Supplementary Materials for

EigenRF: an improved metabolomics normalization method with scores for reproducibility evaluation on importance rankings of different metabolites

Chencheng Tang[1,2], Dongfang Huang[1], Xudong Xing[1,*] and Hua Yang[1,*]

[1]State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 211198, China
[2]School of Science, China Pharmaceutical University, Nanjing 211198, China

*Corresponding author. State Key Laboratory of Natural Medicines, China Pharmaceutical University, Nanjing 211198, China. E-mail: yanghuacpu@126.com (H.Y.)

Table S1 The information of ten other normalization methods.

Type of method	Name	Annotation	Implementation with R package	Description
Method based on QC samples	RLSC	Robust locally weighted scatter plot smoothing.[23]	NormalyzeMets[3]	Correct for batch effects by applying robust locally weighted scatter plot smoothing to quality control samples.
	RSC	Robust spline correction.[24]	pmp	Fit a smoothing spline to quality control samples and apply it to normalize the data across different batches.
	SVR	Support vector regression.	MetNormalizer[25]	Apply support vector regression to model and correct for intensity drift in metabolomics data based on quality control samples.
	SERRF	Systematic error elimination employing random forest.	SERRF[17]	Utilize random forest to model and correct for systematic errors in large-scale untargeted lipidomics data based on quality control samples.
Method based on Internal Standard	SIS	Single internal standard.[26]	CCMN[28]	Apply a single internal standard to normalize data, assuming the standard's response is indicative of overall technical variability.
	NOMIS	Optimal normalization factor based on multiple internal standards.[27]		Determine the optimal normalization factor for each metabolite by utilizing the variability information from multiple internal standards and their correlation to the metabolites.
	CCMN	Cross-contribution compensation based on multiple internal standards.[28]		Use multiple internal standards, and apply linear regression and PCA to account for cross-contribution effects.
Method based on biological samples	ber	Linear fitting based on position/scale.	ber[6]	Employ a two-stage regression approach to adjust for location and scale batch effects.
	WaveICA	Wavelet transform algorithm based on independent component analysis.	WaveICA[19]	Utilize wavelet transform and independent component analysis to remove batch effects by decomposing data into multi-scale components and eliminating non-biological variations.

Combination method	ISWSVR	Based on the combination of B-MIS and SVR.	Norm-ISWSVR[4]	Integrate internal standard normalization with support vector regression to remove systematic errors.

Note S1 The information of the other evaluation metrics.

a) RSD of QCs

The relative standard deviation (RSD) of features in QC samples is a commonly used metric to evaluate the reproducibility of biomarker discovery studies. Here, we use RSD to represent RSD of QCs. In theory, the signal value of a feature on all QC samples should be the same. The higher the RSD, the more unwanted variation and the lower the reproducibility. It is generally accepted that the number of features with RSD less than 0.3 should account for at least 70% of the total number of features. Additionally, the median RSD of each batch is required to account for systematic errors due to batch effects. The ability of the normalization method to eliminate systematic errors can be evaluated by comparing the median RSD of each batch before and after normalization. The RSD of the i-th feature is calculated as follows:

$$RSD_i = \frac{S_i^{qc}}{\bar{y}_i^{qc}}, \tag{13}$$

where RSD_i, S_i^{qc} and $\bar{y}_i^{qc}$ denote the relative standard deviation, standard deviation, and mean of the i-th feature in the QC samples, respectively.

b) Run plots

The run plot of the feature is represented by a scatter plot, with the signal values as the vertical axis and the injection sequence or injection time as the horizontal axis. The run plot of the internal standards and non-differential metabolites can effectively visualize the presence of batch effects and signal drift. In theory, the signal values of an internal standard should remain consistent across all samples. However, the batch effects may cause variations between batches, while signal drift may result in a discernible trend along the injection sequence.

c) PCA plots

Principal component analysis (PCA) is performed on all features to generate PCA plots.

Batch-based PCA plots: The sample points are color-coded based on batch information, illustrating the batch effect through the clustering of samples from the same batch.

Group-based PCA plots: The sample points are color-coded based on classification information, allowing for the assessment of normalization methods by observing alterations in the clustering of samples within each category before and after normalization.

d) OPLS-DA score plots

Orthogonal partial least squares discriminant analysis is performed on all features of biological samples corresponding to specific categories, resulting in the generation of score plots. The performance of the normalization method can be evaluated by examining the classification shifts of various sample categories before and after normalization.

e) ROC and PR curves

The data are stratified and then randomly sampled at a ratio of 7:3 to be divided into a training set and a test set. For the training set, differential metabolites are screened according to the criteria of VIP $\geq$ 1, which is derived from the OPLS-DA model, and the FDR-adjusted *p*-value $<$ 0.05, which is obtained from the Mann-Whitney U test. A classification model is subsequently constructed with differential metabolites as predictors and grouping information as response.

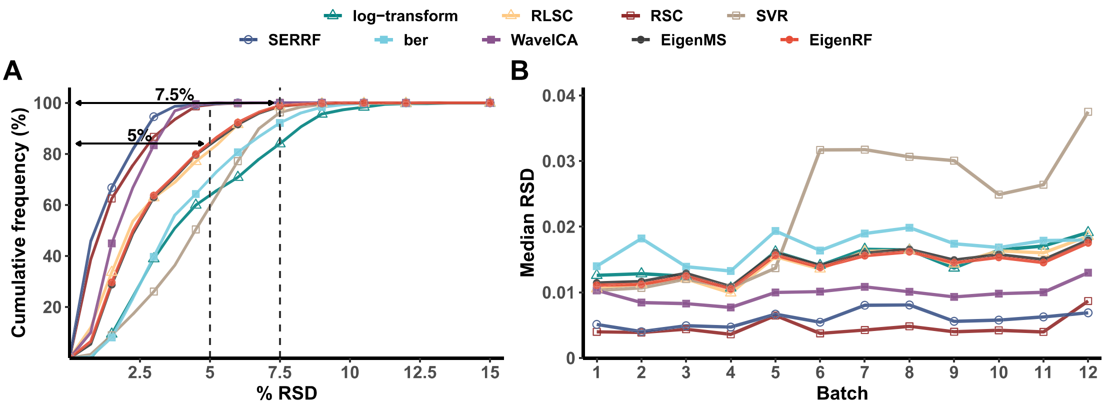
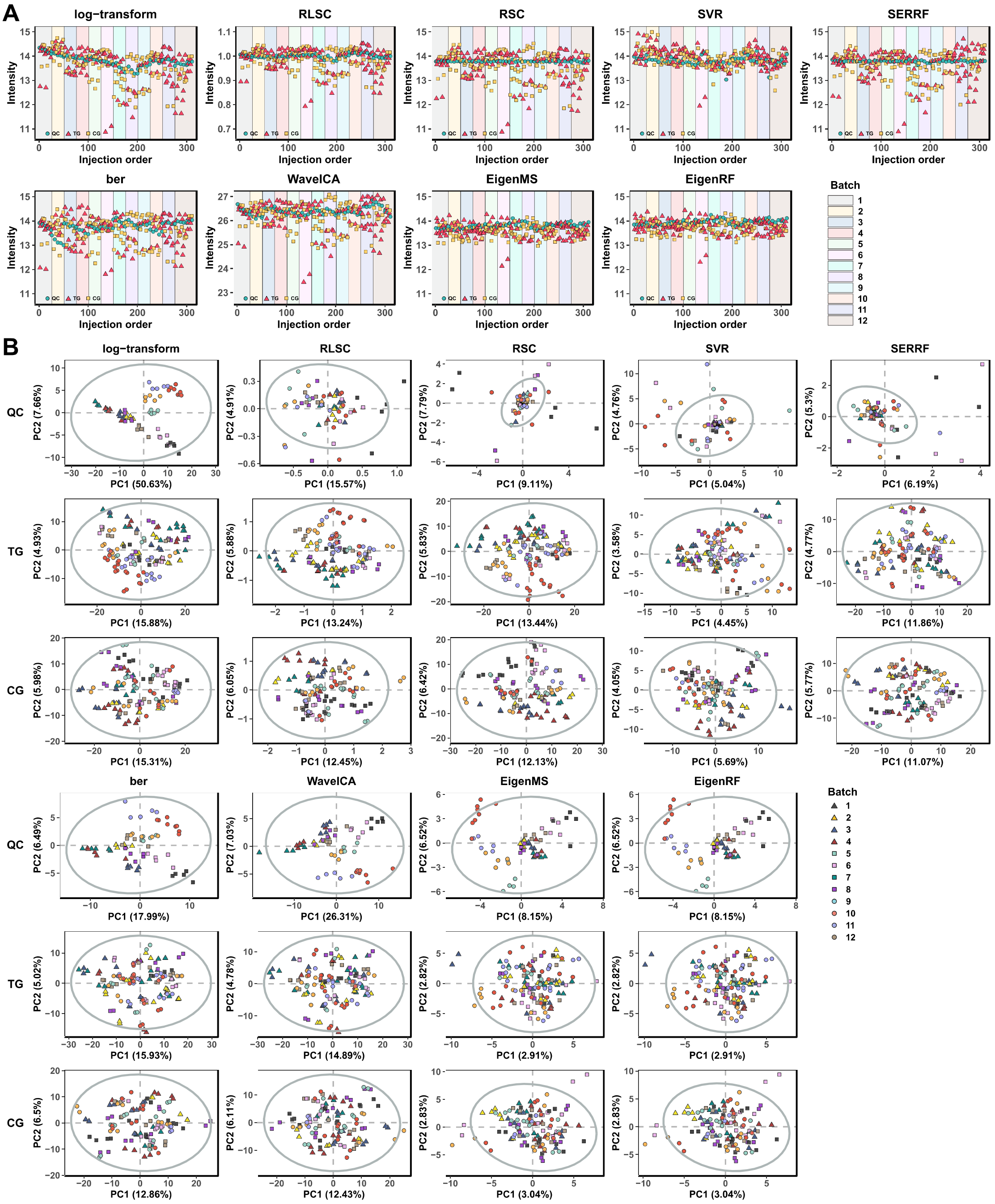


Fig. S1 Cumulative frequency distribution of the RSDs of QC samples (A) and the median RSD of each batch of QC samples (B) in the BCPUM dataset.


A
log−transform
RLSC
RSC
SVR
SERRF
ber
WaveICA
EigenMS
EigenRF
Intensity
Injection order
QC
TG
CG
Batch
1
2
3
4
5
6
7
8
9
10
11
12
B
log−transform
RLSC
RSC
SVR
SERRF
QC
TG
CG
PC1 (50.63%)
PC2 (7.66%)
PC1 (15.57%)
PC2 (4.91%)
PC1 (9.11%)
PC2 (7.79%)
PC1 (5.04%)
PC2 (4.76%)
PC1 (6.19%)
PC2 (5.3%)
PC1 (15.88%)
PC2 (4.93%)
PC1 (13.24%)
PC2 (5.88%)
PC1 (13.44%)
PC2 (5.83%)
PC1 (4.45%)
PC2 (3.58%)
PC1 (11.86%)
PC2 (4.77%)
PC1 (15.31%)
PC2 (5.98%)
PC1 (12.45%)
PC2 (6.05%)
PC1 (12.13%)
PC2 (6.42%)
PC1 (5.69%)
PC2 (4.05%)
PC1 (11.07%)
PC2 (5.77%)
ber
WaveICA
EigenMS
EigenRF
PC1 (17.99%)
PC2 (6.49%)
PC1 (26.31%)
PC2 (7.03%)
PC1 (8.15%)
PC2 (6.52%)
PC1 (8.15%)
PC2 (6.52%)
PC1 (15.93%)
PC2 (5.02%)
PC1 (14.89%)
PC2 (4.78%)
PC1 (2.91%)
PC2 (2.82%)
PC1 (2.91%)
PC2 (2.82%)
PC1 (12.86%)
PC2 (6.5%)
PC1 (12.43%)
PC2 (6.11%)
PC1 (3.04%)
PC2 (2.83%)
PC1 (3.04%)
PC2 (2.83%)
Batch
1
2
3
4
5
6
7
8
9
10
11
12

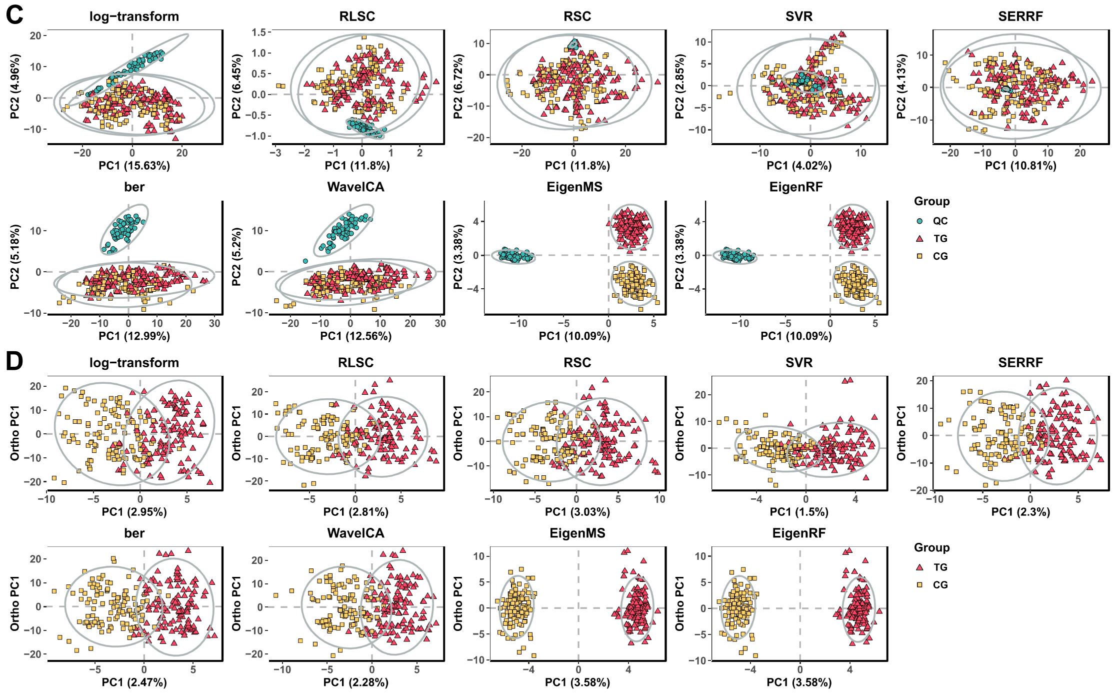


Fig. S2 Run plots of the first feature (A), batch-based PCA plots of the QC, TG, and CG samples (B), group-based PCA plots (C), and group-based OPLS-DA score plots (D) in the BCPUM dataset.

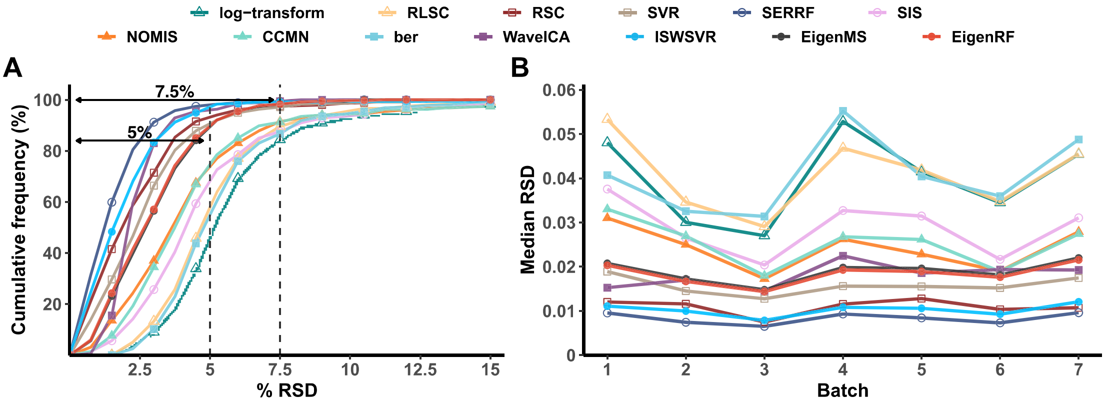
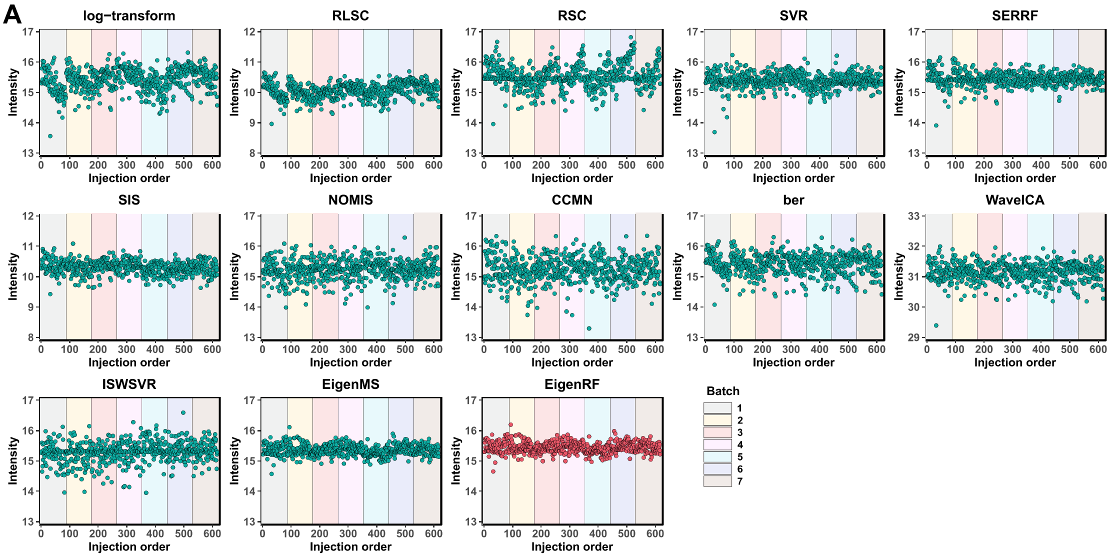
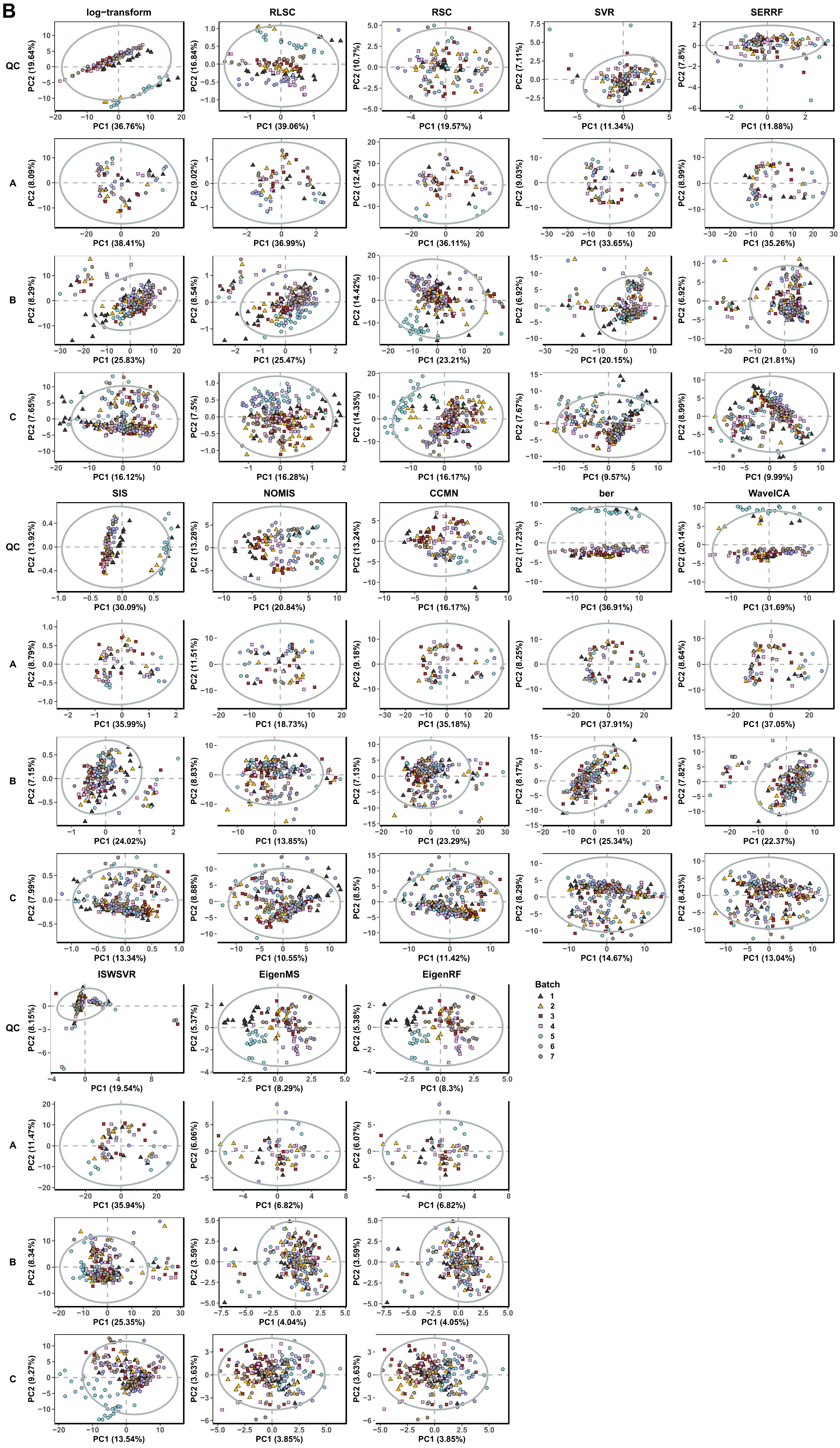


Fig. S3 Cumulative frequency distribution of the RSDs of QC samples (A) and the median RSD of each batch of QC samples (B) in the GCPPM dataset.


A
log−transform
RLSC
RSC
SVR
SERRF
SIS
NOMIS
CCMN
ber
WaveICA
ISWSVR
EigenMS
EigenRF
Intensity
Injection order
0
100
200
300
400
500
600
13
14
15
16
17
8
9
10
11
12
29
30
31
32
33
Batch
1
2
3
4
5
6
7


B
log−transform
RLSC
RSC
SVR
SERRF
QC
PC2 (19.64%)
PC1 (36.76%)
PC2 (16.84%)
PC1 (39.06%)
PC2 (10.7%)
PC1 (19.57%)
PC2 (7.11%)
PC1 (11.34%)
PC2 (7.8%)
PC1 (11.88%)
A
PC2 (8.09%)
PC1 (38.41%)
PC2 (9.02%)
PC1 (36.99%)
PC2 (12.4%)
PC1 (36.11%)
PC2 (9.03%)
PC1 (33.65%)
PC2 (8.99%)
PC1 (35.26%)
B
PC2 (8.29%)
PC1 (25.83%)
PC2 (8.54%)
PC1 (25.47%)
PC2 (14.42%)
PC1 (23.21%)
PC2 (6.92%)
PC1 (20.15%)
PC2 (6.92%)
PC1 (21.81%)
C
PC2 (7.65%)
PC1 (16.12%)
PC2 (7.5%)
PC1 (16.28%)
PC2 (14.35%)
PC1 (16.17%)
PC2 (7.67%)
PC1 (9.57%)
PC2 (8.99%)
PC1 (9.99%)
SIS
NOMIS
CCMN
ber
WaveICA
QC
PC2 (13.92%)
PC1 (30.09%)
PC2 (13.28%)
PC1 (20.84%)
PC2 (13.24%)
PC1 (16.17%)
PC2 (17.23%)
PC1 (36.91%)
PC2 (20.14%)
PC1 (31.69%)
A
PC2 (8.79%)
PC1 (35.99%)
PC2 (11.51%)
PC1 (18.73%)
PC2 (9.18%)
PC1 (35.18%)
PC2 (8.25%)
PC1 (37.91%)
PC2 (8.64%)
PC1 (37.05%)
B
PC2 (7.15%)
PC1 (24.02%)
PC2 (8.83%)
PC1 (13.85%)
PC2 (7.13%)
PC1 (23.29%)
PC2 (8.17%)
PC1 (25.34%)
PC2 (7.82%)
PC1 (22.37%)
C
PC2 (7.99%)
PC1 (13.34%)
PC2 (8.88%)
PC1 (10.55%)
PC2 (8.5%)
PC1 (11.42%)
PC2 (8.29%)
PC1 (14.67%)
PC2 (8.43%)
PC1 (13.04%)
ISWSVR
EigenMS
EigenRF
QC
PC2 (8.15%)
PC1 (19.54%)
PC2 (5.37%)
PC1 (8.29%)
PC2 (5.38%)
PC1 (8.3%)
A
PC2 (11.47%)
PC1 (35.94%)
PC2 (6.06%)
PC1 (6.82%)
PC2 (6.07%)
PC1 (6.82%)
B
PC2 (8.34%)
PC1 (25.35%)
PC2 (3.59%)
PC1 (4.04%)
PC2 (3.59%)
PC1 (4.05%)
C
PC2 (9.27%)
PC1 (13.54%)
PC2 (3.63%)
PC1 (3.85%)
PC2 (3.63%)
PC1 (3.85%)
Batch
1
2
3
4
5
6
7

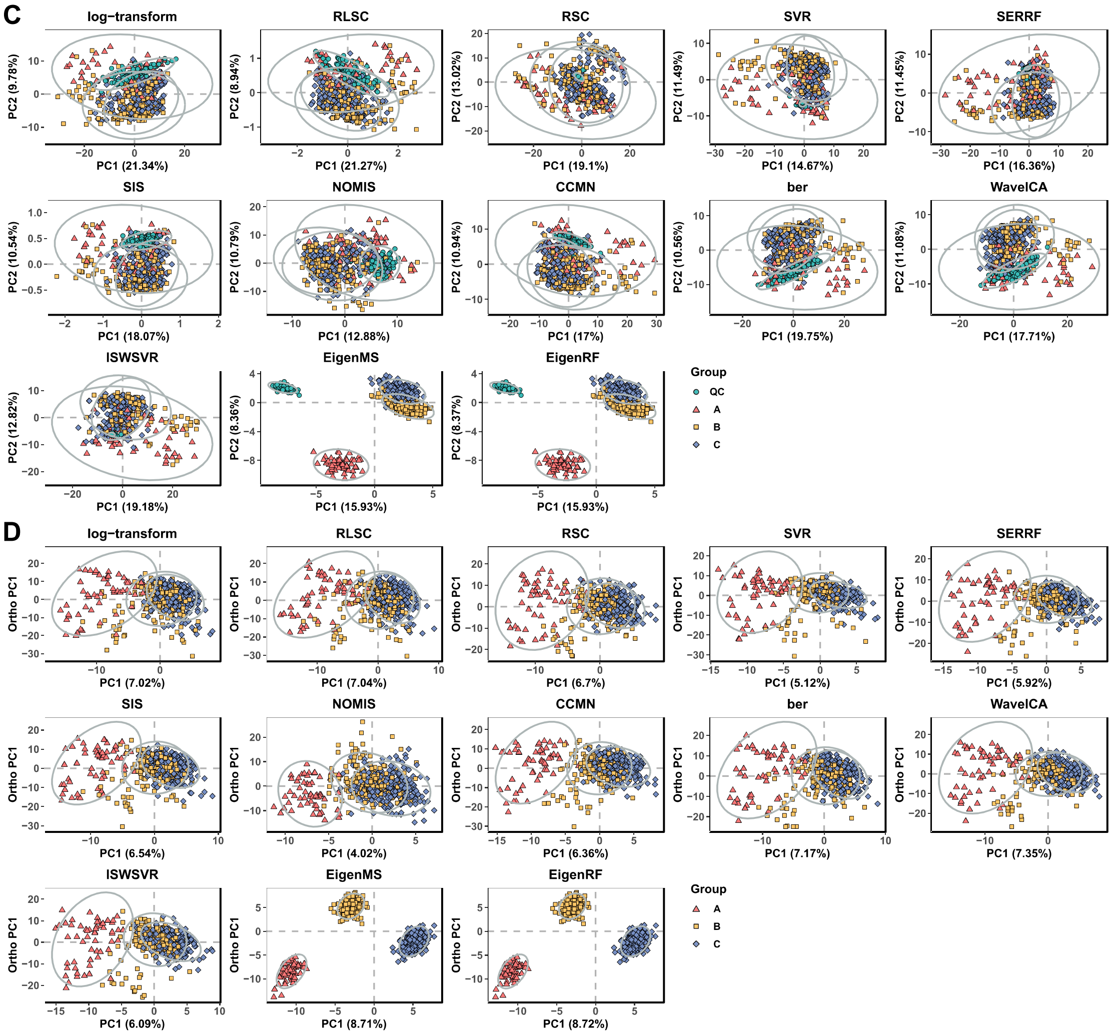
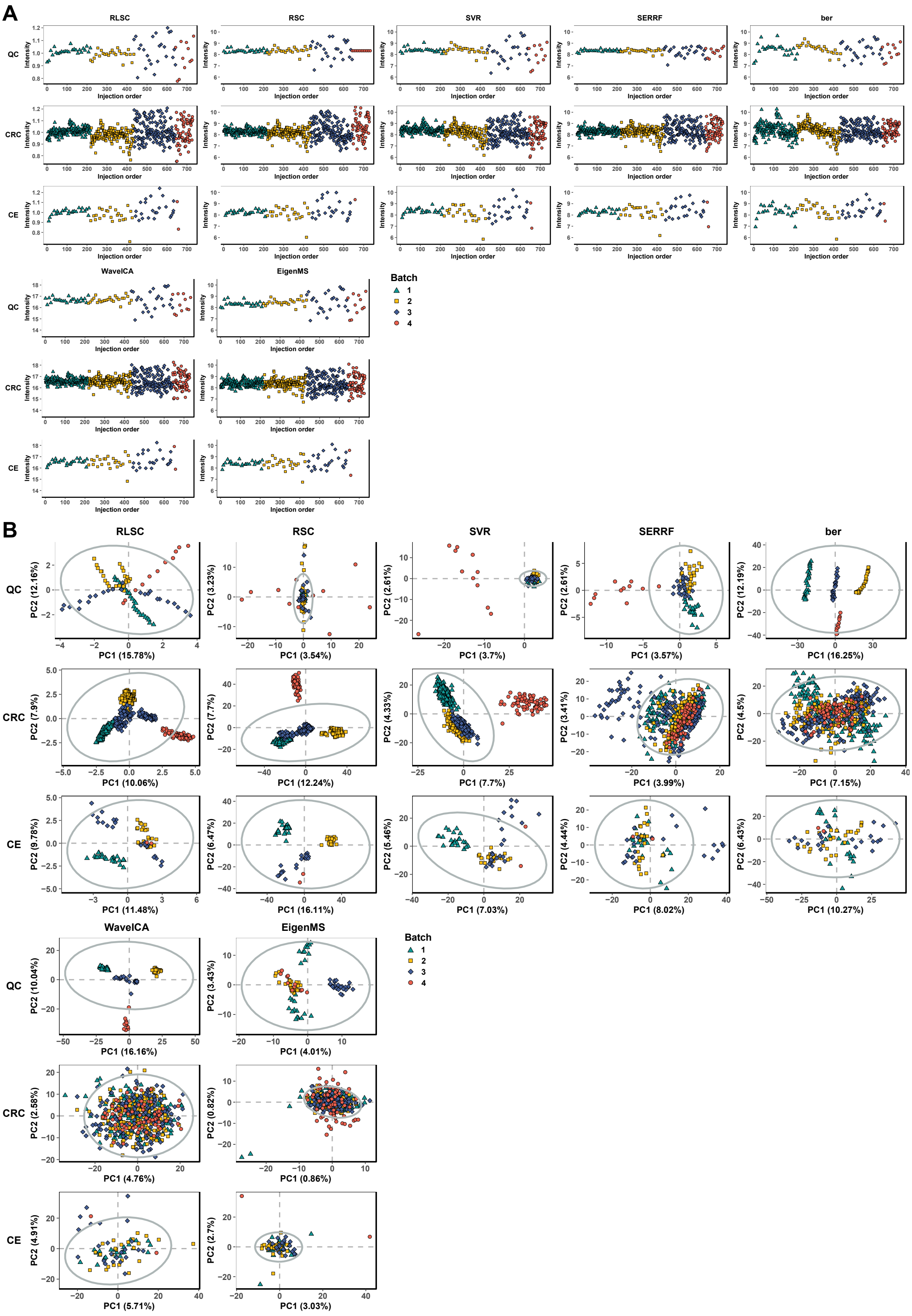


Fig. S4 Run plots of the internal standard for all samples (A), batch-based PCA plots of the QC, A, B, and C samples (B), group-based PCA plots (C), and group-based OPLS-DA score plots (D) in the GCPPM dataset.


A
RLSC
RSC
SVR
SERRF
ber
QC
CRC
CE
Intensity
Injection order
WaveICA
EigenMS
Batch
1
2
3
4
B
RLSC
RSC
SVR
SERRF
ber
QC
PC1 (15.78%)
PC2 (12.16%)
PC1 (3.54%)
PC2 (3.23%)
PC1 (3.7%)
PC2 (2.61%)
PC1 (3.57%)
PC2 (2.61%)
PC1 (16.25%)
PC2 (12.19%)
CRC
PC1 (10.06%)
PC2 (7.9%)
PC1 (12.24%)
PC2 (7.7%)
PC1 (7.7%)
PC2 (4.33%)
PC1 (3.99%)
PC2 (3.41%)
PC1 (7.15%)
PC2 (4.5%)
CE
PC1 (11.48%)
PC2 (9.78%)
PC1 (16.11%)
PC2 (6.47%)
PC1 (7.03%)
PC2 (5.46%)
PC1 (8.02%)
PC2 (4.44%)
PC1 (10.27%)
PC2 (6.43%)
WaveICA
EigenMS
QC
PC1 (16.16%)
PC2 (10.04%)
PC1 (4.01%)
PC2 (3.43%)
CRC
PC1 (4.76%)
PC2 (2.58%)
PC1 (0.86%)
PC2 (0.82%)
CE
PC1 (5.71%)
PC2 (4.91%)
PC1 (3.03%)
PC2 (2.7%)
Batch
1
2
3
4

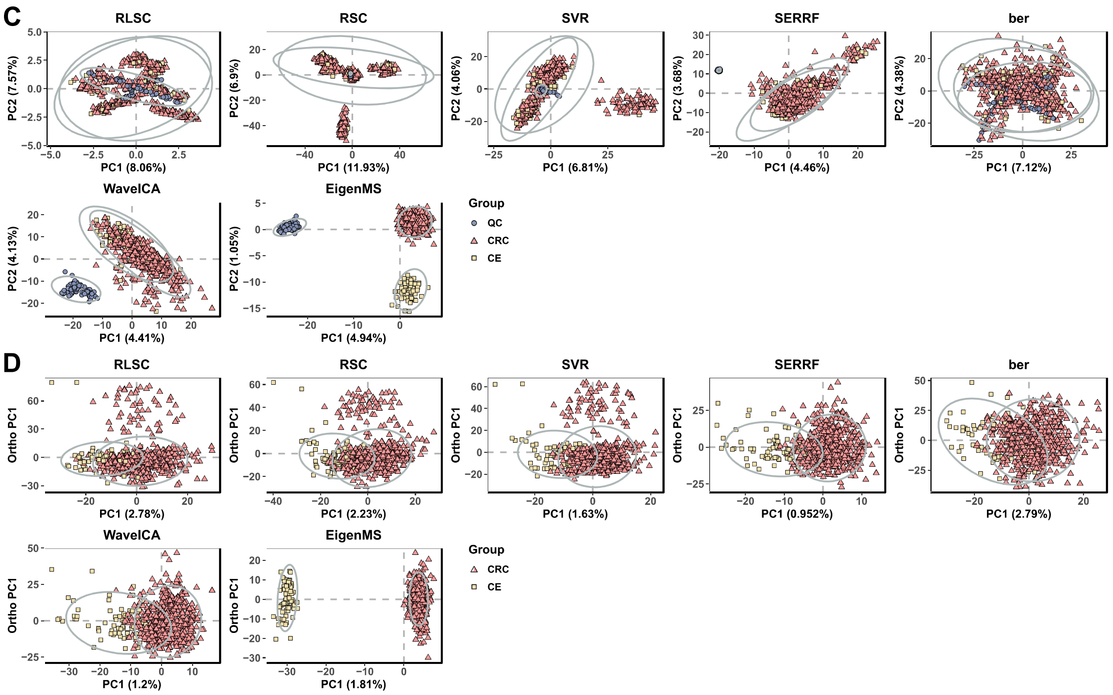


Fig. S5 Run plots of the first feature (A), batch-based PCA plots of the QC, CRC, and CE samples (B), group-based PCA plots (C), and group-based OPLS-DA score plots (D) in the ACPPM dataset.

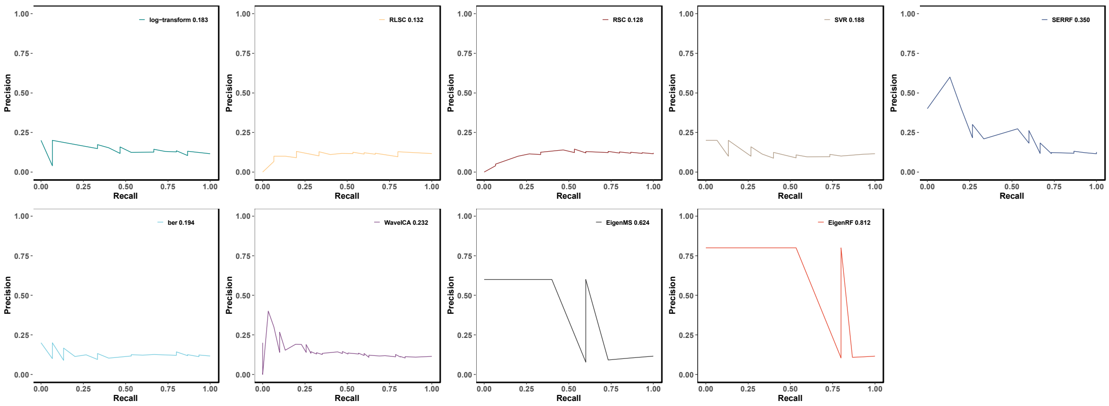


Fig. S6 PR plots for CE samples of the SVM classification model based on the differential metabolites between the CRC and CE samples in the ACPPM dataset. In the PR plots, the horizontal axis and the vertical axis are respectively the mean recall rate and the mean precision rate of ten-fold cross-validation, and the mean AUC values in the legends.

References

23 W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, A. Knowles, J.D. Halsall, J.N. Haselden, I.D. Wilson, A.W. Nicholls, D.B. Kell, R. Goodacre and HUSERMET Consortium, *Nat. Protoc.*, 2011, **6**, 1060-1083.

24 J.A. Kirwan, D.I. Broadhurst, R.L. Davidson and M.R. Viant, *Anal. Bioanal. Chem.*, 2013, **405**, 5147-5157.

25 X. Shen, X. Gong, Y. Cai, Y. Guo, J. Tu, H. Li, T. Zhang, J. Wang, F. Xue and Z. Zhu, *Metabolomics*, 2016, **12**, 89.

26 J. Gullberg, P. Jonsson, A. Nordström, M. Sjöström and T. Moritz, *Anal. Biochem.*, 2004, **331**, 283-295.

27 M. Sysi-Aho, M. Katajamaa, L. Yetukuri and M. Orešič, *BMC Bioinformatics*, 2007, **8**, 93.

28 H. Redestig, A. Fukushima, H. Stenlund, T. Moritz, M. Arita, K. Saito and M. Kusano, *Anal. Chem.*, 2009, **81**, 7974-7980.