metabolites

MDPI

Article

Using Out-of-Batch Reference Populations to Improve Untargeted Metabolomics for Screening Inborn Errors of Metabolism

Michiel Bongaerts [1,*], Ramon Bonte [1], Serwet Demirdas [1], Edwin H. Jacobs [1], Esmee Oussoren [2], Ans T. van der Ploeg [2], Margreet A. E. M. Wagenmakers [3], Robert M. W. Hofstra [1], Henk J. Blom [1], Marcel J. T. Reinders [4] and George J. G. Ruijter [1,*]

1 Department of Clinical Genetics, Erasmus Medical Centre, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands; r.bonte@erasmusmc.nl (R.B.); s.demirdas@erasmusmc.nl (S.D.); e.jacobs@erasmusmc.nl (E.H.J.); r.hofstra@erasmusmc.nl (R.M.W.H.); h.j.blom@erasmusmc.nl (H.J.B.)
2 Department of Pediatrics, Center for Lysosomal and Metabolic Diseases, Erasmus Medical Centre, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands; e.oussoren@erasmusmc.nl (E.O.); a.vanderploeg@erasmusmc.nl (A.T.v.d.P.)
3 Department of Internal Medicine, Center for Lysosomal and Metabolic Diseases, Erasmus Medical Centre, Dr. Molewaterplein 40, 3015 GD Rotterdam, The Netherlands; m.wagenmakers@erasmusmc.nl
4 Faculty of Electrical Engineering, Mathematics and Computer Science, TU Delft, Van Mourik Broekmanweg 6, 2628 XE Delft, The Netherlands; M.J.T.Reinders@tudelft.nl
* Correspondence: m.bongaerts@erasmusmc.nl (M.B.); g.ruijter@erasmusmc.nl (G.J.G.R.)

Abstract: Untargeted metabolomics is an emerging technology in the laboratory diagnosis of inborn errors of metabolism (IEM). Analysis of a large number of reference samples is crucial for correcting variations in metabolite concentrations that result from factors, such as diet, age, and gender in order to judge whether metabolite levels are abnormal. However, a large number of reference samples requires the use of out-of-batch samples, which is hampered by the semi-quantitative nature of untargeted metabolomics data, i.e., technical variations between batches. Methods to merge and accurately normalize data from multiple batches are urgently needed. Based on six metrics, we compared the existing normalization methods on their ability to reduce the batch effects from nine independently processed batches. Many of those showed marginal performances, which motivated us to develop *Metchalizer*, a normalization method that uses 10 stable isotope-labeled internal standards and a mixed effect model. In addition, we propose a regression model with age and sex as covariates fitted on reference samples that were obtained from all nine batches. *Metchalizer* applied on log-transformed data showed the most promising performance on batch effect removal, as well as in the detection of 195 known biomarkers across 49 IEM patient samples and performed at least similar to an approach utilizing 15 within-batch reference samples. Furthermore, our regression model indicates that 6.5–37% of the considered features showed significant age-dependent variations. Our comprehensive comparison of normalization methods showed that our *Log-Metchalizer* approach enables the use out-of-batch reference samples to establish clinically-relevant reference values for metabolite concentrations. These findings open the possibilities to use large scale out-of-batch reference samples in a clinical setting, increasing the throughput and detection accuracy.

Keywords: untargeted metabolomics; inborn errors of metabolism; normalization; internal standards; batch effects

Citation: Bongaerts, M.; Bonte, R.; Demirdas, S.; Jacobs, E.H.; Oussoren, E.; Ploeg, A.T.v.; Wagenmakers, M.A.E.M.; Hofstra, R.M.W.; Blom, H.J.; Reinders, M.J.T.; Ruijter, G.J.G. Using Out-of-Batch Reference Populations to Improve Untargeted Metabolomics for Screening Inborn Errors of Metabolism. *Metabolites* **2021**, *11*, 8. https://dx.doi.org/10.3390/metabo11010008

Received: 27 October 2020
Accepted: 18 December 2020
Published: 25 December 2020

1. Introduction

The screening of patients suspected for inborn errors of metabolism (IEM) is currently based on measuring panels of specific groups of metabolites, like amino acids or organic acids using a number of different tests, and techniques, such as ion-exchange chromatography, liquid chromatography mass spectrometry (LC-MS) and gas chromatography mass

spectrometry (GC-MS). This targeted approach with several different tests is time consuming and limited in the number of metabolites being analyzed. Untargeted metabolomics using high resolution accurate mass liquid chromatography mass spectrometry (HRAM LC-MS) can detect hundreds to thousands of metabolites within one test and, as a consequence, receives increasing interest to be used in IEM screening [1–5]. Moreover, untargeted metabolomics can also reveal new biomarkers or increase our understanding of disease mechanism when exploited in epidemiological studies [6].

In traditional targeted diagnostic laboratory tests, hundreds of reference samples are required for establishing robust reference intervals. When using untargeted metabolomics, the establishment of reference values is complicated, due to the semi-quantitative nature of the data, owing to several sources of variation, like injection volume, retention time, temperature, or ionization efficiency in the mass spectrometer that cannot easily be amended. Moreover, these variations are even larger between different measurement runs in which a batch of samples is being measured simultaneously, hampering the resemblance between different batches. Consequently, the within-batch variation is smaller than between-batch variation. Targeted metabolomics generally deals with these technical variations by using internal standards that are chosen such that they are chemically identical (an isotope) or similar to the metabolite of interest, thereby making it possible to accurately measure its quantity. However, since untargeted metabolomics involves the measurement of a large number of different metabolites/features, it becomes unfeasible to add an internal standard for each metabolite. Therefore, in order to conquer these batch effects, the current approaches include reference samples in each single batch of measurements [1–5] to improve detection sensitivity (due to tighter reference values as a result of lower variation in the within-batch reference samples).

Clearly, this reduces the throughput efficiency of IEM screening, as the number of patient samples that can be included in a batch is considerably lower when reference samples also need to be measured. However, more importantly, the number of reference samples in one batch might fall short in the establishment of adequate reference ranges as variations in certain metabolites are not captured well enough in the relatively small reference panel. For example, factors, like age, sex, and BMI, can affect abundancies of metabolites and, to establish reliable reference ranges, one thus needs to correct for these factors by using a large number of reference samples [7–10]. Consequently, for reliable untargeted metabolomics in clinical testing, a large set of reference samples is needed, while, for throughput efficiency, a small set is preferred. Altogether, this calls for an approach that can establish reference values that are based on reference samples being measured in several batches (out-of-batch controls).

When relying on reference samples from different batches, one needs to correct for the batch effects in order to obtain reliable estimates for the reference ranges. This is generally solved by normalization methods and some have already been proposed within the context of untargeted metabolomics and mass spectrometry [11–13]. Only a few groups have used out-of-batch reference samples to determine the reference values and used relatively simple normalization techniques, like median scaling [1], a reference internal standard per metabolite [3], or anchor samples [6]. However, there has not been an extensive exploration of normalization techniques within the context of diagnostic testing for IEM's.

We explore several known normalization methods for their ability to remove batch effects and detect biomarkers from patients with known IEM. Furthermore, we introduce a new normalization method, which we called *Metchalizer*, which uses internal standards and a mixed effect model to remove batch effects. As this allows for a large set of (out-of-batch) reference samples, we also explore a regression model that uses age and sex as covariates to correct for potential age and sex effects on the reference values. Using the regression model combined with the *Metchalizer* normalization, we achieve similar performances in biomarker detection as compared to the use of within-batch controls. Hence, this opens the possibility to increase the throughput of untargeted metabolomics in IEM screening as well as including more complex confounder strategies. *Metchalizer* and the regression model are available at https://github.com/mbongaerts/Metchalizer.

2. Results

2.1. Data and Batch Characteristics

Using ultra-high performance liquid chromatography-Orbitrap-MS (UHPLC-Orbitrap-MS), nine untargeted metabolomics runs/batches were measured containing 261 control samples and 58 IEM patient samples, together having 35 unique IEMs. All nine batches were measured on a single mass spectrometer (Thermo Scientific Q Exactive Plus), while three separate Kinetex F5 columns for ultra-high performance liquid chromatography (UHPLC) were used. Using in-house developed software, features across the nine batches were matched and accordingly merged into a single dataset (see Section 4.2). After merging, 446 positively ionized features were obtained, among which 114 were annotated, and 328 negatively ionized features were attained with 82 annotated features. We only included features that were merged across all nine batches to ensure consistency among the findings. This resulted in the loss of IEM biomarkers, and the full list of the lost biomarkers per IEM can be found in Appendix E. Intra-batch coefficients of variation (CV) on 17 (internal and external) standards were smaller (median CV = 14%, see Appendix A Figure A1) than inter-batch CV's (median CV = 27%, see Appendix A Table A1), indicating that batch effects were present. Principle Component Analysis (PCA) further demonstrated the presence of batch effects, as shown in Figure 1A,B, showing the first three PCs for the raw abundancies (*Raw* and *Log-Raw*).

2.2. Comparing Normalization Methods

We investigated the performance of several normalization methods on batch effect removal by evaluating multiple metrics that are based on quantitative measurements, the Quality Control (QC) samples and PCA analysis (Section 4.4.5).

Reduced batch effects: from the PCA plots, we observe that most normalization methods reduced batch effects, since batch clustering seemed to be reduced after normalization (Figure 1), which is confirmed when looking at the *batch prediction score* (Figure 2A), showing lower scores for normalized abundancies when compared with the raw data (*Raw* or *Log-Raw*). *BC-Metchalizer*, *Log-Metchalizer* had the lowest *batch prediction scores*, with median scores of 0.12 (0.11), 0.12 (0.12) for positive (negative) ion mode, respectively (see Appendix B Table A2 for all medians).

Most methods conserve separation of QC samples: QC samples were included in every batch and thet were expected to segregate from the human plasma samples in the first four principle components (PC) due to overall abundancy differences for several metabolites (see Figure 1 for the first 3 PCs). Normalization should maintain this separation, which was quantified by the *QC prediction score* (Figure 2B). We observe that, for most normalization methods, the median *QC prediction score* was about 1.00. Although, *Log-NOMIS* scored relatively well when considering the *batch prediction scores*, with a median score of 0.20 (0.17) for positive (negative) ion mode, it performed poor on the *QC prediction score*, with a median of 0.33 (0.88) for positive (negative) ion mode. Therefore, it is likely that this method removed variations other than batch related variation.

Resemblance with quantitative measurements: to further quantify batch effect removal, we calculated the *Spearman score* and R^2 *score* between quantitative plasma concentrations (in µmol/L) and the normalized abundancies of our evaluation set of amino acids and (acyl)carnitines (Section 4.4.5). In order to ensure high signal-to-noise ratio's in the quantitative measurements, we selected only metabolites having a population average concentration above 1 µmol/L. Matching the evaluation set with the annotated features in the untargeted metabolomics data resulted in 15 and 10 metabolites in positive and negative ion mode, respectively. Figure 2C,D shows both metrics for the investigated normalization methods. Again, for most normalization methods, both of the metrics improved when compared to the raw data (*None-Raw*). *BC-Metchalizer*, *Log-Metchalizer* and *None-Anchor* appeared to perform the best on these metrics with median R^2 scores of 0.66 (0.64), 0.61 (0.68), 0.63 (0.57), and median *Spearman scores* of 0.78 (0.79), 0.78 (0.83), 0.75 (0.74), for positive (negative) ion mode.

Figure 1. Principle Component Analysis (PCA) plots for raw data and normalized data as indicated by the title of each panel. Each batch is indicated with a unique color. PCA was performed on 431 features (excluding the internal and external standards) in positive ion mode. The squares indicate QC samples, whereas the circles indicate patient and control samples. (**A**) PCA plot for *Raw*, (**B**) *Log-Raw*, (**C**) *BC-Metchalizer*, (**D**) *Log-Metchalizer*, (**E**) *None-Anchor*, (**F**) *Log-NOMIS*, (**G**) *Log-CCMN*, (**H**) *Log-EigenMS*, (**I**) *Log-Fast Cyclic Loess*, (**J**) *Log-RUVrand*, (**K**) *None-VSN*, (**L**) *None-PQN*, (**M**) *None-Best correlated IS*.

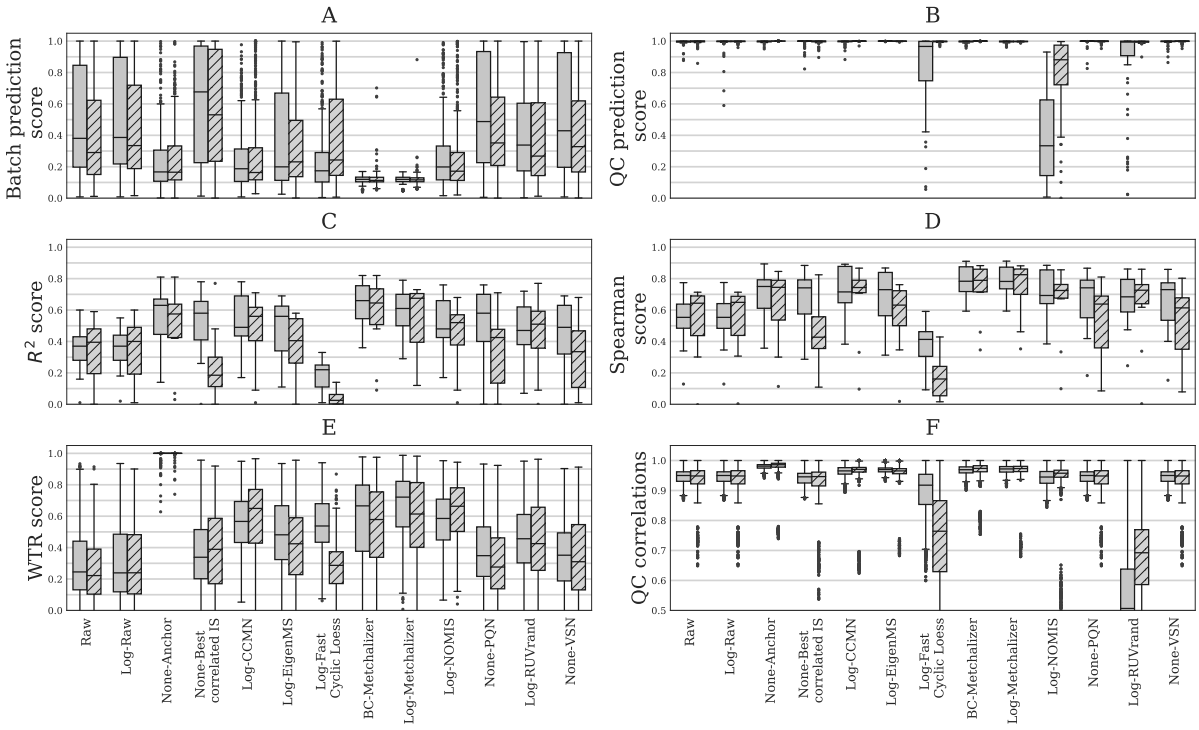


Figure 2. Six different performance metrics for batch effect removal (Section 4.4.5). Data from positive – or negative ion mode is indicated by plain and striped boxplots, respectively. (**A**) *Batch prediction score* measures the presence of batch effects in the first four principal components (PCs) from PCA analysis. (**B**) Quality Control (QC) prediction score measures how well QC samples are separated from human plasma sample in the first four PCs. (**C**) R^2 score between (normalized) abundancies and quantitative measurements. (**D**) *Spearman score* of (normalized) abundancies with quantitative measurements. (**E**) The *WTR score* measuring the overall within batch variation with respect to the total variance using the QC samples. (**F**) *QC correlations* measuring the resemblance of all QC samples among each other. Each data point represents a pair-wise Spearman correlation between two QC samples.

Reduced between-batch variation in QC samples: next, we compared the within-batch variance of the QC samples with respect to the total variance which is expressed by the *WTR score* (Section 4.4.5) for each normalization method. *WTR scores* close to 1 indicate the absence of batch effects. *None-Raw* and *Log-Raw* had low WTR scores and after normalizing these scores increased (Figure 2E). *BC-Metchalizer* and *Log-Metchalizer* scored among the highest on this WTR score. *None-Anchor* had high WTR scores, which was expected, since *None-Anchor* uses the QC samples for normalization and, consequently, the *WTR scores* are biased towards higher values.

Preserved resemblance of QC samples: removal of variation results in higher WTR scores, but also potentially removes variation(s) of interest. Therefore, we investigated whether the resemblance of all QC samples among each other was conserved after normalization using the Spearman correlation. Lower Spearman correlations indicate that variation of interest might also be lost, since the resemblance between the QC samples is reduced. Figure 2F shows the *QC correlations* for each normalization method. These results show that *Log-Fast Cyclic Loess* and *Log-RUVrand* also removed the non-batch related variations, even while having relatively good *QC prediction scores*.

Additionally, we investigated whether normalization improved the resemblance with patients sharing the same IEM and, likewise, reduced the resemblance between patients having a different IEM. This analysis shows that *BC-Metchalizer* and *Log-Metchalizer* were among the best when considering two different resemblance scores (see Appendix K for more details).

Taken together, *BC-Metchalizer*, *Log-Metchalizer*, and *None-Anchor* showed the optimal normalization characteristics across the evaluation metrics and were evaluated in more detail (see Section 2.4).

2.3. Confounder Effects of Age and Sex

We developed a regression model with sex as covariate and age as a polynomial (p = 1, 2, 3) covariate in order to explore confounding effects of age and sex on metabolite abundancies (see Section 4.5.2). After normalization, we fitted the model parameters for every feature while using all of the samples that are present in the nine batches and determined the significance of the coefficients in the regression model (Section 4.5.2). The obtained p-values were corrected for multiple testing per coefficient and ion mode using the Benjamin–Hochberg procedure (FWER = 0.05). Table 1 shows the percentages of significant coefficients ($\alpha = 0.05$) per ion mode and (selected) normalization method. Our findings suggest that 6.5–37% of all features showed age dependency when looking at coefficient $\hat{\beta}_1^{\text{Age}}$ (i.e., the linear term in the model). It is noteworthy that more age-related features were found in the negative ion mode.

Table 1. The percentage of significant coefficients in the regression model for a given ion mode and normalization method.

Coefficient	Ion Mode	None-Anchor	BC-Metchalizer	Log-Metchalizer
$\hat{\beta}^{\text{Intercept}}$	−	98.4	100.0	100.0
$\hat{\beta}^{\text{Intercept}}$	+	100.0	100.0	100.0
$\hat{\beta}^{\text{Age}_1}$	−	22.9	36.9	31.8
$\hat{\beta}^{\text{Age}_1}$	+	6.5	12.5	13.2
$\hat{\beta}^{\text{Age}_2}$	−	4.5	17.8	16.9
$\hat{\beta}^{\text{Age}_2}$	+	0.5	4.4	5.1
$\hat{\beta}^{\text{Age}_3}$	−	1.6	8.3	8.3
$\hat{\beta}^{\text{Age}_3}$	+	0.2	0.2	0.5
$\hat{\beta}^{\text{Sex}}$	−	0.0	0.0	0.0
$\hat{\beta}^{\text{Sex}}$	+	0.0	0.0	0.0
$\hat{\beta}^{\text{Sex,Age}}$	−	0.0	0.0	0.0
$\hat{\beta}^{\text{Sex,Age}}$	+	0.5	0.7	0.0

Although the significance of the regression coefficient indicates whether the determined coefficient is a true finding, the (relative) magnitude of the coefficient determines the effect size. While selecting only significant coefficients $\hat{\beta}_1^{\text{Age}}$ with an effect size larger than 2% per year (see Appendix C Figure A3 for explanation), we found that around 1–7% of all features in positive ion mode, and 5–22% of all features in negative ion mode, showed (strong) age-dependency (Appendix C Table A5). Moreover, age-dependent features have the tendency to increase/decrease in abundancy faster at younger and older ages, which implies that a matching reference population for these age groups are more important (see Appendix C Figure A2).

When using normalization by *BC-Metchalizer*, age-dependent metabolites (Appendix C Table A3), include known IEM biomarkers, such as: guanidinoacetic acid(+), homoarginine (−), 2-ketoglutaric acid (−), C3 propionylcarnitine (+), phenylacetic acid (+), and uridine (−). As an example, we plotted the regression model for guanidinoacetic acid (Figure 3), illustrating that the Z-score for a fixed abundancy depends on age (and slightly on sex at later ages). This also shows a non-linear trend with age. Our analyses showed that more metabolites have significant non-linear trends over age ($\hat{\beta}_2^{\text{Age}}$ and $\hat{\beta}_3^{\text{Age}}$ in Table 1).

No significant gender-related features were found (Table 1) and just 0.5%, 0.7% of all features in positive ion mode showed significant sex-age interaction ($\hat{\beta}^{\text{Sex,Age}}$), for *None-Anchor* and *BC-Metchalizer*, respectively. Among these features are biomarkers: guanidinoacetic acid(+) and ornithine(−). See Appendix C Table A4 for more details.

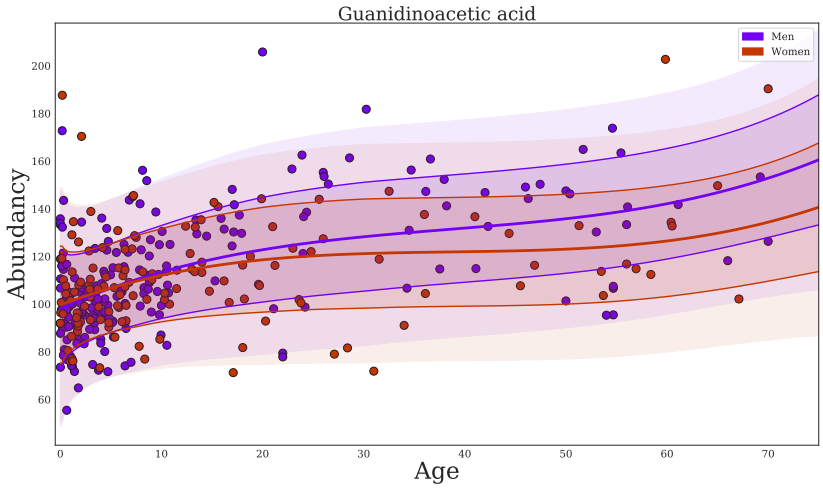


Figure 3. Regression of guanidinoacetic acid when using *BC-Metchalizer* normalized data. The different colors indicate the sex as shown in the legend. The thick red/blue line indicates the average obtained from the fit on all samples for a given sex. The first standard deviation is indicated by the thin(ner) line whereas the second standard deviation ends at the shaded region.

2.4. Detection of the Expected IEM Biomarkers

Next, we investigated the impact of normalization and using out-of-batch reference samples on expected biomarker detection in the 49 (/58) IEM patients (see Section 4.6 and Appendix D Table A6) by plotting the number of detected expected biomarkers (expected true positives) against the average number of positives (true plus false positives) per patient at various Z-score or *p*-value thresholds (Section 4.6), similar to a Receiver Operator Curve (ROC). Untargeted metabolomics did not allow for us to make a distinction between false positives and true positives, due to unannotated features and even unknown disease related features/biomarkers. When assuming that the majority of the positives per patient are false positives, we used the average number of positives per patient as a proxy for the false positives. Improved performance was considered to increase the number of detected expected biomarkers (true positives of which we are certain) while lowering the average number of positives per patient, thereby increasing the Area Under the Curve (AUC) (see Section 4.6 for more explanation).

We decided to take the method that uses 15 within-batch reference samples and raw abundancies (*15in&None-Raw*) as the reference approach. Performance was expressed as a percentage of this reference AUC, as indicated by $AUC^{x}_{15in\&None\text{-}Raw}$ (where x indicates if the AUC was created from the average Z-scores or *p*-values). These *p*-values were obtained from the Welch's *t*-test, which tests whether the average Z-score of an expected biomarker or feature across the triplicate significantly differs from the average Z-score of the reference population (Section 4.5.4).

Log-transform improves biomarker detection for *p*-values: our first observation is that, when considering the Z-scores, the log-transformed raw abundancies (*15in&Log-Raw*) have an AUC approximately equal to $AUC^{Z}_{15in\&None\text{-}Raw}$ (Figure 4), implying that this transformation hardly affected this performance metric. However, when using the *p*-values, the log-transformation improved the detection of the expected biomarkers, as $AUC^{p}_{15in\&Log\text{-}Raw}$ is 8% higher than the $AUC^{p}_{15in\&None\text{-}Raw}$ (Figure 4).

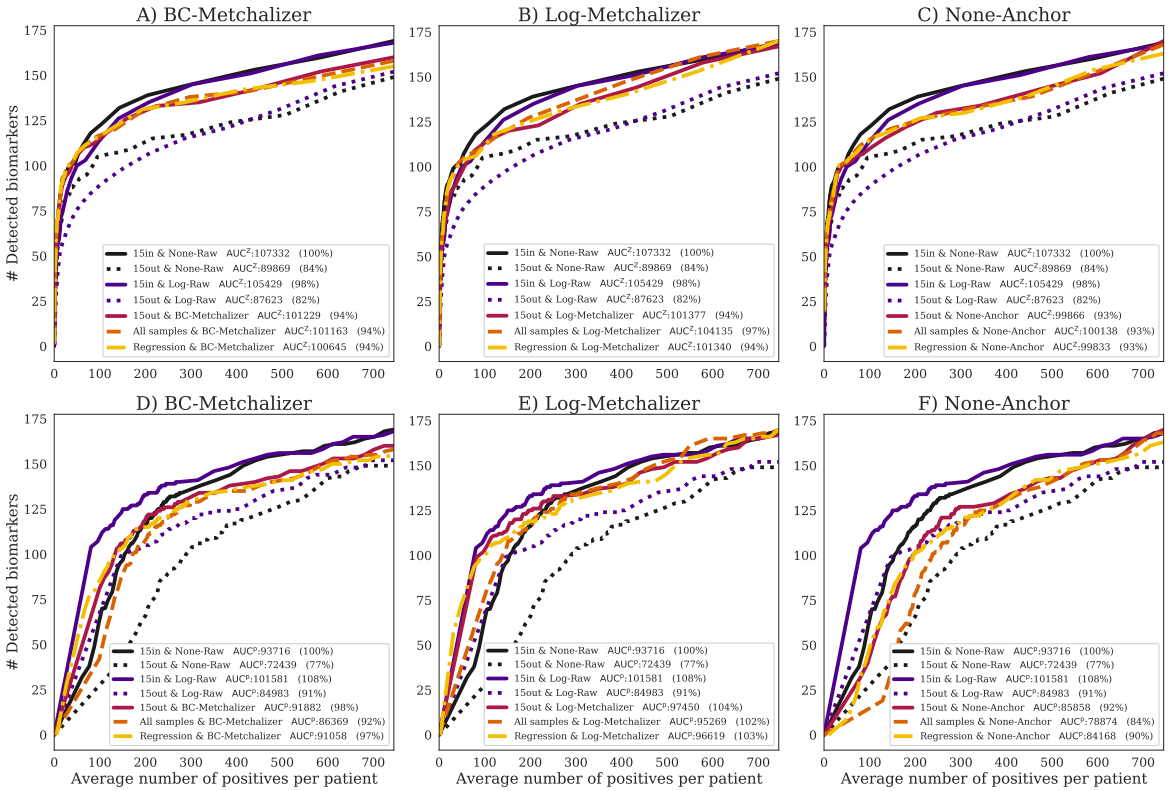


Figure 4. The number of detected expected biomarkers versus the average number of positives per patient. A curve in each (sub)figure was formed by increasing the Z-score or *p*-value threshold (Z_{abnormal} , Methods). The legend indicates (per curve) the methods used to determine Z-scores and how data was normalized, the AUC and AUC expressed as percentage of the $\text{AUC}^{x}_{\text{15in\&}\textit{None-Raw}}$. Performances using (**A**) *BC-Metchalizer* and Z-scores, (**B**) *Log-Metchalizer* and Z-scores, (**C**) *None-Anchor* and Z-scores, (**D**) *BC-Metchalizer* and *p*-values, (**E**) *Log-Metchalizer* and *p*-values, and (**F**) *None-Anchor* and *p*-values.

Reduced performance with age/sex matched out-of-batch references: when comparing the performance of using 15 out-of-batch samples (*15out&None-Raw*) to the *15in&None-Raw* reference, the performance for *15out* was clearly reduced (Figure 4A), achieving only 84% of the reference $\text{AUC}^{Z}_{\text{15in\&}\textit{None-Raw}}$. This difference was also present when looking at the *p*-values, which resulted in a clear reduction of the $\text{AUC}^{p}_{\textit{15out\&None-Raw}}$ (77%). Hence, the potential improved age/sex matching for *15out*, due to the increased number of available reference samples (Appendix H), did not result in improved performance, most likely due to the dominance of batch effects.

Normalization improves performance of age/sex matched out-of-batch controls: after normalizing with *BC-Metchalizer*, *Log-Metchalizer*, or *None-Anchor* and using 15 out-of-batch controls (*15out*), the performance increased when compared to *15out&None-Raw* (Figure 4A–C), and it came closer to $\text{AUC}^{Z}_{\text{15in\&}\textit{None-Raw}}$; for *BC-Metchalizer* 94%, *Log-Metchalizer* 94%, and *None-Anchor* 93%. Interestingly, when considering biomarker detection performance using the *p*-values, the *BC-Metchalizer* performed on par with *15in&None-Raw* (98%), *Log-Metchalizer* improved over *15in&None-Raw* (104%), while *None-Anchor* was 92%. *Log-Metchalizer* performed similarly to *15in&Log-Raw* (104% and 108%, respectively), indicating that out-of-batch samples can be used instead of within-batch samples to determine reference values.

Regression model effectively models age and sex effects: the performance for AUC^{Z} using the regression model (*Regression*) remained the same for all considered normalization methods with respect to *15out*, see also Figure 4A–C. When considering the *p*-values, AUC^{p}, the performance was also similar to *15out*; *BC-Metchalizer* (−1%), *Log-Metchalizer* (−1%), *None-Anchor* (−2%) (Figure 4 D–F). Interestingly, when we took all of the reference samples to determine the Z-scores (*All samples*, Methods), similar AUC^{Z} performances

were observed when compared to *Regression*, i.e., +0% for *BC-Metchalizer* and +3% *Log-Metchalizer* and +0% for *None-Anchor*. When considering the *p*-values the difference were larger, i.e., −5% for *BC-Metchalizer* and −1% *Log-Metchalizer*, and −6% for *None-Anchor*, suggesting an influence of age- and sex effects on the detection of biomarkers.

3. Discussion

Targeted measurements of metabolites in body fluids using various platforms, such as HPLC, GC-MS, and LC-MS/MS are traditionally applied for laboratory diagnosis of IEM. For each individual metabolite, age- and, sometimes, sex-dependent reference ranges are established while using hundreds of reference samples. Untargeted metabolomics is a promising alternative by enabling the determination of many metabolites in one analysis. This can speed up the diagnostic process and extend the number of different IEMs that can be screened in a single run. A major drawback of current approaches is that reference samples need to be included in the same experimental batch in order to ensure proper reference ranges (or Z-score transformations). Some methods do exist that use reference samples that were measured in different batches (out-of-batch samples) to determine age and sex corrected Z-scores, and they are based on methods that remove the technical variations. There has not been a comprehensive comparison of the different normalization methods with approaches that use out-of-batch samples, which we have set out in this work. Moreover, we developed a new normalization method, *Metchalizer*, which makes use of internal standards, an approach that has been shown to be useful when mapping specific metabolites to specific internal standards [3], and that we generalize to all features measured. Because more reference samples are available when using the out-of-batch samples, we additionally propose a regression model that incorporates sex and age effects as (non-linear) covariates. Altogether, we have shown that our methodology has biomarker detection performances that are at least similar to using 15 within-batch samples.

Typically, around 20,000 features in both negative and positive mode were detected per batch. When we require a feature to be measured (and matched) in all nine batches, we retained 446 positively and 328 negatively ionized features, respectively. Because some normalization methods use a statistical approach (*PQN*, *Fast Cyclic Loess*), the reduction in features might explain the reduced performance of these methods. In addition, the requirement of features being measured (and matched) across all nine batches resulted in the loss of clinically relevant biomarkers (see Appendix E), which is a significant limitation of using out-of-batch samples. This suggests that within-batch references are still required when this limitation cannot be conquered. As an alternative, we could have made the inclusion of features dependent on fewer batches (for example, being present/matched in >5 out of 9 batches). We decided not to do that in this study, as this would have resulted in an unequal number of reference samples for the different features, leading to inconsistent results between the out-of-batch methods. The availability of more batches could have solved this issue, because an equal number of reference samples could likely be obtained per feature, even when these features were not present/matched in some batches. It is interesting to note that our proposed normalization method (*Metchalizer*) showed consistent performances when data from a varying number of batches were used (Appendix G). Some biomarkers, for example, isobutyrylglycine, were only detected in the batches containing patient samples with elevated levels of these specific metabolites. We anticipate that, for this kind of biomarkers, out-of-batch strategies are less useful, since abundancies in (normal) references are (very) low, thereby making out-of-batch Z-score calculation unsuitable.

Anchor uses anchor (fixed) samples, measured in all batches, in order to normalize the features. *Anchor* normalization on none-transformed data performed well when compared to most of the other normalization methods explored, but slightly less than *BC-Metchalizer* and *Log-Metchalizer* when considering the performance metrics *Spearman score*, R^2 *score*, *batch prediction score*, and performance on biomarker detection. We anticipate that the anchor samples may not correlate with all types of variation, like, for example,

injection volume, which is a source of variation at the sample level, whereas the abundancy of the internal standards (used by *Metchalizer*) is directly linked to the injection volume. *Anchor* also assumes that metabolite levels remain constant over time in the anchor samples. As a consequence, if, for example, storage effects take place, *Anchor* is impeded. The use of *Anchor* may also be less practical, because it requires the same anchor samples in every batch. The introduction of a new anchor sample requires an 'overlapping batch' containing a set of both the former anchor sample together with the newly introduced anchor samples.

Metchalizer is based on the basic assumption that all variations which can be explained by the variation in the abundancies of the internal standards (being the latent variables from PLS analysis) are technical variations, including batch effects. It exploits the linear relationship between these latent variables and the feature being measured across all samples, thereby capturing the covariance between the standards and that feature. *Metchalizer* assumes that this relationship holds across batches and with that assumption determines (batch) intercepts that correct for 'unexplained' batch/technical variations. Consequently, *Metchalizer* can correct for large batch effects, but this comes with the potential danger of overcorrection when batches differ from each other due to biological variation, which will then be interpreted as 'unexplained' batch/technical variations. For this reason, it is important to use randomized samples in each batch (in terms of age, sex, etc.) in order to minimize the possibility of biological variations between batches.

Log-Metchalizer log transforms the abundancies before applying *Metchalizer*, whereas the *BC-Metchalizer* uses a less strong Box–Cox transformation. The effect of this stronger transformation on most investigated metrics in this study was small, although we did observe that a stronger initial transformation led to improved biomarker detection performances when considering the *p*-values. *15in&None-Raw* had a lower AUC^p than *15in&Log-Raw* and it could therefore also explain the improved performance of *Log-Metchalizer* over *BC-Metchalizer* on this metric. A simulation showed that log-transforming the raw abundancies indeed caused differences in the obtained Z-scores and *p*-values when compared to the raw abundancies (Appendix I). Positive Z-scores had relatively lower *p*-values (and vice versa) for log-transformed abundancies and this could therefore explain the improved performance on biomarker detection, since most of the considered biomarkers had positive Z-scores, thus biasing this performance metric. Increasing the number of internal standards did not improve the normalization performance when considering metrics that are based on the quantitative measurements, although we observed that certain combinations of internal standards improved the normalization of specific metabolites (Appendix F). This suggests that *Metchalizer* might be improved by matching features/metabolites with a certain set of internal standards (for example, based on retention time).

We were a bit surprised that biomarker detection performance while using the Z-scores (AUC^Z) for the regression model was similar to using all of the samples, as abundancies are known to be dependent on age (and sex). Plotting the differences between the obtained Z-scores in a Bland–Altman plot show that, on average, no differences are present between the two approaches; for all features as well as the IEM biomarkers (see Appendix J). This explains why the AUC^Z performances are similar for both Z-score approaches, since, for a given Z-score cutoff (used to make the ROC curve), the number of positives is approximately the same, and the same holds when looking at the number of biomarkers detected for a given Z-score cutoff. However, this does not necessarily imply that the regression model is less or equally accurate in determining abberations. We anticipate that the performances for *Regression* would outperform *All samples* when more age-dependent IEM biomarkers were included. Additionally, when judging biomarker detection using the *p*-values, we did see that *Regression* slightly outperformed *All samples*.

4. Materials and Methods

4.1. Untargeted Metabolomics Datasets

While using UHPLC-Orbitrap-MS, human plasma samples of 261 control samples and 58 IEM patients were measured over nine batches over the period 10 December

2018 to 10 January 2020 [5], having, in total, 35 unique IEMs. In agreement with national legislation and institutional guidelines, all patients or their guardians approved the possible anonymous use of the remainder of their samples for method validation and research purposes. The study was conducted in accordance with the Declaration of Helsinki. For every patient, a technical triplicate was included, which allows for obtaining more certainty regarding the measured abundancy (or Z-score) by looking at the spread in the triplicate. A QC (Quality Control) sample was included in all nine batches and 5–9 technical replicates were present in every batch. Because the QC sample was a commercial sample, the sample differed in the concentration of several metabolites when compared to the (average) concentrations of the human plasma samples that were analyzed in these datasets. Features were annotated, as described in Bonte et al. [5]. For eight batches, 18–40 normal controls have been measured (no triplicate) to ensure more accurate reference values. These controls were random selected samples in terms of age and sex, and none of the samples (patients and controls) were measured in more than one batch. One batch included 22 triplicates measurements (plus QC samples) and no control samples. In this study, we will refer to 'feature' as being either a single m/z-value (with unique retention time) or a merge of multiple features, where the adduct type and/or isotope was determined with corresponding neutral mass and, consequently, merged to a single feature.

The following internal standards have been added to each batch in order to facilitate normalization that is based on these internal standards: 1,3-^{15}N uracil (+/−) [300 µmol/L], 5-bromotryptophan (+/−) [85 µmol/L], D_{10}-isoleucine (+/−) [500 µmol/L], D_3-carnitine (+/−) [285 µmol/L], D_4-tyrosine (+/−) [230 µmol/L], D_5- phenylalanine (+/−) [600 µmol/L], D_6-ornithine (+) [225 µmol/L], dimethyl-3,3-glutaric acid (+/−) [300 µmol/L], ^{13}C-thymidine (+/−) [300 µmol/L], D_4-glycochenodeoxycholic acid (−) [44 µmol/L], where + indicates positive ion mode, and—indicates the negative ion mode.

4.2. Data Processing

Pre-processing steps (alignment, peak picking etc.) were performed per batch while using Progenesis QI v2.4 (Newcastle-upon-Tyne, UK) [5]. In-house software was developed in order to match features from each batch to a reference batch, which, in this case, was the fifth batch when sorting on chronologically order. Chromatograms between batches were initially aligned to the reference batch by using lowess regression, where the features were matched based on retention time difference, m/z-value, and median abundancy difference similar to the criteria described below.

Matching features was performed based on several criteria:

1. When features were annotated in reference batch and the batch being merged, these features were pooled to the merged dataset.
2. When MS/MS spectra were present for a potential matching pair of features, the cosine similarity metric was calculated and it had to be >0.8.
3. The retention time difference in percentage was calculated between potential matches, and it had to be $<3\%$.
4. Progenesis QI determined per feature an isotope distribution and we required sufficient overlap of these distributions between potential matching pairs. This was determined by calculating a difference in the percentage between each bin of this distribution. The maximum difference of these bins had to be $<25\%$.
5. Despite the batch effects and potential biological differences between batches, we expected the within-batch median of the (raw) abundancies for matching features to be at least similar. We calculated the differences between these medians in percentages, and required that this difference was $<300\%$.
6. When neutral masses were known for the matching pair, but not the MS/MS spectra, the ppm-error had to be <1.
7. When m/z-values were known for the matching pair, but not the MS/MS spectra and neutral masses, the ppm-error of between the m/z-values had to be <1.

When a feature from the reference batch had two or more (potential) matches with the batch being merged, we decided to exclude these matches, since it was not clear which match would be the correct one. Similarly, when a feature from the batch being merged had more than one match with the reference batch, this feature would also be excluded. The resulting merged dataset only contained features that were matched (i.e., fulfilling the above matching criteria) across all nine batches. Consequently, this led to the loss of circa 98% of the number of features that were normally detected within the batch.

4.3. Quantitative Evaluation Set

For the evaluation of the normalization methods, the following 15 metabolites were quantitatively (µmol/L) measured in two separate assays: leucine (+), C0 L-carnitine (+/−), methionine (+/−), C2 acetylcarnitine (+), 5-aminolevulinic acid/4-hydroxyproline (+), citrulline (+/−), aspartic acid (−), glutamine (+/−), (allo)isoleucine (+/−), proline (+), tyrosine (+), phenylalanine (+/−), taurine (+/−), asparagine (+/−), and arginine (+/−). Amino acids were determined by ion-exchange chromatography according to protocols described by the manufacturer (Biochrom). Free carnitine and acylcarnitines analysis was performed, as described by Vreken et al. [14].

4.4. Normalization Methods

4.4.1. Initial Transformations

Prior to normalization, raw abundancies were, for some methods, transformed while using a log-transform or Box–Cox transformation given by $\hat{y} = ((y + \lambda_2)^{\lambda_1} - 1)/\lambda_1$ with $\lambda_1 = 0.5$ and $\lambda_2 = 1$. If an initial transformation was applied, this was indicated in the name of the (normalization) method, where 'BC-' refers to the Box–Cox transformation and 'Log-' to the log transformation. When no transformation was performed, this was indicated with 'None-'.

4.4.2. Normalization by Metchalizer

Metchalizer assumes a linear mixed effect relationship between the abundancies of the internal standards and the feature of interest. Because the internal standards were expected to be correlated, we represented them by an orthogonal set of covariates. These covariates are obtained as the Latent Variables (LV) from the Partial Least Squares (PLS) of the set of internal standard abundancies (represented in matrix **X**) and the (categorical) information regarding which sample belonged to which batch (represented by matrix **Y**). The number of LV's were chosen from the metric $I(K)$:

$$I(K) = \sum_{k=1}^{K} \sum_{b,i} \left(x_{ib}^{\mathrm{LV}_k} - \bar{x}_{ib}^{\mathrm{LV}_k} \right)^2 \quad (1)$$

where $\bar{x}_{ib}^{\mathrm{LV}_k}$ is the center of batch b in the direction of LV_k. We selected that K, for which $I(K)$ reached 75% of its maximum value.

The mixed effect model then considers the LV's as fixed effects and all variations not explained by the LV's are considered as (random) batch effects:

$$\hat{y}_{ijb} = \beta_j^0 + \sum_{k=1}^{\text{selected } K} \beta_j^k x_i^{\mathrm{LV}_k} + \gamma_{jb} + \epsilon_{ijb} \quad (2)$$

with $\hat{y}_{ijb}$ being the estimated abundancy for feature j and sample i in batch b. $x_i^{\mathrm{LV}_k}$ indicates the covariate (score) of the kth Latent Variable (LV) of sample i. γ_{jb} is the (random) batch intercept for feature j. Note that, when the LV's are sufficient in explaining y_{ijb}, the random intercept γ_{jb} will not contribute much. Before fitting the model, we removed the outlier samples per batch b and feature j based on their within-batch Z-score ($|Z| > 2$) determined from all samples in that batch. Note that these Z-scores are different from the Z-scores that are defined in other parts of this study.

The batch corrected abundancy were given by:

$$y_{ijb}^{\text{batch corrected}} = y_{ijb} - \hat{y}_{ijb} + \text{Median}(\hat{y}_{.jb}) \quad (3)$$

4.4.3. Normalization by Best Correlated IS

The internal standard, m, which best correlates with a feature j is being used to normalize the abundancy of feature j. The correlation was measured within each batch while using the Spearman correlation between feature j and each internal standard individually across all samples and subsequently averaged across all nine batches. The internal standard that (positively) correlated the best was used for normalization according:

$$\hat{y}_{ij} = \frac{y_{ij}}{y_{im}} \text{Median}(y_{.m}) \quad (4)$$

with m being the best correlated internal standard.

4.4.4. Normalization Methods from Literature

The following normalization methods were used in this study:

Anchor [6]: *Anchor* assumes a linear response between the features in the anchor samples and samples in the batch. An anchor sample is a fixed sample, which is analyzed in all nine batches, and it was included more than four times in each batch. Normalization was performed per batch by dividing each feature by the average of the anchor samples for that same feature per batch. In this study, we used our QC samples as the anchor samples.
CCMN [15]: we used function normFit from the crmn R package with input argument `'crmn'`. As a design matrix, we chose QC samples versus human plasma's.
EigenMS [16]: QC samples and human plasma samples were treated as two different groups.
Fast Cyclic Loess [17]: we used the `normalize CyclicLoess` function from the limma R package while using the method `'fast'` and `iterations=100`.
NOMIS [18]: we used the function normFit from the crmn R package with input argument `'nomis'`.
PQN [19]: *PQN* was implemented, as described by Filzmoser et al. The reference spectrum was given by the median of every feature j.
RUV [20]: we used the function `RUVRand` from the MetNorm R package.
VSN [21]: we used the vsn R package while using the `vsn2` function.

Some settings were optimized; the reader is referred to Section 4.4.6 for more details.

4.4.5. Evaluation of Normalization Methods

Six metrics were used in order to evaluate the performance of normalization methods.
WTR score: the WTR score (Within variance Total variance Ratio) calculates the ratio between the 'overall' within-batch variance and the total variance from the QC samples:

$$\text{WTR}_j = \frac{\sigma^2_{j,\text{within}}}{\sigma^2_{j,\text{tot}}} = \frac{\sigma^2_{j,\text{tot}} - \sigma^2_{j,\text{between}}}{\sigma^2_{j,\text{tot}}} \quad (5)$$

where $\sigma_{j,\text{between}}$ is the variance of all nine batch averages for metabolite j in the QC samples, and $\sigma_{j,\text{tot}}$ the 'overall' variance based on all QC samples. The WTR score is between 0 and 1. Because we would like batch averages to be similar for the QC samples (resulting in $\sigma_{j,\text{between}}$ approaching zero), we are interested in WTR scores close to one. Note that the coefficient of variation (CV) was considered to be an inadequate metric, as a simple log-transformation of the data already results in a decreased CV. Because the WTR score considers a ratio between two standard deviations, this metric is less sensitive to such initial data transformations.

QC correlations: for all QC samples, the Spearman correlations were calculated on the

(normalized) abundancies. Normalization should increase the resemblance of the QC samples among each other, therefore increasing the Spearman correlations. It is expected that the Spearman correlations decreases when variations other then techical variation are removed.

Spearman score: for the set of 15 quantitatively measured metabolites, we calculated the Spearman correlation between their quantitative measurements and the normalized abundancies. The overall normalization performance could be judged based on the median Spearman score of these 15 scores, having scores $\in [-1, 1]$. Higher values indicate better resemblance with the quantitative measurements.

R^2 **score**: the R^2 between the quantitative measurements and the normalized abundancies of the 15 quantitatively measured metabolites. The overall performance could be judged from the median R^2 score, with scores of $\in [0, 1]$. Higher values indicate better (linear) fits with the quantitative measurements.

QC prediction score: since the QC samples were different from the human plasma samples in terms of concentrations for several metabolites/features, we expect this difference to be observed in the first few principal components (PCs) of a Principal Component Analysis (PCA) analysis applied to all features (excl. standards). We fitted a logistic function while using the first four PCs as covariates and with class labels: 'human plasma' and 'QC'. The fitted model returns per sample a probability of belonging either to the class 'human plasma' or 'QC'. The probabilities for all samples are averaged into the QC prediction score. Increasing normalization performances should result in higher scores, as QC- and human plasma samples should be segregated. We used LogisticRegression from the Python package scikitlearn with parameters `penalty='l1', solver='saga', multi_class='auto', and max_iter=10,000` [22].

Batch prediction score: increasing normalization performances should result in less batch clustering when examining the first few PCs of the PCA analysis (see QC prediction score). We fitted a logistic function for each batch versus all other eight batches while using the first four PCs as covariates and obtained the probability scores for all human plasma's having the correct batch label. These scores were than averaged for all human plasma samples into a batch prediction scores $\in [0, 1]$. Scores that are closer to 1 indicate decreased normalization performances, since batch separation is (still) present.

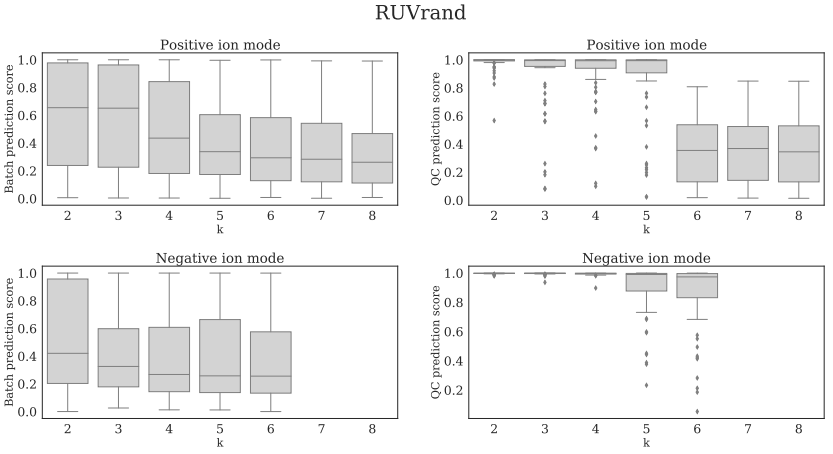
4.4.6. Settings for Normalization Methods from Literature

Based on the *Batch prediction score* and *QC prediction score*, we optimized the parameter settings for the following normalization methods: *CCMN* `ncomp = 8` for both ion modi, *EigenMS* `eigentrends = 6` for positive ion mode and `eigentrends = 4` for negative ion mode, *RUVrand* `k = 5` for positive ion mode and `k = 4` for negative ion mode and *Fast Cyclic Loess* `span = 0.6` for positive ion mode and `span = 0.4` for negative ion mode. The reader is referred to Figure 5a–d for clarification of these choices.

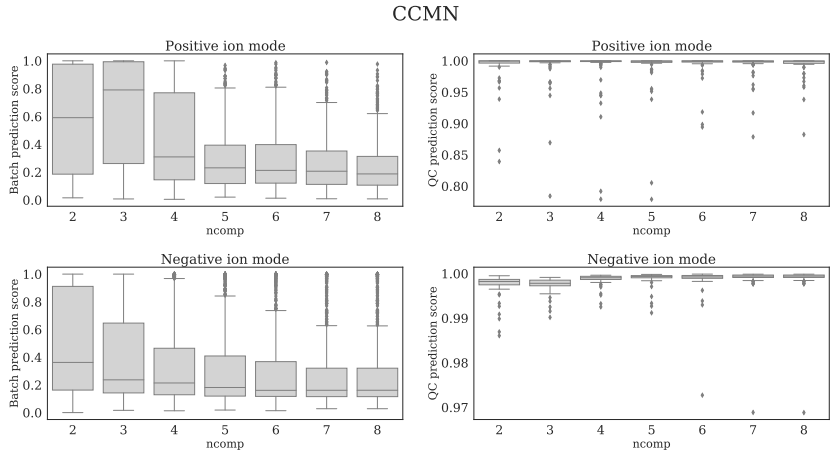
4.5. Methods to Determine Metabolic Abberations

4.5.1. Outlier Removal

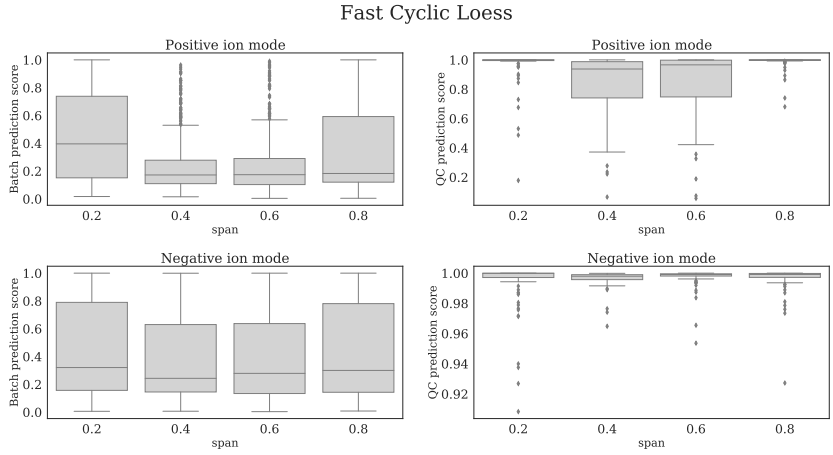
In this study, we used reference samples (controls and patient) to calculated Z-scores. In order to prevent outlier samples (for a given feature/metabolite) to affect the accuracy of the Z-score, we used an iterative procedure to remove outliers before determining the set of samples used for calculating the Z-score. In this procedure, an 'outlier Z-score' was determined based on all of the samples (which samples were taken depends on the given Z-score method, see below), where samples having a $|\text{Z-score}| > 3$ were removed. This was repeated five times and, from the non-outlier samples, a selection was made, depending on the selected Z-score method i.e., *15in*, *15out*, *All samples*, and *Regression* (see below).



(**a**)

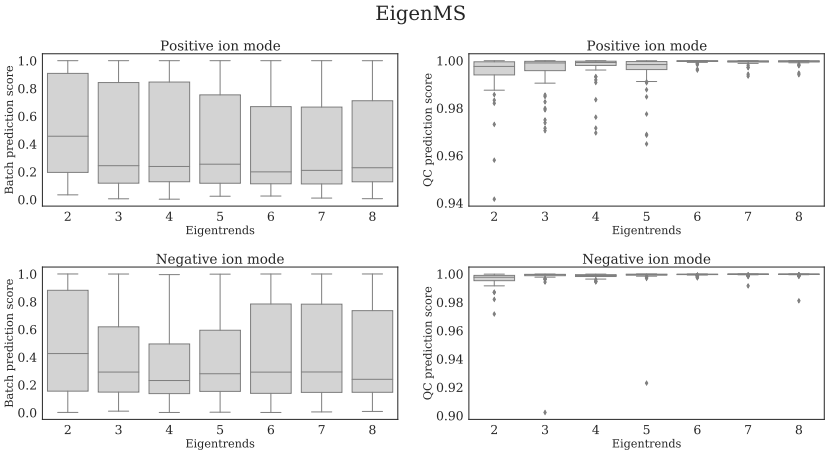


(**b**)



(**c**)

Figure 5. *Cont*.



(**d**)

Figure 5. Each subfigure shows 4 panels which belong to the normalization method stated in the title. The upper, lower panels are the *batch prediction scores* and *QC prediction scores* (vertical axes) for various choices of the parameter (horizontal axis) for positive, negative ion mode, respectively. (**a**) *RUVrand*, (**b**) *CCMN*, (**c**) *Fast Cyclic Loess*, (**d**) *EigenMS*.

4.5.2. Z-Score Methods

Four different methods were used to determine Z-scores.

15in, best matching samples within batch: the Z-scores were calculated by selecting 15 samples originating from the same batch that were matched with the patient based on age and sex, as described in Bonte et al. [5].
15out, best matching samples from other batches: the Z-scores were calculated similarly as in method *15in* while using 15 out-of-batch samples. Note that since there are more out-of-batch samples than within-batch samples the age and sex matching can be done more accurate for *15out* than for *15in*.
All samples: this method used all available reference samples from all nine batches, including within-batch controls, for Z-score calculation, thereby ignoring age- and sex matching.
Regression: we fitted a linear model on all available reference samples excluding outliers that were first removed based on a within-batch $|\text{Z-score}| > 3$. This Z-score is different from other Z-scores mentioned in this study, and it is only used to remove outliers. The regression model is given by:

$$\begin{aligned} \hat{y}_i &= \hat{\beta}^{\text{Intercept}} + \hat{\beta}^{\text{Sex}}\, x_i^{\text{Sex}} + \hat{\beta}^{\text{Sex,Age}}\, x_i^{\text{Sex}} x_i^{\text{Age}} \\ &\quad + \sum_{p=1}^{P} \hat{\beta}_p^{\text{Age}}\, (x_i^{\text{Age}})^p + \hat{\epsilon}_i \\ \hat{y}_i &= \vec{x}_i^T \vec{\hat{\beta}} + \hat{\epsilon}_i \end{aligned} \tag{6}$$

where $\hat{y}_i$ is the predicted (normalized) abundancy of feature j for sample i, $\hat{\beta}^{\text{Intercept}}$ is an intercept. $\hat{\beta}^{\text{Sex}}$, $\hat{\beta}^{\text{Sex,Age}}$ (interaction) and $\hat{\beta}_p^{\text{Age}}$ indicate slopes. P is the degree of the polynomial used for regression on age and set to $P = 3$ in this study. x_i^{Sex} is 1 for women and 0 for men. $\hat{\epsilon}_i$ is the estimated error. The latter expression is the model in vector notation with $\vec{x}_i^T = [1, x_i^{\text{Sex}}, ..., (x_i^{\text{Age}})^P]$.

The coefficients were determined from the OLS estimator:

$$\vec{\hat{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{y} \tag{7}$$

where the rows of $\mathbf{X}$ are given by $\vec{x}_i^T$. The variance in $\hat{y}_i$ is determined by the variance in $\vec{\hat{\beta}}$ and the variance in $\hat{\epsilon}_i$:

$$\begin{aligned} \mathrm{Var}[\hat{y}_i] &= \mathrm{Var}\left[\vec{x}_i^T \vec{\hat{\beta}}\right] + \mathrm{Var}[\hat{\epsilon}_i] \\ &= \vec{x}_i^T \mathrm{Cov}[\vec{\hat{\beta}}]\vec{x}_i + \hat{\sigma}_i^2 \end{aligned} \tag{8}$$

The covariance matrix of $\vec{\hat{\beta}}$ is given by:

$$\begin{aligned} \mathrm{Cov}[\vec{\hat{\beta}}] &= \mathrm{Cov}[\beta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{\epsilon}\,] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\,\mathrm{E}[\vec{\epsilon}\,\vec{\epsilon}^T]\,\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned} \tag{9}$$

We estimated $\mathrm{E}[\vec{\epsilon}\,\vec{\epsilon}^T]$ by:

$$\mathrm{E}[\vec{\epsilon}\,\vec{\epsilon}^T] = \begin{bmatrix} \hat{\sigma}_1^2 & 0 & \dots & 0 \\ 0 & \hat{\sigma}_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \hat{\sigma}_N^2 \end{bmatrix} \tag{10}$$

Because we expected σ_i^2 to be dependent on age (neglecting sex), we estimated $\hat{\sigma}_i^2$ from a weighted mean on the squared residuals:

$$\begin{aligned} \hat{\sigma}_i^2 &= \sum_{k=1}^{N} \frac{w_k(x_i^{\mathrm{Age}})}{\sum_{k'=1}^{N} w_{k'}(x_i^{\mathrm{Age}})}(y_k - \hat{y}_k)^2 \\ w_k(x_i^{\mathrm{Age}}) &= \exp\left(-\frac{|x_i^{\mathrm{Age}} - x_k^{\mathrm{Age}}|}{a + bx_i^{\mathrm{Age}}}\right) \end{aligned} \tag{11}$$

where a and b determine how the weights decay (a) or increase (b) over age (we set $a, b = 1$ years). The Z-scores were obtained by subtracting the predicted average $\hat{y}_i$ and dividing by the variance $\mathrm{Var}[\hat{y}_i]$ (Equation (8)).

The significance of the regression coefficients (Equation (6)) was obtained by considering the statistic:

$$\frac{(\hat{\beta}_i - \beta_i)}{\sqrt{\mathrm{Var}[\hat{\beta}_i]}} \sim \mathcal{N}(0,1) \tag{12}$$

The variances of the coefficients were found in the diagonal elements of $\mathrm{Cov}[\vec{\hat{\beta}}]$ (Equation (9)). We tested the hypotheses that $\beta_i = 0$ with a two-tailed test. A robust p-value was obtained from a bootstrap procedure by taking the median p-value from a series of p-values that were obtained from 50 bootstraps on the above test statistics taking 90% of the data each bootstrap.

4.5.3. Final Z-Scores

Because the patient samples were measured in triplicate, we determined the final Z-scores from the average of these three Z-scores [5]. These average Z-score were determined for all Z-score methods i.e., *15in*, *15out*, *All samples*, and *Regression*.

4.5.4. p-Values from Welch's T-Test

As an alternative to using the (average) Z-scores, we also considered the p-values that were obtained from the Welch's t-test, as it indicates whether the mean of triplicates differs significantly from the population average. Note that the triplicate was expected to only have technical variance, whereas the reference population has variance that consists

of technical- plus biological variance. For each Z-score method (*15in*, *15out*, *All samples*, and *Regression*), these p-values were obtained per feature (and patient).

When using the regression model, we used an adjusted Welch's t-test assuming that the variance in the estimate of the average of the population (which is $Z = 0$) was negligible:

$$t_j = \frac{\text{Mean}(Z_{j.})}{\sqrt{\frac{s_j^2}{3}}} \quad (13)$$

where s_j is the sample standard deviation of the triplicate Z-scores, $\text{Mean}(Z_{j.})$ indicates the average of the triplicate for feature j.

4.6. Detection of the Expected IEM Biomarkers

In order to explore how normalization and the method of determining these Z-scores (*15in*, *15out*, *All samples*, and *Regression*) affected the detection of the expected biomarkers, we plotted the number of abberant biomarkers of the known IEM patients against the average number of abberant features (true plus false positives) per patients for various (final) Z-score and p-value cutoff levels, similar to a ROC curve. Improved biomarker detection was believed to increase the area under the curve (AUC).

Establishing this curve was done by assigning a status for every biomarker (if present and annotated in the MS-data). A database was established containing the expected biomarkers for each IEM, including the expected Z-score sign (up or down regulated), as can be found in Appendix D Table A6. For every IEM patient, we assigned, for all expected biomarkers, the status 'positive' or 'negative'. The status 'positive' was assigned when (1) $|\text{Z-score}| > Z_{\text{abnormal}}$ and (2) the sign of the Z-score corresponded with the expected sign for that biomarker in the IEM patient. When creating this curve while using the p-values, we also required that the sign of the Z-score corresponded with the expected sign for that biomarker, and similarly assigned the 'positive' status when $p\text{-value} < p_{\text{abnormal}}$. When a biomarker was found in both positive and negative ion mode, the Z-score(s) from the mode having the largest population average abundancy was taken. The average number of detected features (per patient) was obtained by considering the features from both ion modes.

Because some biomarkers are only found in a single IEM patient and not in reference samples (or other IEM patients), some of the expected IEM biomarkers were not matched across all nine batches and, therefore, were absent in the merged dataset and analysis in this study. In the merged dataset, we obtained 195 patient-biomarker combinations (one patient could have multiple biomarkers) that were associated with 49 patients.

5. Conclusions

In conclusion, out of all explored normalization methods, the removal of batch effects was best performed by *Log-Metchalizer*. Fitting our regression model on the corresponding normalized data showed that 6.5–37% (Table 1) of all considered features were dependent on age, underlining the need for using age corrected Z-scores. On average, biomarker detection performance using *Log-Metchalizer* using out-of-batch controls was at least similar to the best performing *Log-Raw* approach when using the 15 within-batch controls (*15in&Log-Raw*). We anticipate that the success of *Metchalizer* and age- and sex correcting strategies, such as our regression model, depend on three factors: (1) a feature of interest being measured in a number of other batches (not necessarily all), (2) batch effects containing (only) technical variations, and (3) abundancies being affected by age or other covariates and their effect size. In summary, our proposed approach using out-of-batch reference samples opens new opportunities for improving abnormality detection, especially for age-dependent features/biomarkers.

Author Contributions: R.B. performed all the experimental work. M.B. designed the statistical models, the computational framework and analyzed the data. The manuscript was written by M.B., H.J.B., M.J.T.R. and G.J.G.R., S.D. and E.H.J. contributed in establishing the IEM database, and actively contributed in giving feedback on the methods. M.J.T.R. contributed to in-depth reviewing of the manuscript, all analytical methods and suggested adjustments to initial work. E.O., A.T.v.d.P., M.A.E.M.W. and R.M.W.H. provided resources. The research was under supervision of G.J.G.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Erasmus Medical Centre, department Clinical Genetics.

Informed Consent Statement: In agreement with national legislation and institutional guidelines, all patients or their guardians approved the possible anonymous use of the remainder of their samples for method validation and research purposes. The study was conducted in accordance with the Declaration of Helsinki.

Conflicts of Interest: All authors state that they have no conflict of interest to declare. None of the authors accepted any reimbursements, fees, or funds from any organization that may in any way gain or lose financially from the results of this study. The authors have not been employed by such an organization. The authors have not acted as an expert witness on the subject of the study. The authors do not have any other conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AUC	Area under the curve
ROC	Receiver operating characteristic
IEM	Inborn error of metabolism
CV	Coefficient of variation
QC	Quality control
PCA	Principle component analysis
PC	Principle component
UHPLC	Ultra-high performance liquid chromatography

Appendix A. Variations of Standards

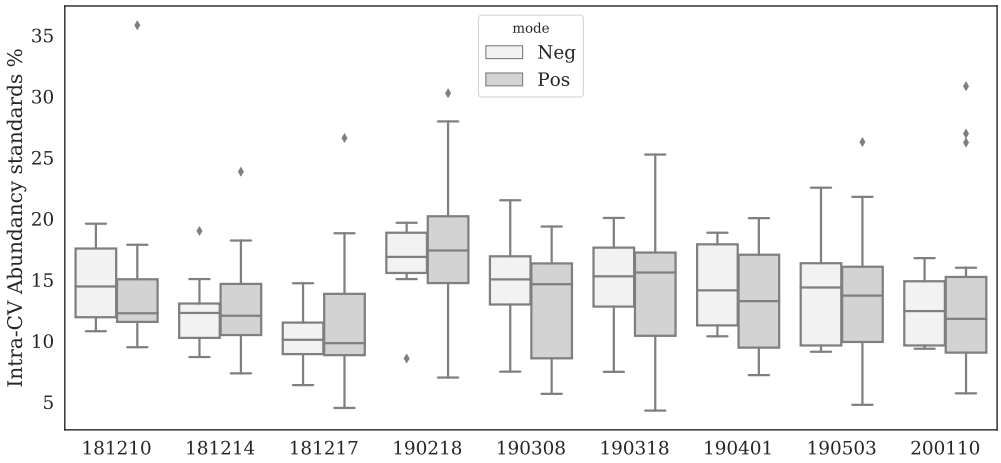


Figure A1. Boxplots per batch and ion mode for intra-batch Coefficients of Variation (CV) for all standards. Area's were determined by using Xcalibur 4.0.

Table A1. Inter-batch Coefficients of Variation per standard based on 9 batches. Note, that some components were detected in both ion modi but the best CV was considered for quality control. Peak areas were determined using Xcalibur 4.0.

Standard	Ion Mode	Inter-Batch CV (%)
3,3-dimethylglutaric acid	−	23
3,3-dimethylglutaric acid	+	72
5-bromotryptophane	−	18
5-bromotryptophane	+	21
Acetylcarnitine D2	+	31
Carnitine D3	+	23
Glycochenodeoxycholic Acid D4	−	20
Hexadecanoylcarnitine D3	+	31
Hexanoylcarnitine D3	+	28
Isoleucine D10	−	40
Isoleucine D10	+	27
Methylmalonic Acid D3	−	52
Ornithine D6	+	26
Phenylalanine D5	−	35
Phenylalanine D5	+	29
Tetradecanoylcarnitine D3	+	27
Thymidine 13C	−	18
Thymidine 13C	+	51
Tyrosine D4	−	27
Tyrosine D4	+	27
Uracil-1,3N15	−	18
Uracil-1,3N15	+	19
Uridine D2	−	27
Uridine D2	+	58
Valine D8	+	41

Appendix B. Batch Effect Removal Performances

Table A2. Median scores for six metrics per normalization method and ion mode. The last column is the average of the former columns. Note, that we modified the *batch prediction score* such that higher values correspond with improved normalization performances.

Method	Ion Mode	QC Prediction Score	R^2 Score	Spearman Score	WTR Score	QC Correlations	1—Batch Prediction Score	Mean
None-Anchor	+	1.00	0.63	0.75	1.00	0.98	0.83	0.77
None-Anchor	−	1.00	0.57	0.74	1.00	0.99	0.83	0.76
Log-Metchalizer	−	1.00	0.68	0.83	0.61	0.97	0.88	0.73
Log-Metchalizer	+	1.00	0.61	0.78	0.72	0.97	0.88	0.73
BC-Metchalizer	+	1.00	0.66	0.78	0.67	0.97	0.88	0.73
BC-Metchalizer	−	1.00	0.64	0.79	0.58	0.97	0.89	0.71
Log-CCMN	−	1.00	0.56	0.74	0.65	0.97	0.84	0.70
Log-NOMIS	−	0.88	0.52	0.72	0.66	0.96	0.83	0.68
Log-EigenMS	+	1.00	0.56	0.73	0.48	0.97	0.80	0.68
Log-CCMN	+	1.00	0.49	0.72	0.57	0.97	0.81	0.68
None-PQN	+	1.00	0.58	0.74	0.35	0.95	0.51	0.66
None-Best correlated IS	+	1.00	0.58	0.74	0.34	0.95	0.32	0.66
None-VSN	+	1.00	0.49	0.73	0.35	0.95	0.57	0.65
Log-EigenMS	−	1.00	0.40	0.63	0.42	0.97	0.77	0.63

Table A2. *Cont.*

Method	Ion Mode	QC Prediction Score	R^2 Score	Spearman Score	WTR Score	QC Correlations	1—Batch Prediction Score	Mean
Log-RUVrand	−	1.00	0.51	0.73	0.43	0.69	0.73	0.62
None-PQN	−	1.00	0.42	0.64	0.28	0.95	0.65	0.61
Log-Raw	−	1.00	0.40	0.65	0.24	0.95	0.67	0.61
Raw	−	1.00	0.40	0.64	0.22	0.95	0.71	0.60
None-VSN	−	1.00	0.34	0.61	0.31	0.95	0.67	0.60
Raw	+	1.00	0.37	0.55	0.25	0.95	0.62	0.59
Log-Raw	+	1.00	0.37	0.55	0.24	0.95	0.61	0.59
Log-RUVrand	+	0.99	0.47	0.68	0.46	0.51	0.66	0.59
Log-Fast Cyclic Loess	+	0.97	0.22	0.41	0.54	0.92	0.83	0.58
Log-NOMIS	+	0.33	0.48	0.69	0.59	0.94	0.80	0.58
None-Best correlated IS	−	1.00	0.18	0.43	0.39	0.95	0.47	0.56
Log-Fast Cyclic Loess	−	1.00	0.02	0.16	0.29	0.76	0.76	0.46

Appendix C. Regression Analysis

Table A3. Significant age related metabolites (corrected p-value $p_1^{Age} < 0.05$) for BC-Metchalizer normalized data. Slashes in the names indicate that chromatographic separation of isomers was not possible. The column "Sign" indicates if the sign of coefficient p_1^{Age} is positive (up) or negative (down).

Metabolite	Ion Mode	Sign	$p^{\text{Intercept}}$	p_1^{Age}	p_2^{Age}	p_3^{Age}	p^{Sex}	$p^{\text{Sex,Age}}$
1-Methyladenosine (1)	+	down	0	5.8×10^{-5}	7.1×10^{-3}	1.8×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
2-Aminoadipic acid	+	down	0	7.3×10^{-5}	9.8×10^{-3}	1.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
2-Aminoadipic acid	-	down	0	3.0×10^{-5}	1.1×10^{-3}	1.9×10^{-2}	8.6×10^{-1}	5.4×10^{-1}
2-Ketoglutaric acid	-	down	0	4.5×10^{-4}	1.9×10^{-2}	1.6×10^{-1}	8.6×10^{-1}	5.2×10^{-1}
3-Methoxytyrosine	+	down	0	4.1×10^{-9}	4.3×10^{-3}	2.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
3-Methylhistidine/1-Methylhistidine	+	up	0	8.1×10^{-4}	2.4×10^{-1}	5.1×10^{-1}	8.6×10^{-1}	7.8×10^{-1}
4-Pyridoxic acid	-	down	0	7.1×10^{-3}	1.6×10^{-1}	4.2×10^{-1}	8.6×10^{-1}	6.4×10^{-1}
Acetoacetic acid/Succinic acid semialdehyde	+	down	0	2.1×10^{-3}	2.5×10^{-1}	6.8×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Adenine	+	down	0	1.1×10^{-2}	1.1×10^{-1}	3.3×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
C2 Acetylcarnitine	+	down	0	1.2×10^{-2}	2.9×10^{-1}	6.4×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
C3 Propionylcarnitine	+	down	0	7.7×10^{-3}	1.4×10^{-1}	5.2×10^{-1}	8.6×10^{-1}	8.3×10^{-1}
C8:1 Octenoylcarnitine	+	down	0	4.8×10^{-2}	3.7×10^{-1}	6.9×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Chenodeoxycholic acid	-	up	0	2.0×10^{-6}	1.7×10^{-1}	6.2×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
Cholesterol	+	up	0	3.2×10^{-2}	2.9×10^{-1}	5.4×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Citrulline	-	up	0	3.7×10^{-3}	7.0×10^{-2}	1.5×10^{-1}	8.6×10^{-1}	2.9×10^{-1}
Creatine	-	down	0	1.3×10^{-4}	3.7×10^{-1}	8.6×10^{-1}	8.6×10^{-1}	5.4×10^{-1}
Creatine	+	down	0	6.0×10^{-5}	4.2×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	8.3×10^{-1}
Creatinine	+	up	0	5.1×10^{-14}	4.2×10^{-3}	2.7×10^{-1}	8.6×10^{-1}	7.8×10^{-1}
Dehydro-epiandrosteronsulfaat (dheas)	-	up	4.4×10^{-8}	6.0×10^{-15}	6.2×10^{-3}	5.2×10^{-1}	8.6×10^{-1}	5.4×10^{-1}

Table A3. *Cont.*

Metabolite	Ion Mode	Sign	$p^{\text{Intercept}}$	p_1^{Age}	p_2^{Age}	p_3^{Age}	p^{Sex}	$p^{\text{Sex,Age}}$
Dihydroxycholanoic acid + gly	-	down	0	1.1×10^{-2}	3.7×10^{-1}	7.6×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
Dimethylarginine (sdma + adma)	+	down	0	1.9×10^{-6}	3.3×10^{-2}	4.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Dimethylarginine (sdma + adma)	-	down	0	1.2×10^{-5}	8.2×10^{-3}	1.4×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
Glycocholic acid	+	down	0	7.7×10^{-3}	4.6×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Glycocholic acid	-	down	0	2.0×10^{-3}	3.5×10^{-1}	8.4×10^{-1}	8.6×10^{-1}	8.0×10^{-1}
Guanidinoacetic acid	+	up	0	3.1×10^{-3}	6.2×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	4.9×10^{-2}
Histidine	-	up	0	1.3×10^{-2}	7.1×10^{-2}	2.6×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
Homoarginine	+	up	0	2.1×10^{-3}	2.0×10^{-1}	5.7×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Indoxylsulfuric acid	-	up	0	4.3×10^{-3}	7.2×10^{-2}	2.0×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
Kynurenin	+	down	0	1.8×10^{-5}	2.6×10^{-2}	3.1×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
L-Rhamnose	-	up	0	7.2×10^{-7}	1.8×10^{-5}	4.4×10^{-4}	8.6×10^{-1}	4.8×10^{-1}
Lactic acid	-	down	0	4.8×10^{-2}	7.1×10^{-1}	8.6×10^{-1}	8.6×10^{-1}	8.0×10^{-1}
N-Acetylaspartic acid	-	down	0	7.9×10^{-10}	3.5×10^{-4}	4.7×10^{-2}	8.6×10^{-1}	7.9×10^{-1}
Pantothenic acid	-	down	0	1.7×10^{-12}	3.7×10^{-6}	6.1×10^{-4}	8.6×10^{-1}	8.5×10^{-1}
Pantothenic acid	+	down	0	4.0×10^{-15}	1.2×10^{-8}	3.5×10^{-5}	8.6×10^{-1}	8.6×10^{-1}
Phenylacetic acid	+	up	0	1.8×10^{-2}	1.2×10^{-1}	2.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Pseudouridine	+	down	0	9.4×10^{-4}	1.8×10^{-2}	1.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Raffinose/Hex3	+	down	0	7.2×10^{-6}	4.2×10^{-3}	1.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Raffinose/Hex3	-	down	0	3.0×10^{-6}	1.6×10^{-3}	4.8×10^{-2}	8.6×10^{-1}	8.5×10^{-1}
Ribose/Xylose/ Arabinose	-	down	0	2.3×10^{-7}	1.0×10^{-3}	5.2×10^{-2}	8.6×10^{-1}	8.5×10^{-1}
Sialic acid	-	down	0	4.1×10^{-7}	1.6×10^{-3}	8.6×10^{-2}	8.6×10^{-1}	6.5×10^{-1}
Sialic acid	+	down	0	1.0×10^{-3}	8.2×10^{-2}	4.6×10^{-1}	8.6×10^{-1}	8.6×10^{-1}
Theophylline/ Paraxanthine	+	up	0	8.3×10^{-3}	8.5×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	5.8×10^{-1}
Trihydroxycholanoic acid + tau	-	down	0	5.6×10^{-4}	1.6×10^{-1}	4.4×10^{-1}	8.6×10^{-1}	7.1×10^{-1}
Uric acid	+	up	0	5.1×10^{-3}	5.4×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	5.7×10^{-1}
Uric acid	-	up	0	2.7×10^{-3}	5.1×10^{-1}	8.6×10^{-1}	8.6×10^{-1}	1.9×10^{-1}
Uridine	-	down	0	3.6×10^{-2}	3.3×10^{-1}	6.6×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
cis-Aconitic acid/trans-Aconitic acid	-	down	0	4.1×10^{-3}	6.8×10^{-1}	7.4×10^{-1}	8.6×10^{-1}	8.5×10^{-1}
cis-Aconitic acid/trans-Aconitic acid	+	down	0	1.4×10^{-3}	2.5×10^{-1}	6.9×10^{-1}	8.6×10^{-1}	8.6×10^{-1}

Table A4. Significant (corrected p-value $p^{\text{Sex,Age}} < 0.05$) interaction of age and sex related metabolites for *BC-Metchalizer* normalized data. An "up" in column "Sign" indicates that the metabolite was increased in women and vice versa. Slashes in the names indicate that chromatographic separation was not possible.

Metabolite	Mode	Sign	$p^{\text{Intercept}}$	p_1^{Age}	p_2^{Age}	p_3^{Age}	p^{Sex}	$p^{\text{Sex,Age}}$
Guanidinoacetic acid	+	down	0	3.1×10^{-3}	6.2×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	4.9×10^{-2}
Ornithine	+	down	0	8.4×10^{-1}	7.1×10^{-1}	8.8×10^{-1}	8.6×10^{-1}	4.9×10^{-2}

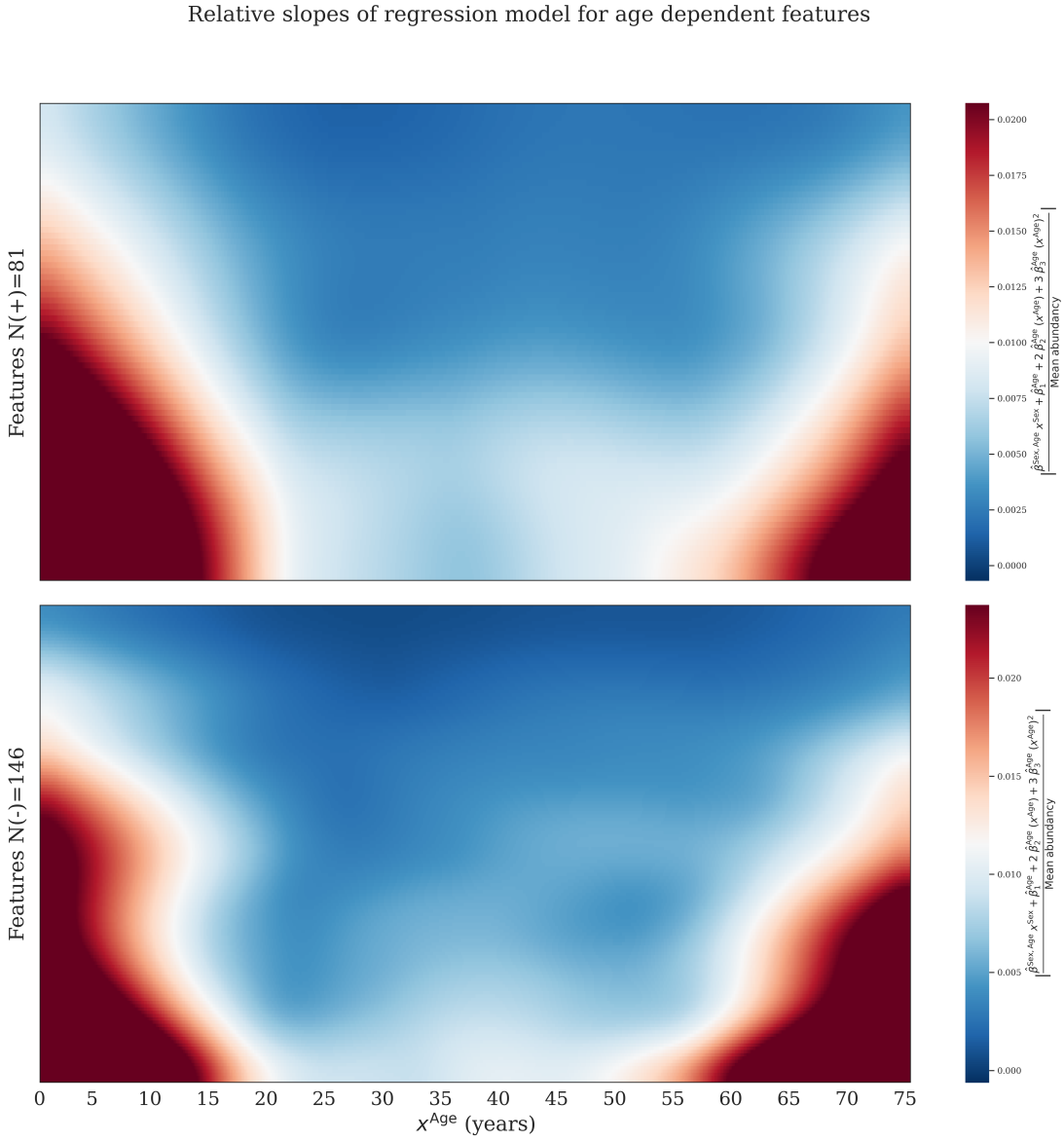


Figure A2. By taking the derivative of the regression model with respect to x_i^{Age} we obtained the slope at a given age. This figure was obtained by plotting these slopes for features where $p_1^{\text{Age}} < 0.05$. Slopes were first divided by the average abundancy to obtain comparable numbers where after the absolute value was taken. Red and orange area's indicates larger (absolute) values of the slopes. Features were ranked based on there median (absolute) slope, and colors were smoothed. The data was normalized using *BC-Metchalizer*. We observe that age dependent features have the tendency to increase/decrease in abundancy faster at younger and older ages, implying that a matching reference population for these age groups are more important.

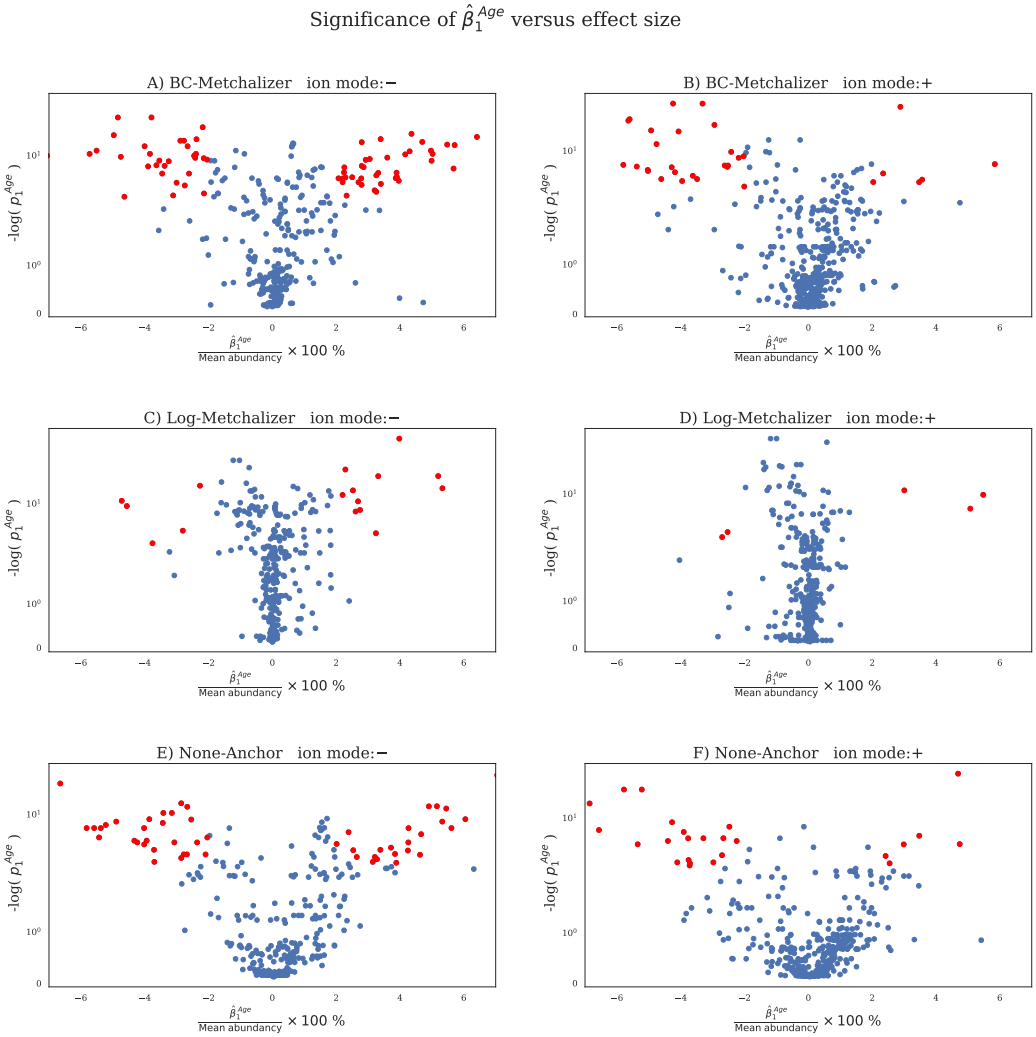


Figure A3. For each normalization method and ion mode the scaled p-value is plotted against the effect size per feature (dot). The effect size is calculated by dividing the coefficient $\hat{\beta}^{\text{Age}_1}$ by the average abundancy of that feature and multiplying the result with 100 %. This effect size measures how the abundancy changes per year (relative to the average abundancy). The red dots indicate features which have a corrected p-value $p^{\text{Age}_1} < 0.05$ and an absolute effect size >2 %. (**A**) *BC-Metchalizer* with negative ion mode data, (**B**) *BC-Metchalizer* with positive ion mode data, (**C**) *Log-Metchalizer* with negative ion mode data, (**D**) *Log-Metchalizer* with positive ion mode data, (**E**) *None-Anchor* with negative ion mode data, (F) *None-Anchor* with positive ion mode data.

Table A5. Per normalization method and ion mode the percentage of features is displayed having a corrected p-value p_1^{Age} <0.05 and effect size >2%. See for explaination of the effect size the caption in Figure A3.

Method	Ion Mode	Percentage
BC-Metchalizer	−	22.03
BC-Metchalizer	+	7.14
Log-Metchalizer	−	5.22
Log-Metchalizer	+	1.08
None-Anchor	−	17.10
None-Anchor	+	5.84

Appendix D. IEM Biomarkers

Table A6. Z-scores obtained from *None-Raw&15in*, *BC-Metchalizer&Regression* and *None-Anchor&Regression* per biomarker and IEM. The stars on the Z-scores indicate that Welch's *t*-test *p*-values < 0.05. The amount of unique patients is indicated by N next to the name of the IEM.

Biomarker (Expected Z-Score Sign) (Ion Mode)	*15in&None-Raw*	*Regression&BC-Metchalizer*	*Regression&None-Anchor*
Aminoacylase I deficiency N = 1			
N-Acetylarginine (up) (+)	0.6	0.0	0.5
N-Acetylglycine (up) (−)	9.3 *	4.2 *	7.4 *
N-Acetylglycine (up) (+)	5.3 *	1.8 *	4.6 *
Argininemia N = 1			
4-Guanidinobutyric acid (up) (+)	28.5 *	11.4 *	27.1 *
Arginine (up) (−)	5.2 *	4.0 *	7.0 *
Arginine (up) (+)	4.1 *	3.1 *	5.3 *
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	0.2	0.0	0.8
Guanidinoacetic acid (down) (+)	3.3 *	2.1 *	2.8 *
Homoarginine (up) (+)	8.7 *	10.7 *	2.8 *
N-Acetylarginine (up) (+)	117.2 *	26.2 *	95.7 *
Uridine (up) (−)	3.0 *	4.1 *	4.1 *
Argininosuccinic aciduria N = 3			
Arginine (down) (−)	−0.9 *, −0.3, 0.8 *	−1.0 *, 0.2, 1.1 *	−1.0 *, 1.0, 0.9 *
Arginine (down) (+)	−1.0 *, −0.0, 1.0 *	−1.2 *, 0.3, 1.4 *	−1.3 *, 1.2 *, 1.3 *
Citrulline (up) (−)	30.3 *, 20.3 *, 12.9 *	9.8 *, 10.2 *, 8.6 *	20.1 *, 17.4 *, 12.0 *
Citrulline (up) (+)	19.9 *, 21.0 *, 11.2 *	7.1 *, 10.8 *, 8.5 *	14.4 *, 16.9 *, 10.6 *
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	1.9 *, 0.8 *, 0.9 *	1.6 *, 1.2, 1.3 *	1.2 *, 1.8 *, 1.7 *
Homocitrulline (up) (+)	12.9 *, 2.1 *, 3.7 *	5.3 *, 1.5 *, 1.9 *	17.5 *, 2.3 *, 4.0 *
Uridine (up) (−)	0.5, −0.4, 5.0 *	0.3, −0.7 *, 4.7 *	0.5 *, −0.6 *, 4.5 *
Beta-ketothiolase deficiency N = 2			
2-Methylacetoacetic acid (up) (+)	0.5, −0.9 *	1.0, 0.7	0.4, 0.5
Carbamoyl Phosphate Synthetase deficiency N = 2			
2-Ketoglutaric acid (up) (−)	−1.2 *, −0.8	−1.5 *, −1.1	−0.8 *, −0.6
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	3.5 *, −0.1	2.5 *, 0.2	4.1 *, 0.3

Table A6. *Cont.*

Biomarker (Expected Z-Score Sign) (Ion Mode)	*15in&None-Raw*	*Regression&BC-Metchalizer*	*Regression&None-Anchor*
Citrullinemia type I N = 1			
Arginine (down) (−)	1.1 *	1.1	0.0
Arginine (down) (+)	1.2 *	1.1 *	0.5
Citrulline (up) (−)	183.0 *	54.3 *	185.0 *
Citrulline (up) (+)	104.9 *	47.8 *	112.5 *
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	−0.5	−1.4	−1.0 *
Uridine (up) (−)	1.4 *	1.5 *	0.7
Glutamate formiminotransferase deficiency N = 1			
Formiminoglutamic acid (up) (+)	1748.8 *	135.6 *	1060.0 *
Glutaric aciduria I N = 2			
C5DC Glutarylcarnitine (up) (+)	360.3 *, 411.3 *	26.9 *, 56.9 *	309.5 *, 529.8 *
Glutaric acid (up) (+)	0.4, 2.3 *	0.3, 2.8 *	1.1, 3.5 *
Glutaric aciduria II N = 2			
Adipic acid (1) (up) (−)	1.5 *, 6.8 *	1.2 *, 5.0 *	1.4 *, 8.2 *
C10 Decanoylcarnitine (up) (+)	114.9 *, 56.5 *	23.5 *, 23.0 *	65.3 *, 52.0 *
C12 Dodecanoylcarnitine (up) (+)	0.9, 66.7 *	0.5 *, 26.0 *	1.2, 27.8 *
C14:1 Tetradecenoylcarnitine (up) (+)	99.4, 68.7 *	20.4 *, 18.5 *	146.7, 75.3 *
C16 Hexadecanoylcarnitine (up) (+)	10.3, 13.0 *	5.3, 5.5 *	13.6, 11.9 *
C16:1 Hexadecenoylcarnitine (up) (+)	200.6, 99.0 *	31.3 *, 23.2 *	188.7, 105.7 *
C18 Octadecanoylcarnitine (up) (+)	11.6, 4.5 *	4.2, 2.1 *	7.4, 5.1 *
C18:2 Linoleoylcarnitine (up) (+)	15.4, 27.4 *	8.5, 8.6 *	33.8, 20.9 *
C4 Butyrylcarnitine (up) (+)	33.5 *, 134.5 *	10.9 *, 35.1 *	28.6 *, 182.3 *
C5 Isovalerylcarnitine (up) (+)	362.7 *, 31.7 *	33.3 *, 17.4 *	172.3 *, 31.4 *
C5DC Glutarylcarnitine (up) (+)	27.7 *, 91.8 *	5.8 *, 24.3 *	25.0 *, 127.0 *
C6 Hexanoylcarnitine (up) (+)	112.9 *, 173.5 *	28.2 *, 47.1 *	92.3 *, 200.3 *
C8 Octanoylcarnitine (up) (+)	154.3 *, 32.1 *	24.6 *, 25.2 *	114.4 *, 38.4 *
Glutaric acid (up) (+)	1.9 *, 1.1	1.0 *, 1.3	2.2 *, 1.8
C18:1 Oleoylcarnitine (up) (+)	16.3, 8.1	5.0, 3.9 *	17.0, 7.0
Homocystinuria N = 3			
Homocysteine (up) (+)	2.5 *, 1.8 *, 0.7 *	2.8 *, 1.2, 0.6 *	1.5, 1.5, 0.8 *
Methionine + Methioninesulfoxide (up) (+)	7.4 *, 8.5 *, 46.9 *	8.8 *, 5.2 *, 16.4 *	8.1 *, 5.6 *, 61.3 *

Table A6. *Cont.*

Biomarker (Expected Z-Score Sign) (Ion Mode)	*15in&None-Raw*	*Regression&BC-Metchalizer*	*Regression&None-Anchor*
Isovaleric acidemia N = 1			
C5 Isovalerylcarnitine (up) (+)	62.6 *	34.7 *	39.0 *
L−2-Hydroxyglutaric aciduria N = 1			
Lysine (up) (+)	5.4 *	3.5 *	5.6 *
Long-chain−3-hydroxyacyl CoA dehydrogenase deficiency N = 2			
3-Hydroxydecanedioic acid (up) (−)	1.2, 1.4 *	1.6, 1.6 *	0.2, −0.1
Adipic acid (1) (up) (−)	0.8, 0.6	0.8, 0.6	0.4, −0.0
Sebacic acid (up) (−)	2.4, 1.2 *	3.8, 1.2	1.4, −0.2
Sebacic acid (up) (+)	2.9, 1.6 *	4.0, 1.3	1.1, −0.3
Suberic acid (up) (−)	1.0, 1.0 *	1.1, 0.8	0.2, −0.3
Suberic acid (up) (+)	1.7, 1.6 *	1.3, 0.9	0.4, −0.1
Lysinuric protein intolerance N = 2			
Arginine (down) (−)	−1.2 *, −1.6 *	−1.1 *, −1.9 *	−0.8 *, −1.7 *
Arginine (down) (+)	−1.2 *, −1.2 *	−1.1 *, −1.9 *	−1.1 *, −1.4 *
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	1.3 *, 3.2 *	0.8, 6.6 *	0.3, 4.6 *
Lysine (down) (+)	−2.1 *, −2.4 *	−2.1 *, −2.8 *	−2.3 *, −1.9 *
Ornithine (down) (+)	−0.6 *, −1.6 *	−1.8 *, −2.9 *	−1.2 *, −2.2 *
Malonyl-Coa decarboxylase deficiency N = 1			
Malonic acid (up) (−)	−0.0	−0.3	−0.7 *
Maple syrup urine disease N = 2			
(allo)Isoleucine (up) (−)	3.1 *, 20.2 *	3.4 *, 9.7 *	4.3 *, 11.3 *
(allo)Isoleucine (up) (+)	3.6 *, 16.4 *	3.9 *, 9.9 *	5.0 *, 11.7 *
2-Keto−3-methylvaleric acid (up) (−)	−1.2 *, 20.5 *	−1.5 *, 14.4 *	−0.8 *, 19.6 *
2-Keto−4-methylvaleric acid (up) (−)	−1.5 *, 0.2	−2.6 *, 0.2	−1.5 *, 0.6
Leucine (up) (+)	3.8 *, 23.8 *	3.3 *, 13.1 *	4.4 *, 17.5 *
Valine (up) (−)	0.3, 3.6	0.8, 2.8	1.7, 3.5
Valine (up) (+)	2.0 *, 7.5 *	2.1 *, 5.0 *	2.0 *, 3.9 *
Medium Chain Acyl-CoA Dehydrogenase Deficiency N=5			
3-Hydroxydecanedioic acid (up) (−)	2.0 *, 1.9, 1.9, 0.5, −0.3	−0.3, −0.3, −0.1, −0.4, −1.8	0.7, 0.8, 0.7, −0.1, −0.1

Table A6. *Cont.*

Biomarker (Expected Z-Score Sign) (Ion Mode)	***15in&None-Raw***	***Regression&BC-Metchalizer***	***Regression&None-Anchor***
Adipic acid (1) (up) (−)	2.4 *, 1.3, 1.8, 0.2, −0.3	0.2, −0.3, 0.2, −0.7 *, −1.0	0.5 *, 0.1, 0.3, −0.5, −0.0
C10:1 Decenoylcarnitine (up) (+)	31.4 *, 24.3 *, 6.2 *, 19.2 *, 12.8 *	22.0 *, 18.3 *, 7.6 *, 15.9 *, 8.3 *	63.3 *, 49.8 *, 13.6 *, 39.4 *, 9.4 *
C6 Hexanoylcarnitine (up) (+)	38.7 *, 32.5 *, 7.7 *, 22.6 *, 5.6 *	21.4 *, 19.2 *, 7.3 *, 15.0 *, 6.9 *	78.0 *, 66.1 *, 16.4 *, 46.1 *, 5.6 *
C8 Octanoylcarnitine (up) (+)	58.1 *, 69.7 *, 29.7 *, 38.3 *, 25.2 *	25.0 *, 27.6 *, 16.5 *, 19.1 *, 12.1 *	148.7 *, 178.8 *, 76.7 *, 98.2 *, 23.9 *
Hexanoic acid/Trans-cyclohexane−1,2-diol (up) (−)	0.4, 0.1, 0.4, 0.6, −0.4	−0.1, −0.9, 0.5, 1.3, −0.9	−0.0, −0.2, 0.0, 0.2, 0.4
Sebacic acid (up) (−)	0.6, 1.5, 1.3, −0.3, −0.2	−0.5, −0.3, −0.1, −0.5, −0.9	0.0, 0.4, 0.3, −0.2, 0.0
Sebacic acid (up) (+)	0.1, 0.5, 0.6, −0.5, −0.4	−0.3, −0.2, 0.1, −0.6, −0.7	0.4 *, 0.6, 0.6 *, 0.0, −0.0
Suberic acid (up) (−)	1.6 *, 1.6, 1.6 *, 0.0, −0.1	−0.0, −0.2, 0.0, −0.5, −0.7	0.3 *, 0.3, 0.3, −0.4, 0.2
Suberic acid (up) (+)	0.9, 1.0, 0.5, −0.2, −0.5	0.1, −0.2, −0.1, −0.6, −0.9	0.5, 0.5, 0.2, −0.2, −0.3
Methylmalonyl-CoA mutase deficiency N = 1			
C3 Propionylcarnitine (up) (+)	85.0 *	39.9 *	234.9 *
Organic cation transporter 2 deficiency N = 1			
C0 L-Carnitine (down) (+)	−2.3 *	−1.3 *	−0.7
Ornithine aminotransferase N = 1			
Guanidinoacetic acid (down) (+)	−2.2 *	−2.2 *	−1.7 *
Ornithine (up) (+)	33.1 *	11.7 *	37.4 *
Ornithine transcarbamylase deficiency N = 2			
Citrulline (down) (−)	2.9 *, 1.0 *	1.4 *, 0.6	2.0 *, 1.1 *
Citrulline (down) (+)	1.4 *, 2.5 *	0.6 *, 1.1 *	1.2 *, 1.5 *
Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid (up) (−)	1.0 *, 1.3 *	0.6, 0.6	0.3, 1.5 *
Uridine (up) (−)	7.7 *, 4.0 *	4.4 *, 5.0 *	6.5 *, 5.4 *
Phenylketonuria N = 4			
N-Acetylphenylalanine (up) (−)	116.5 *, 18.6, 73.9 *, 8.5 *	25.2 *, 11.9 *, 34.6 *, 7.3 *	45.7 *, 14.7, 56.3 *, 6.5 *
Phenylacetic acid (up) (+)	−0.3, −0.1, 1.6, 0.6	−0.4, −0.0, 1.6, 0.6	−0.3, −0.1, 0.1, 1.2
Phenylalanine (up) (−)	87.7 *, 38.5 *, 161.6 *, 14.6 *	34.3 *, 23.4 *, 43.4 *, 9.5 *	94.2 *, 51.6 *, 121.7 *, 26.1 *
Phenylalanine (up) (+)	42.6 *, 24.5 *, 80.8 *, 30.3 *	21.1 *, 16.7 *, 29.9 *, 9.2 *	55.5 *, 38.3 *, 68.9 *, 22.9 *
alpha-N-Phenylacetylglutamine (up) (−)	5.2 *, 1.3 *, 4.7 *, 0.8	2.2 *, 1.6 *, 3.2 *, 0.3	3.7 *, 2.4 *, 5.2 *, −1.0 *
alpha-N-Phenylacetylglutamine (up) (+)	6.1 *, 1.4, 4.3 *, 0.4	2.8 *, 1.9 *, 3.1 *, 0.3	4.9 *, 3.1, 4.7 *, −0.7 *

Table A6. *Cont.*

Biomarker (Expected Z-Score Sign) (Ion Mode)	***15in&None-Raw***	***Regression&BC-Metchalizer***	***Regression&None-Anchor***
Propionic acidemia N = 2			
C3 Propionylcarnitine (up) (+)	124.3 *, 199.2 *	47.8 *, 37.1 *	148.5 *, 200.1 *
Thymidine phosphorylase deficiency N = 1			
Thymidine (up) (−)	44.3 *	103.5 *	35.3 *
Tyrosinemia I N = 2			
4-Hydroxyphenyllactic acid (up) (−)	337.5 *, 1046.6 *	11.4 *, 16.1 *	403.6 *, 582.8 *
Tyrosine (up) (+)	16.6 *, 43.1 *	9.8 *, 23.6 *	25.2 *, 40.4 *
Very Long Chain Acyl-CoA Dehydrogenase Deficiency N = 1			
C14:1 Tetradecenoylcarnitine (up) (+)	211.0 *	23.8 *	213.8 *
C1 6 Hexadecanoylcarnitine (up) (+)	7.0 *	2.3 *	7.3 *
C18 Octadecanoylcarnitine (up) (+)	3.3 *	1.2 *	3.9 *
C18:1 Oleoylcarnitine (up) (+)	5.8 *	2.0 *	9.5 *
Carnitine palmitoyltransferase II N = 2			
Adipic acid (1) (up) (−)	−0.1, 0.1	−0.0, −0.1	0.7, −0.0
C0 L-Carnitine (down) (+)	−0.1, −0.7	−0.1, −1.0 *	1.5 *, −0.4
C16 Hexadecanoylcarnitine (up) (+)	8.5 *, 19.3 *	4.1 *, 6.8 *	8.0 *, 30.9 *
C18 Octadecanoylcarnitine (up) (+)	13.8 *, 53.5 *	5.0 *, 8.2 *	14.4 *, 45.3 *
Sebacic acid (up) (−)	−0.1, −0.3	−0.0, −0.4	0.7, −0.3
Sebacic acid (up) (+)	0.2, 0.1	−0.1, −0.2	0.7, −0.2
Suberic acid (up) (−)	0.1, −0.2	0.3, −0.4	1.0, −0.2
Suberic acid (up) (+)	0.2, −0.5	0.0, −0.5	0.5, −0.2
C18:1 Oleoylcarnitine (up) (+)	6.5 *, 27.7 *	3.5 *, 5.6 *	5.7 *, 33.8 *

Appendix E. Lost Biomarkers Due to Merging of Datasets

58 IEM patients were measured in nine separate batches, having together 35 unique IEM. We obtained 517 biomarker-patient combinations across all batches (for both ion modi). When a biomarker was detected in both ion modi, and we select the ion modus for which that biomarker had on average (across all nine batches) the highest abundancy, we ended up with 473 biomarker-patient combinations. After merging the nine batches, we lost biomarkers since we required each feature/metabolite to be merged across all nine batches. This resulted in 239 biomarker-patient combinations for both ion modi. Additionally, again, after removing biomarkers which are detected in both ion modi (selecting the most abundant one), we ended up with 195 biomarker-patient combinations. These last 195 biomarker-patient combinations were used to make the ROC curves as described in Section 4.6.

The following biomarkers (ion mode) were obtained within the batch, but were lost after merging the nine batches: **2-Methyl-3-hydroxybutyryl-CoA dehydrogenase deficiency**: Tiglylglycine(-). **3-methylcrotonyl-coa-carboxylase deficiency**: C5OH 3-Hydroxyisovalerylcarnitine(+/−), 3-Methylcrotonylglycine(+/−). **Adenylosuccinate lyase deficiency**: Succinyladenosine(+/−), SAICAR(+). **Alkaptonuria**: Homogentisic acid(+). **Alpha-Methylacyl-CoA racemase deficiency**: Dihydroxycholestanoic acid + gly(−), Dihydroxycholestanoic acid + tau(−), Trihydroxycholestanoic acid + gly(−), Trihydroxycholestanoic acid + tau(−). **Aminoacylase I deficiency**: N-Acetylmethionine(+), N-Acetylthreonine(+), N-Acetylalanine(−), N-Acetylarginine(−), N-Acetylglutamic acid(−), N-Acetylglutamic acid(+), N-Acetylleucine(−), N-Acetylleucine(+), N-Acetylmethionine(−). **Argininemia**: 4-Guanidinobutyric acid(−), Uracil (1)(−), Uridine(+), Orotic acid(−), Guanidinoacetic acid(−), N-Acetylarginine(−), Homoarginine(−), Uracil (2)(−). **Argininosuccinic aciduria**: Argininosuccinic acid(+/−), Orotic acid(−), Cytidine(−), Cytidine(+), Uracil (2)(+/−), Uracil (1)(−), Homocitrulline(−), Uridine(+), N-Acetylcitrulline(+/−), Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid(+). **Beta-ketothiolase deficiency**: 2-Methyl-3-hydroxybutyric acid(−), C5:1 Tiglylcarnitine(+), 2-Methylacetoacetic acid(−), Tiglylglycine(−). **Beta-mannosidose**: Glcnac-man(+/−). **Carbamoyl Phosphate Synthetase deficiency**: 2-Ketoglutaric acid(+), Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid(+). **Citrullinemia type I**: N-Acetylcitrulline(+/−), Orotic acid(−), Uracil (1)(−), Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid(+), Citrulline lactam(+), Uridine(+). **Combined malonic and methylmalonic aciduria**: C3DC Malonylcarnitine(+). **Glutamate formiminotransferase deficiency**: Formiminoglutamic acid(−), Hydantion-5-propionic acid(−). **Glutaric aciduria I**: C5DC Glutarylcarnitine(−), 3-Hydroxyglutaric acid(+), Glutaric acid(−), Glutarylglycine(+/−), 3-Hydroxyglutaric acid(−). **Glutaric aciduria II**: Ethylmalonic acid(+), C5DC Glutarylcarnitine(−), 2-Hydroxyglutaric acid(+), Glutaric acid(−), 2-Hydroxyglutaric acid(−), Adipic acid (2)(−), C14 Tetradecanoylcarnitine(+), C10 Decanoylcarnitine(−), Hexanoylglycine(−), Ethylmalonic acid(−), C16:1 Hexadecenoylcarnitine(−), 3-Hydroxyglutaric acid(+), Adipic acid (1)(+), Isobutyrylglycine(−), C14:1 Tetradecenoylcarnitine(−), Hexanoylglycine(+), Isovaleryglycine(−), 3-Hydroxyglutaric acid(−). **Homocystinuria**: Homocysteine(−), Methionine + Methioninesulfoxide(−). **Isovaleric acidemia**: Isovaleryglycine(+/−), C5 Isovalerylcarnitine(−), 3-Hydroxyisovaleric acid(−). **Long-chain-3-hydroxyacyl CoA dehydrogenase deficiency**: C18OH 3-Hydroxyoctadecanoylcarnitine(+), C14OH 3-Hydroxytetradecanoylcarnitine(+), C18:1OH 3-Hydroxyoleoylcarnitine(+), C14:1OH 3-Hydroxytetradecenoylcarnitine(+), C18OH 3-Hydroxyoctadecanoylcarnitine(−), C16:1OH 3-Hydroxyhexadecenoylcarnitine(+), C16OH 3-Hydroxyhexadecanoylcarnitine(+). **Lysinuric protein intolerance**: Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid(+), Lysine (−), Orotic acid(−), Ornithine(−). **Malonyl-Coa decarboxylase deficiency**: C3DC Malonylcarnitine(+), Malonic acid(+). **Maple syrup urine disease**: Leucine(−), 2-Keto-4-methylvaleric acid(+). **Medium Chain Acyl-CoA Dehydrogenase Deficiency**: Heptanoylcarnitine(+), 3-Hydroxydecanedioic acid(+), 7-Hydroxyoctanoic acid(−), nonanoylcarnitine(+), Decenedioic acid(+), Octanoic acid(−), Phenylpropionylglycine(−), Suberylglycine(−), Unde-

canoylcarnitine(+), Octanoylglycine(−), Hexanoylglycine(+/−), Phenylpropionylglycine(+), Hexanoic acid/Trans-cyclohexane-1,2-diol(+), Octanoylglycine(+), C10:1 Decenoylcarnitine(−), Adipic acid (1)(+), C8 Octanoylcarnitine(−). **Methylmalonic acidemia**: Methylcitric acid (1)(−), Methylcitric acid (2)(−), Methylmalonic acid(−), Methylmalonic acid(+), C4DC Methylmalonylcarnitine(+). **Mevalonic aciduria**: Mevalonic acid(−), Mevalonic acid(+). **Organic cation transporter 2 deficiency**: C0 L-Carnitine(−). **Ornithine aminotransferase**: 3-Amino-2-piperidone(+). **Ornithine transcarbamylase deficiency**: Orotic acid(−), Glutamine + Glutamic acid/N-Methyl-D-Aspartic acid(+), Uridine(+). **Phenylketonuria**: Phenylpyruvic acid(−), Phenylacetic acid(−), Phenyllactic acid(−), N-lactoyl-Phenylalanine(−), Phenylalanylphenylalanine(+/−), Glutamylphenylalanine(+), Phenylpyruvic acid(+), Glutamylphenylalanine(−). **Propionic acidemia**: Methylcitric acid (2)(−), C3 Propionylcarnitine(−), Methylcitric acid (1)(+/−), Propionylglycine(+/−), Glycine(+). **Thymidine phosphorylase deficiency**: Deoxyuridine(+), Deoxyuridine(−), Uracil (2)(−), Thymidine(+), Uracil (1)(−). **Tyrosinemia I**: Phenylpyruvic acid(+/−), Tyrosine(−), N-Acetyltyrosine(−), 4-Hydroxyphenylacetic acid(+/−), 4-Hydroxyphenylpyruvic acid(−), N-Acetyltyrosine(+). **Very Long Chain Acyl-CoA Dehydrogenase Deficiency**: C14 Tetradecanoylcarnitine(+), C14:2 Tetradecadienoylcartinine(+). **Carnitine palmitoyltransferase II**: Adipic acid (1)(+), C0 L-Carnitine(−).

Appendix F. Performance BC-Metchalizer for a Different Number of Internal Standards

Because *BC-Metchalizer* uses multiple internal standards we explored the influence of the number of internal standards on normalization by comparing quantitative measurements with the abundancies after normalization. We took 20 random combinations of n internal standards; n being the number of internal standards. Subfigure in A, B in Figure A4 show the overall R^2 score and Spearman score, for the quantified metabolites (see subfigure C). No clear increase/decrease in performance was observed when increasing the number of internal standards. However, looking at the individual normalization performances per metabolite showed that some combinations of internal standards improved or decreased R^2 score. For example, phenylalanine showed most variation in performance and more internal standards did not lead to increased performance. From here, it seemed that some combinations of internal standards normalized certain metabolites better.

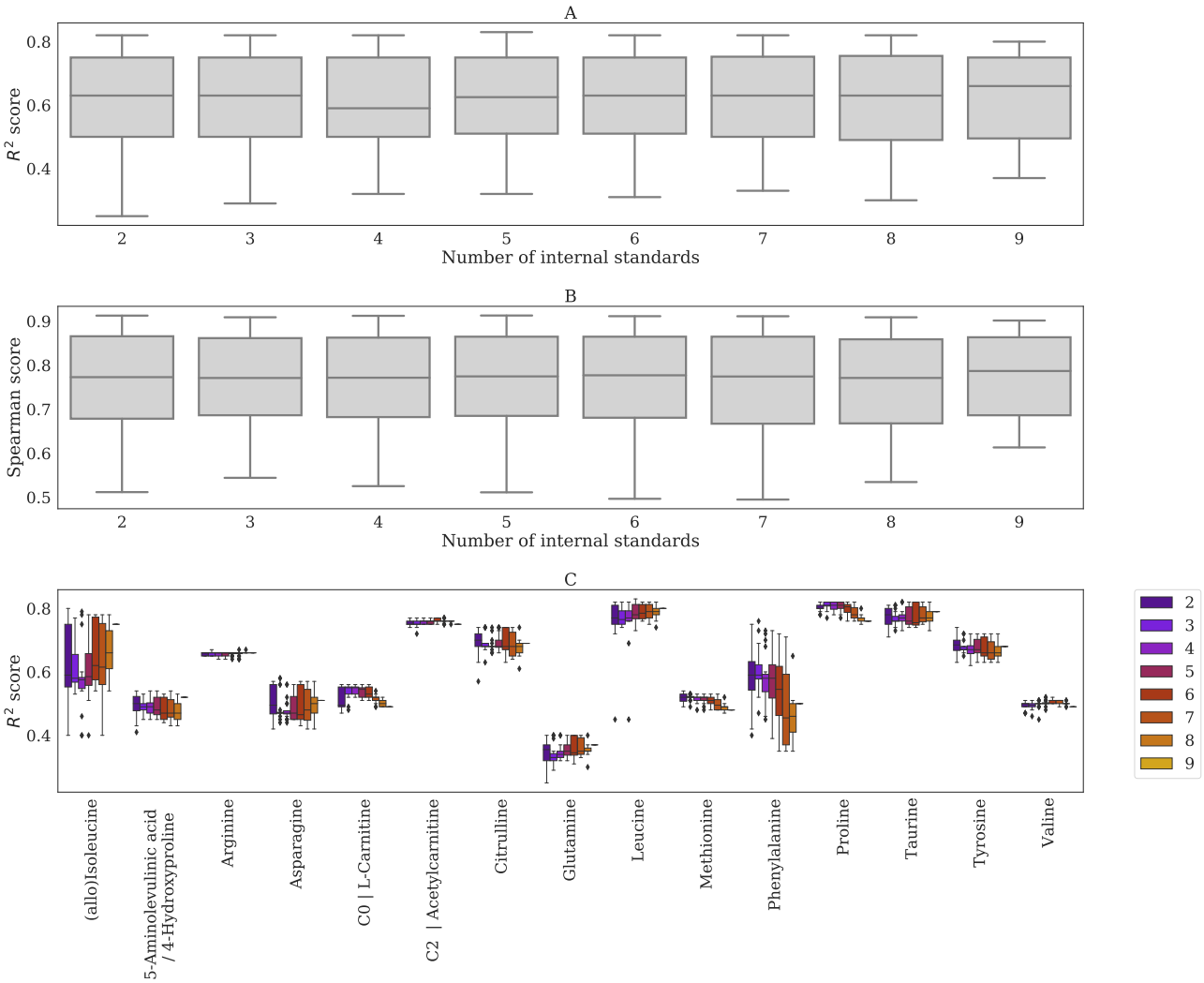


Figure A4. Performance of normalization based on a comparison with quantitative measurements for increasing the number of internal standards with *BC-Metchalizer*. (**A**) R^2 score (**B**) Spearman score (**C**) R^2 score per metabolite and number of internal standards. The legend indicates how many internal standards were used for normalization.

Appendix G. Performance BC-Metchalizer for a Different Number of Batches

Robustness of normalizing with *BC-Metchalizer* was investigated by normalizing a different number of batches, where we took 20 random combinations of n batches; n being the number of batches. Ideally, the performance of normalization should be constant over the number of batches being analyzed. Subfigure in A, B in Figure A5 shows the overall R^2 score and Spearman score for the quantified metabolites (see subfigure C) for increasing batches. It can be observed that its performance was globally constant over the whole range. However, on the metabolite specific-level, we do observe difference between the number of batches being used.

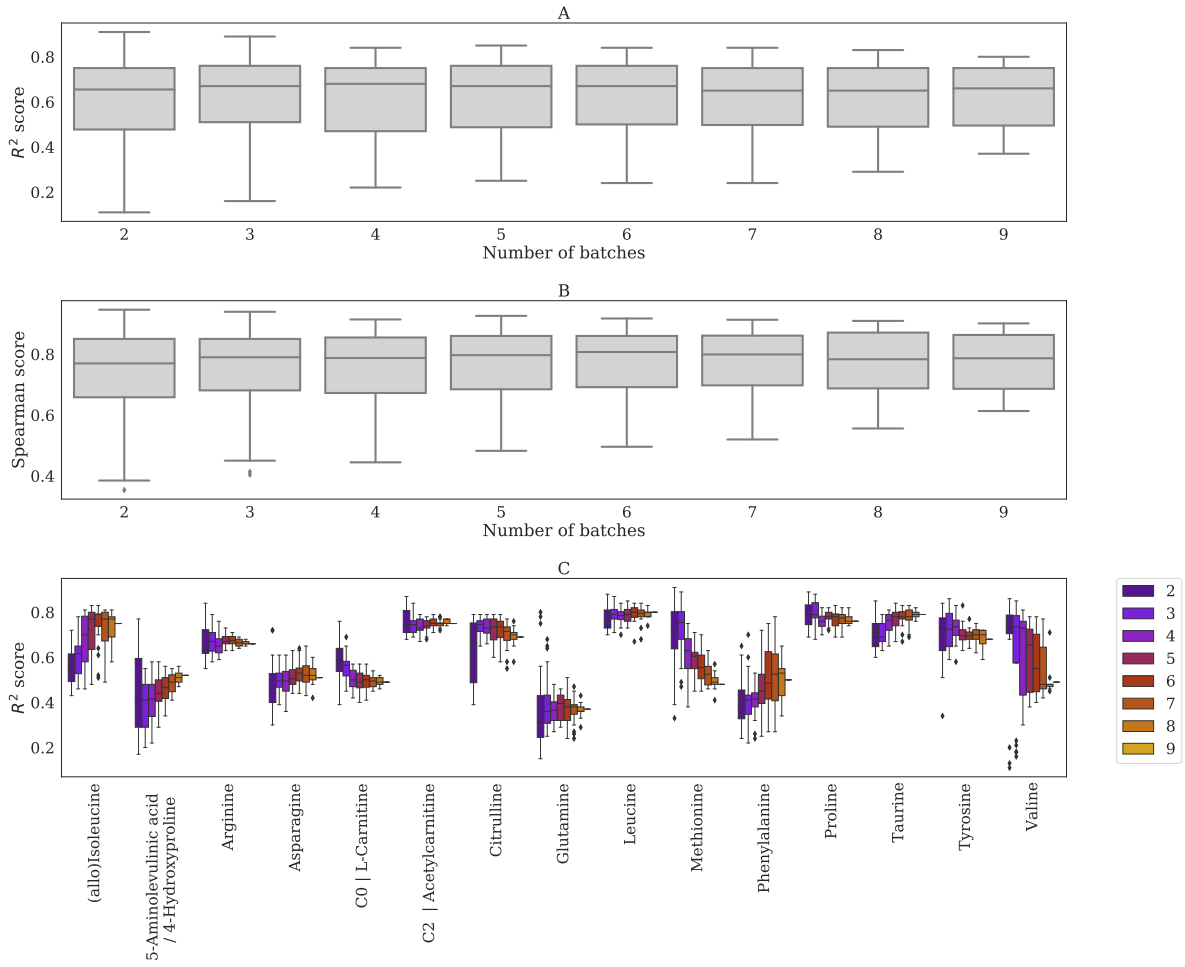


Figure A5. Performance of normalization based on a comparison with quantitative measurements for increasing the number of batches with *BC-Metchalizer*. (**A**) R^2 score (**B**) Spearman score (**C**) R^2 score per metabolite and number of internal standards. The legend indicates how many batches were used for normalization.

Appendix H. Age and Sex Similarity of Reference Samples with Patients for Different Z-Score Methods

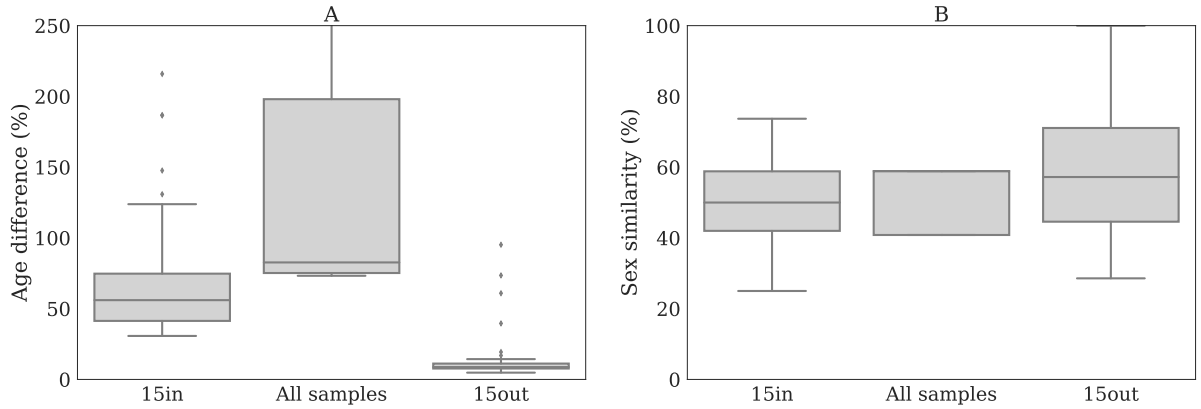


Figure A6. (**A**) Difference in age between patient and the reference samples for every Z-score method. Mean age difference was determined by calculating the difference in age of the patient with the references divided by the age of the patient (times 100%) whereafter the mean was taken. (**B**) Similarity in sex between patient and controls for every method.

Appendix I. The Influence of Log-Transformation on Z-Scores and p-Values

Simulations were performed by randomly drawing a mean μ_{pop} from $\mathcal{N}(\mu = 10{,}000, \sigma = 5000)$ followed by 15 random reference samples from $\mathcal{N}(\mu = \mu_{pop}, \sigma = \frac{\mu_{pop}}{5})$. From the same normal distribution we drew μ_{trip}, and three 'triplicate measurements' were drawn from $\mathcal{N}(\mu = \mu_{trip}, \sigma = \frac{\mu_{trip}}{5})$. Z-scores were calculated from the average en standard deviation obtained from the 15 'controls'.

From Figure A7, we observe that positive Z-scores have relative lower p-values for log-transformed abundancies, and vice versa.

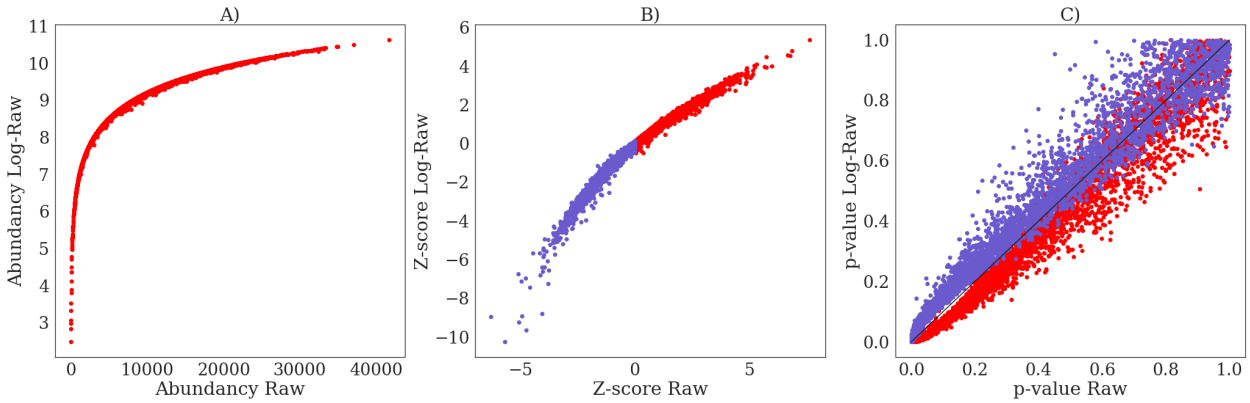


Figure A7. (**A**) The simulated log-transformed abundancies (*Log-Raw*) versus the raw abundancies (*Raw*). (**B**) The difference in Z-scores between *Log-Raw* and *Raw*. (**C**) The difference in Welch's t-test p-values between *Log-Raw* and *Raw*.

Appendix J. Bland–Altman Analysis for Z-Scores Obtained Using All Samples Versus Regression

The Z-scores that were obtained from *Regression* and *All samples* were compared while using a Bland–Altman plot, see Figure A8 (left panels). The vertical axis indicates the difference in Z-scores obtained between the two methods, whereas the horizontal axis indicates the average Z-scores. The middle and right panels indicate the distribution of the Z-score differences for all features and the expected IEM biomarkers, respectively. These distributions show that the Z-score differences are not baised towards one of the two Z-score methods and suggests that for a given Z-score cutoff ($Z_{abnormal}$) the same amount of positives will be obtained.

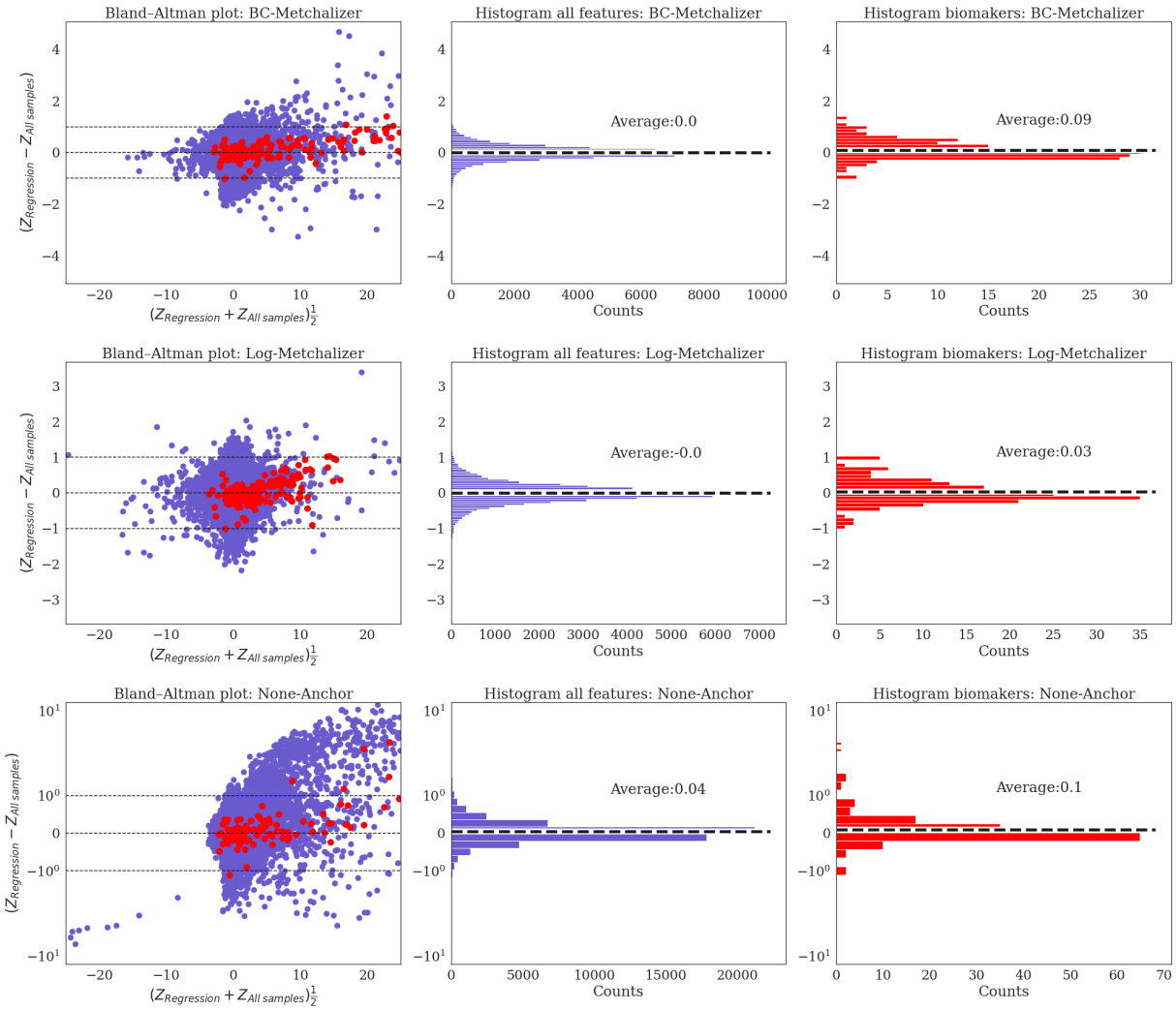


Figure A8. (**Left**) These panels display the Bland- Altman plots where the differences in Z-scores between *Regression* and *All samples* are plotted. Blue dots indicate data points originating from a (random) feature and patient whereas the red dots are the IEM biomarkers. (**Middle**) These panels show the distribution of the Z-score differences between the two approaches for all features and patients. (**Right**) The panels show the distribution of the Z-score differences between the two approaches for all IEM biomarkers.

Appendix K. Resemblance of Patients Sharing the Same IEM

We investigated whether normalization improved resemblance among patients sharing the same IEM without losing the biological information present between non-matching IEM patients. Two metrics were considered to be informative: (1) the *Euclidean distance* between each pair of patients using the Z-transformed data (by subtracting the mean and dividing by the standard deviation using all samples) and (2) a *weighted cosine similarity* between each pair of patients on the Z-transformed data. Triplicates were averaged prior to calculating these metrics.

The *weigthed cosine similarity* between two vectors $\vec{u}$ and $\vec{v}$ is given by:

$$\begin{aligned} \text{Weighted cosine similarity} &= \frac{\sum_i w_i u_i v_i}{\sum_i w_i v_i^2 \ \sum_i w_i u_i^2} \\ \text{with } w_i &= \frac{1}{1 + \exp\left(-2[(|v_i| + |u_i|)0.5 - 1]\right)} \end{aligned} \tag{A1}$$

where u_i (v_i) indicates the Z-score of patient u (v) for feature i. Note that the weight w_i increases when the average $(|v_i| + |u_i|)0.5$ increases, thereby reducing the importance of a feature when this feature has a low |Z-score| for both patients (and vice versa). The *weighted cosine similarity* approaches 1 when two profiles ($\vec{u}$ and $\vec{v}$) are (more) similar.

We included the following IEM for this analysis: Alpha-Methylacyl-CoA racemase deficiency (N = 2), Argininosuccinic aciduria (N = 3), Beta-ketothiolase deficiency (N = 2), Carbamoyl Phosphate Synthetase deficiency (N = 2), Glutaric aciduria I (N = 2), Glutaric aciduria II (N = 2), Homocystinuria (N = 3), Long-chain-3-hydroxyacyl CoA dehydrogenase deficiency (N = 2), Lysinuric protein intolerance (N = 2), Maple syrup urine disease (N = 2), Medium Chain Acyl-CoA Dehydrogenase Deficiency (N = 5), Ornithine transcarbamylase deficiency (N = 2), Phenylketonuria (N = 4), Propionic acidemia (N = 2), Tyrosinemia I (N = 2), Carnitine palmitoyltransferase II (N = 2).

Figure A9 shows the results for both metrics (and ion modi). When compared with *Raw* and *Log-Raw*, we observe that the *Euclidean distances* were generally reduced after normalization for pairs of patients sharing the same IEM, whereas the *weighted cosine similarities* generally increased. The 75th percentiles for the *Euclidean distance* were lowest for *BC-Metchalizer* (22.54), *Log-Metchalizer* (22.85) and *None-Anchor* (22.86) in positive ion mode. For negative ion mode, the 75th percentiles were lowest for *Log-Metchalizer* (20.39), *BC-Metchalizer* (21.93) and *Log-RUVrand* (22.42). When considering the *weighted cosine similarities*, the 25th percentile was highest for *BC-Metchalizer* (0.285), *Log-NOMIS* (0.253) and *Log-Metchalizer* (0.250).

Additionally, we expect the difference between the *None-matching IEM* and *Matching IEM* group to increase for improved normalization performances. We calculated the *p*-value using the Mann–Whitney U test between these two groups for each method (and ion mode), where the -log(*p*-value) is shown in Figure A9. For the *Euclidean distance*, we observe that *Log-EigenMS* and *Log-Metchalizer* separated these groups best in positive ion mode, and *Log-Metchalizer* performed best using negative ion mode. When considering the *weighted cosine similarity*, *Log-EigenMS*, *BC-Metchalizer*, and *Log-Metchalizer* performed best for positive ion mode. For negative ion mode *BC-Metchalizer* and *Log-Metchalizer* scored best. When using the *weighted cosine similarity* for clustering, we observe that indeed *Log-Metchalizer* improved clustering for patient sharing the same IEM when compared with using raw data (see Figures A10 and A11). Together, this shows that *Metchalizer* performances among the best in bringing similar IEM patients closer to each other without reducing the differences between non-matching IEM patients.

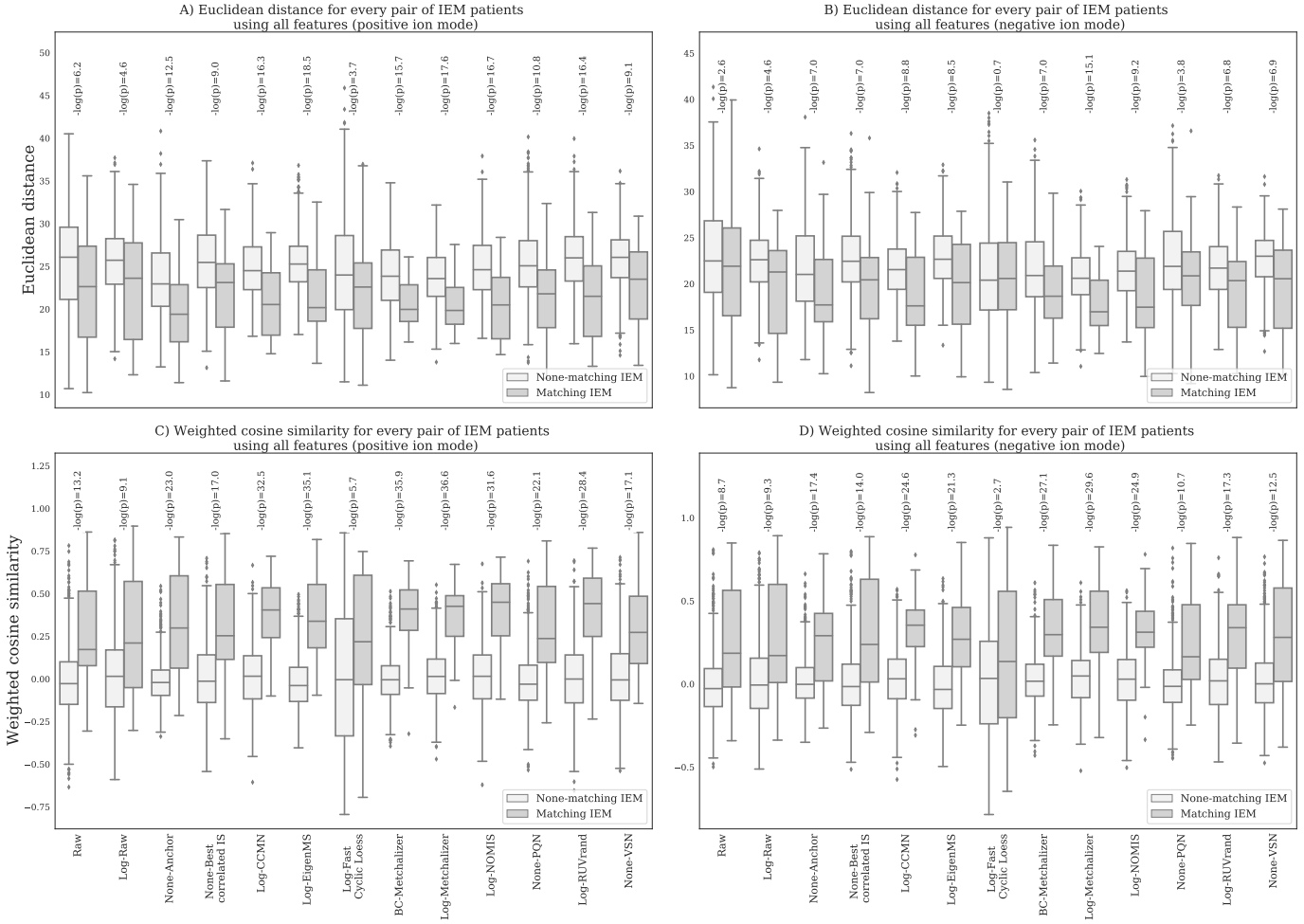


Figure A9. Given a certain normalization method, each boxplot shows the *Euclidean distance/ weighted cosine similarity* for pairs of patients, either sharing or not sharing the same IEM as indicated by the legend. We used the Mann-Whitney U test to test how these two groups differed from each other, and showed the -log(*p*-value) above each corresponding method. (**A**) *Euclidean distances* for positive ion mode (**B**) *Euclidean distances* for negative ion mode. (**C**) *Weighted cosine similarities* for positive ion mode. (**D**) *Weighted cosine similarities* for negative ion mode.

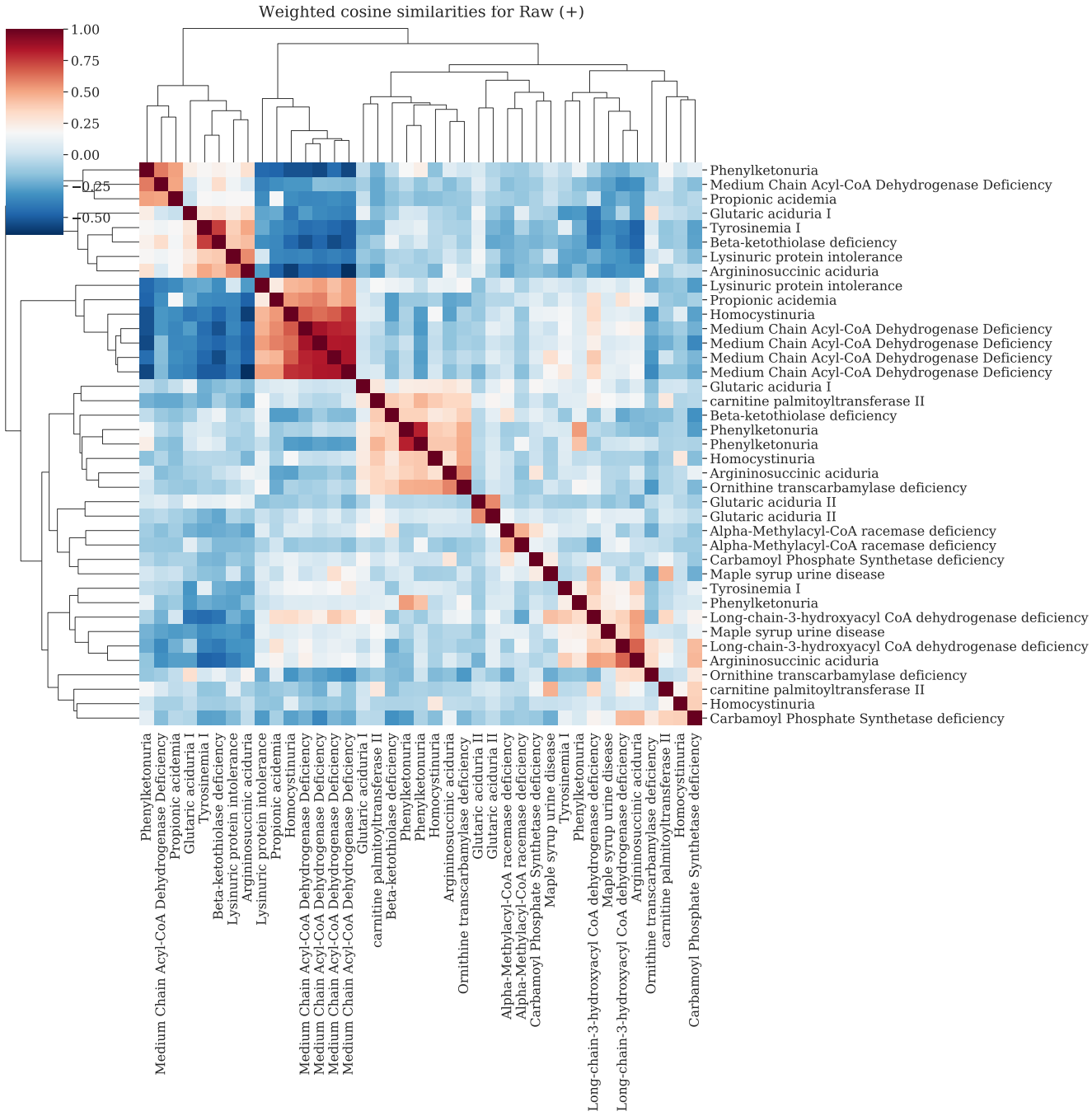


Figure A10. Clustermap of the *weighted cosine similarities* for IEM patients in positive ion mode using *Raw* data.

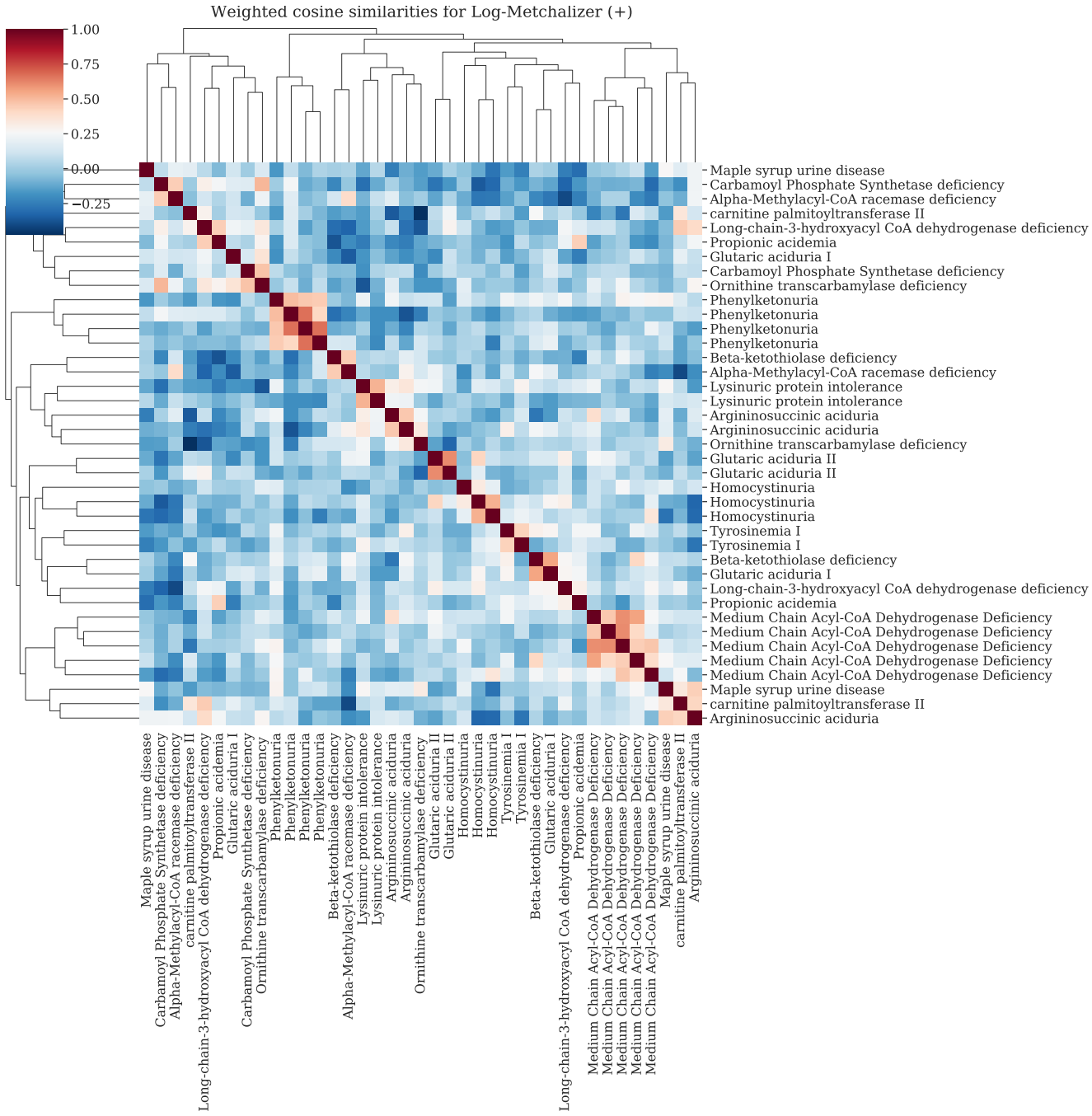


Figure A11. Clustermap of the *weighted cosine similarities* for IEM patients in positive ion mode using *Log-Metchalizer* data.

References

1. Miller, M.J.; Kennedy, A.D.; Eckhart, A.D.; Burrage, L.C.; Wulff, J.E.; Miller, L.A.; Milburn, M.V.; Ryals, J.A.; Beaudet, A.L.; Sun, Q.; et al. Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *J. Inherit. Metab. Dis.* **2015**, *38*, 1029–1039, [CrossRef] [PubMed]
2. Coene, K.L.M.; Kluijtmans, L.A.J.; van der Heeft, E.; Engelke, U.F.H.; de Boer, S.; Hoegen, B.; Kwast, H.J.T.; van de Vorst, M.; Huigen, M.C.D.G.; Keularts, I.M.L.W.; et al. Next-generation metabolic screening: Targeted and untargeted metabolomics for the diagnosis of inborn errors of metabolism in individual patients. *J. Inherit. Metab. Dis.* **2018**, *41*, 337–353, [CrossRef] [PubMed]
3. Körver-Keularts, I.M.L.W.; Wang, P.; Waterval, H.W.A.H.; Kluijtmans, L.A.J.; Wevers, R.A.; Langhans, C.D.; Scott, C.; Habets, D.D.J.; Bierau, J. Fast and accurate quantitative organic acid analysis with LC-QTOF/MS facilitates screening of patients for inborn errors of metabolism. *J. Inherit. Metab. Dis.* **2018**, *41*, 415–424, [CrossRef] [PubMed]

4. Haijes, H.A.; Willemsen, M.; Van der Ham, M.; Gerrits, J.; Pras-Raves, M.L.; Prinsen, H.C.M.T.; Van Hasselt, P.M.; De Sain-van der Velden, M.G.M.; Verhoeven-Duif, N.M.; Jans, J.J.; et al. Direct Infusion Based Metabolomics Identifies Metabolic Disease in Patients' Dried Blood Spots and Plasma. *Metabolites* **2019**, *9*, 12, [CrossRef] [PubMed]
5. Bonte, R.; Bongaerts, M.; Demirdas, S.; Langendonk, J.G.; Huidekoper, H.H.; Williams, M.; Onkenhout, W.; Jacobs, E.H.; Blom, H.J.; Ruijter, G.J.G. Untargeted Metabolomics-Based Screening Method for Inborn Errors of Metabolism using Semi-Automatic Sample Preparation with an UHPLC- Orbitrap-MS Platform. *Metabolites* **2019**, *9*, 289, [CrossRef]
6. Glinton, K.E.; Levy, H.L.; Kennedy, A.D.; Pappan, K.L.; Elsea, S.H. Untargeted metabolomics identifies unique though benign biochemical changes in patients with pathogenic variants in UROC1. *Mol. Genet. Metab. Rep.* **2019**, *18*, 14–18, [CrossRef]
7. Chaleckis, R.; Murakami, I.; Takada, J.; Kondoh, H.; Yanagida, M. Individual variability in human blood metabolites identifies age-related differences. *Proc. Natl. Acad. Sci. USA* **2016**, *113*, 4252–4259, [CrossRef]
8. Rist, M.J.; Roth, A.; Frommherz, L.; Weinert, C.H.; Krüger, R.; Merz, B.; Bunzel, D.; Mack, C.; Egert, B.; Bub, A.; et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS ONE* **2017**, *12*, e0183228, [CrossRef]
9. Yu, Z.; Zhai, G.; Singmann, P.; He, Y.; Xu, T.; Prehn, C.; Römisch-Margl, W.; Lattka, E.; Gieger, C.; Soranzo, N.; et al. Human serum metabolic profiles are age dependent. *Aging Cell* **2012**, *11*, 960–967, [CrossRef]
10. Lawton, K.A.; Berger, A.; Mitchell, M.; Milgram, K.E.; Evans, A.M.; Guo, L.; Hanson, R.W.; Kalhan, S.C.; Ryals, J.A.; Milburn, M.V. Analysis of the adult human plasma metabolome. *Pharmacogenomics* **2008**, *9*, 383–397, [CrossRef]
11. Veselkov, K.A.; Vingara, L.K.; Masson, P.; Robinette, S.L.; Want, E.; Li, J.V.; Barton, R.H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T.M.; et al. Optimized Preprocessing of Ultra-Performance Liquid Chromatography/Mass Spectrometry Urinary Metabolic Profiles for Improved Information Recovery. *Anal. Chem.* **2011**, *83*, 5864–5872.10.1021/ac201065j. [CrossRef] [PubMed]
12. Li, B.; Tang, J.; Yang, Q.; Li, S.; Cui, X.; Li, Y.; Chen, Y.; Xue, W.; Li, X.; Zhu, F. NOREVA: Normalization and evaluation of MS-based metabolomics data. *Nucleic Acids Res.* **2017**, *45*, W162–W170, [CrossRef] [PubMed]
13. Välikangas, T.; Suomi, T.; Elo, L.L. A systematic evaluation of normalization methods in quantitative label-free proteomics. *Briefings Bioinform.* **2016**, *19*, 1–11, [CrossRef] [PubMed]
14. Vreken, P.; van Lint, A.E.M.; Bootsma, A.H.; Overmars, H.; Wanders, R.J.A.; van Gennip, A.H. Rapid Diagnosis of Organic Acidemias and Fatty-acid Oxidation Defects by Quantitative Electrospray Tandem-MS Acyl-Carnitine Analysis in Plasma. In *Current Views of Fatty Acid Oxidation and Ketogenesis*; Springer US: New York, NY, USA, 2002; pp. 327–337, [CrossRef]
15. Redestig, H.; Fukushima, A.; Stenlund, H.; Moritz, T.; Arita, M.; Saito, K.; Kusano, M. Compensation for Systematic Cross-Contribution Improves Normalization of Mass Spectrometry Based Metabolomics Data. *Anal. Chem.* **2009**, *81*, 7974–7980, [CrossRef] [PubMed]
16. Karpievitch, Y.V.; Nikolic, S.B.; Wilson, R.; Sharman, J.E.; Edwards, L.M. Metabolomics Data Normalization with EigenMS. *PLoS ONE* **2015**, *9*, e116221, [CrossRef]
17. Ballman, K.V.; Grill, D.E.; Oberg, A.L.; Therneau, T.M. Faster cyclic loess: Normalizing RNA arrays via linear models. *Bioinformatics* **2004**, *20*, 2778–2786, [CrossRef]
18. Sysi-Aho, M.; Katajamaa, M.; Yetukuri, L.; Orešič, M. Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinform.* **2007**, *8*, 93. [CrossRef]
19. Filzmoser, P.; Walczak, B. What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatogr. A* **2014**, *1362*, 194–205. [CrossRef]
20. Livera, A.M.D.; Sysi-Aho, M.; Jacob, L.; Gagnon-Bartsch, J.A.; Castillo, S.; Simpson, J.A.; Speed, T.P. Statistical Methods for Handling Unwanted Variation in Metabolomics Data. *Anal. Chem.* **2015**, *87*, 3606–3615. [CrossRef]
21. 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* **2002**, *18*, S96–S104. doi:10.1093/bioinformatics/18.suppl_1.s96. [CrossRef]
22. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.