Qing Xia, Jeffrey A. Thompson and Devin C. Koestler*

Batch effect reduction of microarray data with dependent samples using an empirical Bayes approach (BRIDGE)

<https://doi.org/10.1515/sagmb-2021-0020>

Received March 26, 2021; accepted October 29, 2021; published online December 14, 2021

Abstract: Batch-effects present challenges in the analysis of high-throughput molecular data and are particularly problematic in longitudinal studies when interest lies in identifying genes/features whose expression changes over time, but time is confounded with batch. While many methods to correct for batch-effects exist, most assume independence across samples; an assumption that is unlikely to hold in longitudinal microarray studies. We propose Batch effect Reduction of microarray data with Dependent samples using Empirical Bayes (*BRIDGE*), a three-step parametric empirical Bayes approach that leverages technical replicate samples profiled at multiple timepoints/batches, so-called “bridge samples”, to inform batch-effect reduction/attenuation in longitudinal microarray studies. Extensive simulation studies and an analysis of a real biological data set were conducted to benchmark the performance of *BRIDGE* against both *ComBat* and *longitudinal ComBat*. Our results demonstrate that while all methods perform well in facilitating accurate estimates of time effects, *BRIDGE* outperforms both *ComBat* and *longitudinal ComBat* in the removal of batch-effects in data sets with bridging samples, and perhaps as a result, was observed to have improved statistical power for detecting genes with a time effect. *BRIDGE* demonstrated competitive performance in batch effect reduction of confounded longitudinal microarray studies, both in simulated and a real data sets, and may serve as a useful preprocessing method for researchers conducting longitudinal microarray studies that include bridging samples.

Keywords: batch effect correction; COMBAT; longitudinal gene expression; temporal microarray data.

1 Introduction

High-throughput technologies are used to assess the genome, transcriptome, metabolome, proteome, and epigenome and have led to an unprecedented understanding of the role of genetic variants, gene and protein expression and epigenetic modifications involved in various biological processes and response to treatment/exposures. The reliability and accuracy of measurements obtained from such technologies are contingent upon standardized experimental protocols, well-defined treatment conditions, the quality and integrity of both the biological material and any reagents used in the experiment, and highly trained personnel that conduct the experiments (Leek et al. 2010). Variability in the previously mentioned factors leads to both biological and technical variation in downstream measurements of the molecular mechanism of interest (Irizarry et al. 2005; Scherer 2009). In the ideal experimental setting, all samples are subjected to the

*Corresponding author: Devin C. Koestler, Department of Biostatistics & Data Science, University of Kansas Medical Center, 3901 Rainbow Blvd., Kansas City, KS 66160, USA, E-mail: dkoestler@kumc.edu

Qing Xia and Jeffrey A. Thompson, Department of Biostatistics & Data Science, University of Kansas Medical Center, 3901 Rainbow Blvd., Kansas City, KS 66160, USA

same exact experimental conditions, with only the biological/treatment conditions varied across study subjects. This helps to ensure that any differences in the measurement of interest (e.g., gene/protein expression, metabolite concentration, DNA methylation, etc.) arise as a result of variation in the underlying biological/treatment conditions. Due to constraints on the number of samples that can be processed in parallel or within the same batch – for example, usually only 96 or 384 samples can be loaded on well plates in RNA amplification and labeling, and in the scanning step, 48 or 172 samples can be scanned in each batch (Müller et al. 2016) – samples are frequently analyzed in different batches. This type of technical variability, commonly referred to as batch effects, is often a major source of variability in high-throughput experiments (Leek et al. 2010), with batch effects having been reported across a variety of different high-throughput technologies, including: microarray (Li et al. 2009; Scherer 2009), next-generation sequencing (Taub et al. 2010), single-cell sequencing (Hicks et al. 2018) and DNA methylation (Sun et al. 2011). If left uncorrected, batch effects introduce noise that can dilute the detection of biological signals, leading to a reduction in statistical power or worse, misleading, biased, or non-reproducible results (Goh et al. 2017). These issues are magnified when there are either flaws in the initial study design or when the study involves longitudinal profiling and batch is confounded with time, such as in studies (Banchereau et al. 2016; Tasaki et al. 2018; Wong et al. 2020).

An understanding of the statistical ramifications of batch-effects across a wide range of different technologies (Akey et al. 2007; Baggerly et al. 2004) has motivated the development of many statistical methods for correcting/mitigating batch effects in downstream statistical analyses (Beer et al. 2020; Fei and Yu 2020; Johnson et al. 2007; Wang and Lê Cao 2020; Zhang et al. 2018). Johnson et al. proposed *ComBat* which estimates and adjusts the gene-wise location and scale parameters in a location/scale (LS) model using an empirical Bayes procedure for batch effects correction (Johnson et al. 2007). Location and scale parameters represent additive and multiplicative batch effects, respectively. Briefly, *ComBat* is a three-step empirical Bayes based approach that involves the following steps: (1) data standardization with the goal of increasing the amount of systematic batch information that can be borrowed across genes, (2) obtaining empirical Bayes based estimates of the prior distribution hyperparameters from the standardized data using methods of moment estimation, followed by the estimation of the location/scale parameters, and (3) adjusting the data for batch effects. Numerous independent studies have benchmarked the performance of *ComBat* to competing batch-effect adjustment/attenuation approaches, with many of these studies demonstrating favorable performance of *ComBat* (Chen et al. 2011; Müller et al. 2016; Zhou et al. 2019). Although *ComBat* assumes independence across samples, its performance for correcting batch effects in longitudinal gene expression data was recently benchmarked and compared to several other batch effect correction/attenuation approaches, where a combination of quantile normalization and *ComBat* exhibited the best performance (Müller et al. 2016). Despite its many strengths, *ComBat* was not designed for studies with dependent samples and does not account for within-subject repeated measures. As a result, *ComBat* may overcorrect additive and multiplicative batch effects in studies involving longitudinal repeat measures sampling on study subjects (Beer et al. 2020).

Longitudinal studies are important for measuring within-subject temporal change due to genetic, environmental, and stochastic effects, illustrating individual trends over time, and to compute within-subject variability (Hejblum et al. 2015). Several approaches have been recently proposed to correct batch effects in longitudinal studies using a linear mixed model (LMM) framework. For example, Venkatraman et al. (2015) estimated fixed effects for batch using LMM applied to longitudinal data and then applied a mean estimate correction to new data (Venkatraman et al. 2015). Accounting for within-subject variability led to improved batch-effect estimates in longitudinal compared to cross-sectional data. However, this method does not adjust the multiplicative batch effects. Beer et al. proposed *longitudinal ComBat*, which is an extension of the *ComBat* methodology from a cross-sectional to a longitudinal setting, and which accounts for within-subject repeated measures by including a subject-specific random effects (Beer et al. 2020). *Longitudinal ComBat* demonstrated improved statistical power for detecting longitudinal changes, as well as improved type I error rate control, compared to *ComBat*. Despite its strengths, *longitudinal ComBat* was primarily validated assuming a subject-specific random intercept model formulation and thus, its performance for more complicated study designs, assuming other types of random effects structures, remains unclear.

To circumvent the limitations of existing batch-effect correction methods for studies involving the longitudinal collection of high-dimensional ‘omic data, we propose *BRIDGE*. *BRIDGE* corrects batch effects for a completely batch-confounded longitudinal study with the help of “bridging samples” – technical replicate samples that are measured in multiple batches. After correcting for batch effects with *BRIDGE*, adjusted data can be used for downstream statistical analyses as if all samples are measured in a single batch. In principle, *BRIDGE* is applicable to different ‘omics platforms, such as microarray gene expression, DNA methylation or neuroimaging studies of neurodevelopment, as long as the transformed data are high-dimensional and approximately normally distributed. In what follows, we outline the steps of the proposed *BRIDGE* methodology and then present the results from an extensive set of simulation studies aimed at benchmarking the performance of *BRIDGE* against both *ComBat* and *longitudinal ComBat* under a variety of different scenarios. The performance of *BRIDGE* is further compared to *ComBat* and *longitudinal ComBat* using a real biological data set from a longitudinal gene expression study of rheumatoid arthritis.

2 Methods

2.1 Study description and nomenclature

Figure 1A shows the general framework of the type of studies the *BRIDGE* methodology was created to address. Suppose a batch-confounded longitudinal study with a total of N participants is conducted to understand the temporal effects of molecular markers (e.g., gene expression (Banchereau et al. 2016; Karlovich et al. 2009; Tasaki et al. 2018), DNA methylation (Bjornsson et al. 2008), or RNA-seq (Wong et al. 2020)). To simplify the problem, we assume there are two time points and two batches in the study, which are confounded with one another. Samples from the N participants collected at time point 1 are profiled in batch 1 and samples collected from the same N participants at time point 2 are profiled in batch 2. To strip the time effects of the molecular markers from completely confounded batch effects, a bridge between two batches is needed. Suppose there exist samples that are collected as technical replicates for $M < N$ of the study participants at time point 1 and that these samples are profiled for expression, DNA methylation, etc., in both batch 1 and batch 2. We hereafter refer to these M samples as “bridging samples” as they serve as a bridge between batch 1 and batch 2. Correspondingly, we refer to samples collected from the remaining $N - M$ participants collected at time-point 1 and profiled in batch 1, but not re-analyzed as technical replicates in batch 2, as “non-bridging samples”. We refer to the paired measurements from the M participants whose samples were run as technical replicates in both batches as the “bridging data”. Figure 1B lists the factors affecting observed gene expression differences among data.

2.2 Model specification

The classic location-and-scale adjustment model (L/S) assumes that batch effects exert both additive effects, which shift the mean gene expression, and multiplicative effects, which scale the variance of gene expression (Johnson et al. 2007). We augment the classic L/S model for batch effects reduction in a longitudinal microarray study. Due to the difficulty in estimation when there are two sources of within-subject variation, arising here from both batch and time, the L/S model used here only accounts for dataset-dependency (the dataset created from each batch or each time point share common effects because of the same experimental technique and environmental conditions), ignoring within-subject correlation. Denoting $Y_{bij}^{(t)}$ as the expression levels of gene $j = 1, 2, \dots, J$ for sample $i = 1, 2, \dots, N$ in batch $b = 1, 2$ at time point $t = 1, 2$, the L/S model is formulated as:

$$Y_{bij}^{(t)} = \mu_j^{(t)} + \varphi_{bj} + \delta_{bj}\epsilon_{ij}^{(t)}$$

where $\mu_j^{(t)}$ is the overall time effect of gene j at time t , φ_{bj} is the batch effect for batch b gene j , δ_{bj} is the multiplicative batch effect for batch b gene j , and $\epsilon_{ij}^{(t)} \sim N(0, \sigma_j^{2(t)})$ is the error term of gene j in subject i . For simplicity, we absorb multiplicative batch effects and standard deviation of gene j into $\omega_{bj}^{(t)}$, where $\omega_{bj}^{(t)} = \delta_{bj}\sigma_j^{(t)}$. The L/S model given above can be further augmented to include additional covariates (e.g., sex, subject age, etc.). See Additional File S1 for details, along with how parameter estimates obtained from the L/S model that includes covariates are used in downstream data standardization steps.

2.3 Batch effects adjustment

Figure 1C shows the workflow of the *BRIDGE* method. First, the bridging data is standardized to increase the amount of systematic batch information that can be borrowed across genes. Next, a Bayesian linear regression is fit gene/feature-wise using the

Figure 1: Framing the assumed study design and workflow. (A) General framework of the type of studies that *BRIDGE* is suitable for. (I) Sample collection is preformed following the same standard operating procedures at time point 1 & 2. (II) Sample processing. All N samples collected at time point 1 (blue) are profiled in batch 1 (orange) and all N samples collected at time point 2 (red) are profiled in batch 2 (purple). A subset of samples ($M < N$) collected at time point 1 are profiled in both batches as technical replicates (aka “bridging samples”). Correspondingly, the samples collected from the remaining $N - M$ participants collected at time point 1 and profiled in batch 1, but not re-analyzed in batch 2, are termed “non-bridging samples”. We refer to the paired measurements from the M bridging samples as “bridging data”. (B) The expected factor effects within the data. (C) Workflow of the *BRIDGE* methods. Step 1: parameter estimation of time effects $\mu^{(t)}$, standard deviation $\sigma^{(t)}$, additive batch effect φ_2 , and multiplicative batch effects δ_2 for later use in data standardization, step 2: Empirical Bayes (EB) based estimation of linear regression parameters $\alpha^*, \beta^*, \tau^{*2}$ and calculation of EB estimates of additive and multiplicative batch effects φ_2^*, δ_2^* and step 3: Direct adjustment the data for batch effects (*BRIDGE-1*) or alternatively, obtaining prediction/imputation-based data for non-bridging samples (*BRIDGE-2*) for downstream statistical analysis.

standardized bridging data treating the batch 2 standardized data as the response and the batch 1 standardized data as the explanatory variable. Alternatively, ordinary least squares (OLS) estimation could be applied to standardized bridging data, however we have found that estimates obtained from a Bayesian linear regression approach tend to be more reliable (data not shown). Regardless of approach, within-subject variation due to batch is reflected through the regression coefficients of the described regression models, which characterize the nature of batch effects for a given gene/feature. Because batch-effects are believed to impact genes/features in a systematic way (Johnson et al. 2007), and thus similar regression coefficients would be expected across genes/features, information can be borrowed across all genes/features as a means toward obtaining more stable estimates. To accomplish this task, an empirical Bayes procedure is applied to the gene/feature-wise regression coefficients obtained from the previous step in order to obtain empirical Bayes-based estimates of regression coefficients, and subsequently, empirical Bayes-based estimates of the additive and multiplicative batch effect parameters. These estimates are then used to directly adjust the data to remove/attenuate batch-effects (BRIDGE-1), or to predict/impute the data for the non-bridging samples (BRIDGE-2) for downstream statistical analyses.

Step 1: Standardize the data

Across the genes/features, expression values could differ due to the differences in $\mu_j^{(t)}$ or $\sigma_j^{(t)}$. Differences in the underlying distribution of expression across genes undermine our goal of borrowing information across genes via empirical Bayes. To address this issue, we first standardize the gene expression data to harmonize expression data across all genes; that is, the expression levels of all genes share similar mean and variance. Assuming a study with two batches and two time points, we first estimate $\mu_j^{(1)}, \mu_j^{(2)}, \varphi_{2j}, \sigma_j^{(1)}$ and $\sigma_j^{(2)}$, constraining $\varphi_{1j} = 0$ and $\delta_{1j} = 1$ to obtain unique estimates for each parameter. Refer to Additional File S2 for derivations of estimates. Batch 1 is set as the reference batch, whose additive and multiplicative batch effects are assumed to be absent. With estimates of the aforementioned parameters, the standardized data $Z_{bij}^{(t)}$ can be calculated by:

$$Z_{bij}^{(t)} = \frac{Y_{bij}^{(t)} - \hat{\mu}_j^{(t)} - \hat{\varphi}_{bj}}{\hat{\sigma}_j^{(t)}} + \hat{\varphi}_{bj},$$

where the marginal distribution of $Z_{bij}^{(t)} \sim N(\varphi_{bj}, \delta_{bj}^2)$.

Step 2: Bayesian linear regression

We assume that the standardized bridging data $\mathbf{Z}_{M \times 2}$ follows a bivariate normal distribution,

$$\mathbf{Z}_{M \times 2} = (\mathbf{Z}_{1j}^{(1)}, \mathbf{Z}_{2j}^{(1)}) = \begin{pmatrix} Z_{11j}^{(1)} & Z_{21j}^{(1)} \\ \vdots & \vdots \\ Z_{1sj}^{(1)} & Z_{2sj}^{(1)} \\ \vdots & \vdots \\ Z_{1Mj}^{(1)} & Z_{2Mj}^{(1)} \end{pmatrix}$$

More specifically, for each gene/feature, the measurements obtained from the two batches are assumed to follow:

$$(Z_{1j}^{(1)}, Z_{2j}^{(1)}) \sim N_2 \left(\begin{bmatrix} \varphi_{1j} \\ \varphi_{2j} \end{bmatrix}, \begin{bmatrix} \delta_{1j}^2 & \rho_1 \delta_{1j} \delta_{2j} \\ \rho_1 \delta_{1j} \delta_{2j} & \delta_{2j}^2 \end{bmatrix} \right)$$

The mean of the conditional distribution of $Z_{2j}^{(1)}$ given $Z_{1j}^{(1)}$, $E(Z_{2j}^{(1)} | Z_{1j}^{(1)}) = \varphi_{2j} - \rho_1 \frac{\delta_{2j}}{\delta_{1j}} (Z_{1j}^{(1)} - \varphi_{1j})$, is a linear function of $Z_{1j}^{(1)}$, while the variance $\text{Var}(Z_{2j}^{(1)} | Z_{1j}^{(1)}) = (1 - \rho_1^2) \delta_{2j}^2$ is constant. Here ρ_1 denotes the correlation coefficient of the standardized expression data between batch 1 and batch 2. For each gene/feature, we assume the following linear model:

$$Z_{2sj}^{(2)} | Z_{1sj}^{(1)} = \alpha_j + \beta_j Z_{1sj}^{(1)} + \varepsilon_{sj}$$

where $\varepsilon_{sj} \sim N(0, \tau_j^2)$, $\tau_j^2 = (1 - \rho_1^2) \delta_{2j}^2$. The linear coefficients α_j, β_j are given by:

$$\alpha_j = \varphi_{2j} - \frac{\rho_1 \delta_{2j} \varphi_{1j}}{\delta_{1j}} = \varphi_{2j} \quad \text{and} \quad \beta_j = \frac{\rho_1 \delta_{2j}}{\delta_{1j}} = \rho_1 \delta_{2j},$$

assuming $\varphi_{1j} = 0$ and $\delta_{1j} = 1$. Solving the system equations above in terms of $\varphi_{2j}, \delta_{2j}$, we have:

$$\varphi_{2j} = \alpha_j \quad \text{and} \quad \delta_{2j} = \sqrt{\beta_j^2 + \tau_j^2} \quad (1)$$

Parameters α_j, β_j and τ_j are estimated based on $\mathbf{Z}_{M \times 2}$ using Bayesian linear regression with an empirical prior. The empirical Bayes estimates of regression model parameters α_j^*, β_j^* and τ_j^* are given by the conditional posterior means, and empirical Bayes estimators of batch effect φ_{2j}^* and δ_{2j}^* are obtained from Eq. (1). Further elaboration and derivations of the empirical Bayes based estimates are given in Additional File S3.

Step 3: batch effects adjusted data

We propose two methods to adjust data for batch-effects depending on which batch is of superior quality. As previously described, both *BRIDGE-1* and *BRIDGE-2* set batch 1 as reference in **Step 1**. *BRIDGE-1* uses empirical Bayes estimates of the additive and multiplicative batch effect parameters φ_{2j}^* and δ_{2j}^* , respectively, to directly adjust the data for batch-effects. Specifically, for *BRIDGE-1*, the adjusted data is given by:

$$y_{bij}^{*(t)} = \frac{\hat{\sigma}_j^{(t)}}{\delta_{bj}^*} \left(Z_{bij}^{(t)} - \varphi_{bj}^* \right) + \hat{\mu}_j^{(t)} \quad (2)$$

Conversely, *BRIDGE-2* uses empirical Bayes estimates of regression model parameters, α_j^* and β_j^* , to predict/impute the standardized data for the non-bridging samples in batch 2, e.g., $\hat{Z}_{2qj}^{(1)}$, $q = 1, 2, \dots, N - M$, where $Z_{2qj}^{(1)} = \alpha_j^* + \beta_j^* Z_{1qj}^{(1)}$. Back-transformation of $\hat{Z}_{2qj}^{(1)}$ to obtain predicted expression values on original scale for the non-bridging samples, $\hat{y}_{2qj}^{(1)}$, is then obtained by:

$$\hat{y}_{2qj}^{(1)} = \left(\hat{Z}_{2qj}^{(1)} - \hat{\varphi}_{2j} \right) \hat{\sigma}_j^{(1)} + \hat{\varphi}_{2j} + \hat{\mu}_j^{(1)} \quad (3)$$

assuming $\varphi_{1j} = 0$ and $\delta_{1j} = 1$. Here $\hat{\sigma}_j^{(t)}$, $\hat{\mu}_j^{(t)}$ in Eq. (2) and $\hat{\mu}_j^{(1)}$, $\hat{\varphi}_{2j}$ in Eq. (3) reflect the estimates obtained in **Step 1**.

2.4 Downstream analysis – statistical inference on time effects

To test the null hypothesis of no difference in the mean expression of gene j between time point 1 and time point 2, that is $H_0: \Delta_j = \mu_j^{(2)} - \mu_j^{(1)} = 0$, we propose a paired t -test using the adjusted (*BRIDGE-1*) or predicted/imputed data (*BRIDGE-2*) obtained from **Step 3**. Here Δ_j represents the difference in the mean expression of gene j at time point 2 and the mean expression of gene j at time-point 1. Alternatively, a linear mixed model or an alternative modeling framework, could be used when there are covariates that one wishes to adjust/control for, or if data are collected at more than two time points.

2.5 Simulation studies

A preliminary simulation study was conducted to determine the number of iterations needed for Markov Chain Monte Carlo (MCMC) convergence for the coefficients in the Bayesian linear regression model. Our results showed that the MCMC algorithm converges very quickly; after 150 iterations as burn-in samples, we collect the next 350 iterations for posterior inference. We next compared the performance of *BRIDGE*, *ComBat* (Johnson et al. 2007) and *longitudinal ComBat* (Beer et al. 2020) in correcting batch effects in Simulation I & II. In our simulations, *ComBat* was fit using all samples, including both bridging and non-bridging samples as previously (Müller et al. 2016), with time as a single covariate. Similarly, *longitudinal ComBat* was fit using all samples and assumed a subject-specific random intercept model formulation to account for within-subject variation in repeated measurements. The simulation procedure and corresponding parameter settings for the simulated data sets are described in detail in Additional File S4. Additionally, Table 1 contains the parameter settings for both Simulation I & II. Briefly, for both Simulation I & II, we started by simulating standardized gene expression levels $Z_{bij}^{(t)}$ from a multivariate normal distribution, where the marginal distribution of $Z_{bij}^{(t)} \sim N(\varphi_{bj}, \delta_{bj}^2)$, the within-subject correlation structure containing within-subject correlation of batch ρ_1 , and within-subject correlation of time ρ_2 , assuming ρ_1 and ρ_2 are independent. Of the genes simulated, 20% were randomly selected to exhibit a time effect, $|\Delta_j| \neq 0$, and the remaining 80% of genes were simulated to exhibit no time effect, $|\Delta_j| = 0$. Hereafter, we use G to refer to the number of genes simulated to exhibit a time-effect.

In Simulation I, we simulated data with weak, moderate and strong time effects, each with a small or large multiplicative batch effect, and set batch 1 as reference batch (Table 1). We chose a relatively modest number of genes/features, $J = 100$, and bridging sample size, $M = 30$, to demonstrate the difference in batch effect removal/attenuation across methods *BRIDGE-1*, *BRIDGE-2*, *ComBat*, and *longitudinal ComBat*. Comparisons of methods were based on a principal component analysis (PCA) of the uncorrected and batch-corrected data; partial redundancy analysis (pRDA); the root mean square error (RMSE) of genes simulated to exhibit a time effect; and density plots of the time-effect estimates and their standard errors based on data that was corrected by *BRIDGE-1*, *BRIDGE-2*, *ComBat*, and *longitudinal ComBat*. In PCA, we visualized the uncorrected and corrected data by plotting principal component 1 (PC1) against principal component 2 (PC2) to visually assess the predominant sources of variation in our simulated data. pRDA is a multivariate method that was used here to globally assess variability in our simulated data attributable to both batch and time. With respect to pRDA, a smaller proportion of variance explained by batch and larger proportion of variance explained by time in the corrected data as compared uncorrected data, indicates a successful batch effects removal/attenuation. Our use of pRDA as a performance metric was primarily motivated by the study by Muller et al. (Müller et al. 2016), wherein a similar methodology, principal variance component analysis (PVCA), was used to estimate the proportion or variation explained by batch and time across several competing batch effect correction methods. As the primary goal of the type of studies we consider are centered on the identification genes whose expression changes over time, extra variability due to batch (e.g., batch effects) stands as an impediment to this objective. Further, since batch is confounded with time for the studies we consider, we also want to ensure that a given method does not completely dilute or wash-out time-effects, which provides justification for comparing the proportion of variance attributable to time after correction across the various methodologies.

Table 1: Parameter settings for simulation I & II.

Simulation I	Multiplicative Batch Effects	Time Effect Size	
	Large $\delta_2 = 2$	Weak $\Delta_0 = 1$ Moderate $\Delta_0 = 1.2$ Strong $\Delta_0 = 1.5$	
	Small $\delta_2 = 0.5$	Weak $\Delta_0 = 0.5$ Moderate $\Delta_0 = 0.6$ Strong $\Delta_0 = 0.8$	
	$\sigma_0^{(1)} = 0.8, \sigma_0^{(2)} = 1, G = 20, J = 100, \rho_1 = 0.8, \rho_2 = 0.4, \varphi_2 = 2, M = 30, N = 100, \mu_0^{(1)} = 2$		
Simulation II	Multiplicative batch effects	Within-subject correlation	Variance structure
	Large $\delta_2 = 2$	Weak $\rho_1 = 0.3$ Strong $\rho_1 = 0.9$	$\sigma_0^{(1)} = 0.2$ or 0.8 $\sigma_0^{(1)} = 0.2$ or 0.8
	Small $\delta_2 = 0.8$	Weak $\rho_1 = 0.3$ Strong $\rho_1 = 0.9$	$\sigma_0^{(1)} = 0.2$ or 0.8 $\sigma_0^{(1)} = 0.2$ or 0.8
	$\sigma_0^{(2)} = 1, G = 200, J = 1000, \rho_2 = 0.4, \varphi_2 = 5, M = \{5, 10, 20, 40\}, N = 100, \mu_0^{(1)} = 2$		

Note that $\mu_0^{(1)}$ is the overall time effect mean at time point 1, φ_2 is the additive batch effect mean for batch 2, $\sigma_0^{(1)}$ and $\sigma_0^{(2)}$ are the standard deviations of gene expression of time point 1 and time point 2, ρ_1 is within-subject correlation between batches, ρ_2 is within-subject correlation between time, M is bridging sample size, N is the number of participants in study, J is the number of genes/features and G is the number of truly differential genes/features, e.g., genes/features with a time-effect.

Our simulation studies were repeated 20 times for calculating pRDA, RMSE, and obtaining the density plot of estimated time effects. The estimated time effects were calculated as the subject-specific difference in expression between timepoints 1 and 2 for each of the G genes that were simulated to exhibit a “true” effect of time (e.g., genes whose expression changed over time). This was done based on the corrected expression data obtained from each of the four methods listed above, and across all of the 20 simulated data sets. For a given method, we report the mean time-effect of the G genes simulated to exhibit a “true” effect of time, as the mean of the $N \times G \times 20$ vector of subject-specific time-effects, where N denotes the total number of subjects. For a given method, the standard error was taken to be the standard deviation of the $N \times G \times 20$ vector of subject-specific time-effects. RMSE was calculated between the estimated time effects in the corrected, batch-effect adjusted data and the known, simulated time effects, separately for each method across all G differentially expressed genes and across all samples.

Simulation II was conducted to evaluate each batch-effect correction method’s performance in terms of downstream statistical inference. In contrast to *ComBat*, which assumes independence across the samples, *longitudinal ComBat* and the *BRIDGE* methods take within-subject variation into account. However, unlike *longitudinal ComBat*, the *BRIDGE* methods also allow nonequivalent variance structure and explicitly make use of bridging samples. The performance of *BRIDGE* methods depends on multiplicative batch effects (δ_2), which indicates which batch has superior quality. Along these lines, we set multiplicative batch effects for batch 1 as reference, $\delta_1 = 1$. A lower multiplicative batch effect parameter for batch 2 ($\delta_2 < 1$) indicates that batch 2 has superior quality over batch 1, e.g., batch 2 does not inflate the variance of gene expression compared to batch 1. Comparisons across the *BRIDGE* methods, *ComBat* and *longitudinal ComBat* were conducted for different assumed bridging sample size (M) and by assuming:

- simulated data with weak within-subject correlation of batch ($\rho_1 = 0.3$) and strong within-subject correlation of batch ($\rho_1 = 0.8$);
- simulated data with the assumed gene-specific standard deviations at time-points 1 and 2 ($\sigma^{(1)} = 0.8, \sigma^{(2)} = 1$) and ($\sigma^{(1)} = 0.2, \sigma^{(2)} = 1$);
- simulated data with large multiplicative batch effects ($\delta_2 = 2$) and small multiplicative batch effects ($\delta_2 = 0.8$).

Type I error was calculated as the proportion of genes/features (simulated to have no time effect) that had a Benjamini-Hochberg corrected FDR < 0.05 . Statistical power was estimated by computing the proportion of genes/features (simulated to exhibit a time effect) that had a Benjamini-Hochberg corrected FDR < 0.05 .

2.6 Application dataset description

We used a previously published whole blood gene expression data set of rheumatoid arthritis (RA), which aimed to longitudinally monitor sustained clinical remission (CR), with or without drug treatment, in the peripheral blood of patients with RA and healthy

controls (Tasaki et al. 2018). A total of 122 observations were collected from 90 participants at two time points, which were profiled in two batches. Among them, 60 observations were obtained from samples collected at time point 1 and profiled in batch 1; 22 observations were from replicates samples profiled in batch 2; and 40 observations were obtained from samples collected at time point 2 and profiled in batch 2. Among the 90 participants, 30 participants were healthy controls and 60 participants had rheumatoid arthritis. The distribution of the healthy controls and RA patients across time points and batches is shown in Additional Files, Figure S1. As only RA patients were followed up at a second time point and because the majority of RA patients (43 out of 60) were treated with one or a combination of three different disease-modifying antirheumatic drugs (DMARDs): methotrexate (MTX), infliximab (IFX), or tocilizumab (TCZ), time effects are most likely to represent the effect of treatment. The samples collected from patients with a follow-up visit at 18–28 weeks are recognized as samples collected at time point 2. If the same patient had more than one visit within 18–28 weeks, we only considered their last visit. Gene expression data collected from this study were preprocessed as described in the original report describing this study (Tasaki et al. 2018). The data are publicly available in the Gene Expression Omnibus (GEO) Data Bank under the accession code GSE93272. *BRIDGE*, *ComBat* and *longitudinal ComBat* were applied to the preprocessed data to correct/attenuate batch effects. In order to adjust for health status (Healthy Control versus RA), the L/S model described in Additional File S1 was fit, treating health status as a covariate. Parameter estimates for the data standardization step were obtained via ordinary least square (OLS) estimation. *ComBat* was fit using all 122 samples in the RA data set, including time and health status as covariates. *Longitudinal ComBat* was also fit using all 122 samples and assumed a subject-specific random intercept model formulation to account for within-subject correlation in expression measurements. PCA and PVCA were conducted to visualize batch effect removal/attenuation and to quantify the extent of technical and biological variation in the corrected and uncorrected data. For PCA, we visualized the uncorrected and corrected data by plotting PC1 and PC2, similar to Simulation I.

3 Results

3.1 Simulation studies

To gauge the operating characteristics and performance of the proposed *BRIDGE* methodologies relative to *ComBat* and *longitudinal ComBat*, we first conducted a series of simulation studies. Figure 2 contains plots of PC1 by PC2 based on principal components estimated from the uncorrected and corrected data by: *BRIDGE*-1, *BRIDGE*-2, *ComBat*, and *longitudinal ComBat*, for small multiplicative batch effects (Figure 2A, $\delta_2 = 0.5$) and large multiplicative batch effects (Figure 2B, $\delta_2 = 2$), assuming relatively weak time effects, e.g., $\Delta_0 = 0.5$ or 1.0. In the PCA plots, minimal clustering by batch but clustering by time, is indicative of a successful removal/attenuation of batch effects. When multiplicative batch effects are small (e.g., $\delta_2 = 0.5$), the *BRIDGE*-1, *ComBat* and *longitudinal ComBat* adjusted data show some overlap between samples from different time points, either due to remaining or residual batch effects or attenuation of the time effects arising as a result of batch adjustment, while the *BRIDGE*-2 adjusted data represents a better separation of samples collected from different time points (Figure 2A). In contrast, when multiplicative batch effects are large, *BRIDGE*-1 appears to outperform *BRIDGE*-2, *ComBat* and *longitudinal ComBat* in attenuating batch effects whilst preserving time-effects (Figure 2B). PCA plots based on simulated data exhibiting moderate and strong time effects (e.g., moderate $\Delta_0 = 0.6$ and 1.2; strong $\Delta_0 = 0.8$ and 1.5) are shown in Additional File Figures S2 and S3, and show, irrespective of method, improved performance of batch effects removal/attenuation as the magnitude of the time effects increase. We further note that data subjected to correction by the various batch-effect methods tend to be strongly correlated with one another (Pearson correlation >0.90 , between the corrected data between each pair of methods), regardless of the multiplicative batch effect parameter or magnitude of the time effect (Additional File, Figure S4).

In an attempt to further quantify each method's performance in removing batch effects while preserving time effects, we conducted partial redundancy analysis (pRDA), which allowed us to estimate the proportion of variation in gene expression explained by both batch and time before and after the application of a batch effect correction method. Ideally, batch effect correction methods should result in a smaller proportion of variance explained by batch and larger proportion of variance explained by time compared to the uncorrected data. Figure 3 shows that all four methods successfully decrease the proportion of variance explained by batch (Figure 3A, ~40% in the uncorrected data to 0.70–3.3% based on the corrected data across the four methods

Figure 2: PCA plot matrix to visualize batch effects removal for small multiplicative batch effects (A), $\delta_2 = 0.5$ and large multiplicative batch effects (B), $\delta_2 = 1.0$ when time effect sizes are small (A), $\Delta_0 = 0.5$; (B), $\Delta_0 = 1.0$). Columns of the PCA plot matrix (left to right) reflect: Uncorrected, BRIDGE-1, BRIDGE-2, ComBat, and longitudinal ComBat. Samples collected at time point 1 are marked as red and samples collected at time points 2, in blue. Cross symbols represent samples that were analyzed in batch 1. Samples profiled in batch 2 are marked as filled circles. Note that BRIDGE-2 imputes/predicts data for non-bridging samples which are marked as open squares. The degree of overlap between samples profiled in different batches, and the degree of separation between samples collected at different time points are indicative of the effectiveness of batch effects removal/attenuation.

Figure 3: Comparison of the proportion of variance in gene expression explained by batch and time before and after batch effect correction by *BRIDGE-1*, *BRIDGE-2*, *ComBat*, and *longitudinal ComBat* for small ((A), $\delta_2 = 0.5$) and large ((B), $\delta_2 = 2$) multiplicative batch effects. Columns represent varying assumed time-effects, ranging from: weak (left), moderate (middle), and strong (right). The partitioned variance explained by batch (Red Box) and time (Blue Box) are estimated by pRDA. Filled stars mark the method that obtains the highest proportion of variance explained by time. Note, that *BRIDGE-2* imputes/predicts data for non-bridging samples. Before pRDA was implemented on the *BRIDGE-2* corrected data, we removed all data from batch 1 and marked the predicted data as “imputed” in the batch factor. The simulation studies were repeated 20 times.

and Figure 3B ~16–19% in the uncorrected data to 0.30–6.0% based on the corrected data across the four methods). *BRIDGE-2* preserves the highest proportion of variance by time when multiplicative batch effects are small (Figure 3A) and *BRIDGE-1* preserves the highest proportions of variance by time when multiplicative batch effects are large (Figure 3B). Although *BRIDGE-1* preserves highest proportion of variance explained by time when multiplicative batch effects are large, it also has the highest proportion of variance explained by batch.

The mean and standard error of time effects estimate are labeled around the peak of each density line in Figure 4B. *BRIDGE-2* obtained the most accurate estimates in both small and large multiplicative batch effects (green dashed in Figure 4B). In terms of variance, estimates obtained from *BRIDGE-2* had the largest standard error (2.59) when multiplicative batch effects are large (Figure 4B left), but the smallest standard error (0.64) when multiplicative batch effects are small as compared to the other three methods (Figure 4B right). Root mean square error (RMSE) was calculated based on a comparison of predicted time effects and observed time effects across genes and samples. Given that all four methods (*BRIDGE-1*, *BRIDGE-2*, *ComBat* and *longitudinal ComBat*) were observed to result in unbiased time effects (Figure 4B), differences in the RMSE across methods mainly arose out of differences between in the precision of time-effect estimates associated with each method. Thus, it is not surprising to find that RMSE calculated from *BRIDGE-1* adjusted data is the lowest when multiplicative batch effects are large (Figure 4A) as *BRIDGE-1* obtains the minimum-variance estimate of time effects (1.37) in this situation. Similarly, the *BRIDGE-2* adjusted data achieves the minimum RMSE when

Figure 4: Comparison of time effect estimates of time-varying genes based on RMSE (A) and density plots (B) for large multiplicative batch effects $\delta_2 = 2$ ((B) left with effect size $\Delta_0 = 1.0$) and small multiplicative batch effects $\delta_2 = 0.5$ ((B) right with effect size $\Delta_0 = 0.5$). The time effect estimates of the G genes simulated to exhibit an effect of time were obtained after batch effects correction by *BRIDGE-1*, *BRIDGE-2*, *ComBat*, and *longitudinal ComBat*, and were compared to their true simulated values. (A) RMSE was calculated separately for each of the four methods across the G genes simulated to exhibit an effect of time and across the N samples. (B) Density plots for effect estimates of the G differential genes based on *BRIDGE-1* (red solid line), *BRIDGE-2* (green dashed), *ComBat* (blue long dashed), and *longitudinal ComBat* (purple dashed). The estimated time effects were calculated as the subject-specific difference in expression between timepoints 1 and 2 for each of the G genes that were simulated to exhibit a “true” effect of time (e.g., genes whose expression changed over time). This was done based on the corrected expression data obtained from each of the four methods listed above, and across all of the 20 simulated data sets. The mean $\pm$ se of time effects estimates are labeled around the peak of each density line.

multiplicative batch effects are small (Figure 4A) as *BRIDGE-2* obtains the minimum-variance estimate of time effects (0.64) in this situation.

Simulation II was conducted to evaluate each method's performance in relation type I error rates and statistical power in downstream statistical inference performed on the batch corrected data. As indicated in Figure 5, the general pattern of type I error rates as a function of M , number of bridging samples, is similar across the three methods. In particular, as bridging size M increases from 5 to 10, type I error rates drop quickly and stabilize thereafter. Figure 5A provides a comparison of the four methods in controlling type I error rates when multiplicative batch effects are large ($\delta_2 = 2$). In this scenario, *BRIDGE-2* demonstrates improved control of the type I error rate as compared to *BRIDGE-1*, *ComBat* and *longitudinal ComBat*. When the within-subject correlation ρ_1 increases from 0.3 to 0.9, *BRIDGE-1*, *BRIDGE-2*, and *longitudinal ComBat* result in lower type I error compared to *ComBat*, and type I errors from all methods are controlled when M increased to 40. When the multiplicative batch effects are small ($\delta_2 = 0.8$) and within-subject correlation is weak ($\rho_1 = 0.3$), *ComBat* and *longitudinal ComBat* obtain the lowest type I error rates (Figure 5B). However, as the sample dependency from batch gets stronger ($\rho_1 = 0.9$), type I error rates observed for *BRIDGE-1*, *BRIDGE-2*, and *longitudinal ComBat* decrease as a function of increasing M , whereas the decrease in type I error for increasing M is less pronounced for *ComBat*. We also investigated how the assumed variance structure affects the comparison between the *BRIDGE* methods, *ComBat*, and *longitudinal ComBat*. Figure 5 shows that as $\sigma_0^{(1)}$ increases from 0.2 to 0.8, type I error rates increase in all methods except *BRIDGE-1* (when the multiplicative batch effect parameter is large) and *BRIDGE-2* (when the multiplicative batch effect parameter is small), assuming strong sample dependency ($\rho_1 = 0.9$). We also assessed statistical power between the *BRIDGE* methods, *ComBat* and *longitudinal ComBat* (Figure 6). As bridging size M increases, power remains relatively stable, indicating somewhat counterintuitively, that larger M has a small effect on increasing statistical power for detecting genes whose expression change over time. Figure 6A shows *BRIDGE-1* is the most powerful method in detecting genes with a time-effect when multiplicative batch effects are large ($\delta_2 = 2$), followed by *ComBat*, and finally *BRIDGE-2* and *longitudinal ComBat*. In contrast to the observed relationship between within-subject correlation ρ_1 and type I error rates (Figure 5), ρ_1 does not appear to influence statistical power across the three methods (Figure 6). Further, we observed that *BRIDGE-2* has improved power over the other methods when multiplicative batch effects are small ($\delta_2 = 0.8$, Figure 6B).

3.2 Rheumatoid arthritis dataset (RA data set)

To further compare the performance of the *BRIDGE* methods, *ComBat*, and *longitudinal ComBat*, we applied each of these batch-effect correction methods to a publicly available data set consisting of longitudinal gene expression measurements. In the RA data set, the empirical Bayes estimate of the multiplicative batch effect is 0.93, indicating that data from batch 2 has better quality than batch 1, albeit only slightly. Figure 7A contains a PCA plot to visualize the batch effects removal in the RA data set, where principal components were estimated from the uncorrected data and the adjusted data across the four methods. *BRIDGE-1*, *BRIDGE-2*, *ComBat* and *longitudinal ComBat* all showed minimal clustering of samples by batch after batch-effect correction, indicating that each method was successful in attenuating batch effects. Figure 7B depicts the proportion of variation explained by health status (health control/RA), batch, and time for the uncorrected and corrected data. As observed in Figure 7B, all methods successfully decrease the proportion of variance by batch from 29.75% in uncorrected data to 0.41% using *BRIDGE-1*, 3.34% using *BRIDGE-2*, 2.38% using *ComBat*, and 1.62% using *longitudinal ComBat*. *BRIDGE-2* preserves the highest proportions of variance by time, the proportion of variance accounted for by time increases from 0.02% to 7.38%, compared to *ComBat* (0.21%), *BRIDGE-1* (0.54%), and *longitudinal ComBat* (0.216%). *BRIDGE-2* also preserves the highest proportion of variance by health status, which increased from 12.44% to 24.0%, compared to *ComBat* (17.0%), *BRIDGE-1* (17.36%), and *longitudinal ComBat* (15.56%).

Figure 5: Type I error rates across batch-effect correction methods. The effects of within-subject correlation of batch (Columns) and gene expression variance (Rows) on type I error rate as a function of M (x-axis) for large multiplicative batch effects (A), $\delta_2 = 2$ and small multiplicative batch effects (B), $\delta_2 = 0.8$. The parameters are set as $M = 5, 10, 20$, or 40 , $N = 100$, $\sigma_0^{(1)} = 0.2$ (row 1) or 0.8 (row 2), $\sigma_0^{(2)} = 1$, $G = 200$, $J = 1000$, $\rho_1 = 0.3$ (column 1) or 0.9 (column 2), $\rho_2 = 0.4$, $\varphi_2 = 5$ and $\mu_0^{(1)} = 2$. Results based on 20 simulations.

Figure 6: Statistical power across batch-effect correction methods. The effects of within-subject correlation of batch (Columns) and gene expression variance (Rows) on statistical power across different M (x-axis) for large multiplicative batch effects (A), $\delta_2 = 2$ and small multiplicative batch effects (B), $\delta_2 = 0.8$. The parameters are set as $M = 5, 10, 20$, or 40 , $N = 100$, $\sigma_0^{(1)} = 0.2$ (row 1) or 0.8 (row 2), $\sigma_0^{(2)} = 1$, $G = 200$, $J = 1000$, $\rho_1 = 0.3$ (column 1) or 0.9 (column 2), $\rho_2 = 0.4$, $\varphi_2 = 5$ and $\mu_{o_j}^{(1)} = 2$. Results based on 20 simulations.

Figure 7: Visualization of batch effects removal in the rheumatoid arthritis (RA) Dataset. (A) Plots of PC1 versus PC2 based on the uncorrected data and data resulting from batch effect correction via the four methods. (B) PVCA based estimates of the proportion of variance attributable to batch, time, and health status, before and after batch effect correction for uncorrected, *BRIDGE-1*, *BRIDGE-2*, *ComBat*, and *longitudinal ComBat*. Note that before implementation of PCA and PVCA, we removed samples collected at time point 1 but run at batch 2 for the *ComBat*, *BRIDGE-1*, and *longitudinal ComBat* adjusted data.

4 Discussion

As there is an increasing interest on how molecular features change over time, we expect longitudinal genomic studies become more commonplace over the coming years. The use of randomization for assigning samples collected different time points to specific batches mitigates confounding between time and batch, however complete confounding between batch and the outcome/phenotype of interest does happen in practice. Here, we considered the situation where time is completely confounded by batch, such as what might occur in a longitudinal study of gene expression, however complete confounding can also occur in non-longitudinal studies when the variable of interest (e.g., disease status, exposure status, etc.) is completely confounded with batch. In the completely confounded situation, our results suggest that even a small number of samples run as technical replicates across batches (e.g., bridging samples), may help in attenuating technical variability due to batch, whilst retaining the biological signal of interest. While our work is not the first to examine batch effect correction in study settings with bridging samples, where time is completely confounded with batch, to our knowledge, the *BRIDGE* methods proposed here are the first methods that make explicit use the bridging samples as a means toward estimating and ultimately attenuating technical variability due to batch effects.

Here we presented *BRIDGE*, a method specifically designed to remove batch effects using technical replicates (bridging samples) in longitudinal study involving the collection of high-throughput molecular data, when within-subject covariance is complex. *BRIDGE* separates within-subject covariation between batches by fitting a Bayesian linear regression using the standardized bridging data, obtaining empirical Bayes based estimates of the linear coefficients, and then explicitly correcting for batch. Therefore, the data adjusted by *BRIDGE* methods are ready to be used for downstream analysis as if they were profiled in a single batch. *BRIDGE* has the following three assumptions. First, bridging data are assumed to be bivariate normally distributed. Second, there is no interaction between batch and time, that is, within-subject correlation of batch ρ_1 is invariant to time, and within-subject correlation of time ρ_2 is also invariant to batch. Third, additive batch effects and multiplicative batch effects share a similar pattern across genes, leading to a similar ρ_1 and ρ_2 across genes.

Numerous studies have shown that *ComBat* demonstrates consistently good performance in terms of correcting batch effects in the context of cross-sectional studies (Chen et al. 2011; Zhou et al. 2019) and even longitudinal studies (Müller et al. 2016). Some studies have even applied *ComBat* to bridging samples to correct batch effects, ignoring the within-subject covariance structure (Banchereau et al. 2016; Tasaki et al. 2018; Wong et al. 2020). However, two potential problems exist when *ComBat* is used in batch-confounded longitudinal studies. First, it is problematic to use *ComBat* to correct for batch effects when samples are not independent, as *ComBat* does not account for within-subject variation, and as a result fails to completely remove additive and multiplicative batch effects (Beer et al. 2020). Furthermore, *ComBat* may not be able to remove batch effects when batches are confounded with variables of interest and could even introduce false signals (Akey et al. 2007; Nygaard et al. 2016; Price and Robinson 2018; Zindler et al. 2020). To solve these problems, *ComBat* was extended to *longitudinal ComBat* by Beer et al. (2020) to account for within-subject dependency by introducing a subject-specific random intercept in a linear mixed model to estimate batch effects (Beer et al. 2020). Although *longitudinal ComBat* was primarily validated using a subject-specific random intercept model formulation, in theory, it can be applied to any type of random effects structure. As a result, *longitudinal ComBat* may be particularly well suited to unbalanced designs, however more work is needed to assess its performance in study settings with bridging samples, such as the types of study designs considered here.

After fitting a Bayesian linear regression model on bridging data, we proposed two slightly different methods – *BRIDGE-1* and *BRIDGE-2*. *BRIDGE-1* uses the estimated coefficients from the Bayesian linear regression models to arrive at estimates of the additive and multiplicative batch effects, and then models out the batch effects from raw data. On the other hand, *BRIDGE-2* imputes/predicts data of the non-bridging samples based on the parameter estimates from the Bayesian linear regression model fits. The performance of the two *BRIDGE* methods depends heavily on which batch is with superior quality, which is reflected by the multiplicative batch effect parameter for batch 2 ($\hat{\delta}_2$). When we assume that batch 1 is absent of batch effects, a large multiplicative batch effects of batch 2 ($\delta_2 > 1$) indicates that data obtained by batch 2 is less stable and of lower quality. Results from PCA and pRDA show that *BRIDGE-1* is the preferred approach to correct batch effects in such situations. After batch effect correction, the standard error of time effect estimates from *BRIDGE-1* was the lowest across the methods considered here. However, when multiplicative batch effects of batch 2 are small ($\delta_2 < 1$), data obtained from batch 2 is more stable and of better quality, and in this situation, *BRIDGE-2* performs the best in batch effect removal and time effects estimation. In terms of the performance of *BRIDGE* in relation to statistical inference on time effects, as bridging sample size M increases, power was observed to increase, but only slightly. Despite that somewhat counterintuitive finding, larger M does appear to associate with more favorable type I error rates. We also find that *BRIDGE* and *longitudinal ComBat* perform better in controlling type I error rate compared to *ComBat* when the data exhibits a strong within-subject correlation ($\rho_1 = 0.9$). Based on the results from Simulation II, we found that a bridging sample size of as little as $M = 5$, resulted in adequate statistical power (>0.8) when the multiplicative batch effect parameter of batch 2 is small ($\delta_2 < 1$) and *BRIDGE-2* is used for batch effect correction/attenuation. Similarly, when the multiplicative batch effect parameter of batch 2 is large ($\delta_2 > 1$) and *BRIDGE-1* is used for batch-effect

correction/attenuation, a bridging sample size of as little as $M = 5$, resulted in adequate statistical power (>0.8). In both situations however, the type I error rates are unacceptably large when the bridging sample size is small and additional samples ($M = 20\text{--}40$) may be needed in order to control type I error <0.05 . Application of the three methods to the RA data set showed that *BRIDGE-2* performed the best in terms of removing/attenuating batch effects whilst preserving biological effects. This is perhaps not surprising given the results of our simulation studies and the fact that the empirical Bayes estimate of the multiplicative batch effect of batch 2 is 0.93, indicating data from batch 2 has better quality than batch 1.

While the proposed *BRIDGE* performed favorably as compared to *ComBat* and *longitudinal ComBat* in both our simulation study and real data analysis, it is not free of limitations. First, while we developed a model that allows one to correct batch effects when covariates need to be considered (Additional File S1), this model was not used when benchmarking *BRIDGE* to *ComBat* and *longitudinal ComBat* in our simulation studies and was only applied here in the real dataset application where health status was controlled for as a covariate. A second limitation is that the *BRIDGE* methods were developed for very specific study designs in which there are two batches that are completely confounded with time, with bridging samples run as technical replicates between batches. However, given the declining cost of high-throughput technologies for interrogating the genome, epigenome, transcriptome, etc., coupled with an increasing interest in the dynamic/temporal behavior of such mechanisms, we expect longitudinal studies of such molecular data types to become more commonplace over the coming years. For logistical and/or practical reasons, such studies may need to be designed and implemented such that samples collected at certain timepoints are arrayed together (e.g., storage constraints, sample degradation, etc.), resulting in the so-called “completely confounded by time” scenario our *BRIDGE* methodologies were developed to address. Finally, given the complex nature of the assumed study design, the expected proportion of variance attributable to time in our simulation study was not obtained, thus precluding a direct comparison of the expected and observed proportions of variance attributable to time based on the data corrected by each of the considered methods. However, we feel that the comparisons presented here provide nevertheless provide valuable insight since our simulation studies and real data analysis benchmark *BRIDGE* against the other two major competing methods that have been suggested and used for the types of designs we consider, *ComBat*, and *Longitudinal ComBat*.

Opportunities for future work include further exploration how bridging sample size, M , influences the performance of *BRIDGE* in terms of controlling type I error and achieving a target statistical power, based user supplied parameters. Along the same lines, it would be valuable to develop an open-source application to guide researchers in determining an appropriate bridging sample size before they conduct their study. Furthermore, more research into methods for estimating parameters used in the standardization step, including the potential use of linear mixed effects models, has potential to mitigate the inflated type I error and increase power. Finally, it would be worthwhile to explore how to scale up the approach described here for longitudinal studies with more than two time-points and two batches.

5 Conclusions

In conclusion, the results of our study suggest that *BRIDGE* may serve as a useful preprocessing method for researchers conducting longitudinal microarray studies where batch is completely confounded with time. Compared to a widely used batch-effect correction method, *BRIDGE* achieved improved performance in the removal of batch-effects, while preserving a greater proportion of variance by time. As a result, *BRIDGE* was observed to have improved statistical power in downstream statistical inference at an acceptable type I error rates.

Acknowledgments: We would like to extend our gratitude to: Lisa Neums, Richard Meier, Shachi Patel, Duncan Rotich, Shelby Bell-Glenn, Whitney Shae, Bo Zhang, Samuel Boyd, Emily Nissen, Jonah Amponsah,

Dr. Prabhakar Chalise, Dr. Jinxiang Hu, Dr. Lynn Chollet-Hinton, Dr. Nanda Yellapu, and Dr. Dong Pei, for their constructive feedback on this manuscript.

Author contribution: Q.X. conceived the statistical method, performed the statistical analyses to evaluate and assess the proposed method, developed the R package for the *BRIDGE* method, and participated in the drafting and review of the manuscript. J.A.T. helped direct the statistical analyses, contributed to the interpretation of study findings, and assisted in manuscript writing and development. D.C.K helped to conceive the statistical method, provided guidance and direction in the evaluation and assessment of the method, and assisted in manuscript writing and development.

Research funding: Research reported was supported by: the National Cancer Institute (NCI) Cancer Center Support Grant P30 CA168524; the Kansas IDEa Network of Biomedical Research Excellence Bioinformatics Core, supported by the National Institute of General Medical Science award P20 GM103418; the Kansas Institute for Precision Medicine COBRE, supported by the National Institute of General Medical Science award P20 GM130423.

Conflict of interest statement: The authors declare no conflicts of interest regarding this article.

Application availability: The *BRIDGE* R Package can be found and installed at GitHub: <https://github.com/qingxiaa/brg>.

References

- Akey, J.M., Biswas, S., Leek, J.T., and Storey, J.D. (2007). On the design and analysis of gene expression studies in human populations. *Nat. Genet.* 39: 807–808.
- Baggerly, K.A., Edmonson, S.R., Morris, J.S., and Coombes, K.R. (2004). High-resolution serum proteomic patterns for ovarian cancer detection. *Endocr. Relat. Cancer* 11: 583–585.
- Banchereau, R., Hong, S., Cantarel, B., Baldwin, N., Baisch, J., Edens, M., Cepika, A.-M., Acs, P., Turner, J., and Anguiano, E. (2016). Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 165: 551–565.
- Beer, J.C., Tustison, N.J., Cook, P.A., Davatzikos, C., Sheline, Y.I., Shinohara, R.T., and Linn, K.A. (2020). Longitudinal ComBat: a method for harmonizing longitudinal multi-scanner imaging data. *Neuroimage* 220: 117129.
- Bjornsson, H.T., Sigurdsson, M.I., Fallin, M.D., Irizarry, R.A., Aspelund, T., Cui, H., Yu, W., Rongione, M.A., Ekstrom, T.J., Harris, T.B., et al. (2008). Intra-individual change over time in DNA methylation with familial clustering. *J. Am. Med. Assoc.* 299: 2877–2883.
- Chen, C., Grennan, K., Badner, J., Zhang, D., Gershon, E., Jin, L., and Liu, C. (2011). Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* 6: e17238.
- Fei, T. and Yu, T. (2020). scBatch: batch-effect correction of RNA-seq data through sample distance matrix adjustment. *Bioinformatics* 36: 3115–3123.
- Goh, W.W.B., Wang, W., and Wong, L. (2017). Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol.* 35: 498–507.
- Hejblum, B.P., Skinner, J., and Thiebaut, R. (2015). Time-course gene set analysis for longitudinal gene expression data. *PLoS Comput. Biol.* 11: e1004310.
- Hicks, S.C., Townes, F.W., Teng, M., and Irizarry, R.A. (2018). Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics* 19: 562–578.
- Irizarry, R.A., Warren, D., Spencer, F., Kim, I.F., Biswal, S., Frank, B.C., Gabrielson, E., Garcia, J.G., Geoghegan, J., Germino, G., et al. (2005). Multiple-laboratory comparison of microarray platforms. *Nat. Methods* 2: 345–350.
- Johnson, W.E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 8: 118–127.
- Karlovich, C., Duchateau-Nguyen, G., Johnson, A., McLoughlin, P., Navarro, M., Fleurbaey, C., Steiner, L., Tessier, M., Nguyen, T., Wilhelm-Seiler, M., et al. (2009). A longitudinal study of gene expression in healthy individuals. *BMC Med. Genom.* 2: 33.
- Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., and Irizarry, R.A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* 11: 733–739.
- Li, J., Bushel, P.R., Chu, T.M., and Wolfinger, R.D. (2009). Batch Effects Noise in Microarray Experiments: sources Solutions: principal variance components analysis: estimating batch effects in microarray gene expression data. In: *Batch effects noise in microarray experiments: sources solutions*. John Wiley & Sons, Ltd, West Sussex, United Kingdom, pp. 141–154.
- Müller, C., Schillert, A., Röthmeier, C., Trégouët, D.-A., Proust, C., Binder, H., Pfeiffer, N., Beutel, M., Lackner, K.J., and Schnabel, R.B. (2016). Removing batch effects from longitudinal gene expression-quantile normalization plus combat as best approach for microarray transcriptome data. *PLoS One* 11: e0156594.

- Nygaard, V., Rødland, E.A., and Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics* 17: 29–39.
- Price, E.M. and Robinson, W.P. (2018). Adjusting for batch effects in DNA methylation microarray data, a Lesson Learned. *Front. Genet.* 9: 83.
- Scherer, A. (2009). *Batch effects and noise in microarray experiments: sources and solutions*. John Wiley & Sons, West Sussex, United Kingdom.
- Sun, Z., Chai, H.S., Wu, Y., White, W.M., Donkena, K.V., Klein, C.J., Garovic, V.D., Therneau, T.M., and Kocher, J.P. (2011). Batch effect correction for genome-wide methylation data with Illumina Infinium platform. *BMC Med. Genom.* 4: 84.
- Tasaki, S., Suzuki, K., Kassai, Y., Takeshita, M., Murota, A., Kondo, Y., Ando, T., Nakayama, Y., Okuzono, Y., Takiguchi, M., et al. (2018). Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat. Commun.* 9: 2755.
- Taub, M.A., Bravo, H. C., and Irizarry, R.A. (2010). Overcoming bias and systematic errors in next generation sequencing data. *Genome Med.* 2: 87.
- Venkatraman, V.K., Gonzalez, C.E., Landman, B., Goh, J., Reiter, D.A., An, Y., and Resnick, S.M. (2015). Region of interest correction factors improve reliability of diffusion imaging measures within and across scanners and field strengths. *Neuroimage* 119: 406–416.
- Wang, Y. and Lê Cao, K.-A. (2020). A multivariate method to correct for batch effects in microbiome data. *bioRxiv*.
- Wong, C.J., Wang, L.H., Friedman, S.D., Shaw, D., Campbell, A.E., Budech, C.B., Lewis, L.M., Lemmers, R., Statland, J.M., van der Maarel, S.M., et al. (2020). Longitudinal measures of RNA expression and disease activity in FSHD muscle biopsies. *Hum. Mol. Genet.* 29: 1030–1043.
- Zhang, Y., Jenkins, D.F., Manimaran, S., and Johnson, W.E. (2018). Alternative empirical Bayes models for adjusting for batch effects in genomic studies. *BMC Bioinf.* 19: 262.
- Zhou, L., Sue, A.C.-H., and Goh, W.W.B. (2019). Examining the practical limits of batch effect-correction algorithms: when should you care about batch effects? *J. Genet. Genom.* 46: 433–443.
- Zindler, T., Frieling, H., Neyazi, A., Bleich, S., and Friedel, E. (2020). Simulating ComBat: how batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. *BMC Bioinf.* 21: 1–15.

Supplementary Material: The online version of this article offers supplementary material (<https://doi.org/10.1515/sagmb-2021-0020>).