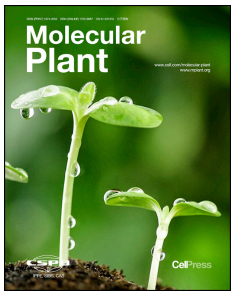
Accepted Manuscript



Discrimination and quantification of true biological signals in LC-MS-based metabolomics analysis

Lixin Duan, István Molnár, John Hugh Snyder, Guo-an Shen, Xiaoquan Qi

PII: S1674-2052(16)30065-X
DOI: 10.1016/j.molp.2016.05.009
Reference: MOLP 304

To appear in: *MOLECULAR PLANT*
Accepted Date: 17 May 2016

Please cite this article as: **Duan L., Molnár I., Snyder J.H., Shen G.-a., and Qi X.** (2016). Discrimination and quantification of true biological signals in LC-MS-based metabolomics analysis. Mol. Plant. doi: 10.1016/j.molp.2016.05.009.

Discrimination and quantification of true biological signals in LC-MS-based metabolomics analysis

Running title: Filtering rules and relative calibration models for metabolomics

Lixin Duan[1], István Molnár[2],John Hugh Snyder[1], Guo-an Shen[1] and Xiaoquan Qi[1]*

[1]The Key Laboratory of Plant Molecular Physiology, Institute of Botany, Chinese Academy of Sciences, 20 Nanxincun, Xiangshan Road, Beijing 100093, China

[2] Natural Product Center, School of Natural Resources and the Environment, The University of Arizona, 250 East Valencia RD., Tucson, Arizona 85706, United States

***To whom correspondence should be addressed. E-mail: xqi@ibcas.ac.cn, tel. +86-10-6283 6671, fax +86-10-8259 9701**

Dear Editor,

Metabolomics is a rapidly emerging field of post-genomic research that aims to comprehensively analyze all metabolites in biological samples. Potential biomarkers that distinguish prostate cancer samples were successfully identified through metabolomics analysis (Sreekumar et al., 2009). Metabolome quantitative trait loci (mQTL) and genome-wide association studies coupled with metabolomics analysis (mGWAS) also became efficient tools to decipher the genetic basis of complex metabolic traits in large populations (Gong et al., 2013; Chen et al., 2014).

Liquid chromatography-mass spectrometry (LC-MS) techniques widely used for metabolomics analysis allow for the highly sensitive, high throughput detection of thousands of metabolites (Chen et al., 2013). However, LC-MS unavoidably yields a large number of false positive signals mixed with true biological signals (Yu et al., 2013; Broadhurst and Kell, 2007). Without the means to confidently discriminate and evaluate the detected signals, biomarker discovery turns out to be misleading, or downright impossible. Metabolomics is routinely used to compare the relative concentrations of metabolites amongst different samples. For this, the monitored signals must fall within the quantitative dynamic range and show a good quantitative correlation with the amounts of the compounds of interest. Thus, it is necessary to systematically evaluate the quantitative performance of each peak. Isotope-labeling of internal standards or the whole metabolome (Giavalisco, 2009) are powerful tools to improve compound annotation and relative quantification. Standard mixtures (Phinney et al., 2013) can also be used to quantitatively analyze a selected set of known metabolites. Artificial biological gradients (Redestig et al., 2011) may allow the exploration of matrix effects and quantification performance. Here, we report a novel strategy for a comprehensive LC-MS-based metabolomics analysis that enables the unambiguous and facile discrimination of biological and non-biological signals, and improves the quantification accuracy of metabolites without labeling or other specialized techniques.

For a LC-MS-based metabolomics experiment, we prepare a blank sample, a Quality Control mixture (QC_mix) that combines all samples in equal proportions, and a dilution series of the QC_mix (Figure 1A). First, three to six replicates of the blank sample are

analyzed to balance the instrument. This also precludes contaminating the system with the biological samples. Next, the QC_mix is analyzed in six or more technical replicates to identify peaks that can be reproducibly detected. Third, the dilution series of the QC_mix is analyzed, proceeding from the most dilute (16 times dilution, DS_1/16x) to the most concentrated (two times concentration, DS_2X) (Figure 1A). Finally, the individual samples are analyzed separately in a random order. This data acquisition pipeline enables the discernment of true biological signals, and the building of calibration curves for the quantification of all metabolites, including unknown peaks.

Figure 1B illustrates the principles of distinguishing signals derived from true metabolites from those that have a non-biological origin. Peak 3 is absent from the blank and its signal intensity displays a good quantitative response in the dilution range. Such peaks are considered to originate from the biological source and have good quantitative performance. Peak 2 is detected in the blank sample and its signal intensity is independent of the dilution. Such peaks may derive from the chromatography solvent, or from column contaminants. Peak 1 is also present in the blank sample, but its signal intensity is dependent on the dilution. Such peaks may represent contaminations introduced during biological extract preparation, or impurities from labware (Supplemental Table S1).

We developed a hierarchical five-step filtering approach which applies these principles to comprehensive metabolomics experiments that often present thousands of peaks (Supplemental Results). To validate this LC-MS-based metabolomics strategy, we prepared and analyzed two groups of artificial samples, each including 20 standard compounds. In total, 1,342 peaks were enumerated from these artificial samples after standard peak extraction and alignment (Figure 1C). Step 1, the reproducibility check, eliminated 1,053 peaks as these were detectable less than five times in the six replicates of the QC_mix. Step 2, the variation check, filtered out a further 40 peaks since their relative standard deviation (RSD) was >20% in the six QC_mix replicates. Another 104 peaks failed to satisfy Step 3, the blank check, since they showed a peak area ratio of blank to sample ($Ratio_{B/S}$) of > 1%. Step 4, the response check, eliminated an additional 28 peaks because these had an unsatisfactory quantitative correlation (r<0.9) in the QC_mix dilution series. Finally, re-extractionand manual inspection of the peaks with an r between 0.9 and 0.99 (Step 5)

disqualified a further 17 signals, resulting in a final set of 102 peaks. Remarkably, all 20 standard compounds in the artificial samples passed this strict, hierarchical filtering process, while 92.4% of the peaks were eliminated as false positives or peaks with insufficient quantitative performance. Of the final set of 102 filtered peaks, 53 were identified to derive from the 20 standard compounds and their adductor fragment ions. A further 36 peaks were de-replicated to 21 unknown metabolites. However, these peaks were also deduced to originate from the 20 standards, because their concentration ratios were very similar to those of the standards in the artificial samples(Supplemental Figure S5, Supplemental information and Additional supplementary Data 1). These unknown compounds may be minor impurities present in the standards. Only 13 peaks assigned to 11 compounds had an unknown origin.

In a biological sample, the absolute concentration of most compounds represented by a peak will remain necessarily unknown due to the absence of standards. To overcome this limitation, our strategy introduces a relative concentration index (RCI) as an arbitrary concentration. To calculate the RCI, we build a relative quantitative model for the calibration of each analyte in the QC_mix dilution series, using a custom Python script (Supplemental Methods). We assume that the RCI of any compound is 3,200, 1,600, 800, 400, 200, and 100 arbitrary units in a 2x, 1x,1/2x, 1/4x, 1/8x, and 1/16x sample dilution, respectively. Using these assumptions, the coumarin peak, for example, yielded a relative calibration model of y = 435.07x + 24,301 for the dilution series (Supplemental Figure S1A). Since the peak area of coumarin in one of the artificial samples was 672,060 units, its RCI in that sample was calculated to be 1,489 arbitrary units.

Among the 102 filtered peaks in our validation experiment, 72.0% fitted a linear model, 21.8% fitted a binomial model, and 3.4% fitted a logarithmic model. A combination of a linear and a binomial model was manually built for one particular peak (Supplemental Figure S1D). To evaluate the quantitative accuracy of our strategy, we compared the RSDs of the ratios of the concentrations to peak areas on one hand and to RCIs on the other (Supplemental Table S2). Indeed, for 75% of the standards (15 out of 20), the quantification was more reliable when using the RCI. In addition, because all peaks are calibrated using the same scale range, we can use RCI instead of the peak area to compare changes amongst

various samples in a more precise manner (Supplemental Results). Moreover, by following our metabolomics analysis strategy, metabolite peaks that pass the five-step filtering process and quantified by RCI can be used as a targeted metabolomics dataset. A metabolite report is also generated displaying the final metabolite list, with each filtered peak annotated with the unique *m/z*, the retention time and other evaluation parameters (Supplemental Figure S2B).

We further applied our strategy for the analysis of the metabolomes of the seeds of two typical rice cultivars, 9311 (*Oryza sativa* L. ssp. *indica*) and Nipponbare (*O. sativa* L. ssp. *japonica*). A total of 2,162 peaks were enumerated from all samples, but 71.3% of these initial peaks were eliminated using the five-step filtering approach (Supplemental Figure S3). Principal component analysis (PCA) of the variations between the two rice samples yielded a model with both a better explanatory performance and a higher predictive power for our strategy, compared to that for the traditional, non-targeted metabolomics analysis method (Figure 1D, F and G, and Supplemental Figure S4).

Our new strategy also significantly reduced the number of false positive differential peaks, filtering out 444 of the 565 peaks that a traditional metabolomics analysis may have considered as potential biomarkers for the two rice cultivars (Figure 1E, Supplemental Table S3). For example, the area of the peak 636.2171_0.6162 (*m/z_Rt*) is 6.3 times larger in the 9311 cultivar sample than in the Nipponbare sample. This peak satisfied the reproducibility and the variation criteria, and was absent in the blank samples. However, it showed an unacceptable quantitative performance in the response check (Step 4), with a correlation coefficient of -0.2415 between the peak areas and the RCI. Thus, although peak 636.2171_0.6162 may be flagged as a distinguishing biomarker by a traditional metabolomics pipeline, it is revealed by our analysis to represent a compound of non-biological origin. Even more remarkably, the metabolite report containing our final list of differential peaks allowed the putative identification of 30 metabolites by database comparisons, including 12 lipids, nine flavonoids, two amino acids, two phenolics, two nucleosides, vitamin B6, hydroxylamine and a diterpenoid (Supplemental Table S4). Many of these are validated biomarkers critically important for the yield and the seed quality of the two rice cultivars (Supplemental Results). Nevertheless, possible interactions between the

biological matrix and some compounds in a complex metabolome may complicate the identification of filtered biomarkers.

In conclusion, our new strategy greatly reduces the number of false positive peaks, enhances quantitative accuracy, and allows for a more meaningful analysis of comparative metabolomics datasets.

SUPPLEMENTAL INFORMATION

Supplemental information is available at Molecular Pant Online.

AUTHOR CONTRIBUTIONS

L.X.D., X.Q. and J.H.S designed the research. L.X.D. performed the experiments. G.A.S. wrote the script to build the optimized regression models. L.X.D., X.Q. and I.M. analyzed the data and wrote the article.

FUNDING

This work was supported by the Chinese National Key Programs for Research and Development (JFYS2016ZY03002153), the Key Project of Chinese National Programs for Fundamental Research and Development (2013CB127000), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA08020104), the National High-tech R&D Program of China (2012AA10A304-3) and the National Natural Science Foundation of China (31200227).

ACKNOWLEDGMENTS

No conflict of interest declared.

REFERENCES

Broadhurst, D.I., and Kell, D.B. (2007). Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics 2:171-196.

Chen, W., Gao, Y., Xie, W., Gong, L., Lu, K., Wang, W., Li, Y., Liu, X., Zhang, H., Dong, H., Zhang, W., Zhang, L., Yu, S., Wang, G., Lian, X., and Luo, J. (2014). Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 46:714-721.

Chen, W., Gong, L., Guo, Z., Wang, W., Zhang, H., Liu, X., Yu, S., Xiong, L., and Luo, J. (2013). A novel integrated method for large-Scale detection, identification, and quantification of widely targeted metabolites: Application in the study of rice metabolomics. Mol Plant 6: 1769-1780.

Giavalisco, P., Kohl, K., Hummel, J., Seiwert, B., and Willmitzer, L. (2009).^{13}C isotope-labeled metabolomes allowing for improved compound annotation and relative quantification in liquid chromatography-mass spectrometry-based metabolomic research. Anal Chem81: 6546-51.

Gong, L., Chen, W., Gao, Y., Liu, X., Zhang, H., Xu, C., Yu, S., Zhang, Q., and Luo, J. (2013). Genetic analysis of the metabolome exemplified using a rice population. Proc Natl Acad Sci U S A 110:20320-20325.

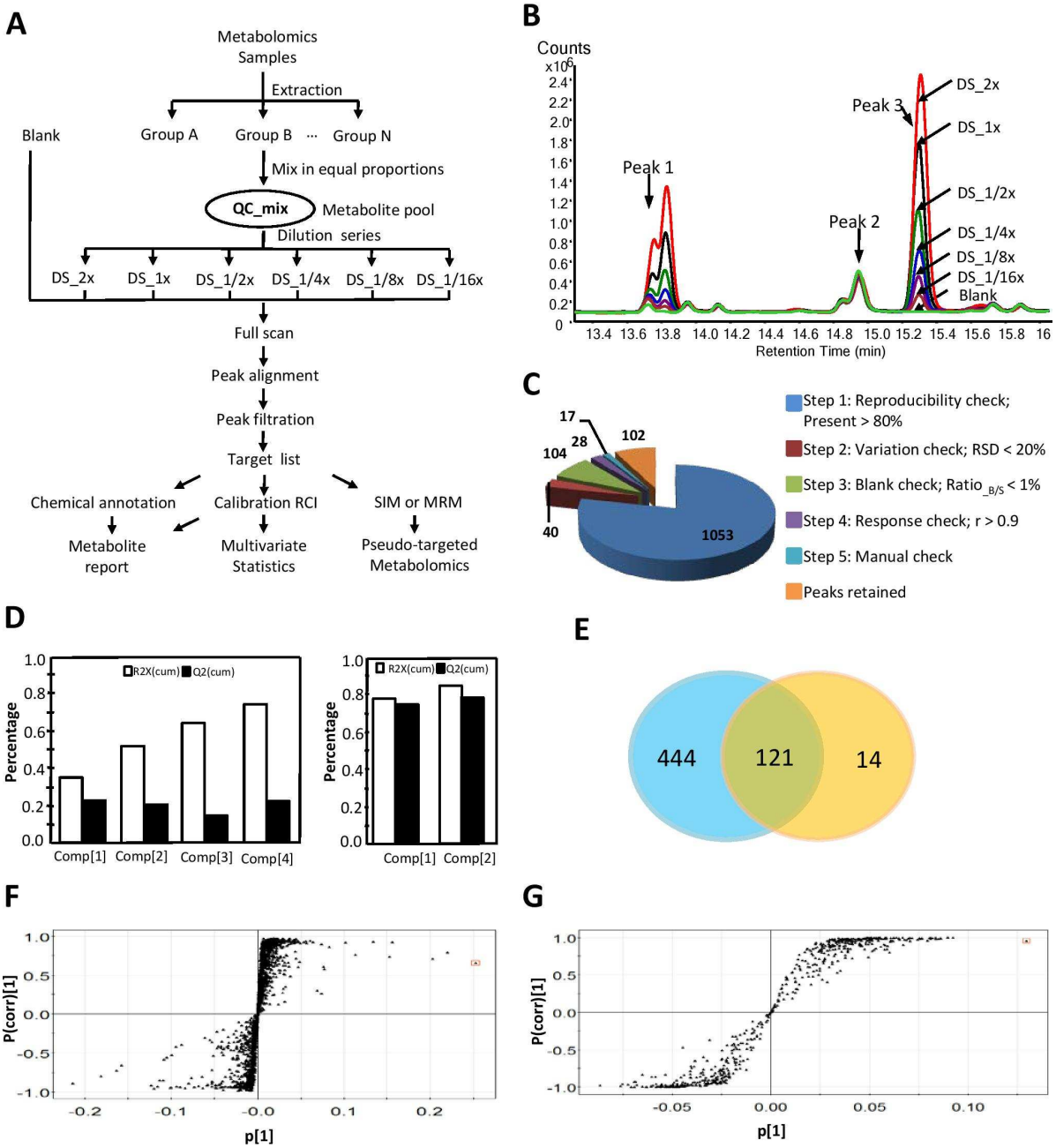
Phinney, K.W., Ballihaut, G., Bedner, M., Benford, B.S., Camara, J.E., Christopher, S.J., Davis, W.C., Dodder, N.G., Eppe, G., Lang, B.E., Long, S.E., Lowenthal, M.S, McGaw, E.A, Murphy, K.E., Nelson, B.C., Prendergast, J.L., Reiner, J.L., Rimmer, C.A., Sander, L.C., Schantz, M.M., Sharpless, K.E., Sniegoski, L.T., Tai, S.S., Thomas, J.B., Vetter, T.W., Welch, M.J., Wise, S.A., Wood, L.J., Guthrie, W.F., Hagwood, C.R., Leigh, S.D., Yen, J.H., Zhang, N.F., Chaudhary, W.M., Chen, H., Fazili, Z., LaVoie, D.J., McCoy, L.F., Momin, S.S., Paladugula, N., Pendergrast, E.C., Pfeiffer, C.M., Powers, C.D., Rabinowitz, D., Rybak, M.E., Schleicher, R.L., Toombs, B.M., Xu, M., Zhang, M., and Castle, A.L.. (2013). Development of a Standard Reference Material for metabolomics research. Anal Chem 85:11732-11738.

Redestig, H., Kobayashi, M., Saito, K., and Kusano, M. (2011). Exploring matrix effects and quantification performance in metabolomics experiments using artificial biological gradients. Anal Chem 83:5645-5651.

Sreekumar, A., Poisson, L.M., Rajendiran, T.M., Khan, A.P., Cao, Q., Yu, J., Laxman, B., Mehra, R., Lonigro, R.J., Li, Y., Nyati, M.K., Ahsan, A., Kalyana, S.S., Han, B., Cao, X., Byun, J., Omenn, G.S., Ghosh, D., Pennathur, S., Alexander, D.C., Berger, A., Shuster, J.R., Wei, J.T., Varambally, S., Beecher, C., and Chinnaiyan, A.M. (2009). Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. Nature 457:910-914.

Yu, T., Park, Y., Li, S., and Jones, D.P. (2013). Hybrid Feature Detection and Information Accumulation Using High-Resolution LC–MS Metabolomics Data. J Proteome Res 12:1419-1427.

Figure 1.The strategy of LC-MS-based metabolomics analysis. (**A**) The design of the metabolomics experiment. DS_Nx, a sample in the dilution series with an Nx dilution factor; RCI, relative concentration index; SIM, selected ion monitoring; MRM, multiple reaction monitoring. (**B**) Overlay of LC-MS total ion chromatograms for a dilution series of a sample and the blank samples. (**C**) Benchmarking the five steps of peak filtering using artificial samples. (**D**) Principle components analysis (PCA) models for rice samples. Left, the traditional method; right, the new strategy. R^2X (cum), cumulative explained variation; Q^2 (cum), cumulative cross validated R^2; Comp, principal component identified in the model.(**E**) Venn diagram of the number of differential peaks identified in the traditional (circle on the left) versus the new metabolomics approach (circle on the right). (**F**) and (**G**) The loading S-plots for PCA with the traditional metabolomics method (**F**) and the new strategy using RCI (**G**).


A
Metabolomics Samples
Extraction
Blank
Group A
Group B
Group N
Mix in equal proportions
QC_mix
Metabolite pool
Dilution series
DS_2x
DS_1x
DS_1/2x
DS_1/4x
DS_1/8x
DS_1/16x
Full scan
Peak alignment
Peak filtration
Target list
Chemical annotation
Calibration RCI
SIM or MRM
Metabolite report
Multivariate Statistics
Pseudo-targeted Metabolomics
B
Counts
Peak 1
Peak 2
Peak 3
DS_2x
DS_1x
DS_1/2x
DS_1/4x
DS_1/8x
DS_1/16x
Blank
Retention Time (min)
C
17
28
102
104
40
1053
Step 1: Reproducibility check; Present > 80%
Step 2: Variation check; RSD < 20%
Step 3: Blank check; Ratio B/S < 1%
Step 4: Response check; r > 0.9
Step 5: Manual check
Peaks retained
D
R2X(cum)
Q2(cum)
Percentage
Comp[1]
Comp[2]
Comp[3]
Comp[4]
E
444
121
14
F
P(corr)[1]
p[1]
G
P(corr)[1]
p[1]