Supporting Information

# Systematic Error Removal using Random Forest (SERRF) for Normalizing Large-Scale Untargeted Lipidomics Data

Sili Fan,[†] Tobias Kind,[†] Tomas Cajka,[†,‡] Stanley L. Hazen,[▽,○] W. H. Wilson Tang,[▽,○] Rima Kaddurah-Daouk,[–] Marguerite R. Irvin,[§] Donna K. Arnett,[⊥] Dinesh K. Barupal,[†] and Oliver Fiehn[*,†]

†West Coast Metabolomics Center, UC Davis Genome Center, University of California, Davis, 451 Health Sciences Drive, Davis, California 95616, United States

‡Department of Metabolomics, Institute of Physiology CAS, Videnska 1083, 14220 Prague, Czech Republic

§Department of Epidemiology, University of Alabama at Birmingham, 1720 2nd Ave S, Birmingham, Alabama, 35294, United States

⊥College of Public Health, University of Kentucky, Lexington, 121 Washington Ave, Kentucky, 40508, United States

–Department of Psychiatry and Behavioral Sciences; Department of Medicine and the Duke Institute for Brain Sciences, Duke University, Durham, NC, 27708 USA

▽ Department of Cellular & Molecular Medicine, Cleveland Clinic, Cleveland, OH 44195, USA

O Department of Cardiovascular Medicine, Cleveland Clinic, Cleveland, OH 44195, USA

slfan@ucdavis.edu
tkind@ucdavis.edu
tcajka@ucdavis.edu
HAZENS@ccf.org
TANGW@ccf.org
rima.kaddurahdaouk@duke.edu
irvinr@uab.edu
donna.arnett@uky.edu
dinkumar@ucdavis.edu
ofiehn@ucdavis.edu

## Supporting Files

## Supporting Text T 1

Code to achieve the following conclusion (taken from the main manuscript) is given below: *“Every increment of 5% of standard deviation for a metabolite with a small effect size needs 41 more samples to achieve 80% statistical power”.*

```
# Load R library pwr
library(pwr)
# Simulate a variable with small effect size of 0.2 (Cohen's d)
# set mean difference of 1.
mean_diff = 1
# standard deviation of 5
sd = 5
# calculate the sample size difference when standard deviation increases 5%
ceiling(pwr.t.test(d = mean_diff/(sd*1.05), sig.level = 0.05, power = 0.8, type = c("two.sample"))$n - pwr.t.test(d = mean_diff/sd, sig.level = 0.05, power = 0.8, type = c("two.sample"))$n)
# 41 # 41 SAMPLES NEEDED TO MAINTAIN A 80% POW
```

## Supporting Figure S1

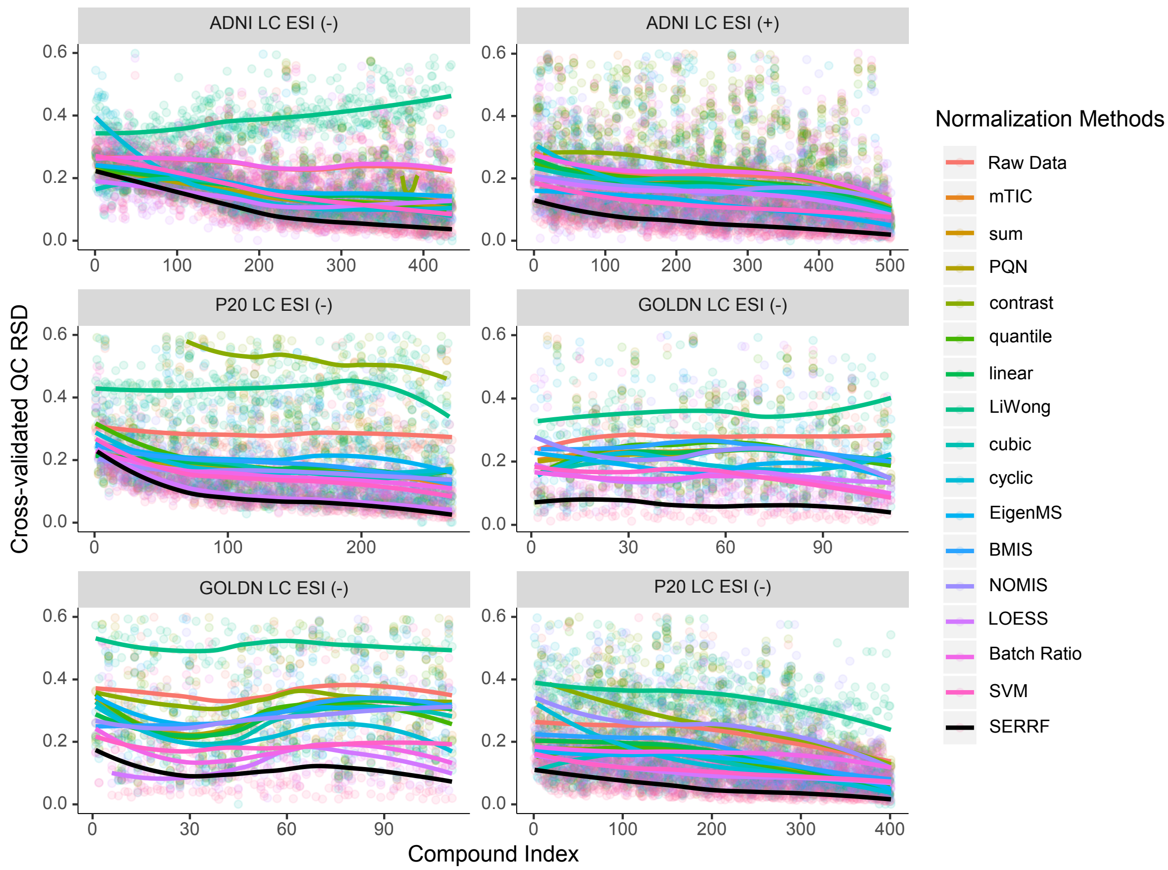


***Figure S1*** Cross-validated Relative Standard Deviations for QC samples (cvRSD) for each compound in six datasets. The x-axis is the compound index sorted by the average intensity in a descending order. The y-axis is the cross-validated RSD for each compound. The cvRSD after SERRF normalization is almost uniformly lower than those of other methods, indicating that SERRF normalization can reduce systematic variation independently from the peak intensities in all the datasets.

## Supporting Figure S2

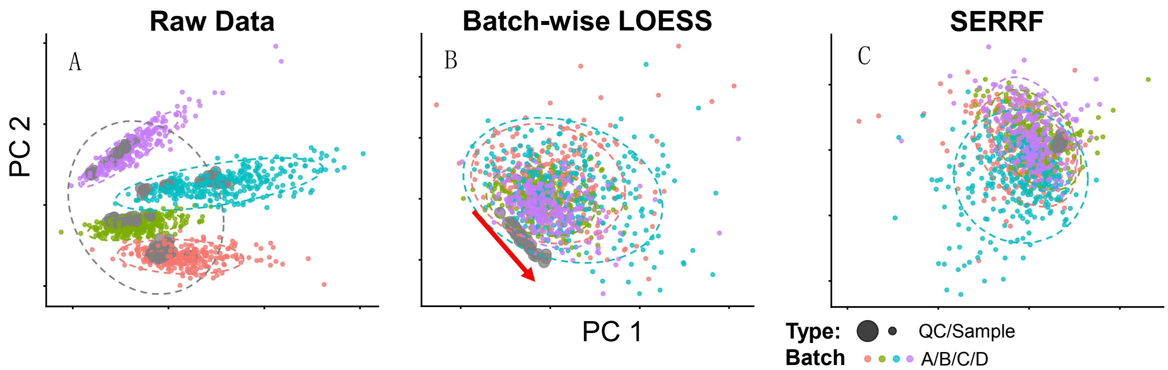


***Figure S2*** Principal component analysis score plot of raw data, batch-wise LOESS and SERRF normalized data. Dataset: P20 CSH-QTOF MSMS_lipidomics, ESI (-). Raw data, left panel. QC samples are given in grey. Mid panel: Without utilizing the information given by correlation between compounds, a technical drift remains after batch-wise LOESS normalization (QC samples, grey). Right panel: SERRF eliminates the residual drift as shown in the score plot (QC samples, grey, are clustered tightly)

## Supporting Table S1

Summary of 15 commonly used sample normalization approaches.

| Normalization method | Description | R package | Type |
|---|---|---|---|
| **mTIC normalization** | Normalize compounds by the sum of all identified metabolites | metabox | data-driven normalization |
| **Sum normalization** | Normalize compounds by the sum of all compounds. The sums of the intensities for each experimental run are forced to be equal. | metabox | data-driven normalization |
| **Median normalization** | Normalize compounds by the median average intensity of all the compounds. The use of the median method is found to be more practical than the sum method, especially in situations where several saturated abundances may be associated with some of the factors of interest | metabox | data-driven normalization |
| **Contrast Normalization** | It originated from the integration of MA-plots and logged Bland-Altman plots, which assumes the presence of non-linear biases | affy | data-driven normalization |
| **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 | affy | data-driven normalization |
| **Linear baseline normalization** | Maps each sample spectrum to the baseline based on the assumption of a constant linear relationship21. However, this assumption of a linear correlation among sample spectra may be oversimplified | affy | data-driven normalization |
| **Li-Wong normalization (non-linear baseline normalization)** | aiming at removing unwanted sample-to-sample variations. This method is first used to analyze oligonucleotide arrays based on a multiplicative parametrization, and currently adopted to improve NMR-based metabolomics analysis | affy | data-driven normalization |
| **Cubic Splines normalization** | non-linear baseline methods assuming the existence of non-linear relationships between baseline and individual spectra | affy | data-driven normalization |
| **Cyclic Locally Weighted Regression (Cyclic Loess)** | comes also from the combination of MA-plot and logged Bland-Altman plot by assuming the existence of non-linear bias21. However, cyclic loess is the most time-consuming one among those studied normalization methods, and the | affy | data-driven normalization |

| | | | |
|---|---|---|---|
| | amount of time grows exponentially as the number of sample increases | | |
| **eigenMS** | A singular value decomposition-based method originally designed for LC-MS metabolomics dataset. | eigenMS | QC-based normalization |
| **Best-Matched Internal Standard Normalization (B-MIS)** | Normalizes peak areas using a batch-specific normalization process, which matches measured metabolites with isotope-labeled internal standards that behave similarly during the analysis. | B-MIS-normalization[1] | IS-based normalization |
| **Normalization method for metabolomics data using optimal selection of multiple internal standards (NOMIS)** | Linearly combine the intensity of multiple internal standards to optimize the normalization factor for each individual molecular species. | metabolomics | IS-based normalization |
| **Batch-wise LOESS normalization** | Fit LOESS curve for each batch using QC samples to calibrate the systematic variation | stats | QC-based normalization |
| **Batch-ratio normalization** | Normalize compound by the median average intensify of QC samples for each batch | bapred | QC-based normalization |
| **SVM** | To reduce unwanted variations and integrate multiple batches in large-scale metabolomics studies | StatTarget | QC-based normalization |

[1] GitHub repository: https://github.com/IngallsLabUW/B-MIS-normalization