Accepted Manuscript

statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data

Hemi Luan, Fenfen Ji, Yu Chen, Zongwei Cai

PII: S0003-2670(18)30939-5

DOI: 10.1016/j.aca.2018.08.002

Reference: ACA 236176

To appear in: *Analytica Chimica Acta*

Received Date: 5 May 2018

Revised Date: 19 July 2018

Accepted Date: 2 August 2018

Please cite this article as: H. Luan, F. Ji, Y. Chen, Z. Cai, statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data, *Analytica Chimica Acta* (2018), doi: 10.1016/j.aca.2018.08.002.

statTarget: a streamlined tool for signal drift correction and interpretations of quantitative mass spectrometry-based omics data

Hemi Luan[1#], Fenfen Ji[1#], Yu Chen[2], Zongwei Cai[1*]

1. State Key Laboratory of Environmental and Biological Analysis (SKLEBA), Hong Kong Baptist University, Kowloon Tong, Hong Kong SAR, China

2. Max Planck Institute for Terrestrial Microbiology & LOEWE Research Center for Synthetic Microbiology (SYNMIKRO), Marburg, Germany

# These authors contributed equally to this work.

* corresponding author

Abstract

Large-scale quantitative mass spectrometry-based metabolomics and proteomics study requires the long-term analysis of multiple batches of biological samples, which often accompanied with significant signal drift and various inter- and intra- batch variations. The unwanted variations can lead to poor inter- and intra-day reproducibility, which is a hindrance to discover real significance. The use of quality control samples and data treatment strategies in the quality assurance procedure provides a mechanism to evaluate the quality and remove the analytical variance of the data. The statTarget we developed is a streamlined tool with an easy-to-use graphical user interface and an integrated suite of algorithms specifically developed for the evaluation of data quality and removal of unwanted variations for quantitative mass spectrometry-based omics data. A novel quality control-based random forest signal correction algorithm, which can remove inter- and intra- batch unwanted variations at feature-level was implanted in the statTarget. Our evaluation based on real samples showed the developed algorithm could improve the data precision and statistical accuracy for mass spectrometry-based metabolomics and proteomics data. Additionally, the statTarget offers the streamlined procedures for data imputation, data normalization, univariate analysis, multivariate analysis, and feature selection. To conclude, the statTarget allows user-friendly the improvement of the data precision for uncovering the biologically differences, which largely facilitates quantitative mass spectrometry-based omics data processing and statistical analysis.

1. INTRODUCTION

Mass spectrometry (MS)-based omics techniques, including metabolomics and proteomics, are rapidly growing fields in system biology combining both analytical and statistical methodologies for a high throughput analysis of multiplex molecule profiles. Large-scale experiments coupled with varieties of statistical tools have been used to identify specific biological changes, leading to the understanding of biomarkers and multi-biochemical pathways [1, 2]. However, the omics data obtained from MS-based experiments is subject to various forms of unwanted variations including both within-batch and between-batch variations introduced by signal drift/attenuation and multiplicative noise [3] (e.g., temperature changes within the instrument, accumulated contamination, or loss of instrument performance during a long run of samples). These inherent biases and variations in metabolomics and proteomics data were challenges for quantitative comparative analysis [3-5].

To reduce the above variations, one of the most common analytical approaches is to add one or several internal standard into the biological subject samples, which is based on the assumption that the additive internal standards are representative of the entire metabolome or proteome in samples. However, it is not an ideal approach and may in fact skew the data because of the invalid assumption [6]. Recently, a controlled experiment involved the repetitive use of the same pooled quality control (QC) samples throughout the data collection process was developed based on the fact that the variation shown in the QC data is a summed representative of the analytical behavior in the real sample data [7]. This kind of QC strategy

allows us to evaluate the signal drift and other systematic noise through mathematical algorithms in liquid or gas chromatography (LC or GC) hyphenated to long-term MS-based metabolomics and proteomics studies [8-10]. However, for open-source software metaboAnalyst and metaX developed for statistical analysis [11, 12] , the QC procedure has been overlooked. In 2016, we firstly developed and released statTarget (http://bioconductor.org/packages/statTarget/) [13], an interface/ R/Bioconductor package for signal correction and statistical analysis of QC based experiments. The removal of unwanted variations and integration of multi-batch metabolomics data was implemented by using quality control-robust LOESS signal correction methods. Since then, the popular NOREVA [14] and NormalizeMets [15] softwares have used our early released statTarget for handling the inter- and intra- batch effects from given large-scale metabolomics data. Herein, we further developed the statTarget package. The new edition of statTarget represents an upgraded and improved version over what was described in 2016. The statTarget now provides a point-and-click interface and a novel QC-based random forest signal correction (QC-RFSC) algorithm to remove unwanted variations at feature-level, as well as new capabilities in feature selection for interpretations and visualization of quantitative mass spectrometry-based omics data.

2. MATERIALS and METHODS

2.1. Study designs

In this work, two non-targeted metabolomics and one quantitative selected reaction monitoring (SRM)-based proteomics studies were involved to evaluate the performance of QC-RFSC method and the statTarget software. A quality assurance strategy based on the periodic analysis of standard QC samples together with the real samples was involved in the above studies. In details, the standard QC samples were prepared by mixing equal volumes of real samples before sample preparation when they were aliquoted for analysis. The QC samples were processed the same as real samples and used to estimate a "mean" profile representing all the peaks detected during the MS analysis. At the beginning of analysis, QC samples were advisable to equilibrate the analytical platform and then injected at regular intervals throughout the analysis in order to provide data. The running order of subjects was randomized. A diagram of quality assurance strategy was shown in **Fig. S1**.

All studies were approved by the ethics committee of Hong Kong Baptist University's Institutional Review Board, and written informed consents were obtained in these studies. Firstly, a non-targeted LC-MS based metabolomics experiment were designed for the investigation of human urine metabolic profiles (Dataset 1). Two urine samples were collected from one person at different two days, and equally pooled as the standard QC sample. The first day sample was aliquoted and utilized as the real samples. The QC samples were analyzed at the start of analysis to equilibrate the analytical platform, and after every five aliquoted real samples. In total, we produced 4549 features, which were consistently detected in 46 aliquoted QC samples and 175 aliquoted real samples. Secondly, we collected a published metabolomics data (Dataset 2) from MetaboLights databases

(https://www.ebi.ac.uk/metabolights/MTBLS146) to evaluate the performance of QC-RFSC method in multi-batches metabolomics data. The detailed description of Dataset 2 was provided in Supplementary file. Briefly, pregnant women were recruited and divided into six groups (A-F) according to their gestational weeks. The aliquoted QC samples were analyzed after every six real plasma samples in two batches. In total, we produced 1312 features, which were consistently detected in 38 aliquoted QC samples and 179 real plasma samples. Thirdly, we designed a SRM-based proteomics study for quantitative analysis of proteins in the serum of Parkinson's disease (PD) patients and normal controls. A total of 34 PD patients together with 34 normal control persons were recruited. Venous blood was collected with blood collection tubes in the morning before breakfast from all the participants, and then serum samples were separated at room temperature from vein blood and stored at −80 °C until use. The QC samples were analyzed by LC-MS after every eight persons' samples in the entire batch. In total, we produced 174 features, which were consistently detected in 13 aliquoted QC samples and 68 real serum samples. This study was used to evaluate the performance of the QC-RFSC method for removal of variations in the proteomics data (Dataset 3).

2.2. Sample preparation and LC-MS analysis

Urine preparation for metabolomics analysis was performed as previously reported [16]. A Dionex U3000 LC system coupled online to Q Exactive Focus instrument (Thermo Fisher Scientific, MA, USA) was employed to detect the urinary metabolic profile. On the other hand, serum preparation including protein denaturation, digestion, and alkylation in

proteomics analysis were performed as previously reported [17]. A total of 174 transitions response to 69 proteins were monitored by using a Dionex U3000 LC system coupled online to TSQ Quantiva instrument (Thermo Fisher Scientific, MA, USA) with selected reaction monitoring (SRM) mode. The details of the analytical experiments were described in Supplemental materials.

2.3. Feature-based signal drift correction

Correction factor is the ratio between the response of an analyte being analyzed and the response of reference standard, being utilized to eliminate the irreproducibility in sample measurements. Ideally, the reference standards should be physically and chemically identical to the analyte (e.g., the stable labeled compounds). However, to have hundreds to thousands of metabolites or proteins corresponding reference standards in omics study is insupportable cost. On the other hand, QC samples-based signal drift correction provides the low cost and efficient approach to improve the data quality for metabolomics and proteomics.

In principle, the QC dependent correction factor of each feature was obtained by learning an ensemble of regression trees each grown on a random sub - QC sample to infer a noisy variant during the experimental analysis, and calculated using the regression fit derived from the intensities of features in the QC samples. For any feature j in any real sample i, the intensity value in the preprocessed data is represented by x_{ij} and the corrected value by x'_{ij}. Due to the different degree of drift in each feature j in a real sample i, a QC dependent correction factor F_{ij} is assigned according to eq 1.

$$x'_{ij} = \frac{x_{ij}}{F_{ij}} \qquad (1)$$

For a feature j in the QC samples, a random forest model is trained using the intensity of QC samples with respect to the order of injection (k). Then, the F_{ij} is predicted by the built random forest model according to the injection order k of the features j in real samples i as defined by eq 2:

$$F_{ij} = \Phi(k_j) \quad (2)$$

Φ represents machine learning model. Additionally, Seven common used methods in this study, including four QC samples independent normalization methods (total sum, variance stabilizing normalization (VSN), probabilistic quotient normalization (PQN), Quantiles) and three QC-based signal correction methods (QC-RLSC, QC-SVR, QC-RFSC) were compared in this study. The correction factors in QC-RLSC and QC-SVR methods were calculated using the nonlinear local polynomial regression (LOESS) and support vector regression (SVR) fit derived from the intensities of the QC samples, respectively.

2.4. Permutation-based false discovery rates

To give a measure of the statistical significance of features derived from Benjamini-Hochberg adjusted P-value [18], Mann-Whitney U-test, a permutation test was introduced [19]. We assumed that there was no difference between two groups that were randomly formed in Dataset 1. The labels of the samples were randomly permuted 999 times and the adjusted P-value was calculated between the randomly formed two groups. The permutation-based false discovery rates (FDR) was calculated as $(100 \times FP)/(FP + TP)\%$, in which *FP* is the number of false positive results (adjusted P-value < 0.05) and *TP* is the number of true positive results (adjusted P-value > 0.05).

2.5. Data analysis

MS raw data files were converted to mzML format using ProteoWizard. The XCMS [20] and Skyline [21] were used for the extraction of peak abundances of metabolites and tryptic digested peptides, respectively. The statTarget was employed for signal drift correction at feature-level and statistical analysis.

3. RESULTS and DISCUSSION

3.1. Software scheme of statTarget

The statTarget is a free and open source language R package, and supported by the RGTK2 graphical library[22]. The graphical user interface (GUI) implementations of statTarget in the Windows and MacOS operating systems allow users to easily analyze the data without the requirement of programing experiences (**Fig. S2**). It provides the streamlined analysis approaches for MS-based omics data. The utilizations of statTarget were elaborated in the following subsections according to the statTarget workflow (**Fig. 1**).

The statTarget software, the user guides, and an illustrative example can be downloaded from the statTarget Website (https://stattarget.github.io). Briefly, ***Data Import*** statTarget supports comma-delimited peak abundance data of mass spectra as input file formats. The results reports from widely used programs, i.e. XCMS, MZmine2 [23], SKYLINE, could be easily converted into input file formats by using *transX()* function in statTarget. ***Data Preprocessing*** statTarget provides streamlined procedures for data normalization and transformation. The initial procedures are to estimate the zero and missing values for feature intensities occurred occasionally in the QC and real samples, and imputed zero and missing

values with four methods, such as K-Nearest Neighbor, minimum, half of minimum or median values. The following procedures are to consider the data normalization (i.e. total sum, probabilistic quotient normalization (PQN) or none), variance-stabilizing transformations (generalized log transformation or none) and data scaling (pareto scaling, vast scaling, range scaling, and auto-scaling). Data scaling procedure is applicable for Principal component analysis (PCA) and partial least squares discriminant analysis (PLSDA) only. ***Signal correction*** The QC-RFSC and QC-RLSC are provided for estimation of data quality and removal of unwanted variation at feature-level, and performed after the data imputation procedures. ***Statistical analysis*** PCA, PLSDA, and random forest model were integrated for interpretation, visualization of high dimensional data and selection of important features. The Welch's t-test, Mann-Whitney-Wilcoxon test, false positive rate test, receiver operating characteristic curve, odds ratio, and volcano plot analysis are implemented for feature selection between any two groups. The functionalities of statTarget were illustrated using real-world datasets in supplemental file.

3.2. Signal drift correction

Supervised machine learning allows for systematic pattern identification from high-dimensional features in the training dataset and minimizes manual tuning for optimal model generalization. It is well-suited for an unbiased analysis of MS data, and has been used for identification of metabolite fingerprinting, biomarkers and protein sequences.

Twelve well-known supervised machine learning algorithms were evaluated using all features in the QC samples from MS-based dataset 1 [24], such as random forest (RF),

bayesian, eXtreme gradient boosting, support vector machines, LOESS, linear regression, etc. The caret package (version 6.0-7.1) in R version 3.3.3 was operated for models training and parameter tuning with 10-fold cross-validation. The features in QC samples data were used for the model training and the relative standard deviation (RSD) calculation. Our results showed the RF performed best predictive accuracy, as 37.5% of the features proportion achieved the high R-squared values between 0.5 and 1.0. (**Fig. 2A**). Given that RF performed well in the real dataset, RF was therefore selected for QC-based signal correction [25, 26]. In this study, we further compared the performance of seven normalization methods for three datasets, and our results were consistent with the former reports (**Table S1**), showing the performance of QC-based signal correction were better than the normalization methods without QC procedures. The features adjusted with QC-based signal correction showed the lower RSD% than the features with QC samples independent methods.

QC-RFSC algorithm integrates the RF based ensemble learning approach to learn the unwanted variations from QC samples, and predict the correction factor in the neighboring real samples responses. To evaluate the performance of QC-RFSC algorithms in the statTarget software, we compared available algorithms that were common used for QC-based signal correction, including RF, SVR [27] and LOESS [2]. PCA score plot analysis showed the QC samples were clustered tightly after raw data corrected by any of the three algorithms, which indicated the three algorithms have the ability to remove the unwanted variations in the dataset 1. However, the RF and LOESS showed better performance than SVR as shown in **Fig. S3**. In the raw data, the proportion of peaks within 15 % RSDs was only 1.97 % of the

total peaks in dataset 1. After adjusted by QC-RFSC method, there was a 12.7-fold increase (25.1%) in the number of peaks within 15 % RSDs. Meanwhile, the proportion of peaks with RSDs less than 30 % increased significantly from 43.9% to 86.2 % after QC-RFSC method. The QC-RLSC and QC-SVR algorithms also increased the proportion of peaks with 30% RSDs to 65.5% and 61.9 %, respectively (**Fig. 2B**). Moreover, the features within 30% RSDs in our QC-RFSC method overwrote the features of the other two methods (**Fig. 2C**). We further calculated the cumulative frequency of RSD% of all features in QC samples, which showed that 4517 out of 4549 features, as much as 99.9 %, had decreased RSDs after QC-RFSC method ((**Fig. 2D**). We further performed the permutation-based FDR estimation to check the reliability for available algorithms (See the materials and methods). As a results, the QC-RFSC showed lowest FDR and highest statistical accuracy (**Fig. 2D**). Our results demonstrated that the QC-RFSC robustly increased precision of metabolomics data and statistical accuracy through removal of unwanted variations, which had better performance than the other two methods.

3.3. Precision improvement for metabolomics and targeted proteomics data

Precision is one of the most important criteria in the assessment of an analytical method, and achieved by monitoring quality control samples during analysis[8]. The percentage RSD% of each peak in QC samples was usually used for precision evaluation. To evaluate the precision improvement that can be achieved by our software, we evaluated the cumulative RSD distribution of all features in two batch metabolomics data (dataset 2) and targeted proteomics data (dataset 3) with pre- and post- QC-RFSC, QC-RLSC and QC-SVR (**Fig. 3A**

and 3C). As a comparison, the RSDs of peaks in metabolomics data were significantly decreased across the entire range using the QC-RFSC, QC-RLSC and QC-SVR methods (**Fig. 3A**). The proportion of peaks with inter- and intra-batch RSDs less than 30 % increased significantly to 90.9% after QC-RFSC compared with the raw data (52.1%) (**Table S2**). The QC-RLSC and QC-SVR also increased the proportion of peaks within 30% RSDs to 80.1% and 72.3%, respectively. The PCA score plots showed the inter- and intra-batch QC samples were tightly clustered due to QC-RFSC correction. Beside metabolomics data analysis, the QC-RFSC could also significantly improve the data quality for the targeted proteomics. The proportion of peaks within 30 % RSDs increased significantly to 87.9% after QC-RFSC compared with the QC-RLSC (68.4%) and QC-SVR (47.7%) (**Fig. 3C**). The clustered QC samples in PCA score plots indicated the data precision was significant improved. (**Fig. 3D**). The results demonstrated that the QC-RFSC significantly increased precision of metabolomics and targeted proteomics data, and had better performance than the two other methods.

3.4. Comparison of differentially expressed features

The performances of QC-based signal correction on identifying features differentially expressed in the metabolomics data (Group A vs Group B, dataset 2) and targeted proteomics data (PD vs normal controls, dataset 3) were evaluated, respectively. From statTarget, the visualization image analysis of the quality improvement procedures for every feature can be obtained. As showed in quality control images for the typical features, the QC samples exceeded of the quality control limits (mean ± SD) were detected. After the samples from

metabolomics (**Fig. 4A**) and targeted proteomics data (**Fig. 4D**) were corrected by using QC-RFSC methods, QC samples fell back into the quality control limits. Then for each peak reproducibly detected and remained by passing the acceptable quality control level (RSD% < 30%), Mann–Whitney U test or Welch-test was applied to measure the significance of each peak between two groups. We further utilized the volcano plot function in statTarget to select the differentially expressed features (**Fig. 4B** and **4E**) with fold changes greater than 1.2 or less than 0.8, and an adjusted p-value less than 0.05. As a result, we found 290, 126 and 90 differentially expressed features (**Fig. 4C**) in the metabolomics data corrected by using QC-RFSC, QC-RLSC, QC-SVR methods, respectively. A total of 79 differentially expressed features were shared by three methods, and moreover, there were 174 unique differentially expressed features in QC-RFSC method for dataset 2. In contrast, there were only 19 differentially expressed features screened out by directly analyzing the uncorrected data. As showed in **Fig. 4F** for targeted proteomics data, we found the features selected from QC-RFSC could overwrote the features from the other two methods, and there were 7 unique differentially expressed features from 6 targeted proteins in QC-RFSC method (**Table S3**). Our results demonstrated that QC-RFSC was more efficient to find differentially expressed features than the other two methods, which has the best power to enlarge the number of differentially expressed features both for metabolomics data and targeted proteomics data.

Our results showed that the QC-RFSC is an efficient method to remove inter- and intra-batch of unwanted variations at feature-level, and improve the data precision and statistical accuracy for quantitative mass spectrometry-based omics data. However, the method also has

some drawbacks. Firstly, to collect the QC samples data is somewhat laborious and time consuming. Also for some non-targeted proteomics study, the complicated sample preparation procedures (i.e. pre-fraction, digestion, desalting) may produce random variations that cannot be corrected. The QC samples-based method is recommended for the projects without complete sets of reference standards, and can be effectively improve the data precision and statistical accuracy.

4. CONCLUSION

The statTarget is a user-friendly tool for integration, visualizations and statistical analysis of quantitative mass spectrometry-based omics data. The developed QC-RFSC algorithm is a highly efficient approach to remove inter- and intra-unwanted variations, to improve the data quality of quantitative mass spectrometry-based omics data, and to further enlarge the number of differentially expressed features. Due to the substantial similarities among different types of expression data from system biology analysis (e.g., high dimension, analytical bias, significance analysis and so on), it was also feasible to extend the scope of statTarget from quantitative metabolomics data to other system biology data such as protein or peptide expression data. The statTarget is available as a graphical user interface that makes the application easy to use and can be installed on any computer by standard R installation (R ver. 3.3. +).

ACKNOWLEDGMENTS

The authors would like to thank the financial supports from Hong Kong Baptist University (IRMC/13-14/03-CHE) and the National Sciences Foundation of China (NSFC21675176 and NSFC91543202).

Conflict of Interest: none declared.

REFERENCES

[1] H. Keshishian, M.W. Burgess, H. Specht, L. Wallace, K.R. Clauser, M.A. Gillette, S.A. Carr, Quantitative, multiplexed workflow for deep analysis of human blood plasma and biomarker discovery by mass spectrometry, Nat Protoc, 12 (2017) 1683-1701.

[2] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D.B. Kell, R. Goodacre, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat Protoc, 6 (2011) 1060-1083.

[3] F. Fernandez-Albert, R. Llorach, M. Garcia-Aloy, A. Ziyatdinov, C. Andres-Lacueva, A. Perera, Intensity drift removal in LC/MS metabolomics by common variance compensation, Bioinformatics, 30 (2014) 2899-2905.

[4] M.P. Molloy, E.E. Brzezinski, J. Hang, M.T. McDowell, R.A. VanBogelen, Overcoming technical variation and biological variation in quantitative proteomics, Proteomics, 3 (2003) 1912-1919.

[5] D.L. Tabb, Quality assessment for clinical proteomics, Clin Biochem, 46 (2013) 411-420.

[6] T.D. Veenstra, Metabolomics: the final frontier?, Genome Med, 4 (2012) 40.

[7] W.B. Dunn, I.D. Wilson, A.W. Nicholls, D. Broadhurst, The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans, Bioanalysis, 4 (2012) 2249-2264.

[8] E. Zelena, W.B. Dunn, D. Broadhurst, S. Francis-McIntyre, K.M. Carroll, P. Begley, S.

O'Hagan, J.D. Knowles, A. Halsall, H. Consortium, I.D. Wilson, D.B. Kell, Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum, Anal Chem, 81 (2009) 1357-1364.

[9] H. Luan, N. Meng, P. Liu, Q. Feng, S. Lin, J. Fu, R. Davidson, X. Chen, W. Rao, F. Chen, H. Jiang, X. Xu, Z. Cai, J. Wang, Pregnancy-induced metabolic phenotype variations in maternal plasma, J Proteome Res, 13 (2014) 1527-1536.

[10] A. Beasley-Green, D. Bunk, P. Rudnick, L. Kilpatrick, K. Phinney, A proteomics performance standard to support measurement quality in proteomics, Proteomics, 12 (2012) 923-931.

[11] J. Xia, I.V. Sinelnikov, B. Han, D.S. Wishart, MetaboAnalyst 3.0--making metabolomics more meaningful, Nucleic Acids Res, 43 (2015) W251-257.

[12] B. Wen, Z. Mei, C. Zeng, S. Liu, metaX: a flexible and comprehensive software for processing metabolomics data, BMC Bioinformatics, 18 (2017) 183.

[13] H. Luan, statTarget: Statistical Analysis of Metabolite Profile, R package version 1.2.2, (2016).

[14] B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, Y. Chen, W. Xue, X. Li, F. Zhu, NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res, (2017).

[15] A.M. De Livera, M. Sysi-Aho, L. Jacob, J.A. Gagnon-Bartsch, S. Castillo, J.A. Simpson, T.P. Speed, Statistical methods for handling unwanted variation in metabolomics data, Anal Chem, 87 (2015) 3606-3615.

[16] H. Luan, L.F. Liu, Z. Tang, M. Zhang, K.K. Chua, J.X. Song, V.C. Mok, M. Li, Z. Cai,

Comprehensive urinary metabolomic profiling and identification of potential noninvasive marker for idiopathic Parkinson's disease, Sci Rep, 5 (2015) 13888.

[17] E. Scheerlinck, M. Dhaenens, A. Van Soom, L. Peelman, P. De Sutter, K. Van Steendam, D. Deforce, Minimizing technical variation during sample preparation prior to label-free quantitative mass spectrometry, Anal Biochem, 490 (2015) 14-19.

[18] Y. Benjamini, Discovering the false discovery rate, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72 (2010) 405-416.

[19] Y. Xie, W. Pan, A.B. Khodursky, A note on using permutation-based false discovery rate estimates to compare different analysis methods for microarray data, Bioinformatics, 21 (2005) 4280-4288.

[20] R. Tautenhahn, G.J. Patti, D. Rinehart, G. Siuzdak, XCMS Online: a web-based platform to process untargeted metabolomic data, Anal Chem, 84 (2012) 5035-5039.

[21] B. MacLean, D.M. Tomazela, N. Shulman, M. Chambers, G.L. Finney, B. Frewen, R. Kern, D.L. Tabb, D.C. Liebler, M.J. MacCoss, Skyline: an open source document editor for creating and analyzing targeted proteomics experiments, Bioinformatics, 26 (2010) 966-968.

[22] M. Lawrence, D. Temple Lang, RGtk2: A Graphical User Interface Toolkit for R, 2010, 37 (2010) 52.

[23] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic, MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data, BMC Bioinformatics, 11 (2010) 395.

[24] M. Kuhn, Building Predictive Models in R Using the caret Package, Journal of

Statistical Software, 28 (2008) 1-26.

[25] L. Breiman, Random forests, Machine learning, 45 (2001) 5-32.

[26] W.G. Touw, J.R. Bayjanov, L. Overmars, L. Backus, J. Boekhorst, M. Wels, S.A. van Hijum, Data mining in the Life Sciences with Random Forest: a walk in the park or lost in the jungle?, Brief Bioinform, 14 (2013) 315-326.

[27] J. Kuligowski, A. Sanchez-Illana, D. Sanjuan-Herraez, M. Vento, G. Quintas, Intra-batch effect correction in liquid chromatography-mass spectrometry using quality control samples and support vector regression (QC-SVRC), Analyst, 140 (2015) 7810-7817.

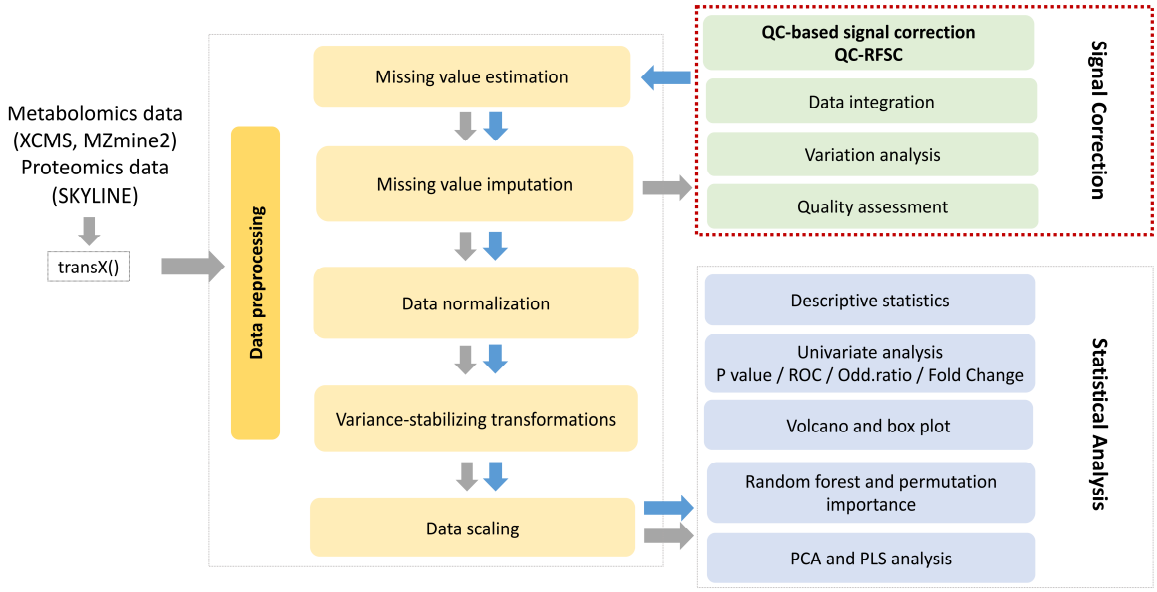
Figure captions

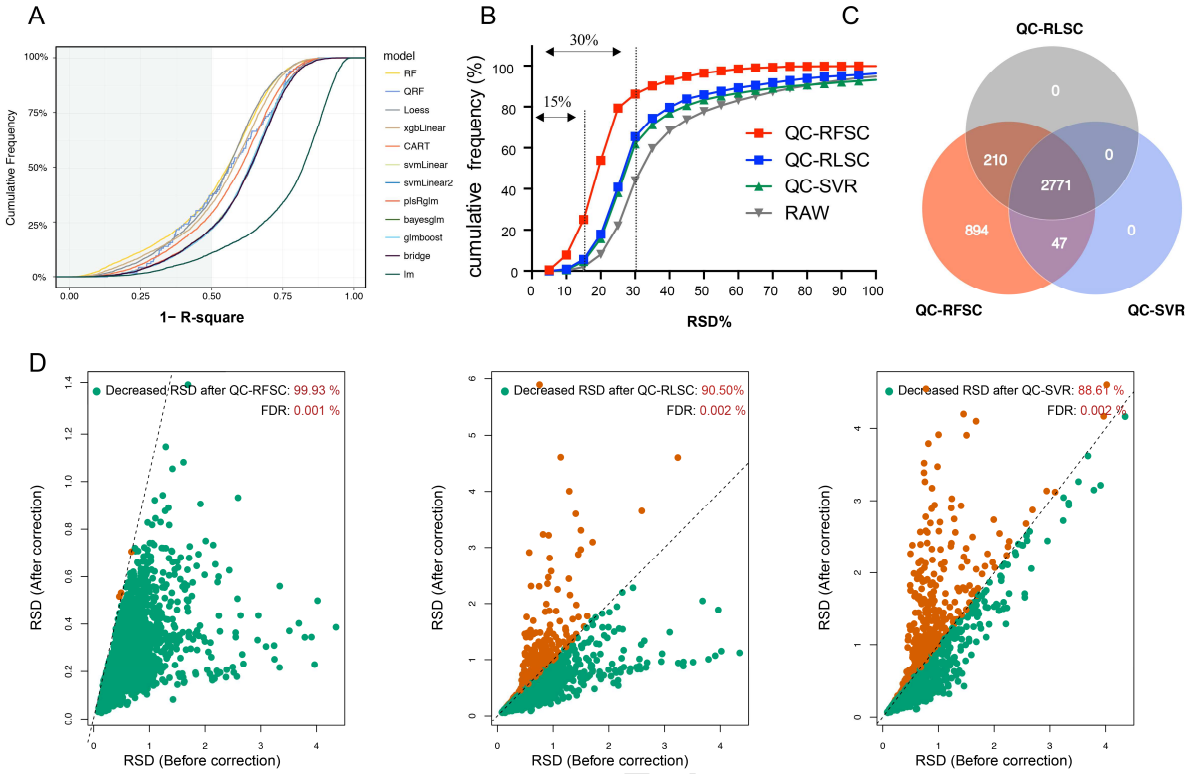
Figure 1 Flowchart of statTarget. The metabolomics and proteomics data preprocessed by using XCMS, or SKYLINE software, respectively, were then converted into input file formats by *transX()* function. Following the data preprocessing including estimation and imputation of missing value or zero value in the input files, the signal correction module was performed to remove the unwanted variations. Alternatively, the imputed data was normalized, g-log transformed or not, and submitted to statistical analysis modules (Direction of grey arrow). The corrected data files could also be submitted to data preprocessing and statistical analysis modules (Direction of blue arrow).

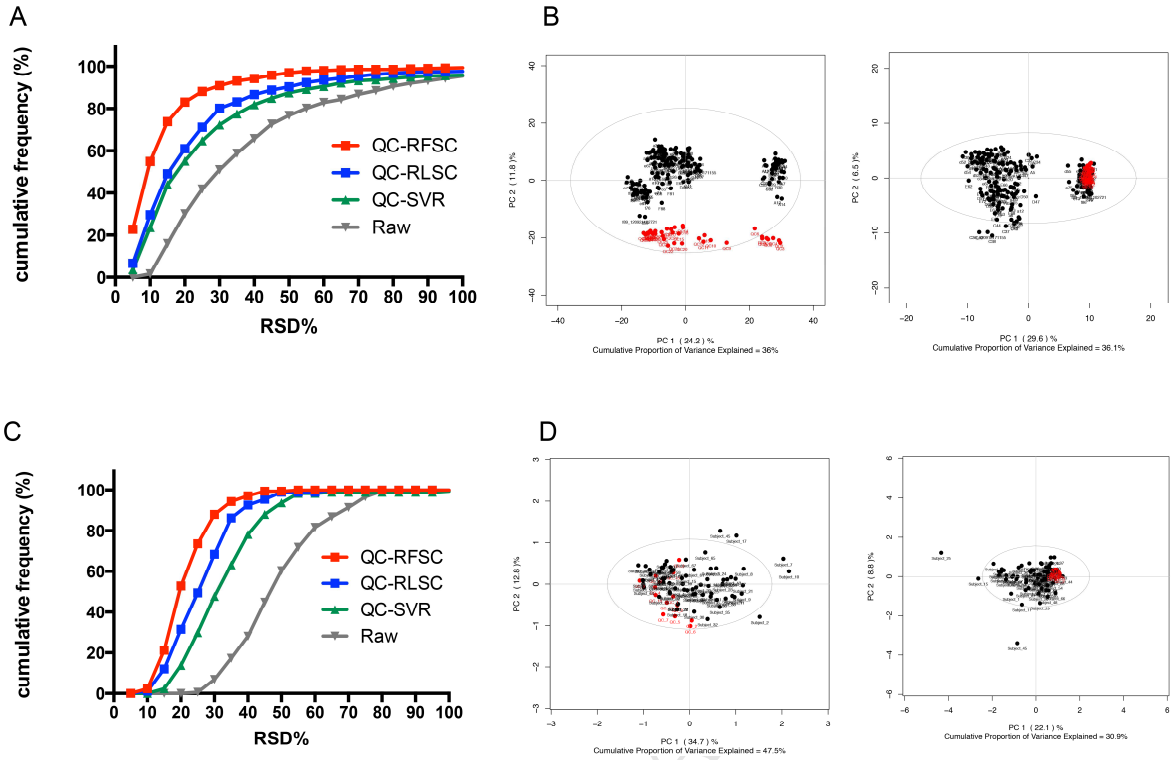
Figure 2 Comparison of the performance of machine learning models. A, cumulative frequency of the predictive accuracy of twelve machine learning models on dataset 1. B, comparison of the cumulative frequency of RSD% of all features with three correction methods. C, Venn plot of features with deceased RSDs from three correction methods. D, Scatter plot of features with deceased RSDs and permutation-based false discovery rates.

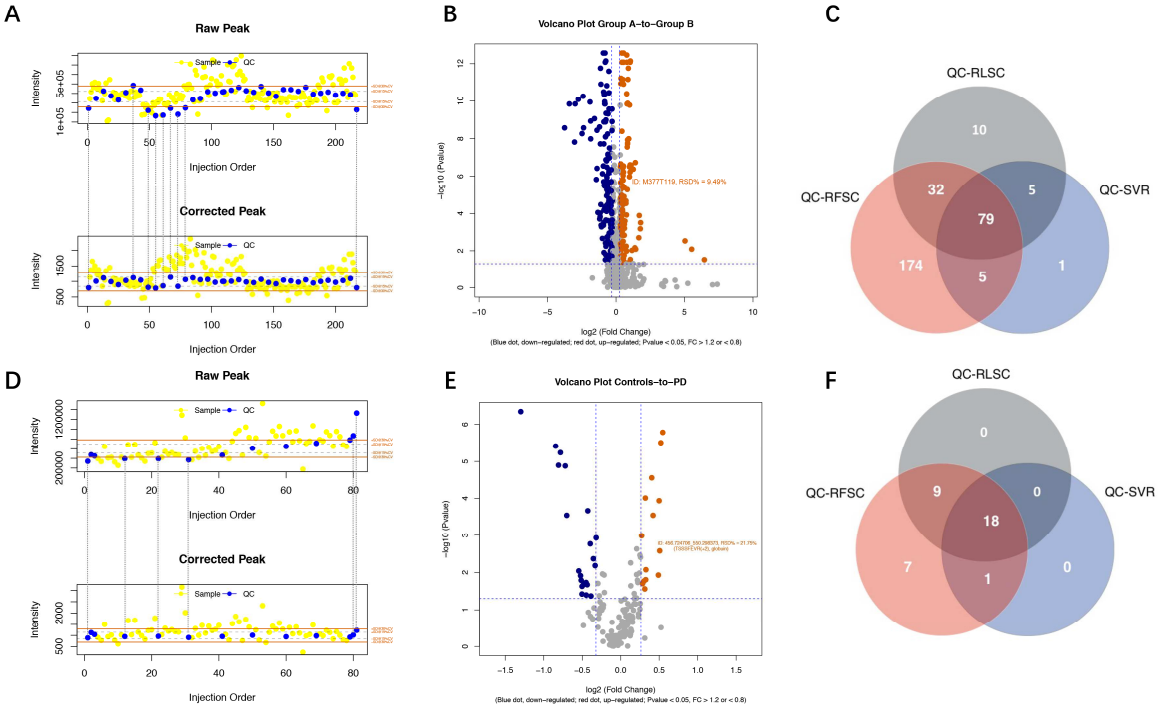
Figure 3 The performance of QC-RFSC method for metabolomics and targeted proteomics data. A, comparison of the cumulative frequency of RSD% of features in metabolomics data (dataset 2). B. PCA score plots of the metabolomics data with pre- (Left) and post- (right) correction. Red dots denote the QC samples; Black dot, real samples. C, comparison of the cumulative frequency of RSD% of features in targeted proteomics data (dataset 3). D. PCA score plots of the targeted proteomics data with pre- (Left) and post- (right) correction. Red dots denote the QC samples; Black dot, real samples.

Figure 4 Comparison of differentially expressed features. A, the visualization image analysis of the quality improvement procedures for typical features (ID: M377T119) in metabolomics data (dataset 2). The quality control limits was calculated with mean ± SD and SD = $RSD \times mean$. The red full line denotes the first quality control limits at 30% RSD threshold and the grey dashed line denotes the second quality control limits at 15% RSD threshold. B, volcano plot of differentially expressed features in dataset 2 (QC-RFSC application). C, venn plot of differentially expressed features from QC-RFSC, QC-RLSC, QC-SVR methods for dataset 2. D, the visualization image analysis of the quality improvement procedures for typical features (ID: 456.724706_550.298373) in targeted proteomics data (dataset 3). E, volcano plot of differentially expressed features in dataset 3 (QC-RFSC application). F, venn plot of differentially expressed features from QC-RFSC, QC-RLSC, QC-SVR methods for dataset 3.









Highlights

- A streamlined tool for quantitative MS-based omics data.
- Allowing user-friendly the improvement of the data quality of MS-based omics data.
- A novel QC-RFSC algorithm to remove unwanted variations at feature-level
- Facilitating quantitative MS-based omics data processing and statistical analysis.
- An alternative for controlling the data quality in MS-based omics studies.