with a median difference of 9%. The lowest peak intensity fraction showed the greatest error with a median difference of 21%. That fraction included intensity data from empty spectral regions containing only background noise (i.e., the wild type data in Figure 6). To produce this figure, the pairwise percent intensity differences were calculated for all 3071 peak groups in each of 238 pairs of duplicate serum samples. The percent intensity difference is defined as 100(absolute intensity difference)/(mean intensity). The resulting 730 898 percentages were divided into five equally sized fractions across the dynamic range of intensities. Mean peak intensities were normalized to a maximum of 1.

In particular, Figure 6 shows metabolites potentially regulated by FAAH that were not identified by the authors' original manual data analysis.

To help begin the identification of unknown metabolites showing differential regulation, XCMS provides links to the Metlin Metabolite Database (http://metlin.scripps.edu) and produces a list of potential metabolite identities based on exact mass. For instance, in the FAAH knockout data set, the most statistically significant ($p = 5.0 \times 10^{8}$) differentially regulated ion was 300.2 m/z and eluted at 56.6 min. Given a mass uncertainty of ± 0.15 m/z, there were 10 metabolites recorded in Metlin (Supporting Information Figure S-1). *N*-(2-Hydroxyethyl)palmitamide, an *N*-acyl ethanolamine and known substrate of FAAH, was among those. It is important to note that *N*-(2-hydroxyethyl)palmitamide was added to Metlin after publication of the initial study. As more and more metabolites are characterized and placed into public databases, the number of successful matches, as in this case, will increase. In addition, higher accuracy mass measurement will further narrow the number of metabolites identified based on mass alone.

CONCLUSION

Three important aspects of XCMS are its design, availability, and flexibility. It is written in the R statistical programming language[20] and is freely available under an open-source license. It is distributed through both the Metlin Metabolite Database[21] and the Bioconductor bioinformatics project,[22] which itself provides a wealth of easily integrated tools for analysis of preprocessed, high-throughput data. The openness of the software allows it to be easily customized for different data analysis needs or optimized for a particular application. Alterations can be undertaken without requiring significant investment developing the infrastructure and algorithms that XCMS already provides. This type of open software architecture will be increasingly important for encouraging the growth and development of data-intensive research.[23]

Much of the software currently available for LC/MS analysis is either proprietary or restricted to a particular vendor's instruments. By contrast, the software described here has been designed to be independent of instrument vendor. That enables identical data analysis to be undertaken and repeated using instruments from a variety of manufacturers. While other programs include tools for peak identification and visualization, they may be designed in a very rigid manner that does not allow automated analysis of many samples. On the other hand, the design of XCMS inherently allows for many complex tasks to be programmed and performed without user intervention. A key example of that flexibility is how it integrates peak detection, statistical analysis, and subsequent visualization of raw data for verification purposes. Additionally, while the software has been designed for LC/MS metabolite data, it could be adapted to work with other types of data, such as peptide digests, or other instrumentation, such as GC/MS.

ACKNOWLEDGMENT

The authors appreciate funding from NIH grants 5P30EY012598-04 and 5R24EY01474-04 for supporting this effort. Zhouxin Shen provided useful discussions and insights at the outset of development. The authors thank Alan Saghatelian and Benjamin Cravatt for providing the original raw data from their FAAH knockout experiment for use as a benchmark during algorithm development. We also thank the Bioconductor project for providing software distribution and testing infrastructure.

SUPPORTING INFORMATION AVAILABLE

Figure S-1, table of 10 possible metabolite identities for positive-mode ion at 300.2 m/z. This material is available free of charge via the Internet at http://pubs.acs.org.

Received for review August 10, 2005. Accepted November 23, 2005.

AC051437Y

(20) R Development Core Team. *R Foundation for Statistical Computing*; Vienna, Austria, 2005.

(21) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. *Ther. Drug Monit.* **2003**, *27*, 747−751.

(22) Gentleman, R. C.; Carey, V. J.; Bates, D. M.; Bolstad, B.; Dettling, M.; Dudoit, S.; Ellis, B.; Gautier, L.; Ge, Y.; Gentry, J.; Hornik, K.; Hothorn, T.; Huber, W.; Iacus, S.; Irizarry, R.; Leisch, F.; Li, C.; Maechler, M.; Rossini, A. J.; Sawitzki, G.; Smith, C.; Smyth, G.; Tierney, L.; Yang, J. Y.; Zhang, J. *Genome Biol.* **2004**, *5*, R80.

(23) Wiley: H. S.; Michaels, G. S. *Nat. Biotechnol.* **2004**, *22*, 1037−1038.