Supporting Information

One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks

Owen D. Myers,[†] Susan J. Sumner,[‡] Shuzhao Li,[¶] Stephen Barnes,[§] and Xiuxia Du[*,†]

[†]*University of North Carolina at Charlotte, Charlotte, NC 28223, USA*

[‡]*University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA*

[¶]*Emory University, Atlanta, GA 30322, USA*

[§]*University of Alabama at Birmingham, Birmingham, AL 35294, USA*

E-mail: xiuxia.du@uncc.edu

Phone: (704) 687-7307

9201 University City Blvd., Charlotte, NC 28223, USA

Contents

1 ADAP in MZmine 2

The most up to date ADAP algorithms for EIC construction and EIC peak picking can be found in the official MZmine 2 release at `http://mzmine.github.io/`. The algorithms are implemented in Java to be easily incorporated into MZmine 2 and benefit from Java's platform independence and other positive qualities. However, the ADAP algorithms were originally developed in C++ and were incorporated into MZmine 2 using Java Native Interface (JNI). Because we changed the implementation in a few simple ways when changing from the C++ version to Java, the results obtained with the current ADAP in MZmine 2 may different slightly from those in the paper. To allow users to exactly reproduce our results, **we have made the original version of the code used to produce these results avaliable** at `http://www.du-lab.org/publications.html`. This older version is not actively supported, unlike the version found in the current release of MZmine 2, but it should still be usable out of the box. The C++ library called through JNI was compiled on OS X El Capitan and so we can only guarantee that it runs on an apple machine.

2 Experimental Procedures

2.1 DCSM

DCSM is a standard mixture file that was generated from a mixture of 21 standard compounds. Section "Compounds Manually Confirmed in the DCSM File" of the Supporting information lists these 21 compounds.

The standard compounds were purchased from Sigma Aldrich (St. Louis, MO). HPLC grade water (Sigma Aldrich, St. Louis, MO), dimethyl solfoxide (DMSO, Sigma Aldrich St. Louis, MO), and methanol (Sigma Aldrich St. Louis, MO) were used in dissolving standards. Mobile phase A: Water with 0.1% formic acid (Sigma Aldrich). Mobile phase B: Methanol with 0.1% formic acid (Sigma Aldrich, St. Louis, MO).

Stock solutions of the metabolite standards were prepared by dissolving each compound (1mg/mL) in DMSO. A working mixture of metabolites ($100\,\mu g/mL$ of each) was prepared in ethanol. The working mixture was further diluted in water to have a final concentration of $20\,\mu g/mL$.

A $10\,\mu L$ of standard mixture ($20\,\mu g/mL$ each standard concentration) was injected into an Orbitrap Velos mass spectrometer (Thermo Scientific, CA) coupled to an Acquity UPLC (Waters Corporation, MA). The metabolites were separated on a Waters Acquity HSS T3 column ($2.1\,mm \times 100\,mm$, $1.8\,\mu m$ particle size) operating at $50\,°C$ using a reverse phase chromatographic method. A gradient mobile phase consisting of water with 0.1% formic acid (A) and methanol with 0.1% formic acid (B) were used as previously described (Dunn, et al., 2011). All MS data were collected over $120 - 1000$ m/z in ESI positive ion mode.

2.2 YP01, YP02, and VT001

YP01, YP02, and VT001 were all generated from NIST Standard Reference Material (SRM) 1950, a representation of human plasma. Metabolomics analysis was performed similarly as previously described in Jin et al.[1] and Li et al.[2] Briefly, for each sample, $65\,\mu L$ of plasma was used and acetonitrile containing a mixture of 14 stable isotope internal standards was added to the aliquot at 2:1 in order to precipitate proteins. The samples were kept on ice for $30\,min$ and then centrifuged for $10\,min$ at $13,400$ rpm at $4\,°C$. The supernatant was then removed and placed into autosampler vials. Mass spectral data were collected with a $10\,min$ gradient on a Dionex UltiMate 3000 rapid separation LC system coupled with: 1) Thermo Q Exactive HF Orbitrap system (Thermo Fisher Scientific, San Diego, California) that generated the **VT001** data file. Ions were scanned in the m/z range from $85 - 1275$ in the positive ionization mode with a resolution of $120,000$; or 2) a LTQ Orbitrap Velos system, with m/z scan range $85 - 2000$, resolution $30,000$ for generating the **YP01**, **YP02** data files . The chromatography was performed using either a reverse phase C18 column (**YP02**) or an anion exchange column (**YP01**). Mass spectra were acquired in profile mode and converted to CDF format using the Xcalibur file converter software (ThermoFisher Scientific).

We manually checked for compounds certified by NIST as being in the samples. A certified compound is defined by NIST as "... a value for which NIST has the highest confidence of its accuracy in that all known or suspected sources of bias have been investigated or taken into account."[3] In Sections "Compounds Manually Confirmed in ..." of the Supporting Information we list the compounds found by manual investigation in all data files.

2.3 MAR17

This data file is part of the study named ST000045 in the Metabolomics Workbench.[4] Both the raw data file and the experimental information are available on the website. Briefly, plasma samples (200 µL) were thawed on ice at 4 °C followed by deproteinization with methanol (1:4 ratio of plasma to methanol) and vortexed for 10 s, followed by incubation at −20 °C for 2 h. The samples were then centrifuged at 15,871g for 30 min at 4 °C. The supernatants were lyophilized (Savant, Holbrook, NY) and stored at −20 °C prior to analysis. The samples were reconstituted in 50% H_2O/acetonitrile and passed through a Microcon YM3 filter (Millipore Corporation). The supernatants were transferred to analytical vials, stored in the autosampler at 4 °C, and analyzed within 48 h of reconstitution in buffer.

The liquid platform consisted of an Acquity UPLC system (Waters, Milford, MA). Plasma metabolite separation was achieved using hydrophilic interaction chromatography (HILIC). The run time was 20 min at a flow rate of 400 $\mu L\,min^{-1}$. The HILIC gradient was as follows: 0 min, 100% B; 1 min, 100% B; 5 min, 90% B; 13.0 min, 0% B; 16 min, 0% B; 16.5 min, 100% B; and 20 min, 100% B. Other LC parameters were injection volume 5 µL and column temperature 50 °C.

A 6220 TOF MS (Agilent Technologies) was operated in negative electrospray ionization modes using a scan range of 50 to 1,200 m/z. The mass accuracy and mass resolution were less than 5 (ppm) and $\sim$ 20,000, respectively. The instrument settings were as follows: nebulizer gas temperature 325 °C, capillary voltage 3.5 kV, capillary temperature 300 °C, fragmentor voltage 150 V, skimmer voltage 58 V, octapole voltage 250 V, cycle time 0.5 s, and run time 15.0 min.

3 ADAP Implementation of EIC Construction

One key aspect of the ADAP implementation is the way in which the search is performed to determine if a data point belongs to an EIC. Each EIC is initialized with the highest point, i.e. maximum intensity of that EIC. The tolerance range associated with that EIC will not change. We use a (Google's Guava) range object to represent that range and map it directly to the EIC it defines with a hash table. The range objects are stored in a tree (TreeRangeSet). Searching the tree can be done in $\mathcal{O}(\log n)$ time and accessing the EIC object from the range object is done in time $\mathcal{O}(1)$. To preserve the ordering of the ranges it takes $\mathcal{O}(\log n)$ time to add a new EIC. However, most of the operations performed in the building process are searching for the correct EIC to add a new data point to, so the extra cost associated with adding to the tree is acceptable.

4 Example Baseline Removal Creating False Positive

In the top panel of Figure S1 we show a large broad feature in the DCSM data set created by chemical noise. With the parameters shown in Section 12 for the DCSM data set this, peak is successfully *not* detected. After a baseline correction (details in next paragraph) is applied to the data this feature is distorted as shown in the bottom panel of Figure S1. The distortion makes this feature into something that appears very much like a real peak, colored in red, and is detected as a peak by the ADAP algorithm. There are many such examples in the DCSM data set but we only show one here.

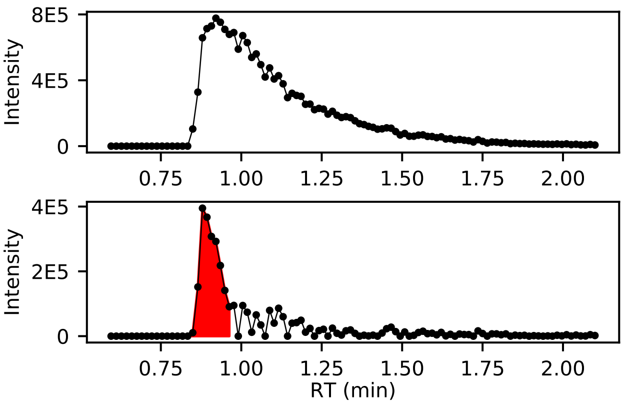

Figure S1: (*Top*) A broad feature in the raw data corresponding to chemical noise. (*Bottom*) The same feature shown after a baseline correction algorithm is applied to the data

The baseline correction used for the example shown in Figure S1 can be found in the MZmine 2 software package. We use the "Asymmetric baseline corrector" with the following parameters:

- *Chromatogram type* base peak intensity
- *MS level*: 1
- *Use m/z bins*: true
- *m/z bin width*: 0.01
- *Smoothing*: 2
- *Asymmetry*: 0.01

5 ADAP Determining Local Maxima in Wavelet Coefficients

Here we show an example of the first two iterations in the process for determining local maxima in the wavelet coefficients. In Figure S2 the top panel shows a portion of an example EIC. In the middle panel we show the wavelet coefficients created from the CWT of the above EIC at the fifth scale. For this scale, the fist step of the coefficient maxima detection process is to find the largest coefficient which is the maximum point of the red colored curve. The red curve itself shows the coefficients that will be removed before searching for the next maxima. The points in the red curve are determined from the point of the maxima out to ±2.5×scale in number of scans (rounded), in this case ±13 scans. The bottom panel of Figure S2 shows the coefficients after the points in red from the middle panel have been removed. The next maxima is the maximum point in the red curve of the bottom panel. The points that will be removed before the next maximum search are again colored red and are determined the same way.

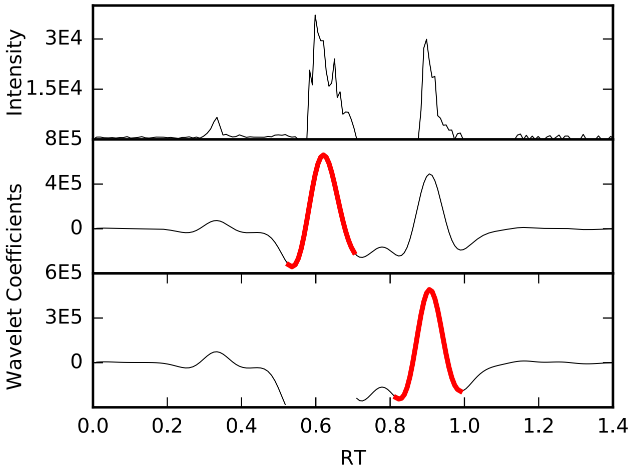

Figure S2: (*Top*) Example EIC. (*Middle*) Wavelet coefficients of the fifth scale from the CWT of the EIC shown in the *top* panel. The red portion of the curve denotes the current maximum in the coefficients as well as the points that will be removed before the next maximum search. (*Bottom*) The set of coefficients after the points from the previous maxima detection are removed. Red colors the next maximum and the next set of points to be removed.

6 Details Regarding Random Sampling and Sorting of Detected Peaks

A random sample of 400 peaks, with replacement, is chosen from each lobe of each Venn diagram and visually inspected. We look at each peak's shape and the boundaries defining the peak to sort the random sample of peaks into three categories. The first category is peaks that correspond possibly to real compounds, the second are peaks that clearly do not correspond to real compounds, and the third are peaks that can not easily be placed into the other two categories. From here on we refer to these three categories as good, bad, and uncertain in that order. More specifically, a peak will be considered good if it meets the following criteria:

(1) The boundaries of the peak appear to encapsulate the majority of the peak.

(2) The boundaries of the peak encapsulate the maximum of the peak.

(3) The peak is not immediately surrounded by peaks of similar intensity and shape which make the peak look like noise.

(4) The peak has a good shape or, if not, strongly meets criteria (3).

We recognize that these criteria and the choices based off of them are subjective. However, we believe the results presented here do not strongly depend on the person performing the categorization and have made all of the images from which the visual inspection was performed available for scrutiny at `http://www.du-lab.org/publications.html`. We must state that the person sorting the peaks was not blind from the methods used to detect them.

7 Important Details of Venn Diagram Counting

When we compare the detected peaks from the different algorithms we use a retention time range and an m/z range, creating a rectangle around each peak in the RT and m/z plane. If the rectangle corresponding to a peak detected by one algorithm overlaps a rectangle corresponding to a peak detected by a different algorithm they are considered the same. In the three way comparison three overlapping rectangles, one from each detection method, are counted as the same peak and contribute to the fully overlapping region of the Venn diagram. However, what would be found as two separate peaks in a two way comparison might be found to be “the same peak” when the third method is considered if the third rectangle overlaps the other two (that are themselves non-overlapping). There are two consequences. First, the ADAP/XCMS/MZmine 2 overlapping region, plus the MZmine 2/XCMS overlapping region will be slightly larger than the XCMS/MZmine 2 overlapping region if only XCMS and MZmine 2 are considered. Second, the ADAP/XCMS plus the XCMS region and the ADAP/MZmine 2 plus MZmine 2 region will be be smaller than the respective independent XCMS and MZmine2 regions if only XCMS and MZmine 2 are considered.

8 Comparing ppm and m/z

8.1 Example Calculation of ppm From Mass Spectrometry Data

In Figure 4 in the main text we show three examples of EICs and their corresponding m/z values. Here, for each example EIC shown in the figure we will find the largest difference in the m/z values (i.e. the mass range) in m/z and ppm. Figure 4's left panel has an m/z range of 0 in Dalton and ppm. Middle panel: The mass range is approximately 0.006, the representative m/z value is 128.038 (m/z of highest point), and the mass range in ppm is found to be

$$ppm = \frac{0.006}{128.038} \times 10^6 \approx 46.9. \tag{1}$$

Right panel: the mass range is approximately 0.014, the representative m/z value is 157.899, and the mass range in ppm is found to be

$$ppm = \frac{0.014}{157.899} \times 10^6 \approx 88.7. \tag{2}$$

8.2 Spectra Data Point Spacing

In Figure 6(A) in the main text we show a single mass spectrum's m/z sampling interval as a function of m/z for the Mar17 data file. In Figure S3 we show the MAR17 plot again (top) as well as the sampling interval m/z relationship for two other data files: YP01 and SCLC. Each scan used was the 200th scan from each file. For these data files the sampling interval increases with m/z, but the exact dependence of the interval on m/z depends on the instrument being used to collect the data. The instrument used to collect the MAR17 data was Agilent 6220 TOF-MS, the instrument used to collect YP01 was ThermoFisher Scientific LTQ Orbitrap Velos, the instrument used to collect the SCLC data was Waters Synapt-G2 Si.

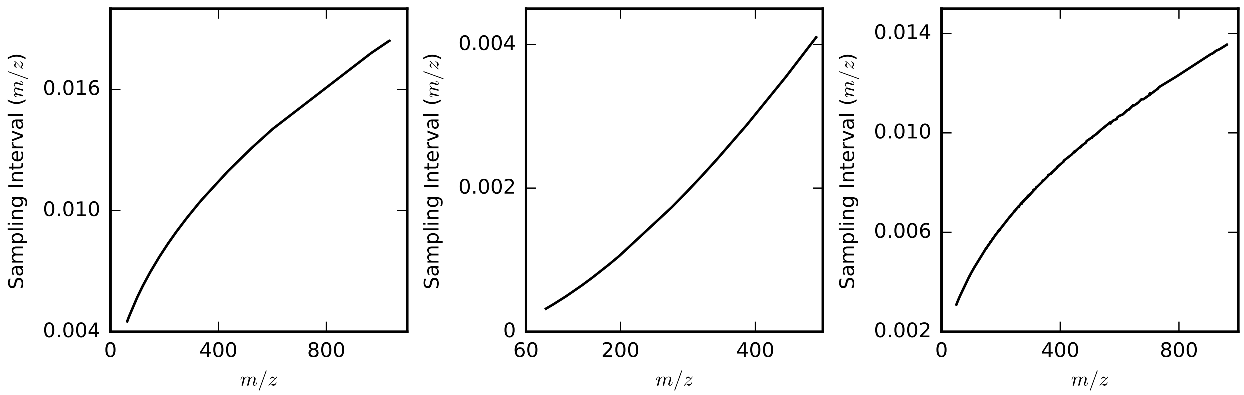

Figure S3: m/z sampling interval as a function of m/z from the 200th scan of data file (top) 17March10-11-r001, (middle) YP_111212_01, and (bottom) SCLC_16HV_1_POS_RP01.

9 Summary of Compound Detection Results

For compounds that have been manually confirmed to be present in the data files (including the isomers), Table S1 shows the number of compounds detected by ADAP, XCMS, and MZmine 2, respectively. Clearly, all three software packages are able to detect the majority of the known compounds.

Table S1: Number of known compounds detected by ADAP, XCMS, or MZmine 2.

Data File	Num. Known Compounds	Num. Found by ADAP	XCMS	MZmine 2
DCSM	23	22	22	22
YP01	19	19	19	19
YP02	18	16	16	17
VT001	14	14	13	14

10 Compounds Manually Confirmed in the DCSM File

Table S2 lists the standard compounds that have been manually confirmed to be present in the DCSM file that is generated from a standard mixture sample. For easy reference, we give each compound a short reference name.

Table S2: Compounds that have been manually confirmed to be present in the DCSM data file.

Compound	**Reference Name**
Bicine	c1
N-Acetylglutamine	c2
N-(Hydroxymethyl)nicotinamide	c3
Nicotinuric acid	c4
6-Hydroxynicotinic acid	c5
6-O-Methylguanine	c6
3-Chloro-L-tyrosine	c7
Acetazolamide	c8
Pantothenic acid	c9
L-Tryptophan	c10
4-Acetamidobenzaldehyde	c11
Glafenine	c12
3-(4-Fluorobenzoyl)propionic acid	c13
3-(2-Thienyl)acrylic acid (cis or trans)	c14
2,5-Dimethoxycinnamic acid (cis or trans)	c15
Bufexamac	c16
2-Amino-5-nitrobenzophenone	c17
Probenecid	c18
Dodecanedioic acid	c19
1,11-Undecanedicarboxylic acid	c20
DL-threo-1-Phenyl-2-palmitoylamino-3 -morpholino-1-propanol	c22

To determine if a compound is in the data file, we checked the mass error (i.e., the difference between the exact mass and the measured mass) and the similarity between the theoretical and experimental isotopic distribution. Table S3 lists the information known about each compound. XCMS was not able to detect 1,11-Undecanedicarboxylic acid. MZmine 2 was not able to detect Bufexamac. ADAP detected all.

Table S3: Data file DCSM compound information. Parentheses next to a reference name distinguish possible isomers. Abbreviated column names: Man. → Manually, Theo. → Theoretical, Mass Err. → Mass Error, Iso. Dist. → Isotopic Distribution.

Compound	**Man. Detected**	**Theo.** m/z	**Data** m/z	**Mass Err. (ppm)**	**RT**	**Iso. Dist.**
c1	✓	164.0923	164.0909	5.3	0.65	yes
c2	✓	189.0875	189.0853	6.3	1.17	yes
c3	✓	153.0664	153.0649	6.4	1.94	yes
c4	✓	181.0613	181.0596	5.2	2.31	yes
c5	✓	140.0348	140.0332	7.8	2.65	yes
c6	✓	166.0728	166.0712	6.1	2.94	yes
c7	✓	216.0427	216.0411	3.6	3.25	yes
c8	✓	222.9960	222.9946	2.7	3.90	yes
c9	✓	220.1185	220.1168	3.5	4.00	yes
c10	✓	205.0977	205.0958	4.5	4.26	yes
c11	✓	164.0712	164.0698	4.9	6.08	yes
c12(1)	✓	373.0955	373.0937	1.3	6.90	yes
c12(2)	✓	373.0955	373.0937	1.3	7.15	yes
c13	✓	197.0614	197.0600	3.7	7.97	yes
c14	✓	155.0167	155.0153	5.6	8.40	yes
c15	✓	209.0814	209.0796	4.1	9.68	yes
c16	✓	224.1287	224.1270	3.4	9.88	yes
c17	✓	243.0770	243.0756	2.3	10.83	yes
c18	✓	286.1113	286.1095	2.2	11.18	yes
c19	✓	231.1596	231.1580	3.1	11.54	yes
c20(1)	✓	245.1753	245.1741	2.0	11.56	yes
c20(2)	✓	245.1753	245.1741	2.0	12.28	yes
c22	✓	475.3900	475.3881	0.8	14.25	yes

11 Compounds Manually Confirmed in the YP01, YP02, and VT001 Files

Data files YP01, YP02, and VT001 were generated from NIST Standard Reference Material 1950 (frozen human plasma). The compounds manually found in YP01, YP02, and VT001 are displayed in Table S4. All compounds which were not found in any of the data files manually are omitted from the table entirely.

Table S4: Compounds that have been manually confirmed to be present in the YP01, YP02, and VT001 data files.

Files	YP01	YP02	VT001
Alanine	✓	✓	✓
Histidine	✓		
Isoleucine	✓	✓	
Leucine	✓	✓	
Lysine	✓	✓	✓
Methionine	✓	✓	✓
Proline	✓	✓	✓
Serine	✓	✓	✓
Threonine	✓	✓	✓
Tyrosine	✓	✓	
Valine	✓	✓	
Cysteine	✓		✓
Phenylalanine	✓	✓	✓
Creatinine	✓	✓	
Glucose	✓	✓	
Uric Acid	✓	✓	
Homocysteine	✓	✓	
Testosterone		✓	
Arginine		✓	✓
(Z,Z)-9,12-Octadecadienoic Acid			✓
Lutein			✓

Manually detected compounds in each data file which are missed by either XCMS or MZmine 2 are shown in Table S5. A blank space in the table means that all manually detected compounds for that data file were detected by that software package.

Tables S6, S7, and S8 show detailed information about compounds that have been manually

Table S5: Compounds from the respective data file that were not detected by the respective software package.

	YP01	**YP02**	**VT001**
XCMS		Glucose Progesterone	Cysteine
MZmine 2		Progesterone	

confirmed to be present in the YP01, YP02, and VT001 files, respectively.

Table S6: Data file YP01 compound information.

Compound	**Theo.** m/z	**Data** m/z	**Mass Err. (ppm)**	**RT**	**Iso. Dist.**
Alanine	90.0558	90.0541	21.0	0.90	yes
Histidine	156.0768	156.0706	25.7	0.91	no
Isoleucine	132.1028	132.1007	12.0	0.89	yes
Leucine	132.1028	132.1007	12.0	0.89	yes
Lysine (1)	147.1138	147.1119	10.8	0.68	no
Lysine (2)	147.1138	147.1119	10.8	0.87	no
Methionine	150.0588	150.0572	7.1	0.91	yes
Proline	116.0708	116.0696	9.3	0.92	yes
Serine	106.0508	106.0489	17.2	0.92	yes
Threonine	120.0658	120.0646	8.3	0.92	maybe
Tyrosine	182.0818	182.0793	7.6	0.90	yes
Valine	118.0868	118.0851	12.1	0.89	yes
Cysteine	122.0278	122.0215	42.8	0.95	maybe
Phenylalanine	166.0868	166.0848	7.3	0.90	yes
Creatinine	114.0668	114.0651	13.1	0.86	yes
Glucose	181.0708	181.0707	0.5	1.00	yes
Uric Acid	169.0358	169.0342	5.5	1.03	yes
Homocysteine	136.0428	136.0474	24.8	0.86	yes

Table S7: Data file YP02 compound information.

Compound	Theo. m/z	Data m/z	Mass Err. (ppm)	RT	Iso. Dist.
Alanine	90.0558	90.0541	21.0	0.85	no
Histidine	156.0768	156.0757	4.6	0.68	yes
Isoleucine	132.1028	132.1007	12.0	0.87	yes
Leucine	132.1028	132.1007	12.0	0.87	yes
Lysine	147.1138	147.1119	8.8	0.67	no
Methionine	150.0588	150.0572	7.1	0.86	no
Proline	116.0708	116.0696	9.3	0.87	yes
Serine	106.0508	106.0489	17.2	0.87	no
Threonine	120.0658	120.0646	8.3	0.85	no
Tyrosine	182.0818	182.0802	4.9	0.88	no
Valine	118.0868	118.0851	12.1	0.87	yes
Phenylalanine	166.0868	166.0848	7.3	0.88	yes
Creatinine	114.0668	114.0651	13.1	0.86	yes
Glucose	181.0708	181.0707	0.5	1.00	yes
Uric Acid	169.0358	169.0342	5.5	0.81	no
Homocysteine	136.0428	136.0468	21.5	0.72	no
Testosterone	289.2168	289.2190	2.6	1.72	no
Arginine	175.1198	175.1177	7.0	0.67	yes

Table S8: Data file VT001 compound information.

Compound	Theo. m/z	Data m/z	Mass Err. (ppm)	RT	Iso. Dist.
Alanine	90.0558	90.0551	8.8	3.95	yes
Lysine	147.1138	147.1131	3.2	5.36	yes
Methionine	150.0588	150.0587	0.5	3.86	yes
Proline	116.0708	116.0707	0.7	4.31	yes
Serine	106.0508	106.0501	6.8	4.05	no
Threonine	120.0658	120.0658	0.3	3.98	yes
Cysteine	122.0278	122.0227	34.2	5.30	no
Phenylalanine (1)	166.0868	166.0864	1.5	3.08	yes
Phenylalanine (2)	166.0868	166.0864	1.5	3.37	yes
Phenylalanine (3)	166.0868	166.0864	1.5	3.56	yes
Arginine	175.1198	175.1194	1.4	5.26	yes
(Z,Z)-9,12-Octadecadienoic Acid	281.2478	147.1131	0.2	0.75	no
Lutein	569.4358	569.4307	1.6	0.80	yes

12 Preprocessing Parameters used by ADAP, XCMS, and MZmine 2

ADAP in the MZmine 2 framework contains an option to output centroiding results to a CDF file allowing us to use the MZmine 2 wavelet centroiding for all files. In table S9 we show the MZmine 2 centroiding parameters used for all data files.

Table S9: Parameters used for the peak detection in the mass spectrometry data.

Parameters	**DCSM**,**YP01**,**YP02** and **VT001**
Mass Detection	
Noise level	100
Scale level	20
Wavelet window size	30

Up to now we have mentioned user define values in the context of their role in the ADAP algorithms, but now we explicitly state where in the ADAP program a user will find the parameters along with a brief description of each of the parameters. *ADAP EIC building* can be found in the *Raw data methods* menu. Parameters:

- *Min group size in # of scans*: The entire EIC must have at least this number of sequential scans with points above the *Group intensity threshold* set by the user.
- *Group intensity threshold*: See above.
- *Min highest intensity*: There must be at least one point in the EIC that has an intensity greater than or equal to this value.
- m/z *tolerance*: The tolerance used in constructing the EICs, ϵ. Twice the *m/z tolerance* set by the user is the maximum width of an EIC. This value can be set in Dalton or ppm, whichever value is set to 0.0 will not be used.

ADAP EIC chromatographic peak picking can be found in the *Peak list methods* menu in the *Chromatogram deconvolution* option. Select *Wavelets (ADAP)* in the drop down box.

- *S/N threshold*: The minimum signal to noise ratio a peak must have to be considered a real feature.
- *Min feature height*: The smallest intensity a peak can have and be considered a real feature.

- *Coefficient/area threshold*: The best coefficient divided by the area under the curve of the feature, $C_{\max}/A$.

Table S10: ADAP parameters used for EIC building and peak picking

Parameters	**DCSM**	**YP01** and **YP02**	**VT001**
ADAP chromatogram builder			
Min group size	4	5	4
Group intensity threshold	500	500	500
Min highest intensity	5000	1000	1000
m/z *tolerance*	0.01	0.01	0.01
Chromatogram peak picking			
S/N *threshold*	10	10	10
Min feature height	5000	1000	1000
Coefficient/area threshold	120	130	300
Peak duration range	0.02-0.60	0.05-1.00	0.01-0.5

Table S11: XCMS parameters used for EIC building and peak picking

Parameters	**DCSM**	**YP01** and **YP02**	**VT001**
method	centWave	centWave	centWave
mzTolerance	0.01	0.01	0.01
peakwidth	(1.2,36.0)	(1.2,60.0)	(0.6,30.0)
snthresh	10	10	10
prefilter	(1,5000)	(1,1000)	(1,1000)
mzCenterFun	wMean	wMean	wMean
integrate	2	2	2
mzdiff	-0.001	-0.001	-0.001
fitgauss	F	F	F
noise	100	100	100
sleep	0	0	0

Table S12: MZmine2 parameters used for EIC building and peak picking

Parameters	**DCSM**	**YP01** and **YP02**	**VT001**
Chromatogram builder			
Min time span	0.04	0.05	0.01
Min height	5000	1000	1000
m/z *tolerance*	0.01	0.01	0.01
Chromatogram deconvolution			
S/N *threshold*	10	10	10
Wavelet scales	0.02-0.80	0.02-1.20	0.01-0.5
Peak duration range	0.02-0.60	0.05-1.00	0.01-0.5

13 Parameters used by ADAP for Constructing EICs and Detecting EIC Peaks from File MAR17

Table S13: parameters used for the peak detection MAR17

Centroiding:	
Noise level	*Scale level*
100	10
Wavelet window size	
30	
ADAP EIC builder:	
Min group size	*Group intensity threshold*
5	100
Min highest intensity	m/z *tolerance*
5000	0.02
Chromatogram peak picking:	
S/N *threshold*	*Min feature height*
10	5000
Coef./area threshold	*Peak duration range*
350	0.02-0.6

References

(1) Jin, R.; Banton, S.; Tran, V. T.; Konomi, J. V.; Li, S.; Jones, D. P.; Vos, M. B. *The Journal of Pediatrics* **2016**, *172*, 14 – 19.e5.

(2) Li, S.; Park, Y.; Duraisingham, S.; Strobel, F. H.; Khan, N.; Soltow, Q. A.; Jones, D. P.; Pulendran, B. *PLOS Computational Biology* **2013**, *9*, 1–11.

(3) NIST SRM 1950. https://www-s.nist.gov/srmors/certificates/1950.pdf, [Accessed January 30, 2017].

(4) http://www.metabolomicsworkbench.org/.