analytical chemistry

Joint Bounding of Peaks Across Samples Improves Differential Analysis in Mass Spectrometry-Based Metabolomics

Leslie Myint,[†] Andre Kleensang,[‡] Liang Zhao,[‡] Thomas Hartung,[‡,§] and Kasper D. Hansen*,[†,∥]

[†]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States

[‡]Center for Alternatives to Animal Testing (CAAT), Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland 21205, United States

[§]University of Konstanz, CAAT-Europe, 78457 Konstanz, Germany

[∥]McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, United States

S Supporting Information

ABSTRACT: As mass spectrometry-based metabolomics becomes more widely used in biomedical research, it is important to revisit existing data analysis paradigms. Existing data preprocessing efforts have largely focused on methods which start by extracting features separately from each sample, followed by a subsequent attempt to group features across samples to facilitate comparisons. We show that this preprocessing approach leads to unnecessary variability in peak quantifications that adversely impacts downstream analysis. We present a new method, bakedpi, for the preprocessing of both centroid and profile mode metabolomics data that relies on an intensity-weighted bivariate kernel density estimation on a pooling of all samples to detect peaks. This new method reduces this unnecessary quantification variability and increases power in downstream differential analysis.
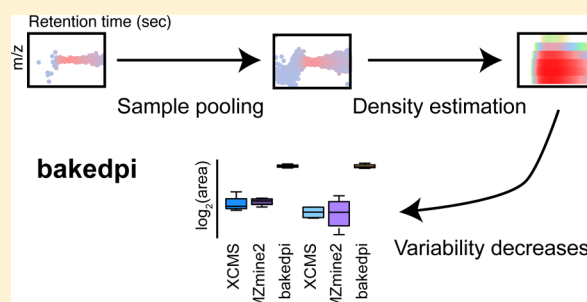
As mass spectrometry-based metabolomics becomes a more mature and popular means of scientific investigation,[1−3] it is important to revisit existing data analysis paradigms. Existing approaches to preprocessing metabolomics data focus on a two-step approach which starts by extracting features (peaks) separately from each sample, followed by a subsequent attempt to group features across samples to facilitate comparisons.[4] In particular, there has been considerable attention in the literature on individual stages of preprocessing, including peak detection[5−14] and alignment.[15−18] Additional work has been done on specific issues with downstream differential analysis such as missing information or dependence structures.[19−21] Single sample processing methods tend to focus on reducing bias. The bias-variance trade-off[22] shows that the overall performance of a method also depends on its noise and experience from gene expression studies suggests that noise can be removed by processing samples jointly.

In this work, we investigate the consequences of traditional sample-specific preprocessing on the quality of differential analysis. We show that the retention time (RT) bounds that arise from preprocessing samples individually cause unnecessary variability in peak quantifications (based on integrated peak area) which leads to under-powered differential analysis. We propose a relative quantification method, called bakedpi, which addresses this shortcoming by jointly detecting and bounding peaks in the two-dimensional m/z-RT space, across all samples simultaneously. The backbone of our method is an intensity-weighted bivariate kernel density estimation that is computed on a pooling of all samples. We show that this approach reduces unnecessary quantification variability and increases power in downstream differential analysis. Our method is open source and freely available as part of the yamss package through the Bioconductor project under Artistic License 2.0.

RESULTS AND DISCUSSION

Excess Variability with Sample-Specific Processing. To demonstrate issues with sample specific detection and bounding of peaks, we consider the widely used software packages XCMS[23] and MZmine2.[24] Output for one peak from a QTOF data set with two sample groups is shown in Figure 1 (additional examples from other data sets in Supplementary Figures S1 and S2). The shape, width, and location of this peak do not appear to vary across samples. Despite this, the XCMS and MZmine2 RT bounds for this peak, indicated by blue and purple rectangles respectively, are highly heterogeneous between samples (Figure 1c). To a first approximation, the retention time (RT) bounds can be grouped into narrow and wide bounds; this grouping is not associated with the two
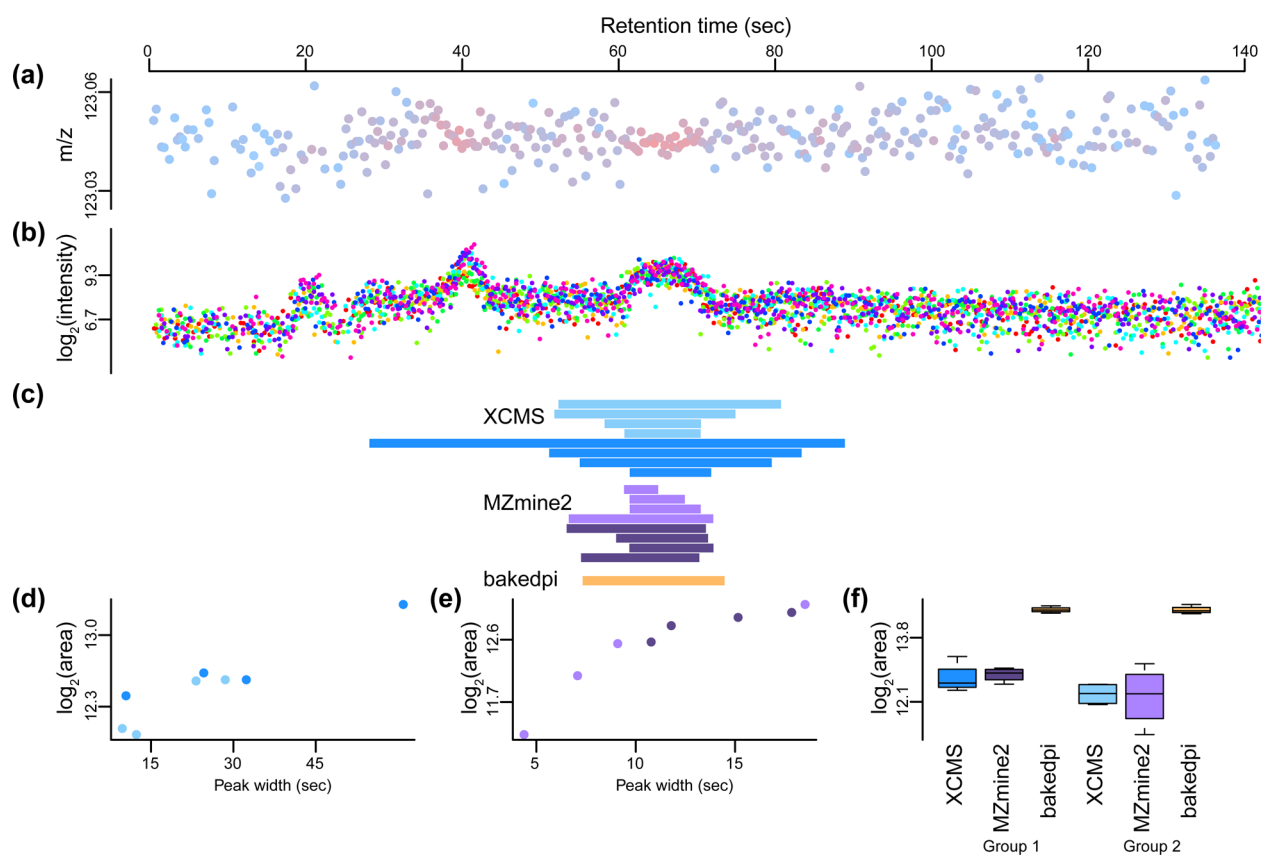
Figure 1. Problems with sample-specific processing in XCMS and MZmine2. Peak detection and bounding for a single peak in the MTBLS2_rep1 data set. (a) The *m/z*-RT space surrounding this peak for a single sample, color is used to depict intensity (red is high). (b) Overlaid extracted ion chromatograms from all 8 samples in the experiment. Different colors denote different samples. (c) The peak bounds for all samples for XCMS (blue), MZmine2 (purple) and bakedpi (orange; all samples have same bounds). This experiment compares two groups of samples indicated with different color shades. (d) XCMS peak quantification vs peak width. (e) Like part d but for MZmine. (f) Distribution of peak quantifications, based on the peak bounds in part c. Substantial heterogeneity in the sample-specific bounds leads to excess variability in the quantifications; this is addressed by using the same RT bound for all samples.
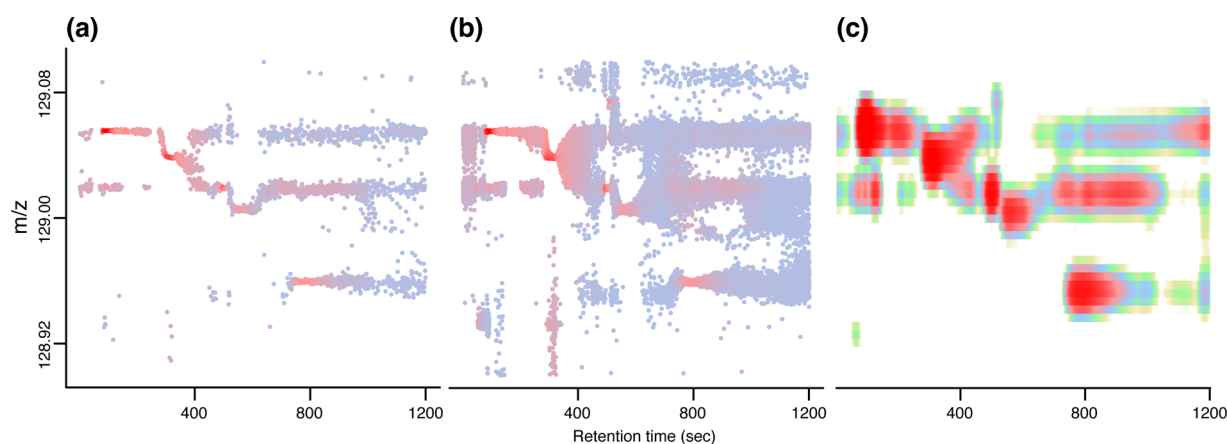

Figure 2. Weighted bivariate kernel density estimation. We depict a selected rectangle in *m/z*-RT space for (a) one sample and (b) the pooled metasample. *m/z* values with higher intensity are shown in red, lower with blue. (c) The weighted bivariate density estimate.

sample groups (light and dark rectangles). As a consequence, the integrated peak area is completely determined by whether the RT bounds are narrow or wide (Figure 1d,e), and this leads to high variability in the peak quantifications (Figure 1f). If instead, we use the same RT bound across all samples (Figure 1c, orange rectangle), we substantially reduce the between-sample variability in the peak quantifications (Figure 1f). Excess variability results in loss of power in a differential analysis.

Joint Sample Processing with bakedpi. To address the problem of excess variability, we propose a method which jointly detects and bounds peaks across all samples in an experiment (see Methods); an important feature of our method is the use of homogeneous RT bounds across all samples. We pool the data from all samples into a single metasample, on which we detect and bound peaks (Figure 2a,b). To do this, we use intensity-weighted bivariate kernel density estimation in the

two-dimensional m/z-RT space. By using the intensities as weights, we differentiate between groups of detected m/z values (data points) with high and low intensities. The output is a smooth density in the m/z-RT space, where peaks in the density correspond to clusters of high-intensity points (Figure 2c). To detect and bound peaks, we slice the density using a single global threshold and form a set of contiguous regions based on the density slices. By performing this procedure on a single metasample, we ensure the same peak bounds across all samples. Like XCMS and MZmine2, we quantify the peaks by integrating the extracted ion chromatogram (EIC) for each sample across the peak's RT bounds. We can optionally perform RT alignment prior to density estimation. Our method has 3 parameters: 2 of these parameters control the bandwidth in the m/z and RT domains and are easy to set based on the resolution of the instrument. The last parameter, the only significant tuning parameter, is the global density threshold. We call our method bakedpi for bivariate approximate kernel density estimation for peak identification.

Joint Sample Processing Reduces Excess Variability. We applied bakedpi to 10 different data sets from 7 different experiments. Features of these data sets are summarized in Table 1. All data sets were subset (if necessary) to only contain

Table 1. Characteristics of Evaluation Datasets[a]

name (source)	MS instrument column	no. samples (group 1, 2)
ASD_hirisk (C)	QTOF HPLC-HILIC	20, 20
timecourse_4 h (C)	QTOF HPLC-HILIC	6, 6
timecourse_24 h (C)	QTOF HPLC-HILIC	6, 6
MTBLS2_rep1 (M)	QTOF UPLC-reverse phase	4, 4
MTBLS2_rep2 (M)	QTOF UPLC-reverse phase	4, 4
CAMERA_pos (M)	QTOF UPLC-reverse phase	3, 3
CAMERA_neg (M)	QTOF UPLC-reverse phase	3, 3
MTBLS103 (M)	QTOF UPLC-HILIC	14, 12
MTBLS213 (M)	QTOF UPLC-reverse phase	6, 6
MTBLS126 (M)	Orbitrap HPLC-HILIC	3, 3

[a]C = CAAT, M = Metabolights.

two sample groups, to keep the experimental design simple and constant. For the Orbitrap data set (MTLS216) we expect little to no differences between the sample groups, based on the design of the experiment.[25] We ran XCMS, MZmine2, and bakedpi on the 10 data sets. XCMS parameters were optimized using the IPO package available on Bioconductor[26] using recommended starting values for most data sets (Methods). MZmine2 parameters were set based on optimized XCMS parameters where possible (Methods). When running bakedpi, we use the higher of a fixed quantile cutoff and a data-driven cutoff to set the global tuning parameter (Methods).

To compare the quantification variability between bakedpi and XCMS and between bakedpi and MZmine2, we first identified peaks which overlapped between bakedpi and XCMS

and between bakedpi and MZmine2. We will call these shared peaks. The number of peaks detected by both methods as well as the percentage of peaks that are common to both methods are shown in Supplementary Figure S3; for many data sets the overlap is around 60−80% of the peaks. On these overlapping peaks, we computed the residual standard deviation of the log-abundances to assess their variability. We used residual standard deviation to avoid being influenced by changes in the log-abundances between the two sample groups in the different experiments. Figure 3 shows the distribution of differences in
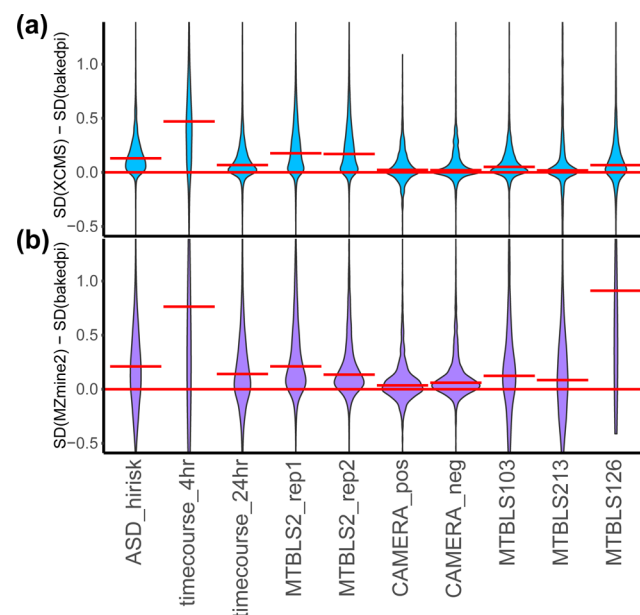

Figure 3. Variability comparison of peak quantifications. (a) For peaks that are detected both by bakedpi and XCMS, the distribution of the differences in residual standard deviation for all data sets are shown as violin plots. Each violin is a mirrored density plot; the median is indicated by a horizontal red line. (b) Like part a but for MZmine. For all data sets, the majority of peaks detected by both methods have quantifications that are less variable when quantified with bakedpi.

residual standard deviation (XCMS or MZmine2 minus bakedpi) for each data set. Values greater than zero indicate that bakedpi has smaller variability than the other method. For all data sets examined, more than half of the peaks detected by both methods had lower variability when quantified by bakedpi; for some data sets it was substantially higher.

Joint Processing Improves Power in a Differential Analysis. We next sought to determine if the decrease in residual standard deviation of the peak quantifications leads to increased power in a differential analysis. We used the limma[27] differential analysis pipeline as it has been shown to provide robust and powerful inference for proteomics data.[28] This method was originally developed to analyze microarray expression studies and uses empirical Bayes techniques to shrink feature (adduct)-wise variances toward a common underlying value to provide more stable inference. The resulting p-value distributions for the shared peaks in the timecourse_4 h data set are shown in Figure 4a (additional data sets in Supplementary Figure S4). For the majority of the data sets, bakedpi has a p-value distribution that is more peaked around zero than XCMS and MZmine2, indicating that bakedpi detects more significant peaks among the overlapping peaks. When comparing with XCMS, the timecourse_24 h data set is
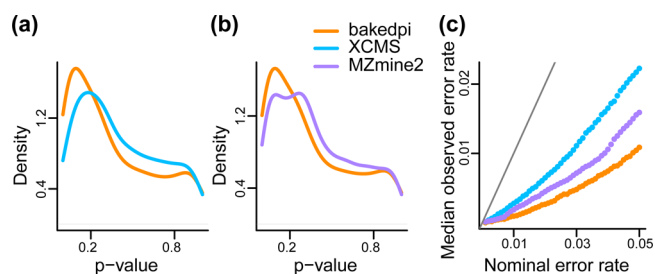
Figure 4. Comparison of differential analysis quality and type I error control in the timecourse_4 h data set. (a) Distribution of *p*-values for peaks detected by both bakedpi and XCMS. (b) Like part a but for MZmine, (c) median error rate over null permutations as a function of the nominal error rate.

the only one in which XCMS has a taller peak around zero. When comparing with MZmine2, only for the CAMERA_pos data set does MZmine2 have a taller peak around zero.

Higher detection rates alone do not necessarily indicate an increase in power. To assess power, we also evaluated the type I error control of the methods. We performed a permutation experiment in which we shuffled the sample group labels so that each of the new comparison groups were composed half of cases and half of controls. For example, in an experiment with eight cases and eight controls, the new permuted "case" group would include four true cases and four true controls, as would the new permuted "control" group. In this way, we created null data sets in which no abundance differences are expected. With data sets containing a sufficient number of samples, we performed 1000 permutations. Otherwise we enumerated all permutations satisfying the balancing characteristic just described. We again used limma to perform differential testing. Results of the permutation experiment for the timecourse_4 h data set are shown in Figure 4c (additional data sets in Supplementary Figure S5). For a range of nominal type I error rates, we computed the median observed error rate over all permutations. For all 10 data sets, all methods are quite conservative, showing a markedly lower error rate than the nominal value for the entire range. For most of the data sets, bakedpi is the most conservative of the three methods. The combination of more conservative type I error control and a higher detection rate indicates that bakedpi has higher power to detect differences than the sample-by-sample processing procedures of XCMS and MZmine2.

Retention Time Alignment. It is well established that RT deviations between experimental runs can complicate the matching of peaks across samples. We investigated the impact of correcting RT drift on the variability improvements of our method using multiple strategies. First, we used the RT warping function computed by XCMS to align the raw data before computing the density estimate. Second, we computed local sample-specific RT shifts that maximized the correlation of the chromatograms between samples and used these shifts to align the raw data. Third, we used correlation-optimal shifts to align peaks already detected from the density estimate before quantification. None of these RT alignment strategies had a large impact on the variability of detected features. The proportion of peaks detected by both bakedpi and XCMS or MZmine2 that had lower variability with bakedpi did not change appreciably with these RT corrections (Supplementary Figure S6).

Parameter Choices. Because the detection of peaks and their bounds depend on the cutoff applied to the density estimate, it is important to investigate the sensitivity of method performance to this cutoff. We performed a sensitivity analysis by varying the density cutoff and examining the *p*-value distribution resulting from the detected peaks (Supplementary Figures S7 and S8). Raising the cutoff to be more stringent or lowering the cutoff to be more inclusive generally does not have a substantial impact on the global pattern of inference as assessed by *p*-value distributions.

Method-Specific Peaks. There are a number of peaks that are detected only by one method (Supplementary Figure S3). As comprehensive gold standard information on the true peaks corresponding to compounds was not available, we examined the characteristics of these method-specific peaks to assess their quality (Supplementary Figures S9−S12). For more than half of the data sets, XCMS-specific peaks tend to have more extreme *t*-statistics and lower *p*-values. For half the data sets, MZmine2 peaks have higher *p*-values than bakedpi. For nearly all data sets, bakedpi-specific peaks have greater peak heights than XCMS- and MZmine2-specific peaks with comparable peak widths. Peaks specific to bakedpi are also more likely to be supported by all samples in the experiment. The last two observations are sensible given that bakedpi relies on an intensity-weighted density estimation; a peak is more likely to be detected when a large number of high-intensity points are close together. Based on observations about *t*-statistics and *p*-values, it is not clear that one of the two sets of methods-specific peaks is best. If peaks with greater heights or greater numbers of samples supporting them are more likely to be of scientific interest, then bakedpi-specific peaks seem to be of higher quality than XCMS- or MZmine2-specific peaks. Given the lack of gold standard data on peak identities, evaluation of method-specific peaks is less clear than evaluation of peaks common to multiple methods. On peaks common to both bakedpi and MZmine2 or XCMS, bakedpi shows a clear reduction in quantification variability and an increase in statistical power.

■ CONCLUSIONS

We have proposed a method for the joint processing of metabolomics data across samples, which reduces variability in peak quantification across samples, leading to increased power in a differential analysis. We take the position that the most important task in metabolomics is the identification of differentially abundant peaks, in contrast to, e.g., identifying all peaks in a sample. Our method compares favorably to XCMS and MZmine2 across 10 data sets and will be useful for drawing better and more substantiated inferences from untargeted metabolomics studies. We do note that the commercial software Progenesis CoMet also uses the idea of pooling samples into a metasample for processing. However, details on CoMet method are not available, making it impossible to comment further on differences between the two approaches.

A limitation of our approach is that peaks that are only truly present in a small fraction of the samples are unlikely to be detected. Such metabolites may be of interest but are by definition less well supported by the observed data. In developing bakedpi, we have chosen to focus on peaks with sufficient information across all samples and on obtaining for those peaks the best quality quantifications for the purposes of differential analysis.

It is important to note that the benefit of our method is dependent on using peak areas for quantification rather than peak height. As we show, the variability in quantification of a particular peak across samples is driven entirely by the variability in peak width. If peak height is used instead of peak area, our method will show the same quantification as XCMS and MZMine2, provided the sample-specific RT bounds contain the mode of the peak; this is true for two of our three examples.

In our evaluation of bakedpi, we have used both centroid-mode and profile-mode data sets with fairly stable chromatography. The RT drift we observe in these data sets is not so large that corresponding peaks from different samples do not overlap. However, stable chromatography is not required for bakedpi to work because we do implement RT alignment procedures. Our evaluation data sets also come from mass spectrometers with a range of mass accuracies from 5 ppm on Q-TOF instruments to less than 1 ppm on the Orbitrap, so bakedpi is able to handle data from a representative range of instruments. We expect lower mass accuracy to make peak merging more likely and to cause peak m/z bounds to be wider than necessary, but this is mostly a feature of low mass accuracy in general. Currently, our method is implemented as the standalone yamss package as part of the Bioconductor project.

■ METHODS

Data. Also see Table 1.

ASD_hirisk: Prenatal serum samples from 40 mothers participating in the EARLI study whose infants had the highest (n = 20) and lowest (n = 20) Autism Observation Scale for Infants (AOSI) at the time of the experiment.[29]

timecourse_4h, timecourse_24 h: Six MCF-7 cell line samples exposed to estradiol (E2) and six control samples unexposed to E2 for up to 72 h.[30]

MTBLS2: Four wild-type and four *cyp79b2 cyp79b3* knockout *Arabidopsis thaliana* leaves exposed to silver nitrate.[31,32]

CAMERA: Spike-ins of 39 known compounds at varying concentrations on methanolic extracts of *Arabidopsis thaliana* leaves.[33] Three samples with a spike-in concentration of 20 μM were compared to three samples with a spike-in concentration of 5 μM in both positive and negative ion modes.

MTBLS103: Serum profiling of 12 adolescent girls with hyperinsulinaemic androgen excess and 14 healthy controls matched on age, weight, and ethnicity.[34]

MTBLS213: Human retinal pigment epithelium cell line (ARPE-19) batches grown labeled and unlabeled glucose media.[35]

MTBLS126: Liver concentrations of resveratrol (RESV) metabolites after application of a mixture of RESV in hydrophilic ointment to mouse skin (3 samples) compared to liver concentrations of resveratrol (RESV) metabolites after application of hydrophilic ointment without RESV to mouse skin (3 samples).[25]

Processing with XCMS and MZmine2. XCMS parameters were optimized using the IPO package available on Bioconductor[26] using recommended starting values for most data sets. Because optimization for the MTBLS2 and MTBLS213 data sets required significant computational time (we terminated the optimization after 11 days), we either fixed parameters that could be reasonably inferred beforehand (such as ppm) or set a smaller range of values over which to optimize. MZmine2 parameters were set based on optimized XCMS

parameters where possible. In particular, the "prefilter", "mzdiff", minimum and maximum peakwidth, and ppm parameters from XCMS had near equivalents in MZmine2 parameters. For XCMS, we used the "centWave" algorithm[9] for the nine centroid-mode data sets and the "matchedFilter" algorithm[23] for the profile-mode MTBLS126 data set. We used the density method for peak grouping, the obiwarp method for retention time alignment, and the fillPeaks method to fill in information for peaks missing from certain samples. For MZmine2, we used the GridMass module for peak detection,[36] the join aligner for retention time alignment, and the same-range gap filler module. Details on optimization and parameter settings for XCMS and MZmine2 are provided in the Supporting Information.

Processing Workflow. Our processing procedure consists of three steps. First is background correction which increases the signal-to-noise ratio of true peaks. Second is RT alignment, which aligns the raw data to correct for drifts in compound elution times between samples; this is optional. Third is density estimation to detect peaks.

Background Correction. Background correction is performed on each sample separately. We divide the m/z-RT space into bins and estimate background separately for each bin; this is arbitrarily done for bins of width 10 m/z units and 40 scans in the RT domain. We observe that each grid region exhibits a multimodal intensity distribution with 2 or more modes (Supplementary Figure S13) and reason that the lowest mode is background. We estimate the location of the mode with the first peak of the kernel density estimate of the intensity distribution and subtract this value from all observations in the grid region.

Retention Time Alignment. We investigated two RT alignment procedures that could be applied to the raw data before peak detection and one procedure that could be applied after peak detection. The first prepeak detection approach was to use the sample-specific corrected RTs reported by XCMS to define a RT warping function that could be applied to the raw data to yield aligned RTs. In the second approach, we found tentative m/z regions containing peaks using univariate kernel density estimation and computed EICs in these regions for all samples. For each region and sample, we then found the shift that would maximize the correlation between the EICs in each sample and a reference sample (the sample with the largest area beneath the EIC). These local and sample-specific shifts were applied to the raw data to yield aligned RTs. We also investigated a correction procedure that could be applied to peaks that had already been detected. For each detected peak, we computed the sample-specific shifts that would maximize the correlation between the EICs in each sample and a reference sample (the sample with the largest area beneath the EIC). We then recomputed the peak quantifications using original RT bounds and shifted EICs.

Bivariate Density Estimation. To detect peaks, we pool all samples into a single metasample by concatenating the spectral information from all of the samples. For example, the spectral information for the first scan of the metasample is formed by concatenating the first scan's spectral information from the individual samples. We use this metasample to estimate a two-dimensional density in the m/z-RT space. We represent the input data as a set of data points (M_j, T_j, I_j), where M_j is the mass over charge (m/z) of the jth data point (all samples are pooled), T_j is the scan number (RT in seconds divided by number of scans per second), and I_j is the intensity.

Per sample, T typically has up to a few thousand unique values depending on the scan rate of the mass spectrometer and the duration of the experiment, and M has on the order of one hundred observations per scan in centroid-mode data and several hundred in profile-mode data. Thus, the data consists of tens of thousands of data points such triples for each sample.

The bivariate intensity-weighted density estimator using a Gaussian kernel at a point (m, t) in m/z-RT space is given by

$$\hat{f}(m, t) = \frac{1}{h_M h_T \sum_{j=1}^{n} I_j} \sum_{j=1}^{n} I_j \phi_2 \left(\frac{m - M_j}{h_M}, \frac{t - T_j}{h_T} \right)$$

where $j = 1, ..., n$ indexes the n data points, h_M and h_T are the bandwidths in m/z and RT space, respectively, and ϕ_2 is a bivariate Gaussian density. The density estimate is not highly sensitive to the RT bandwidth, and a default of bandwidth of 10 scans is recommended. The m/z bandwidth should be set based on the type of mass spectrometer used and is recommended to be 0.005 for TOF and 0.002 for Orbitrap instruments. Because the density estimate involves a sum over all n data points at each value of (m, t), we use various approaches to make this computationally tractable. First, we use a diagonal covariance matrix for the bivariate kernel; this implies the factorization

$$\phi_2 \left(\frac{m - M_j}{h_M}, \frac{t - T_j}{h_T} \right) = \phi_1 \left(\frac{m - M_j}{h_M} \right) \phi_1 \left(\frac{t - T_j}{h_T} \right)$$

We do this because our focus is on identifying regions of interest rather than on highly exact estimation of the density.[37] Second, we use a simple binning strategy[38] where the m/z-RT space is binned and a single representative value for each bin is chosen. In the RT domain, the default bin width is 1 scan, and in the m/z domain the default bin width is set to be equal to the bandwidth (0.005 for TOF and 0.002 for Orbitrap). Third, we use a Gaussian kernel truncated at ± 3, effectively only including points close to (m, t) in the summation.[38] Fourth, in our implementation, we make use of sparse linear algebra as well as efficient data structures for selecting points close to (m, t) as implemented in the data.table package.[39]

After obtaining the density estimate, we select a cutoff using information from the strongest (most intense) features in the data. The m/z domain is divided into bins of a default width of 2 m/z. Within each bin, the most intense data point is selected. We assume that this data point belongs to a true feature and use local univariate density estimation in the m/z and RT domains to define a m/z and RT window for this feature. We compute quantiles of the density estimate values in these regions and compute the mode of this quantile distribution for various quantile values. For example, we compute the 99th percentile for each of the approximately 500 strong feature regions and compute the mode of this distribution. We repeat this for a wide range of percentiles. We then order these modes and select the first mode substantially different from zero as a cutoff. To ensure reasonable peak bounds, we enforce that this cutoff should be greater than or equal to the 99th percentile of nonzero density values. Applying the cutoff to the density estimate matrix yields a binary matrix that denotes peak and nonpeak regions. In order to obtain m/z and RT bounds for these peak regions, we use a connected components labeling algorithm.[40]

Software Availability. Our method is implemented in the yamss package, available from the Bioconductor project at https://www.bioconductor.org/packages/yamss.

ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.anal-chem.6b04719.

> Additional details on parameter settings for XCMS and MZmine2 as well as additional figures of data analysis (PDF)

AUTHOR INFORMATION

Corresponding Author

*E-mail: khansen@jhsph.edu.

ORCID Ⓞ

Thomas Hartung: 0000-0003-1359-7689
Kasper D. Hansen: 0000-0003-0086-0687

Notes

Disclaimer: The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors declare no competing financial interest.

REFERENCES

(1) Bouhifd, M.; Hartung, T.; Hogberg, H. T.; Kleensang, A.; Zhao, L. J. Appl. Toxicol. 2013, 33, 1365−1383.

(2) Bouhifd, M.; et al. ALTEX 2015, 32, 319−326.

(3) Ramirez, T.; et al. ALTEX 2013, 30, 209−225.

(4) Aberg, K. M.; Alm, E.; Torgrip, R. J. O. Anal. Bioanal. Chem. 2009, 394, 151−162.

(5) Hastings, C. A.; Norton, S. M.; Roy, S. Rapid Commun. Mass Spectrom. 2002, 16, 462−467.

(6) Vivó-Truyols, G.; Torres-Lapasió, J. R; van Nederkassel, A. M.; Vander Heyden, Y.; Massart, D. L. J. Chromatogr. A 2005, 1096, 133−145.

(7) Du, P.; Kibbe, W. A.; Lin, S. M. Bioinformatics 2006, 22, 2059−2065.

(8) Noy, K.; Fasulo, D. Bioinformatics 2007, 23, 2528−2535.

(9) Tautenhahn, R.; Böttcher, C.; Neumann, S. BMC Bioinf. 2008, 9, 504.

(10) Chen, S.; Li, M.; Hong, D.; Billheimer, D.; Li, H.; Xu, B. J.; Shyr, Y. Bioinformatics 2009, 25, 808−814.

(11) Nguyen, N.; Huang, H.; Oraintara, S.; Vo, A. Bioinformatics 2010, 26, i659−i665.

(12) Shalliker, R. A.; Stevenson, P. G.; Shock, D.; Mnatsakanyan, M.; Dasgupta, P. K.; Guiochon, G. J. Chromatogr. A 2010, 1217, 5693−5699.

(13) Vivó-Truyols, G. Anal. Chem. 2012, 84, 2622−2630.

(14) Fu, H.-Y.; Guo, J.-W.; Yu, Y.-J.; Li, H.-D.; Cui, H.-P.; Liu, P.-P.; Wang, B.; Wang, S.; Lu, P. J. Chromatogr. A 2016, 1452, 1−9.

(15) Tomasi, G.; van den Berg, F.; Andersson, C. J. Chemom. 2004, 18, 231−241.

(16) Podwojski, K.; Fritsch, A.; Chamrad, D. C.; Paul, W.; Sitek, B.; Stühler, K.; Mutzel, P.; Stephan, C.; Meyer, H. E.; Urfer, W.; Ickstadt, K.; Rahnenführer, J. *Bioinformatics* **2009**, *25*, 758−764.

(17) Hoffmann, N.; Keck, M.; Neuweger, H.; Wilhelm, M.; Högy, P.; Niehaus, K.; Stoye, J. *BMC Bioinf.* **2012**, *13*, 214.

(18) Jeong, J.; Shi, X.; Zhang, X.; Kim, S.; Shen, C. *BMC Bioinf.* **2012**, *13*, 27.

(19) Tekwe, C. D.; Carroll, R. J.; Dabney, A. R. *Bioinformatics* **2012**, *28*, 1998−2003.

(20) Zhan, X.; Patterson, A. D.; Ghosh, D. *BMC Bioinf.* **2015**, *16*, 77.

(21) Taylor, S. L.; Ruhaak, L. R.; Weiss, R. H.; Kelly, K.; Kim, K. *Bioinformatics* **2017**, *33*, 17−25.

(22) Hastie, T. J.; Tibshirani, R. J.; Friedman, J. H. *The Elements of Statistical Learning: Data Mining, Inference, And Prediction*; Springer: New York, 2011.

(23) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, *78*, 779−787.

(24) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinf.* **2010**, *11*, 395.

(25) Murakami, I.; Chaleckis, R.; Pluskal, T.; Ito, K.; Hori, K.; Ebe, M.; Yanagida, M.; Kondoh, H. *PLoS One* **2014**, *9*, e115359.

(26) Libiseller, G.; Dvorzak, M.; Kleb, U.; Gander, E.; Eisenberg, T.; Madeo, F.; Neumann, S.; Trausinger, G.; Sinner, F.; Pieber, T.; Magnes, C. *BMC Bioinf.* **2015**, *16*, 118.

(27) Smyth, G. K. *Stat. Appl. Genet. Mol. Biol.* **2004**, *3*, 1−26.

(28) Kammers, K.; Cole, R. N.; Tiengwe, C.; Ruczinski, I. *EuPa Open Proteomics* **2015**, *7*, 11−19.

(29) Newschaffer, C. J.; et al. *J. Neurodev. Disord.* **2012**, *4*, 7.

(30) Kleensang, A.; Vantangoli, M. M.; Odwin-DaCosta, S.; Andersen, M. E.; Boekelheide, K.; Bouhifd, M.; Fornace, A. J., Jr; Li, H.-H.; Livi, C. B.; Madnick, S.; Maertens, A.; Rosenberg, M.; Yager, J. D.; Zhaog, L.; Hartung, T. *Sci. Rep.* **2016**, *6*, 28994.

(31) Böttcher, C.; Westphal, L.; Schmotz, C.; Prade, E.; Scheel, D.; Glawischnig, E. *Plant Cell* **2009**, *21*, 1830−1845.

(32) Neumann, S.; Thum, A.; Böttcher, C. *Metabolomics* **2013**, *9*, 84−91.

(33) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84*, 283−289.

(34) Samino, S.; Vinaixa, M.; Díaz, M.; Beltran, A.; Rodríguez, M. A.; Mallol, R.; Heras, M.; Cabre, A.; Garcia, L.; Canela, N.; de Zegher, F.; Correig, X.; Ibáñez, L.; Yanes, O. *Sci. Rep.* **2015**, *5*, 11496.

(35) Capellades, J.; Navarro, M.; Samino, S.; Garcia-Ramirez, M.; Hernandez, C.; Simo, R.; Vinaixa, M.; Yanes, O. *Anal. Chem.* **2016**, *88*, 621−628.

(36) Treviño, V.; Yañez-Garza, I.-L.; Rodriguez-López, C. E.; Urrea-López, R.; Garza-Rodriguez, M.-L.; Barrera-Saldaña, H.-A.; Tamez-Peña, J. G.; Winkler, R.; Díaz de-la Garza, R.-I. *J. Mass Spectrom.* **2015**, *50*, 165−174.

(37) Duong, T. *J. Stat. Softw.* **2007**, *21*, 1−16.

(38) Wand, M. P. *J. Comput. Graph. Stat.* **1994**, *3*, 433−445.

(39) Dowle, M.; Srinivasan, A.; Short, T.; Lianoglou, S.; Antonyan, E. *data.table: Extension of Data.frame; R package*, version 1.9.6; 2015

(40) Pau, G.; Fuchs, F.; Sklyar, O.; Boutros, M.; Huber, W. *Bioinformatics* **2010**, *26*, 979−981.