hRUV: Hierarchical approach to removal of unwanted variation for large-scale metabolomics data

Taiyun Kim
The University of Sydney https://orcid.org/0000-0002-5028-836X
Owen Tang
University of Sydney
Stephen Vernon
Kolling Institute for Medical Research
Katharine Kott
Kolling Institute for Medical Research
Yen Chin Koay
Heart Research Institute
John Park
The University of Sydney
David James
The University of Sydney https://orcid.org/0000-0001-5946-5257
Terence Speed
Walter and Eliza Hall Institute of Medical Research https://orcid.org/0000-0002-5403-7998
Pengyi Yang
University of Sydney https://orcid.org/0000-0003-1098-3138
John O´Sullivan
Heart Research Institute, Newtown, New South Wales
Gemma Figtree
University of Sydney
Jean Yang (jean.yang@sydney.edu.au)
University of Sydney https://orcid.org/0000-0002-5271-2603

Article

Keywords: adaptive data normalization, dynamic batch correction, metabolomics, cardiovascular, removing unwanted variation

Posted Date: February 4th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-156243/v1

Version of Record: A version of this preprint was published at Nature Communications on August 17th, 2021. See the published version at https://doi.org/10.1038/s41467-021-25210-5.

hRUV: Hierarchical approach to removal of unwanted variation for large-scale metabolomics data

Taiyun Kim[1,2,9,^], Owen Tang[1,3,4,5,^], Stephen T Vernon[1,3,4,5], Katharine A Kott[1,3,4,5], Yen Chin Koay[1,5], John Park[1,3,4,5], David James[1,6], Terence P Speed[7,8], Pengyi Yang[1,2,5,9,+], John F. O'Sullivan[1,5,+], Gemma A Figtree[1,3,4,5,+] , Jean Yee Hwa Yang[1,2,+,*]

[1] Charles Perkins Centre, The University of Sydney, Sydney, NSW, Australia
[2] School of Mathematics and Statistics, The University of Sydney, Sydney, NSW, Australia
[3] Department of Cardiology, Royal North Shore Hospital, Sydney, NSW, Australia
[4] Cardiovascular Discovery Group, Kolling Institute of Medical Research, The University of Sydney, Sydney, NSW, Australia
[5] Faculty of Medicine and Health, The University of Sydney, Sydney, NSW, Australia
[6] School of Life and Environmental Sciences, The University of Sydney, Sydney, NSW, Australia
[7] Bioinformatics Division, Walter Eliza Hall Institute, Parkville, VIC, Australia
[8] School of Mathematics and Statistics, University of Melbourne, Parkville, VIC, Australia
[9] Computational Systems Biology Group, Children's Medical Research Institute, Westmead, NSW, Australia
[+],[^] Equal contribution
[*] Corresponding author (jean.yang@sydney.edu.au)

Keywords: adaptive data normalization / dynamic batch correction / metabolomics / cardiovascular / removing unwanted variation

Abstract

Liquid chromatography-mass spectrometry-based metabolomics studies are increasingly applied to large population cohorts, which run for several weeks or even years in data acquisition. This inevitably introduces unwanted intra- and inter-batch variations over time that can overshadow true biological signals and thus hinder potential biological discoveries. To date, normalisation approaches have struggled to mitigate the variability introduced by technical factors whilst preserving biological variance, especially for protracted acquisitions. Here, we propose a study design framework with an arrangement for embedding biological sample replicates to quantify variance within and between batches and a novel workflow that uses these replicates to remove unwanted variation in a hierarchical (hRUV) manner. We use this design to produce a dataset of more than 1,000 human plasma samples run over an extended period of time. We demonstrate significant improvement of hRUV over existing methods in preserving biological signals whilst removing unwanted variation for large scale metabolomics studies. Our novel tools not only provide a strategy for large scale data normalization, but also provides guidance on the design strategy for large omics studies.

Introduction

Liquid chromatography-tandem mass spectrometry (LC-MS/MS) is a preferred method of metabolomic acquisition given its high sensitivity and dynamic range. Typically, a range of metabolites can be separated on a single high performance LC column and their relative abundance quantified in MS/MS. This enables capture of fingerprints of specific biological processes that are critical in precision medicine applications such as studying complex metabolic diseases, and discovering new therapeutic targets and biomarkers[1]. There are a number of large-scale cohort studies that have performed metabolomic analyses, such as the Consortium of Metabolomics Studies (COMETS)[2], and the Framingham Heart Study (FHS)[3].

Despite a rapid increase in the number of large-scale metabolomics studies, the normalisation of metabolomics data remains a key challenge[4]. Due to the data acquisition time of studies with large sample size, prolonged study recruitment and potentially multiple samples at various time points for each participant, the data acquisition process may require the samples be divided into multiple batches, and may span anywhere from months to years[4,5]. Signals often drift over extended periods due to multiple factors including buffer changes, pooled quality control (QC) sample solutions, instrument cleanliness, and machine scheduled maintenance[6]. Common intra-batch variations include changes in LC-MS/MS performance due to instrument-dependent factors such as component failure or inconsistency, and fouling of the column, LC or MS source. Common inter-batch variations include time-dependent instrument variations such as instrument cleaning, tuning, column change, or inconsistent sample preparation factors including change in equipment and operator. These technical factors have substantial impact in downstream analytics and need to be appropriately accounted for to maximise the opportunity to identify true biological signals.

Several workflows have been designed for analysing metabolomics data (e.g. MetaboAnalyst[7] and NormalyzerDE[8]). However, most of them adapt common normalisation methods developed for other omics platforms and do not account for signal drift across extended time or inter-batch variations which are distinct unwanted variations commonly observed in metabolomics studies. A limited number of metabolomics specific normalisations methods have been developed (Table 1). These include: Support Vector Regression (SVR)[5], Systematic Error Removal using Random Forest (SERRF)[9], and Removal of Unwanted Variation based approaches[10,11]. These approaches utilise pooled QC samples to control signal drift by fitting non-linear models or by estimating inter-batch variations. The common assumption of these approaches is that the pooled QC samples are identical over an extended period of time. Whilst, this is appropriate for a short period of time the assumption may not hold for large-scale metabolomics data over months or years and currently there are no existing methods to account for normalisation over an extended period.

In this paper, we present a novel experimental sample arrangement strategy to embed biological sample replicates throughout large scale experiments to facilitate the estimation of unwanted variation within and between batches with RUV-III[12], which we will refer to as RUV in this paper. We propose a novel hierarchical approach to removing unwanted variation by harnessing information from sample replicates embedded in the sequence of experimental runs/batches and applying signal drift correction with robust linear or non-linear smoothers. An in-house targeted metabolomics study was performed on a hospital-based cohort of patients with atherosclerosis (BioHEART- CT) was conducted based on the proposed sample arrangement strategy, and we utilise this to assess the normalisation on a number of criteria including retention of biological signal, low variability among replication, and reproducibility of results in comparison to other existing methods. The hRUV method is accessible as an R package and also as a shiny application at https://shiny.maths.usyd.edu.au/hRUV/.

Results

Replicate arrangement strategy in large scale metabolomics study. We developed a series of technical replications designed as a framework to enable effective data harmonisation in large cohorts studies or studies over extended periods of time. 1,000 samples were manually divided into 88 sample batches including 80 individual samples and 8 pooled samples per batch. Our design includes three types of replicates within each 88 sample batches (in a 8×11 array format), these are the (i) classical pooled QC samples, (ii) single sample replicates in each row of a batch from a random selection of non-replicated samples in previous rows which we call 'short replicates', and (iii) five randomly selected non-short replicated samples from each batch are replicated to the next batch, which we call 'batch replicates'. Fig. 1a and b illustrates a schematic layout of the sample replicates design.

The classical pooled QC consists of a mixture of 10 μL of each of the 1,002 samples, pooled together into a single tube. The pooled QC sample was aliquoted and frozen, and a fresh aliquot was thawed for each batch to minimize the impact of repeated freeze thaw cycles. The spacing of the technical short replicates approximately 10 samples apart capture variation within a short time (approximately 5 hours, based on 30 minutes of run time per sample). This is a good measure of the variation of the metabolomics experiment. In contrast to pooled QC samples, where one sample is repeated many times, short technical replicates are duplicates of different samples; this increase in heterogeneity of samples for the estimation of unwanted variation is more robust compared to estimation with pooled QC. Finally, the batch replicates measure the variation that occurs across different batches. These replicates are typically 60-70 samples apart, capturing variation over a longer time period of 48-72 hours.

This design was used to generate data for a large metabolomics study consisting of 1,002 individuals from the BioHEART-CT biobank and quantification of 100 metabolites per individual. After preprocessing, 53 metabolites were detected at adequate levels in plasma to be included in the analyses. The exact sample designs are given in Supplementary Table 1. In total we had 164 replicates

from one pooled QC sample, 230 duplicates from samples across 15 batches, and 140 'batch' duplicates from 70 samples. As expected, variation between replicates within a batch tends to be smaller compared to replicates between batches, as demonstrated in Supplementary Fig 1. A shiny application was developed to enable easy generation of the replicate design upon input of the desired batch size and the desired number of inter-batch replicates. The experimental design with appropriate numbers and assignment of replicates inserted is then exported as a Comma Separated Values (CSV) file. The extra replicate tubes were prepared during the aliquoting and inserted into the appropriate positions during sample processing.

In this current study, we used 80 individual samples and 8 pooled samples per batch with consideration of blanks needed to be included on the auto sampler sample tray (100 in capacity) as well as to minimise the overall number of pooled sample used. In practice, one could select any number of samples, subjects to the tray capacity of the auto sampler as the unit for a batch. The notions of short and long (batch) replicates can be applied to any batch size to assess variation over a variety of distances.

A novel hierarchical method to remove unwanted variation (hRUV) in large scale omics experiments. To enable effective data harmonisation across large cohort studies or across an extended period of time, we propose a novel hierarchical approach to correct for the unwanted variation between smaller subsets of batches individually, and to sequentially expand to the next set of batches. The two key components of the hRUV can be summarised as follows: (i) signal drift correction within batches with a robust smoother that captures the irregular patterns affecting each metabolite; and (ii) a scalable hierarchical approach to removal of unwanted variation between batches with the use of carefully assigned sample replicates.

The signal drift within each batch was corrected using a robust smoother that captures the trends visible by run order. We explored linear (robust linear model) and non-linear (local regression) model fitting smoothers to capture and remove the run order effects in the data. This is because, due to their chemical and physical properties (Supplementary Fig 2a), each metabolite is affected differently across runs within each batch. These unique changes in signal for each metabolite need to be treated separately.

The concept of sequential batch correction is introduced here to enable scalability for large scale cohort studies. This is a clear contrast to the conventional data integration for normalisation that involves estimating unwanted variation across all batches as whole (Fig. 1d). Supplementary Fig 2b shows the inter-batch variation and the differences in the corresponding adjustment factors over time, highlighting the need for dynamic normalisation. To this end, we propose two tree structured approaches to estimating the different forms of unwanted variation across a large-scale cohort study, and to dynamically modify the batch effect removal across time. Fig. 1d illustrates the two approaches: balanced tree and the concatenating tree. The balanced tree approach requires $\log_2(n)$ RUV adjustments to deal with n batches, while the concatenating tree approach requires n-1 RUV adjustments. The concatenating approach requires more computation than the balanced tree approach but has an advantage when future integration with new batches is necessary. For once the initial batches are normalised, the additional RUV adjustments are needed are only as many as the number of new batches. While the balanced tree approach is quicker for large n, if m new batches are introduced in the future, the data will require additional $\log_2(n+m)$ RUV adjustments from individual batch level.

The details of hRUV are included in the Methods section. The final output of hRUV is a single normalised and batch-corrected matrix with all input matrices merged and ready for downstream analysis.

Implementation of a smoother and RUV with sample replicates enables effective adjustment of within batch variation. We assess the performance of signal drift correction by comparing the results of smoothers against themselves and against the commonly used approaches (see Methods). Here, we applied both linear and non-linear smoothers to two sets of sample types; pooled QC samples, or all biological samples within a batch. In general, all four adjustment approaches (loess, rlm, loessSample, rlmSample) give adjusted values that have effectively removed any signal drift associated with experimental run order (Supplementary Fig. 3d-f). The intra-batch correction with all biological samples performs comparably to adjustments performed with pooled QC samples (Supplementary Fig. 3). This suggests a possible reduction in pooled QC samples during experimental design, and thus reducing the total run cost.

In addition to using robust smoothers, the use of RUV with short replicates within each batch after a robust smoother further reduced the sample variations as demonstrated in Supplementary Fig. 2c. Thus using a robust smoother and RUV with short replicates provides effective removal of various unwanted intra-batch variations (Fig. 2) and highlights the value of intra-batch sample replicates.

hRUV is more effective in removing unwanted variation compared to other existing methods. Across an extended period of time, there are many different types of unwanted variation. Figure 3a shows that across 1,000 samples, we observe constant or irregular signal drift or abrupt jumps in signals. The run plots (Fig. 3a) illustrate the removal of technical variations introduced between batches and from run time effects for the metabolite glutamate.

Next, in comparison, we note that the hierarchical RUV normalisation was better at removing unwanted between-batch variation than the original single-step RUV. We compared the standard deviation (SD) between all sets of replicates, with lower values indicating better performance as the replicates should theoretically be identical. Figure 3b highlights lower SD between hierarchical normalisation methods (coloured in orange and red) and single-step ones (coloured in blue and green), suggesting that the hierarchical approach is more effective in batch-correction across extended periods of time. Additionally, hierarchical approaches following intra-batch RUV (coloured red) showed even lower sample replicate variation.

hRUV retains biological signals and outperforms existing normalisation methods. To examine the extent to which our method removes only unwanted variation and retains known biological signals, we performed supervised machine learning to illustrate our ability to identify known biological signals for disease prediction. Here we have chosen hypertension as a response variable and performed supervised machine learning classification to detect hypertension status from metabolomics abundance. We anticipate that a normalisation method that retains biological signals has a higher classification accuracy. The differential expression (DE) analysis to identify corresponding biomarkers (DE metabolites) measures the interpretability of the signal.

We observed that the average accuracy of hierarchical based normalisation methods were generally higher compared to one-step methods (Fig 3c). The 'loessAllShort_batch_Hc' method showed the best performance in prediction accuracy. This approach first adjusted for signal drift by fitting a loess line through all the samples and RUV corrected with short technical replicates within a batch, followed by applying RUV in a hierarchical fashion using batch technical replicates.

Proline, valine, cyclic adenosine monophosphate and dimethylguanidino valeric acid are metabolites known to be associated with individual's hypertension[13–15]. We observed all these metabolites to be significantly differentially expressed in association with the hypertension status of over 1,000 patients only in hRUV normalised data. Other methods were only capable of identifying subsets of these metabolites.

In general, we found that hRUV performed favorably in terms of maintaining strong biological signal and reducing unwanted variation such as signal drift and batch specific noise in this large study (Fig. 4). Our evaluation metrics capture the trade-off between these two broad objectives. hRUV manages the trade-off between removing unwanted variation and retaining known clinical features of interest. We observed that the ratio, SVR, SERRF and RLSC methods have removed batch effects and reduced sample and pooled QC replicate variance, but as a trade-off these methods result in a loss of biological signals, as evident by the low AUC and prediction accuracy values. Visualizing all these quantities on a heatmap, we find that hRUV methods are ranked in the top 5-10 in most of the evaluation criteria (Fig. 4). The hRUV normalised data show the least variation across the different types of replicate samples and correctly removed batch driven technical noise, whilst maintaining a strong biological signal (Fig. 3b and Supplementary Fig 5a-b).

hRUV is robust to the key decisions and the types of hierarchical structure and choice of negative controls. We have investigated a number of parameters under the hRUV framework, including the various kinds of technical replication, types of negative control metabolite and types of hierarchical structure. We have explored different combinations of replication including pooled QC, inter-batch, and intra-batch replicates. We found that corrections with only pooled QC sample replicates over-estimates the unwanted variation and thus removes biological signals from the data, see Supplementary Fig 6. This highlights the value of using sample replicates as opposed to the pooled QC samples in the estimation of intra- and inter-batch unwanted variation.

In contrast, the two hierarchical approaches in our hRUV show only a small differences in many of the evaluation measures. Both the balanced trees and the concatenating approaches performed adjustments between two sets of batches with 5 inter-batch replicates at each layer. As summarised in Fig 4, the normalisation results are very similar between the two types of tree structure.

We explored several approaches to obtaining negative controls for RUV, including a data driven iterative procedure to select stable metabolites and the selection of all metabolites, and saw little difference between the normalised data outputs. We saw that the use of sample replicates demonstrated the greatest impact on the final normalised data (Supplementary Fig. 7).

Discussion

In this manuscript, we introduce a design strategy and present hRUV, a novel hierarchical approach to remove unwanted variations and batch correct for large scale metabolomics data, where there is a substantial unmet need. We illustrated the performance of this approach using metabolomics data derived from over 1,000 patients in the BioHEART-CT biobank that was run over 15 batches across 44 days.

The careful arrangement of sample replicates on each batch is an important design consideration for large-scale mass-spectrometry studies. Here we believe that systematic arrangements perform better with hRUV normalisation than fully random ones. In our current design, the samples to be replicated were selected randomly from the previous batch and the corresponding repeats (batch replicates) were placed at the start of the current batch. Ideally we'd expect to select samples with this strategy whose positions were evenly distributed across the batch, but it is possible by chance to select samples whose positions are from only the first half or only the second half of the previous batch (Supplementary Table 1). This unintended clustering of selected samples was observed between batches 6 and 7, and also between batches 13 and 14, where replicates are selected only from the second half of the previous batch. This limits our ability to capture the unwanted variability across the whole batch, and as a result, we saw a slight shift in signal between these two batches for selected metabolites (Supplementary Fig 5c).

While the proposed hRUV algorithm expects data without missing values, this is often not possible in large-scale metabolomics data due to the nature of the mass spectrometry technology pragmatic issues related to real world clinical studies. To this end, we include an option for users in which the missing values are first imputed prior to applying hRUV and the missing values can be replaced back after hRUV integration. This allows many more sparse metabolites to be incorporated for downstream analyses which is an important aspect in large-scale metabolomics studies and may improve our chance of identifying novel metabolites from the data.

The negative controls are used in RUV to estimate the unwanted variation. The challenge with metabolomics is that the signals of the metabolites are dependent on their individual chemical properties[4,11] and thus the selection of appropriate negative controls to correct for batch effect is a challenge. Whilst hRUV function accepts a user-defined set of negative controls, in our exploration of data driven negative control metabolites compared to all metabolites as negative control, we have

found no significant differences between the two approaches in removal of unwanted variation and utility of inter-batch sample replicates were more effective for batch correction (Supplementary Fig. 7).

In summary, hRUV uses sample replicates to integrate data from many batches in large-scale metabolomics studies. We show the value of suitably located sample replicates for estimating unwanted variation and guiding the design of future studies. While several other existing methods exist to correct large numbers of batches for intra-batch signal drift and inter-batch unwanted variation, hRUV performs consistently better than them in retaining biological variation whilst at the same time removing unwanted variation within and across batches.

Methods

Clinical samples. The samples used were from the BioHEART-CT discovery cohort, which has been described in detail previously[16]. The study was approved by the Northern Sydney Local Health District Human Research Ethics Committee (HREC/17/HAWKE/343) and all participants provided informed, written consent. Briefly, patients undergoing clinically indicated CT coronary angiogram for suspected coronary artery disease were recruited from multiple sites in Sydney, Australia. Blood samples were taken at time of recruitment, and after appropriate processing, plasma samples including replicates were aliquoted and stored at -80°C until analysis.

Metabolites extraction and analysis were performed as previously described[17,18]. In brief, 10µl plasma was mixed with 90µl HILIC sample buffer, an acetonitrile: methanol: formic acid mix (75:25:0.2, v:v:v) to precipitate plasma proteins. The resulting mixture was vortexed and spun at 14,000 rpm for 20 minutes to separate debris. The metabolite containing supernatant was then transferred to a glass HPLC sample vial and resolved on an Agilent 1260 Infinity HPLC System, and m/z was determined

306 by Qtrap5500 (Sciex)[17,18]. Each sample was eluted over a 25 minute period, and each batch of
307 samples took 40 hours to complete. A total of 15 batches were completed over 44 days.
308
309 **Technical replicate design.** For each batch, three pooled QC samples were run to fine adjust the
310 acquisition setting of the HPLC/MASS-Spec system, then a single pooled QC sample was repeated
311 after every 10 runs. For each row of the 11 samples (10 individual samples + 1 pooled QC sample)
312 from the 12th sample onwards (second row), we randomly selected a single biological sample from
313 the previous row to be replicated at random position in a current row (short replicate). For each batch,
314 after the first three pooled QC samples, a random selection of 5 biological samples from the previous
315 batch was repeated (batch replicate) and short replicates are embedded at each row. All randomisation
316 was performed using the sample function in R. The replicate design is available as a function
317 `expDesign` in the *hRUV* package, and also in our shiny application
318 http://shiny.maths.usyd.edu.au/hRUV.
319

Pre-processing of metabolomics data. Targeted metabolomics based on scheduled multiple reaction monitoring optimised to the metabolite of interest using authentic standards was applied in this study. Metabolite abundance peaks were integrated using the area under the curve for calibrated peaks from MultiQuant™ version 3.0.3 (SCIEX), with manual peak integration was performed when necessary. This ensures the consistency of all the peaks integrated. The signal intensity of the ions were $\log_2$ transformed and metabolites that were not present in at least 50% of the samples were filtered out. Missing values were imputed using k-nearest neighbour with default parameters implemented in *DMwR2*[19] package in R. We examined the three consecutive QC samples embedded at the start of each batch and removed any outlying measurements.

Hierarchical approach to removal of unwanted variation (hRUV). The hRUV algorithm was designed for experiments with a large number of batches, and consists of two key components, including (i) within batch signal drift adjustment with robust smoothers and RUV; and (ii) the adjustment of the datasets with unwanted variation using an RUV in a hierarchical approach. The main inputs to hRUV consist of a list of raw signal matrix, with rows corresponding to metabolites and columns to samples as `SummarizedExperiment`[20] objects in R, a specific intra- and inter-batch normalisation method, structure of the tree and the parameters for RUV. hRUV performs repeated RUV procedures to sequentially adjust the data over a large collection of batches, with the number of unwanted variation factors (k) defaulting to 5.

Part I: Signal drift adjustment within a batch

In the present setting, batch refers to one 88 sample run. However, this can be any pre-specified number of samples.

(i) Standard adjustment (ratio) - The signal ratio was calculated by dividing the sample signal to its nearest pooled QC sample run. Let us denote P_L as the early run pooled QC sample at run L and

P_{L+M} as the next pooled QC sample in a batch at run L+M, where M denotes the number of run gaps between P_L and P_{L+M}. Then the signal ratio is defined as follows:

$$\text{If } L < l \leq L + \frac{M}{2},$$

$$\widehat{s_{il}} = \frac{s_{il}}{P_L}$$

$$\text{else if } L + \frac{M}{2} < l \leq L + M,$$

$$\widehat{s_{il}} = \frac{s_{il}}{P_{L+M}}$$

where s_{il} denotes a signal of a sample with metabolite i at run number l.

(ii) Loess line - A loess line was fitted to all the biological samples for each metabolite within a batch with the default span parameter of 0.75. The differences between the fitted line to the median of each metabolite across all samples per batch were calculated for adjustment of each samples as follows:

$$\widehat{y_{ij}} = y_{ij} + (\tilde{y}_i - \widehat{y_{ij}^*})$$

where y_{ij} represents a log_2 transformed signal for sample j in metabolite i in a batch and $\widehat{y_{ij}^*}$ denote a loess fitted value of y_{ij} and $\tilde{y}_i$ denote the median of y_{ij} for all j. Here, the loess line uses the `loess` function in the *stats*[21] package.

(iii) Linear line - Robust linear model (rlm) was fitted on a log_2 transformed signal against the run index using the `rlm` function from the *MASS*[22] package to all the biological samples with a maximum iteration set to 100. The adjustments of each sample were calculated as with to the loess approach where $\widehat{y_i^*}$ would denote the predicted value of y_i.

(iv) Loess line fitting with pooled QC samples - Similar to (ii), we fit the loess line to the pooled QC samples only. The adjustment values of each sample were calculated using the predicted values from the model ($\widehat{y_{pij}^{*}}$).

$$\widehat{y_{ij}} = y_{ij} + (\tilde{y}_i - \widehat{y_{pij}^{*}})$$

(vii) Linear line fitting with pooled QC samples - Likewise, we fit the rlm to the pooled QC samples. The adjustment values of each sample were calculated using the predicted values from the model.

(viii) RUV based approaches - We incorporated sample replicates into our design matrix of RUV introduced by Molania et al[12]. These sample replicates are utilised to estimate the unwanted variation as the signals of these replicate samples should theoretically be identical. All metabolites were selected as the negative controls for RUV and the number of unwanted factors to use (*k*) was taken as 5.

Part II: Hierarchical batch integration design

(i) Balanced tree. The balanced tree approach to normalisation is to perform removal of unwanted variation measured between pairs of different batch groups at different levels of the tree. In this approach, we began by removing unwanted variation between pairs of neighbouring batches. In the next layer of adjustment, we paired the two neighbouring groups of integrated batches (sets of 2 batches) and repeated the process to expand the number of batches per set until the last layer, where we had a single group of all the batches which was normalised, as illustrated in schematics in Figure 1d. For a study with *n* batches, this will requires $log_2(n)$ RUV adjustments.

(ii) Concatenation. As with the balanced tree approach, the concatenating approach performs removal of unwanted variation measured between pairs of batches, but in a sequential progression. We began with the first two batches for batch correction and sequentially introduced the next batch to remove unwanted variation as illustrated in schematics in Figure 1d. For a study with *n* batches, this will require an *n*-1 number of RUV adjustments.

For both balanced tree and concatenation methods, we apply RUV at each layer as follows:

Let us denote by *B* the pair of batches of interest, *M* as the number of metabolites, and S as the number of samples in batches B. The mean adjusted sample Z_{mbc} can be calculated as:

$$Z_{mbs} = Y_{mbs} - Y_{m..},$$

where $Y_{m..}$ is the average expression of metabolite *m* across samples *S* and batches *B* calculated by:

$$Y_{m..} = \frac{1}{S}\sum_{bs} Y_{mbs};$$

The mean adjusted data $Z_{S\times M}$ can be fitted to the model underlying the RUV model, which is formulated as:

$$Z_{S\times M} = X_{S\times p}\beta_{p\times M} + W_{S\times k}\alpha_{k\times M} + \epsilon_{S\times M}$$

where *X* is the matrix of factor of interest; *p* is the number of factors of interest; *W* is the unobserved design matrix corresponding to the unwanted factors; *k* is the linear dimension of the unwanted factors, which is unknown; ϵ denotes the random error. Thus the RUV normalised data can be represented as:

$$\hat{Z}_{S\times M} = Z_{S\times M} - \widehat{W}_{S\times k}\hat{\alpha}_{k\times M}$$

After the RUV, $Y_{m..}$ is returned back to the mean adjusted RUV normalised data as follows:

$$\hat{Z}^*_{mbs} = \hat{Z}_{mbs} + Y_{m..}$$

Data driven negative metabolite selection

We explored adaptive data driven selection of negative control metabolites in comparison to a selection of all metabolites in an RUV method. The adaptive selection was performed by ranking non-differentially expressed metabolites by *p*-values per batch for the hypertension response variable. We utilised differential expression analysis with the *limma*[23] package (version 3.46) in R.

Performance evaluation / evaluation metric processing

We evaluate hRUV methods including 13 publicly available metabolomics data normalisation methods (Table 1). Details of the method abbreviations is explained in Table 2. These packages were installed either through the official CRAN or Bioconductor website where available, or from GitHub pages. For all 13 existing methods, we used the default settings and parameters as described in the package README or vignette for training each model.

Evaluation metrics and plots

(i) Skewness - The skewness of samples were calculated with `skewness` function from *e1071*[24] package in R.

Let us denote x_j for the non-missing elements of **x**, n for the number of samples, $\mu$for the sample mean, s for the sample standard deviation, and $m_r = \sum_j (x_j - \mu)^r / n$ for the sample moments of order r. The skewness then can be calculated as:

Skewness = m_3/s^3,

(ii) Normality metric - The normality tests were performed with Shapiro-Wilk normality test implemented in `shapiro.test` function from the *stats*[21] package in R.

(iii) Predictability with accuracy - To assess the predictability of a normalised dataset, we utilised a binary diagnosis of hypertension as the response variable. This was chosen as it had a reasonably balanced class distribution, as 39% of the cohort had hypertension. We used a Support Vector Machine (SVM) implemented in the *e1071*[24] package to predict the hypertension status of participants of the study. We measured the average accuracy via a 30-repeated 10 folds cross-validation strategy.

(iv) Signal strength with AUC - We use the same prediction model from (iii) and calculate the area under the ROC curve (AUC) values.

(v) Standard deviation of replicates (SD replicates) - To demonstrate the variation between the replicate samples after normalisation, for each set of replicated sample, we calculated the standard deviation for each metabolite and visualised the results as a boxplot. A low standard deviation indicates a small variability between the replicates and thus illustrates that the replicates are close to identical.

(vi) Clustering by batch (Reduction in batch effect) - To assess the removal of batch effects, we performed unsupervised hierarchical and *k*-means clustering (`hclust` and `kmeans` in *stats*[21] package in R respectively) where we set the number of clusters *k* to the number of batches. The cluster output is evaluated using adjusted rand index (ARI):

$$\mathrm{ARI} = \frac{2\,(ad - bc)}{(a+b)(b+d) + (a+c)(c+d)}$$

where a is the number of pairs of samples partitioned into the same batch group by the clustering method, b is the number of pairs of samples partitioned into the same cluster but does not belong to the different batch group, c is the number of pairs of samples partitioned into different clusters but belongs to the same batch group and d is the number of pairs of samples correctly partitioned into different clusters. A low ARI value indicates lower concordance with the batch information and thus demonstrates removal of batch effect in the data.

(vii) Differential expression (DE) analysis of hypertension - To assess the biological signal in the normalised data, we performed DE analysis with the R package *limma*[23]. We identified a set of metabolites with a 5% level of significance and verified their association with hypertension from the literature.

Diagnostic Plots

To graphically assess whether the normalisation method or the choice of parameters of hRUV has effectively corrected the batch effect, we have provided three kinds of diagnostic plots: (1) PCA plots; (2) relative log expression (RLE) plots[19]; (3) metabolite run plots.

1. PCA. PCA plots were generated using all metabolites. We show the first and second principal components.

2. Relative Log Expression (RLE) plot. RLE plots are a useful tool to visualize unwanted variation. RLA plots are boxplots of RLA for each sample, calculated as $Y_{ij} - \tilde{Y}_i$, where $\tilde{Y}_i$ = median$\{Y_{ij}: j = 1,2,\ldots\}$, and Y_{ij} is the log signal value of metabolite *i* in sample *j*. The samples from different batches should have a similar distribution, and the medians of the boxplots should be close to zero if the unwanted variations are removed.

497

498 ***3. Metabolite run plot.*** Metabolite plots are a useful diagnostic visualisation to visualise the signal
499 drifts. The run plots are a scatter plot of signals for each metabolite against the run order of all the
500 samples. The overall shape of the scatter plot should be a flat horizontal bar. All other shapes of trend
501 in the scatter plot is an indication of a signal drift.

502

503 **Author contributions** JYHY, GF and JO conceived and designed the study. JY designed the sample
504 replicate framework with input and discussion from TPS, JO, DJ and TK implemented the design.
505 The BioHEART study was designed by GF and the metabolomics assays were performed by OT and
506 JP supervised under YCK and JO. OT and JP processed and curated the data with input from JO, YK,
507 JY. GF, SV & KK provided the clinical and pathology input and guidance with BioHEART clinical
508 data. JY, PY and TK led the hRUV method development with input from TPS. JY and TO lead the
509 evaluation and data analysis with input from GF, JO, TP, PY, and OT. TK implemented the R
510 package with help from PY and JY. All authors wrote and reviewed the manuscript have approved
511 the final version of the manuscript.

512

513 **Acknowledgements** The authors thank all their colleagues, particularly at The University of Sydney,
514 School of Mathematics and Statistics, Charles Perkins Center, Kolling Institute and Royal North
515 Shore Hospital for their support and intellectual engagement. The following sources of funding for
516 each author, and for the manuscript preparation, are gratefully acknowledged: Australian Research
517 Council Discovery Project grant (DP170100654) to JYHY; Judith and David Coffey Lifelab
518 Scholarship to TK. KK is supported by an Australian Commonwealth Government Research Training
519 Program Stipend Scholarship; SV is supported by a University of Sydney Postgraduate Research
520 Scholarship funded by Heart Research Australia. GF is supported by a National Health and Medical
521 Research Council Practitioner Fellowship (grant number APP11359290), Heart Research Australia,

and the New South Wales Office of Health and Medical Research. PY is supported by a National Health and Medical Research Council Investigator Grant (APP1173469).

The funding source had no role in the study design; in the collection, analysis, and interpretation of data, in the writing of the manuscript, and in the decision to submit the manuscript for publication.

Declaration of competing interests

None

Data availability

The sample designed is made available on Supplementary Table 1. The processed mass spectrometry metabolomics data is available on the GitHub repository, https://github.com/SydneyBioX/BioHEART_metabolomics. All other data are available from the corresponding author on reasonable request.

Code availability

The hRUV implementation is available as an R package stored at the GitHub, https://github.com/SydneyBioX/hRUV and as a web shiny application at http://shiny.maths.usyd.edu.au/hRUV.

References

1. Clish, C. B. Metabolomics: an emerging but powerful tool for precision medicine. *Mol. Case Stud.* **1**, a000588 (2015).
2. Yu, B. *et al.* The Consortium of Metabolomics Studies (COMETS): Metabolomics in 47 Prospective Cohort Studies. *Am. J. Epidemiol.* **188**, 991–1012 (2019).
3. Tsao, C. W. & Vasan, R. S. Cohort Profile: The Framingham Heart Study (FHS): overview of milestones in cardiovascular epidemiology. *Int. J. Epidemiol.* **44**, 1800–1813 (2015).
4. Misra, B. B. Data normalization strategies in metabolomics: Current challenges, approaches, and tools. *Eur. J. Mass Spectrom.* **26**, 165–174 (2020).
5. Shen, X. *et al.* Normalization and integration of large-scale metabolomics data using support vector regression. *Metabolomics* **12**, 89 (2016).
6. Dunn, W. B., Wilson, I. D., Nicholls, A. W. & Broadhurst, D. The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **4**, 2249–2264 (2012).
7. Xia, J. & Wishart, D. S. Web-based inference of biological patterns, functions and pathways from metabolomic data using MetaboAnalyst. *Nat. Protoc.* **6**, 743–760 (2011).
8. Willforss, J., Chawade, A. & Levander, F. NormalyzerDE: Online Tool for Improved Normalization of Omics Expression Data and High-Sensitivity Differential Expression Analysis. *J. Proteome Res.* **18**, 732–740 (2019).
9. Fan, S. *et al.* Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal. Chem.* **91**, 3590–3596 (2019).
10. De Livera, A. M. *et al.* Normalizing and Integrating Metabolomics Data. *Anal. Chem.* **84**, 10768–10776 (2012).

11. Livera, A. M. D. *et al.* Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **87**, 3606–3615 (2015).
12. Molania, R., Gagnon-Bartsch, J. A., Dobrovic, A. & Speed, T. P. A new normalization for Nanostring nCounter gene expression data. *Nucleic Acids Res.* **47**, 6073–6083 (2019).
13. Flores-Guerrero Jose L. *et al.* Concentration of Branched-Chain Amino Acids Is a Strong Risk Marker for Incident Hypertension. *Hypertension* **74**, 1428–1435 (2019).
14. Teymoori, F. *et al.* Various proline food sources and blood pressure: substitution analysis. *Int. J. Food Sci. Nutr.* **71**, 332–340 (2020).
15. Siffert, W. G Protein Polymorphisms in Hypertension, Atherosclerosis, and Diabetes. *Annu. Rev. Med.* **56**, 17–28 (2004).
16. Kott, K. A. *et al.* Biobanking for discovery of novel cardiovascular biomarkers using imaging-quantified disease burden: protocol for the longitudinal, prospective, BioHEART-CT cohort study. *BMJ Open* **9**, e028649 (2019).
17. Koay, Y. C. *et al.* Ingestion of resistant starch by mice markedly increases microbiome-derived metabolites. *FASEB J.* **33**, 8033–8042 (2019).
18. Koay, Y. C. *et al.* Effect of chronic exercise in healthy young male adults: a metabolomic analysis. *Cardiovasc. Res.* (2020) doi:10.1093/cvr/cvaa051.
19. Torgo, L. *Data Mining with R, learning with case studies, 2nd edition*. (Chapman and Hall/CRC, 2016).
20. Morgan, M., Obenchain, V., Hester, J. & Pagès, H. *SummarizedExperiment: SummarizedExperiment container*. (2020).
21. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, 2020).
22. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S*. (Springer, 2002).
23. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47–e47 (2015).

24. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. & Leisch, F. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. (2020).
25. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **6**, 1060–1083 (2011).
26. De Livera, A. M., Olshansky, G., Simpson, J. A. & Creek, D. J. NormalizeMets: assessing, selecting and implementing statistical methods for normalizing metabolomics data. *Metabolomics* **14**, 54 (2018).
27. Chawade, A., Alexandersson, E. & Levander, F. Normalyzer: A Tool for Rapid Evaluation of Normalization Methods for Omics Data Sets. *J. Proteome Res.* **13**, 3114–3120 (2014).
28. Wei, X. *et al.* MetPP: a computational platform for comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry-based metabolomics. *Bioinformatics* **29**, 1786–1792 (2013).
29. Veselkov, K. A. *et al.* Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Anal. Chem.* **83**, 5864–5872 (2011).
30. Lee, J. *et al.* Quantile normalization approach for liquid chromatography-mass spectrometry-based metabolomic data from healthy human volunteers. *Anal. Sci. Int. J. Jpn. Soc. Anal. Chem.* **28**, 801–805 (2012).
31. t'Kindt, R., Morreel, K., Deforce, D., Boerjan, W. & Van Bocxlaer, J. Joint GC–MS and LC–MS platforms for comprehensive plant metabolomics: Repeatability and sample pre-treatment. *J. Chromatogr. B* **877**, 3572–3580 (2009).
32. Wang, W. *et al.* Quantification of Proteins and Metabolites by Mass Spectrometry without Isotopic Labeling or Spiked Standards. *Anal. Chem.* **75**, 4818–4826 (2003).
33. Roberts, L. D., Souza, A. L., Gerszten, R. E. & Clish, C. B. Targeted Metabolomics. *Curr. Protoc. Mol. Biol.* **98**, 30.2.1-30.2.24 (2012).

Figures

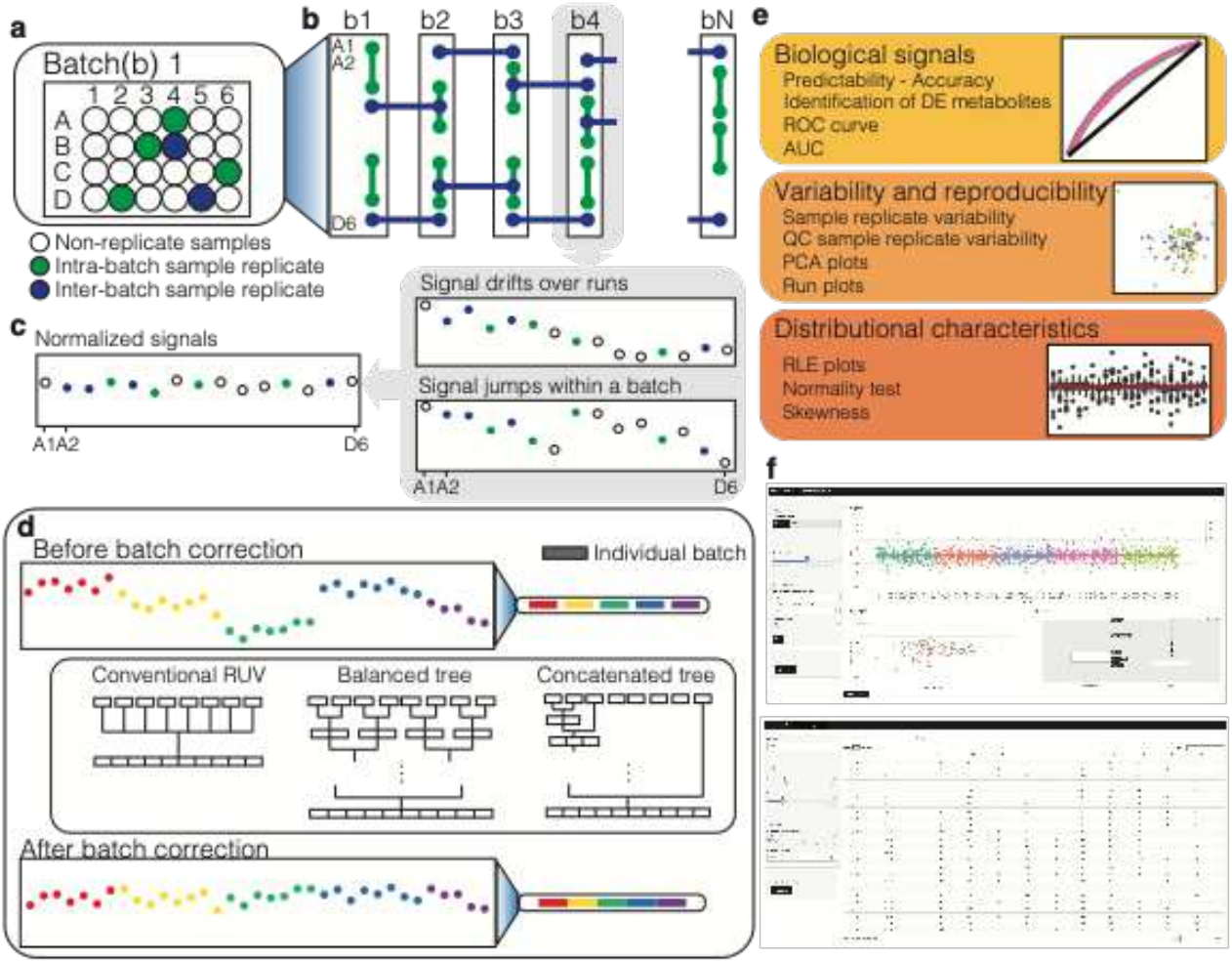


Fig. 1 Schematic overview of the hRUV pipeline. a A schematic illustration of the plasma sample arrangement for the experimental batch in an array format where intra-batch sample replicates (green circles) and inter-batch sample replicates (blue circles) are embedded. **b** A schematic illustration of overall sample replicate design and arrangements. **c** Continuing the colour scheme from **b**, two illustrative run plots requiring intra-batch correction, with signal drift and other variations illustrated in the grey boxes. **d** A demonstration of signal variation before and after inter-batch correction in hRUV. A common approach to inter-batch correction is illustrated as conventional RUV and the proposed hierarchical approaches are illustrated. **e** A list of evaluation criteria to assess hRUV performances grouped into categories of biological signals, variability and reproducibility and distributional characteristics. **f** A screenshot of the user-friendly shiny application.

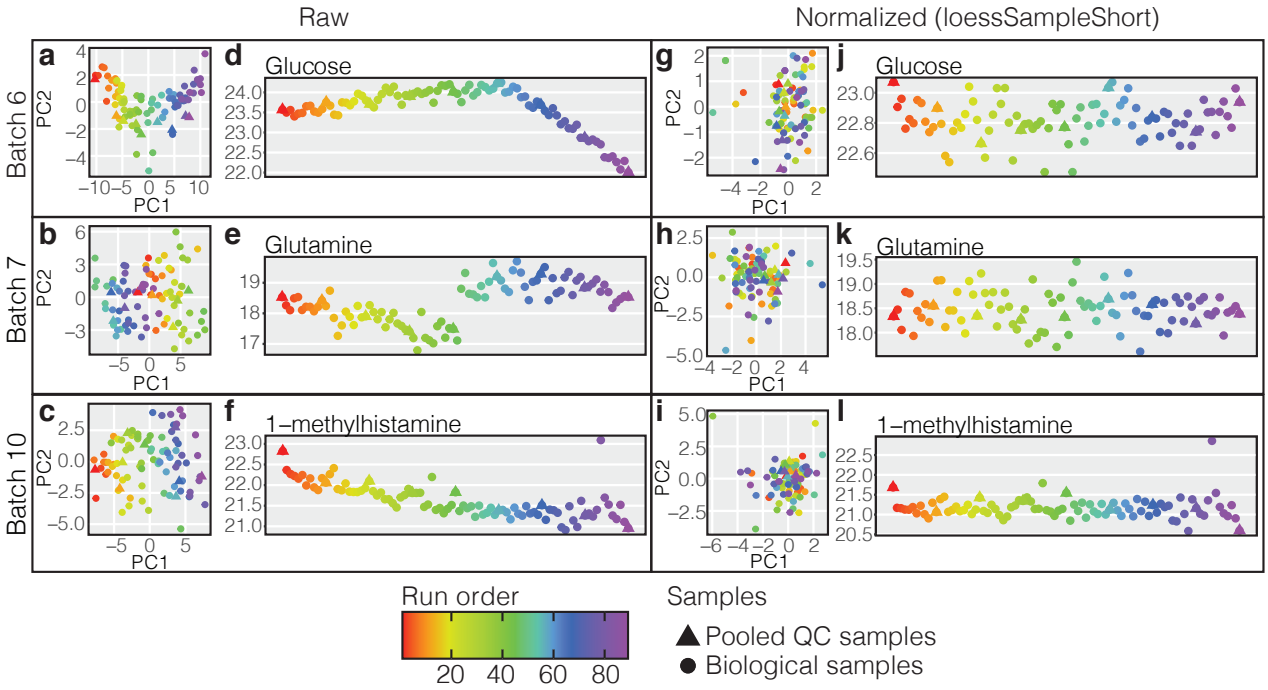
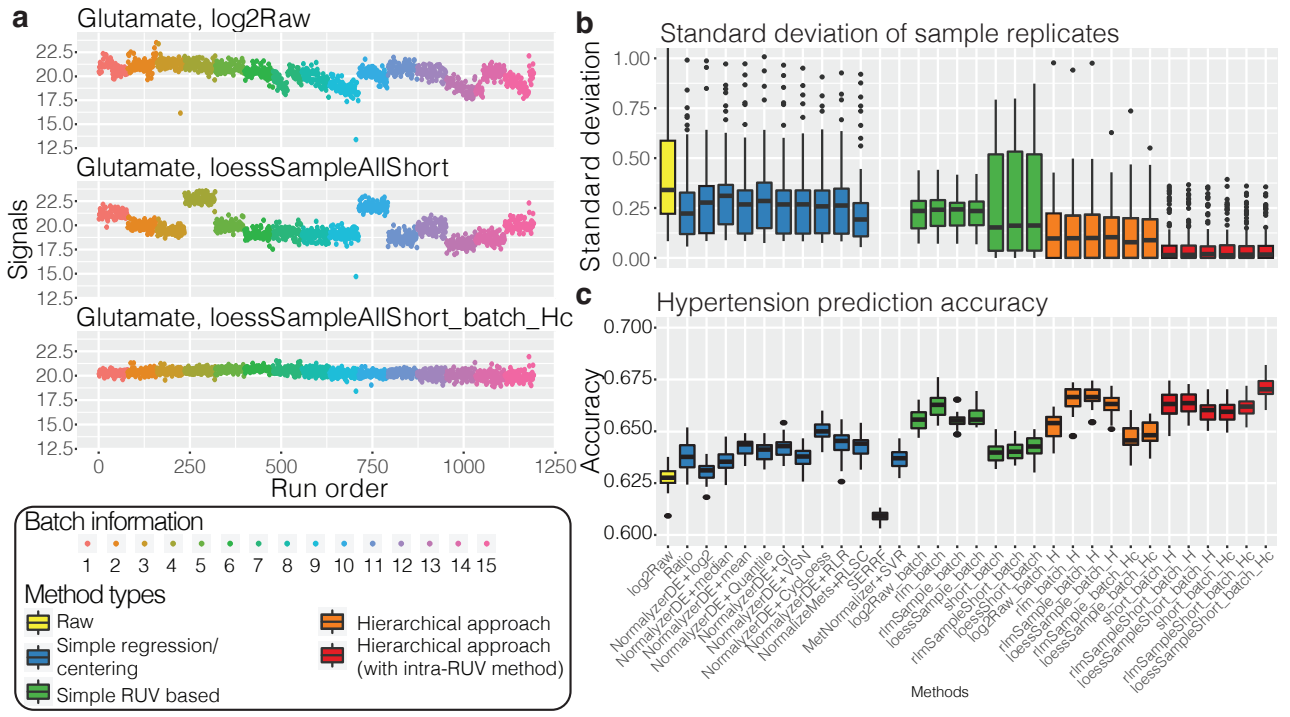


Fig. 2 Examples of typical technical variations and signal drifts within each batch for different metabolites and comparison to normalised data. (Note vertical scale change.) **a-c** PCA plots of raw data in batches 6, 7 and 10 respectively, each marker is coloured by the sample run order. **d-f** Run plots exhibiting metabolite signal drift for glucose, glutamine, and 1-methylhistamine respectively, for their respective batches. **g-i** PCA plots from the same batches as **a-c** respectively with intra-batch normalised data using loessSampleShort method. **j-l** Run plot for the metabolites and batches illustrated in **d-f** but with intra-batch data normalised using loessSampleShort method.



643
644 **Fig. 3 Key assessments of hRUV performances. a** A run plot of raw, intra-batch corrected and final
645 hRUV normalised data in all 15 batches of the BioHEART-CT cohort. The x-axis indicates the run
646 order, the y-axis indicates the signal of glutamate, and samples are coloured by the batch numbers. **b**
647 Boxplots of all sample replicate standard deviations, where lower values indicate better performance.
648 The boxes are coloured by the approach taken to normalize the data. The y-axis of the plot is restricted
649 to a range between 0 to 1 to highlight the differences between the majority of the methods. SERRF
650 and MetNormalizer+SVR's median sample replicate SD was greater than 1 and thus is not shown. **c**
651 Boxplots of hypertension prediction accuracies for all methods. Higher prediction accuracy indicates
652 better performance. The colouring of the boxes are consistent with that in **b**.

653
654 **Fig. 4**

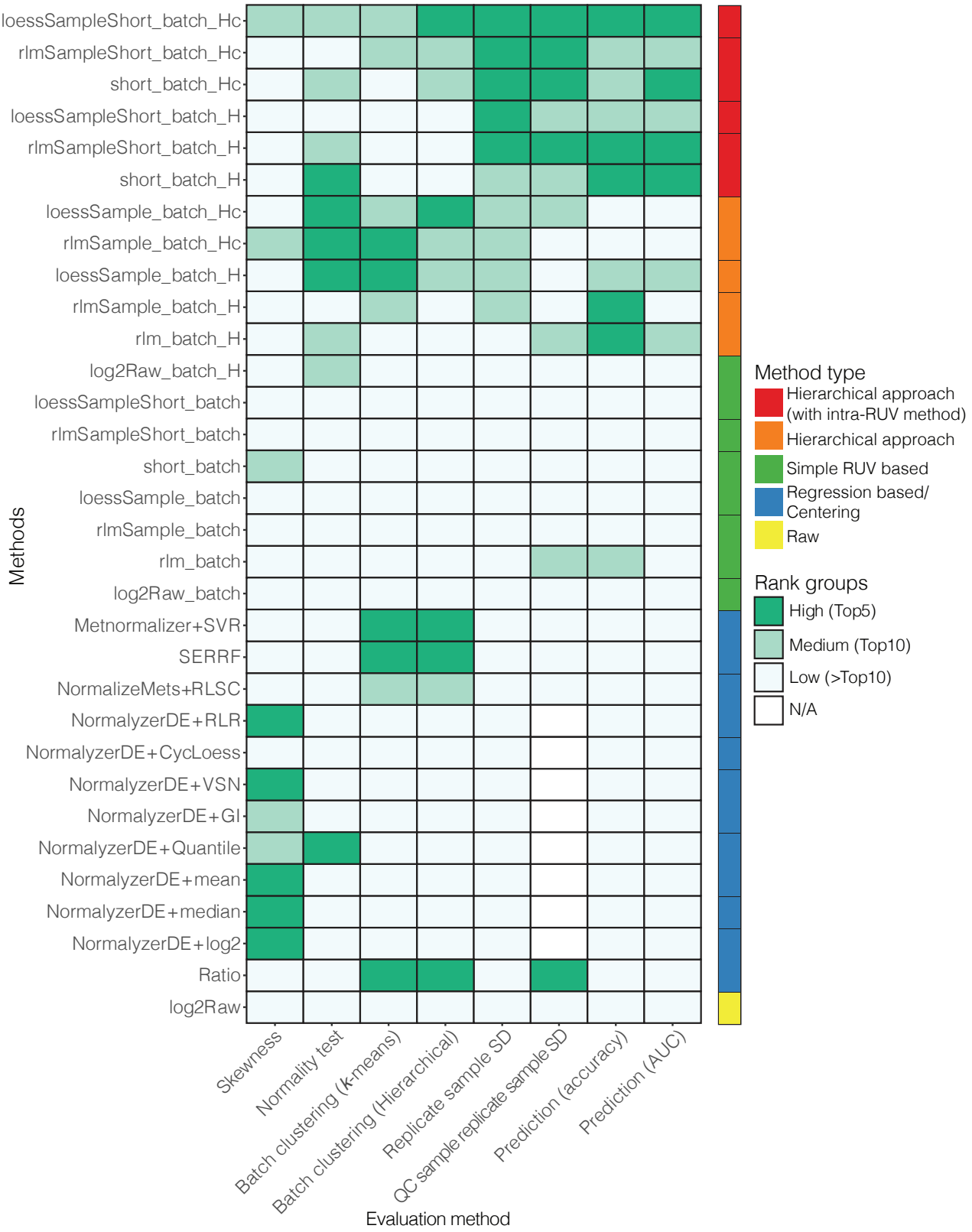


Fig. 4 Heatmap of rankings in all evaluations criteria. The y-axis represents all the methods explored in this study and the x-axis represents all the qualitative evaluation metrics used for evaluation of the integrated data. In each category, the evaluation scores are ranked and categorised

into 3 groups, high, medium and low. The coloured bar on the right indicates the categorised method approaches consistent with **Fig. 3b-c**.

Tables

Table 1 List of existing normalisation methods.

Tags	Method	Resource/implementation
log2Raw_batch	RUV-III[12]	R package ruv version 0.9.7.1
MetNormalizer+SVR	Support vector regression	R package MetNormalizer[5] version 1.3.02
NormalizeMets+RLSC	Robust locally estimated scatterplot smoothing[25]	R package NormalyzeMets[26] version 0.24
SERRF	Systematic error removal using random forest[9]	Online: https://slfan.shinyapps.io/ShinySERRF/
NormalyzerDE_RLR	Global robust linear regression[26]	R package NormalyserDE[8] version 1.7.0
NormalyzerDE_CycLoess	Cyclic loess[27]	
NormalyzerDE_VSN	Variance-stabilising normalisation[28]	
NormalyzerDE+GI	Global intensity	
NormalyzerDE+Quantile	Quantile normalisation[29]	
NormalyzerDE+mean	Mean[30]	
NormalyzerDE+median	Median[31]	
NormalyzerDE+log2	$\log_2$ transformation	
Ratio	Ratio[32]	

Table 2 A normalisation method abbreviation dictionary.

Tags	Definition
X_Y	Methods separated by '_' indicates 2 levels of adjustments applied. In this example, X is the intra batch adjustment applied and Y indicates the inter batch adjustment method applied.
X+Y	Methods separated by '+' denotes a method Y implemented in an R package X.

loess	A loess line fitting method with pooled QC samples.
rlm	A robust linear model fitted to pooled QC samples.
loessSample	A loess line fitting method only on biological samples.
rlmSample	A robust linear model fitted only on biological samples.
short	RUV with short (intra-batch) sample replicates
batch	RUV with batch (inter-batch) sample replicates
_H	A hierarchical balanced tree approach
_Hc	A hierarchical concatenating tree approach

669
670
671

Figures

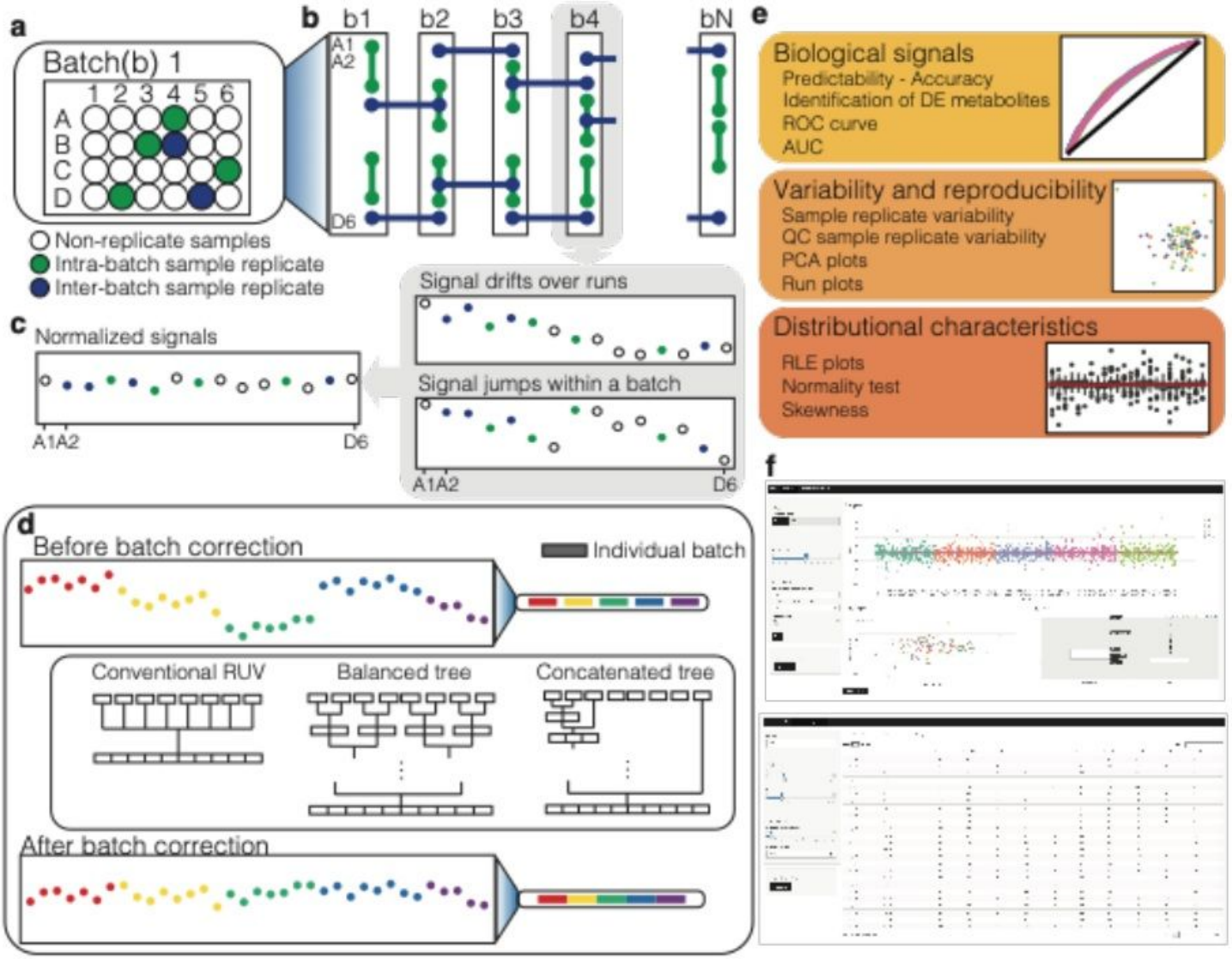


Figure 1

Schematic overview of the hRUV pipeline. a A schematic illustration of the plasma sample arrangement for the experimental batch in an array format where intra-batch sample replicates (green circles) and inter-batch sample replicates (blue circles) are embedded. b A schematic illustration of overall sample replicate design and arrangements. c Continuing the colour scheme from b, two illustrative run plots requiring intra-batch correction, with signal drift and other variations illustrated in the grey boxes. d A demonstration of signal variation before and after inter-batch correction in hRUV. A common approach to inter-batch correction is illustrated as conventional RUV and the proposed hierarchical approaches are illustrated. e A list of evaluation criteria to assess hRUV performances grouped into categories of biological signals, variability and reproducibility and distributional characteristics. f A screenshot of the user-friendly shiny application.

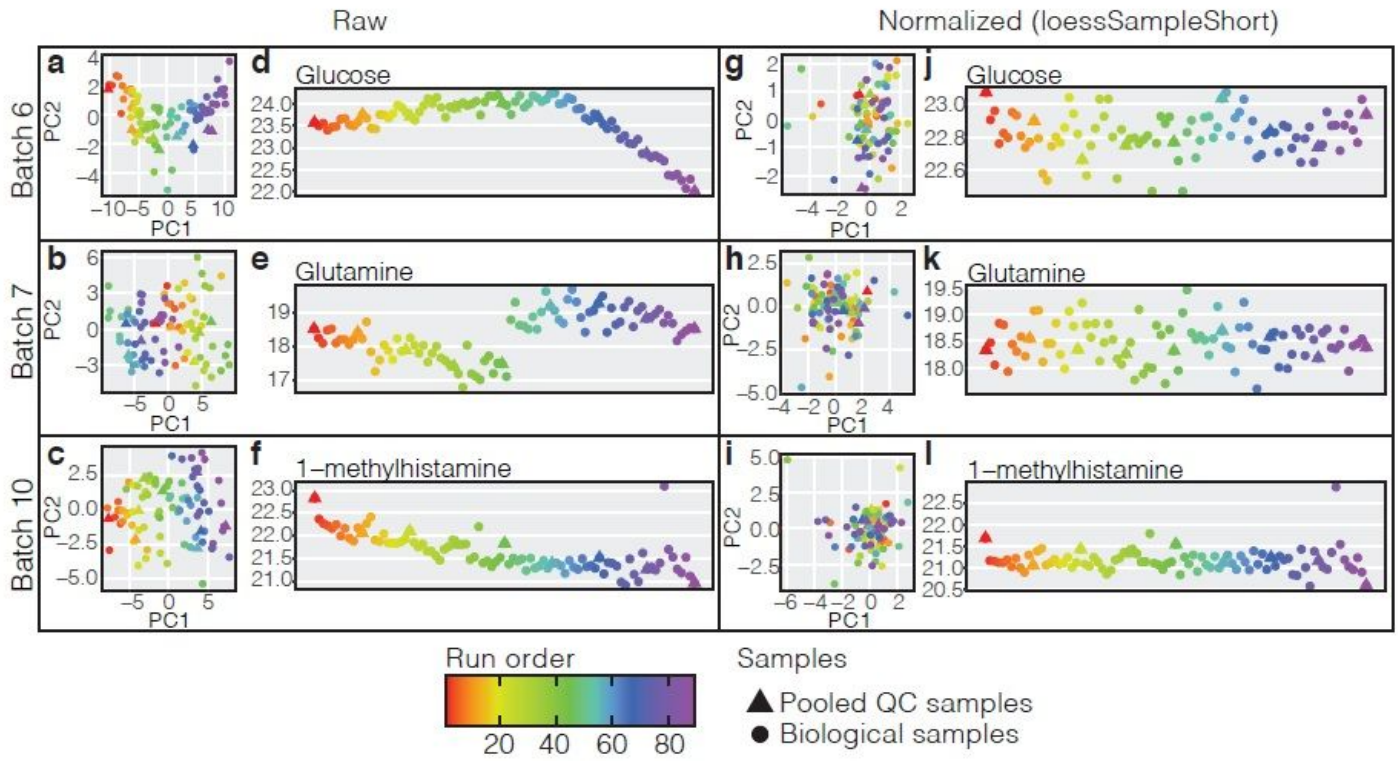


Figure 2

Examples of typical technical variations and signal drifts within each batch for different metabolites and comparison to normalised data. (Note vertical scale change.) a-c PCA plots of raw data in batches 6, 7 and 10 respectively, each marker is coloured by the sample run order. d-f Run plots exhibiting metabolite signal drift for glucose, glutamine, and 1-methylhistamine respectively, for their respective batches. g-i PCA plots from the same batches as a-c respectively with intra-batch normalised data using loessSampleShort method. j-l Run plot for the metabolites and batches illustrated in d-f but with intra-batch data normalised using loessSampleShort method.

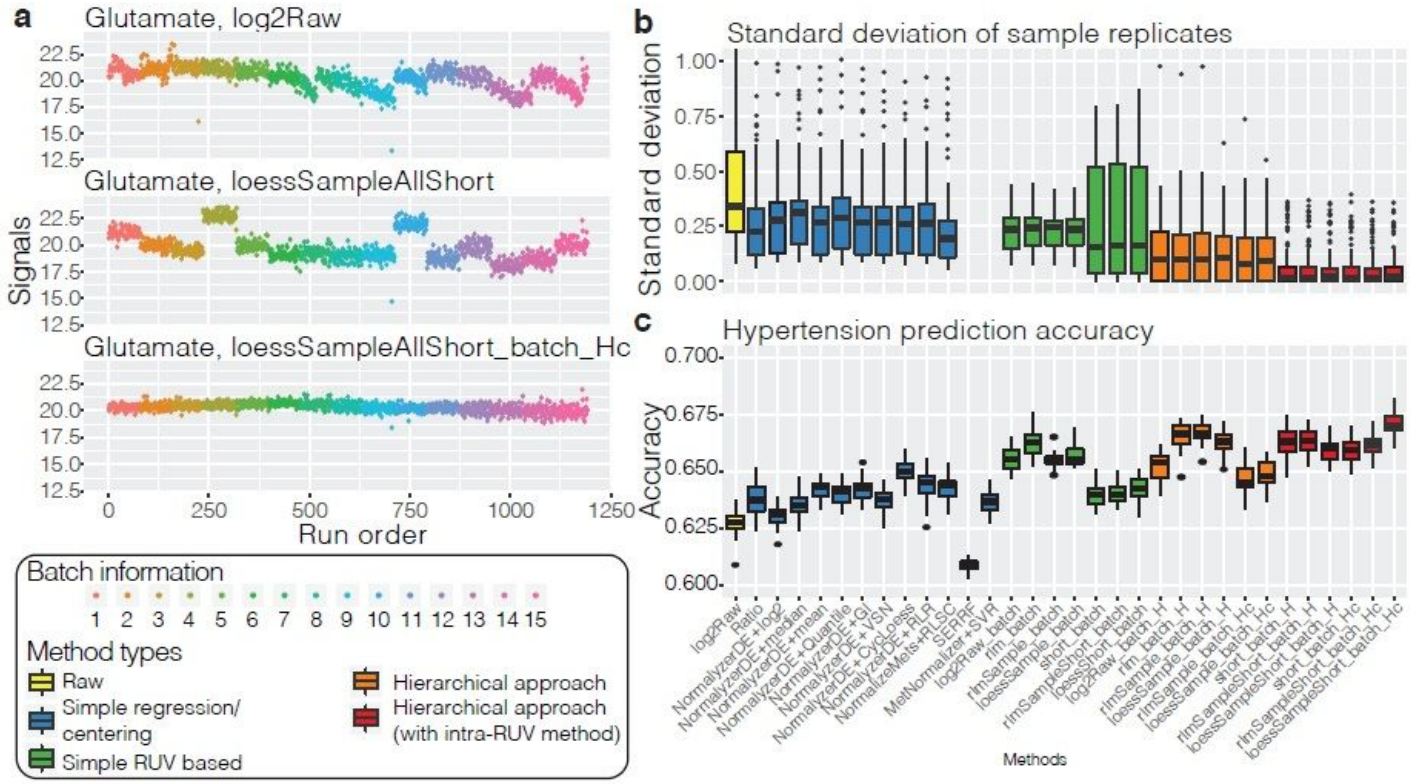


Figure 3

Key assessments of hRUV performances. a A run plot of raw, intra-batch corrected and final hRUV normalised data in all 15 batches of the BioHEART-CT cohort. The x-axis indicates the run order, the y-axis indicates the signal of glutamate, and samples are coloured by the batch numbers. b Boxplots of all sample replicate standard deviations, where lower values indicate better performance. The boxes are coloured by the approach taken to normalize the data. The y-axis of the plot is restricted to a range between 0 to 1 to highlight the differences between the majority of the methods. SERRF and MetNormalizer+SVR's median sample replicate SD was greater than 1 and thus is not shown. c Boxplots of hypertension prediction accuracies for all methods. Higher prediction accuracy indicates better performance. The colouring of the boxes are consistent with that in b.

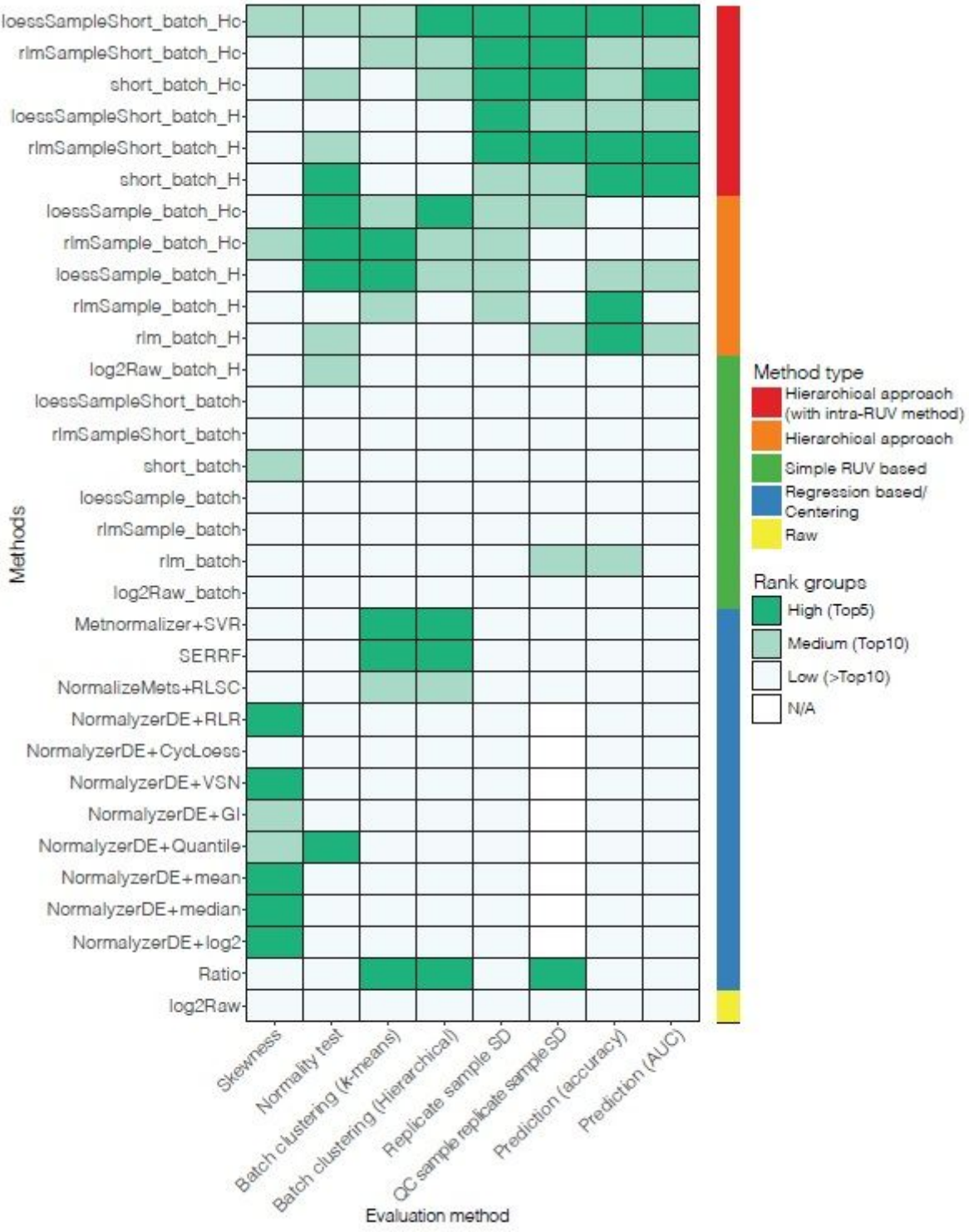


Figure 4

Heatmap of rankings in all evaluations criteria. The y-axis represents all the methods explored in this study and the x-axis represents all the qualitative evaluation metrics used for evaluation of the integrated data. In each category, the evaluation scores are ranked and categorised log2Raw into 3 groups, high, medium and low. The coloured bar on the right indicates the categorised method approaches consistent with Fig. 3b-c.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- hRUVsupplementaryinformation.docx
- hRUVSupplementaryTable1.xlsx