**Manuscript Draft**

|  |  |
| --- | --- |
| Full title | Doing mega-analysis right: class-specific quantile normalization coupled with ComBat as most optimal approach to remove batch effects |
| Abstract | **Background**  Mega-analysis involves pooling data from different experimental sources. This inevitably results in batch effects, which are technical bias that confound with inherent biological variability and can cause reproducibility issues. Although there are many batch effect removal methods (BERMs) used conventionally, careless application of these approaches at best do not properly remove batch effect and at worse introduce false positives and false negatives.  **Results**  We applied seven BERMs on a mega-dataset created by combining four ageing-related datasets. We demonstrate via principal component analysis (PCA) that the combination of class-specific quantile normalisation (CSQN) with subsequent ComBat performs the best at removing batch effects while retaining biological variation. Bootstrapping and Jaccard scoring show that this method is highly robust. Meaningful features were obtained from feature selection conducted on the dataset treated with this approach.  **Conclusion**  Although BERMs are commonly used to remove batch effects in mega-analysis studies, they may not eradicate batch effects properly and hence result in unreproducible results. We thus recommend CSQN with downstream ComBat as the best and most reliable approach for processing of datasets with batch effects. |
| Author | Joan Jong Jing Yi |

**Table of Contents**

[Introduction 3](#_Toc55307400)

[Materials and Methods 6](#_Toc55307401)

[Results 13](#_Toc55307402)

[Discussion 22](#_Toc55307403)

[Conclusion 24](#_Toc55307404)

[References 25](#_Toc55307405)

1. Introduction

High-throughput omics technologies such as genomics, transcriptomics and proteomics allow deep assaying (identification and quantification) of biological moieties such as DNA, RNA and protein in cellular tissues. By being able to identify and quantify biological moieties within samples, comparative analysis between multiple tissues can be conducted. This facilitates detection of differential expression states amongst biological moieties, which in turn informs on what makes these tissues unique or different.

However, such high-throughput assays are not without issues. Typically, high-throughput assays are underpowered and suffer from ‘curse of dimensionality’ issues. That is, there are many more biological moieties being measured than actual samples, making it difficult to correctly estimate meaningful variations, which in turn causes both Type 1 and Type 2 errors. Issues with Type 1 errors are of particular concern in recent times, where study reproducibility has been shown to be highly problematic. Ioannidis et al. found that in scientific fields with smaller studies, such as research of molecular predictors, the positive predictive value (PPV) is low [1]. This means that research findings are less likely to be true, compared to the findings in large-scale randomised trials, due to the small statistical power of the data used. Other factors that can increase the possibility that a published study is unreproducible include small effect sizes, data dredging and conflict of interests. Many metascience studies from different fields concur that reproducibility is subpar [2] [3]. The severity of this problem has led to the inception of the Reproducibility Project led by Brian Nosek, which is a collaborative attempt to replicate previously published psychology studies [4] and has been expanded to also include cancer studies [5].

Study reproducibility issues are at times also referred to as the ‘Winner’s Curse’ phenomenon [6]. This describes situations where a particular moiety is observed to be significant in one study but disappears in other studies. While such moieties are likely to be false positives in themselves, this is not always true; subsequent studies may be simply underpowered or may suffer from random sampling bias, technical bias or covers a different subpopulation for which a different set of causes are relevant.

With decreasing running costs and the ever-growing collection of datasets submitted to online repositories, data pooling is a plausible approach towards improving study reproducibility. The idea is simple: to improve the coverage on the study population and to increase statistical power. This strategy works on expanding the sample size by consolidating data from similar studies. Statistical power is improved relative to that in each individual study, thus making it possible to detect subtle trends otherwise undetectable in smaller studies. Conversely, suppose a signal is not small but suffers from high variability, increasing sample size will also help to stabilize the signal. Finally, data pooling is more cost-effective compared to conducting larger randomized trials.

There are two main approaches for data pooling in biomedical research: meta-analysis of aggregate patient data and mega-analysis based on individual patient data. Meta-analysis involves combining the summary results of separate studies published in the medical literature, while mega-analysis requires the collection of raw individual-level clinical and biological data from multiple studies. The latter strategy demands considerably more resources, time and energy devoted to acquiring and processing raw data from researchers responsible for the studies examined. Thus, it is unsurprising that meta-analysis is typically more popular than mega-analysis in biomedical research. Despite its popularity, meta-analysis may not be the best method of data pooling and is considered less reliable than mega-analysis. While both methods increase statistical power, mega-analysis confers unique advantages such as permitting data-checking, data-updating and exploration of heterogeneity between subgroups at the patient level. Mega-analysis has thus been considered since the early days of systemic review to be the ‘gold standard’ despite the arduous effort that entails this strategy.

Mega-analysis inevitably generates batch effects

In mega-analysis, the combination of data from different studies will inevitably generate technical variations such as general noise and batch effects. In fact, some studies deliberately join technical replicates to simulate batch effects. Study-dependent differences in sample preparation, types of microarray chip and experiment protocols may contribute to shifts in measurements of samples and are observed as batch effects. As batch effects are idiosyncratic in nature with no biological basis, they can have profound effects in analysis outcome. In the relatively benign case, they simply lead to more variability and thus decreased power in detecting a true biological signal. This certainly would have implications for reproducibility. However, the more concerning repercussions are false effects that may be induced due to confounding between batch effects and the results, leading to inaccurate clinical conclusions.

Batch effects may manifest non-uniformly

Further complicating batch effects is their possibly non-uniform nature. Leek et al. examined datasets from 9 superficial transitional cell carcinoma (sTCC) studies and found that, across these datasets, different proportions (32.1-99.5%) of measured features showed statistically significant correlation with processing date, regardless of biological phenotype [7]. This suggests that batch effects within a study may be non-uniformly distributed among features. Furthermore, the proportion of features associated with batch effects is not homogeneous among different studies. This non-uniform nature can make batch effects challenging to deal with.

Synergy between normalization techniques and batch effect removal algorithms not adequately explored

To tackle batch effects, several approaches can be used. Normalization is an intervention that encompasses many different techniques, each with its own assumptions about the data distribution. The goal of normalization is to remove or minimize technical variation in a dataset. Generally, normalization techniques aim to re-distribute signal intensities across samples so that the distributions of signal intensities for all samples are the same. For example, quantile normalization (QN) ensures that all samples have the same mean and standard deviation, while linear scaling (also known as min-max scaling) standardizes the range of the signal intensities. Other common normalization techniques include Z-normalization and rank-scaling. Such generic normalization techniques typically do not work well alone without prior modification. It has been shown that class-specific QN, where QN is conducted on dataset split by phenotype classes, outperforms whole-data QN by 1.5 times in removing batch effects while preserving biological variation. There are also algorithms developed specifically to deal with batch effects known as batch effect-correcting algorithms (BECAs), such as ComBat [8] and Surrogate Variable Analysis. Several mega-analyses have attempted to combine generic normalization techniques with BECAs as an approach to remove batch effects. Notably, Tylee et al. utilized Z-normalization as the first cleaning step prior to ComBat [9] and Muller et al. evaluated that combining whole-data QN with ComBat serves as the best approach for batch effect removal [10]. Zhao et al, however, demonstrated that whole-data QN is a flawed approach, and that accounting for other co-variates such as class and batch information, can dramatically improve reproducibility and also overcome technical bias such as batch effects. As for synergies between normalization and BECAs, there are some contextual dependencies as well. Hence, further research exploring how normalization methods, especially improved ones taking into account important confounding co-variates such as the class and batch factors, and their interplay with BECAs is needed.

Using ageing datasets for exploring synergies between normalization and Batch Effect Correction Algorithms (BECAs)

Ageing and longevity are of increasing concern in many parts of the world, especially in developed countries where populations are becoming progressively more aged. By 2030, it is projected that 34 countries will be classified as ‘super-aged’, with more than 20% of the population being over 65 years old [11]. Such a phenomenon may over-burden healthcare systems and decrease economic growth. Thus, there is a pressing need for biotechnological research to devise novel drugs and healthcare products for healthier ageing. Age is also a major risk factor influencing the development of many illnesses, such as Alzheimer’s disease. These concerns culminate in the need to gain novel insights into the ageing process and identify potentially crucial genes and biomarkers.

Unfortunately, the relationship between the ageing process and its molecular etiology is not fully understood. Gene expression studies have been typically noisy, oftentimes with few genes found to be differentially expressed with age, and even fewer overlapping genes across tissues and species. The noise may originate from the accumulation of random mutations leading to aberrant activation or repression of promoters, or it may be age-linked, for example in the case of DNA damage in human brain samples [12]. In order to accurately distinguish conserved age-linked gene expression patterns from idiosyncratic mutations, there is a need to increase statistical power by expanding sample size. Mega-analysis is therefore a crucial approach to examine age-associated gene expression patterns and should be conducted properly; batch effects arising from combining data from different studies must be well-mitigated. This makes ageing datasets an appropriate and relevant benchmark for evaluating batch effect removal in mega-analysis.

In this study, we will explore the batch effects present in mega-analysis of ageing data using Principal Component Analysis (PCA). As there have been many prior benchmark studies indicating that the BECA, ComBat, is the best performer across a wide variety of scenarios, we will not consider other BECAs. Instead, we will determine whether there is synergy between various normalization techniques with ComBat. And recommended the best combination using reproducibility and PCA as the primary indicators of success.

Materials and Methods

Data mining and filtering

We built our corpus using datasets retrieved from the Gene Expression Omnibus (GEO) repository. The query passed into the search bar was “aging” [All fields] OR “ageing” [All fields] AND “vastus lateralis” [All fields] AND “homo sapiens” [Organism]. Only datasets that were generated from expression profiling by microarray were considered. These datasets were required to include patients with a wide age range. This was to ensure that both young and old patients were represented. Datasets that used myoblast cells from the vastus lateralis muscle were excluded as the gene expression levels would be confounded by the degree of differentiation of these cells. As the contributions of normalization methods towards batch correction was being evaluated, the datasets also had to be free of any prior normalization. However, many datasets available for download were normalized and they did not include the processed but non-normalized version. After filtering through the search results, GEO Series GSE98613, GSE87105, GSE83352 and GSE40645 were found to meet all inclusion criteria. Their non-normalized data were directly downloaded from the Supplementary Files table in GEO.

For the datasets, we only included samples which did not undergo intervention or were controls. The participants ranged from 28 to 89 years old. We separated them into 2 classes, young and old, using a threshold age at 50 years old. The summary of the metadata of the datasets is compiled in ***Table 1***. Each GSE dataset is considered a batch and we have included the respective batch labels in ***Table 1***. We included a total of 110 samples; 44 samples were allocated to the young class and 66 samples were assigned to the old class.

**Table 1:** Table summarising the metadata for each dataset used in our mega-analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Batch label** | **GSE Name** | **Platform** | **Total samples** | **Young** | **Old** | **Female** | **Male** |
| Raz | GSE40645 | Illumina HumanHT-12 V3.0 | 29 | 14 | 15 | 14 | 15 |
| Hubal | GSE83352 | Illumina HumanHT-12 V4.0 | 42 | 16 | 26 | 23 | 19 |
| Mercken | GSE87105 | Illumina HumanHT-12 V4.0 | 16 | 10 | 6 | 0 | 16 |
| Gonzalez | GSE98613 | Illumina HumanHT-12 V4.0 | 23 | 4 | 19 | 8 | 15 |
| Total | | | 110 | 44 | 66 | 45 | 65 |

Mean-probe processing of expression data

After the downloaded files were decompressed, columns containing signal intensities for each sample were extracted. A new matrix containing the expression data was generated for each GSE, where the columns represented the samples in the dataset and the rows represented the microarray probes used to test for gene expression.

The probe IDs linked to each row were then substituted with symbols for the gene that each probe detects for. These probe IDs and their respective gene symbols were retrieved using the IlluminaHuman4 library from the Bioconductor package [13]. For probe IDs that did not have any matching gene symbol, the corresponding row in the GSE was removed. It was discovered that there were multiple probes detecting for the same gene and thus data compression was performed to ensure that every row in each GSE corresponded to one gene. All the GSEs were then combined to create a mega-dataset.

Benchmarking methods

In this study, we examined 7 different batch effect removal methods (BERMs). These methods were either normalisation, ComBat, or a combination of both (***Table 2***).

**Table 2:** Table showing the 7 batch effect removal methods (BERMs) examined in this study. The ‘None’ approach is a negative control that consists of neither normalisation nor batch effect correction algorithm (BECA).

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **BERM Label** | **Normalisation** | **BECA** |
| 1 | None | - | - |
| 2 | QN | Quantile normalisation | - |
| 3 | ZN | Z-normalisation | - |
| 4 | ComBat | - | ComBat |
| 5 | CSQN | Class-specific quantile normalisation | - |
| 6 | QN + ComBat | Quantile normalisation | ComBat |
| 7 | ZN + ComBat | Z-normalisation | ComBat |
| 8 | CSQN + ComBat | Class-specific quantile normalisation | ComBat |

Z-normalization

We conducted Z-normalisation (ZN) in a gene-wise manner. The mean of the expression values for each gene was standardised to zero and the variance across all genes were equalised.

Quantile and class-specific quantile normalization

The quantile normalization (QN) used in this study is a simple procedure commonly used in microarray studies. We first ranked the gene in each sample by magnitude and calculated the average value of the genes in the same rank. The values for all the genes in the rank were then substituted with this average value. The final step involves reordering the genes in each sample in their original order.

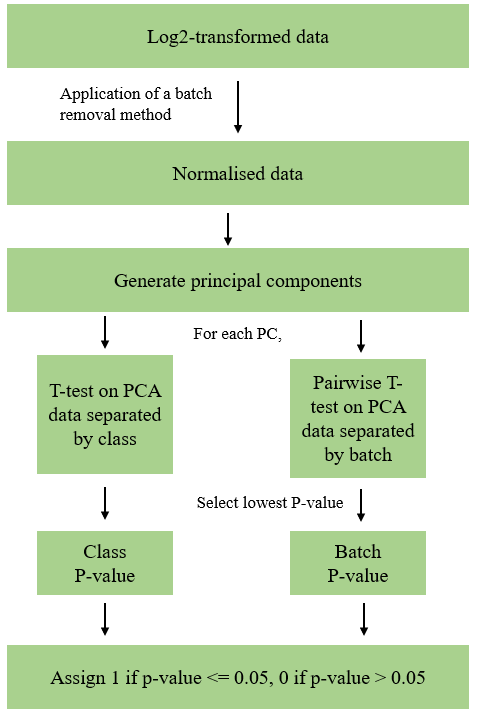
Class-specific QN (CSQN) is a modified version of QN. Instead of doing QN on the whole dataset, we first divided the samples into their respective classes before conducting QN on each class separately. This approach has been demonstrated to be effective in preserving class-specific information, and exhibits some robustness against batch effects.

ComBat

ComBat is an empirical Bayes batch effect correction algorithm that combines both location and scale adjustment to remove batch effect [8]. The mean and variance are estimated for each batch and gene independently. ComBat requires the specification of batches in the data.

PCA analyses for checking batch effects

The combined dataset that was cleaned and harmonised was subject to log2-transformation. Subsequently, one of the seven BERMs was used to normalise the dataset. Principal Component Analysis (PCA) was then conducted on each of the normalised datasets (**Figure *1***). This analysis was used to determine how much class or batch effect contribute most to the variation in the dataset. It is an indication of how effective a certain batch removal method is at alleviating batch effect. In each dataset, the eigenvalues for each principal component (PC) was separated by class and Student’s T-test was done between the two classes based on the null hypothesis that there is no significant difference in the means and any difference was due to chance. A similar procedure was performed for the same PCs separated by batch. Since there are more than two batches, pairwise T-test was done, and the lowest p-value was considered. The p-values for the T-test were collated and then binarized; all p-values that are less than or equal to 0.05 were denoted as 1 and the remaining were substituted by 0 (**Table *3***). Since the PCs with higher rankings capture more variation in the dataset, when most of them are correlated with batch, it is indicative that the variation within the dataset is contributed mostly by batch effect. Thus, PCA can be used to quantify the amount of batch or class effect in the normalised dataset as a way of evaluating the performance of BERMs.



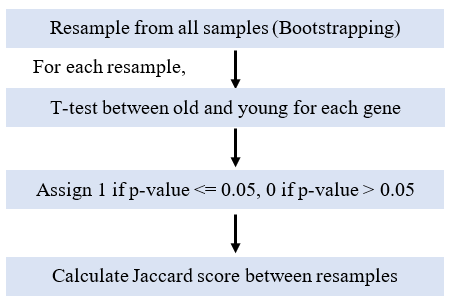
**Figure 1:** Flowchart of the PCA process to determine the amount of batch and class effects present in a normalised dataset.

**Table 3:** An example output from PCA conducted on the log2-transformed dataset. The higher the PC ranking, the more variation is captured in the PC. Batch effect contributes to the most variation in this example dataset since the top three PCs are correlated with batch.

|  |  |  |
| --- | --- | --- |
| **PC** | **Class** | **Batch** |
| 1 | 0 | 1 |
| 2 | 0 | 1 |
| 3 | 0 | 1 |
| 4 | 0 | 0 |
| 5 | 0 | 1 |
| 6 | 0 | 1 |
| … | … | … |

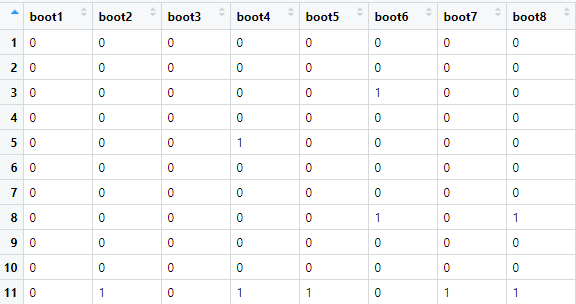
Bootstrap analysis

To determine which BERM is the most consistent at alleviating batch effect in terms of feature select reproducibility, bootstrapping and concomitant calculation of Jaccard score were conducted for each normalised dataset (**Figure *2***). For each BERM-treated dataset, we resampled from all samples and for each resample, we conducted T-test on the genes between the old and young classes. The p-values for each resample were binarized using a threshold of 0.05 (**Table *4***).



**Figure 2:** Flowchart of the process involving bootstrapping and calculation of Jaccard scores. Bootstrapping was conducted 1000 times for each normalised dataset.

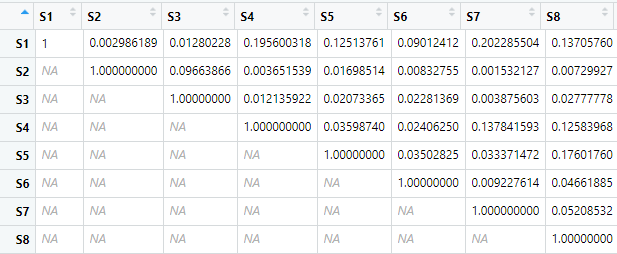
**Table 4:** A portion of the output after binarization of p-values in each resample. Each column represents a bootstrap sample while each row represents a gene. When a gene is labelled 0 in a resample, it indicates that the gene was considered insignificant in contributing to the class effect in the resampled dataset.



After binarization of p-values, each bootstrap was compared to another sample to calculate how similar their binary patterns were. Jaccard score quantifies the similarity between two sets (**Equation *1***). The higher the score, the greater the similarity between the sets compared. When the bootstraps in a dataset are similar, it indicates that the identities of the significant and insignificant genes are similar after numerous rounds of random resampling. Such an outcome implies that the BERM used on the dataset is highly reliable in generating resamples that are reproducible. The output of the Jaccard calculation conducted on the dataset in **Table *4*** is shown in **Table *5***.

**Equation 1:** Calculation of Jaccard score, an index used to determine the degree of similarity between sample sets.

**Table 5:** A portion of the output after Jaccard scores were calculated for each comparison between the bootstrap samples. For example, the cell S2-S1 contains the Jaccard score for the comparison between bootstraps 1 and 2 in **Table 4**.

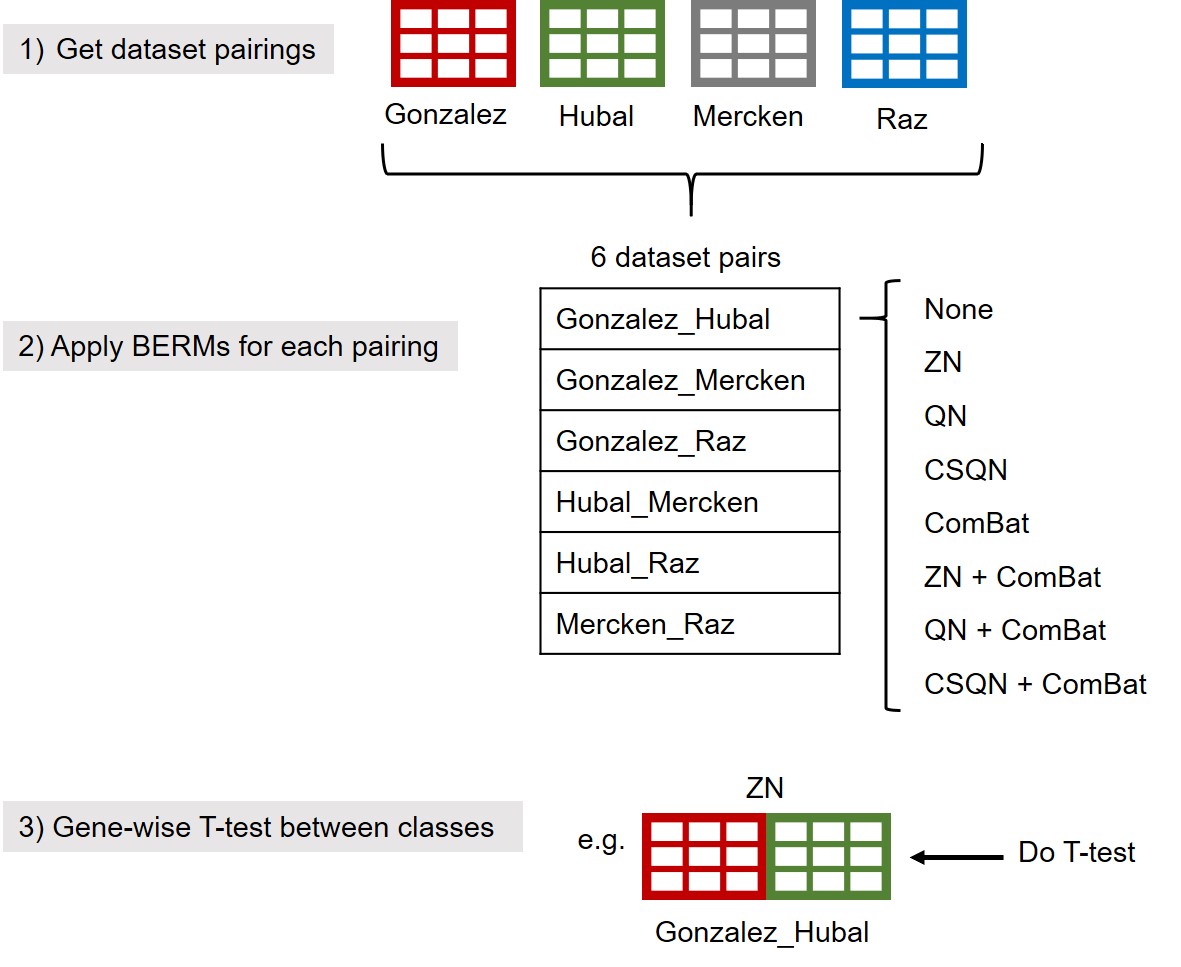


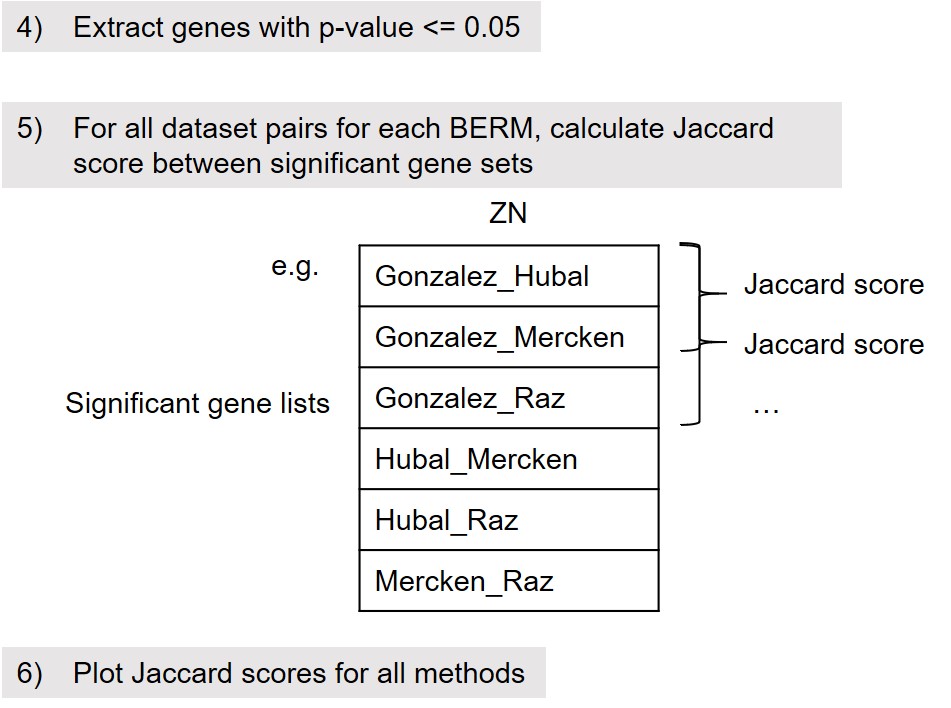
After the Jaccard scores were obtained for each normalised dataset, a violin plot was generated to compare the scores for each batch removal method. The method with the highest Jaccard score was deemed to result in the greatest reproducibility.

Using **Table *4***, the reproducibility of a batch removal method can also be determined by summing up the binarized p-values across all bootstraps for each gene. If there is a high number of genes that are consistently significant for the majority of the bootstraps, it means that the batch removal method is able to produce similar bootstraps even after numerous rounds of resampling. Furthermore, the class effect is implied to be more predominant than batch effect in the dataset since more genes are consistently significant across all bootstraps. This test can thus evaluate both the reproducibility of a batch removal method as well as gauge the degree of class effect in the dataset after normalisation by a certain method.

Pairwise cross comparison of batches

We investigated if a BERM is reproducible in terms of yielding stable feature selection results. 6 dataset pairs were created from pairing 4 batches. For each dataset pair, we applied the 7 BERMs and ‘None’ method to obtain 8 treated paired datasets. We then conducted T-test for each gene to test if the difference between old and young classes is significant. Genes that had p-values lower than 0.05 were extracted. For all the dataset pairs treated with the same BERM, we calculated the Jaccard scores between the pairs and plotted the results. This workflow is outlined in ***Figure 3***.

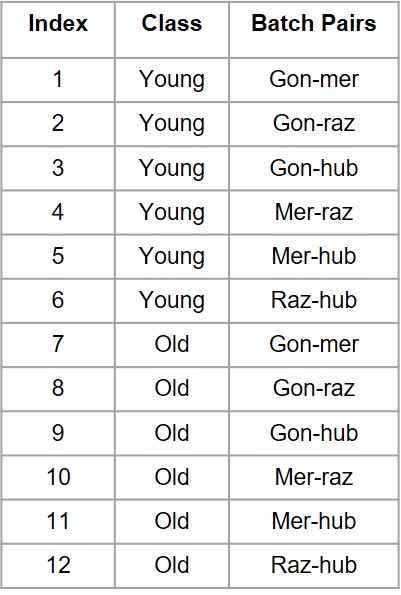




**Figure 3:** Workflow for paired dataset T-test and Jaccard analysis.

Probe-batch effect correlation checks

We checked for the existence of a conserved batch-correlated gene set by conducting PCA on 12 pairwise combinations of same-class-different-batch (SCDB) samples.



**Figure 4:** Table showing the 12 same-class-different-batch (SCDB) combinations used to check for a conserved batch-correlated set of genes.

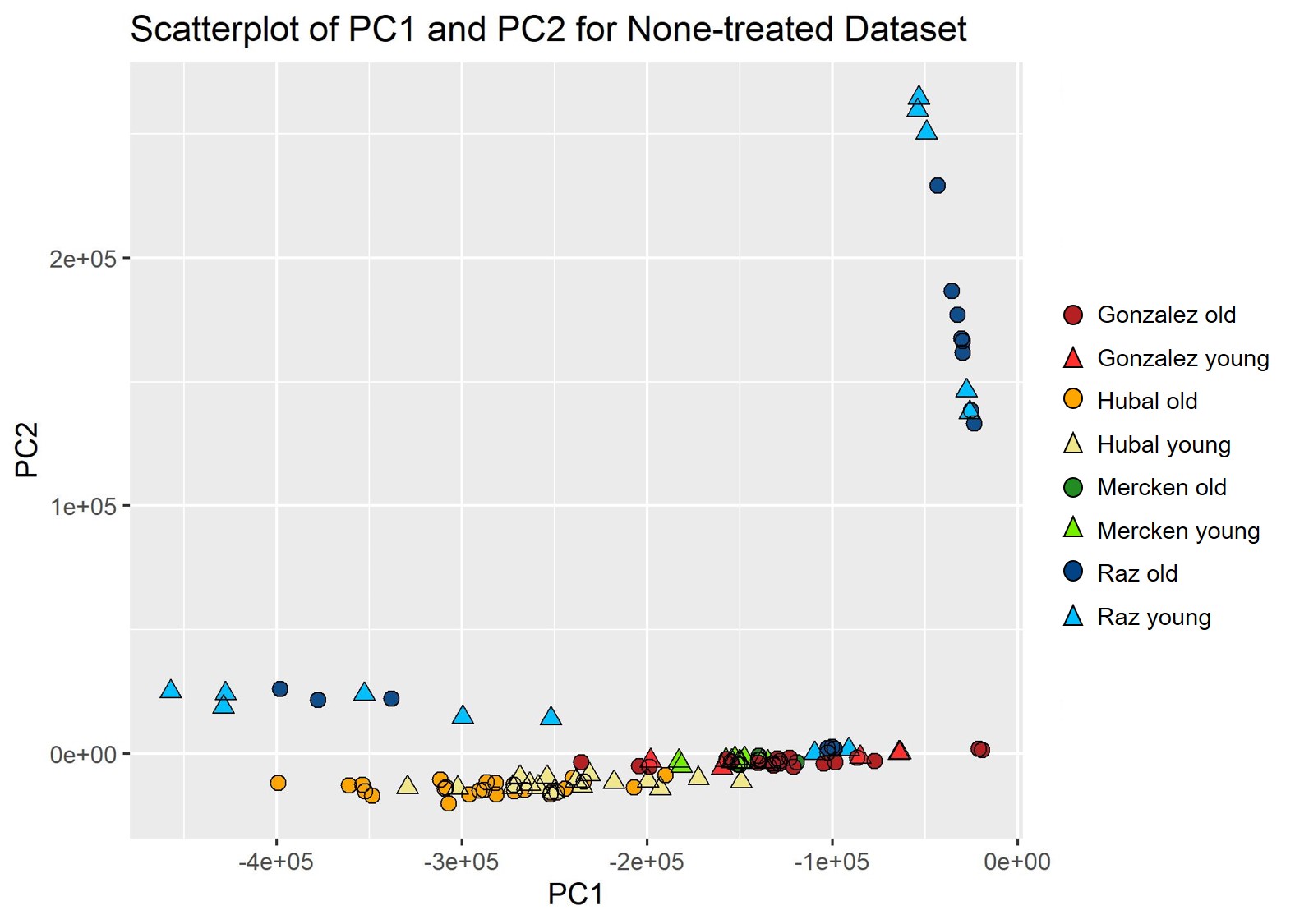
Since each SCDB combinations contain samples from the same class but different batches, the source of variation in the combination arises mostly from batch effect. The loadings of genes in each SCDB combination were calculated and we plotted these values to observe the distribution. Batch effects may be manifested non-uniformly if there appears to be a distinct group of genes with higher loadings.

We subsequently examined if there exists a correlation between the number of probes matching to a gene and the tendency of that gene to be associated with batch effect. We combined the sets of significant genes loading to batch effect from all 12 SCDB combinations. The frequency at which a gene appears in this combined gene set is calculated. We plotted this data against the number of probes matching to a gene to observe if there is a correlation.

Results

Batch effects are generated when datasets from different experiments are combined

We first plotted the PCA scatterplot for the log2-transformed combined dataset to observe the distribution of samples before BERMs were applied (***Figure 5***).



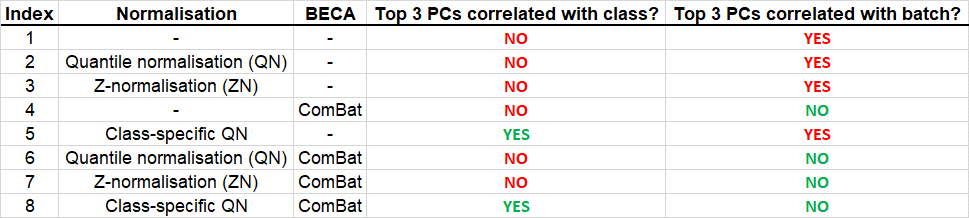
**Figure 5:** PCA scatterplot for None-treated dataset. Triangles represent samples classified as young and circles represent samples assigned as old. The colours denote the batch which each sample originates from.

The scatterplot shows that the samples are clustered by batches and the separation between classes is not distinct. This indicates the presence of batch effects in the dataset that arose from combining batches from different experiments.

CS quantile + ComBat gives the best outcome

In our methods section, we mentioned that PCA can be used to rank how strongly a data factor such as batch, class and gender influences the variation in a dataset. We applied eight different BERMs on our mega-dataset and evaluated the association with data factors using PCA (**Figure *6***).

The negative control (‘None’) yielded top 3 PCs correlated with batch and none with class. As expected, this result is similar for conventional normalisation methods i.e. ‘QN’ and ‘ZN’. Both approaches gave rise to datasets that were disproportionately correlated with technical bias. When the mega-dataset was treated with ComBat, even though the PCs correlated with batch were ideally relegated to lower ranks, the resultant dataset did not show improvement in preserving class effect. Combination with upstream QN or ZN remained ineffective at boosting the PCs correlated with class effect. This outcome corroborates with our expectation that an effective upstream normalisation is critical for the preservation of biological signal.



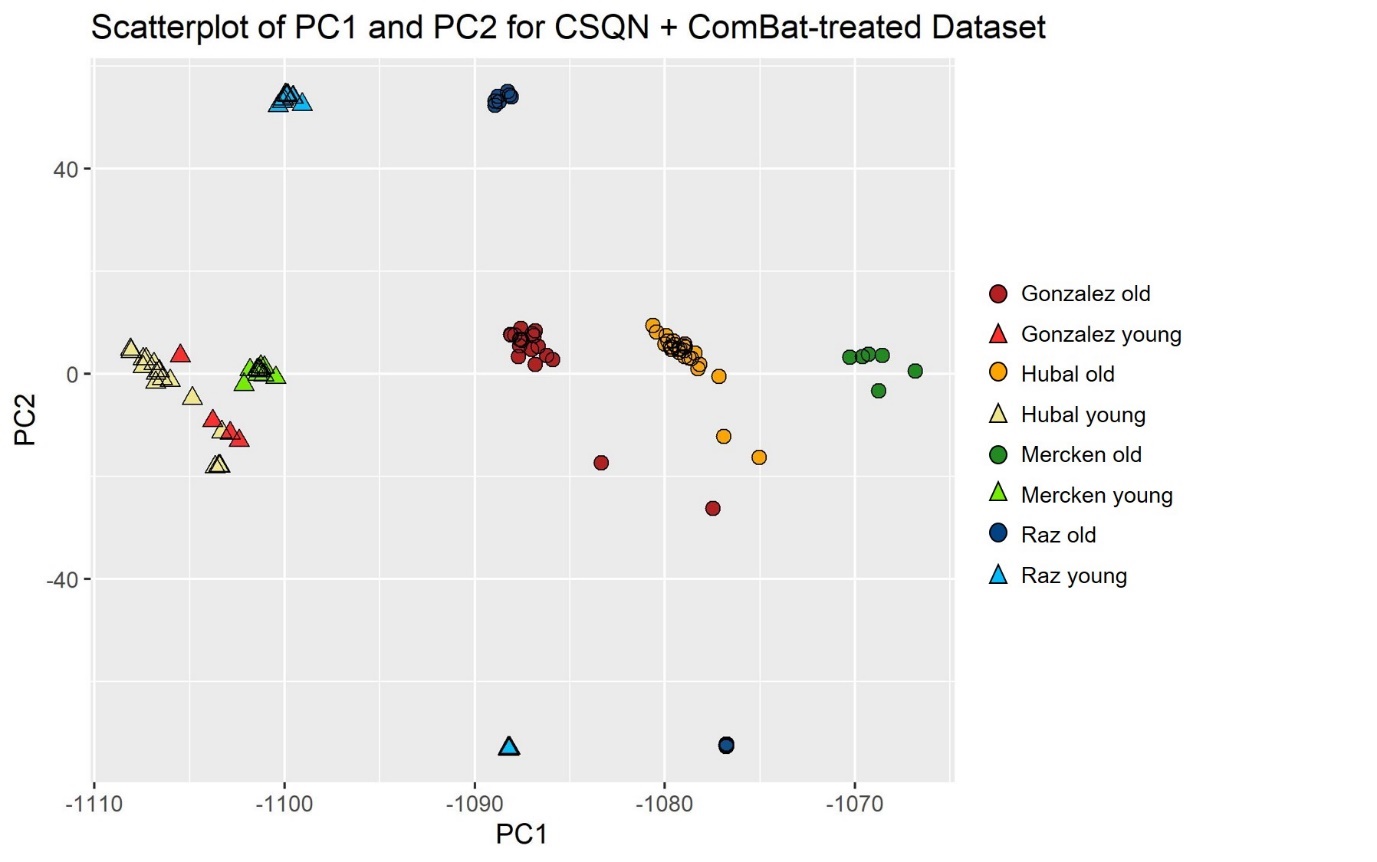
**Figure 6:** Summary of the performance of all eight methods. Method 1 is the negative control where no BERM has been applied to the dataset. Methods 2 to 5 do not combine the use of normalisation and ComBat as compared to methods 6 to 8. The top 3 PCs were examined and their correlation with batch and class factors was noted as either significant (‘YES’) or not significant (‘NO’) depending on a p-value threshold of 0.05. From the table, it appears that only the CSQN + ComBat approach managed to preserve biological variation while simultaneously alleviating the batch effect in the dataset.

Notably, CSQN was the only normalisation approach that yielded high association with class effect as class-correlated variation was observed among the top 3 PCs of CSQN-treated mega-dataset. However, a significant proportion of the variation in the dataset was also associated with batch effect. To combine the signal-preserving strength of CSQN and the batch-alleviating benefit of ComBat, we applied CSQN followed by ComBat on our mega-dataset. This approach appeared to be effective as class effect was apparent while batch effect was made insignificant in the top 3 PCs.

**Table 6:**Table of significant association between class and batch factors for ComBat, CSQN and CSQN + ComBat methods for the top 10 PCs. The p-values derived from T-test of each PC are shown.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| PC | Significant association between data factors and PCs | | | | | |
| ComBat | | CSQN | | CSQN + ComBat | |
| Class | Batch | Class | Batch | Class | Batch |
| PC1 | 0.61 | 0.91 | **1.16E-27** | **1.88E-03** | **1.45E-33** | 0.73 |
| PC2 | 0.71 | 0.36 | 0.99 | **2.21E-05** | 0.83 | 0.73 |
| PC3 | 0.63 | 0.80 | 0.21 | **8.34E-37** | 0.20 | 0.67 |
| PC4 | 0.10 | 0.69 | 0.49 | **2.32E-44** | 0.67 | 0.75 |
| PC5 | 0.06 | 0.91 | 0.80 | **1.85E-20** | 0.05 | 0.91 |
| PC6 | 0.13 | 0.73 | 0.08 | **5.66E-22** | 0.33 | 0.61 |
| PC7 | 0.21 | 0.22 | 0.30 | **1.78E-09** | 0.45 | 0.64 |
| PC8 | 0.31 | 0.25 | **0.23** | **0.03** | **1.26E-04** | 0.81 |
| PC9 | 0.38 | 0.85 | **0.85** | **8.17E-07** | **0.01** | 0.69 |
| PC10 | **0.02** | 0.49 | 0.01 | **0.01** | 0.48 | 0.82 |

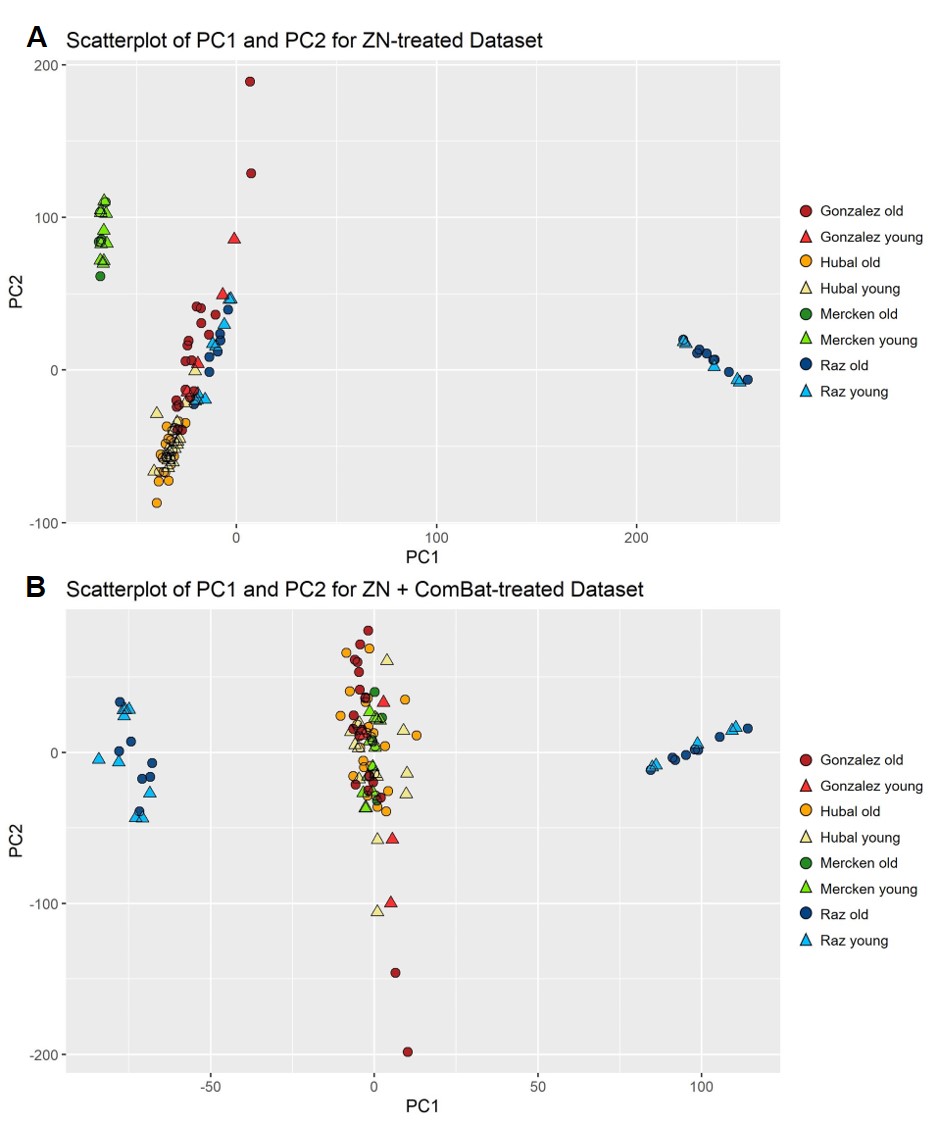
A higher resolution view of the PCs and significant associations of ComBat, CSQN and CSQN + ComBat-treated datasets further exemplifies the synergy between ComBat and CSQN (**Table 6**). Each method has their downsides. The class-correlated PC in ComBat was ranked lowly at 10 even though the batch effect was removed from the top 10 PCs. For the CSQN approach, even though the top PC is correlated with class, all top 10 PCs in the CSQN approach were also associated with batch effect. After combining both approaches, the class-correlated PC was elevated to rank 1 and association with batch effect was made insignificant in all top 10 PCs. This indicates that the individual strengths of CSQN and ComBat can be harnessed when we apply CSQN as the upstream normalisation technique and ComBat as the batch effect cleaning method. The CSQN + ComBat approach thus appears to be the most optimal combination for improving signal-to-noise ratio. From the PCA scatterplot, we observed that the samples are clearly separated by classes. However, even though batch effects were reported as insignificant from T-test, the old samples seem to still be aggregated by batches.



**Figure 7:** PCA scatterplot of CSQN + ComBat-treated dataset. Triangles represent samples classified as young and circles represent samples assigned as old. The colours denote the batch which each sample originates from.

Row-wise Z-normalization does not eradicate batch effects in mega-analysis

Row-wise ZN has been used in a number of microarray studies in the data pre-processing step. We hypothesise that even though ZN can centre gene expression values, it has minimal effect on eradicating batch effect even when coupled with downstream ComBat. Plots of individual samples against the top 2 PCs indicate that after treatment with ZN or ZN + ComBat, the separation by samples classes remained ambiguous (**Figure *6***).



**Figure 8:** PCA scatterplots for ZN (**A**) and ZN + ComBat-treated (**B**) datasets. Triangles represent samples classified as young and circles represent samples assigned as old. The colours denote the batch which each sample originates from.

After applying ZN on the dataset, the samples along the first and second PC still show distinct separation by batch. The samples were clustered together by batches and the division between old and young samples was not apparent. ZN + ComBat approach also exhibited sample clustering, but the aggregation by batches was attenuated. Nonetheless, T-test of the top 3 PCs still showed significant association with batch effects (**Figure *6***). Similar to ZN, the separation by class remained indistinct. In both approaches, the Raz batch showed large deviation from the rest of the batches, which may imply that it is still not ideal for cross-comparison with other batches in mega-analysis after applying ZN or ZN + ComBat. Both BERMs thus appear to not effectively alleviate batch effect and are subpar at preserving biological variation in the dataset.

Bootstrap analyses show that CS quantile + ComBat confers high reproducibility

To evaluate the overall reproducibility of batch effect removal method on total combined data, we conducted bootstrapping and Jaccard scoring on each dataset treated with a BERM. Bootstrapping was done by resampling samples in the dataset 1000 times followed by carrying out gene-wise T-test between classes to obtain a binary matrix of significance. We then determined how similar a bootstrap is to another by calculating the Jaccard score. A higher Jaccard score indicates that there is greater fidelity between two bootstraps in reporting if a certain gene is significant. When many higher Jaccard scores are obtained, the variation in the dataset is predominantly influenced by class effect, rather than idiosyncratic noise such as batch effect.

Chart

Description automatically generated

**Figure 9:** Violin plot of Jaccard score for all BERMs.

We found that when no BERM was applied (‘None’), the Jaccard scores obtained among bootstraps were the lowest across all BERMs. Conventional normalisation methods, i.e. ZN and QN showed yielded higher reproducibility than ComBat. Combination of ComBat and ZN or QN only marginally improved reproducibility. Even though ComBat displayed promising capability in reducing batch effect, the lack of signal preservation may have led to low reproducibility. All these six BERMs resulted in Jaccard scores with medians from 0.05 to 0.20. Expectedly, CSQN appeared to be the performance differential as the CSQN and CSQN + ComBat methods generated significantly higher Jaccard scores (medians ranging from 0.6 to 0.8) with CSQN + ComBat producing the highest median (0.78) and the smallest variance. This suggests that the batch-effect cleaning performance of the CSQN + ComBat method is the most stable across all the other BERMs examined. Combined with the optimal PCA result, CSQN + ComBat appears to be the most effective and reliable method to remove batch effect while preserving biological variation in mega-datasets.

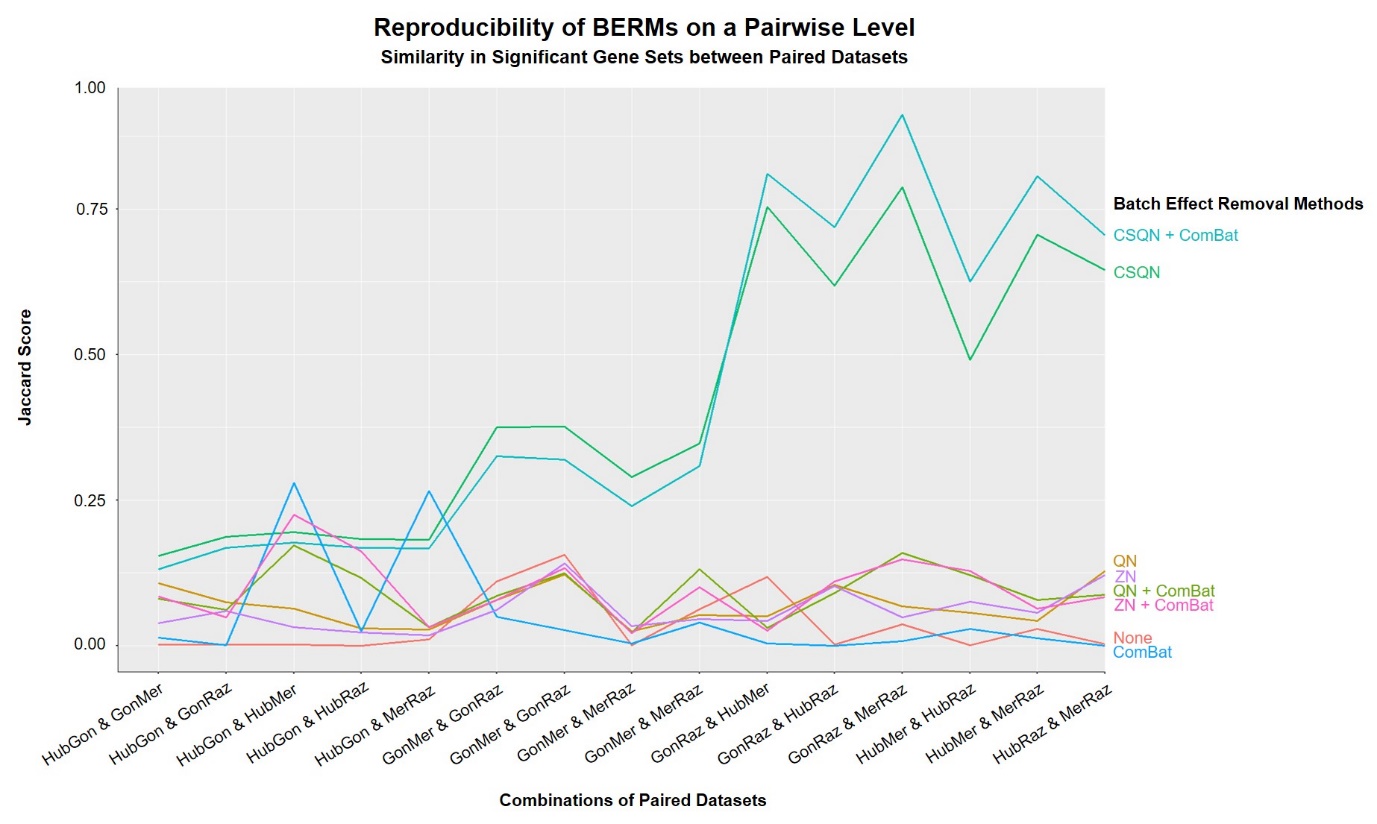
Pairwise independent analyses show that CS quantile + ComBat confers highest agreement rates

We subsequently investigated the performance stability of BERMs on pairwise combinations of datasets, instead of a whole mega-dataset. By combining any two batches in the mega-dataset, we obtained a dataset pair on which the 7 BERMs were applied. A total of 6 dataset pairs were generated from 4 batches (Gonzalez, Hubal, Mercken and Raz). We conducted T-test on each dataset pair to determine the significant genes associated with class effect. The significant gene sets for every 2 pairings were then cross-compared and their similarity calculated via Jaccard scoring (**Figure *11***).

A screenshot of a cell phone

Description automatically generated

**Figure 10:** Boxplot showing the Jaccard scores when significant gene sets from BERM-treated dataset pairs were compared.

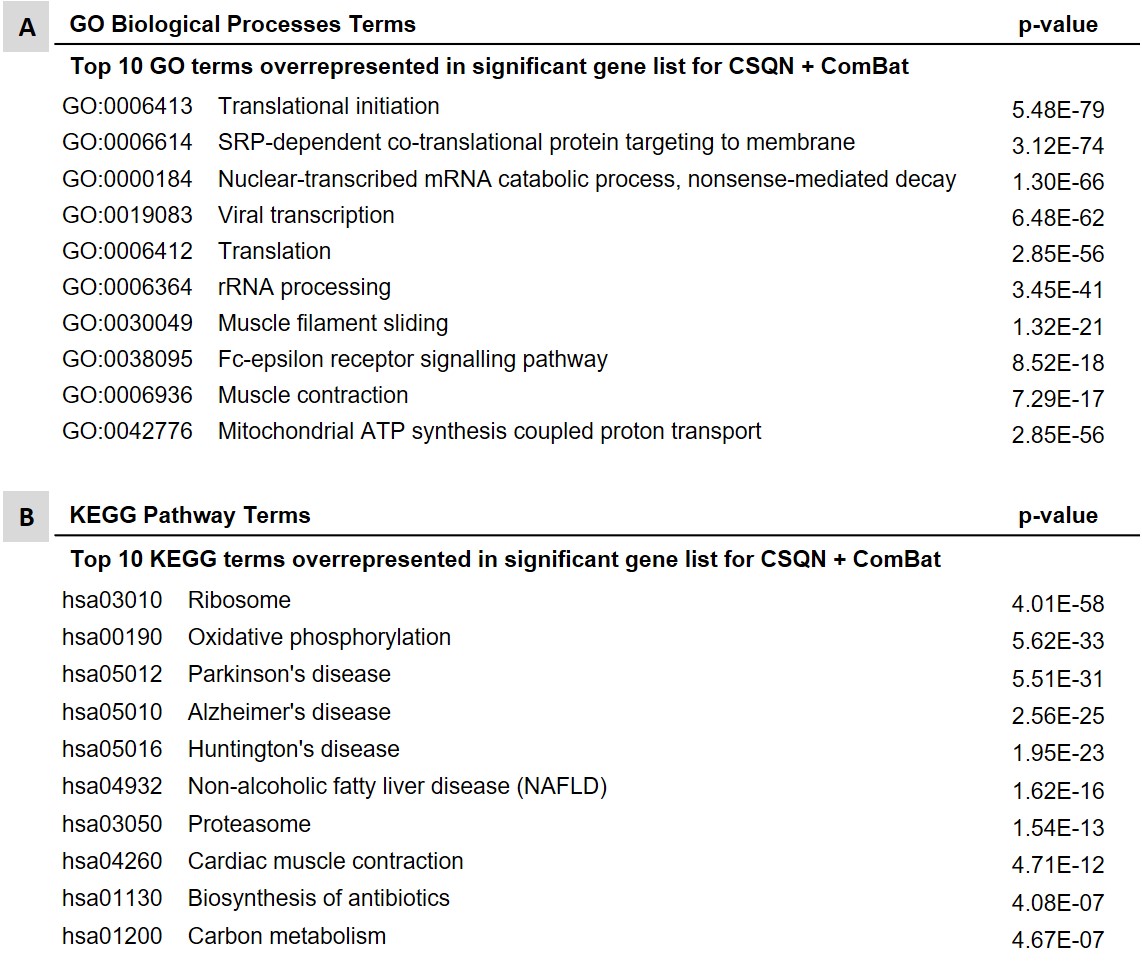


**Figure 11:** The Jaccard scores from were examined on the paired dataset level. Each line represents one BERM.

As expected, QN, ZN, and their combination with ComBat displayed consistently low Jaccard scores across all dataset pairs (0.00 to 0.20). However, the variances for CSQN and CSQN + ComBat were abnormally high; thus, we investigated the Jaccard scores on a pairwise level (**Figure *11***). Although the reproducibility of ComBat similarly remained low, there were peaks in Jaccard scores for the comparison between HubGon and HubMer, and between HubGon and MerRaz. Unsurprisingly, out of all 7 BERMs, CSQN and CSQN + ComBat show the highest similarity in significant gene sets between dataset pairings, reaching a score of 0.6 for CSQN + ComBat. However, the variance between the Jaccard scores for both approaches was large as the lowest Jaccard scores were close to 0.20. We suspect this might be due to the disparity in sample sizes of the dataset pairs compared.

Functional analysis of significant genes in CSQN + ComBat-treated dataset

Functional analysis of the significant genes in the CSQN + ComBat-treated dataset was conducted using the Database for Annotation, Visualization, and Integrated Discovery (DAVID) Tool [14].



**Figure 12:** DAVID functional annotation for the significant gene list for CSQN + ComBat-treated dataset. The top 10 gene ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) terms and the corresponding p-values are shown.

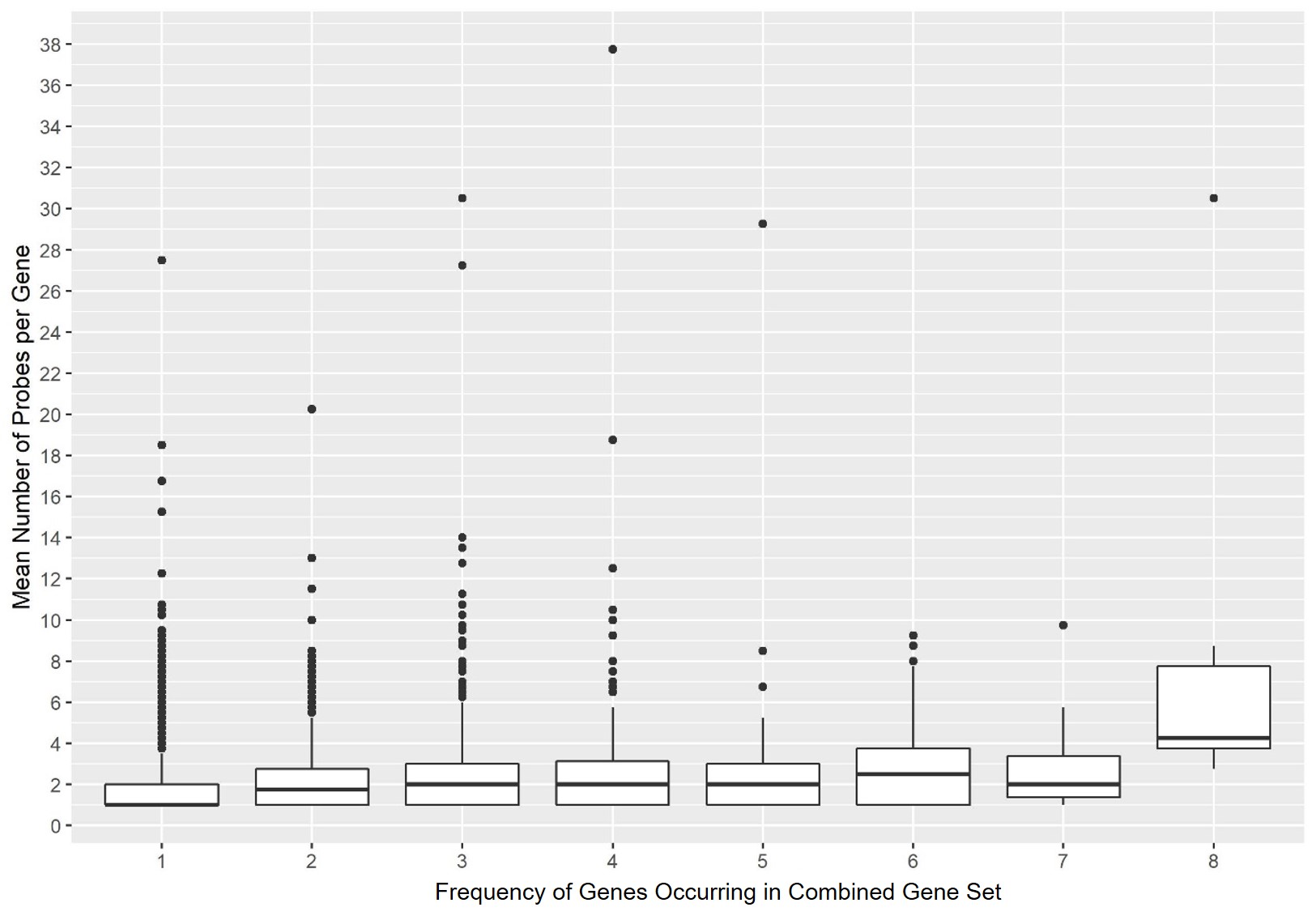
Many of the gene ontology (GO) terms include housekeeping biological functions such as translation and rRNA processing. Terms relevant to muscle contraction are also overrepresented and this may be because the samples were all collected from the vastus lateralis muscle. The top 10 KEGG terms include many pathways for degenerative diseases commonly associated with ageing such as Parkinson’s, Alzheimer’s and Huntington’s disease.

There are no inherent biases suggesting existence of batch-correlated genes/probes

We investigated how batch effect manifests among genes to find out if there exists a conserved gene set that disproportionately loads more to batch effect. By combining samples from the same class but different batches, we obtained 12 same-class-different-batch (SCDB) dataset combination. As these samples are of the same class, we expect that the variation between the samples originated from technical bias such as batch effect. We conducted PCA on each SCDB combination. We found out that the majority of the SCDB dataset pairs show that the distribution of batch effects among genes is not uniform as some had much greater loadings, leading to plots that are skewed towards the left. However, when we cross-compared the top batch-correlated genes across these SCDB combination, we found that these gene sets were not similar as the median Jaccard score was 0.18. This suggests that even though there exists a disproportionate number of genes that influence batch effect more strongly, this set of genes changes from dataset to dataset.

To investigate whether the number of probes matching to a gene affects the likelihood of it being reported as batch-correlated, we extracted the top significant genes associated with batch effect for each SCDB pair. We then combined these gene sets across all 12 SCDB pairs into a union set and calculated the frequency of each gene being reported as significant (**Figure *13***). A higher probe count for a gene does not seem to increase the frequency of it being batch-correlated in more SCDB dataset pairs. There thus appears to be no apparent correlation between the number of probes matching to a gene and the tendency for that gene to be reported as batch-correlated. This suggests that the number of probes is not a factor influencing how strongly a gene is associated with batch effect.

These results imply that although batch effect in a dataset manifests non-uniformly among genes, the genes loading highly to batch effect are not consistent and this batch-correlation is likely not due to the number of probes matching to these genes.



**Figure 13:** Boxplot of number of probes matching to a gene against the frequency of occurrence in the combined gene set.

Discussion

Whole-data and gene-wise normalization do not eradicate batch effects

Multiple mega-analysis studies incorporate gene-wise standardization in the data pre-processing step [9] in order to normalize the range and variance of expression values. We have shown that such standardization conducted gene-wise is not effective at priming datasets for mega-analysis, even if ComBat is applied downstream. The use of ZN showed insignificant improvement at eradicating batch effect, and this performance was minimally enhanced when ComBat was applied as a subsequent step. Both approaches were not effective at preserving biological signals. For ZN, due to under-correction of batch effect, batch effect remained the predominant variation in the mega-dataset. PCA analysis supported this by showing the distinct delineation of batches in the top 2 PCs (**Figure *8***). ZN + ComBat was rather effective at removing batch effect, but the biological variation remained suppressed. This can be seen from the aggregation of samples from different batches in **Figure *8*** but indistinct separation by class. Such similar results were observed for whole-data QN and QN + ComBat.

This sub-optimal batch-correction performance may be due to the nature of applying normalization methods in a whole-data manner. By doing so, we assume that all samples in the dataset follow identical feature value distribution, regardless of class. Thus, ZN and QN serve to normalize each sample in the dataset to the same target distribution, instead of normalizing samples to their respective class distribution. This may have suppressed the variation by class. Therefore, even though the subsequent use of ComBat can correct some batch effect, the biological variation remains indistinguishable.

ComBat by itself is not good enough to retain biological variability

We have shown that ComBat can significantly alleviate batch effect in the mega-dataset (**Figure *6***). This result has been widely proven in many studies. For example, Chen et al found ComBat to outperform 5 other programs designed for batch effect correction (distance-weighted discrimination, support vector machines, mean-centering, surrogate variable analysis and Ratio\_G) in multiple metrics [15]. Thus, the superior batch effect correction of ComBat is certain. However, due to its subpar ability in preserving biological signal, it does not fare well at reproducing similar significant gene sets (**Figure *9***), thus making the ComBat-only method an unreliable data pre-processing approach for subsequent feature-based mega-analysis.

A similar downside of ComBat was found by Muller et al who aimed to identify an appropriate method for batch effect removal in a large-scale longitudinal microarray study [10]. This study included RNA collection of subjects at the base level and after 5 years. The samples collected from individuals and analyzed at base level were allocated to group BL, samples collected from these same individuals at base level but analyzed after 5 years were grouped as BLFU and samples collected and analyzed from the same individuals after 5 years were considered in the FU group. It was expected that any variation between the BL and BLFU groups would be due to technical bias, while variation between the BLFU and FU groups would be due to biological variability. Muller et al combined BLFU and FU to create one batch and BL was another batch. ComBat was applied based on this batch delineation. Using PCA, Muller et al discovered that ComBat was able to remove clustering between BL and BLFU samples, indicating that batch effects have been reduced. However, Bland-Altman plots for evaluation of agreement between repeated probe measures for ComBat method showed that even though reproducibility was improved compared to the uncorrected data, differences could still be observed. Hence, this result agrees with our study that while ComBat can effectively reduce batch effects, it performs mediocrely at retaining class effect.

An optimal configuration based on CSQN + ComBat for mega-analysis

Batch-correction performance and reproducibility can be significantly boosted when ComBat is coupled with upstream CSQN. In this study, we have shown that CSQN performs well at preserving the biological signals due to the application of QN on two classes of samples separately. The CSQN method was shown to outperform other QN methods in terms of batch effect correction and statistical feature selection in proteomics data with simulated class and batch effects [16]. Hence, we recommend the usage of CSQN over whole-data normalization methods as a critical upstream technique to preserve class effect before batch-correction. When coupled with downstream ComBat, the technical bias in the dataset was reduced and CSQN + ComBat emerged as the best method to reduce batch effect while retaining biological signals. The signal-preserving effect of CSQN proved to be useful for subsequent feature selection as we were able to extract meaningful differential genes influencing ageing from the CSQN + ComBat-treated mega-dataset.

Muller et al also examined the use of QN prior to the application of ComBat [10]. Unlike our study, they explored batch-wise QN coupled with subsequent ComBat and mentioned that this method led to a slight enhancement in reducing technical bias compared to the ComBat-only approach. QN strategies have also been investigated in a methylation microarray study by Sun et al [17]. They found that while the commonly used “lumi” method of QN outperformed other QN methods to reduce batch effect, it became less effective when batch effects were more severe. ComBat was required as subsequent cleaning step to effectively remove batch effects for better signal detection. These studies suggest that the general framework of a QN method followed by ComBat is effective for proper batch effect correction. In this study, we recommend the use of CSQN as the prior normalization approach before the application of ComBat.

Our functional analysis of the top significant genes obtained from feature selection in CSQN + ComBat dataset show that they are relevant and meaningful. Most of the GO and KEGG terms are housekeeping biological functions and closely related to the ageing phenotype. This suggests that CSQN + ComBat can effectively prime mega-datasets for subsequent meaningful feature selection.

There is no evidence for a conserved set of batch-correlated genes

In this study, we investigated how batch effects manifest in datasets and checked if there exists a distinct group of genes contributing strongly and consistently to batch variation. Our PCA results show that the batch-correlated group of genes differs from dataset to dataset. We also examined the possible correlation between the number of probes matching to a gene and the tendency of the gene to load highly to batch effect. Our results indicate that this correlation is spurious. Hence, we conclude that batch effect generally does not manifest uniformly among genes. However, we are still unsure of what causes this non-uniformity.

Study limitations

Even though we evaluated multiple BERMs, our study is limited to transcriptomics data. Due to inter-platform differences (e.g. different number of variables measured, different sources of technical bias, and different orders of magnitudes for features measured), the statistical performance of the BERMs evaluated in this study may change when data from another microarray platform is used. We can consider comparing the performances of BERMs across different platforms and prescribe optimal batch effect removal techniques unique to the properties of each platform.

In this study, we only used real transcriptomics data and thus we cannot quantify the exact amount of batch and class effect before and after application of BERMs. This hindered us from evaluating true performance in terms of precision and recall (if such a thing exists). We could only determine the relative effectiveness of batch effect removal by doing relative comparisons among the BERMs examined.

Another limitation of this study includes the possible lack of sensitivity in detecting batch effect as we based our PCA correlations using p-values and t-statistics. As seen from the scatterplot for CSQN + ComBat method, some clustering by batch still remained even though T-test conducted on the top PCs indicate that they are not significantly correlated with batch. P-values have also shown to be highly unstable [18] and this may affect the reliability of the standard statistical tests we conducted.

Efforts in microarray data analysis are shifting to focus on subnetworks of genes with related functions [19] rather than identifying individual genes or probes. More feature engineering possibilities are also unleashed when microarray data analysis is combined with machine learning. Future work on the performance of BERMs should thus be benchmarked in the context of such paradigms.

Conclusion

The combination of datasets from different experiments inevitably creates batch effect that confounds with true biological signal. Such data pooling (mega-analysis) creates reproducibility issues if batch effects are not adequately corrected. Batch effect correcting algorithms and normalisation methods are commonly used to remove technical bias in mega-analysis. However, careless application of such methods is at best ineffective and at worst, introduce false positives and false negatives in the dataset. In this study, we examined 7 BERMs in terms of effectiveness at removing batch effect in a combined aging dataset while retaining biological variation. We also evaluated the reproducibility of their performance. Our results show that while ComBat can eradicate batch effect, it is subpar at retaining biological signals. Preservation of biological variability was boosted the most when ComBat was coupled with prior CSQN and we managed to derive a gene signature biological relevant to aging using the CSQNB + ComBat-treated dataset. We thus recommend CSQN + ComBat as the best and most reliable approach for removing batch effects in combined datasets for mega-analysis studies.

References

1. Ioannidis, J.P., *Why most published research findings are false.* PLoS Med, 2005. **2**(8): p. e124.

2. Begley, C.G. and J.P. Ioannidis, *Reproducibility in science: improving the standard for basic and preclinical research.* Circulation research, 2015. **116**(1): p. 116-126.

3. Chan, A.-W., et al., *Increasing value and reducing waste: addressing inaccessible research.* The Lancet, 2014. **383**(9913): p. 257-266.

4. Nosek, B.A., et al., *Estimating the reproducibility of psychological science.* Science, 2015. **349**(6251): p. aac4716-aac4716.

5. Nosek, B.A. and T.M. Errington, *Making sense of replications.* eLife, 2017. **6**: p. e23383.

6. Halsey, L., et al., *The fickle P value generates irreproducible results*. Vol. 12. 2015. 179-85.

7. Leek, J.T., et al., *Tackling the widespread and critical impact of batch effects in high-throughput data.* Nature Reviews Genetics, 2010. **11**(10): p. 733-739.

8. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-27.

9. Tylee, D.S., et al., *Blood transcriptomic comparison of individuals with and without autism spectrum disorder: A combined-samples mega-analysis.* American Journal of Medical Genetics Part B: Neuropsychiatric Genetics, 2017. **174**(3): p. 181-201.

10. Müller, C., et al., *Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data.* PloS one, 2016. **11**(6): p. e0156594-e0156594.

11. *Aging will reduce economic growth worldwide in the next two decades*. 2014, Moody's Investors Service.

12. Lu, T., et al., *Gene regulation and DNA damage in the ageing human brain.* Nature, 2004. **429**(6994): p. 883-91.

13. Dunning M, L.A., Eldridge M, *illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4)*. 2015.

14. Dennis Jr, G., et al., *DAVID: Database for Annotation, Visualization, and Integrated Discovery Genome Biol 4 (9): R60–R60. 11.* Find this article online, 2003.

15. Chen, C., et al., *Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods.* PLOS ONE, 2011. **6**(2): p. e17238.

16. Zhao, Y., L. Wong, and W.W.B. Goh, *How to do quantile normalization correctly for gene expression data analyses.* Scientific Reports, 2020. **10**(1): p. 15534.

17. Sun, Z., et al., *Batch effect correction for genome-wide methylation data with Illumina Infinium platform.* BMC Medical Genomics, 2011. **4**(1): p. 84.

18. van Helden, J., *Confidence intervals are no salvation from the alleged fickleness of the P value.* Nature Methods, 2016. **13**(8): p. 605-606.

19. Lim, K. and L. Wong, *Finding consistent disease subnetworks using PFSNet.* Bioinformatics, 2014. **30**(2): p. 189-96.