

RESEARCH

Open Access



# Propensity scores as a novel method to guide sample allocation and minimize batch effects during the design of high throughput experiments

Patrick M. Carry<sup>1,2\*</sup>, Tim Vigers<sup>3,4</sup>, Lauren A. Vanderlinden<sup>2,3</sup>, Carson Keeter<sup>1</sup>, Fran Dong<sup>4</sup>, Teresa Buckner<sup>2</sup>, Elizabeth Litkowski<sup>2</sup>, Ivana Yang<sup>5</sup>, Jill M. Norris<sup>2†</sup> and Katerina Kechris<sup>3†</sup>

<sup>†</sup>Jill M. Norris and Katerina Kechris: Co-senior authors

\*Correspondence: patrick.carry@cuanschutz.edu

<sup>1</sup> Colorado Program for Musculoskeletal Research, Department of Orthopedics, University of Colorado Anschutz Medical Campus, 12631 E. 17Th Ave, Room 4602, Mail Stop B202, Aurora, CO 80045, USA

<sup>2</sup> Department of Epidemiology, Colorado School of Public Health, Aurora, CO, USA

<sup>3</sup> Department of Biostatistics and Informatics, Colorado School of Public Health, Aurora, CO, USA

<sup>4</sup> Barbara Davis Center for Diabetes, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

<sup>5</sup> Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA

## Abstract

**Background:** We developed a novel approach to minimize batch effects when assigning samples to batches. Our algorithm selects a batch allocation, among all possible ways of assigning samples to batches, that minimizes differences in average propensity score between batches. This strategy was compared to randomization and stratified randomization in a case–control study (30 per group) with a covariate (case vs control, represented as  $\beta_1$ , set to be null) and two biologically relevant confounding variables (age, represented as  $\beta_2$ , and hemoglobin A1c (HbA1c), represented as  $\beta_3$ ). Gene expression values were obtained from a publicly available dataset of expression data obtained from pancreas islet cells. Batch effects were simulated as twice the median biological variation across the gene expression dataset and were added to the publicly available dataset to simulate a batch effect condition. Bias was calculated as the absolute difference between observed betas under the batch allocation strategies and the true beta (no batch effects). Bias was also evaluated after adjustment for batch effects using ComBat as well as a linear regression model. In order to understand performance of our optimal allocation strategy under the alternative hypothesis, we also evaluated bias at a single gene associated with both age and HbA1c levels in the ‘true’ dataset (CAPN13 gene).

**Results:** Pre-batch correction, under the null hypothesis ( $\beta_1$ ), maximum absolute bias and root mean square (RMS) of maximum absolute bias, were minimized using the optimal allocation strategy. Under the alternative hypothesis ( $\beta_2$  and  $\beta_3$  for the CAPN13 gene), maximum absolute bias and RMS of maximum absolute bias were also consistently lower using the optimal allocation strategy. ComBat and the regression batch adjustment methods performed well as the bias estimates moved towards the true values in all conditions under both the null and alternative hypotheses. Although the differences between methods were less pronounced following batch correction, estimates of bias (average and RMS) were consistently lower using the optimal allocation strategy under both the null and alternative hypotheses.



**Conclusions:** Our algorithm provides an extremely flexible and effective method for assigning samples to batches by exploiting knowledge of covariates prior to sample allocation.

**Keywords:** Batch effects, Propensity scores, Batch effect adjustment, ComBat

## Background

Batch effects represent non-biological sources of variation in high throughput experiments [1–3]. Batch effects are often associated with a greater magnitude of differential expression than the primary biological covariate(s) of interest [2]. Correlation between these technical factors and biological effects has the potential to mask true associations or lead to spurious findings [2, 4–6]. Leek et al. [2] identified batch effects in nine large, publicly available datasets. Moreover, up to 99.5% of the features in the datasets were significantly correlated with one or more of the technical variables within the individual datasets [2]. Batch effect concerns extend beyond the individual studies. Failure to account for batch effects can confound meta-analyses and decrease reproducibility of study results leading to an inefficient use of resources [2, 4].

Following data acquisition, numerous batch correction methods exist, ranging from simple options including explicitly modeling known technical factors in a multivariable analytic framework to more complex algorithms [1, 3, 5]. Surrogate variable analysis (sva) is a popular method in which, consistent with other approaches, batch effects are estimated from the data [7], capturing both known and unknown technical factors. The factors estimated using this data-driven approach are then corrected for in subsequent analyses [7]. Removing unwanted sources of variation is another method that is conceptually similar to the sva approach except that it exploits biological features assumed to be independent of the covariate of interest, colloquially termed ‘housekeeping genes’, to remove batch effects [8]. Known batch effects can be specified and removed using a parametric or a non-parametric empirical Bayes approach [9]. This popular method is often implemented using the ComBat function within the sva R package [7]. Less commonly used batch correction methods include the machine learning based Distance-weighted discrimination method [10], mean centering [11], geometric and arithmetic ratio methods [12], and location/scale adjustment models [13]. However, even when the experimental outcome (i.e. case in a case–control study) is equally distributed across batches, imbalance in the distribution of biologically relevant covariates across batches has the potential to complicate batch correction methods [14] as it is challenging to differentiate between confounding effects and true biological effects. Therefore, sound experimental designs using methods that minimize imbalance in covariates across batches should be prioritized.

Propensity scores are used in observational studies to address imbalance in important covariates across levels of the exposure/treatment variable. In the context of a treatment study (active treatment vs control), propensity scores represent the probability of receiving active treatment conditional upon a set of covariates [15]. Within the causal inference framework, propensity scores are used to address the ‘exchangeability assumption’ or the assumption that the potential outcomes are independent of the treatment assignment and thus, the treated group would have developed the same average outcome as the untreated group had they not been assigned to the treatment group (we

refer the reader to Hernan and Robins [16] for more additional information about the causal inference framework). In this way, propensity score matching, stratification and/or weighting are used to ensure the probability of being assigned to the treatment group given a complete set of clinically relevant covariates is balanced between groups [16–20], potentially satisfying one of many assumptions required for estimation of causal treatment effects. Propensity scores are also useful as a dimension reduction technique, providing a single value, ranging between 0 and 1, that represents the overall balance in the distribution of a set of covariates between groups. In the context of batch assignments, differences in propensity scores between batches can be used to screen for imbalance in the overall distribution of a set of covariates across two or more batches. Therefore, we propose using propensity scores to guide allocation of samples to batches in order to identify the optimal allocation iteration that minimizes the differences in propensity scores between the batches. In this way, we build on the methods for implementing covariate constrained randomization in cluster randomized controlled trials [21] to provide a more optimal methodology for allocating samples to batches during the design of high throughput experiments.

**Results**

**Study population**

In order to create a biologically plausible scenario, we downloaded a publicly available microarray gene expression dataset (GSE50397), that includes gene expression data from human pancreas islet cells. Biologically relevant variables in the phenotype file (age and HbA1c) were used to represent potential confounding variables that differ between cases and controls in a typical case–control design. Consistent with the experimental design, there was a significant difference in both age (49.3 yrs. ± 10.9 vs. 62.9 yrs. ± 6.2,  $p < 0.0001$ ) and HbA1c (5.2 ± 0.4 vs. 5.8 ± 0.3,  $p < 0.0001$ ) levels in cases ( $n = 30$ ) versus controls ( $n = 30$ ).

**Bias prior to batch correction**

Under each experimental scenario (Table 1), we estimated bias as the difference between each of the beta values ( $\beta_1$ , cases and controls,  $\beta_2$ , increase in expression for every 1 unit increase in age, and  $\beta_3$ , increase in expression for every 1 unit increase in HbA1c levels) and the ‘true’ beta values. The ‘true’ beta values were calculated using gene expression values from GSE50397 before the simulated batch effects were added to the expression matrix. The beta value for the case variable,  $\beta_1$ , was set at zero and thus represents the effectiveness of the batch allocation strategies under the null hypothesis, no difference between cases and controls. Prior to batch correction, (no adjustment), average maximum absolute bias was lowest, closest to the true  $\beta_1$  values, in the optimal condition

**Table 1** Summary of experimental scenarios

Batch allocation strategy	Adjustment method		
Optimal	No adjustment	ComBat adjustment	Regression adjustment
Randomization	No adjustment	ComBat adjustment	Regression adjustment
Stratified randomization	No adjustment	ComBat adjustment	Regression adjustment

**Table 2** Average maximum absolute bias or difference between true\* and batch effect conditions prior to batch adjustment

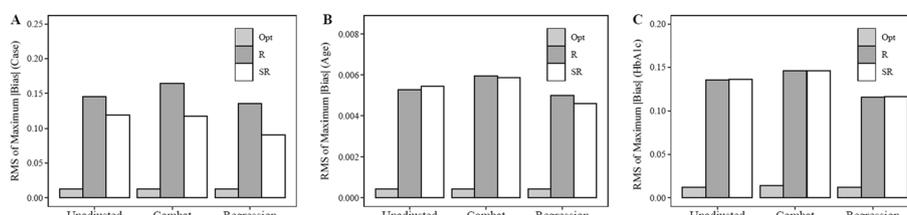
		Average max	RMS
Case ( $\beta_1$ )	Randomization versus True	3.06E-01	1.45E-01
	Stratified Randomization versus True	2.81E-01	1.18E-01
	Optimal versus True	<b>1.72E-01</b>	<b>1.25E-02</b>
Age ( $\beta_2$ )	Randomization versus True	1.12E-02	5.29E-03
	Stratified Randomization versus True	1.12E-02	5.45E-03
	Optimal versus True	<b>6.18E-03</b>	<b>4.28E-04</b>
HbA1c ( $\beta_3$ )	Randomization versus True	2.96E-01	1.35E-01
	Stratified Randomization versus True	2.95E-01	1.36E-01
	Optimal versus True	<b>1.65E-01</b>	<b>1.21E-02</b>

True\* condition represents gene expression values from GSE50397 before batch effects were added to the expression sets in each of the 1000 simulation iterations

Within each experimental iteration, absolute bias was calculated as the absolute value of the difference between observed beta coefficient and the ‘true’ coefficient across the 10,000 most variable genes in the expression dataset, maximum bias represents maximum absolute bias across these genes

Average Max mean of maximum absolute bias across all simulation iterations; RMS root mean square of maximum absolute bias value across all simulation iterations

The emboldened values represent the smallest, lowest bias, values under each experimental condition



**Fig. 1** Root mean square (RMS) of maximum absolute bias across the experimental scenarios. Within each experimental iteration, absolute bias was calculated as the absolute value of the difference between observed beta coefficient and the ‘true’ coefficient across the 10,000 most variable genes in the expression dataset. Maximum absolute bias represents the maximum absolute difference between the true beta (prior to adding batch effects) and the beta estimates from each of the experimental scenarios (Table 1). The root mean square or standard deviation of the maximum absolute bias estimates across the experimental simulations was estimated for  $\beta_1$  (panel A, case vs control parameter that was set at 0 to represent the null hypothesis),  $\beta_2$  (panel B, parameter estimate representing the association between age and gene expression), and  $\beta_3$  (panel C, parameter estimate representing the association between HbA1c and gene expression). Opt = optimal allocation strategy, R = simple randomization strategy, SR = stratified randomization strategy

(Table 2). In many cases, bias values achieved using the optimal allocation strategy were closest to the ‘true’ values by an order of magnitude relative to the other approaches (Table 2). Variability in maximum absolute bias as represented by RMS of maximum absolute bias for  $\beta_1$  was also lowest using the optimal condition Table 2 and Fig. 1). This finding was also consistent across  $\beta_2$  (age) and  $\beta_3$  (HbA1c).

**Bias after batch correction**

Most experimental pipelines recommend batch effect adjustment prior to final analysis. We evaluated two batch adjustment methods, ComBat and a regression adjustment technique. Following batch correction, average maximum absolute bias for  $\beta_1$  was similar across all methods but lowest in the optimal condition following regression adjustment

for all beta parameters (Table 3). RMS of maximum absolute bias was also lower for  $\beta_1$  under the optimal allocation scenario following regression adjustment compared to the other randomization methods and batch correction methods (Table 3 and Fig. 1). This finding was also consistent across  $\beta_2$  (age) and  $\beta_3$  (HbA1c).

**Bias at CAPN13: alternative hypothesis condition ( $\beta_2 \neq 0$  and  $\beta_3 \neq 0$ )**

In the results section above,  $\beta_2$  and  $\beta_3$  are expected to be associated with some but not all genes in the dataset. This is analogous to inclusion of adjustment variables in the statistical modelling framework that are expected to reduce bias in the phenotype gene expression associations, but the associations between these variables and gene expression may not be of interest to the investigator. Therefore, we also evaluated bias at a single gene, CAPN13, that was associated with age and HbA1c and thus, is more likely to be of interest to the investigator. Expression of this gene was used to understand the effectiveness of the optimal allocation method under the alternative hypothesis. Consistent with the primary analysis, mean absolute bias, maximum absolute bias, and root mean square (RMS) of absolute bias were lowest in the optimal allocation group prior to batch correction (Table 4 and Fig. 2). Following batch correction, mean absolute bias, maximum absolute bias, and RMS of absolute bias were consistently lower in the optimal condition relative to the randomization conditions (Table 4 and Fig. 2).

**Minimum standard error**

In addition to minimizing bias, it is desirable to estimate standard errors for the beta parameters that are close to the true error. Overestimation of error increases likelihood of a type II error while underestimation increases the likelihood of a type I error. We evaluated the average of the mean standard error as well as the minimum of the mean standard error across all experimental iterations. Prior to batch

**Table 3** Maximum absolute bias or difference between true\* and batch effect conditions following batch adjustment

		Combat adjustment		Regression adjustment	
		Average max	RMS	Average max	RMS
Case ( $\beta_1$ )	Randomization versus True	3.16E-01	1.64E-01	2.89E-01	1.35E-01
	Stratified Randomization versus True	2.74E-01	1.18E-01	2.46E-01	9.05E-02
	Optimal versus True	<b>1.76E-01</b>	<b>1.31E-02</b>	<b>1.71E-01</b>	<b>1.25E-02</b>
Age ( $\beta_2$ )	Randomization versus True	1.16E-02	5.95E-03	1.05E-02	4.99E-03
	Stratified Randomization versus True	1.14E-02	5.86E-03	1.03E-02	4.62E-03
	Optimal versus True	<b>6.34E-03</b>	<b>4.47E-04</b>	<b>6.17E-03</b>	<b>4.28E-04</b>
HbA1c ( $\beta_3$ )	Randomization versus True	3.00E-01	1.46E-01	2.68E-01	1.15E-01
	Stratified Randomization versus True	3.01E-01	1.46E-01	2.72E-01	1.16E-01
	Optimal versus True	<b>1.72E-01</b>	<b>1.44E-02</b>	<b>1.64E-01</b>	<b>1.20E-02</b>

True\* condition represents gene expression values from GSE50397 before batch effects were added to the expression sets in each of the 1000 simulation iterations

Within each experimental iteration, absolute bias was calculated as the absolute value of the difference between observed beta coefficient and the 'true' coefficient across the 10,000 most variable genes in the expression dataset, maximum bias represents maximum absolute bias across these genes

Average Max mean of maximum absolute bias values across all simulation iterations. RMS root mean square of maximum absolute bias value across all simulation iterations

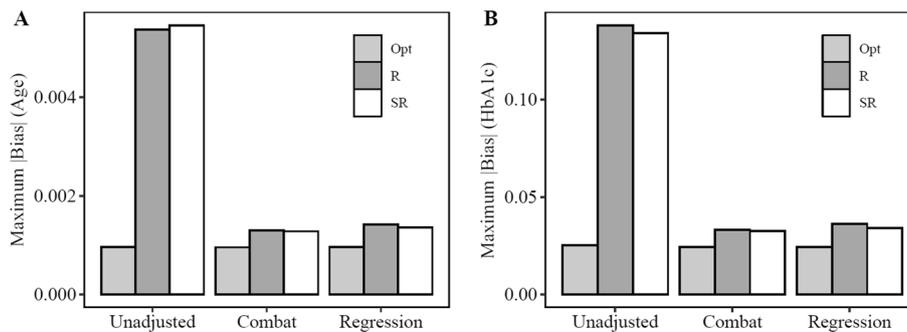
The emboldened values represent the smallest, lowest bias, values under each experimental condition

**Table 4** Absolute bias under alternative hypothesis\* in the ‘true’ expression dataset

	No adjustment			ComBat adjustment			Regression adjustment		
	Mean	Max	RMS	Mean	Max	RMS	Mean	Max	RMS
<i>Age</i> ( $\beta_2$ )									
Randomization versus True	5.52E-03	5.38E-03	4.05E-02	1.66E-03	1.30E-03	8.59E-03	1.72E-03	1.42E-03	8.54E-03
Stratified Randomization versus True	5.54E-03	5.46E-03	4.07E-02	1.65E-03	1.28E-03	6.26E-03	1.70E-03	1.36E-03	7.73E-03
Optimal versus True	<b>1.25E-03</b>	<b>9.59E-04</b>	<b>5.37E-03</b>	<b>1.26E-03</b>	<b>9.54E-04</b>	<b>5.18E-03</b>	<b>1.25E-03</b>	<b>9.57E-04</b>	<b>5.24E-03</b>
<i>HbA1c</i> ( $\beta_3$ )									
Randomization versus True	1.45E-01	1.38E-01	1.04E+00	4.25E-02	3.30E-02	2.06E-01	4.48E-02	3.61E-02	2.29E-01
Stratified Randomization versus True	1.43E-01	1.34E-01	1.49E+00	4.25E-02	3.25E-02	1.93E-01	4.26E-02	3.40E-02	1.89E-01
Optimal versus True	<b>3.27E-02</b>	<b>2.51E-02</b>	<b>1.28E-01</b>	<b>3.21E-02</b>	<b>2.42E-02</b>	<b>1.30E-01</b>	<b>3.20E-02</b>	<b>2.43E-02</b>	<b>1.31E-01</b>

\* Alternative hypothesis based on genes associated with both age and HbA1c in the true expression dataset

Mean mean of absolute bias values across all simulation iterations; RMS root mean square of absolute bias across all simulation iterations; Max max absolute bias value across all simulation iterations; the emboldened values represent the smallest, lowest bias, values under each experimental condition.



**Fig. 2** Maximum absolute bias under the alternative hypothesis. In order to understand performance of the optimal allocation algorithm under the alternative hypothesis, bias was estimated at a gene (CAPN13) that was associated with both age and HbA1c in the true dataset. Maximum absolute bias represents the maximum absolute difference between the true betas (prior to adding the batch effects) and the beta estimates under each experimental scenario (Table 1) across all of the 1000 simulation iterations. Maximum absolute bias was estimated for  $\beta_2$  (panel A, parameter estimate representing the association between age and gene expression), and  $\beta_3$  (panel B, parameter estimate representing the association between HbA1c and gene expression). Opt = optimal allocation strategy, R = simple randomization strategy, SR = stratified randomization strategy

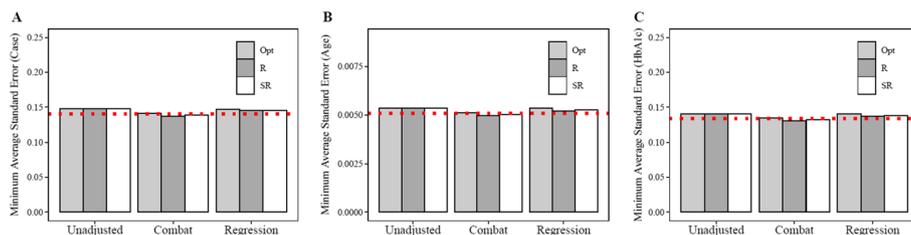
**Table 5** Standard error before and after batch adjustment

	No adjustment		ComBat adjustment		Regression adjustment	
	Mean	Min	Mean	Min	Mean	Min
Case ( $\beta_1$ )   True Value = 1.40E-01						
Randomization Condition	2.37E-01	1.48E-01	1.42E-01	<b>1.37E-01</b>	1.50E-01	1.45E-01
Stratified Randomization Condition	2.38E-01	1.47E-01	1.42E-01	<b>1.39E-01</b>	1.49E-01	1.45E-01
Optimal Condition	2.41E-01	1.47E-01	1.42E-01	1.41E-01	1.48E-01	1.47E-01
Age ( $\beta_2$ )   True Value = 5.09E-03						
Randomization Condition	8.59E-03	5.35E-03	5.13E-03	<b>4.97E-03</b>	5.44E-03	5.21E-03
Stratified Randomization Condition	8.64E-03	5.34E-03	5.13E-03	<b>5.02E-03</b>	5.44E-03	5.26E-03
Optimal Condition	8.73E-03	5.35E-03	5.14E-03	5.13E-03	5.35E-03	5.34E-03
HbA1c ( $\beta_3$ )   True Value = 1.34E-01						
Randomization Condition	2.26E-01	1.40E-01	1.35E-01	<b>1.31E-01</b>	1.43E-01	1.37E-01
Stratified Randomization Condition	2.27E-01	1.40E-01	1.35E-01	<b>1.32E-01</b>	1.43E-01	1.38E-01
Optimal Condition	2.29E-01	1.40E-01	1.35E-01	1.35E-01	1.41E-01	1.40E-01

Within each experimental iteration, standard error calculated as the average standard error across the 10,000 most variable genes in the dataset

Mean mean of average standard error across all experimental iterations; Min minimum of average standard error across all experimental iterations

Red highlighting identifies standard error values less than standard error values estimated in the 'true' gene expression dataset (before batch effects were added). The emboldened values represent standard error estimates that are less than the corresponding true value



**Fig. 3** Average minimum standard error of the case parameter estimate. The standard error of the parameter estimates was estimated under each of the experimental scenarios (Table 1) across all 10,000 gene expression features. We estimated the average minimum standard error across all of the 1000 simulation iterations. The average minimum standard error was estimated for  $\beta_1$  (panel A, case vs control parameter that was set at 0 to represent the null hypothesis),  $\beta_2$  (panel B, parameter estimate representing the association between age and gene expression), and  $\beta_3$  (panel C, parameter estimate representing the association between HbA1c and gene expression. Opt = optimal allocation strategy, R = simple randomization strategy, SR = stratified randomization strategy

correction, standard error estimates for all batch allocation strategies were similar (Table 5 and Fig. 3). After batch correction, standard error estimates decreased in all three batch allocation strategies. Following Combat adjustment, the minimum average standard error was less than the 'true' average standard error value for  $\beta_2$  (age) and  $\beta_3$  (HbA1c) under the randomization and stratified randomization conditions. Using the optimal allocation method, the minimum average standard error was greater than the 'true' average standard error value for all parameter estimates before and after batch correction (Table 5 and Fig. 3).

## Discussion

Batch effects represent an unwanted source of variation in high throughput experiments that have the potential to bias the estimation of biological effects. In the current study, propensity scores were an effective methodology for allocating samples to batches such that all relevant covariates were equally distributed across the batches. The average of maximum absolute bias, as well as the variability of maximum absolute bias across simulations were consistently lower in the optimal, propensity guided allocation strategy compared to alternative randomization strategies.

We compared the effectiveness of our novel batch allocation algorithm relative to simple randomization and stratified randomization batch allocation strategies. In each experimental scenario, prior to batch correction, maximum absolute bias was lower in the optimal allocation conditions compared to the randomization conditions (Table 2). This finding was consistent for all beta estimates representing the null hypothesis (case beta value,  $\beta_1=0$ ). Variability in maximum absolute bias, as represented by the RMS of the maximum bias metrics revealed the largest separation between conditions both pre- and post-batch correction among all probes (See Tables 2, 3 and Fig. 1). This indicates bias is lower and more constrained using the optimal sample allocation strategy. We also evaluated the performance of the algorithm under the alternative hypothesis ( $\beta_2$ , association with age) and ( $\beta_3$ , association with HbA1c). Not all genes are expected to be associated with both age and HbA1c. Therefore, we evaluated the bias in expression of the CAPN13 gene, a gene that was significantly associated with both age and HbA1c levels in the 'true' gene expression dataset. Consistent with the null hypothesis results, maximum absolute bias under the alternative hypothesis was consistently lowest in the optimal allocation condition before batch correction (Table 4 and Fig. 2).

The stratified randomization method was the next most effective allocation strategy, outperforming the simple randomization design, especially under the null hypothesis (Tables 3–4 and Fig. 1). In randomized control trials, stratified randomization designs are used to prevent imbalance in the distribution of key confounders across the treatment groups [22]. This is accomplished by independently randomizing treatment assignments within each strata of a key confounder, often sex, age category, site, or clinician. Our study results reinforce the importance of considering stratified randomization designs in high throughput experimental designs.

In cluster randomized trials, covariate constrained randomization provides additional protections against imbalance by considering cluster level factors, such as site location, site volume, etc., during randomization design [21, 23]. However, these designs in RCTs are inherently limited by the prospective nature of enrollment. It is not possible to know all subject level covariates and/or temporally variant cluster level factors prior to treatment assignment. In contrast, in high throughput experiments, it is most cost effective to run samples at one time, after all subject level data has been collected. Therefore, measurable or observed subject level confounders are often known prior to the allocation of samples to batches. The methodology proposed in this study exploits this knowledge by selecting the sample allocation option that minimizes imbalance in confounders across batches. Our approach is conceptually very similar to the OSAT Bioconductor package [24]. However, in contrast to the OSAT package as well as cluster constrained

approaches, our approach allows for consideration of multiple continuous variables as well as more complicated interactions between all covariates.

We also evaluated bias after removal of batch effects using the empirical Bayes based ComBat function implemented using the *sva* R package as well as a linear model regression framework. After adjusting for batch effects, bias estimates from all methods were small, indicating the adjustment methods performed well. Adjustment for batch effects using ComBat had more of an impact on the randomization and stratified randomization conditions due to the higher levels of bias present prior to batch correction. However, a key finding was that the optimal allocation scenario consistently outperformed the other strategies in regard to the mean maximum absolute bias as well as the variability of maximum absolute bias. On average, the randomization methods will produce unbiased estimates. However, running multiple experiments in order to obtain average and thus, unbiased estimates is very costly. The lower variability in maximum absolute bias as well as the consistency in maximum absolute bias before and after batch effect adjustment using our novel optimal allocation method suggests the investigator is less likely to encounter more extreme bias using the optimal allocation approach. This is an important point as batch correction methods do not perform as well in situations where randomization results in an imbalance in key covariates across batches [6].

Under the alternative hypothesis (expression of the CAPN13 gene), the ComBat method slightly outperformed the regression adjustment batch correction method (Table 4) based on RMS of absolute bias. This finding is consistent with Jiao et al., in which the authors concluded that ComBat was the superior method for correcting technical effects including batch and position relative to other batch correction methods evaluated in an analysis of methylation data [25]. In microarray expression data, implementation of ComBat using parametric and non-parametric methods outperformed other commonly used batch correction methods in terms of minimizing proportion of variance attributable to batch effect and improving correlation between technical replicates [1]. ComBat methods were also superior to other techniques according to an overall performance metric, area under the curve, based on sensitivity and specificity of the batch correction methods [1].

Batch adjustment using ComBat has been criticized due to its potential to create artificial associations between high throughput features and biological exposures during the batch correction stage of the analysis [14]. One possible source of bias is due to the underestimation of standard errors. Batch effect correction is performed in two steps and thus, uncertainty in estimating batch effects is not accounted for in the downstream analysis. Therefore, we reviewed minimum average standard errors (in each iteration, standard errors were averaged across all 10,000 features, the minimum average standard error represents the minimum average standard error value across all 1000 of the simulation iterations). As expected, standard errors prior to batch correction were higher in the randomization conditions as well as the optimal randomization condition compared to true biological variation (Table 5). After removal of batch effects, the standard errors for biological exposure (case vs control) as well as the confounders were decreased relative to the unadjusted condition (Table 5 and Fig. 3). Consistent with concerns outlined in previous studies [14, 26], minimum average standard error following ComBat adjustment in the randomization and stratified randomization conditions were less than

true average standard error values for the case ( $\beta_1$ ), age ( $\beta_2$ ), and HbA1c ( $\beta_3$ ) parameters (Table 5 and Fig. 3). In contrast, the minimum average standard error values for all parameters estimated using the optimal allocation method were greater than the true standard errors after ComBat adjustment.

### Limitations

The current simulation study has several potential limitations. Performance of our optimal allocation algorithm was based on a single experiment available in the GEO data repository, GSE50397. Parameters for the simulation were based on an array platform. The effectiveness of optimal batch allocation methods under different distributions as well as different normalization strategies should be confirmed in subsequent studies. We only considered one source of batch effects. In high throughput experiments, batch effects are possible during numerous stages of the experiment [2]. We used differences in average propensity scores as our balancing metric. More sophisticated balancing metrics may achieve superior results, especially when there is substantial heterogeneity in propensity score variability across batches. Our simulation framework also assumed no missing covariate data and furthermore, that all covariates are known and measured (no unmeasured confounding). In practice, as with all observational study designs, these assumptions should be rigorously evaluated by the investigator before and after batch correction. Finally, our study assumes that batch effects are known and a balanced group design is achievable (balance in covariate distribution across batches). We acknowledge there are situations where this may not be feasible due to identification of batch effects that were not anticipated during the initial experimental design. Inclusion of technical replicates is a powerful tool to aid in diagnosing and managing these unanticipated batch effects. Batch effects in the context of an unbalanced design is beyond scope of current study. We refer the interested reader to Li et al. [6] for an elegant review of the impact of unbalanced design on batch correction and a potential solution for addressing this concern.

Overall, we developed and tested our sample allocation algorithm under an idealized scenario. The algorithm was based on a relatively small number of samples. We acknowledge that there are situations in which the algorithm may not be computationally feasible. In these more complicated situations, we recommend an adaptation of our algorithm that reviews a computationally feasible number of randomization iterations, based on user input, and selects the randomization iteration that minimizes imbalance in key covariates identified by the investigator across batches. This function is available at <https://github.com/carryp/PS-Batch-Effect>. We also created a detailed vignette to accompany the functions that describes in detail how to use the functions as well as additional applications of the functions to situations such as random sampling in which balance in key covariates is desirable.

### Conclusions

The current study reinforces the importance of proper experimental design for high throughput studies. Average of maximum absolute bias and variability of maximum absolute bias were consistently minimized in the optimal allocation method. All three allocation strategies performed well following adjustment for batch effects. However, in

contrast to the randomization conditions, levels of bias in the optimal allocation condition prior to batch correction were similar to levels of bias after batch correction, indicating the optimal allocation method sets up a robust downstream analysis that minimizes the need for subsequent batch correction. This is an important finding because although batch correction performs well, on average, randomization has the potential to produce sample allocations where covariates are not equally distributed between batches and consequently, batch correction methods do not perform as well. Our optimal sample allocation strategy avoids this scenario by selecting the sample allocation strategy that minimizes imbalance in covariates. As indicated by the lower variability in bias, this strategy is expected to represent a more consistent method for minimize bias. Furthermore, standard error values estimated using the optimal method were never less than the true standard error values. Building on covariate constrained randomization methods [21], our sample allocation methodology allows the investigator to exploit knowledge of subject level confounders to optimize experimental design. Relative to other methods, our algorithm provides additional flexibility to handle continuous variables as well as interactions between covariates. We used the average propensity score as our balancing metric, selecting the sample allocation iteration that minimized differences in average propensity score between batches. Although we focused on an expression set representing array data, we expect the results to generalize to other -omics datasets where features have been transformed to approximate a normal distribution.

## Methods

### Gene expression dataset

We illustrate our novel batch allocation algorithm (Additional file 1) in a hypothetical case-control study. In order to create a biologically plausible scenario, we downloaded a publicly available microarray gene expression dataset from NCBI GEO, GSE50397. The dataset includes gene expression data from 89 human pancreas islet donors. The samples were obtained from Nordic Islet Transplantation Programme, Uppsala University, see <http://www.nordicislets.org> for more information about islet processing and isolation. Data processing is described in greater detail in previous publications [27–29]. In brief, microarray profiling was performed using the Affymetrix GeneChip® Human Gene 1.0 ST whole transcript platform. Using the oligo R package, the Robust Multi-array Analysis (RMA) method was used to summarize and normalize the array data. Batch correction was performed using COMBAT function from SVA package [7]. The top 10,000 most variable genes from the batch corrected dataset (downloaded from GEO) were used for all subsequent analyses as the ‘true’ gene expression values.

### Phenotype selection

Using the phenotype information available from GSE50397, we selected two variables known to be important in pancreas islet cell gene expression: (1) age and (2) hemoglobin A1c (HbA1c). We focused on individuals with complete age and HbA1c data ( $n=65$ ). In order to create a scenario where these variables were differentially distributed across the case and control groups, we used a logistic regression model to assign subjects a probability of selection. First, age and BMI were normalized into Z scores so that age and BMI contributed equally to probability of selection. Next, subjects were assigned a

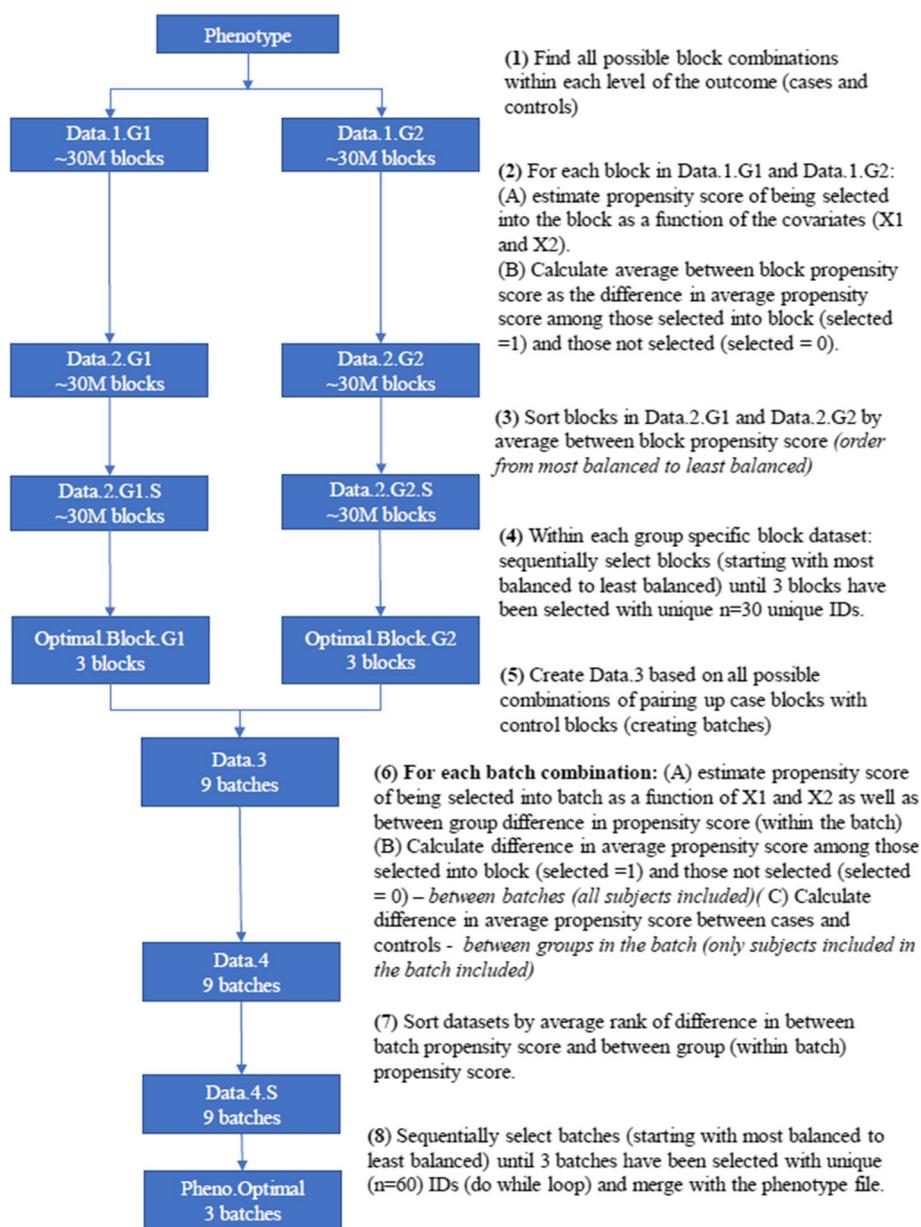
probability ( $P$ ) based on the following:  $P_i = \exp(2.5 + \ln(1.5) * Z_{Age_i} + \ln(1.5) * Z_{HbA1c_i}) / 1 + \exp(2.5 + \ln(1.5) * Z_{Age_i} + \ln(1.5) * Z_{HbA1c_i})$ . Subjects that were assigned a higher probability were more likely to be older (odds ratio per 1 SD increase in age: 1.5) and have higher HbA1c levels (odds ratio per 1 sd increase in HbA1c level: 1.5). Subjects were then ranked 1 through 65 based on this probability. We then selected the bottom 60 individuals (ranks 6–65) for the study. We preferentially selected lower probabilities to avoid including extreme or outlier age and/or HbA1c values. In this subset of 60 subjects, the bottom 30 were assigned to the case group and the remaining 30 were assigned to the control group. In this model, 2.5 represents the intercept. Given that the subjects were ranked based on the probability of selection ( $P$ ), the intercept parameter has no influence on sample selection, the same 60 subjects would have been selected and furthermore the same subjects would have been allocated to the case vs control groups regardless of the intercept value.

### Batch assignment

We assigned subjects to batches using three strategies: simple randomization, stratified randomization, and optimal allocation. The simple randomization strategy was implemented by randomly sampling, without replacement, all of the batch labels using the `sample()` function in `r`. The stratified randomization used a similar strategy; however, batches were randomly assigned within the case and control strata in order to preserve a balanced design (equal number of cases and controls within each batch). Stratified randomization was executed using the `experiment R` [30] package (v1.2.0).

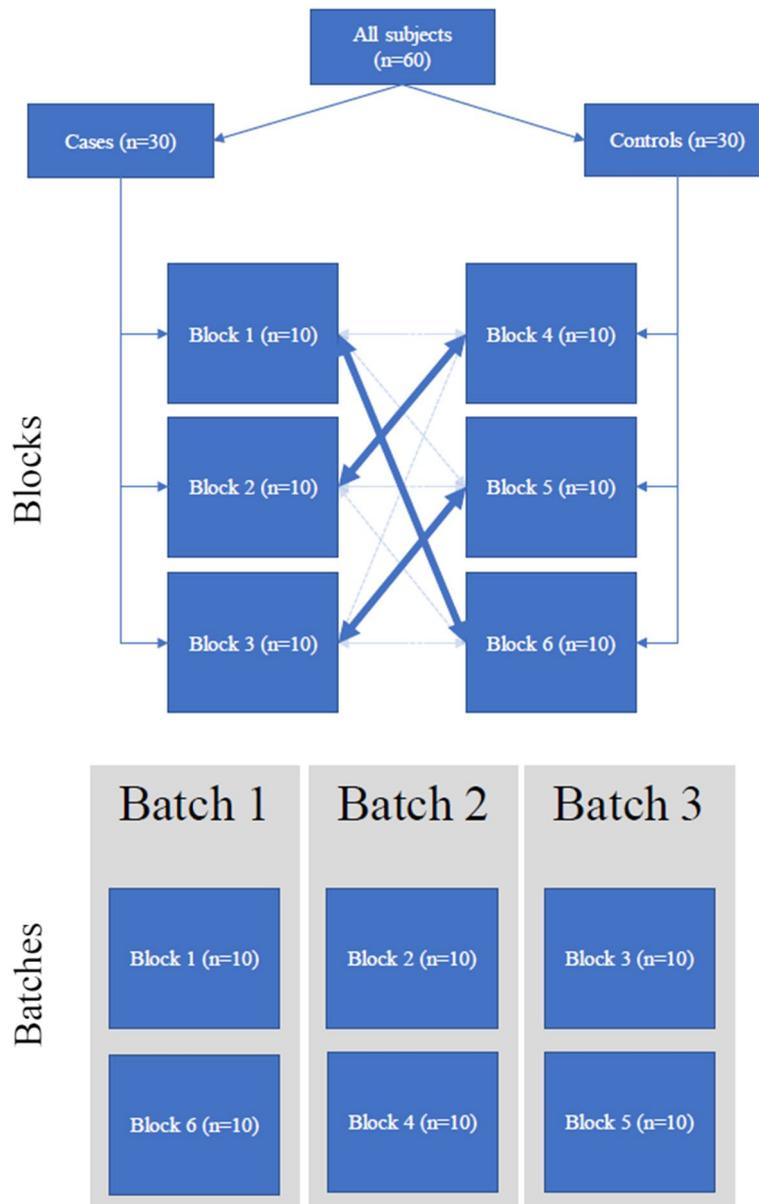
For the optimal allocation strategy, we used our novel propensity score allocation algorithm (See data availability section, Additional file 1) that considers all possible ways of assigning samples to batches and then selects the allocation that minimizes the difference in average between batch propensity score, see Figs. 4 and 5. In step 1, to maintain a balanced design, 10 subjects from each group (case or control) are assigned to each batch. This creates 3 blocks of 10 subjects per group. We reviewed all possible block combinations. The blocking steps were performed separately for cases and controls to minimize the computational burden. In step 2, once all block combinations were determined, within the case strata, we reviewed the probability (propensity score) of being assigned to each individual block as a function of the covariates,  $\text{pr}(\text{block selection} = 1 | \text{age, HbA1c})$ . Propensity scores for each block were estimated using a logistic regression model in which batch assignment was the outcome (1 vs 0) and age and HbA1c were included as covariates. For each block, we calculated the difference in average propensity score among subjects assigned to the block versus the average propensity score among all other case subjects not assigned to the block. In step 3, we then sorted all possible blocks by this difference in propensity score. Using this approach, it is possible for the same subject to be present in multiple blocks. Therefore, in step 4, we iteratively selected blocks, starting with the lowest propensity score difference, until 3 blocks of unique subjects had been selected. The same process was repeated in the control strata.

In steps 5 and 6, we considered all possible ways of assigning the three case blocks and the three control blocks to the three batches (see Fig. 5. For each batch combination (one case block and one control block together), we calculated the probability



**Fig. 4** Batch allocation algorithm based on propensity scores. Figure describes the algorithm used to assign subjects to batches. Using a balanced design, the algorithm uses propensity scores in two steps to minimize imbalance in confounders across the study batches

of being selected into the batch using a logistic model,  $\text{pr}(\text{batch selection} = 1 | \text{Age, HbA1c})$ . Case status was not considered in this step because we used a balanced design (See Fig. 5). We calculated between batch and between group (within a batch) differences in propensity scores. Between batch propensity scores for each batch combination were calculated as the average propensity score among all subjects assigned to the batch compared to the average propensity score among all other subjects. The between group propensity score was calculated as the difference in average propensity score among cases within the selected batch and the average propensity score among



**Fig. 5** Selection of blocks and batches. Figure provides a visual representation of the batch selection algorithm, highlighting the distinction between blocks and batches. All subjects were split into a case and a control population. All possible blocks of 10 subjects within each population were evaluated and the top 3 blocks were considered. Next, we evaluated all ways of pairing the top 3 case blocks with the top 3 control blocks in order to form the three non-overlapping batches

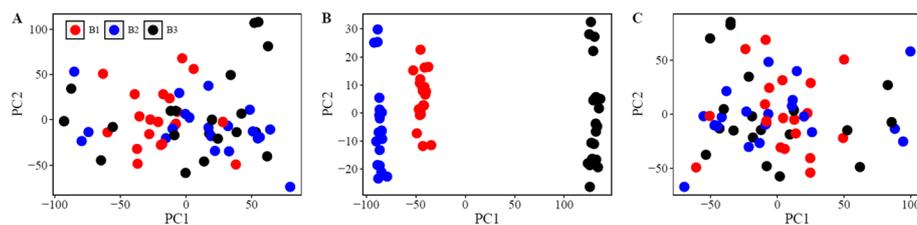
controls within the selected batch. The within batch propensity score difference (case vs control in a given batch) and the between group propensity score (selected in batch vs not selected into batch) differences were assigned a rank. In step 7, we then sorted the batch combinations by the average of the between batch propensity score rank and the between group batch propensity score rank. In the final step, we sequentially selected batch combinations starting with the batch combination having the lowest average rank, until 3 batches of unique subjects had been selected. In summary,

the propensity scores represent probability of selection conditional on all covariates included in model, providing one metric that captures balance in the distribution of all covariates. Therefore, by selecting the allocation iteration that minimizes (i) the difference in average propensity between batches and, (ii) within a given batch, the iteration that minimizes average propensity scores between cases and controls, we are able to identify the allocation that achieves balance in all covariates.

### Batch effect simulation

In order to simulate batch effects, we used the top 10,000 most variable genes from the gene expression dataset. This dataset represented the ‘true’ gene expression values. We elected to focus on genes with a higher underlying variance based on the assumption that higher variance features in array designs are, in general, are more reproducible [31].

For each allocation strategy (Table 1), batch effects were simulated by adding a random normal variable to the ‘true’ gene expression values. In each iteration the batch effect was simulated for the three batches using random normal variables,  $\mu_1 \sim (0, 0.57)$ ,  $\mu_2 \sim (0, 0.57)$ , and  $\mu_3 \sim (0, 0.57)$ . The variability was set to be approximately twice the median biological variability in the ‘true’ gene expression set. The batch effects were then added to the ‘true’ expression values in addition to a random variable (noise), using  $Y_s = Y_T + \mu_1 * B1_s + \mu_2 * B2_s + \mu_3 * B3_s + \omega$ , where  $Y_T$  is a  $n \times 10,000$  matrix representing true gene expression values, B1, B2, and B3 are vectors of length  $n$  of indicator variables (0 or 1) indicating whether subjects are vs are not assigned to batches 1,2, or 3 under each of the scenarios, and  $\omega_s$  is an  $n \times 10,000$  matrix of random error  $\sim N(0, 0.05)$  under each of the experimental scenarios ( $S$ =optimal allocation, simple random allocation, or stratified randomization allocation scenario). The random error was simulated to be small, less than 1/5 of biological error. Next, we considered two strategies for accounting for batch effects: ComBat, implemented through the sva R package [7] (v3.36.0), and a linear regression approach where batches were adjusted for as fixed effects in a linear regression modelling framework. Using principal component analysis of the 10,000 gene expression features included in the microarray gene expression dataset, Fig. 6 provides a visual



**Fig. 6** Example of batch effect simulation. A principal component analysis was used to visualize the addition of batch effects to the ‘true expression’ dataset. Each dot represents a subject. In each plot, the y-axis represents PC2 and the x-axis represents PC1 from the principal component analysis of all 10,000 genes included in the microarray gene expression dataset. The panel illustrates the distribution of batches across samples (random) prior to adding batch effects (A), after adding batch effects (B), and again after batch effects were removed using ComBat (C)

representation of batch effects before and after batch effect adjustment under the randomization condition.

### Outcome measures

We considered 9 experimental scenarios to assess the differences in the allocation methods before and after downstream adjustment (see Table 1). Under each experimental scenario, linear regression models were used to estimate the beta parameters for the case variable ( $\beta_1$ ), as well as the two confounders ( $\beta_2$  for age and  $\beta_3$  for HbA1c, ). Separate models were fit for each of the genes in the expression dataset. Absolute bias was calculated as the absolute difference between the parameter estimates in each experimental scenario relative to the ‘true’ parameter estimates,  $\text{absolute bias}_{ie} = |\beta_{ie}^* - \beta_{ie}|$  where \*represents the ‘true’ condition, i represents gene feature within the expression set ( $i = 1$  to 10,000 features) and e represents simulation iteration number ( $e = 1$  to 1,000 iterations). We also estimated the average standard error for each beta across all of the expression features. Next, for each simulation iteration, we calculated, (1) average maximum absolute bias (mean of the maximum absolute bias across all iterations), (2) root mean square (RMS) of absolute maximum bias (standard deviation of the maximum absolute bias across all iterations), (3) mean standard error (mean of the average standard error estimates across all iterations), and (4) minimum average standard error (minimum of the average standard error estimates across all iterations).

We also identified a single transcript that was significantly related to both age and HbA1c levels in the true gene expression dataset. This transcript represented expression of the CAPN13 gene, a protein coding gene that maps to chromosome 2, location 30,945,637 to 31,030,311. We looked at mean absolute bias, maximum absolute bias, and RMS of absolute bias at this gene across all the experimental iterations. Expression of this gene was used to evaluate the performance of our optimal allocation strategy under the alternative hypothesis. The simulation experiment was repeated using larger batch effects (3x and 4x biological variation). The results of these additional simulation experiments are consistent with the primary results and thus, are included as a supplementary files only (see “Additional file 2”).

### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05202-6>.

**Additional file 1.** This supplementary file includes the code used to develop the optimal allocation algorithm described in the methods section. The following programs are included in this file: (A) data processing, (B) testing association with gene expression, (C) propensity score based sample allocation algorithm, (D) simulating batch effects and testing performance, (E) summarizing performance.

**Additional file 2.** Results of the simulation experiment under different batch effects.

### Acknowledgements

Not applicable

### Author contributions

PC designed the study, developed the algorithm, performed the data analysis, interpreted the data, drafted the manuscript, and prepare the figures. TV, KK, and LV helped to develop the algorithm, interpreted the data, and edited the manuscript. PC and CK developed a more computationally feasible version of the algorithm. JN aided in study development, data analysis interpretation, and edited the manuscript. FD, TB, EL, CK, and IY edited the manuscript. All the author have read and approved the final version of the manuscript.

### Funding

This work was funded by National Institutes of Health (Grant No. 5R21AI142483-02).

**Availability of data and materials**

Source coding for the algorithm included in the Additional file 2. Microarray gene expression data and complete phenotype data are available for download through NCBI GEO, GSE50397. Datasets used in the simulations also included in the Additional file 2. We have also developed a more computationally feasible version of our algorithm to assist in experimental design, the code and accompanying vignette are available here: <https://github.com/carryp/PS-Batch-Effect>

**Declarations****Ethics approval and consent to participate**

Not applicable.

**Consent for publication**

Not applicable.

**Competing interests**

The authors declare that they have no competing interests.

Received: 18 August 2022 Accepted: 22 February 2023

Published online: 07 March 2023

**References**

- Chen C, Grennan K, Badner J, Zhang D, Gershon E, Jin L, Liu C. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*. 2011;6(2):e17238.
- Leek JT, Scharpf RB, Bravo HC, Simcha D, Langmead B, Johnson WE, Geman D, Baggerly K, Irizarry RA. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat Rev Genet*. 2010;11(10):733–9.
- Lazar C, Meganck S, Taminau J, Steenhoff D, Coletta A, Molter C, Weiss-Solis DY, Duque R, Bersini H, Nowe A. Batch effect removal methods for microarray gene expression data integration: a survey. *Brief Bioinform*. 2013;14(4):469–90.
- Scholes AN, Lewis JA. Comparison of RNA isolation methods on RNA-Seq: implications for differential expression and meta-analyses. *BMC Genom*. 2020;21(1):249.
- Goh WWB, Wang W, Wong L. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol*. 2017;35(6):498–507.
- Li T, Zhang Y, Patil P, Johnson WE. Overcoming the impacts of two-step batch effect correction on gene expression estimation and inference. *Biostatistics*. 2021.
- Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–3.
- Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 2012;13(3):539–52.
- Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–27.
- Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS. Adjustment of systematic microarray data biases. *Bioinformatics*. 2004;20(1):105–14.
- Sims AH, Smethurst GJ, Hey Y, Okoniewski MJ, Pepper SD, Howell A, Miller CJ, Clarke RB. The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets—improving meta-analysis and prediction of prognosis. *BMC Med Genom*. 2008;1:42.
- Luo J, Schumacher M, Scherer A, Sanoudou D, Megherbi D, Davison T, Shi T, Tong W, Shi L, Hong H, et al. A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. *Pharmacogenom J*. 2010;10(4):278–91.
- Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc Natl Acad Sci U S A*. 2001;98(1):31–6.
- Nygaard V, Rodland EA, Hovig E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*. 2016;17(1):29–39.
- Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika*. 1983;70(1):41–55.
- Hernán MA, Robins J. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC; 2020.
- Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000;11(5):550–60.
- Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat Med*. 2008;27(12):2037–49.
- Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. *Stat Med*. 2015;34(28):3661–79.
- Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci*. 2010;25(1):1–21.
- Moulton LH. Covariate-based constrained randomization of group-randomized trials. *Clin Trials*. 2004;1(3):297–305.
- Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol*. 1999;52(1):19–26.
- Ivers NM, Halperin IJ, Barnsley J, Grimshaw JM, Shah BR, Tu K, Upshur R, Zwarenstein M. Allocation techniques for balance at baseline in cluster randomized trials: a methodological review. *Trials*. 2012;13:120.

24. Yan L, Ma C, Wang D, Hu Q, Qin M, Conroy JM, Sucheston LE, Ambrosone CB, Johnson CS, Wang J, et al. OSAT: a tool for sample-to-batch allocations in genomics experiments. *BMC Genom.* 2012;13:689.
25. Jiao C, Zhang C, Dai R, Xia Y, Wang K, Giase G, Chen C, Liu C. Positional effects revealed in Illumina methylation array and the impact on analysis. *Epigenomics.* 2018;10(5):643–59.
26. Buhule OD, Minster RL, Hawley NL, Medvedovic M, Sun G, Viali S, Deka R, McGarvey ST, Weeks DE. Stratified randomization controls better for batch effects in 450K methylation analysis: a cautionary tale. *Front Genet.* 2014;5:354.
27. Taneera J, Fadista J, Ahlqvist E, Atac D, Ottosson-Laakso E, Wollheim CB, Groop L. Identification of novel genes for glucose metabolism based upon expression pattern in human islets and effect on insulin secretion and glycemia. *Hum Mol Genet.* 2015;24(7):1945–55.
28. Taneera J, Mohammed AK, Dhaiban S, Hamad M, Prasad RB, Sulaiman N, Salehi A. RORB and RORC associate with human islet dysfunction and inhibit insulin secretion in INS-1 cells. *Islets.* 2019;11(1):10–20.
29. Taneera J, Mohammed I, Mohammed AK, Hachim M, Dhaiban S, Malek A, Duner P, Elemam NM, Sulaiman N, Hamad M, et al. Orphan G-protein coupled receptor 183 (GPR183) potentiates insulin secretion and prevents glucotoxicity-induced beta-cell dysfunction. *Mol Cell Endocrinol.* 2020;499:110592.
30. Imai K, Imai MK, Boot D. Experiment: R package for designing and analyzing randomized experiments. 2007.
31. Logue MW, Smith AK, Wolf EJ, Maniates H, Stone A, Schichman SA, McGlinchey RE, Milberg W, Miller MW. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics.* 2017;9(11):1363–71.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

