Supporting Information

A time-saving design of experiment protocol for optimization of LC-MS data processing in metabolomic approaches

Abstract

In supporting information, we describe a detailed protocol for optimization of LC-MS-based metabolomics data processing by using XCMS or any other software (e.g. MZmine), and six tables and four figures are provided for supporting our optimization study.

Table of contents

Protocol of optimization

Materials

Reagents

- Deionized (18.2 MΩ) filtered (0.22 μm) water
- Acetonitrile (LC-MS CHROMASOLV, FLUKA Sigma-Aldrich, cat. no. 34967)

!CAUTION Acetonitrile is harmful and volatile and should be handled in a fume hood.

- Acetic acid (LC-MS CHROMASOLV, FLUKA Sigma-Aldrich, cat. no. 49199)

Samples

- Human urine **#COMMENT** Other samples (e.g. plant and animal tissues, blood plasma and serum, feces, and others) can be used, which depends on your studies.

Equipment

- Agilent 1200 series capillary HPLC (Agilent, CA, USA) **#COMMENT** Other systems can be used for sample analysis.
- MicrOTOF-Q MS (Bruker Daltonics, Bremen, Germany) **#COMMENT** Other systems can be used for sample analysis.
- ZORBAX 300SB C18 column (0.5 × 150 mm, 3.5 μm, Agilent Technologies, Waldbronn, Germany) **#COMMENT** Other columns can be used for LC separation.
- CompassXport software (Bruker Daltonics, Bremen, Germany) **#COMMENT** Other software can be used for data format transformation.
- R software (v.2.15.2; http://www.r-project.org/) **#COMMENT** XCMS software is performed under R environment.
- XCMS software (v.1.34.0; http://www.bioconductor.org/) **#COMMENT** Other software can be used and optimized for LC-MS data processing.

- Microsoft Excel (Microsoft Corp., Redmond, WA) **#COMMENT** Other software can be used for correlation analysis.
- SAS 9.2 software (SAS Institute Inc, Cary, NC) **#COMMENT** Other software can be used for design of experiment, such as Design-Expert, MODDE, STATISTICA, and others.

Procedure

- Preparation of dilution series **#COMMENT** Dilute samples with solvent, but the number of dilution series is optional (n>3). In this study, 5 levels were prepared, requiring in total 20 min.
- LC-MS analysis **#COMMENT** 35 min per sample, but it depends on LC-MS methods.
- Screening design **#COMMENT** 1. Choose parameters which you want to optimize and set their ranges for optimization; 2. Determine the responses, which should be a good criteria for evaluating data quality; 3. Screening design (e.g. full factorial, fractional fractorial, Plackett-Burman, D-optimal designs and the gradient search based on local main-effect designs). Factors, responses and design methods are optional according to the purpose of the study. In our study, in total 17 parameters were designed to optimize, and default parameters were set as the base level and then the range of parameters increased and decreased from the base level were used for screening design. The reliability index as a response was applied in Plackett-Burman design. A total of 4 h was required.
- XCMS analysis **#COMMENT** Use the parameter setting designed by screening design to analyze LC-MS data and output peak tables. A total of 10 h was required, but this step can be automatic.
- Correlation analysis **#COMMENT** The linearity (*r value*) between peak areas from dilution series and the dilution time of samples were calculated by using data analysis in Microsoft

Excel, and then peaks can be graded according to *r values*. Here, peaks *with* $r > 0.9$ and $r < 0.1$ were classified as reliable and unreliable peaks (RP and UP), respectively, and the reliability index was calculated as RP^2/UP. The threshold of *r value* is optional. A total of 5 h was required.

- Identification of significant parameters **#COMMENT** Model factors and responses with screening designs and identify by ANOVA the factors that significantly affect the response. The significance level is optional, and $P<0.05$ was used in our optimization. A total of 1 h was required.
- Optimization design **#COMMENT** Set the range of significant parameters and design experiments for optimizing parameters. Design methods are optional such as central composite, Box-Behnken and D-Optimal designs. In this study, central composite design was used. A total of 30 min was required.
- XCMS analysis **#COMMENT** Use the parameter setting designed by optimization design to analyze LC-MS data and output peak tables. A total of 6 h was required, but this step can be automatic.
- Correlation analysis **#COMMENT** As described above. A total of 1.5 h was required.
- Optimization of significant parameters **#COMMENT** Model significant factors and responses with optimization designs and discover factor values that are needed to achieve a desired response. A total of 1 h was required.
- Calculation of B/S values **#COMMENT** 1. Blank and sample data files were simultaneously processed by XCMS with optimal parameters; 2. The ratio of peak areas of blanks to samples (B/S) was calculated for each peak. A total of 30 min was required.

- Optimization design **#COMMENT** Set the range of B/S and peak intensity and design experiments for optimization. Design methods are optional such as central composite, Box-Behnken and D-Optimal designs. In this study, central composite design was used. A total of 30 min was required.
- Calculation of data quality parameters **#COMMENT** Calculate the reliability index and count the number of peaks under different levels of B/S and peak intensity, and a total of 1.5 h was required.
- Optimization of B/S and peak intensity **#COMMENT** By maximizing the reliability index and minimizing the loss of peaks, the threshold of intensity and B/S was optimized by optimization design. A total of 30 min was required.
- Peaks with higher B/S and lower intensity than their optimal values were removed from data sets. **#COMMENT** Optimized peak tables for further analysis (e.g. PCA, PLS-DA, OPLS-DA, and others).

Table S-1. List of chemicals used in this study.

Chemical	Purity (%)	Source	Chemical	Purity (%)	Source
(−)-Epicatechin	≥90	Sigma	L-Histidine	≥99	Sigma-Aldrich
Inosine	≥99	Sigma	L-Leucine	≥98	Sigma
2-Hydroxycinnamic acid	97	Aldrich	L-Lysine	≥98	Sigma
2-Oxoglutaric acid	≥99.0	Fluka	L-Methionine	≥98	Sigma-Aldrich
2-Aminobutyric acid	≥99.0	Fluka	L-Proline	≥99	Sigma-Aldrich
Allantoin	≥98	Sigma	L-Rhamnose	≥99	Sigma
Catechin	≥98	Sigma	L-Serine	≥99	Sigma
D-(−)-Fructose	≥99	Sigma	L-Threonine	≥98	Sigma-Aldrich
D-(+)-Galactose	≥99	Sigma-Aldrich	L-Tryptophan	≥98	Sigma-Aldrich
δ-Gluconolactone	≥99.0	Sigma	L-Tyrosine	≥98	Sigma-Aldrich
D-(+)-Glucose	≥99.5	Sigma	L-Valine	≥98	Sigma-Aldrich
D-(+)-Xylose	≥99	Sigma-Aldrich	L-Phenylalanine	≥98	Sigma-Aldrich
D-Glucurone	≥99	Sigma	Maltose	≥99	Sigma-Aldrich
D-Glucuronic acid	≥99.5	Sigma	Maltotriose	≥95	Sigma
DL-Malic acid	≥99	Aldrich	Mannose	≥99	Sigma
D-Mannitol	≥98	Sigma-Aldrich	Myo-Inositol	≥99	Sigma
Dulcitol	≥99	Sigma	Protocatechuic acid	97	Aldrich
Fumaric acid	≥99	Aldrich	Salicylic acid	≥99	Aldrich
Guanosine	≥98	Sigma	Sorbitol	≥98	Sigma
L-(+)-Arabinose	≥99	Sigma	Sucrose	≥99.5	Sigma
L-Alanine	≥98	Sigma	Syringic acid	≥95	Sigma
L-Arginine	≥98	Sigma-Aldrich	Umbelliferone	99	Aldrich
L-Asparagine	≥98	Sigma	Uracil	≥99.0	Fluka
L-Cysteine	97	Aldrich	Uridine	≥99	Fluka
L-Glutamine	≥99	Sigma	Naringin	≥95	Spectrum Chemicals & Laboratory Products

Table S-2. The correlation of reliable peaks (RP) [a], unreliable peaks (UP) [b], total peaks (TP) and related parameters.

	RP	UP	TP	RP/TP	UP/TP	RP/UP
UP	0.88					
TP	0.93	0.99				
RP/TP	-0.61	-0.77	-0.77			
UP/TP	0.45	0.69	0.63	-0.82		
RP/UP	-0.51	-0.67	-0.65	0.92	-0.93	
RP^2/UP	0.58	0.19	0.30	0.14	-0.38	0.28

[a] The peak with a high correlation ($r > 0.9$) between peak area and concentration of dilution series; [b] The peak with a low correlation ($r < 0.1$) between peak area and concentration of dilution series.

Table S-3. The default and Plackett-Burman design parameters of XCMS software.

Parameters	Default	Design I	Design II
Filter and Identify Peaks – centWave			
ppm	25	5→25	25→100
peakwidth_min	20	5→20	20→45
peakwidth_max	50	25→50	50→100
snthresh	10	1→10	10→50
prefilter_k	3	1→3	3→10
prefilter_I	100	5→100	100→500
mzCenterFun	wMean	wMean→wMeanApex3	wMean→apex
integrate	1	1→2	1→2
mzdiff	-0.001	-0.01→-0.001	-0.001→-0.0001
fitguass	FALSE	FALSE→TRUE	FALSE→TRUE
Match Peaks Across Samples			
bw	30	5→30	30→100
minfrac	0.5	0.1→0.5	0.5→1.0
mzwid	0.25	0.05→0.25	0.25→1.0
max	50	5→50	50→100
Retention Time Correction			
smooth	linear	linear→loess	linear→loess
span	0.2	0.05→0.2	0.2→1.0
family	symmetric	symmetric→gaussian	symmetric→gaussian

Table S-4. Plackett-Burman design of XCMS optimization.

Run	X1[a]	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17
1	1[b]	-1[c]	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1
2	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1
3	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1
4	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1
5	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	-1	1
6	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1	-1
7	1	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1	1
8	1	1	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1	-1
9	-1	1	1	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1	1
10	1	-1	1	1	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1	-1
11	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	1	1	-1	-1	-1
12	1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	1	1	-1	-1
13	-1	1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	1	1	-1
14	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	1	1
15	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1	1
16	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1	1	1	-1
17	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1	1	1
18	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1	1
19	-1	1	1	-1	-1	-1	-1	1	-1	1	-1	1	1	1	1	-1	-1
20	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1

[a] X1, ppm; X2, peakwidth_min; X3, peakwidth_max; X4, snthresh; X5, prefilter_k; X6, prefilter_I; X7, mzCenterFun; X8, integrate; X9, mzdiff; X10, fitguass; X11, bw; X12, minfrac; X13, mzwid; X14, max; X15, smooth; X16, span; X17, family; [b] default parameters; [c] designed parameters.

Table S-5. Optimal XCMS settings and threshold values of synthetic and urine samples.

Parameters	Optimal setting of synthetic sample		Optimal setting of urine sample	
Filter and Identify Peaks – centWave				
ppm	25		25	
peakwidth_min	20		20	
peakwidth_max	82.5		50	
snthresh	10		10	
prefilter_k	3		3	
prefilter_I	100		100	
mzCenterFun	wMean		wMean	
integrate	1		1	
mzdiff	-0.001		-0.001	
fitguass	FALSE		FALSE	
Match Peaks Across Samples				
bw	8.25		8.25	
minfrac	0.5		0.39	
mzwid	0.25		0.25	
max	50		50	
Retention Time Correction				
smooth	loess		loess	
span	0.2		0.2	
family	gaussian		gaussian	
Threshold Values				
	Default XCMS	Optimal XCMS	Default XCMS	Optimal XCMS
B/S	0.75	0.25	0.25	0.25
Intensity	3250	3250	7750	7750

Table S-6. Results of synthetic samples analyzed by different methods.

		Reliability index	Group index	Reliable peaks	Unreliable peaks	Total peaks	Identified percentage[a]
Default Setting		735.0	44.3	443	267	883	14.0
Threshold Method	P[b]	4782.9	52.5	305	-[d]	429	-
	V[c]	3513.5	54.9	308	27	386	26.2
Optimal Setting	P	948.5	50.3	-	-	-	-
	V	877.4	53.4	493	277	964	16.6
Optimal+Threshold[e]	P	12400.5	69.5	330	-	409	-
	V	6960.9	68.6	344	17	407	35.9

[a] The percentage of identified peaks in total peaks; [b] Prediction of CCD models; [c] Validation of CCD models; [d] Non-prediction; [e] The approach combining optimal parameter setting and threshold method.

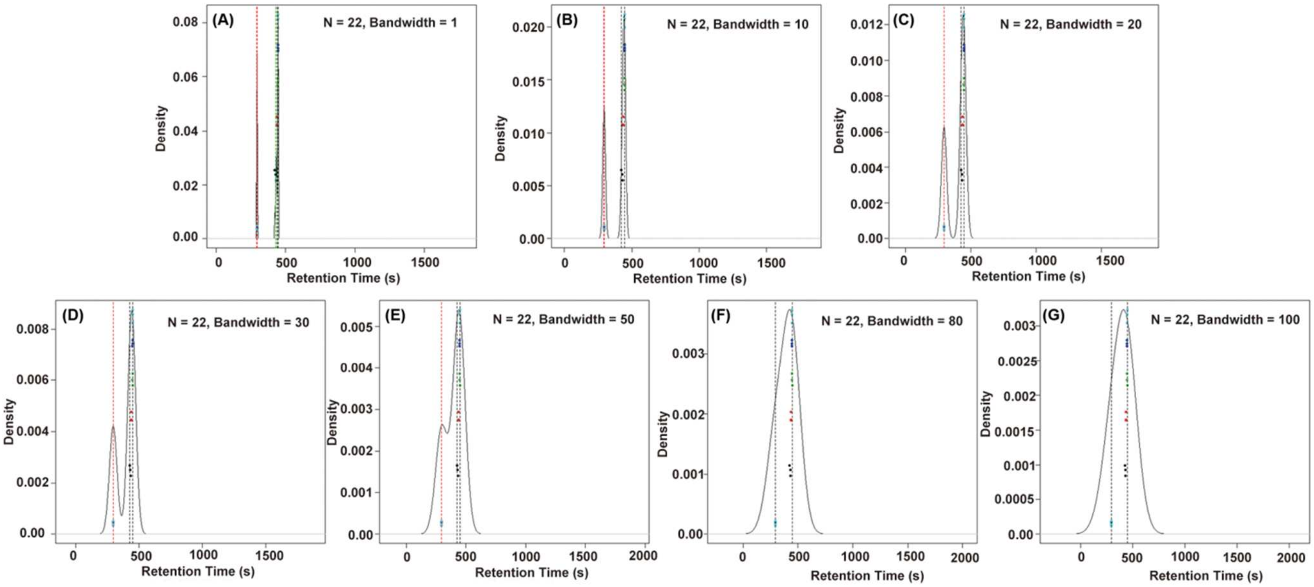

Figure S-1. Effect of parameter ‘bw’ on the peaks group during XCMS processing: A, bw = 1; B, bw = 10; C, bw = 20; D, bw = 30; E, bw = 50; F, bw = 80; G, bw = 100.

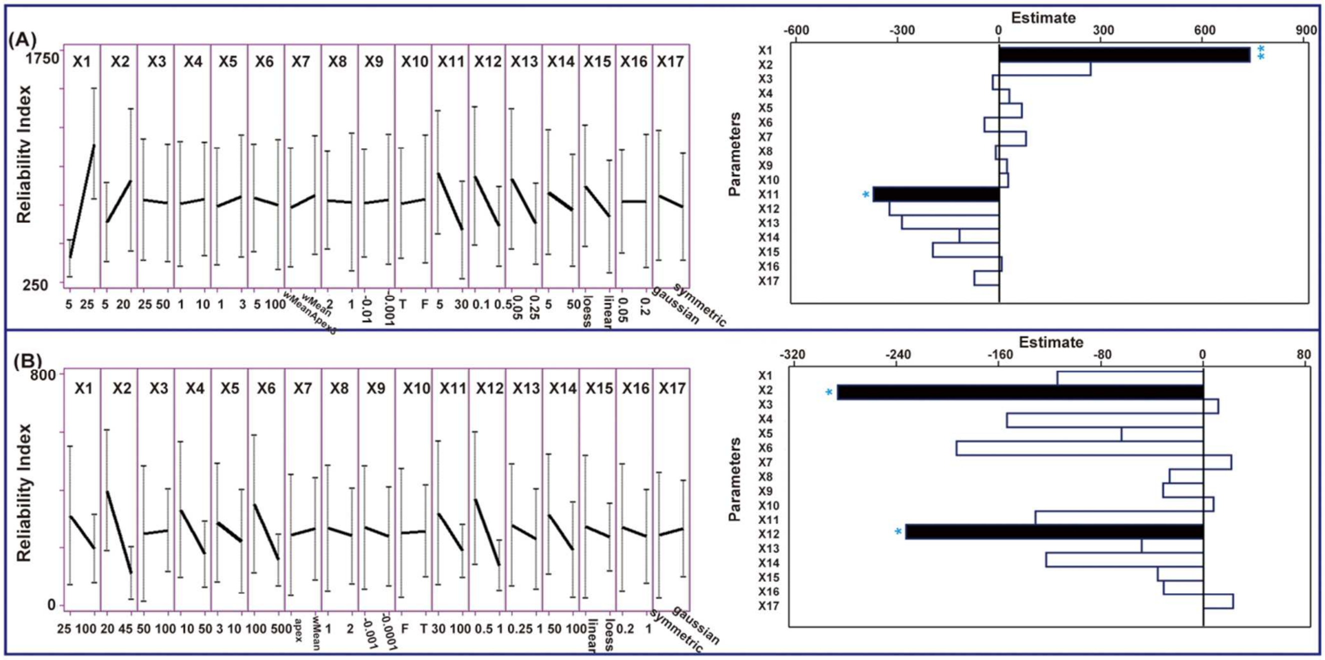

Figure S-2. The main effect and regression coefficient of 17 parameters in Plackett-Burman design (A, Design I; B, Design II): X1, ppm; X2, peakwidth_min; X3, peakwidth_max; X4, snthresh; X5, prefilter_k; X6, prefilter_I; X7, mzCenterFun; X8, integrate; X9, mzdiff; X10, fitguass; X11, bw; X12, minfrac; X13, mzwid; X14, max; X15, smooth; X16, span; X17, family. The black bars represent significant parameters selected by ANOVA (*, P < 0.05; **, P < 0.01).

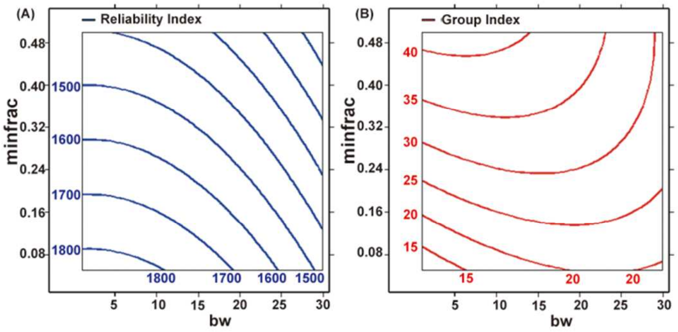

Figure S-3. The effect of ‘bw’ and ‘minfrac’ on reliability index (A) and group index (B).

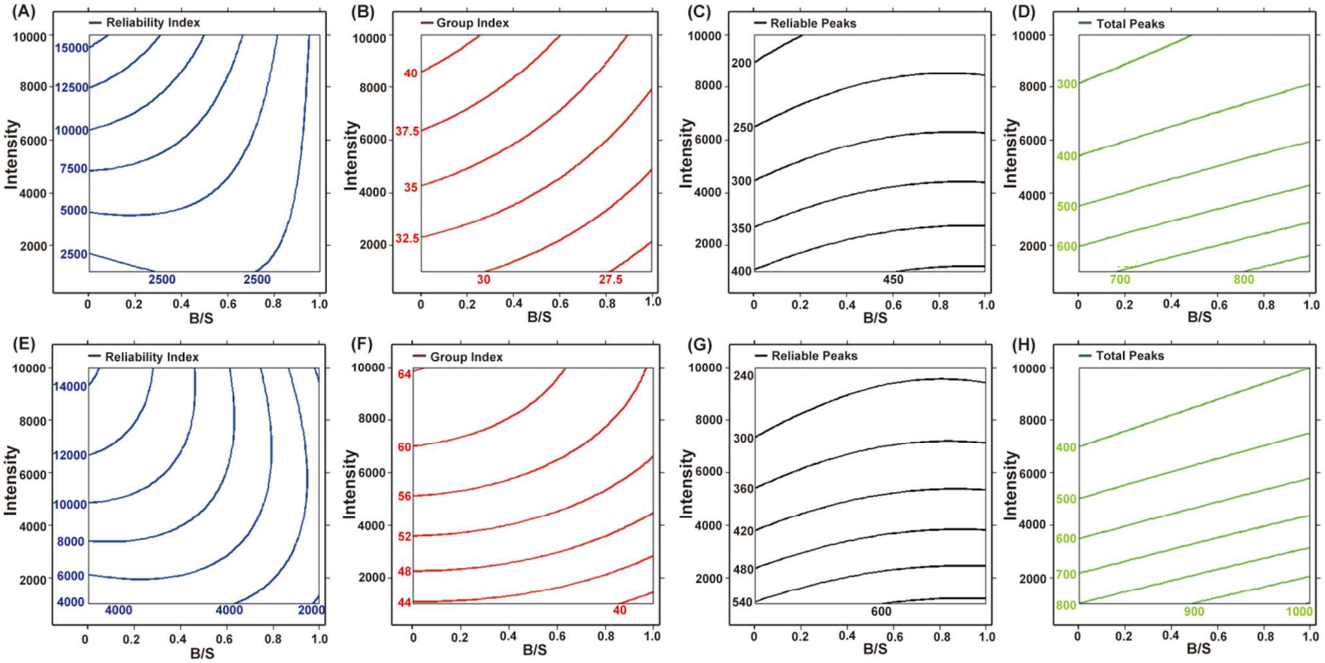

Figure S-4. The effect of peak intensity and B/S value on reliability index, group index, reliable peaks and total peaks in the peaks table produced by XCMS with default (A, B, C and D) and optimal (E, F, G and H) settings.