Doing Mega-Analysis Right: Class-Specific Quantile Normalization Coupled with Combat as Most Optimal Approach for Batch Effect Removal

by

Joan Jong Jing Yi

Supervisor: Dr Wilson Goh

School of Biological Sciences

Submitted to the School of Biological Sciences in partial fulfilment of the requirements for Final Year Project

Nanyang Technological University

April 2021

**DECLARATION**

I declare that,

in accordance with School requirements this thesis is under 6000 words in length;

all presented work was performed within the official project time frame as stated below;

all presented work was performed by me, unless otherwise specified;

all relevant work experience gained before the Final Year Project is stated below;

the input by my supervisor, or delegated supervisor, into this thesis was limited to reviewing of up to two hard copy drafts;

this thesis is my own work, unless otherwise referenced, as defined by the NTU policy on plagiarism and I have read the NTU Honour Code and Pledge;

the included Abstract, Introduction, Results, Discussion and Conclusions sections will be submitted to Turnitin no later than 24 hours after hardcopy submission.

Final Year Project start date: 7 December 2020

Final Year Project submission date: 26 April 2021

Total number of weeks: 20 weeks

Pre-Final Year Project experience

I completed a research internship at Karolinska Institutet from January 2020 to June 2020 as part of an overseas exchange programme.

Student’s signature: Joan Jong Date: 26 April 2021

Contents

[Abbreviations 6](#_Toc69957820)

[Abstract 7](#_Toc69957821)

[Introduction 8](#_Toc69957822)

[Data pooling improves study reproducibility 9](#_Toc69957823)

[Mega-analysis inevitably generates batch effects 9](#_Toc69957824)

[Batch effects may manifest non-uniformly 10](#_Toc69957825)

[Synergy between normalization techniques and batch effect removal algorithms not adequately explored 10](#_Toc69957826)

[Using ageing datasets for exploring synergies between normalization and Batch Effect Correction Algorithms (BECAs) 11](#_Toc69957827)

[Materials and Methods 13](#_Toc69957828)

[Literature search and filtering 13](#_Toc69957829)

[Data processing 15](#_Toc69957830)

[Benchmarking methods 15](#_Toc69957831)

[Z-normalization 15](#_Toc69957832)

[Quantile and class-specific quantile normalization 16](#_Toc69957833)

[ComBat 16](#_Toc69957834)

[PCA for assessment of batch effect removal 16](#_Toc69957835)

[Whole-data bootstrap and Jaccard analysis 17](#_Toc69957836)

[Pairwise cross comparison of batches 18](#_Toc69957837)

[Probe-batch effect correlation check 19](#_Toc69957838)

[Check for biological relevance of differential genes 20](#_Toc69957839)

[Results 22](#_Toc69957840)

[Strong batch effects were generated after combining datasets from different studies 22](#_Toc69957841)

[CSQN-ComBat gave the best outcome for batch effect removal 24](#_Toc69957842)

[Row-wise Z-normalization does not eradicate batch effects in mega-analysis 27](#_Toc69957843)

[Bootstrap analyses showed that CSQN-ComBat conferred high reproducibility 28](#_Toc69957844)

[Pairwise independent analyses showed that CSQN-ComBat conferred highest agreement rates 29](#_Toc69957845)

[Functional analyses of significant genes in CSQN + ComBat-treated dataset 31](#_Toc69957846)

[There were no inherent biases suggesting existence of batch-correlated genes or probes 34](#_Toc69957847)

[Discussion 37](#_Toc69957848)

[Whole-data and gene-wise normalization methods did not alleviate batch effect 37](#_Toc69957849)

[ComBat alone was not good enough to retain biological variability 37](#_Toc69957850)

[An optimal configuration based on CSQN-ComBat for mega-analysis 38](#_Toc69957851)

[There was no evidence for a conserved set of batch-correlated genes 39](#_Toc69957852)

[Study limitations 39](#_Toc69957853)

[Conclusions 41](#_Toc69957854)

[References 42](#_Toc69957855)

**ACKNOWLEDGEMENTS**

I would like to thank Dr Wilson Goh, the project supervisor, for the opportunity to work on this fascinating project. Dr Goh provided valuable guidance and insight throughout the research and offered feedback for further improvement in areas of data analysis, programming and manuscript writing.

# Abbreviations

|  |  |
| --- | --- |
| BECA | Batch Effect Correction Algorithm |
| BERM | Batch Effect Removal Method |
| CSQN | Class-Specific Quantile Normalisation |
| DAVID | Database for Annotation, Visualization, and Integrated Discovery |
| GEO | Gene Expression Omnibus |
| GSE | Gene Expression Omnibus Series |
| HGNC | HUGO Gene Nomenclature Committee |
| HUGO | Human Genome Organisation |
| PC | Principal Component |
| PCA | Principal Component Analysis |
| QN | Quantile Normalisation |
| SCDB | Same Class Different Batch |
| ZN | Z-Normalisation |

# Abstract

|  |  |
| --- | --- |
| Project Title: | Doing Mega-Analysis Right: Class-Specific Quantile Normalization Coupled with Combat as Most Optimal Approach for Batch Effect Removal |
| Student Name: | Joan Jong Jing Yi |

**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.

# 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. Due to the ability to identify and quantify biological moieties within samples, comparative analysis between multiple tissues can be performed. This facilitates detection of differential expression states amongst biological moieties, which in turn informs on what makes these tissues unique.

However, high-throughput assays are not without issues. High-throughput assays are typically underpowered and suffer from ‘curse of dimensionality’ issues (Clarke et al., 2008). That is, there are many more biological moieties being measured than actual samples, making it difficult to correctly estimate meaningful variations. This in turn produces 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. Ionaddis et al. found that in the scientific fields with smaller studies, such as research of molecular predictors, the positive predictive value (PPV) is low (Ioannidis, 2005). This indicates that findings from smaller studies, compared to that in large-scale randomised trials, are less likely to be true due to the small statistical power of the data used. Other factors that can increase the possibility of a published study being unreproducible include small effect sizes, data dredging and conflict of interests. Many metascience studies from different fields concur that reproducibility is subpar (Begley et al., 2015) (Chan et al., 2014). 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 (Brian A Nosek et al., 2015). It has since expanded to include cancer studies (Brian A. Nosek et al., 2017).

Study reproducibility issues are at times also referred to as the ‘Winner’s Curse’ phenomenon (Halsey et al., 2015). This occurs when 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.

## Data pooling improves study reproducibility

Decreasing running costs and the ever-growing collection of datasets submitted to online repositories have made data pooling a plausible approach towards improving study reproducibility. The purpose is to improve coverage on the study population and to enhance statistical power. Data pooling expands sample size by consolidating data from similar studies. Statistical power is improved relative to that in each individual study. Thus, it becomes 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.

Two main approaches for data pooling exist in biomedical research: meta-analysis of aggregate patient data and mega-analysis based on individual patient data. Meta-analysis involves combining summary results from different studies published, while mega-analysis entails collection of raw individual-level clinical and biological data from multiple studies. The latter strategy demands considerably more resources and time 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 is considered less reliable than mega-analysis and may not be the best data pooling method. 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 (Stewart et al., 1995).

## Mega-analysis inevitably generates batch effects

In mega-analysis, 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 (Zhao et al., 2020). 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 (Goh, Wang, et al., 2017). As batch effects are idiosyncratic in nature with no biological basis, they can have profound effects on 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 (Peixoto et al., 2015). However, more concerning repercussions are false effects that may be induced due to confounding between batch effects and the results, leading to inaccurate clinical conclusions (Goh et al., 2018).

## Batch effects may manifest non-uniformly

Further complicating batch effects is their possibly non-uniform nature. Leek et al. examined datasets from nine superficial transitional cell carcinoma (sTCC) studies and found that, across these datasets, different proportions (32.1 to 99.5%) of measured features showed significant association with processing date, regardless of biological phenotype (Leek et al., 2010). 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 (Zhou et al., 2019) (Peixoto et al., 2015). 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 (Zhao et al., 2020). There are also algorithms developed specifically to deal with batch effects known as batch effect-correcting algorithms (BECAs), such as ComBat (Johnson et al., 2007) and Surrogate Variable Analysis (Leek et al., 2007). Several mega-analyses have attempted to combine generic normalization techniques with a BECA as an approach to remove batch effects. Notably, Tylee et al. utilized Z-normalization as the first cleaning step prior to ComBat (Tylee et al., 2017) and Müller et al. evaluated that combining whole-data QN with ComBat serves as the best approach for batch effect removal (Müller et al., 2016). 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 overcome technical bias such as batch effects (Zhao et al., 2020). As for synergies between normalization and BECAs, there are some contextual dependencies as well (Zhou et al., 2019). Hence, further research exploring how normalization methods, especially improved ones considering important confounding co-variates such as 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 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 (Moody's Investors Service, 2014). 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 (Lu et al., 2004). 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 explored 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 did not consider other BECAs (Müller et al., 2016) (C. Chen et al., 2011). Instead, we determined whether there was synergy between various normalization techniques with ComBat. We hypothesized that the combination of CSQN and downstream ComBat would be the most effective and reliable at removing batch effects while retaining biological variation.

# Materials and Methods

## Literature search and filtering

We explored the Gene Expression Omnibus (GEO) repository for microarray-based studies examining the effect of ageing on gene expression. We used the query terms “ageing” and “homo sapiens” and yielded 506 relevant studies. The criteria for study inclusion and exclusion were as follows: (1) we included only studies using expression profiling by microarray, (2) only studies examining human vastus lateralis tissue were included, (3) studies involving subjects of a wide age range were incorporated to ensure that young and old patients were represented, (4) datasets examining myoblast cells were excluded as the degree of cell differentiation would confound with gene expression levels, making comparisons with somatic cells incompatible (Collas et al., 2007), and (5) only datasets without prior normalization were incorporated in our mega-analysis. These criteria narrowed our search to four studies GSE40645 (Raz et al., 2014), GSE83352 (Hubal et al., 2016), GSE87105 (Mercken et al., 2017), and GSE98613 (Gonzalez-Freire et al., 2018). All datasets were Illumina BeadChip arrays. The subjects’ ages in all datasets ranged from 28 to 89 years old.

Each dataset was then subjected to the following sample criteria: (1) we only included samples from subjects who did not undergo intervention or were controls and (2) samples without age data were excluded. We separated eligible samples into two classes, young and old, based on a threshold of 50 years old. The metadata of the studies included in our mega-analysis is summarized inTable 1. The dataset from each study was considered a batch and labelled accordingly (Table 1). Ultimately, our mega-analysis included a total of 110 samples from four batches; 44 samples were allocated to the young class and 66 samples were assigned to the old class (Table 1).

**Table 1: Table summarising the metadata for each study dataset included in mega-analysis after data processing.** A total of 110 samples were included in our mega-analysis, out of which 44 were from young subjects and 66 were from old subjects. All studies were performed using Illumina microarray chips. However, GSE40645 differed slightly from the other datasets as the Illumina bead-chip used was of a newer version.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Study** | **Batch Label** | **GSE Name** | **Microarray Platform** | **Tissue Sampled** | **Total Samples** | **Young/Old** | **Female/Male** |
| (Raz et al., 2014) | Raz | GSE40645 | GPL6947a | Vastus lateralis | 29 | 14/15 | 14/15 |
| (Hubal et al., 2016) | Hubal | GSE83352 | GPL10558b | Vastus lateralis | 42 | 16/26 | 23/19 |
| (Mercken et al., 2017) | Mercken | GSE87105 | GPL10558b | Vastus lateralis | 16 | 10/6 | 0/16 |
| (Gonzalez-Freire et al., 2018) | Gonzalez | GSE98613 | GPL10558b | Vastus lateralis | 23 | 4/19 | 8/15 |
| Total | | | | | 110 | 44/66 | 45/65 |

a = Illumina HumanHT-12 Version 4.0

b = Illumina HumanHT-12 Version 3.0

## Data processing

We processed and normalised each study dataset independently. Covariate and expression data were extracted. Using the IlluminaHuman4 library from the Bioconductor package (Dunning M, 2015), we mapped probes to their respective HUGO Gene Nomenclature Committee (HGNC) gene symbols (Povey et al., 2001). Expression data from probes with no matching HGNC gene symbol was removed. We collapsed expression data from multiple probes matching to the same gene by calculating the mean of the expression values. We log2-transformed the four processed datasets before combining them in series to create a mega-dataset consisting of covariate data and expression values from 21,053 genes.

## Benchmarking methods

In this study, we evaluated seven different combinatorial batch effect removal methods (BERMs). These BERMs were either normalisation, ComBat, or a combination of both techniques. We applied each BERM to our mega-dataset independently and obtained seven different BERM-treated mega-datasets. The BERM-free mega-dataset was used as the negative control (Table 2).

**Table 2. Table summarising the seven BERMs evaluated in this study.** Each BERM was either a normalisation technique, BECA or a combination of upstream normalisation and BECA. The BECA used in this study was ComBat. The application of the seven BERMs on the log2-transformed mega-dataset yielded seven BERM-treated mega-datasets. The log2-transformed mega-dataset without application of a BERM was used as the negative control and named “Control”.

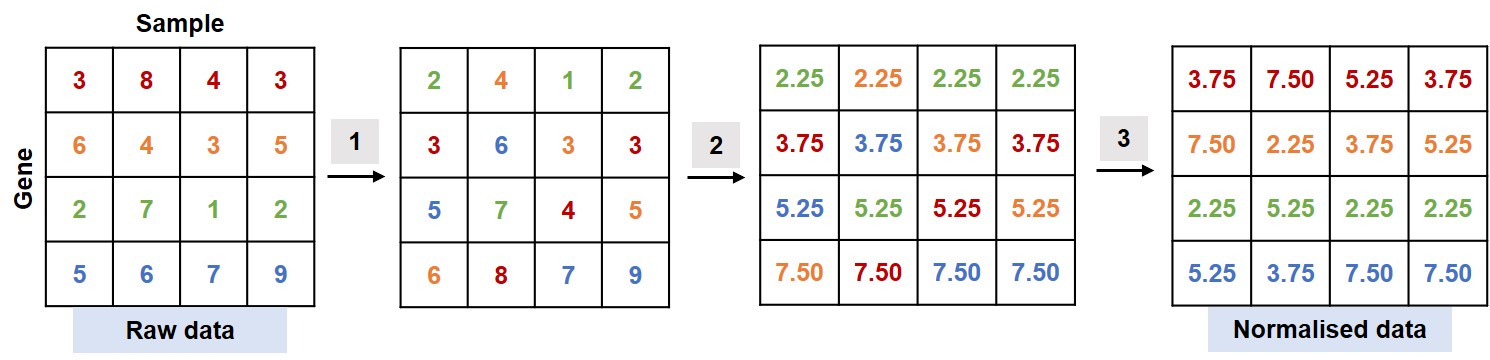
|  |  |  |
| --- | --- | --- |
| **BERM Name** | **Normalisation** | **BECA** |
| (Control) | - | - |
| QN | Quantile normalisation | - |
| ZN | Z-normalisation | - |
| ComBat | - | ComBat |
| CSQN | Class-specific quantile normalisation | - |
| QN-ComBat | Quantile normalisation | ComBat |
| ZN-ComBat | Z-normalisation | ComBat |
| CSQN-ComBat | Class-specific quantile normalisation | ComBat |

## Z-normalization

Z-normalisation (ZN) is a classical technique for normalisation of microarray data so that gene expression values generated from different microarray studies become comparable across experiments. The mean and variance of expression for each gene were standardised across samples in the mega-dataset.

## Quantile and class-specific quantile normalization

Quantile normalization (QN) is a simple procedure conventionally used in microarray studies (Amaratunga et al., 2001). The workflow for QN is shown in Figure 1.



**Figure 1: QN workflow for normalisation of microarray data.** Each dataset column denotes a sample, and each row denotes a gene. (1) Within each sample, expression values are ranked by magnitude. (2) The expression values across samples with the same rank are substituted by the row mean. (3) The mean expression values are resorted according to their original order denoted by the colours.

Class-specific QN (CSQN) is a modified version of QN. Rather than applying QN on the whole mega-dataset, samples were first split according to their respective classes before applying QN on each dataset class separately.

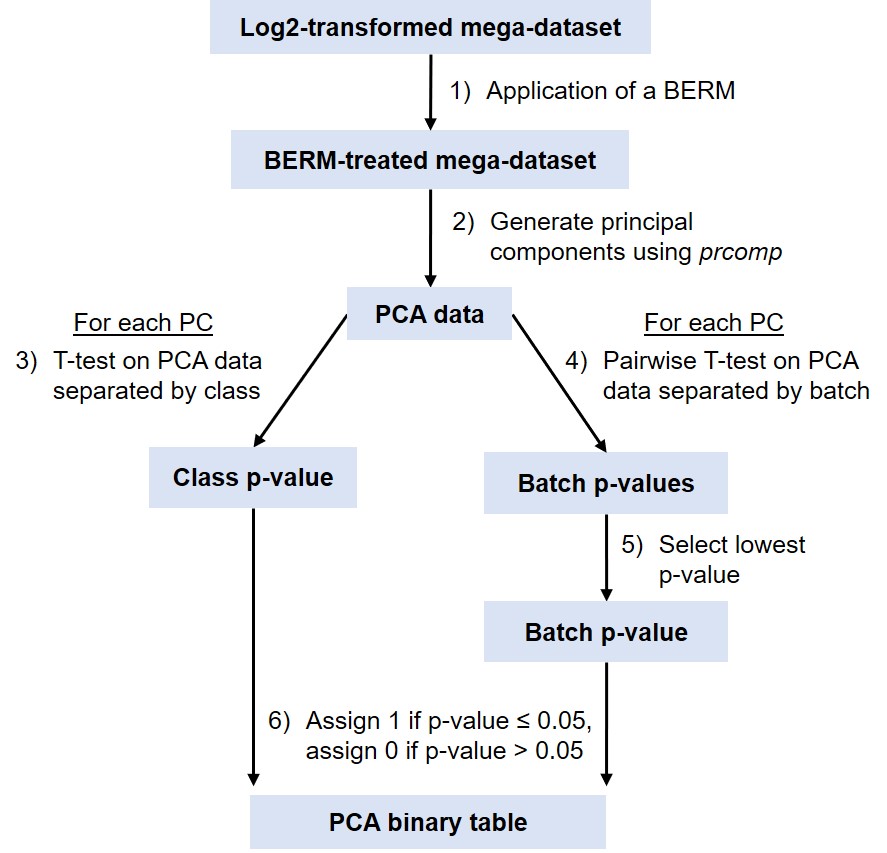
## ComBat

ComBat is an empirical Bayes BECA which combines both location and scale adjustments for batch effect removal (Johnson et al., 2007). The mean and variance are estimated for each batch and gene independently. Specification of batches is required for the use of ComBat.

## PCA for assessment of batch effect removal

Principal Component Analyses (PCA) were performed using the *prcomp* function in R to split overall data variance into independent principal components (PCs). Higher-ranked PCs captured more variance in the dataset and this variance was contributed by biological or technical variability. PCA thus enabled the assessment of batch or class effect in a BERM-treated mega-dataset to evaluate BERM performance (Goh, Sng, et al., 2017).

The workflow for PCA is shown in Figure 2. A PCA binary table was obtained for each BERM-treated mega-dataset (Table 3).



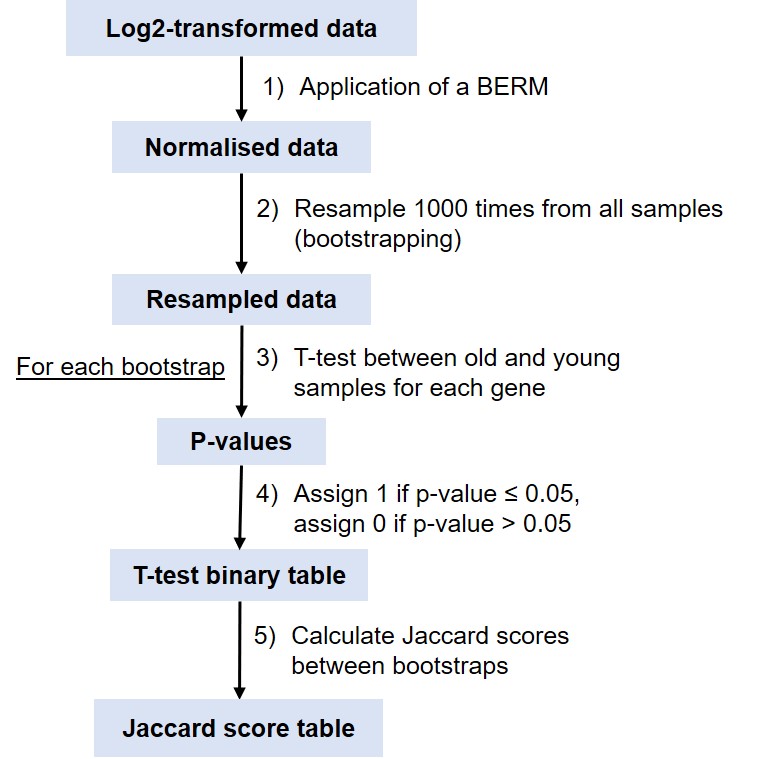
**Figure 2. PCA workflow to determine the amount of batch and class effects present in a BERM-treated mega-dataset.** (1) A BERM was applied on the log2-transformed mega-dataset. (2) Using the prcomp function in R, PCA data was generated for each BERM-treated mega-dataset. (3) Eigenvalues for each PC were separated by class. Student’s T-test was applied between the two classes based on the null hypothesis that there was no significant difference in the means and any difference was due to chance. (4) A similar procedure was performed for the same PCs with the eigenvalues separated by batch. (5) Since there were more than two batches, pairwise T-test was conducted, and the lowest p-value was considered. (6) P-values for the T-test were collated and binarized using a threshold of 0.05.

**Table 3: PCA binary table for Control up to the sixth PC.** The table shows the binarized p-values for each PC in the Control mega-dataset. In binary table, the top three PCs were significantly correlated with batch effect. This indicated that batch effect contributed more to the variance in Control mega-dataset than class effect.

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

## Whole-data bootstrap and Jaccard analysis

We determined which BERM was most consistent at removing batch effect based on feature selection reproducibility. Bootstrapping and concomitant calculation of Jaccard score were conducted for each BERM-treated mega-dataset and Control. A Jaccard score quantifies the similarity between two sets (Equation 1). A higher score indicates a higher degree of similarity. The Jaccard scores for all BERM-treated mega-datasets and Control were then compiled and plotted to compare the reproducibility of each BERM.

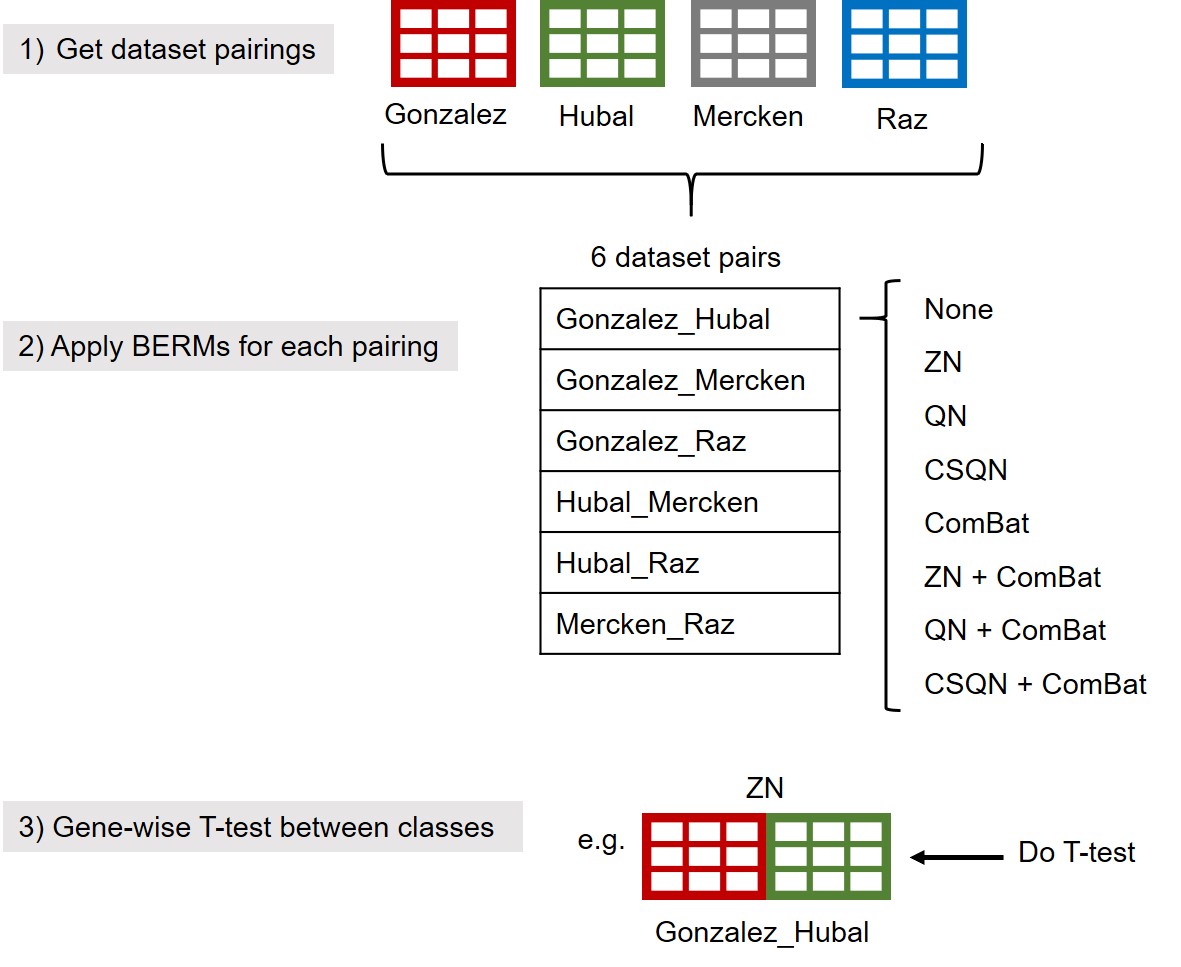


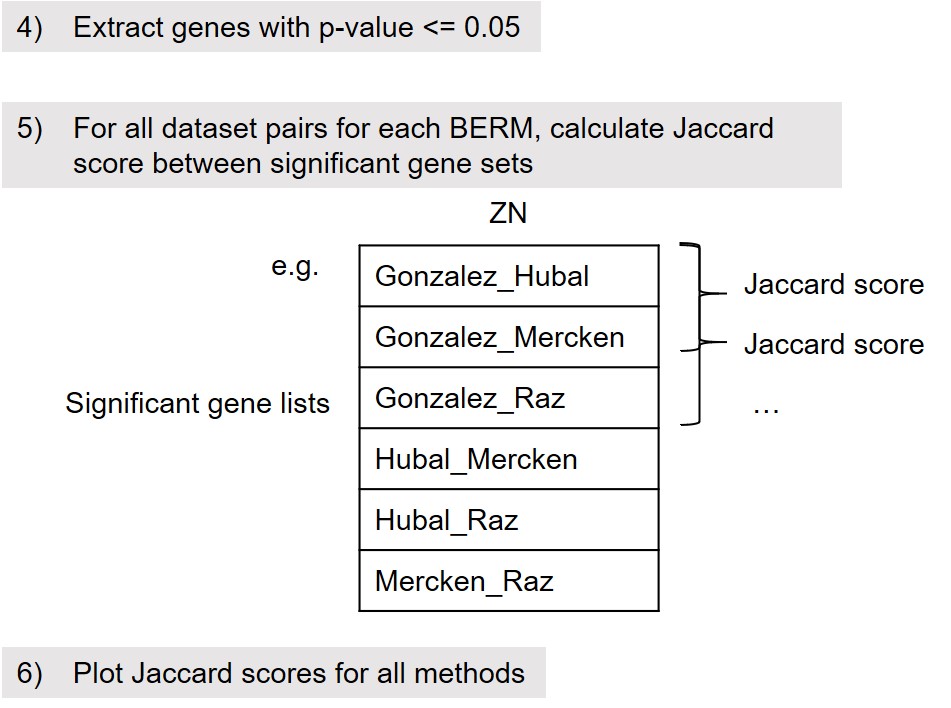
**Figure 3: Bootstrapping and Jaccard scoring workflow.** (1) A BERM was applied on the log2-transformed mega-dataset. (2) Each BERM-treated mega-dataset was resampled 1000 times to yield 1000 bootstraps. (3) For each bootstrap, Student’s T-test was performed on each gene between old and young samples to yield p-values. (4) The p-values for the genes in each bootstrap were binarized using a threshold of 0.05. (5) Each bootstrap in the binary table was compared to the other bootstraps to determine their similarity using Jaccard scoring.

**Equation 1: Jaccard score formula.** Jaccard score was used to determine the degree of similarity between sample sets. A Jaccard score is calculated by dividing the number of elements in the intersection set by the number of elements in the union set.

## Pairwise cross comparison of batches

We further investigated BERM reproducibility by examining the stability of feature selection in dataset pairs. Six dataset pairs were created from four batches. A significant gene list was generated for each dataset pair using Student’s T-test. Jaccard scores were calculated to determine the similarity between the gene lists.





**Figure 4: Workflow for pairwise cross-comparison of batches.** (1) Six dataset pairs were obtained from four batches. (2) For each dataset pair, we applied BERMs to obtain seven BERM-treated dataset pairs and one control. (3) Student’s T-test for each gene was conducted between old and young samples to obtain p-values. (4) Genes which had p-values less than or equal to 0.05 were extracted and allocated to the significant gene list. (5) Jaccard scores were calculated to determine the similarity between significant gene lists of each BERM-treated mega-dataset.

## Probe-batch effect correlation check

We examined how batch effect manifested and checked for the existence of a conserved batch-correlated gene set.12 pairwise combinations of same-class-different-batch (SCDB) samples were derived (Table 4). Since each SCDB combination contained samples from the same class but different batches, any variation in each combination arose mostly from batch effect. PCA was performed on each SCDB combination to investigate the pattern of gene loadings to the PCs. Using Student’s T-test on samples separated by batch, we obtained significant gene lists from each SCDB combination. We first calculated the Jaccard scores between the significant gene lists to determine the degree of similarity. This was to check if the SCDB combinations reported similar sets of genes as significantly correlated with batch effect.

We then used the SCDB combinations to examine if there was a correlation between the number of probes matching to a gene and the tendency of that gene to be correlated with batch effect. The significant gene lists were combined in union and the frequency at which a gene appeared in this combined gene set was calculated. We plotted gene frequency against the number of probes matching to the gene to observe if there was a correlation.

**Table 4: Table showing the 12 SCDB combinations used to check for existence of a conserved batch-correlated gene set.** Each SCDB combination contained samples from the same class, and they originated from two batches. For example, the first SCDB combination shown in this table comprised of young samples from both the Gonzalez and Mercken batches.

|  |  |  |
| --- | --- | --- |
| **SCDB Combination** | **Class** | **Batch Pairs** |
| 1 | Young | Gonzalez\_Mercken |
| 2 | Young | Gonzalez\_Raz |
| 3 | Young | Gonzalez\_Hubal |
| 4 | Young | Mercken\_Raz |
| 5 | Young | Mercken\_Hubal |
| 6 | Young | Raz\_Hubal |
| 7 | Old | Gonzalez\_Mercken |
| 8 | Old | Gonzalez\_Raz |
| 9 | Old | Gonzalez\_Hubal |
| 10 | Old | Mercken\_Raz |
| 11 | Old | Mercken\_Hubal |
| 12 | Old | Raz\_Hubal |

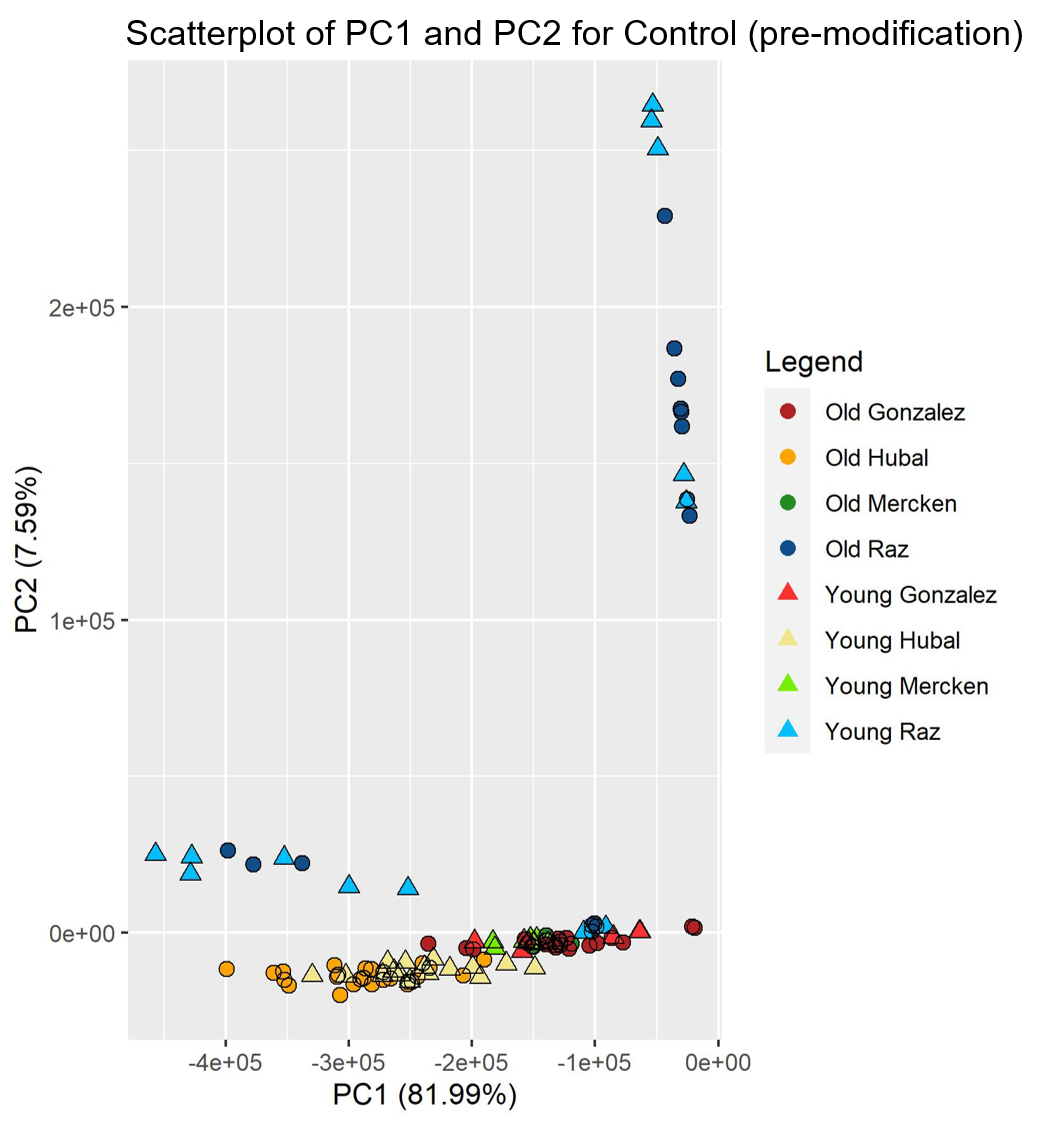
## Check for biological relevance of differential genes

We performed Student’s T-test to extract significant genes (p-value less than 0.05) from the BERM-treated mega-datasets and passed them into the Database for Annotation, Visualization, and Integrated Discovery (DAVID) tool (Dennis Jr et al., 2003). This was conducted to check if the differential features extracted were meaningful. The lack of biologically relevant functional annotations could indicate low biological variability in the mega-dataset, which might render it meaningless for feature selection-based mega-analyses.

# Results

## Strong batch effects were generated after combining datasets from different studies

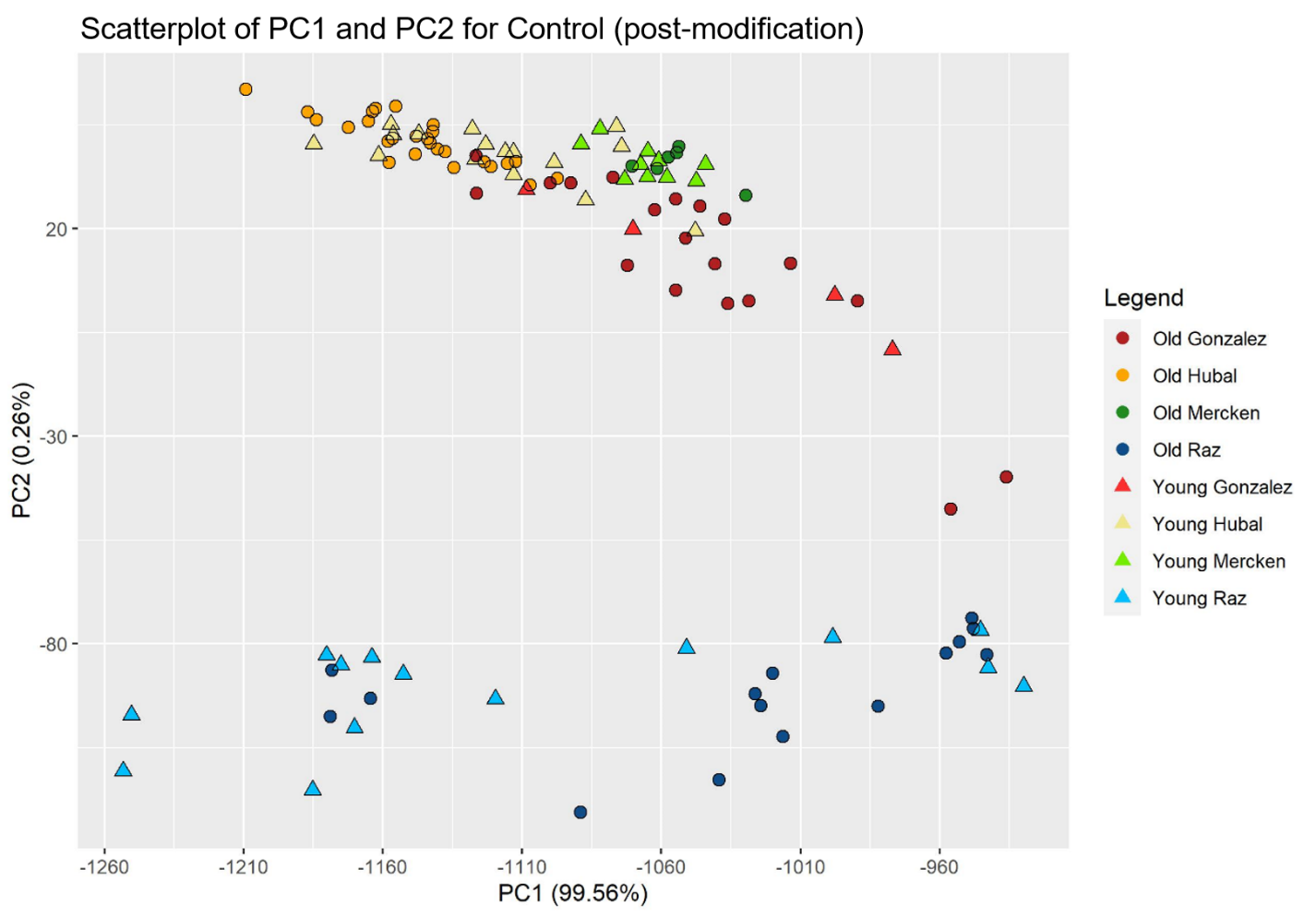
We visualised the components of variance in each BERM-treated mega-dataset as scatterplots. The overall variance explained by the top PCs (PC1 and PC2) was displayed as percentages along the axes. The significance of batch and class correlation for each PC was summarised as binary tables. The PCA scatterplot for Control is shown in Figure 5.



**Figure 5: PCA scatterplot for mega-dataset after log2-transformation and prior to BERM treatment.** Triangles represent samples classified as young and circles represent samples assigned as old. The hues denote the batch which each sample originates from. Light and dark red points represent samples from the Gonzalez batch; light and dark orange points represent samples from the Hubal batch; light and dark green points represent samples from the Mercken batch; light and dark blue points represent samples from the Raz batch.

Figure 5 showed clear intra-sample separation within the Raz batch along PC1 and PC2 that was not attributed to age. 15 Raz samples were clustered closely with samples from the other batches while the remaining 14 Raz samples segregated distinctly from other samples. This was surprising as the other batches did not display such intra-sample separation. We suspected the presence of batch effect manifesting among Raz samples. However, metadata for the Raz batch did not yield a satisfactory explanation as the Raz samples were not split according to gender, which was the only available phenotype data aside from age. Hence, we modified our protocol and applied ComBat to Raz dataset to remove this unknown intra-sample batch effect before combining the study datasets to form our mega-dataset.

The PCA scatterplot and binary table of the modified mega-dataset are shown in Figure 6 and Table 5 respectively. Most samples clustered markedly by batches and separation between classes was indistinct. Nearly all (99.56%) of the variance in Control was explained by PC1, which was correlated with batch effect. No top PCs were correlated with class effect (Table 5). This indicated that combining datasets from different studies resulted in strong inter-study batch effects in the mega-dataset.



**Figure 6: PCA scatterplot for Control after protocol modification.** After the protocol modification, the Raz batch did not exhibit prominent intra-sample batch effect. Triangles represent samples classified as young and circles represent samples assigned as old. The hues denote the batch from which each sample originated. Light and dark red points represent samples from the Gonzalez batch; light and dark orange points represent samples from the Hubal batch; light and dark green points represent samples from the Mercken batch; light and dark blue points represent samples from the Raz batch.

**Table 5: PCA binary table for Control displaying up to the tenth PC.** ‘0’ represents a Student’s T-test p-value of more than 0.05 while ‘1’ represents a p-value of less than or equal to 0.05. All top three PCs were correlated significantly with batch effect and none were correlated with class effect.

|  |  |  |
| --- | --- | --- |
| **PC** | **Class** | **Batch** |
| 1 | 0 | **1** |
| 2 | 0 | **1** |
| 3 | 0 | **1** |
| 4 | 0 | **1** |
| 5 | 0 | 0 |
| 6 | **1** | **1** |
| 7 | 0 | **1** |
| 8 | 0 | 0 |
| 9 | 0 | **1** |
| 10 | 0 | 0 |

## CSQN-ComBat gave the best outcome for batch effect removal

To compare the effectiveness of batch effect removal across all seven BERMs, we summarised the PCA binary tables of the BERM-treated mega-datasets in Table 6.

**Table 6: Summary of PCA binary tables of the seven BERM-treated mega-datasets.** Depending on the values in the binary table, the correlation of the top three PCs with class and batch in each mega-dataset was summarised as ‘TRUE’ or ‘FALSE’. For example, none of the top three PCs in the QN-treated mega-dataset were significantly correlated with class effect, thus this was denoted as ‘FALSE’. At least one of the top three PCs was significantly correlated with batch effect and this was denoted as ‘TRUE’. Ideal results were bolded; it was ideal for at least on of the top three PCs to be correlated with class effect and it was ideal for none of the top three PCs to be correlated with batch effects.

|  |  |  |
| --- | --- | --- |
| **BERM** | **Top three PCs were class-correlated** | **Top three PCs were batch-correlated** |
| QN | FALSE | TRUE |
| ZN | FALSE | TRUE |
| ComBat | FALSE | **FALSE** |
| QN-ComBat | FALSE | **FALSE** |
| ZN-ComBat | FALSE | **FALSE** |
| CSQN | **TRUE** | TRUE |
| CSQN-ComBat | **TRUE** | **FALSE** |

As expected, conventional normalisation methods such as QN and ZN yielded mega-datasets with variance mostly contributed by technical bias. All top three PCs in both mega-datasets were correlated with batch effect (Table 6). The biological variability in both mega-datasets was not preserved as none of the top PCs were correlated with class.

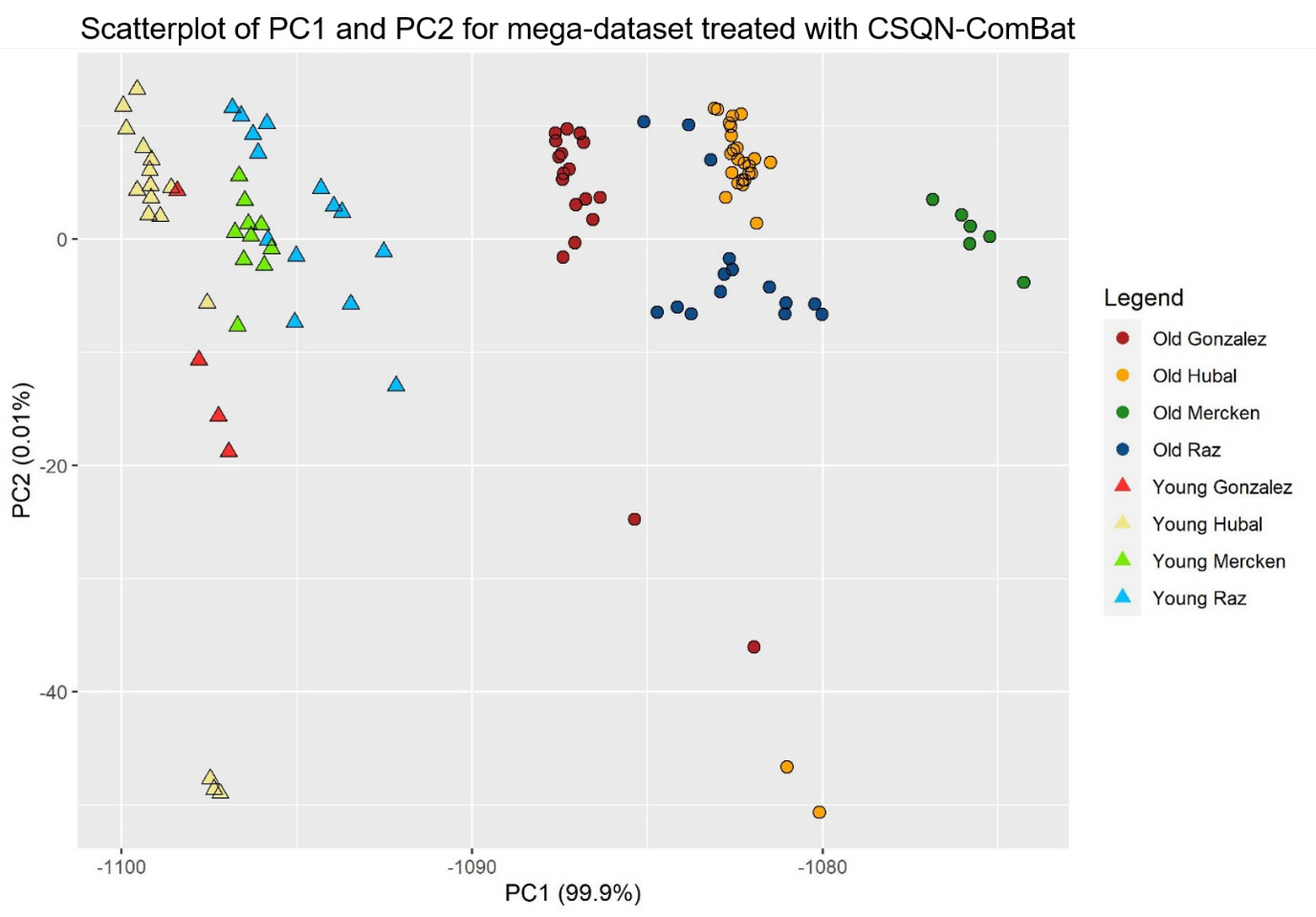
When the mega-dataset was treated with ComBat, even though batch-correlated PCs were ideally relegated to lower ranks, there was no improvement in preserving biological variability (Table 6). Combination of upstream QN or ZN with ComBat (QN-ComBat and ZN-ComBat respectively) remained ineffective at boosting class-correlated PCs (Table 6). This outcome corroborated with our expectation that an effective upstream normalisation technique is critical for preserving biological signal.

Notably, CSQN was the only normalisation technique that preserved biological variability as class correlation was observed in PC1 of the CSQN-treated mega-dataset. However, a significant proportion of the variance in the dataset was also associated with batch effect (Table 6). This meant that the biological signal in the mega-dataset was confounded with technical bias.

Combination of CSQN and ComBat proved to be the most ideal. At least one of the top three PCs for the mega-dataset treated with CSQN-ComBat correlated with class effect and none was significantly associated with batch effect (Table 6). The synergy between CSQN and ComBat is illustrated in Table 7. Although batch effect was removed from the top 10 PCs, the class-correlated PC for ComBat was ranked lowly at 10. For the CSQN approach, while PC1 was correlated with class, all top ten PCs were also associated with batch effect. After combining both BERMs, the class-correlated PC was elevated to the first rank and association with batch effect was made insignificant in all top 10 PCs. This suggested that to improve signal-to-noise ratio, individual strengths of CSQN and ComBat should be harnessed by applying CSQN as the upstream normalisation technique and ComBat as the batch effect cleaning method. This was further exemplified in the PCA scatterplot for the mega-dataset treated with CSQN-ComBat (Figure 7); all samples were perfectly delineated by classes along PC1. The majority (99.9%) of the data variance was explained by the class-correlated PC1, compared to 95.56% explained by batch-correlated PC1 in Control (Figure 6). This indicated that batch effect was removed successfully by CSQN-ComBat while preserving biological variability.

**Table 7: PCA binary table for ComBat-, CSQN- and CSQN-ComBat-treated mega-datasets up to the top 10 PCs.** Although ComBat could effectively remove batch effect judging from the lack of significant correlation of the PCs with batch, class effect was not apparent. CSQN displayed the opposite result, as class effect was captured in PC1, but all the PCs were batch-correlated. Combining both ComBat and CSQN harness the strength of each BERM and resulted in good batch cleaning as well as retention of class effect.

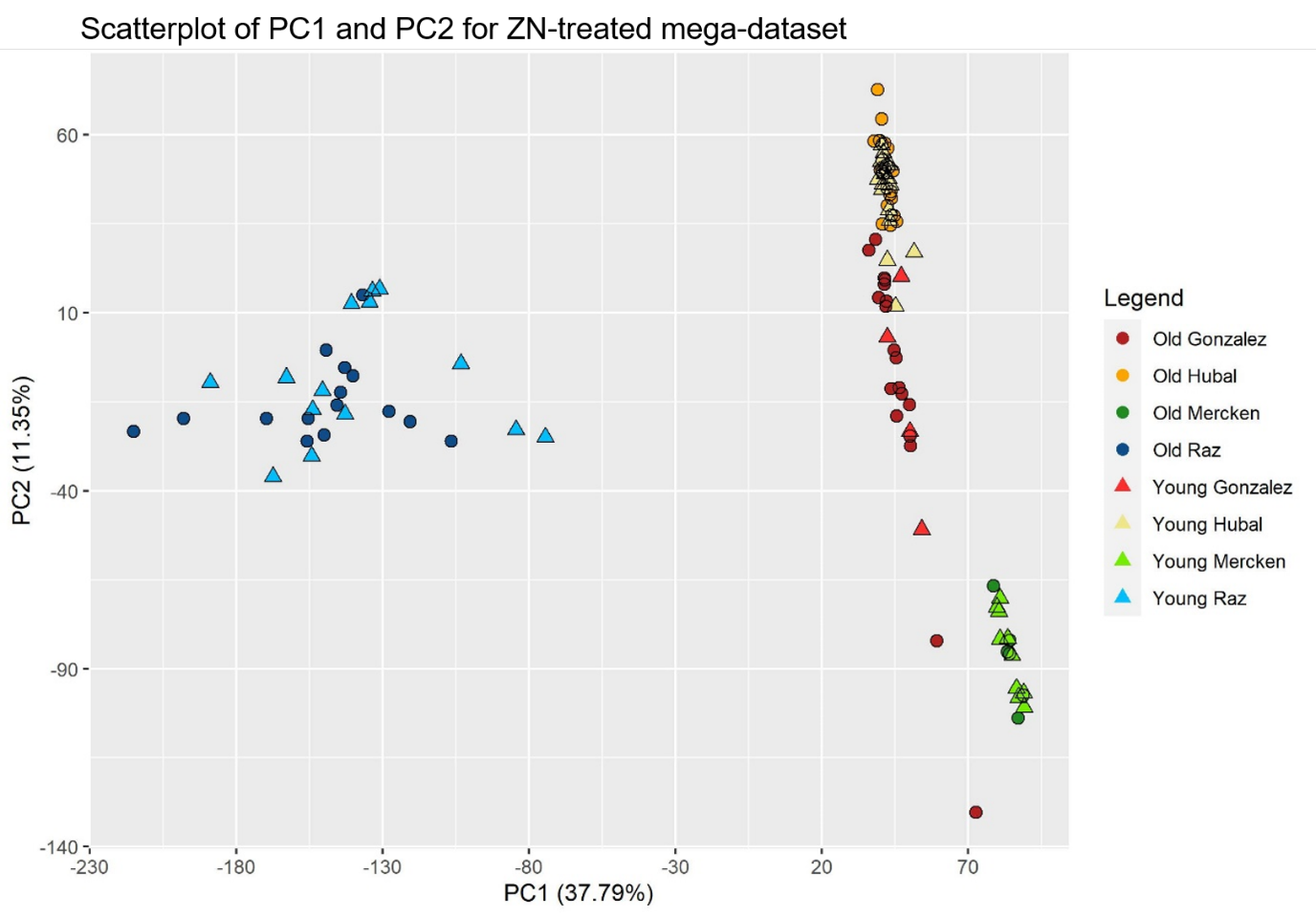
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **PC** | **Significant association between data factors and PCs** | | | | | |
| **ComBat** | | **CSQN** | | **CSQN + ComBat** | |
| **Class** | **Batch** | **Class** | **Batch** | **Class** | **Batch** |
| 1 | 0 | 0 | **1** | **1** | **1** | 0 |
| 2 | 0 | 0 | 0 | **1** | 0 | 0 |
| 3 | 0 | 0 | 0 | **1** | 0 | 0 |
| 4 | 0 | 0 | 0 | **1** | 0 | 0 |
| 5 | 0 | 0 | 0 | **1** | 0 | 0 |
| 6 | 0 | 0 | 0 | **1** | 0 | 0 |
| 7 | 0 | 0 | 0 | 0 | **1** | 0 |
| 8 | 0 | 0 | **1** | **1** | **1** | 0 |
| 9 | **1** | 0 | **1** | 0 | 0 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 |



**Figure 7: PCA scatterplot for mega-dataset treated with CSQN-ComBat.** There was a perfect separation between young and old samples along PC1. However, some batch clustering remained among old samples. Triangles represent samples classified as young and circles represent samples assigned as old. The hues denote the batch from which each sample originated. Light and dark red points represent samples from the Gonzalez batch; light and dark orange points represent samples from the Hubal batch; light and dark green points represent samples from the Mercken batch; light and dark blue points represent samples from the Raz batch.

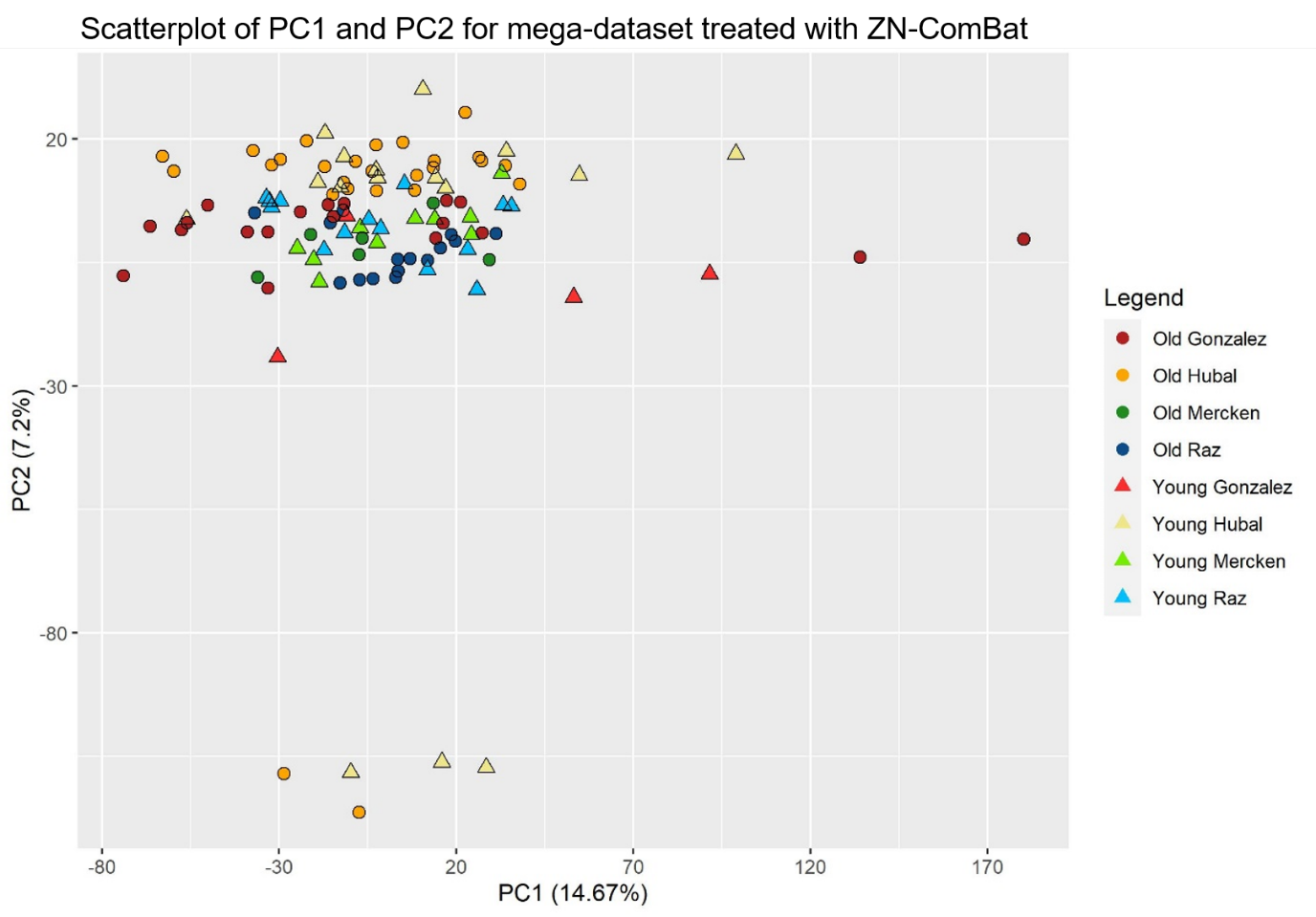
## Row-wise Z-normalization does not eradicate batch effects in mega-analysis

Row-wise ZN has been used in a number of microarray studies for data pre-processing (Tylee et al., 2017) (Hess et al., 2016). We tested our hypothesis that even though ZN can centre gene expression values, it has minimal effect at eradicating batch effect. We applied our PCA workflow (Figure 2) on the ZN-treated mega-dataset. Overall variance explained by PC1 decreased considerably from 95.56% (Figure 6) to 37.79% (Figure 8). Samples along PC1 and PC2 still showed distinct separation by batch. Along PC1, Raz and Mercken samples were clearly separated from other batches. Mercken samples were also segregated from other batches along PC2. In general, division between old and young samples was not apparent.



**Figure 8: PCA scatterplot for ZN-treated mega-dataset.** A pronounced segregation between batches was seen along PC1 and PC2. There was minimal separation by classes. Triangles represent samples classified as young and circles represent samples assigned as old. The hues denote the batch from which each sample originated. Light and dark red points represent samples from the Gonzalez batch; light and dark orange points represent samples from the Hubal batch; light and dark green points represent samples from the Mercken batch; light and dark blue points represent samples from the Raz batch.

The PCA scatterplot for the mega-dataset treated with ZN-ComBat showed attenuation in batch effect as samples were not visually clustered by batches (Figure 9). However, T-test of the top three PCs showed that significant association with batch effects still existed (Table 6). Similar to the PCA scatterplot of ZN-treated mega-dataset (Figure 8), the separation by class remained indistinct. Both BERMs thus were not effective at alleviating batch effect and subpar at preserving biological variation in the mega-dataset.

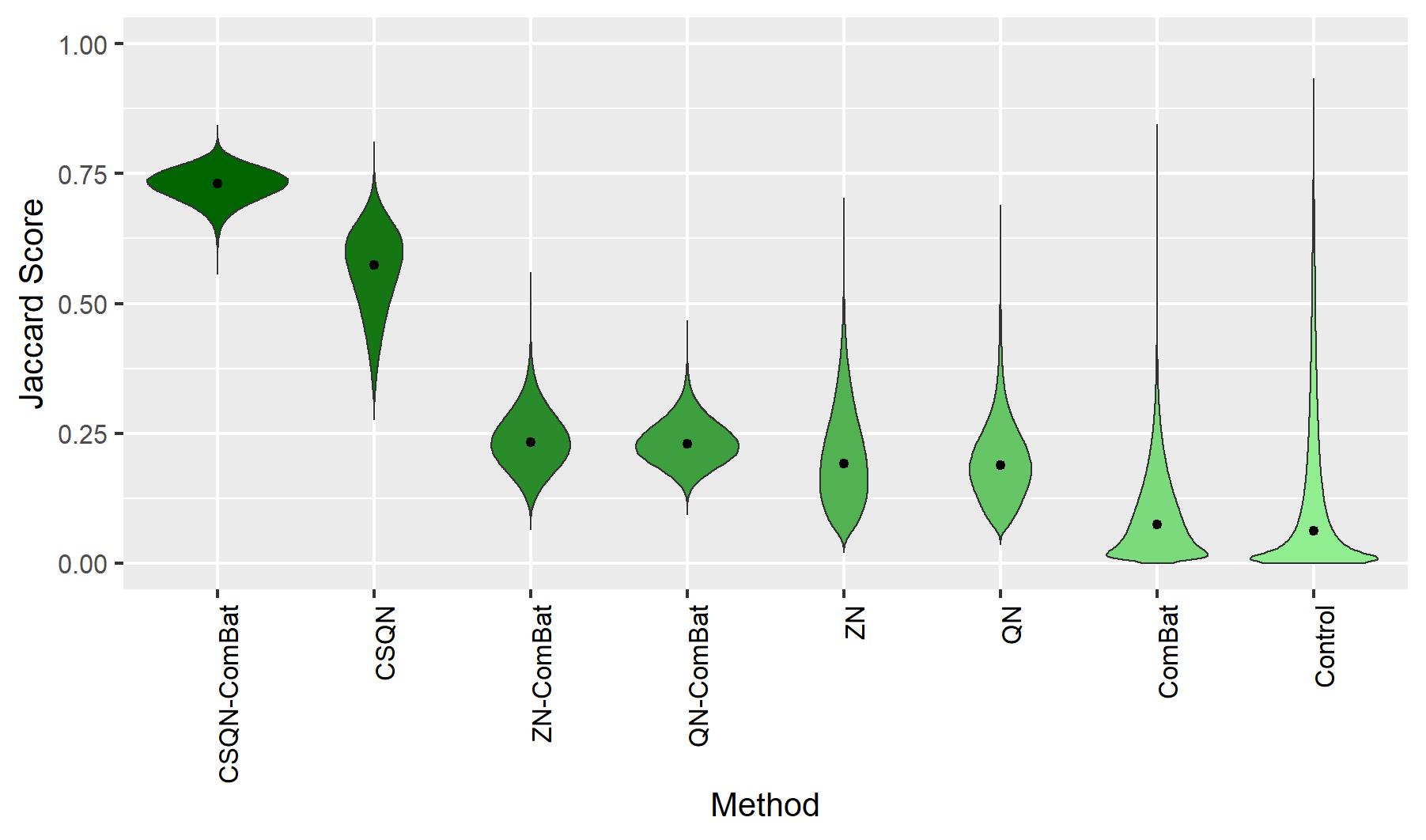


**Figure 9: PCA scatterplot for mega-dataset treated with ZN-ComBat.** Even though batch separation was alleviated, the samples still did not aggregate by classes. Triangles represent samples classified as young and circles represent samples assigned as old. The hues denote the batch from which each sample originated. Light and dark red points represent samples from the Gonzalez batch; light and dark orange points represent samples from the Hubal batch; light and dark green points represent samples from the Mercken batch; light and dark blue points represent samples from the Raz batch.

## Bootstrap analyses showed that CSQN-ComBat conferred high reproducibility

To evaluate overall reproducibility of BERMs on total combined data, we conducted bootstrapping and Jaccard scoring (Figure 3) on Control and on BERM-treated mega-dataset. The resultant violin plot is shown in Figure 10.

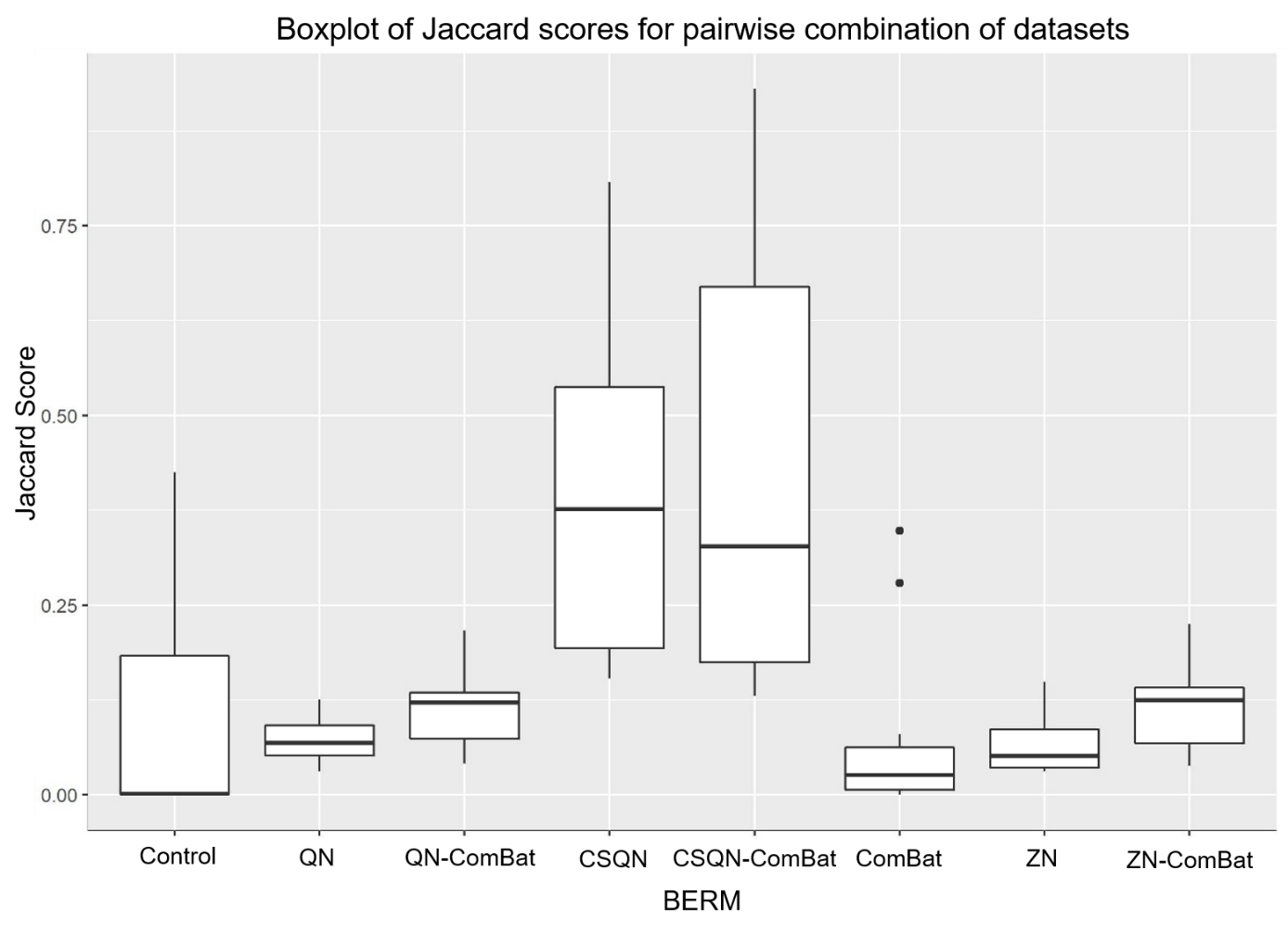
For Control, Jaccard scores obtained among bootstraps were the lowest with a median score of 0.06 and largest standard deviation of 0.16. This was followed by ComBat with a median score of 0.07. Even though ComBat displayed promising capability in reducing batch effect (Table 6), the lack of signal preservation may have translated to low reproducibility. Conventional normalisation methods, ZN and QN, yielded higher reproducibility than ComBat. Combination of ComBat with ZN or QN only marginally improved reproducibility. Expectedly, CSQN appeared to be the performance differential; CSQN and CSQN-ComBat generated significantly higher Jaccard scores than other BERMs, with CSQN-ComBat generating the highest median (0.73) and smallest standard deviation (0.03). This suggested that the batch-effect cleaning performance of the CSQN-ComBat method was the most stable across all BERMs. Combined with the optimal PCA result, CSQN-ComBat appeared to be the most effective and reliable method to remove batch effect while preserving biological variation in mega-datasets.



**Figure 10: Violin plot of Jaccard score for all BERMs.** The median of each BERM was shown as a black dot. The BERM with the highest median of Jaccard score was CSQN-ComBat while ComBat yielded the lowest median score.

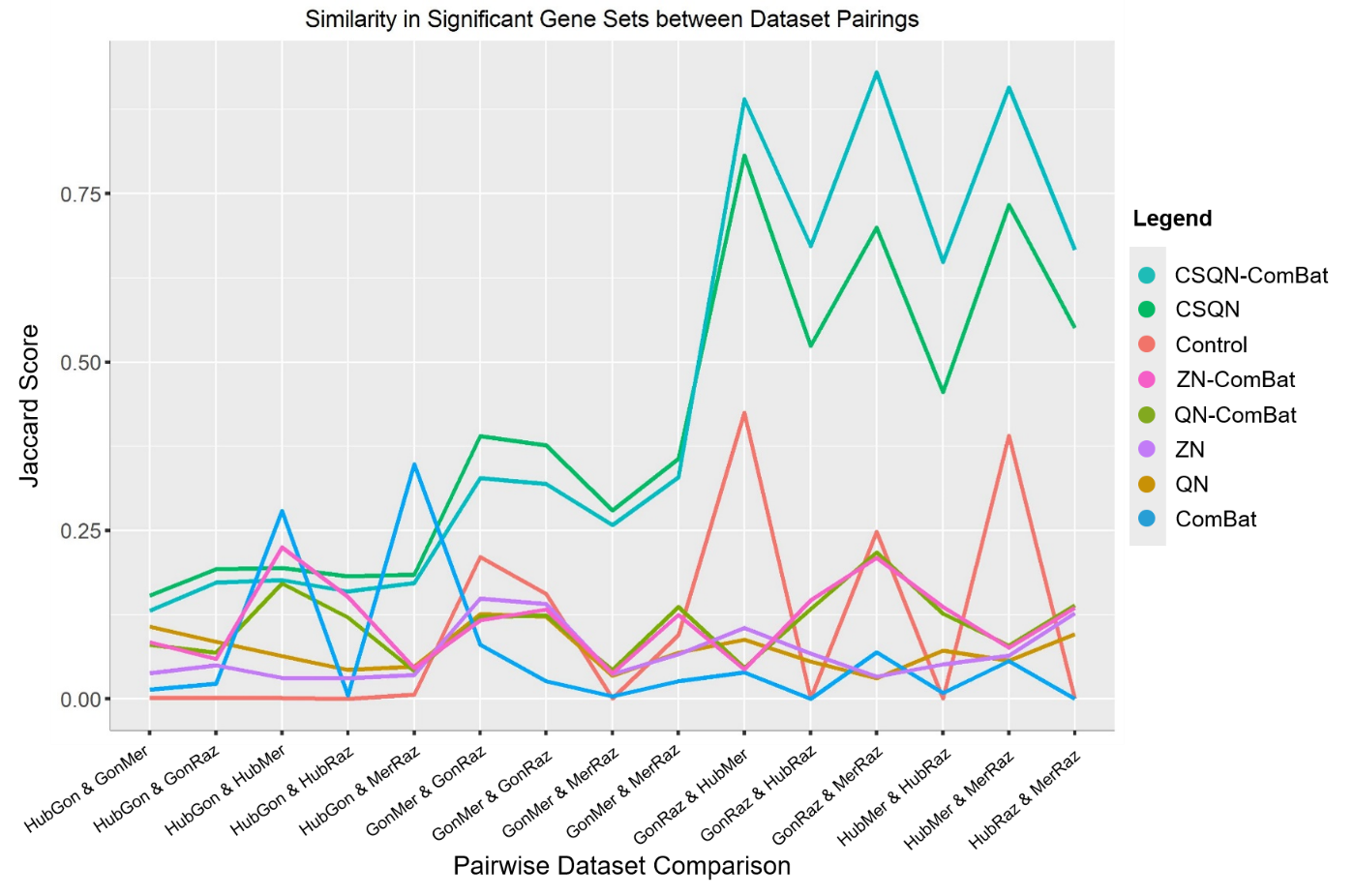
## Pairwise independent analyses showed that CSQN-ComBat conferred highest agreement rates

We subsequently investigated the performance stability of BERMs on pairwise combinations of datasets instead of the four-batch mega-dataset using the workflow in Figure 4. This was to further explore how the BERMs performed when applied on different mega-datasets. The significant gene sets for every two pairings were cross-compared and their similarity calculated via Jaccard scoring (Figure 11).



**Figure 11:** **Boxplot showing the Jaccard scores when significant gene sets from BERM-treated dataset pairs were compared.** The BERM with the lowest Jaccard scores was ComBat. BERMs comprising of ZN and QN also yielded low Jaccard scores. Although CSQN and CSQN-ComBat had significantly higher scores than other BERMs, their variances were much larger.

The Jaccard scores for pairwise analyses followed a similar trend as the whole data analyses. The median score for Control was the lowest, followed by ComBat. As expected, QN, ZN, QN-ComBat and ZN-ComBat displayed consistently low Jaccard scores for all dataset pairs. CSQN and CSQN-ComBat exhibited significantly higher Jaccard scores than the other BERMs due to non-overlapping of interquartile ranges. Surprisingly, their variances were abnormally high. For example, CSQN-ComBat had scores ranging from 0.13 to 0.95. We suspected that this was due to a large disparity of scores between pairwise comparisons. Thus, we reorganised the plot and separated Jaccard scores by pairwise combinations (Figure 12).



**Figure 12:** **The Jaccard scores were further separated by paired dataset comparisons.** Each line represents the Jaccard scores of a BERM. While CSQN and CSQN-ComBat mostly yielded consistent higher Jaccard scores than other BERMs, the scores were highly variable based on the paired dataset comparisons.

The Jaccard scores for mega-datasets treated with CSQN and CSQN-ComBat varied widely based on the combination of pairwise datasets. For example, comparing the similarity of significant gene lists between Hubal-Gonzalez and Gonzalez-Mercken paired datasets gave a low Jaccard score of 0.13 for CSQN-ComBat. The comparison between Gonzalez-Raz and Mercken-Raz instead yielded a high score of 0.90 for the same BERM. This variability was less obvious in other BERMs such as QN and ZN as Jaccard score remained relatively constant.

## Functional analyses of significant genes in CSQN + ComBat-treated dataset

Functional analyses of significant genes from BERM-treated mega-datasets were conducted using the DAVID tool. This was to validate the biological relevance of differential features extracted from BERM-treated mega-datasets. We compared functional analyses of mega-datasets treated with ZN, one of the worst-performing BERMs, and CSQN-ComBat, which was the best-performing BERM. The tables showing the gene ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) pathway terms in the top three annotation clusters are shown in Table 8 and Table 9.

**Table 8: DAVID functional annotation for the significant gene list for CSQN-ComBat-treated dataset.** 972 significant genes were extracted from the mega-dataset treated with CSQN-ComBat. They were then passed into the DAVID platform for functional analysis. Similar GO and KEGG terms were clustered, and their biological importance was ranked using the Enrichment Score. Only the top three clusters are shown in this table. Gene count represents the number of genes enriched for a certain functional annotation.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Annotation Cluster** | **Enrichment Score** | **Top functional annotation terms overrepresented in cluster for CSQN-ComBat** | | | **Gene Count** | **P-value** |
| 1 | 61.75 | GO:0006413 | Translational initiation | 90 | | 5.5E-79 |
| GO:0006614 | SRP-dependent co-translational protein targeting to membrane | 74 | | 3.1E-74 |
| GO:0000184 | Nuclear-transcribed mRNA catabolic process, nonsense-mediated decay | 77 | | 1.3E-66 |
| GO:0019083 | Viral transcription | 72 | | 6.5E-62 |
| GO:0005840 | Ribosome | 84 | | 4.0E-58 |
| 2 | 22.27 | GO:0006119 | Oxidative phosphorylation | 62 | | 5.6E-33 |
| hsa05012 | Parkinson's disease | 62 | | 5.5E-31 |
| hsa05010 | Alzheimer's disease | 61 | | 2.6E-25 |
| hsa05016 | Huntington's disease | 63 | | 2.0E-23 |
| GO:0006120 | Mitochondrial electron transport, NADH to ubiquinone | 24 | | 7.9E-17 |
| 3 | 11.99 | GO:0042776 | Mitochondrial ATP synthesis coupled proton transport | 17 | | 7.3E-17 |
| GO:0006754 | ATP biosynthetic process | 16 | | 3.4E-12 |
| GO:0042776 | ATP synthesis coupled proton transport | 12 | | 4.4E-9 |

**Table 9: DAVID functional annotation for the significant gene list for ZN-treated dataset.** Feature selection was performed on the ZN-treated mega-dataset and 777 genes were extracted. These genes were passed into the DAVID platform for functional analysis. Similar GO and KEGG terms were clustered, and their biological importance was ranked using the Enrichment Score. Only the top three clusters are shown in this table. Gene count represents the number of genes enriched for a certain functional annotation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Annotation Cluster** | **Enrichment Score** | **Top Functional Annotation Terms Overrepresented in Cluster for ZN** | | **Gene Count** | **P-value** |
| 1 | 1.92 | GO:0060079 | Excitatory postsynaptic potential | 6 | 3.1E-3 |
| GO:0051930 | Regulation of sensory perception of pain | 5 | 1.4E-2 |
| GO:0019233 | Sensory perception of pain | 6 | 4.1E-2 |
| 2 | 1.15 | GO:0061337 | Cardiac conduction | 7 | 5.6E-3 |
| GO:0086091 | Regulation of heart rate by cardiac conduction | 4 | 1.3E-1 |
| hsa04261 | Adrenergic signaling in cardiomyocytes | 6 | 4.8E-1 |
| 3 | 1.12 | hsa04024 | cAMP signaling pathway | 12 | 6.1E-2 |
| hsa04020 | Calcium signaling pathway | 11 | 7.0E-2 |
| GO:0001975 | Response to amphetamine | 4 | 1.0E-1 |

Gene clusters, which were ranked by enrichment scores, showed the key biology of interest in the datasets examined. The top and third gene clusters for the mega-dataset treated with CSQN-ComBat were relevant to constitutive cellular functions such as RNA transcription and mitochondrial ATP synthesis, respectively (Table 8). The second gene cluster was significantly associated with KEGG pathway terms for neurodegenerative disorders such as Parkinson’s, Alzheimer’s, and Huntington’s disease (Table 8).

Contrarily, functional annotation terms for the top gene cluster of ZN-treated mega-dataset were relevant to pain perception, followed by clusters associated with cardiac function and amphetamine response. However, some terms in the clusters might have been spurious due to a low number of genes matching to the same terms. This might have exhibited as high p-values (more than 0.05). For example, in Cluster 2, only four genes out of 777 genes were related to the regulation of heart rate by cardiac conduction (Table 9). This biological ambiguity among significant genes in the ZN-treated mega-dataset might have been caused by the minimal class effect present after ZN treatment.

## There were no inherent biases suggesting existence of batch-correlated genes or probes

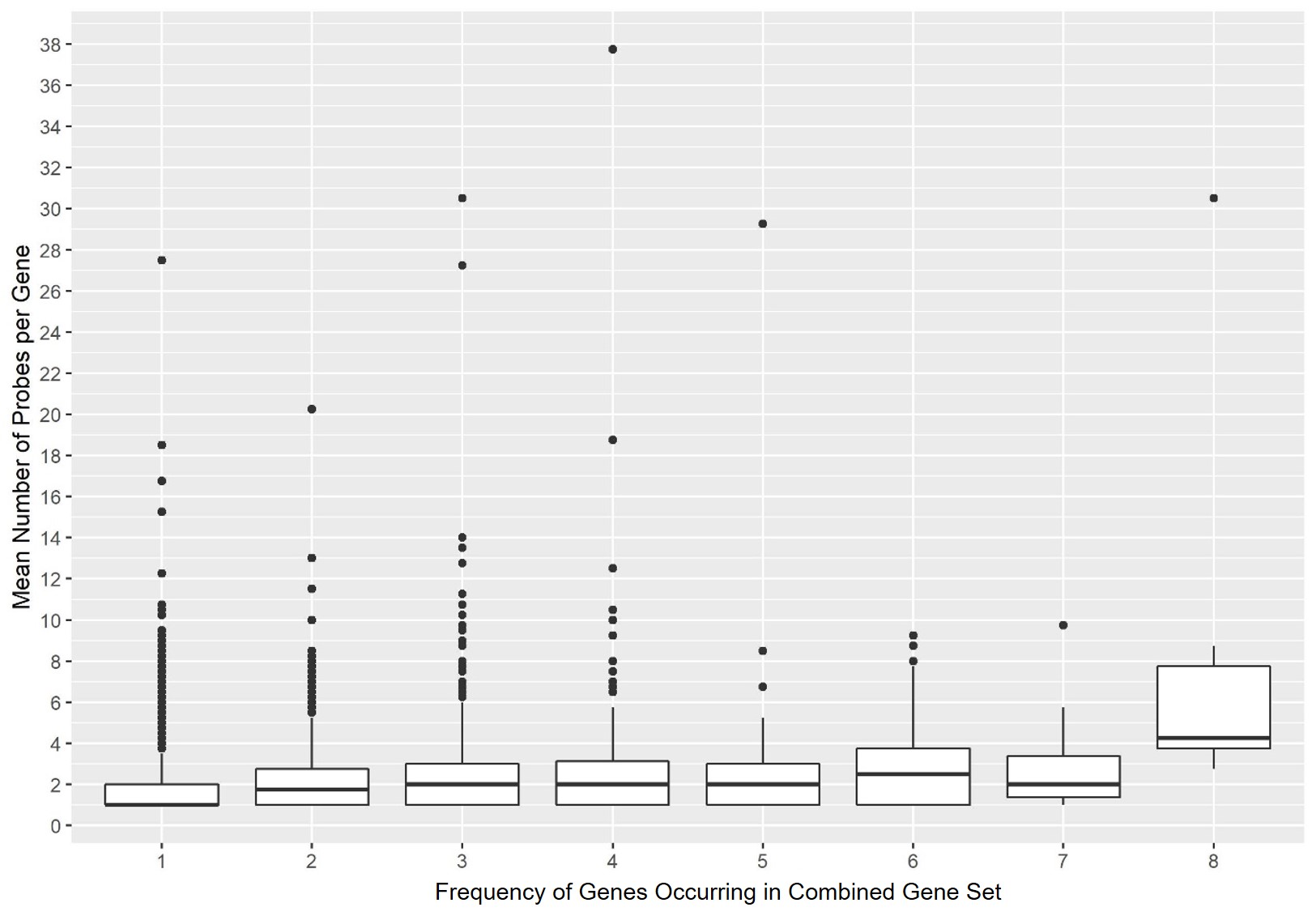
We investigated how batch effect manifested among genes to find out if there existed a conserved gene set that disproportionately loaded more to batch effect. PCA analyses showed that in the majority of SCDB combinations, distribution of batch effects among genes was not uniform. Some genes loaded greater than others to data variance, leading to plots skewed towards the left (Figure 13). This implied a possible distinct set of batch-correlated genes. However, we later cross-compared the genes reported as batch-correlated across the 12 SCDB combinations and found that these gene sets were dissimilar as the median Jaccard score was 0.18. This suggested that even though there existed a distinct set of genes which influenced batch effect more strongly, interestingly, the identity of these genes changed from dataset to dataset.

Shape

Description automatically generated

**Figure 13: An example plot of gene loadings for the SCDB combination comprising of young samples from Gonzalez and Hubal batches.** Gene index refers to the numbering of the gene in the dataset. 21053 genes were examined. The loadings skewed to the left, which indicated that gene loadings were not uniform and there was a set of genes that loaded more to the batch-correlated PC.

We then investigated whether the number of probes matching to a gene affected the likelihood of it being reported as batch-correlated. We combined the significant gene sets across all SCDB pairs into a union set and calculated how frequent each gene appeared. The data was visualized in a boxplot (Figure 14). A higher probe count for a gene did not seem to be correlated with a higher frequency of this gene to appear in the union set. There thus appeared to be no correlation between the number of probes matching to a gene and the tendency for that gene to be reported as batch-correlated. This suggested that the probe number was not a factor influencing how strongly a gene is associated with batch effect.



**Figure 14:** **Boxplot of number of probes matching to a gene against the frequency of occurrence in the combined gene set.** There was no visible correlation between probe number and gene frequency. For example, genes with more than five probes matched to it did not cause an increase in the frequency of genes appearing in the union gene set.

# Discussion

## Whole-data and gene-wise normalization methods did not alleviate batch effect

Multiple mega-analysis studies incorporated gene-wise standardization in the data pre-processing step in order to normalize the range and variance of expression values (Tylee et al., 2017) (Hess et al., 2016). We have shown that such standardization techniques conducted gene-wise was not effective at priming datasets for mega-analysis, even if ComBat is applied downstream. ZN showed minimal effectiveness at eradicating batch effect; PCA analyses showed that batch effect remained the predominant source of data variance. This batch effect removal performance was minimally enhanced using ZN-Combat (Table 2 and Figure 8). Both ZN and ZN-ComBat were ineffective at retaining biological signal. A similar suboptimal performance was also observed for mega-datasets treated with QN and QN-ComBat.

Such underperformance might be due to the nature of applying normalization methods in a whole-data manner. This whole-data normalization assumes that all samples follow an identical feature value distribution regardless of their class. ZN and QN were applied on the whole mega-dataset and both BERMs normalized each sample to the same target distribution, instead of normalizing samples to their respective class distribution. This might have suppressed the variation by class. Therefore, even though the subsequent use of ComBat could correct some batch effect, the biological variability was diminished. Functional analysis on the ZN-treated mega-dataset showed few biologically meaningful GO and KEGG pathway terms (Table 9). This suggested that using whole-data normalization methods for mega-analysis can lead to a detrimental effect in subsequent selection of differential features.

## ComBat alone was not good enough to retain biological variability

Our PCA analyses showed that ComBat significantly alleviated batch effect (Table 2). This result concurs with current literature. For example, Chen et al found that ComBat outperformed five other algorithms designed for batch effect correction (distance-weighted discrimination, support vector machines, mean-centering, surrogate variable analysis and Ratio\_G) (C. Chen et al., 2011). ComBat excelled in experimental and simulated microarray expression data by robustly improving the correlation among replicates. Optimal cross-platform performance was also observed for ComBat on both the Illumina BeadChip and Affymetrix platforms (Kitchen et al., 2010). Thus, the superior batch effect correction of ComBat is certain. However, due to its subpar ability in preserving biological signal, it did not fare well at reproducing similar significant gene sets (Figure 10). Thus, ComBat is not a reliable BERM for feature selection-based mega-analysis.

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

Batch-correction performance and reproducibility were significantly boosted when ComBat was coupled with upstream CSQN. In this study, we showed that CSQN performed well at preserving the biological signals. This optimal outcome might be due to the application of QN on two classes of samples separately; unlike QN, the feature value distribution was assumed to be different for different biological classes. The robustness of CSQN as a normalization method that retains biological signal has been demonstrated in existing literature. CSQN 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 (Zhao et al., 2020). Hence, we recommend the usage of CSQN over whole-data normalization methods as a critical upstream normalization to preserve biological variability.

The downside of CSQN was its inability to effectively remove batch effect (Table 2). This limitation was overcome by coupling CSQN with ComBat. 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 (Table 8). Functional analysis of the top significant genes showed significant relevance with neurodegenerative disorders such as Parkinson’s and Alzheimer’s disease which are known to disproportionately affect the elderly (Dato et al., 2013). Relevance with mitochondrial function was also elucidated, which was unsurprising as mitochondrial disruptions and oxidative stress have been shown to cause neurodegenerative diseases (Carvalho et al., 2015). This suggested that CSQN-ComBat, with its ability to retain biological signals while removing batch effect, can effectively prime mega-datasets for subsequent feature selection.

The combination of QN methods with ComBat is not a new practice. Müller et al examined the use of QN prior to the application of ComBat (Müller et al., 2016). However, unlike our study which examined class-specific QN, they explored batch-wise QN coupled with ComBat and revealed 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 microarray study by Sun et al (Sun et al., 2011). They found that while the commonly used *lumi* method of QN outperformed other QN methods to reduce batch effect, *lumi* 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 can be 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.

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

Our PCA results showed that the batch-correlated gene set differed among datasets. We also examined a possible correlation between the number of probes matching to a gene and the tendency of that gene to load highly to batch effect. Our results indicated that this correlation was spurious. Similar to Leek et al, we concluded that batch effect generally does not manifest uniformly among genes (Leek et al., 2010). We further discovered that no probe-induced batch effect was present.

## Study limitations

Despite evaluating multiple BERMs, our study was limited to transcriptomics data. Due to inter-platform differences, such as different number of variables measured, sources of technical bias, and 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. For example, a study comparing mitochondrial RNA and protein expression for genes involved in lung adenocarcinomas showed little correlation (r = -0.025) (G. Chen et al., 2002). This indicates that transcriptomics and proteomics data may not be cross-comparable. Thus, we will consider comparing the BERM performance across different platforms and prescribe optimal BERMs unique to the properties of each platform.

In this study, we only used real data and thus could not quantify the absolute amount of batch and class effect before and after BERM application. This hindered us from evaluating performance in terms of precision and recall. We could only determine the relative effectiveness of batch effect removal by doing comparisons between BERMs. Future work could include simulating data and applying a known amount of batch and class effect.

Another study limitation was 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 (Figure 7), some clustering by batch still remained among old samples even though T-test conducted on the top PCs indicated that they were not significantly correlated with batch (Table 7). P-values have also shown to be highly unstable (van Helden, 2016) and this might have affected the reliability of our statistical benchmarking tests.

Efforts in microarray data analysis are shifting to focus on subnetworks of genes with related functions (Lim et al., 2014) rather than individual genes or probes. More feature engineering possibilities can also be unleashed when microarray data analysis is combined with machine learning (Turgut et al., 2018). Future work on the performance of BERMs should thus be benchmarked in the context of such paradigms.

# Conclusions

The combination of datasets from different experiments inevitably creates batch effect that confounds with true biological signal. Such data pooling used in mega-analysis creates reproducibility issues if batch effects are not adequately corrected. BECAs 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 aimed to compare the effectiveness of BERMs in removing batch effect in a combined transcriptomics dataset while retaining biological variation. We also intended to evaluate the reproducibility of their performance. We managed to show that while ComBat could eradicate batch effect, it was subpar at retaining biological signals. We found that combining upstream CSQN with ComBat (CSQN-ComBat) helped to preserve of biological variability. CSQN-ComBat was also found to be the most reproducible among all the BERMs examined. The biological validity of the mega-dataset treated with CSQN-ComBat was confirmed as we could derive a gene signature biological relevant to aging via functional analysis. We thus recommend CSQN-ComBat as the best and most reliable approach for removing batch effects in combined datasets for mega-analysis studies.

# References

Amaratunga, D., & Cabrera, J. (2001). Analysis of Data from Viral DNA Microchips. *Journal of the American Statistical Association, 96*(456), 1161-1170. Retrieved from <http://www.jstor.org/stable/3085879>

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

Carvalho, C., et al. (2015). The role of mitochondrial disturbances in Alzheimer, Parkinson and Huntington diseases. *Expert Review of Neurotherapeutics, 15*(8), 867-884. doi:10.1586/14737175.2015.1058160

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

Chen, C., et al. (2011). Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods. *PloS one, 6*(2), e17238. doi:10.1371/journal.pone.0017238

Chen, G., et al. (2002). Discordant Protein and mRNA Expression in Lung Adenocarcinomas \*. *Molecular & Cellular Proteomics, 1*(4), 304-313. doi:<https://doi.org/10.1074/mcp.M200008-MCP200>

Clarke, R., et al. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature reviews. Cancer, 8*(1), 37-49. doi:10.1038/nrc2294

Collas, P., et al. (2007). Programming the genome in embryonic and somatic stem cells. *Journal of cellular and molecular medicine, 11*(4), 602-620. doi:10.1111/j.1582-4934.2007.00079.x

Dato, S., et al. (2013). Exploring the role of genetic variability and lifestyle in oxidative stress response for healthy aging and longevity. *International Journal of Molecular Sciences, 14*(8), 16443-16472.

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

Dunning M, L. A., Eldridge M. (2015). illuminaHumanv4.db: Illumina HumanHT12v4 annotation data (chip illuminaHumanv4) (Version R package version 1.26.0). Retrieved from <https://bioconductor.org/packages/release/data/annotation/html/illuminaHumanv4.db.html>

Goh, W. W. B., et al. (2017). Can Peripheral Blood-Derived Gene Expressions Characterize Individuals at Ultra-high Risk for Psychosis? *Computational Psychiatry, 1*, 168-183. doi:10.1162/CPSY\_a\_00007

Goh, W. W. B., et al. (2017). Why Batch Effects Matter in Omics Data, and How to Avoid Them. *Trends in Biotechnology, 35*(6), 498-507. doi:<https://doi.org/10.1016/j.tibtech.2017.02.012>

Goh, W. W. B., & Wong, L. (2018). Dealing with Confounders in Omics Analysis. *Trends in Biotechnology, 36*(5), 488-498. doi:10.1016/j.tibtech.2018.01.013

Gonzalez-Freire, M., et al. (2018). Skeletal muscle ex vivo mitochondrial respiration parallels decline in vivo oxidative capacity, cardiorespiratory fitness, and muscle strength: The Baltimore Longitudinal Study of Aging. *Aging Cell, 17*(2). doi:10.1111/acel.12725

Halsey, L., et al. (2015). *The fickle P value generates irreproducible results* (Vol. 12).

Hess, J. L., et al. (2016). Transcriptome-wide mega-analyses reveal joint dysregulation of immunologic genes and transcription regulators in brain and blood in schizophrenia. *Schizophrenia research, 176*(2-3), 114-124.

Hubal, M. J., et al. (2016). Pyruvate Dehydrogenase Phosphatase Regulatory Gene Expression Correlates with Exercise Training Insulin Sensitivity Changes. *Med Sci Sports Exerc, 48*(12), 2387-2397. doi:10.1249/mss.0000000000001041

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Johnson, W. E., et al. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics, 8*(1), 118-127. doi:10.1093/biostatistics/kxj037

Kitchen, R. R., et al. (2010). Correcting for intra-experiment variation in Illumina BeadChip data is necessary to generate robust gene-expression profiles. *BMC Genomics, 11*(1), 134. doi:10.1186/1471-2164-11-134

Leek, J. T., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics, 11*(10), 733-739. doi:10.1038/nrg2825

Leek, J. T., & Storey, J. D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. *PLOS Genetics, 3*(9), e161. doi:10.1371/journal.pgen.0030161

Lim, K., & Wong, L. (2014). Finding consistent disease subnetworks using PFSNet. *Bioinformatics, 30*(2), 189-196. doi:10.1093/bioinformatics/btt625

Lu, T., et al. (2004). Gene regulation and DNA damage in the ageing human brain. *Nature, 429*(6994), 883-891. doi:10.1038/nature02661

Mercken, E. M., et al. (2017). Conserved and species-specific molecular denominators in mammalian skeletal muscle aging. *NPJ Aging Mech Dis, 3*, 8. doi:10.1038/s41514-017-0009-8

Moody's Investors Service. (2014). *Aging will reduce economic growth worldwide in the next two decades*. Retrieved from <https://www.moodys.com/research/Moodys-Aging-will-reduce-economic-growth-worldwide-in-the-next--PR_305951>

Müller, C., et al. (2016). Removing Batch Effects from Longitudinal Gene Expression - Quantile Normalization Plus ComBat as Best Approach for Microarray Transcriptome Data. *PloS one, 11*(6), e0156594-e0156594. doi:10.1371/journal.pone.0156594

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

Nosek, B. A., & Errington, T. M. (2017). Making sense of replications. *eLife, 6*, e23383. doi:10.7554/eLife.23383

Peixoto, L., et al. (2015). How data analysis affects power, reproducibility and biological insight of RNA-seq studies in complex datasets. *Nucleic Acids Research, 43*(16), 7664-7674. doi:10.1093/nar/gkv736

Povey, S., et al. (2001). The HUGO gene nomenclature committee (HGNC). *Human genetics, 109*(6), 678-680.

Raz, V., et al. (2014). Major aging-associated RNA expressions change at two distinct age-positions. *BMC Genomics, 15*, 132. doi:10.1186/1471-2164-15-132

Stewart, L. A., & Clarke, M. J. (1995). Practical methodology of meta-analyses (overviews) using updated individual patient data. Cochrane Working Group. *Stat Med, 14*(19), 2057-2079. doi:10.1002/sim.4780141902

Sun, Z., et al. (2011). Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Medical Genomics, 4*(1), 84. doi:10.1186/1755-8794-4-84

Turgut, S., et al. (2018, 18-19 April 2018). *Microarray breast cancer data classification using machine learning methods.* Paper presented at the 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT).

Tylee, D. S., et al. (2017). 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, 174*(3), 181-201. doi:10.1002/ajmg.b.32511

van Helden, J. (2016). Confidence intervals are no salvation from the alleged fickleness of the P value. *Nature Methods, 13*(8), 605-606. doi:10.1038/nmeth.3932

Zhao, Y., et al. (2020). How to do quantile normalization correctly for gene expression data analyses. *Scientific Reports, 10*(1), 15534. doi:10.1038/s41598-020-72664-6

Zhou, L., et al. (2019). Examining the practical limits of batch effect-correction algorithms: When should you care about batch effects? *Journal of Genetics and Genomics, 46*(9), 433-443. doi:<https://doi.org/10.1016/j.jgg.2019.08.002>