Supplementary Data for:

NOREVA: normalization and evaluation of MS-based metabolomics data

Bo Li[1,+], Jing Tang[1,+], Qingxia Yang[1,+], Shuang Li[1], Xuejiao Cui[1] ,Yinghong Li[1], Yuzong Chen[2], Weiwei Xue[1], Xiaofeng Li[1] and Feng Zhu[1,*]

[1] Innovative Drug Research and Bioinformatics Group, School of Pharmaceutical Sciences and Innovative Drug Research Centre, Chongqing University, Chongqing 401331, China

[2] Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Singapore 117543, Singapore

[+] These authors contribute equally

* Corresponding author: Feng Zhu (zhufeng.ns@gmail.com; zhufeng@cqu.edu.cn)

Supplementary Methods

Normalization Methods Provided in NOREVA

In total, 24 methods were provided including *Auto Scaling*[1], *CCMN*[2], *Contrast*[3], *Cubic Splines*[4], *Cyclic Loess*[5], *EigenMS*[6], *Level Scaling*[7], *Linear Baseline Scaling*[8], *Log-transform*[9], *Mean Normalization*[10], *Median Normalization*[11], *MSTUS*[12], *NOMIS*[13], *Pareto Scaling*[14], *Power Scaling*[15], *PQN*[16], *Quantile*[8], *Range Scaling*[17], *RUV-2*[18], *RUV-random*[19], *SIS*[20], *Total Sum*[11], *Vast Scaling*[21] and *VSN*[22, 23].

Auto Scaling (Unit Variance Scaling, UV) is one of the simplest methods adjusting metabolic variances, which scales metabolic signals based on the standard deviation of metabolomics data[24]. This method scales all metabolites to unit variance, and all metabolites are equally important and comparably scaled[25]. The data is analyzed on the basis of correlations and standard deviation of all metabolites is one after auto scaling[24]. But the disadvantage of auto scaling is that analytical errors may be amplified due to dilution effects[24]. Auto scaling has been used to improve the diagnosis of bladder cancer using gas sensor arrays[26] and to identify urinary nucleoside markers from urogenital cancer patients by mass spectrometry (MS)-based metabolomics[27].

CCMN (Cross-Contribution Compensating Multiple Standard Normalization, CRMN) is applicable to monitor systematic error from randomized and designed experiments using multiple internal standards[2]. CCMN compensates for systematic cross-contribution effects that can be traced back to a linear association with experimental design[2], and is superior at purifying the signal of interest using multiple internal standards[2]. But care needs to be taken when normalizing the data using the factors of interest prior to carrying out unsupervised analysis[19]. CCMN is mainly aimed at MS-based metabolomics data and its inclusion will improve the precision of current metabolite profiling protocols[28].

Contrast (Contrast Normalization) comes from the integration of *MA*-plots and logged *Bland-Altman* plots, which assumes the presence of non-linear biases[24]. The input data is logged and transformed into a contrast space by means of an orthonormal transformation matrix[24]. But the use of a log function in this method may impede the processing of zeros and negative numbers, which requires the conversion of non-positive numbers to an extremely small value[24]. The contrast method has been applied in oligonucleotide arrays to normalizing feature intensities[3] and also employed to reveal the role of polychlorinated biphenyls in non-alcoholic fatty liver disease of MS-based metabolic profiling[29].

Cubic Splines is one of the non-linear baseline methods assuming the existence of non-linear relationships between baseline and individual spectra[24]. Like quantile normalization, cubic splines aims to make the distribution of the metabolite concentrations similar across all samples[30]. The geometric or arithmetic mean of the concentrations of each metabolite across all samples is regarded as the baseline sample[30]. A set of evenly distributed quantiles from both the baseline and target samples is used to fit a smooth cubic spline[30]. Finally, a spline function generator uses the generated set of

interpolated splines to fit the parameters of a natural cubic spline[30]. Cubic splines has been adopted to reduce variability in DNA microarray experiments by normalizing all signal channels to a target array[4]. Moreover, it has been applied in MS-based metabolomics profiling enabling to improve the comprehensiveness of global metabolic profiling of body fluids[31].

Similar to the Contrast, *Cyclic Loess (Cyclic Locally Weighted Regression)* originates also from the combination of *MA*-plot and logged *Bland-Altman* plot by assuming the existence of non-linear bias[24], and can estimate a regression surface using multivariate smoothing procedure[32]. However, cyclic loess is one of the most time-consuming one among the normalization methods, and the amount of time grows exponentially as the number of sample increases[33]. Cyclic loess has been applied in MS-based metabolomics profiling, revealing that this method was able to remove the systematic effect[34].

EigenMS removes bias of unknown complexity from the Liquid Chromatography coupled with Mass Spectrometry (LC/MS)-based metabolomics data, allowing for increased sensitivity in differential analysis. EigenMS normalization aims at preserving the original differences while removing the bias from the data[35]. It works by 3 steps[6]: (1) EigenMS preserves the true differences in the metabolomics data by estimating treatment effects with an ANOVA model; (2) singular value decomposition of the residuals matrix is used to determine bias trends in the data; (3) the number of bias trends is estimated via a permutation test and the effects of the bias trends are eliminated. EigenMS has applied in MS-based quantitative label-free proteomics profiling[35] and MS-based metabolomics analysis[6].

Level Scaling transforms metabolic signal variation into variation relative to the average metabolic signal by scaling according to the mean signal, so the resulting values are changes in percentages compared to the mean concentration[7]. This method is especially suitable for the circumstances when huge relative variations are of great interest (e.g., studying the stress responses)[7]. Level scaling is used for identification of biomarkers focusing on relative response, but the disadvantage of it is the inflation of the measurement errors[7]. Level scaling has been used to identify urinary nucleoside markers from urogenital cancer patients in MS-based metabolomics analysis[27].

Linear Baseline (Linear Baseline Scaling) maps each spectrum to the baseline based on the assumption of a constant linear relationship between each feature of a given spectrum and the baseline[24]. The baseline is the median of each feature across all spectra and the scaling factor is computed as the ratio of the mean intensity of the baseline to the mean intensity of each spectrum[24]. The intensities of all spectra are multiplied by their particular scaling factors[24]. However, this assumption of a linear correlation among sample spectra may be oversimplified[24]. This method has been conducted to identify differential metabolomics profiles among the banana's 5 different senescence stages[36]. Moreover, linear baseline scaling has been applied to normalize nuclear magnetic resonance (NMR)-based metabolomics data[37] and MS-based metabolomics data[34].

Log-transform converts skewed metabolomics data to symmetric by non-linear transformation[7]. This method transforms the relationship of metabolites from multiplication to addition[7]. Log-transform is used to perfectly removes heteroscedasticity when the relative standard deviation is constant[7]. But the disadvantage of log-transform is that it is unable to deal with the value zero[7]. Furthermore, its effect on values with a large relative analytical standard deviation is problematic[7]. Log-transform was used to compare plasma amino acid patterns in LC/MS-based metabolomics analysis[38]. And it was applied to normalize the data in metabolomics analysis based on gas chromatography coupled with mass spectrometry (GC/MS)[39].

Mean Normalization normalizes the data by mean value of all signals to eliminate background effect[10]. Intensity of each metabolite in a given sample is used by the mean of intensity of all variables in the sample[18]. In order to make the samples comparable, the means of the intensities for each experimental run are forced to be equal to one another using this method[34]. For example, each sample is scaled such that the mean of all abundances in a sample equals one[18]. This method has been applied to normalize the MS-based metabolomics data[34].

Median Normalization is based on the assumption that the samples of a dataset are separated by a constant. It scales the samples so that they have the same median. For example, the median of the metabolite abundances in the sample equals one[11]. The median normalization, the commonly used method without the need for internal standards, is more practical than the sum normalization especially in situations where several saturated abundances may be associated with some of the factors of interest[11]. Median normalization has previously been used in MS-based proteomics analysis[40] and metabolomics analysis[34].

MSTUS (MS Total Useful Signal) utilizes the total signals of metabolites that are shared by all samples by assuming that the number of increased and decreased metabolic signals is relatively equivalent[12, 41]. Using MSTUS, the concentration of each metabolite is divided by the sum of the concentrations for all the measured metabolites in a given sample[30]. However, the validity of this hypothesis is questionable since an increase in the concentration of one metabolite may not necessarily be accompanied by a decrease in that of another metabolite[41, 42]. MSTUS is a more recent technique, typical used to normalize NMR-based metabolomics data[43] and LC/MS-based metabolomics data[11].

NOMIS (Normalization using Optimal Selection of Multiple Internal Standards) finds optimal normalization factor to remove unwanted systematic variation using variability information from multiple internal standard compounds[13]. NOMIS method can select best combinations of standard compounds for normalization using multiple linear regression[13] and remove all correlations with the standards[2]. This method has a superior ability to reduce variability across the full spectrum of metabolites[13]. Moreover, the NOMIS method can be used in both supervised and unsupervised analysis[19]. Now NOMIS method has been used to normalize LC/MS-based metabolomics data[13].

Pareto Scaling uses the square root of the standard deviation of the data as scaling factor[14]. Pareto scaling is able to reduce the weight of large fold changes in metabolite signals, which is more significantly than auto scaling[24]. But the dominant weight of extremely large fold changes may still be unchanged[24]. So the disadvantage of pareto scaling is the sensitivity to large fold changes[7]. Pareto scaling was used to reduce the mask effect from the abundant metabolites for LC/MS-based metabolomics dataset[44].

Power Scaling aims at correcting for the heteroscedasticity and pseudo scaling[7]. Power scaling shows a similar transformation pattern as the log-transform, but it is not able to make multiplicative effects additive[7]. Unlike log-transform, power scaling can handle zero values[7]. Power scaling reduces heteroscedasticity without problems with small values, but its disadvantage is that the choice for square root is arbitrary[7]. Power scaling has been used to study the serum amino acid profiles and their variations in colorectal cancer patients for MS-based metabolomics[45].

PQN (Probabilistic Quotient Normalization) transforms the metabolomics spectra according to an overall estimation on the most probable dilution[16]. This algorithm has been reported to be significantly robust and accurate comparing to the integral and the vector length normalizations[16]. There are three steps in the procedure of PQN[24]: (1) perform an integral normalization of each spectrum, then select a reference spectrum such as the median spectrum; (2) calculate the quotient between a given test spectrum and reference spectrum, then estimate the median of all quotients for each variable; (3) all variables of the test spectrum are divided by the median quotient. PQN is a robust method to account for dilution of complex biological mixtures for NMR metabolomics analysis[16]. Recently, PQN is also used to reduce unwanted variance for direct infusion MS metabolomics dataset[46].

Quantile (Quantile Normalization) aims at achieving the same distribution of metabolic feature intensities across all samples, and the quantile-quantile plot in this method is used to visualize the distribution similarity[24]. Quantile normalization is motivated by the idea that the distribution of two data vectors is the same if the quantile-quantile plot is a straight diagonal line[8]. While a common and non-data driven distribution is generated using quantile normalization, an agreed standard could not be reached[8]. Quantile normalization has been adopted for high density oligonucleotide array data based on variance[8], improving NMR-based metabolomics analysis[24] and reducing non-biological systematic variation for LC/MS-based metabolomics data[47].

Range Scaling is applied to put all measured intensities on an equal footing, which means that the measured intensity was divided by the range of those intensities over all samples[17]. The biological range (difference between the minimal and the maximal concentration of a certain metabolite) is used as the scaling factor for range scaling[7]. The advantage of range scaling is that relative concentration for each variable is generated after removing instrumental response factors[17]. Range scaling has a property that all levels of variation for the metabolites are treated equally[17]. But the disadvantage of range

scaling is the sensitivity to outliers because only two values are used to estimate the biological range[7]. Range scaling has been used to fuse MS-based metabolomics data[17].

RUV-2 (Remove Unwanted Variation-2) is based on a linear model designed for identifying differentially abundant metabolites, which requires factors of interest along with the factors of unwanted variation[19]. The advantages of the RUV-2 model include[18]: (1) the biological factors of interest are not removed along with the unwanted variation; (2) the method is applied to datasets without internal standards; (3) all unwanted biological variation can be accommodated; (4) it allows for the systematic integration of datasets from different sources; (5) it removes both observed and unobserved unwanted variations. However, RUV-2 method is not a global normalization method without a complete normalized dataset[28], and it cannot be used prior to unsupervised analyses[19]. RUV-2 method has been used for normalizing and integrating MS-based metabolomics data[18].

RUV-random (Remove Unwanted Variation-Random) is based on a linear mixed effects model utilizing quality control metabolites to obtain normalized data in metabolomics experiments[19]. RUV-random method attempts to remove overall unwanted variation[19]. RUV-random accommodates unwanted biological variation and retains the essential biological variation of interest[19]. Moreover, the unwanted variation component from any undetected experimental or biological variability can be removed[19]. This method is applicable in both supervised and unsupervised analysis[19]. RUV-random is used for removing unwanted variation for MS-based metabolomics data[19].

SIS (Single Internal Standard) provides a normalized data matrix by subtracting the log metabolite abundance of a single internal standard from the log abundances of the metabolites in each sample[18, 20]. The SIS method assumes that every metabolite in a sample is subject to the same amount of unwanted variation and they can be simply measured by a single internal standard[18]. However, the use of a single internal standard may result in highly variable normalized values, which depend on the internal standard[18]. SIS method has been used to identify factors influencing extraction and derivatization of *Arabidopsis thaliana* samples in the GC/MS-based metabolomics analysis[20].

Total Sum is a method normalizing the dataset by the sum of squares[11]. The sum of squares of all variables in a sample equals one, after each sample is scaled using sum normalization method[11, 19]. And total sum normalization relies on the self-averaging property[19]. A sample-specific constant assigns an appropriate weight to each sample, which attempts to minimize possible differences in concentration between samples[19]. Total sum normalization is used to correct for LC/MS-based metabolomics data[48].

Vast Scaling (Variable Stability Scaling) weights each variable according to a metric of its stability and it is an extension of auto scaling[21]. This method focuses on stable variables that do not show strong variation using the standard deviation and the coefficient of variation is as scaling factors[7]. Vast scaling

can be used in unsupervised and supervised analysis, but it is not appropriate for large induced variation without group structure[7]. Moreover, vast scaling is used for enhancing multivariate models for classification and biomarker identification in metabolomics analysis[21], which appears to be stable and robust for NMR and GC/MS-based metabolomics data[25].

VSN (Variance Stabilization Normalization) is one of the non-linear methods aiming to keep the variance constant over the entire data range[22, 24]. VSN approaches the logarithm for large values to remove heteroscedasticity using the inverse hyperbolic sine[24]. For small intensities, it performs linear transformation behavior to make the variance unchanged[24]. VSN was originally developed for normalizing single and two-channel microarray data[49], and currently also used to determine metabolic profiles of liver tissue during early cancer development by GC/MS[39].

Renowned Criteria for Evaluating Normalization Performance Used in NOREVA

(***a***) Method's capability of reducing intragroup variation among samples[50]

The performance of normalization method is evaluated using intragroup variation between samples. Low intragroup variation means high similarity among samples and the reproducibility of analysis[35, 50]. Measures of intragroup variability adopted in NOREVA include *pooled coefficient of variation* (PCV), *pooled estimate of variance* (PEV) and *pooled median absolute deviation* (PMAD). The lower value of PCV, PEV and PMAD shown by boxplots denotes more thorough removal of experimentally induced noise and indicates better performance of the normalization method.

Moreover, *relative log abundance* (RLA)[18] plot is used to inspect the possible variations, clustering tendencies, trends and outliers within or across group(s). The RLA plot across groups is obtained by removing the median from each metabolite across all factors of interest. The boxplots of these scaled metabolites provide a way of comparing the behavior of metabolites between two groups. For RLA plots within group, each metabolite is scaled by removing the median within each factor of interest. Boxplots of RLA can be used to visualize the tightness of the replicates within groups. The RLA plot should have a median close to zero and low variation around the median[11].

In addition, differences across groups are visualized using the *principal component analysis* (PCA)[51], a common method used for dimension reduction and visualization. In NOREVA, the PCA plot allows overall visualization of variation between 2 groups. The more distinct group variations indicate better performance of the applied normalization methods.

(***b***) Method's effect on differential metabolic analysis[35]

The differential significance of metabolites across groups measured by *P*-values is calculated using the *limma* package [52] in *R* software. The distribution of *P*-values and clustering dendrogram and heatmap plots based on differential metabolites are used under this criterion[53]. A method would be recognized

as well performed when uniform distribution of *P*-values and obvious differentiation between groups in both dendrogram and heatmap are achieved.

(***c***) Method's consistency of the identified metabolic markers among different datasets[54]

The consistency score is used to quantitatively measure the overlap of the identified metabolic markers among different dataset[54]. Firstly, random sampling is performed within the whole dataset to generate several sub-datasets. Secondly, all metabolites are ranked based on their significance (*q*-values). If the *q*-values of different metabolites are the same, absolute fold changes would be considered. Thirdly, a group of the most significant metabolites in each sub-dataset is chosen to form a list of differential metabolites. Finally, the consistency score is calculated using the most significant metabolites in each sub-dataset based on the equation as follow:

$$S = \sum_{i=2}^{C} \sum_{S \in I_i} 2^{i-2} \cdot n_S$$

where C is the total number of sub-datasets, I_i indicates a set of significant metabolites containing the intersections of any i sub-datasets, and n_S refers to the number of metabolites in the intersection S. Generally, a normalization method is more robust if it results in more metabolic markers shared by more sub-datasets with a higher consistency score.

(***d***) Method's influence on classification accuracy[18, 25, 53]

In NOREVA, *receiver operating characteristic* (ROC) curve together with *area under the curve* (AUC) value based on the support vector machine (SVM) are provided[55]. Firstly, differential metabolic feature is identified by the *partial least squares discriminant analysis* (PLS-DA). Then, the SVM models are constructed based on these identified differential features. After *k*-folds cross validation, a method with larger area under ROC curve and higher AUC value is recognized as better performed one.

(***e***) Level of correspondence between the normalized data and the reference results[35]

Additional experimental data are frequently generated as a reference to validate or adjust prior result of metabolomics analysis[56]. These reference data can be the spike-in compounds and various molecules detected by quantitative analysis or qRT-PCR[56, 57]. Here, log fold changes (logFCs) of concentration between 2 groups were calculated, and the level of correspondence between the normalized data and the reference ones was estimated based on their variations in logFCs. The normalization performance of each method could be therefore reflected by how well the logFC calculated from the normalized data corresponded to what is expected based on the reference logFC[35]. Moreover, a boxplot illustrating variations in logFCs was provided, and the median of the optimal normalization method would be close to zero and the variation around the median would be low.

Supplementary TABLES

Table S1. 24 normalization methods popular in the analysis of MS-based metabolomics data together with the representative MS-based metabolomics studies adopting each of these methods.

No.	Method	Examaple of MS-based metabolomics studies using each method
1	Auto Scaling	Centering, scaling, and transformations: improving the biological information content of metabolomics data. *BMC Genomics*. 7:142, 2006.
2	CCMN	Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Anal Chem*. 81(19):7974-80, 2009.
3	Contrast	Metabolomic analysis of the effects of polychlorinated biphenyls in nonalcoholic fatty liver disease. *J Proteome Res*. 11(7):3805-15, 2012.
4	Cubic Splines	Characterising and correcting batch variation in an automated DIMS metabolomics workflow. *Anal Bioanal Chem*. 405(15):5147-57, 2013.
5	Cyclic Loess	MetPP: a computational platform for comprehensive two-dimensional GC-TOF mass spectrometry-based metabolomics. *Bioinformatics*. 29(14):1786-92, 2013.
6	EigenMS	Metabolomics data normalization with EigenMS. *PLoS One*. 9(12):e116221, 2014.
7	Level Scaling	Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers. *J Chromatogr A*. 1283:122-31, 2013.
8	Linear Baseline Scaling	Evaluation of normalization methods to pave the way towards large-scale LC-MS based metabolomics profiling experiments. *OMICS*. 17(9):473-85, 2013.
9	Log-transform	Metabolomic Analysis of Liver Tissue from the VX2 Rabbit Model of Secondary Liver Tumors. *HPB Surg*. 2014:310372, 2014.
10	Mean Normalization	Joint GC-MS and LC-MS platforms for comprehensive plant metabolomics. *J Chromatogr B Analyt Technol Biomed Life Sci*. 877(29):3572-80, 2009.
11	Median Normalization	Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal Chem*. 75(18):4818-26, 2003.
12	MSTUS	Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer Res*. 74(12):3259-70, 2014.
13	NOMIS	Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*. 8:93, 2007.
14	Pareto Scaling	A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. *Front Mol Biosci*. 2:4, 2015.
15	Power Scaling	Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics*. 8(4):643-53, 2012.
16	PQN	Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Sci Data*. 1:140012, 2014.
17	Quantile	Quantile normalization approach for LC-mass spectrometry based metabolomic data from healthy human volunteers. *Anal Sci*. 28(8):801-5, 2012.

18	Range Scaling	Fusion of mass spectrometry-based metabolomics data. *Anal Chem*. 77(20):6729-36, 2005.
19	RUV-2	Normalizing and integrating metabolomics data. *Anal Chem*. 84(24):10768-76, 2012.
20	RUV-random	Statistical methods for handling unwanted variation in metabolomics data. *Anal Chem*. 87(7):3606-15, 2015.
21	SIS	Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. *Anal Chem*. 78(2):567-74, 2006.
22	Total Sum	Evaluation of the normalization strategies to correct for urinary output in the HPLC-HRTOFMS metabolomics. *Anal Bioanal Chem*. 408(29):8483-93, 2016.
23	Vast Scaling	The influence of scaling metabolomics data on model classification accuracy. *Metabolomics*. 11(3):684-95, 2015.
24	VSN	Optimized preprocessing of ultra-performance LC-MS urinary metabolic profiles for improved information recovery. *Anal Chem*. 83(15):5864-72, 2011.

Table S2. The coverage of normalization methods popular in MS-based metabolomics analysis in currently available online pipelines. Circle (O) indicated that the method was provided in the corresponding pipeline; cross (×) indicated that the method was not available in the corresponding pipeline; square (□) indicated that the method provided in pipeline was not the same as but related to that used in this study. Those methods highlighted in orange color and bold font were not covered by any of these 8 pipelines, and methods highlighted in blue color and bold font were just covered by only one of these pipelines.

	XCMS online	MetaboAnalyst	Normalyzer	MetaDB	MetaPre	MetDAT	MSPrep	Metabolomics Workbench	Workflow4Metabolomics
Auto Scaling	O	O	×	×	O	×	×	O	×
CCMN	×	×	×	×	×	×	O	×	×
Contrast	×	×	×	×	O	×	×	×	×
Cubic Splines	×	□	×	×	O	×	×	×	×
Cyclic Loess	×	×	O	×	O	×	×	×	×
EigenMS	×	×	×	×	×	×	×	×	×
Level Scaling	×	×	×	×	O	×	×	×	×
Linear Baseline	×	×	×	×	O	×	×	×	×
Log-transform	O	O	□	×	O	×	×	×	×
Mean Normalization	×	O	O	×	×	O	×	O	×
Median Normalization	×	O	O	×	×	×	O	×	×
MSTUS	×	×	O	×	O	×	×	×	×
NOMIS	×	×	×	×	×	×	×	×	×
Pareto Scaling	O	O	×	×	O	O	×	O	×
Power Scaling	×	×	×	×	O	×	×	×	×
PQN	×	×	×	×	O	×	×	×	×
Quantile	×	O	O	×	O	×	O	×	×
Range Scaling	×	O	×	×	O	×	×	O	×
RUV-2	×	×	×	×	×	×	×	×	×
RUV-random	×	×	×	×	×	×	×	×	×
SIS	×	□	×	×	×	×	×	×	×
Total Sum	×	O	×	O	×	×	×	×	×
Vast Scaling	×	×	×	×	O	×	×	O	×
VSN	×	×	O	×	O	×	×	×	×

Table S3. The time costs of each procedure in NOREVA for processing a large-scale metabolomics dataset MTBLS28[58] with > 1,000 samples (469 patients and 536 controls) and 1,807 metabolic features. The time costs used for web connection were evaluated by uploading MTBLS28 to NOREVA from 8 different universities around the world, and the calculation time of different normalization methods for the same dataset was also assessed.

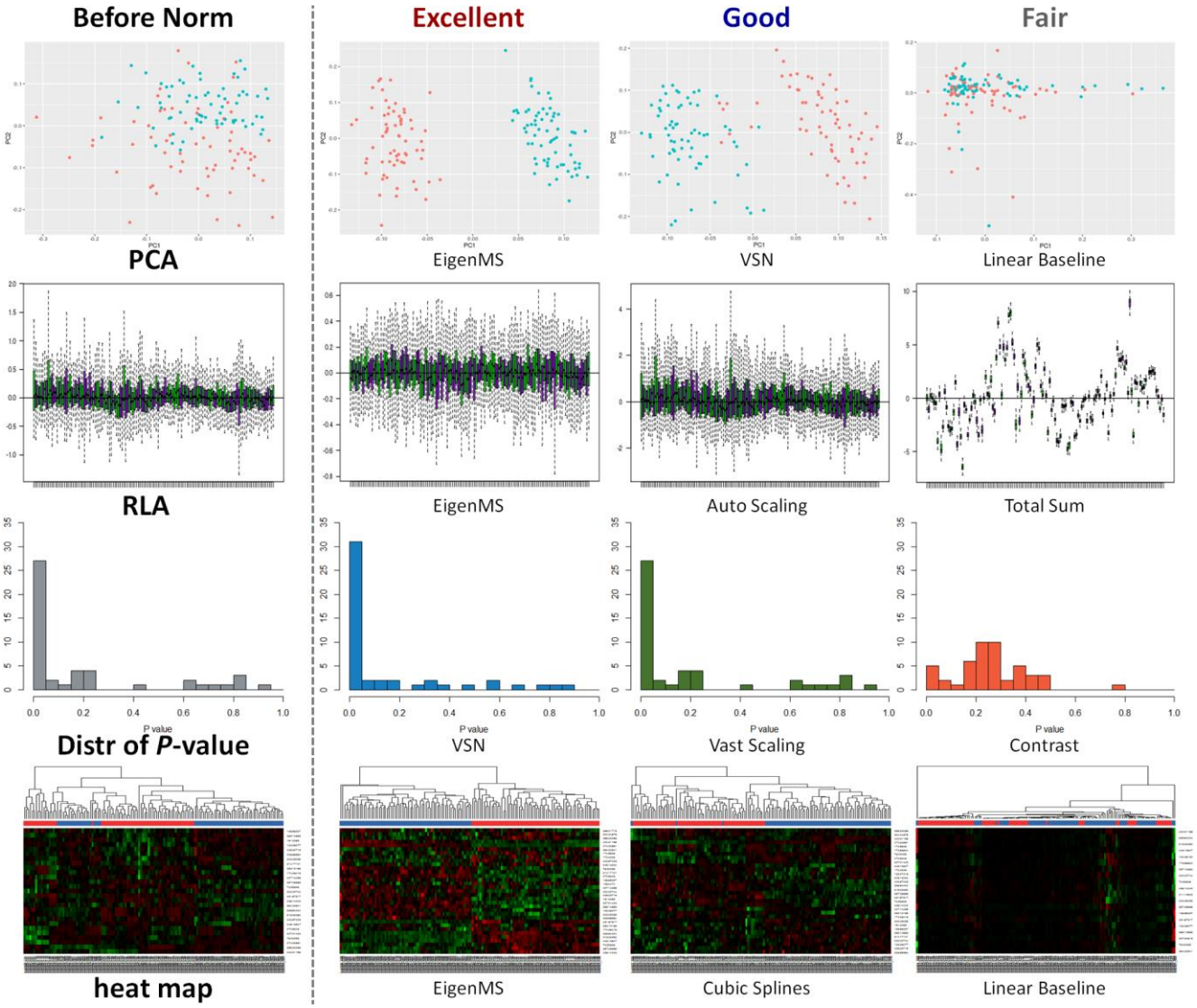
Procedures in NOREVA for Processing Metabolomics Data		Time Cost
Data Uploading	Imperial College, London, United Kingdom	1’15’’
	University of California, Berkeley, United States	2’50’’
	New York University, New York, United States	3’36’’
	Université de Paris VIII, Paris, France	3’41’’
	Goethe University Frankfurt, Frankfurt, Germany	3’10’’
	Novosibirsk State University, Novosibirsk, Russian	4’17’’
	Chongqing University, Chongqing, P. R. China	0’05’’
	Zhejiang University, Hangzhou, P. R. China	0’15’’
Data Preprocessing		0’04’’
Data Normalization	Auto Scaling	2’24’’
	Contrast	9’52’’
	Cubic Splines	2’21’’
	Cyclic Loess	3’01’’
	EigenMS	69’02’’
	Level Scaling	2’36’’
	Linear Baseline	1’37’’
	Log-transform	2’04’’
	Mean	9’05’’
	Median	9’48’’
	MSTUS	2’58’’
	Pareto Scaling	2’43’’
	Power Scaling	3’16’’
	PQN	9’32’’
	Quantile	9’07’’
	Range Scaling	3’22’’
	Total Sum	2’04’’
	Vast Scaling	2’50’’
	VSN	3’25’’
Performance Evaluation		~ 9’00’’

Table S4. Evaluation results of 4 criteria on benchmark dataset MTBLS79 (a full list of results for all measures in each criterion). The way calculating those measures under each criterion was described in **MATERIALS AND METHODS** and **Supplementary Methods**. Besides of quantitative measures, several qualitative ones under criterion ***a*** and ***b*** were also evaluated, and 3 performance levels were provided (Excellent, Good and Fair). Qualitative measures were evaluated by visual inspection, and examples illustrating how 3 performance levels were assigned were shown in **Supplementary Figure S1**.

Criterion	***a***					***b***		***c***	***d***
Measure	PMAD	PEV	PCV	PCA	RLA	distribution of *P*-value	heat map	consistency	AUC
Auto Scaling	0.8360	0.8810	1421.3951	Excellent	Good	Good	Fair	14.6500	0.8344
Contrast	0.7797	68.4340	342.4356	Fair	Fair	Fair	Fair	9.7500	0.6250
Cubic Splines	0.1393	0.0376	262.6381	Good	Excellent	Excellent	Good	13.7500	0.8322
Cyclic Loess	0.3188	0.2226	29.2876	Good	Excellent	Good	Fair	15.6500	0.8356
EigenMS	0.1799	0.0419	208.8599	Excellent	Excellent	Good	Excellent	16.4000	0.8010
Level Scaling	0.2890	0.1231	1421.3951	Good	Good	Good	Good	15.1000	0.8345
Linear Baseline	0.6035	9.4973	3618.3982	Fair	Good	Fair	Fair	6.3000	0.7072
Log-transform	0.1349	0.0242	1032.3457	Good	Excellent	Good	Good	14.7500	0.8168
Mean	0.3100	0.1245	8381.2036	Good	Excellent	Good	Good	14.7500	0.8213
Median	0.3100	0.1275	1033.6318	Good	Excellent	Good	Good	14.5500	0.8177
MSTUS	0.0064	0.0001	32.0893	Good	Excellent	Good	Good	14.3500	0.8405
Pareto Scaling	0.5320	0.3928	1421.3951	Good	Good	Good	Good	14.9500	0.8344
Power Scaling	0.1660	0.0392	1825.9541	Good	Good	Good	Excellent	14.9500	0.8314
PQN	0.3260	0.7871	389.2734	Fair	Good	Good	Fair	13.7000	0.8309
Quantile	0.2989	0.1174	340.6573	Excellent	Good	Excellent	Good	13.8000	0.8119
Range Scaling	0.1573	0.0313	1421.3951	Excellent	Good	Good	Excellent	15.3500	0.8344
Sum	2.4336	7.7602	2148.8569	Good	Fair	Fair	Fair	14.7000	0.7538
Vast Scaling	2.7200	10.3400	1421.3951	Good	Excellent	Good	Fair	15.0000	0.8344
VSN	0.5626	0.3700	285.9184	Good	Excellent	Excellent	Excellent	13.7500	0.8373

Supplementary FIGURES

Figure S1. Examples illustrating how normalization performances were evaluated for those qualitative measures provided in criterion ***a*** and ***b*** based on the benchmark dataset MTBLS79. There were two qualitative measures in each criterion (PCA and RLA in criterion ***a***; distribution of *P*-value and heat map in criterion ***b***), and three performance levels were assigned (Excellent, Good and Fair). These measures were evaluated by visual inspection which could be illustrated by the following examples.



REFERENCES

1. Hu, C.X. & Xu, G.W. Mass-spectrometry-based metabolomics analysis for foodomics. *Trac-Trend Anal. Chem.* **52**, 36-46 (2013).
2. Redestig, H. et al. Compensation for systematic cross-contribution improves normalization of mass spectrometry based metabolomics data. *Analytical chemistry* **81**, 7974-7980 (2009).
3. Astrand, M. Contrast normalization of oligonucleotide arrays. *J. Comput. Biol.* **10**, 95-102 (2003).
4. Workman, C. et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* **3**, research0048 (2002).
5. Dudoit, S., Yang, Y.H., Callow, M.J. & Speed, T.P. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat Sinica* **12**, 111-139 (2002).
6. Karpievitch, Y.V., Nikolic, S.B., Wilson, R., Sharman, J.E. & Edwards, L.M. Metabolomics data normalization with EigenMS. *PloS one* **9**, e116221 (2014).
7. van den Berg, R.A., Hoefsloot, H.C.J., Westerhuis, J.A., Smilde, A.K. & van der Werf, M.J. Centering, scaling, and transformations: improving the biological information content of metabolomics data. *Bmc Genomics* **7**, 142 (2006).
8. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185-193 (2003).
9. Purohit, P.V., Rocke, D.M., Viant, M.R. & Woodruff, D.L. Discrimination models using variance-stabilizing transformation of metabolomic NMR data. *OMICS* **8**, 118-130 (2004).
10. Andjelkovic, V. & Thompson, R. Changes in gene expression in maize kernel in response to water and salt stress. *Plant cell reports* **25**, 71-79 (2006).
11. De Livera, A.M., Olshansky, M. & Speed, T.P. Statistical analysis of metabolomics data. *Metabolomics Tools for Natural Product Discovery: Methods and Protocols*, 291-307 (2013).
12. Warrack, B.M. et al. Normalization strategies for metabonomic analysis of urine samples. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* **877**, 547-552 (2009).
13. Sysi-Aho, M., Katajamaa, M., Yetukuri, L. & Oresic, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC bioinformatics* **8**, 93 (2007).
14. Eriksson, L. et al. Using chemometrics for navigating in the large data sets of genomics, proteomics, and metabonomics (gpm). *Anal. Bioanal. Chem.* **380**, 419-429 (2004).
15. Brodsky, L., Moussaieff, A., Shahaf, N., Aharoni, A. & Rogachev, I. Evaluation of peak picking quality in LC-MS metabolomics data. *Anal. Chem.* **82**, 9177-9187 (2010).
16. Dieterle, F., Ross, A., Schlotterbeck, G. & Senn, H. Probabilistic quotient normalization as robust method to account for dilution of complex biological mixtures. Application in ^{1}H NMR metabonomics. *Anal. Chem.* **78**, 4281-4290 (2006).
17. Smilde, A.K., van der Werf, M.J., Bijlsma, S., van der Werff-van der Vat, B.J. & Jellema, R.H. Fusion of mass spectrometry-based metabolomics data. *Anal. Chem.* **77**, 6729-6736 (2005).

18. De Livera, A.M. et al. Normalizing and integrating metabolomics data. *Analytical chemistry* **84**, 10768-10776 (2012).

19. De Livera, A.M. et al. Statistical methods for handling unwanted variation in metabolomics data. *Analytical chemistry* **87**, 3606-3615 (2015).

20. Gullberg, J., Jonsson, P., Nordstrom, A., Sjostrom, M. & Moritz, T. Design of experiments: an efficient strategy to identify factors influencing extraction and derivatization of Arabidopsis thaliana samples in metabolomic studies with gas chromatography/mass spectrometry. *Analytical biochemistry* **331**, 283-295 (2004).

21. Keun, H.C. et al. Improved analysis of multivariate data by variable stability scaling: application to NMR-based metabolic profiling. *Anal Chim Acta* **490**, 265-276 (2003).

22. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A. & Vingron, M. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18 Suppl 1**, S96-104 (2002).

23. Karp, N.A. et al. Addressing accuracy and precision issues in iTRAQ quantitation. *Mol. Cell Proteomics* **9**, 1885-1897 (2010).

24. Kohl, S.M. et al. State-of-the art data normalization methods improve NMR-based metabolomic analysis. *Metabolomics* **8**, 146-160 (2012).

25. Gromski, P.S., Xu, Y., Hollywood, K.A., Turner, M.L. & Goodacre, R. The influence of scaling metabolomics data on model classification accuracy. *Metabolomics* **11**, 684-695 (2015).

26. Weber, C.M. et al. Evaluation of a gas sensor array and pattern recognition for the identification of bladder cancer from urine headspace. *Analyst* **136**, 359-364 (2011).

27. Struck, W. et al. Liquid chromatography tandem mass spectrometry study of urinary nucleosides as potential cancer markers. *J. Chromatogr. A* **1283**, 122-131 (2013).

28. Jauhiainen, A. et al. Normalization of metabolomics data with applications to correlation maps. *Bioinformatics* **30**, 2155-2161 (2014).

29. Shi, X. et al. Metabolomic analysis of the effects of polychlorinated biphenyls in nonalcoholic fatty liver disease. *J. Proteome Res.* **11**, 3805-3815 (2012).

30. Saccenti, E. Correlation Patterns in Experimental Data Are Affected by Normalization Procedures: Consequences for Data Analysis and Network Inference. *Journal of proteome research* **16**, 619-634 (2017).

31. Contrepois, K., Jiang, L. & Snyder, M. Optimized Analytical Procedures for the Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)-Mass Spectrometry. *Molecular & cellular proteomics : MCP* **14**, 1684-1695 (2015).

32. Cleveland, W.S. & Devlin, S.J. Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American statistical association* **83**, 596-610 (1988).

33. Ballman, K.V., Grill, D.E., Oberg, A.L. & Therneau, T.M. Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics* **20**, 2778-2786 (2004).

34. Ejigu, B.A. et al. Evaluation of normalization methods to pave the way towards large-scale LC-MS-based metabolomics profiling experiments. *OMICS* **17**, 473-485 (2013).

35. Valikangas, T., Suomi, T. & Elo, L.L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings in bioinformatics* (2016).

36. Yuan, Y. et al. Metabolomic analyses of banana during postharvest senescence by ^{1}H-high resolution-NMR. *Food Chem.* **218**, 406-412 (2017).

37. Backshall, A., Sharma, R., Clarke, S.J. & Keun, H.C. Pharmacometabonomic profiling as a predictor of toxicity in patients with inoperable colorectal cancer treated with capecitabine. *Clin. Cancer Res.* **17**, 3019-3028 (2011).

38. Klepacki, J. et al. Amino acids in a targeted versus a non-targeted metabolomics LC-MS/MS assay. Are the results consistent? *Clinical biochemistry* **49**, 955-961 (2016).

39. Ibarra, R. et al. Metabolomic analysis of liver tissue from the VX2 rabbit model of secondary liver tumors. *HPB Surg.* **2014**, 310372 (2014).

40. Ting, L. et al. Normalization and statistical analysis of quantitative proteomics data generated by metabolic labeling. *Molecular & cellular proteomics : MCP* **8**, 2227-2242 (2009).

41. Jacob, C.C., Dervilly-Pinel, G., Biancotto, G. & Le Bizec, B. Evaluation of specific gravity as normalization strategy for cattle urinary metabolome analysis. *Metabolomics* **10**, 627-637 (2014).

42. Chen, Y. et al. Combination of injection volume calibration by creatinine and MS signals' normalization to overcome urine variability in LC-MS-based metabolomics studies. *Anal. Chem.* **85**, 7659-7665 (2013).

43. Craig, A., Cloarec, O., Holmes, E., Nicholson, J.K. & Lindon, J.C. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical chemistry* **78**, 2262-2267 (2006).

44. Yang, J., Zhao, X., Lu, X., Lin, X. & Xu, G. A data preprocessing strategy for metabolomics to reduce the mask effect in data analysis. *Front. Mol. Biosci.* **2**, 4 (2015).

45. Leichtle, A.B. et al. Serum amino acid profiles and their alterations in colorectal cancer. *Metabolomics* **8**, 643-653 (2012).

46. Kirwan, J.A., Weber, R.J., Broadhurst, D.I. & Viant, M.R. Direct infusion mass spectrometry metabolomics dataset: a benchmark for data processing and quality control. *Scientific data* **1**, 140012 (2014).

47. Lee, J. et al. Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. *Analytical sciences : the international journal of the Japan Society for Analytical Chemistry* **28**, 801-805 (2012).

48. Vogl, F.C. et al. Evaluation of dilution and normalization strategies to correct for urinary output in HPLC-HRTOFMS metabolomics. *Analytical and bioanalytical chemistry* **408**, 8483-8493 (2016).

49. Kultima, K. et al. Development and evaluation of normalization methods for label-free relative quantification of endogenous peptides. *Mol. Cell Proteomics* **8**, 2285-2295 (2009).

50. Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets. *Journal of proteome research* **13**, 3114-3120 (2014).

51. Lindon, J., Holmes, E. & Nicholson, J. Pattern recognition methods and applications in biomedical magnetic resonance. *Progress in Nuclear Magnetic Resonance Spectroscopy* **39**, 1-40 (2001).

52. Ritchie, M.E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47 (2015).

53. Risso, D., Ngai, J., Speed, T.P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature biotechnology* **32**, 896-902 (2014).

54. Luan, H. et al. Non-targeted metabolomics and lipidomics LC-MS data from maternal plasma of 180 healthy pregnant women. *GigaScience* **4**, 16 (2015).

55. Zhou, X., Oshlack, A. & Robinson, M.D. miRNA-Seq normalization comparisons need improvement. *Rna* **19**, 733-734 (2013).

56. Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F. & Wehrens, R. A benchmark spike-in data set for biomarker identification in metabolomics. *J Chemometr* **26**, 16-24 (2012).

57. Canales, R.D. et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature biotechnology* **24**, 1115-1122 (2006).

58. Mathe, E.A. et al. Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer. *Cancer research* **74**, 3259-3270 (2014).