Uncovering the consequences of batch effect associated missing values in omics data analysis

Harvard Wai Hann Hui[1,2^] and Wilson Wen Bin Goh[1,2,3*]

[1] Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore 308232, Singapore

[2] School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore

[3] Center for Biomedical Informatics, Nanyang Technological University, Singapore 637551, Singapore

[^] First author(s)

[*] Corresponding author(s)

E-mail: wilsongoh@ntu.edu.sg (WWBG)

Running title: Hui HWH / *Uncovering the consequences of batch effect associated missing values in omics data analysis*

18

ABSTRACT

Statistical analyses in high-dimensional omics data are often hampered by the presence of batch effects (BEs) and missing values (MVs), but the interaction between these two issues is not well-studied nor understood. MVs may manifest as a BE when their proportions differ across batches. These are termed as Batch-Effect Associated Missing values (BEAMs). We hypothesized that BEAMs in data may introduce bias which can impede the performance of missing value imputation (MVI). To test this, we simulated data with two batches, then introduced over 100 iterations, either 20% and 40% MVs in each batch (BEAMs) or 30% in both (control). K-nearest neighbours (KNN) was then used to perform MVI, in a typical global approach (M1) and a supposed superior batch-sensitized approach (M2). BEs were then corrected using ComBat. The effectiveness of the MVI was evaluated by its imputation accuracy and true and false positive rates. Notably, when BEAMs existed, M2 was generally undesirable as the differing application of MV filtering in M1 and M2 strategies resulted in an overall coverage deficiency. Additionally, both M1 and M2 strategies suffered in the presence of BEAMs, highlighting the need for a novel approach to handle MVI in data with BEAMs.

Keywords: Batch effect; Bioinformatics; Data analytics; Missing value imputation; Statistics

Author summary

Data in high-throughput omics data are often combined from different sources (batches), which creates batch effects in the data. Missing values are a common occurrence in these data, and their proportions are assumed to be equal across batches. However, instances exist when these proportions vary between batches, such as one batch having more missing values than another, resulting in batch effect associated missing values. Missing values are often dealt with through missing value imputation, but whether the variation in missing value proportions across batches affects imputation outcomes is unknown. In this paper, we investigate the consequence of performing imputation when this issue persists. We simulated data with equal and unequal missing value proportions, then assessed the performance of k-nearest neighbours imputation by its imputation accuracy and downstream analysis outcomes. This revealed that unequal missing value proportions worsens imputation and establishes the need for smarter imputation strategies to handle this complication.

INTRODUCTION

Batch effects (BEs) and missing values (MVs) are common problems in high-throughput omics data, such as genomics and proteomics(1). MVs refer to missing data that should be present in the acquired dataset, and if not properly handled, can negatively impact downstream analyses(2). MVs can cause a loss of statistical power(3), reducing analysis confidence or distorting data distributions, making the data non-representative of the original population(4).

When the proportion of missing values is low, simple omission of missing data may be sufficient. However, when the proportion of missing values is too high to ignore, complete eradication of missing data can result in significant information loss. In these situations, missing value imputation (MVI) techniques can be used to fill in the missing values. MVI techniques are used to estimate missing values based on the available data, which allows the use of powerful analytical techniques such as Principal Component Analysis (PCA) which require complete data. However, it's important to keep in mind that MVI methods may introduce bias and uncertainty(5), so it's important to choose the appropriate MVI method for the specific data set and the downstream analysis.

MVs can occur for a variety of reasons, including biological and technical reasons. In some cases, missingness is meaningful and should not be imputed, while in other cases it may be due to limited instrument sensitivity or algorithmic/statistical issues. The mechanism behind MVs is important in determining the appropriate handling procedure(6). MVs can be broadly categorized into three types: Missing not at random (MNAR), missing at random (MAR), and missing completely at random (MCAR)(7). MNAR typically refers to MVs that are missing due to having abundances below the limit of detection (LOD), which is a common type of missingness in omics data(8). While MNAR is explained by the MVs themselves, MAR refers to MVs that can be explained by the observed values. MVs that result from no apparent reason or pattern, and are independent of any unobserved or observed values, are known as

MCAR(9). In most omics data, MAR and MCAR are indistinguishable, and thus they are usually assumed to be MAR(10).

MVI covers a wide range of techniques. Simple methods involve replacing an MV with a constant or an estimated value. More complex methods include using statistical correlations or machine learning to transfer information from available data to estimate the missing values. These methods can range from simple mean or median imputation to more sophisticated techniques such as multiple imputation, hot deck imputation, expectation-maximization algorithm, and various machine learning models such as k-nearest neighbors, Random Forest, and neural networks. The choice of the MVI technique will depend on the characteristics of the data, the proportion of missing values, the type of missingness and the downstream analysis(6).

Even though MVI techniques exist, MVs are still considered a problem for several reasons. Determining the specific types and combinations of MVs that plague the data is difficult, making it hard to select the appropriate MVI technique(8). Most omics data are a mixture of MNAR and MAR which makes the selection process even more complex(11). Additionally, the MV filtering threshold, which refers to the proportion of MVs per feature that is acceptable, varies and can range from 5% to 20%(12). This is an important process conducted before MVI that can influence the outcome of downstream analyses. A strict threshold would retain data of higher confidence, whilst losing coverage. A lenient threshold will have greater coverage, but analyses would be more prone to biases. Choosing the appropriate threshold is also a challenging task. Some researchers may choose to avoid MVI as much as possible and instead opt for methods that do not require imputation(13). Ultimately, the choice of MVI technique will depend on the characteristics of the data, the proportion of missing values, the type of missingness, and the downstream analysis(6).

BEs are another common yet complex problem in high-throughput omics data, similar to MVs(1,14). BEs are technical sources of variance between replicates (batches) and are usually caused by factors such as poor experimental design or the combination of data from different studies without sufficient standardization. BEs act as a confounding factor and can have a negative impact on downstream analysis outcomes(1). To address BEs, batch effect correction algorithms (BECAs) are used to estimate and remove BEs from the data(15). Like MVIs, BECAs also make specific assumptions about the BEs, and the choice of a wrong BECA can lead to poor outcomes. A popular BECA is ComBat, which adjusts for batch effects using an empirical Bayes model(16).

Currently, the interaction between MVs and BEs, which are common occurrences in omics data, is not well-studied or understood. Currently, BECAs are designed to deal only on BEs on observable data. They are not designed to handle BEs involving MVs. We termed these Batch Effect Associated Missing values as BEAMs. More specifically, BEAMs are characterized by differential variation of MV proportions across batches, such that one batch can manifest much more MVs than another (Figure 1). In the absence of strategic MVI on a dataset with BEAMs, the different proportions of observed values may cause the MVI to be driven by the batch of least missingness and reduce MVI or subsequent BECA performance. Therefore, it is important to consider both MVs and BEs when analyzing omics data.

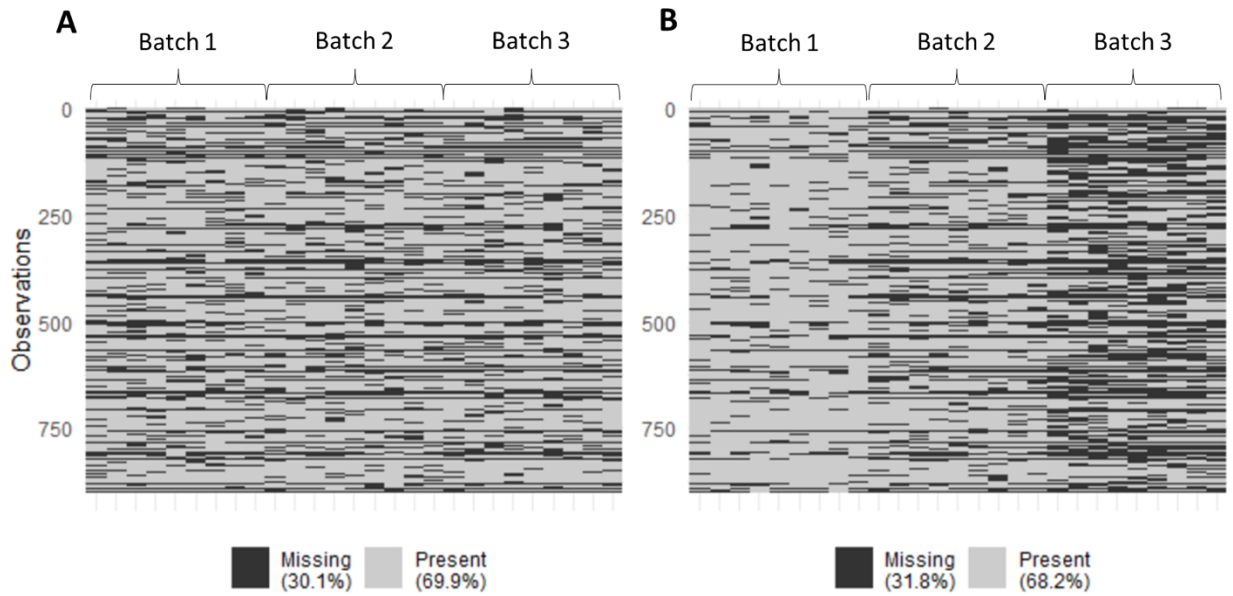
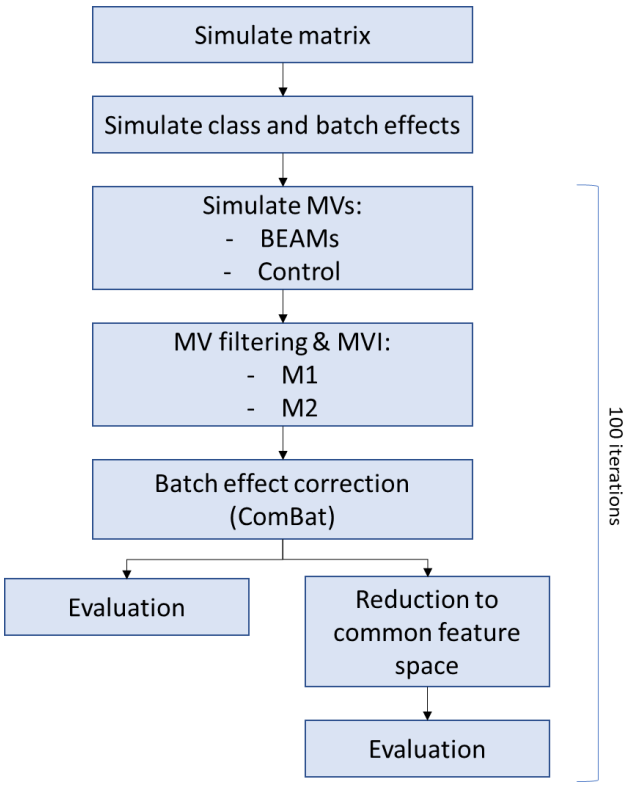


Figure 1. Heatmap visualization of batch effect associated missing values (BEAMs). (A) 30% MVs evenly distributed across batches. (B) 30% MVs disproportionately distributed across batches and are therefore BEAMs.

A previous study evaluated the typical global imputation strategy (termed as M1) against a batch-sensitized imputation approach (termed as M2)(17). It found that an M2 strategy was superior in terms of both imputation accuracy and differential analysis following BE correction. This was due to the retention of pertinent batch information, which improved BECA performances in BE detection and correction. Thus, the general recommendation when performing MVI on data with batches is to employ an M2 approach. However, these tests were conducted on data with evenly distributed MVs, and the utility of M1 and M2 strategies when BEAMs manifest in data is unknown.

In this study, the aim was to investigate the consequences of BEAMs on data when performing MVI using both M1 and M2 strategies. To accomplish this, we first simulated data containing class and batch effects with and without BEAMs. Then, MVI was conducted using the K-nearest neighbors (KNN) algorithm, using

122 both global (M1) and batch-sensitized (M2) strategies. The use of simulated data provides a controlled
123 setting where evaluations based on imputation accuracy and differential analysis can be assessed against
124 a "true" reference. Since varying levels of MVs in each batch between the BEAMs and non-BEAMs data
125 may result in different numbers of features after MV filtering, we also conducted analysis based only on
126 features that were common between both levels of missingness to make fair comparisons. The workflow
127 of the simulation studies is shown in Figure 2.



128

129 ***Figure 2. The workflow for simulation studies. 100 simulations were carried out to increase reliability***
130 ***of results.***

131

Results

Feature retention post-MV filtering is significantly influenced by BEAMs

As discussed, BEAMs are thought to affect the number of features retained in the data after MV filtering, the process which removes features based on a pre-determined missing threshold. The extent of feature coverage is important as it relates to how much information is retained in our data. Furthermore, certain features could be targets of interest which we want to conserve. Thus, it would be desirable to retain as many features as possible. Although a batch-sensitized (M2) approach is generally considered superior over the standard global (M1) approach(17), we wanted to investigate the differences in feature coverage within these strategies with and without BEAMs, and whether this would affect the selection between M1 and M2.

Thus, we compared the number of features retained with both M1 and M2 approaches in BEAMs and non-BEAMs scenarios. It was found that with moderate BEAMs (20:40%), feature coverage was significantly affected, as an M1 approach retained ~3% more features while an M2 approach retained ~7% less features (Figure 3). These differences widen further with severe BEAMs (10:50%), reaching ~13% more features in M1 and ~17% less features in M2 (Figure S2). With M1, the MV threshold for each feature was set across all batches. This means that when BEAMs are present, the batch with the lowest missingness may aid in fulfilling that threshold criterion for that feature, resulting in greater feature retention with M1. In an M2 approach, the MV threshold was set within each batch. Thus, if a feature failed to meet the threshold for even a single batch, the entire feature would be removed from the data. While this implies that an M2 strategy imputes only on the most confident data, it would be at the expense of an overall loss of coverage.

We then reduced the datasets to the common features retained between the imputed BEAMs and control data in each imputation approach, denoted as “M1.r” and “M2.r” (Figure 3). As expected, most features

found in the smaller dataset were found in the larger dataset as well. These reduced datasets were then used to make fairer evaluations between the BEAMs and control imputations. Although comparisons can be made between M1 and M2 with BEAMs in the same feature space, it would not hold much value as the large discrepancy in feature coverage should not be disregarded in the first place.

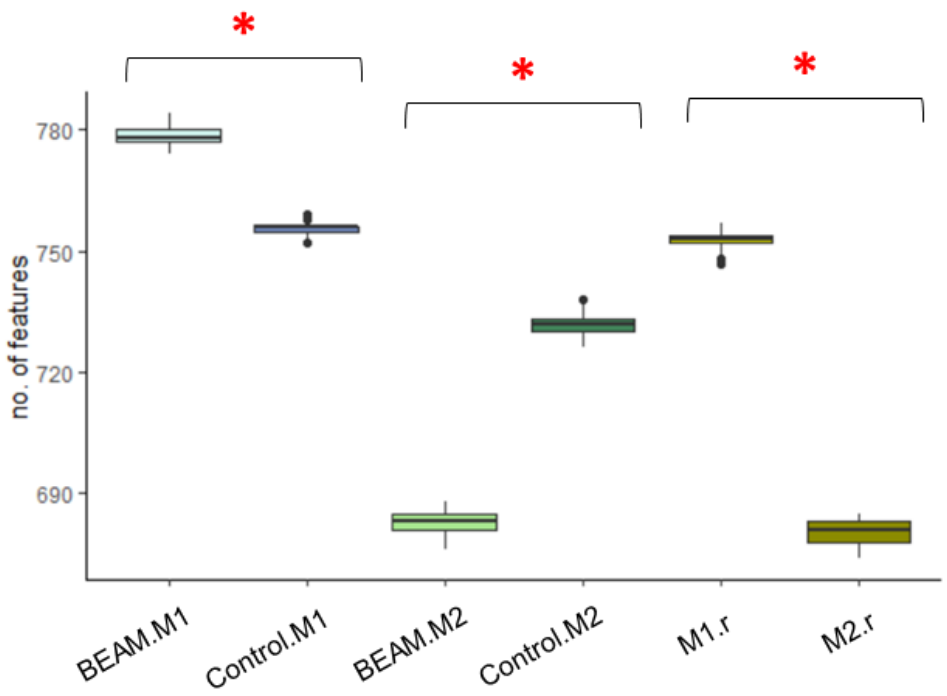


Figure 3. Number of features retained after features with 80% or more MVs were removed in data with moderate BEAMs (20:40% MVs). M1 and M2 with an "r" suffix refer to common features between BEAMs and Control within the same imputation strategy. Asterisks indicate significant difference (p < 0.05).

BEAMs worsen imputation accuracy regardless of imputation strategy (M1 or M2)

A common strategy to evaluate imputation accuracy is the use of the NRMSE metric, which measures the similarity between the imputed values and the true values. A lower NRMSE value indicates higher similarity and thus greater imputation accuracy.

In the initial comparisons between the BEAMs and Control data, the NRMSE rankings appeared to mirror that of the number of features (Figure 4A). Since this was not a fair comparison, we assessed the BEAMs and Control data within the same imputation strategy reduced to the same features (Figure 4B). The subsequent NRMSE analysis then showed that MVI on data with BEAMs significantly reduced imputation accuracy for these selected features in both M1 and M2 strategies, with a greater extent in M1.

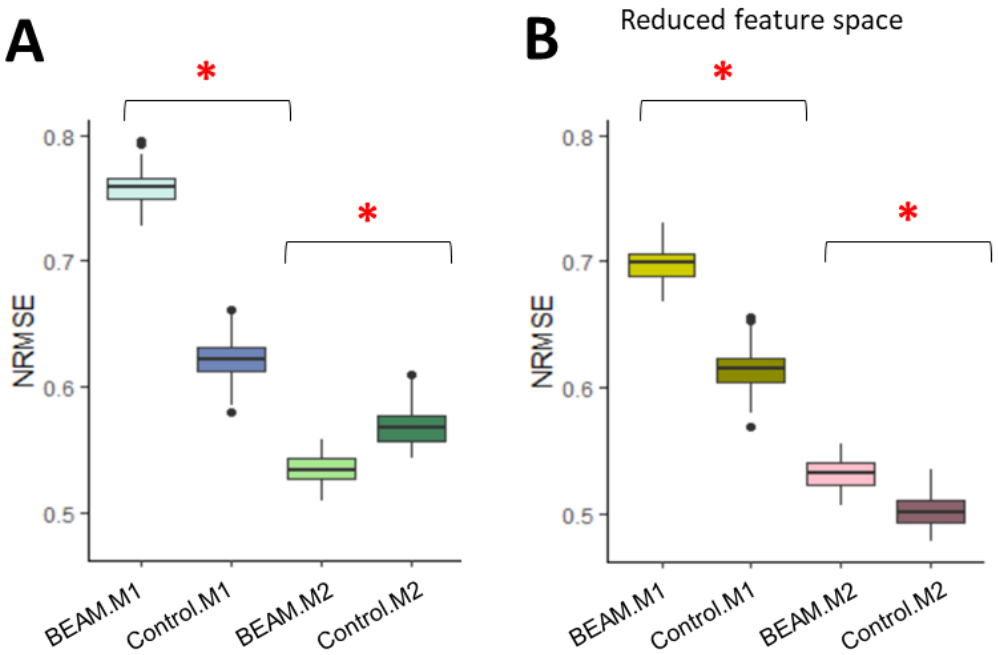


Figure 4. NRMSE outcomes after MVI with and without moderate BEAMs (20:40%). Asterisks indicate significant difference (p < 0.05). (A) Data are analyzed within their own feature space. (B) Data were reduced to the same features within the same imputation strategy (i.e., M1, M2).

In the study, after the 100th iteration, we examined the additional M1 features that were only retained in the BEAMs data. A comparison of the values from the true and imputed datasets before BE correction revealed that the imputation was strongly guided by the batch with lower missingness (Figure 5). This is expected, as when observations are insufficient in one batch, the imputation would inevitably draw information from another batch to form an estimate. Thus, although a BEAMs-infested data may retain more features with an M1 approach, the imputations for these features would be of poor quality and may even mislead analysis.

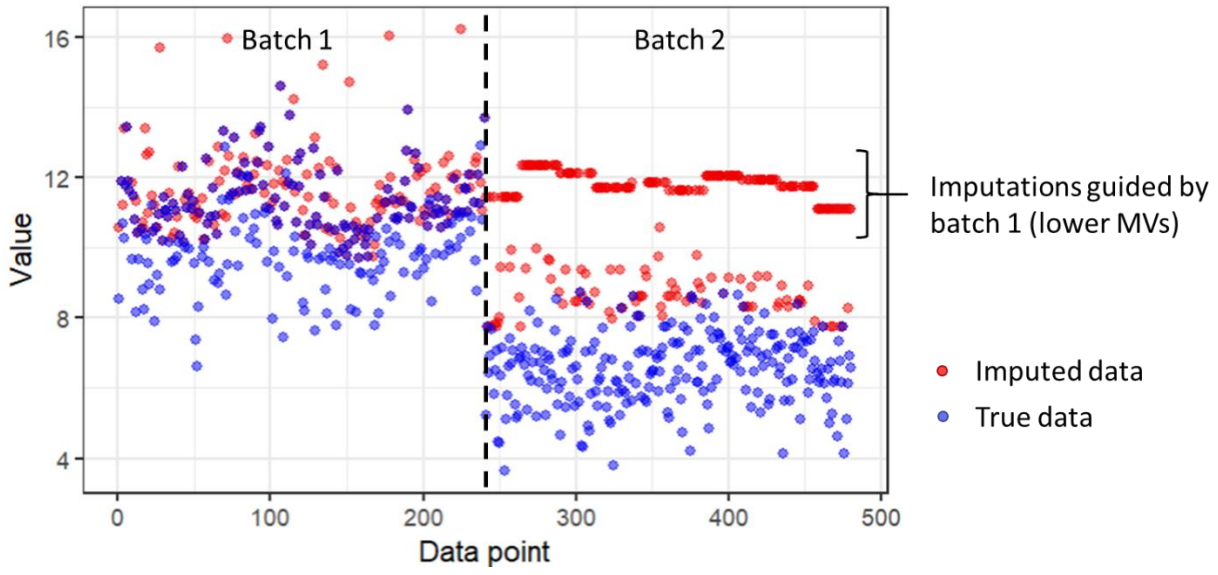


Figure 5. Comparison of values before batch effect correction, in features retained only in M1 with BEAMs data. Red points refer to data points from the imputed data and blue points refer to data points from the original true data. The horizontal intersect separates the data points by their batches.

Poorer imputation accuracy in BEAMS is generally accompanied by worsened True Positive Rates

After imputation accuracy evaluation, we then determined the extent of which the imputations retained class information, through assessing their TPR and FPR (Figure 6). With moderate BEAMs (20:40% MVs), we found that M1 suffered in TPR (Figure 6A) but not FPR (Figure 6B). M2 with BEAMs without feature reduction showed increased TPR (Figure 6A1), though this was at the expense of an increased FPR as well (Figure 6B1). This, however, might have been a result of the discrepancy in feature space, as these differences were abolished after reduction to the same features (Figures 6A2 & 6B2).

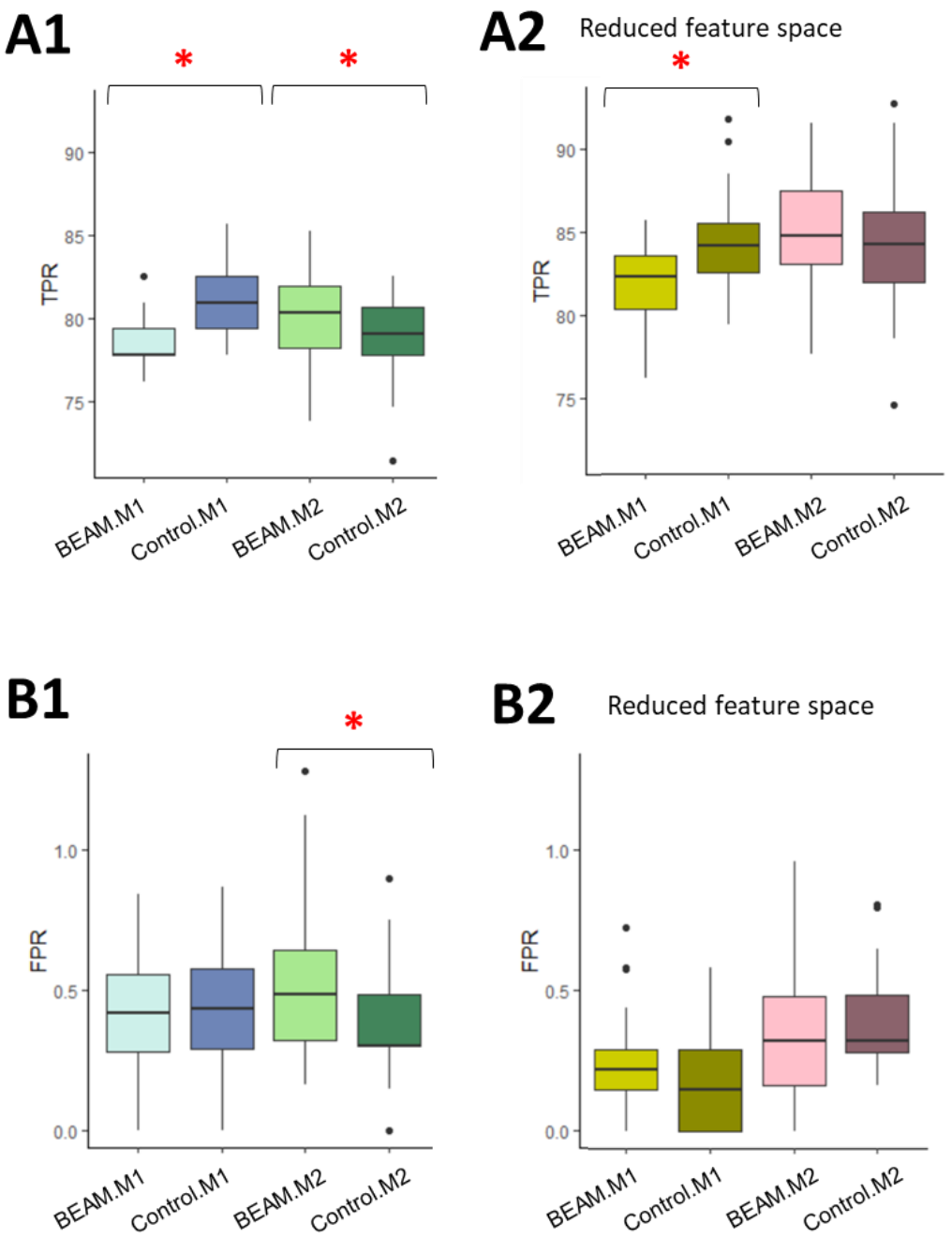


Figure 6. (A) TPR and (B) FPR outcomes after MVI with and without moderate BEAMs (20:40% MVs). Asterisks indicate significant difference (p < 0.05). (1) Data are analyzed within their own feature space. (2) Data were reduced to the same features within the same imputation strategy (i.e., M1, M2).

211 In a more drastic BEAMs scenario (10:50% MVs), the consequences of MVI on data with BEAMs becomes
212 apparent in both M1 and M2 strategies, with every BEAMs scenario producing worse TPR outcomes than
213 the controls (Figure S4). Whilst the presence of BEAMs did not appear to inflate the FPR with an M2
214 approach, an M1 approach with BEAMs produced significantly more false positives.

215 Sample variance was more closely preserved when MVs are evenly distributed than in 216 BEAMs

217 To gain a better understanding of why MVI on data with BEAMs was performing worse, we then assessed
218 the intra-sample variance of the dataset from the final iteration, reduced to the same features between
219 BEAMs and non-BEAMs data. Although imperfect, the sample variances obtained in the control data were
220 visibly closer to the original than that of the BEAMs data (Figure 7). In the latter, sample variances
221 appeared to be lower, which might be an indication that the imputation was influenced by a certain batch,
222 causing the variance to shrink.

223 Earlier, we saw that M2 with moderate BEAMs showed only slight differences compared to the control in
224 the same feature space in TPR and FPR evaluations. When the severity of MV disproportion increased, the
225 consequences of BEAMs became much more apparent. Encouragingly, the sample variances showed
226 consistency with these results, as they were generally similar between the control and moderate BEAMs
227 scenario but deviated from the original when BEAMs were more severe (Figure S5).

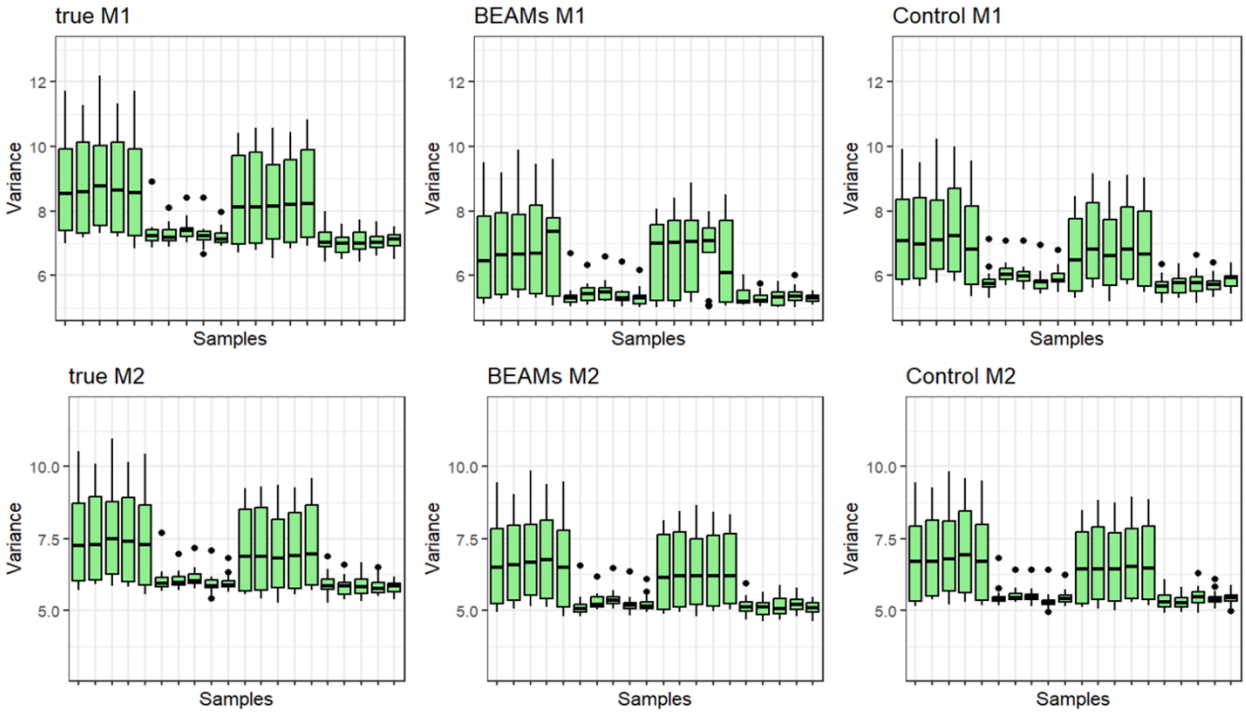


Figure 7. Intra-sample variances of datasets from the final simulation iteration with moderate BEAMs (20:40% MVs). Data were reduced to the same features within the same imputation strategy (i.e., M1, M2).

Case study: CPTAC study 6

A real-world example of BEAMs can be seen in the popular CPTAC study 6 dataset(18,19). When data from three machines were combined (therefore forming three batches), we saw that MVs were unevenly distributed across each batch (Figure 8). Specifically, the missing proportions stand around 64%, 33% and 45% in the respective batches, amounting to 47% MVs in total. Thus, this data contains a rather severe case of BEAMs. With real-world BEAMs, we cannot determine the performance of MVI compared to if BEAMs were not present. However, suppose M2 is expected to be the preferred method, determining the number of features retained between M1 and M2 would give an indication as to whether an M2 approach should be considered in the first place.

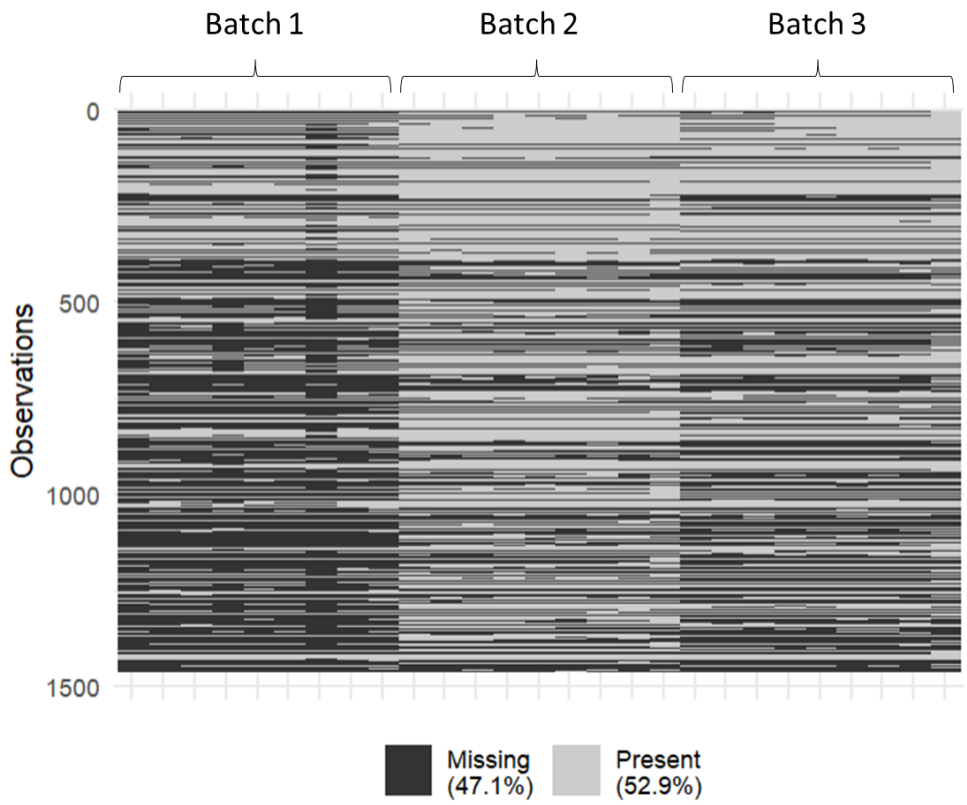


Figure 8. Heatmap of the CPTAC study 6 data from three machines. BEAMs were clearly observed as the machines (batches) displayed clear differences in MV proportions.

Following M1 and M2 MV filtering processes, we find that the resulting data retains only 1024 (30% loss) and 574 (60% loss) features respectively, out of the initial 1462 (Figure 8). In this initial data, 40 spike-ins were the features of interest. M1 retained a total of 30 of these features, whereas M2 retained only 9, demonstrating again the drastic loss of features that can arise with an M2 strategy on data with BEAMs.

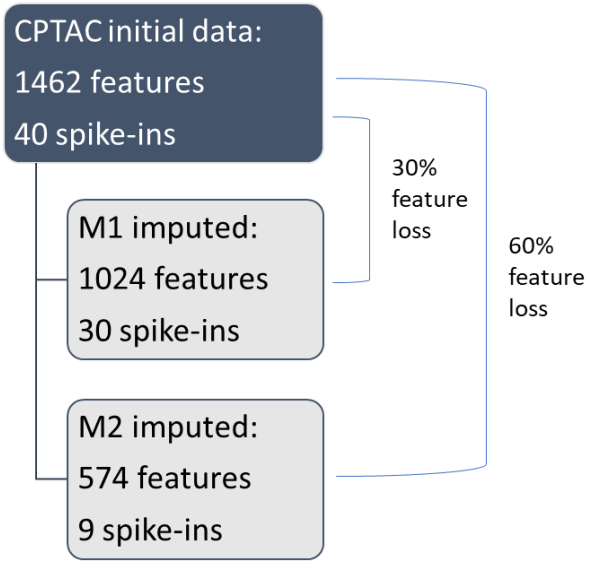


Figure 9. Feature characteristics of the CPTAC study 6 data before and after M1 and M2 MVI.

Discussion

BEAMs worsen KNN MVI regardless of feature retention extent

Post filtering with a predefined MV threshold, the BEAMs data consistently retained more features in M1 and less in M2. Despite having opposing outcomes on M1 and M2 approaches after MV filtering, the negative impact of BEAMs on MVI was clearly seen in both strategies when the same feature space was assessed. This implies that the performance of the MVI was not due to the size of data, but rather, the batch-biased distribution of MVs. Furthermore, after assessing the extra features retained in the BEAMs M1 data, a dilution of batch signal was clearly observed (Figure 5). As the retention of batch signal for BECA detection is important for BE correction, M1 imputation with BEAMs would undoubtedly result in poorer BECA outcomes. Hence, the increased number of features retained in M1 with BEAMs cannot be considered a benefit, as the MVI was not carried out effectively to begin with.

M2 strategies accommodate moderate cases of BEAMs at the cost of feature retention

In our analyses, we saw that an M2 approach when BEAMs were moderate (20:40%) produced similar performances to when MVs were equally distributed. As mentioned earlier, features that fail to meet the MV threshold in either batch are dropped entirely when using an M2 approach. This naturally reduces the extent of BEAMs after MV filtering, as the batch with more MVs would be more likely to contain features that fail to meet the criterion. However, this does not deal with BEAMs perfectly, as we saw that with severe BEAMs, M2 performance plummets in the same feature space as the control. In any case, when BEAMs are considered moderate, it would not be immediately unwise to use an M2 approach, but we must determine the importance of the features that would potentially be lost in the process. In some situations, the loss of features may skew the data and introduce bias, which can affect downstream analyses outcomes. Furthermore, there is the possibility that a biomarker of interest resides in one of these features but was removed as it contained only enough observations in one batch but not the other, as seen in our CPTAC case study (Figure 9).

Neither M1 nor M2 capable of dealing with BEAMs without drawback

When using an M1 approach, we can expect worsened analyses outcomes when BEAMs are present, whether moderate or severe. Whilst M2 may cope with BEAMs innately due to the MV filtering process, the feature loss must be deemed acceptable. Nevertheless, neither method can acceptably handle BEAMs in data without some drawbacks. Thus, a novel approach may be necessary to prevent or at least, reduce the negative impact that BEAMs have on the data.

Limitations and future research

Although we explained the rationale from an efficacy standpoint, we recognize the limitation of studying only one MVI (KNN) on the BEAMs phenomena. In reality, many MVIs exist, and their specific susceptibility

to BEAMs were not investigated in this study. However, we expect that most MVI methods will produce similar outcomes anyway.

Additionally, the MV mechanisms and proportions were arbitrarily determined at 3 MNAR to 1 MCAR, which although not baseless, is not an absolute representation of real-world data. Thus, although our results did demonstrate the negative impact of BEAMs on MVI, we might expect different outcomes with real world data possessing different characteristics. Nonetheless, our intention was to point out that BEAMs can lead analysis astray.

Finally, performing MVI on data with BEAMs with either a global (M1) or batch-sensitized (M2) approach both results in poor outcomes. Hence, future research may be focused on developing a BEAMs-specific strategy that may ideally preserve the performance of M2 with the feature retention of M1.

Conclusion

In summary, this study investigated the consequences of Batch-Effect Associated Missing Values (BEAMs) on data when performing Missing Value Imputation (MVI) using K-Nearest Neighbors (KNN) algorithm. The study found that the manifestation of BEAMs in data can lead to poor performance of MVI in both global (M1) and batch-sensitized (M2) approaches. It was also found that BEAMs lead to worsening performance in terms of either statistical analyses or the size of data, or both. Additionally, the results showed that batch effect correction alone does not eliminate the negative effects of BEAMs on MVI performance. The study highlights the need for the development of new strategies to handle BEAMs in data to improve the performance of downstream analysis.

Materials and Methods

All methods were performed in the R programming language (version 4.2.2) and in the RStudio integrated development environment (version RStudio 2022.12.0+353), following appropriate guidelines and regulations.

K-nearest neighbour (KNN)

There are various MVI methods that have been developed to accommodate different types of data. The simplest form of MVI is mean imputation, which replaces MVs with the mean of the corresponding feature. However, this method is not considered informative enough, so it is not used in this study. On the other end of the spectrum is the Random Forest (RF) algorithm(20,21), which is considered one of the most robust MVI techniques as it can perform well with almost all MV mechanisms(22). However, RF is computationally expensive and time-consuming because it requires numerous iterations.

A middle ground between mean imputation and RF MVI may be found with the K-nearest neighbor (KNN) algorithm. It is a simple machine learning method that replaces MVs with the mean of the K most similar data points(23). The nearest neighbors are typically determined by Euclidean distance, which is the distance between two points. Given two samples 1 and 2 measured on two variables x and y, this may be calculated by the formula:

$$Euclidean\ dist. = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

The KNN can be a good trade-off between the simplicity of mean imputation and the robustness of RF, while still being computationally efficient. KNN is not only fast and simple to perform, but it is also known for being one of the most effective MVI methods(24). An important parameter that influences the performance of KNN is the K value, which determines the number of nearest neighbors used for MV

327 estimation. In our preliminary studies, it was shown that using the square root of the batch size for K was
328 a conservative way to optimize self-batch imputation for retaining batch information to aid batch effect
329 correction. In this study, the same principle was employed, and the square root of the batch size was used
330 as the K value (√10 = 3). The KNN algorithm from the "impute" R package, version 1.72.2 was used(25). By
331 default, the algorithm performs mean imputation when the MVs exceed 50% in a row (in this case,
332 feature). To assess pure KNN imputation, the threshold was set to 100% so that no mean imputation was
333 introduced.

334 Imputation strategies (M1 & M2)

335 In this study, two MVI strategies, M1 and M2, were applied. M1 is a typical global imputation strategy,
336 where batch covariates are not considered. M2 is a batch-sensitized approach, in which MVI is performed
337 within each batch. In a previous study, it was found that an M2 approach was generally superior as it
338 improved BECA performance by retaining the batch covariate within the data for a more reliable
339 correction(17). An MV threshold of 80% was used in this study, meaning that features with 80% or more
340 MVs were removed from the data. This threshold was applied differently in M1 and M2 imputation
341 strategies. With M1, this threshold was set across all batches. With M2, features must fulfill this threshold
342 criterion in both batches to be retained in the data.

343 ComBat

344 Many Batch Effect Correction Algorithms (BECAs) exist, such as ComBat(16), Batch Mean Centering(26),
345 Harman(27), Surrogate Variable Analysis(28), each with different principles. ComBat is the most popular,
346 ranking highly in several BECA evaluation studies(29–32). It uses an Empirical Bayes approach to estimate
347 and standardize the means and variances between batches, effectively removing batch effects from the
348 data(16). The implementation of ComBat used in this study was obtained from the "sva" R package,
349 version 3.46.0(28).

350 ## Data simulations and real-world data

351 To simulate data that may represent real-world data, we employed the method described in Shah et al.
352 2017 which used blockwise correlation and autoregressive correlation to create data containing features
353 with correlations(33). The resulting dataset contained 20 samples with 900 features. For class and batch
354 effects, we used the same method described in our previous study(17). Class effects were simulated for
355 the first 60 features by a 1.5 times multiplication, and batch effects were simulated for the first 10 samples
356 (known as batch 1) by addition and multiplication. This results in a dataset with two batches, each
357 containing two classes. PCA was performed to visualize the artificial batch and class variances, before and
358 after batch effect correction (Figure S1). To simulate the missingness found in real world data, MVs were
359 then artificially inserted at a 3 MNAR to 1 MCAR ratio(11) at three different levels – (1) Control level: 30%
360 MVs in both batches; (2) Moderate BEAMs: 20% MVs in one batch and 40% MVs in the other; and (3)
361 Severe BEAMs: 10% MVs in one batch and 50% MVs in the other (Table 1). The MV simulations were
362 performed 100 times.

363 ***Table 1. The three levels of batch MV proportions simulated***

	Batch 1 MV%	Batch 2 MV%	Total MV%
Control	30%	30%	30%
Moderate BEAMs	20%	40%	30%
Severe BEAMs	10%	50%	30%

364

365 This study used a real-world example, the Clinical Proteomic Tumor Analysis Consortium (CPTAC) study 6
366 benchmarking dataset, which consists of a yeast proteome as the background with human UPS proteins
367 spiked in at different concentrations(18,19). We selected data from three machines at three spike-in
368 concentrations to form a dataset with 3 batches, containing a total of 27 samples. The missing proportions

369 are around 64%, 33% and 45% in the respective batches, amounting to 47% MVs overall (Figure 8). The
370 data is available at:

371 https://pdc.cancer.gov/pdc/browse?currentPath=%2525252FPhase_I_Data%2525252FStudy6&nonav=tr
372 ue.

373 Evaluation methods

374 The following evaluation methods apply only to simulated data.

375 Imputation accuracy

376 The accuracy of the MVI was assessed by the normalized root mean square error (NRMSE) metric, which
377 compares the similarity of the imputed values to the true reference values. A lower NRMSE value suggests
378 that the imputed values are more similar to the true values and therefore implies higher accuracy. The
379 formula for calculating NRMSE is shown below, where x and y refer to the corresponding true and imputed
380 values respectively, and sigma represents standard deviation.

381 $$RMSE = \sqrt{\frac{\Sigma_{i=1}^{n}(x_i - y_i)^2}{n}}$$

382 $$NRMSE = \frac{RMSE}{\sigma(x)}$$

383 Differential analysis

384 In the simulated data, class effects were introduced for the first 60 features. By performing a t-test on the
385 initial data, the features that are truly differential can be determined and we may then identify the true
386 positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). The True Positive Rates
387 (TPR) and False Positive Rates (FPR) can then be calculated from the imputed datasets. A higher TPR and
388 lower FPR indicate better performance. TPR and FPR are calculated as follows:

389 $$TPR = \frac{TP}{TP + FN}$$

390 $$FPR = \frac{FP}{FP + TN}$$

391

392 # Abbreviations

393 BE: Batch effect

394 BEAM: Batch effect associated missing value

395 BECA: Batch effect correction algorithm

396 CPTAC: Clinical Proteomic Tumor Analysis Consortium

397 FPR: False positive rate

398 KNN: K-nearest neighbors

399 LOD: Limit of detection

400 MAR: Missing at random

401 MCAR: Missing completely at random

402 MNAR: Missing not at random

403 MV: Missing value

404 MVI: Missing value imputation

405 NRMSE: Normalized root mean square error

406 PCA: Principal component analysis

TPR: True positive rate

Data availability

The codes used to perform analyses in this study can be found in the GitHub repository:

https://github.com/HarvardHui/BEAMs-paper.

The CPTAC study 6 dataset can be obtained from:

https://pdc.cancer.gov/pdc/browse?currentPath=%2525252FPhase_I_Data%2525252FStudy6&nonav=true.

Author contributions

HWHH performed analyses, developed figures and wrote the manuscript. WWBG conceptualized, supervised, provided critical feedback and wrote the manuscript.

Competing interests

The authors declare no competing interests, financial or otherwise.

Acknowledgements

WWBG acknowledges support from an MOE Tier 1 award (RS08/21).

423 # References

424 1. Goh WWB, Wang W, Wong L. Why Batch Effects Matter in Omics Data, and How to Avoid Them.
425 Trends Biotechnol. 2017 Jun 1;35(6):498–507.

426 2. Aittokallio T. Dealing with missing values in large-scale studies: microarray data imputation and
427 beyond. Brief Bioinform. 2010 Mar 1;11(2):253–64.

428 3. Matafora V, Corno A, Ciliberto A, Bachi A. Missing Value Monitoring Enhances the Robustness in
429 Proteomics Quantitation. J Proteome Res. 2017 Apr 7;16(4):1719–27.

430 4. Zhou L, Wong L, Goh WWB. Understanding missing proteins: a functional perspective. Drug Discov
431 Today. 2018 Mar;23(3):644–51.

432 5. Webb-Robertson BJM, Wiberg HK, Matzke MM, Brown JN, Wang J, McDermott JE, et al. Review,
433 Evaluation, and Discussion of the Challenges of Missing Value Imputation for Mass Spectrometry-
434 Based Label-Free Global Proteomics. J Proteome Res. 2015 May 1;14(5):1993–2001.

435 6. Kong W, Hui HWH, Peng H, Goh WWB. Dealing with missing values in proteomics data.
436 PROTEOMICS. 2022;22(23–24):2200092.

437 7. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to
438 imputation of missing values. J Clin Epidemiol. 2006 Oct 1;59(10):1087–91.

439 8. Karpievitch Y, Stanley J, Taverner T, Huang J, Adkins JN, Ansong C, et al. A statistical framework for
440 protein quantitation in bottom-up MS-based proteomics. Bioinformatics. 2009 Aug 15;25(16):2028–
441 34.

442 9. Rubin DB. Inference and missing data.

443 10. Petrazzini BO, Naya H, Lopez-Bello F, Vazquez G, Spangenberg L. Evaluation of different approaches
444 for missing data imputation on features associated to genomic data. BioData Min. 2021
445 Dec;14(1):44.

446 11. Liu M, Dongre A. Proper imputation of missing values in proteomics datasets for differential
447 expression analysis. Brief Bioinform. 2021 May 20;22(3):bbaa112.

448 12. Bijlsma S, Bobeldijk I, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, et al. Large-Scale Human
449 Metabolomics Studies: A Strategy for Data (Pre-) Processing and Validation. Anal Chem. 2006 Jan
450 1;78(2):567–74.

451 13. Goeminne LJE, Sticker A, Martens L, Gevaert K, Clement L. MSqRob Takes the Missing Hurdle:
452 Uniting Intensity- and Count-Based Proteomics. Anal Chem. 2020 May 5;92(9):6278–87.

453 14. Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, et al. Tackling the widespread
454 and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010 Oct;11(10):733–9.

455 15. Zhou L, Chi-Hau Sue A, Bin Goh WW. Examining the practical limits of batch effect-correction
456 algorithms: When should you care about batch effects? J Genet Genomics. 2019 Sep;46(9):433–43.

457 16. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical
458 Bayes methods. Biostatistics. 2007 Jan 1;8(1):118–27.

459 17. Sun PYQ, Goh WWB. Why Batch Sensitization is Important for Missing Value Imputation [Internet].
460 In Review; 2022 Feb [cited 2023 Jan 28]. Available from:
461 https://www.researchsquare.com/article/rs-1328989/v1

462 18. Paulovich AG, Billheimer D, Ham AJL, Vega-Montoto L, Rudnick PA, Tabb DL, et al. Interlaboratory
463 Study Characterizing a Yeast Performance Standard for Benchmarking LC-MS Platform Performance.
464 Mol Cell Proteomics MCP. 2010 Feb;9(2):242–54.

465 19. Tabb DL, Vega-Montoto L, Rudnick PA, Variyath AM, Ham AJL, Bunk DM, et al. Repeatability and
466 Reproducibility in Proteomic Identifications by Liquid Chromatography−Tandem Mass Spectrometry.
467 J Proteome Res. 2010 Feb 5;9(2):761–76.

468 20. Pantanowitz A, Marwala T. Missing Data Imputation Through the Use of the Random Forest
469 Algorithm. In: Yu W, Sanchez EN, editors. Advances in Computational Intelligence [Internet]. Berlin,
470 Heidelberg: Springer Berlin Heidelberg; 2009 [cited 2023 Jan 28]. p. 53–62. (Kacprzyk J, editor.
471 Advances in Intelligent and Soft Computing; vol. 116). Available from:
472 http://link.springer.com/10.1007/978-3-642-03156-4_6

473 21. Qi Y, Klein-Seetharaman J, Bar-Joseph Z. RANDOM FOREST SIMILARITY FOR PROTEIN-PROTEIN
474 INTERACTION PREDICTION FROM MULTIPLE SOURCES. In: Biocomputing 2005 [Internet]. Hawaii,
475 USA: WORLD SCIENTIFIC; 2004 [cited 2023 Jan 29]. p. 531–42. Available from:
476 http://www.worldscientific.com/doi/abs/10.1142/9789812702456_0050

477 22. Jin L, Bi Y, Hu C, Qu J, Shen S, Wang X, et al. A comparative study of evaluating missing value
478 imputation methods in label-free proteomics. Sci Rep. 2021 Jan 19;11(1):1–11.

479 23. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, et al. Missing value estimation
480 methods for DNA microarrays. Bioinformatics. 2001 Jun 1;17(6):520–5.

481 24. Hrydziuszko O, Viant MR. Missing values in mass spectrometry based metabolomics: an
482 undervalued step in the data processing pipeline. Metabolomics. 2012 Jun;8(S1):161–74.

483 25. Hastie T, Tibshirani R, Narasimhan B, Chu G. impute: impute: Imputation for microarray data
484 [Internet]. Bioconductor version: Release (3.16); 2023 [cited 2023 Jan 29]. Available from:
485 https://bioconductor.org/packages/impute/

486 26. Hastie T, Tibshirani R, Narasimhan B, Chu G. pamr: Pam: Prediction Analysis for Microarrays
487 [Internet]. 2019 [cited 2023 Jan 29]. Available from: https://CRAN.R-project.org/package=pamr

488 27. Oytam Y, Sobhanmanesh F, Duesing K, Bowden JC, Osmond-McLeod M, Ross J. Risk-conscious
489 correction of batch effects: maximising information extraction from high-throughput genomic
490 datasets. BMC Bioinformatics. 2016 Sep 1;17(1):332.

491 28. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and
492 other unwanted variation in high-throughput experiments. Bioinformatics. 2012 Mar 15;28(6):882–
493 3.

494 29. Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, et al. A comparison of batch
495 effect removal methods for enhancement of prediction performance using MAQC-II microarray
496 gene expression data. Pharmacogenomics J. 2010 Aug;10(4):278–91.

497 30. Konstantinopoulos PA, Cannistra SA, Fountzilas H, Culhane A, Pillay K, Rueda B, et al. Integrated
498 Analysis of Multiple Microarray Datasets Identifies a Reproducible Survival Predictor in Ovarian
499 Cancer. PLoS ONE. 2011 Mar 29;6(3):e18202–e18202.

500 31. Kupfer P, Guthke R, Pohlers D, Huber R, Koczan D, Kinne RW. Batch correction of microarray data
501 substantially improves the identification of genes differentially expressed in Rheumatoid Arthritis
502 and Osteoarthritis. BMC Med Genomics. 2012 Jun 8;5(1):23.

503 32. Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, et al. Removing Batch Effects in Analysis of
504 Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. PLoS ONE. 2011 Feb
505 28;6(2):e17238.

506 33. Shah JS, Rai SN, DeFilippis AP, Hill BG, Bhatnagar A, Brock GN. Distribution based nearest neighbor
507 imputation for truncated high dimensional data with applications to pre-clinical and clinical
508 metabolomics studies. BMC Bioinformatics. 2017 Feb 20;18(1):114.

509