Supplemental Material — PLSDA-batch: a multivariate framework to correct for batch effects in microbiome data

Yiwen Wang[1,2], Kim-Anh Lê Cao[2,*]

1 Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, 97 Buxin Rd, Shenzhen, 518000, Guangdong, China.

2 Melbourne Integrative Genomics, School of Mathematics and Statistics, The University of Melbourne, 30 Royal Parade, Melbourne, 3052, VIC, Australia.

* Corresponding author: kimanh.lecao@unimelb.edu.au

S1 Existing Methods

S1.1 removeBatchEffect

It is a location-scale and univariate method [1]. It has been used in a study of human oral microbiome to remove batch effects caused by different experimental times [2]. Let x_{ijcb} denotes the abundance value for the variable j of sample i from the treatment group c and batch b. removeBatchEffect includes batch effects as covariates and models x_{ijcb} as:

$$x_{ijcb} = \mu_j + y_{ic}^{(trt)}\alpha_{jc} + y_{ib}^{(batch)}\beta_{jb} + \epsilon_{ij},$$

where μ_j is the overall abundance of variable j. $y_{ic}^{(trt)}$ and $y_{ib}^{(batch)}$ represent the condition of sample i in the treatment c or batch b respectively, and α_{jc} and β_{jb} represent the corresponding regression coefficient for the variable j in the treatment c or batch b separately. ϵ_{ij} is the error term assumed to follow a normal distribution $N(0, \sigma_j^2)$. Via removeBatchEffect, we first estimate the batch effect coefficients and then calculate the batch effect corrected data as $\hat{x}_{ijcb} = x_{ijcb} - y_{ib}^{(batch)}\hat{\beta}_{jb}$.

S1.2 ComBat

It is a location-scale and univariate method using empirical Bayesian model to estimate parameters [3]. It assumes batch effects are systematic across all variables. ComBat has been applied to a study of human lung microbiome to correct for batch effects caused by different research groups [4] and another study of bowel disease affected gut microbiome [5]. The abundance value x_{ijcb} is formulated using the same notations as removeBatchEffect:

$$x_{ijcb} = \mu_j + y_{ic}^{(trt)}\alpha_{jc} + y_{ib}^{(batch)}\beta_{jb} + \delta_{jb}\epsilon_{ijb},$$

where δ_{jb} represents the multiplicative batch effect of batch b for variable j. Both additive (β_{jb}) and multiplicative batch effects (δ_{jb}) are modelled in ComBat. The final batch effect corrected data are calculated as $\hat{x}_{ijcb} = \hat{\mu}_j + y_{ic}^{(trt)}\hat{\alpha}_{jc} + \hat{\epsilon}_{ijb}$.

S1.3 Surrogate Variable Analysis

It is a hybrid of univariate and multivariate approaches to target differentially expressed variables with biological effects of interest under the control of batch effects but does not aim to generate batch effect corrected data [6]. It estimates unknown batch effects from a subset of variables mostly affected by batch effects and with very little treatment variation using singular value decomposition (SVD), then includes latent factors as estimated batch effects in a linear model. This method has been recently used to assess the presence of batch effects to study microbial interactions with humans in age-related macular degeneration [7], and to account for plate ID in breast cancer microbiota [8]. The abundance value X_{ijcb} uses the same notations as the above methods:

$$X_{ijcb} = \mu_j + y_{ic}^{(trt)}\alpha_{jc} + \sum_{k=1}^{K} y_{kib}^{(batch)}\lambda_{kjb} + \epsilon_{ij},$$

where $\sum_{k=1}^{K} y_{kib}^{(batch)}\lambda_{kjb}$ represents the estimated batch effects $y_{kib}^{(batch)}$ provided by the SVD and the corresponding coefficients λ_{kjb}. The hypothesis test is then performed of whether $\alpha_{jc} = 0$ or not for each variable.

S2 Supporting Information for Weighted PLSDA-batch

The weighted PLSDA-batch approach aims to manage the unbalanced sample sizes within each batch and treatment group using appropriate weights. It was inspired by weighted PCA, which is detailed in [9]. Weighted PCA is used when heterogeneous groups are imbalanced (i.e. they do not contain the same number of observations) by weighting the observations by the inverse of the group size.

The singular value decomposition of PCA is defined as: $\boldsymbol{X} = \boldsymbol{USV}^\top$, where the columns of $\boldsymbol{US}$ are principal components, and the columns of $\boldsymbol{V}$ are principal loadings. For the weighted PCA, the decomposition is $\boldsymbol{WX} = \boldsymbol{U'S'V'}^\top$, where $\boldsymbol{W}$ is a diagonal matrix that includes the inverse of the group size for each observation.

Similarly, for PLSDA-batch we decompose $\boldsymbol{X}^\top\boldsymbol{Y} = \boldsymbol{ADB}^\top$, where the columns of $\boldsymbol{A}$ are the initial loadings of $\boldsymbol{X}$, columns of $\boldsymbol{B}$ are the initial of $\boldsymbol{Y}$. Therefore, following the rationale of weighted PCA, the decomposition of weighted PLSDA-batch becomes $\boldsymbol{X}^\top\boldsymbol{W'Y} = \boldsymbol{A'D'B'}^\top$ where $\boldsymbol{W'}$ is the diagonal matrix that includes the inverse of the group size for each observation. As $\boldsymbol{w}^\top\boldsymbol{w} = \boldsymbol{W'}$, $\boldsymbol{X}^\top\boldsymbol{W'Y} = (\boldsymbol{X}^\top\boldsymbol{w}^\top)\ (\boldsymbol{wY}) = (\boldsymbol{wX})^\top\ (\boldsymbol{wY})$, where $\boldsymbol{w}$ is defined as in the 'Weighted PLSDA-batch' in the Methods section using the inverse of the square root of the group size.

S3 Simulation Results

S3.1 Summary of all simulation results with two batch groups

The accuracy measurements for the simulation scenarios reported in Table 1 were presented in Supplemental Figures S1-S6.

To summarise, in the balanced design, SVA performed best with the highest, and sometimes greater, accuracy measurements than the ground-truth data. For the simulated data including variables with both treatment and batch effects, sPLSDA-batch performed worse than the other methods with $M^{(trt\ \&\ batch)}$ accounting for 50% to 70% of $min(M^{(trt)}, M^{(batch)})$ to reach the worst precision and F1 score. For the other scenarios, all methods except SVA performed similarly with results close to those from the ground-truth data.

In the highly unbalanced design, SVA performed worst among the other methods with results close to those from the original data. When the variability of batch effects among variables $\sigma_{(batch)}$ increased in the simulated datasets, the precision and F1 score of the ComBat corrected data decreased dramatically, while these two measurements from wPLSDA-batch and swPLSDA-batch correction remained stable. When the data simulated with more than 100 (1/3 of the total number of variables) variables with batch effects $M^{(batch)}$, wPLSDA-batch and swPLSDA-batch performed best and their corrected data gradually achieved similar results to the ground-truth data with the increase of $M^{(batch)}$. When $M^{(trt\ \&\ batch)}$ accounted for more than 30% of $min(M^{(trt)}, M^{(batch)})$, wPLSDA-batch might be a preferable choice over swPLSDA-batch as it led to better accuracy measurements.

S3.2 Simulations with Gaussian distribution

S3.2.1 Simulation strategy

We adapted the simulation strategy that is component-based and multivariate from [10]. We assumed the input data after filtering follow a lognormal distribution inspired from [11], thus after Centered Log Ratio (CLR) transformed follow a Gaussian distribution. Thus, we simulated components from a Gaussian distribution across all samples. The data matrix was generated based on the simulated components and corresponding loading vectors for each variable. Different parameters including amount of batch and treatment variability among samples, number of variables with batch and/or treatment effects, balanced and unbalanced batch × treatment designs were considered and summarised in Table S2.

Each simulated dataset included 300 variables and 40 samples grouped according to two treatments

(trt1 and trt2) and two batches (batch1 and batch2). The balanced batch × treatment experimental design included 10 samples from two batches respectively in each treatment group, while the unbalanced design had 4 and 16 samples from batch1 and batch2 respectively in trt1, 16 and 4 samples from batch1 and batch2 in trt2 (see Table 2).

We first generated two base components $\boldsymbol{t}^{(trt)}$ and $\boldsymbol{t}^{(batch)}$ to represent the underlying treatment and batch variation across samples in the datasets. The samples with trt1 or trt2 in the component $\boldsymbol{t}^{(trt)}$ were generated from $N(-\mu_{(trt)}, \sigma'^2_{(trt)})$ and $N(\mu_{(trt)}, \sigma'^2_{(trt)})$ respectively, where $\sigma'^2_{(trt)}$ refers to the variability of treatment effect among samples, and similarly for the batch component. We then sampled the corresponding loading vectors $\boldsymbol{\alpha}^{(trt)}$ and $\boldsymbol{\alpha}^{(batch)}$ from a uniform distribution $[-0.3, -0.2] \cup [0.2, 0.3]$ respectively and scaled them as an unit vector. We subsequently constructed the treatment relevant matrix as $\boldsymbol{X}^{(trt)} = \boldsymbol{t}^{(trt)}(\boldsymbol{\alpha}^{(trt)})^\top$ and similarly for the batch relevant matrix.

We also generated background noise $\boldsymbol{E}$ ($\boldsymbol{E} \in \mathbb{R}^{40\times 300}$), where each element was randomly sampled from $N(0, 0.2^2)$. The final simulated dataset $\boldsymbol{X}_{result}$ was constructed based on the treatment, batch relevant matrices and background noise. Starting with $\boldsymbol{X}_{result} = \boldsymbol{E}$, we then added different types of variables, such that:

$$\boldsymbol{X}_{result}[, \text{variables}^{(trt)}] = \boldsymbol{E}[, \text{variables}^{(trt)}] + \boldsymbol{X}^{(trt)}$$

$$\boldsymbol{X}_{result}[, \text{variables}^{(batch)}] = \boldsymbol{X}_{result}[, \text{variables}^{(batch)}] + \boldsymbol{X}^{(batch)},$$

where variables with treatment or batch effects were randomly indexed in the data.

Finally, we simulated a ground-truth dataset that only included the background noise and treatment but no batch effect to evaluate batch effect corrected datasets.

We simulated different scenarios summarised in Table S2 to verify the influence of different parameters.

S3.2.2 Accuracy measures

We used the same accuracy measures as mentioned in the main text Section Benchmarking and assessment of batch effect removal, namely Precision, Recall and F_1 score for both variables selected from one-way ANOVA (univariate) and sPLSDA (multivariate). In sPLSDA, since we specified the number of variables to select as the number of variables with a true treatment effect, these three measures are equivalent. We thus called this accuracy measure "multivariate selection" to distinguish from the results from one-way ANOVA (see Table S3).

S3.2.3 Simulation results

We measured the accuracy of batch effect corrected data from different methods applied to the simulated data under different scenarios as shown in Figure S18-S23. Here we describe only one scenario that we believe is a representative of real data ($M^{(trt\ \&\ batch)} = 30$, simulation 6 in Table S2).

We first considered the proportion of variance explained by treatment and batch effects before and after batch effect correction across all variables using pRDA. Efficient batch effect correction methods should generate data with a smaller proportion of batch associated variance and larger proportion of treatment variance compared to the original data. Figure S24**A** shows that there was no intersection shared between treatment and batch variation with a balanced batch $\times$ treatment design. All methods successfully removed batch variation, but PLSDA-batch and sPLSDA-batch preserved more proportion of treatment variance than removeBatchEffect and ComBat. In addition, the data corrected by sPLSDA-batch included almost as much proportion of treatment variance as the ground-truth data. With an unbalanced batch $\times$ treatment design (Figure S24**B**), we observed that certain amount of variance was shared (intersection) and explained by both batch and treatment effects. Such intersectional variance should exist even in the ground-truth data with no batch effect, as it originates from treatment variation because of the unbalanced design. Unweighted PLSDA-batch and sPLSDA-batch failed in such design, as their corrected data still included a large amount of batch variation (PLSDA-batch) or not included intersectional variance (sPLSDA-batch), while the other methods removed batch variation successfully. The corrected data from removeBatchEffect and ComBat included less proportion of variance explained by treatment but more intersectional variance compared to the ground-truth data. Although wPLSDA-batch corrected data included the largest treatment variance, swPLSDA-batch outperformed all methods with results similar to the ground-truth data.

We also estimated the proportion of variance explained by treatment and batch effects for each variable respectively using the R^2 value. In the balanced batch $\times$ treatment design (Figure S25**A**), the variables assigned with both treatment and batch effects in the corrected data from removeBatchEffect and ComBat presented less proportion of treatment associated variance than in the ground truth data. This result agrees with the pRDA evaluation that these two methods do not preserve enough treatment variation. After PLSDA-batch correction, variables simulated with only batch effects displayed some amount of treatment variation, but only in the case where the batch effect variability among samples was high. sPLSDA-batch outperformed all methods, with results similar to the ground-truth data. In the unbalanced design (Figure S25**B**), variables assigned with both treatment and batch effects were

segregated into two groups depending on whether their abundance increased or decreased consistently or not according to the two effects. We observed similar results to those obtained from the balanced design (Figure S25**A**).

When considering the measures of accuracy with univariate one-way ANOVA, we observed that for both balanced and unbalanced designs the corrected data from PLSDA-batch, wPLSDA-batch, sPLSDA-batch and swPLSDA-batch led to higher recall and lower precision than the data from removeBatchEffect and ComBat (Table S3). However, the precision of sPLSDA-batch and swPLSDA-batch was competitive to removeBatchEffect and ComBat for each type of design. Moreover, both versions of weighted and unweighted sPLSDA-batch achieved higher F1 scores and multivariate selection scores than removeBatchEffect and ComBat in each design. The standard deviations of the multivariate selection scores were all smaller than the univariate selection scores for the different corrected data, indicating a better stability of the variables selected by multivariate sPLSDA compared to the one-way ANOVA univariate selection.

We observed similar but higher resolution results of accuracy measures for the other simulation scenarios presented in Figures S18-S23. When the variability of batch effects among samples $\sigma'_{(batch)}$ increased, the precision of PLSDA-batch decreased dramatically, but the precision of sPLSDA-batch slightly increased and outperformed removeBatchEffect and ComBat in both designs. In all scenarios with a high variability of batch effects ($\sigma'_{(batch)} = 8$), PLSDA-batch performed the worst among all the methods. The change of mean ($\mu_{(trt)}$) and variability ($\sigma'_{(trt)}$) of treatment effects did not largely affect any accuracy measurement. When the number of variables associated either with treatment or batch effects increased, the precision of sPLSDA-batch increased and was slightly higher than removeBatchEffect and ComBat, especially for the unbalanced design. sPLSDA-batch outperformed the other methods in all scenarios except for the case when a large number of variables were influenced by both treatment and batch effects (greater than half the number of variables with treatment effects), resulting in a lower precision but still higher recall than the other two univariate batch effect correction methods.

S4 Supporting Tables

Table S1. Simulation studies (three batch groups): summary of accuracy measurements before and after batch effect correction. The proportion of correctly identified microbial variables with a true treatment effect was assessed with Precision, Recall, F1 score (using one-way ANOVA as variable selection procedure) and AUC (using sPLSDA as variable selection procedure). Each value is the mean (or standard deviation) over 50 repeats.

		Before correction	Ground-truth data	SVA	removeBatchEffect	ComBat	PLSDA-batch	sPLSDA-batch
Balanced	Precision	**0.986** (0.03)	0.954 (0.07)	0.964 (0.02)	0.949 (0.07)	0.940 (0.08)	0.957 (0.06)	0.856 (0.08)
	Recall	0.667 (0.04)	0.891 (0.03)	**0.934** (0.02)	0.902 (0.03)	0.903 (0.03)	0.896 (0.03)	0.884 (0.03)
	F1	0.795 (0.03)	0.920 (0.04)	**0.948** (0.01)	0.923 (0.04)	0.919 (0.04)	0.924 (0.03)	0.867 (0.04)
	AUC	0.940 (0.02)	0.959 (0.02)	/	**0.964** (0.02)	**0.964** (0.02)	**0.964** (0.02)	0.949 (0.02)
		Before correction	Ground-truth data	SVA	removeBatchEffect	ComBat	wPLSDA-batch	swPLSDA-batch
Unbalanced	Precision	0.637 (0.03)	**0.972** (0.04)	0.648 (0.08)	0.862 (0.12)	0.834 (0.12)	0.915 (0.08)	0.863 (0.10)
	Recall	0.811 (0.03)	0.884 (0.03)	0.872 (0.16)	0.897 (0.03)	**0.904** (0.03)	0.855 (0.04)	0.844 (0.03)
	F1	0.713 (0.03)	**0.925** (0.03)	0.725 (0.11)	0.874 (0.07)	0.863 (0.07)	0.882 (0.04)	0.850 (0.06)
	AUC	0.826 (0.02)	**0.963** (0.02)	/	0.954 (0.02)	0.955 (0.02)	0.948 (0.02)	0.923 (0.02)

Table S2. Summary of simulation scenarios (Gaussian distribution). For a given choice of parameters reported in this table, each simulation was repeated 50 times. $M^{(trt)}, M^{(batch)}$ and $M^{(trt\ \&\ batch)}$ represent the number of variables with treatment, batch, or both effects respectively. **Simulation 6** includes parameters reflective of real data.

Parameters	$\mu_{(trt)}$	$\sigma'_{(trt)}$	$\mu_{(batch)}$	$\sigma'_{(batch)}$	$M^{(trt)}$	$M^{(batch)}$	$M^{(trt\ \&\ batch)}$
Simulation 1	3	1	7	{1,4,8}	60	150	0
Simulation 2	{3,5,7}	1	7	8	60	150	0
Simulation 3	3	{1,2,4}	7	8	60	150	0
Simulation 4	3	2	7	8	{30,60,100,150}	150	0
Simulation 5	3	2	7	8	60	{30,60,100,150}	0
Simulation 6	3	2	7	8	60	150	{0,18,30,42,60}

Table S3. Simulation studies (Gaussian distribution): summary of accuracy measures before and after batch correction. The proportion of correctly identified microbial variables with a true treatment effect was assessed with Precision, Recall, F1 score (using one-way ANOVA as variable selection procedure) and Multivariate selection score (using sPLSDA as variable selection procedure). Each value is the mean (or standard deviation) over 50 repeats.

		Before correction	Ground-truth data	removeBatchEffect	ComBat	PLSDA-batch	sPLSDA-batch
Balanced	Precision	**0.98** (0.02)	0.95 (0.03)	0.94 (0.15)	0.93 (0.16)	0.56 (0.25)	0.86 (0.11)
	Recall	0.74 (0.10)	**1.00** (0.00)	0.87 (0.10)	0.88 (0.10)	**1.00** (0.02)	**1.00** (0.00)
	F1	0.84 (0.06)	**0.98** (0.02)	0.89 (0.12)	0.89 (0.12)	0.68 (0.20)	0.92 (0.07)
	Multivariate selection	0.89 (0.06)	**1.00** (0.00)	0.92 (0.07)	0.92 (0.07)	0.92 (0.12)	**1.00** (0.01)
		Before correction	Ground-truth data	removeBatchEffect	ComBat	wPLSDA-batch	swPLSDA-batch
Unbalanced	Precision	0.52 (0.32)	**0.96** (0.03)	0.85 (0.18)	0.80 (0.23)	0.52 (0.23)	0.80 (0.14)
	Recall	0.72 (0.04)	**1.00** (0.00)	0.86 (0.10)	0.86 (0.10)	0.99 (0.03)	**1.00** (0.00)
	F1	0.55 (0.21)	**0.98** (0.02)	0.84 (0.14)	0.81 (0.18)	0.65 (0.19)	0.88 (0.10)
	Multivariate selection	0.73 (0.05)	**1.00** (0.00)	0.88 (0.07)	0.87 (0.08)	0.89 (0.15)	0.99 (0.02)

S5 Supporting Figures

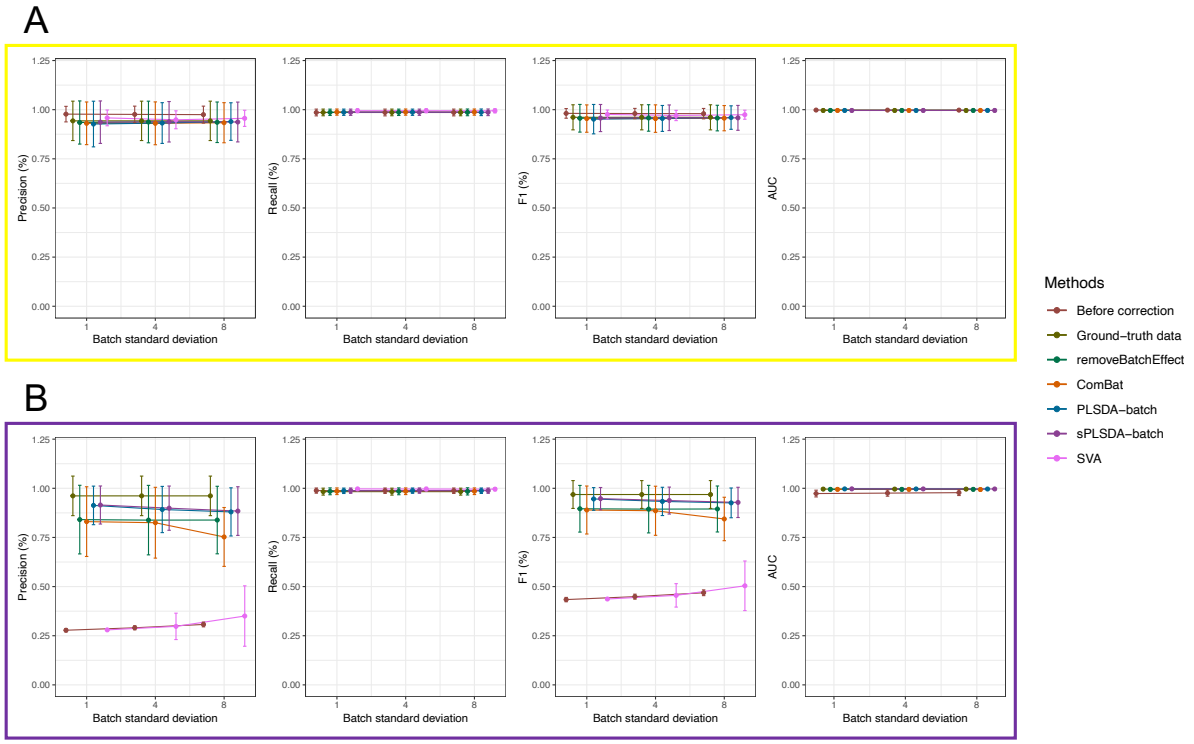

Figure S1. Simulation 1 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different batch effect variability among variables (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Batch effects were generated with three choices of variability $\sigma_{(batch)}$ among variables (x-axis). PLSDA-batch and sPLSDA-batch in the unbalanced design used the weighted versions. The proportion of correctly identified microbial variables with a true treatment effect was assessed with Precision, Recall, F1 score (using one-way ANOVA as variable selection procedure) and AUC (using sPLSDA as variable selection procedure). As SVA is unable to generate batch effect corrected data, there is no AUC value to evaluate this method. Each point was averaged over 50 repeatedly simulated data, with error bars indicating estimated sample standard deviations. In the balanced design, the change of $\sigma_{(batch)}$ did not affect the performance of different batch effect correction methods. SVA controlled data and the original data had slightly higher precision and F1 score compared to the other methods corrected data and the ground-truth data, while the other methods performed similarly. In the unbalanced design, the precision and F1 score of the data corrected from PLSDA-batch and sPLSDA-batch were much higher than from the other methods and the original data, and slowly decreased as $\sigma_{(batch)}$ increased. SVA performed worst among all the methods based on the precision and F1 score. The AUC of the original data with an unbalanced design was lower than the other datasets highlighting the importance of removing batch effects. In the other scenarios, the recall and AUC were similar among different datasets and different choices of $\sigma_{(batch)}$ in both cases of balanced and unbalanced designs.

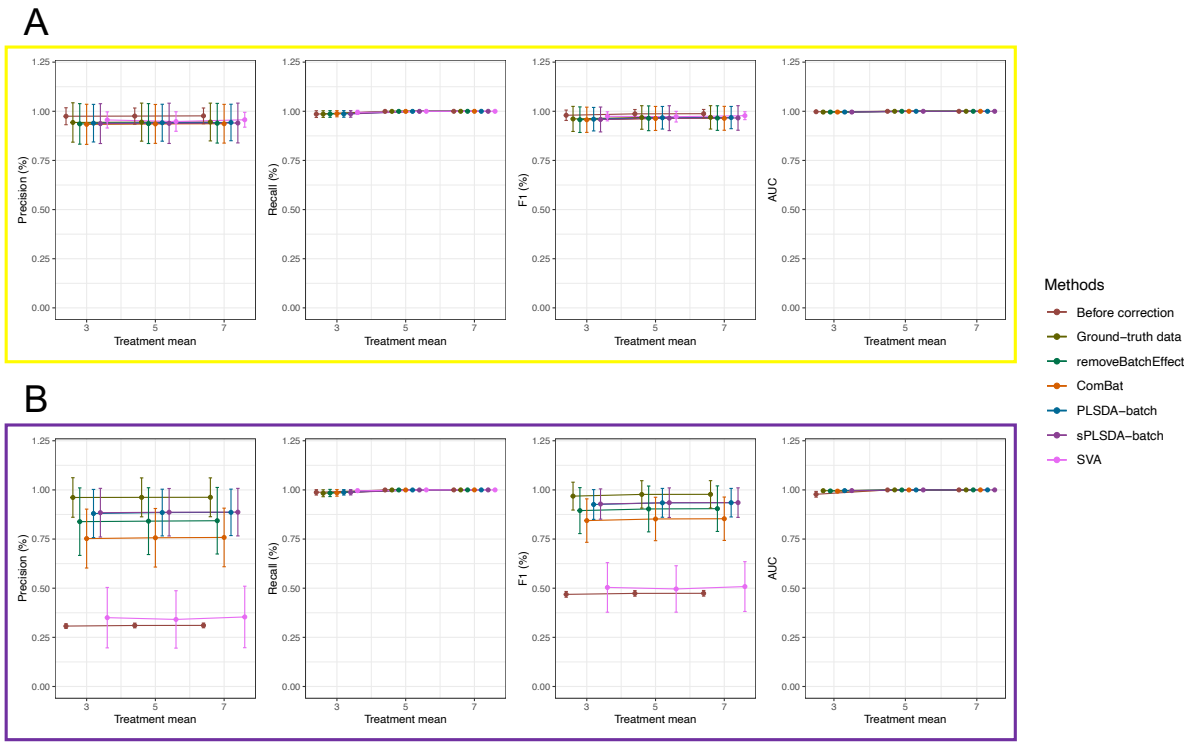

Figure S2. Simulation 2 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different sizes of treatment effects (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Treatment effects were generated with three choices of sizes $\mu_{(trt)}$ (x-axis). The description of these plots is detailed in Figure S1. The change of $\mu_{(trt)}$ did not affect the performance of different batch effect correction methods. In the balanced design, different methods performed similarly. In the unbalanced design, PLSDA-batch and sPLSDA-batch performed much better, while SVA much worse than the other methods based on the precision and F1 score. The recall and AUC were similar among different datasets and different choices of $\mu_{(trt)}$ in both cases of balanced and unbalanced designs.

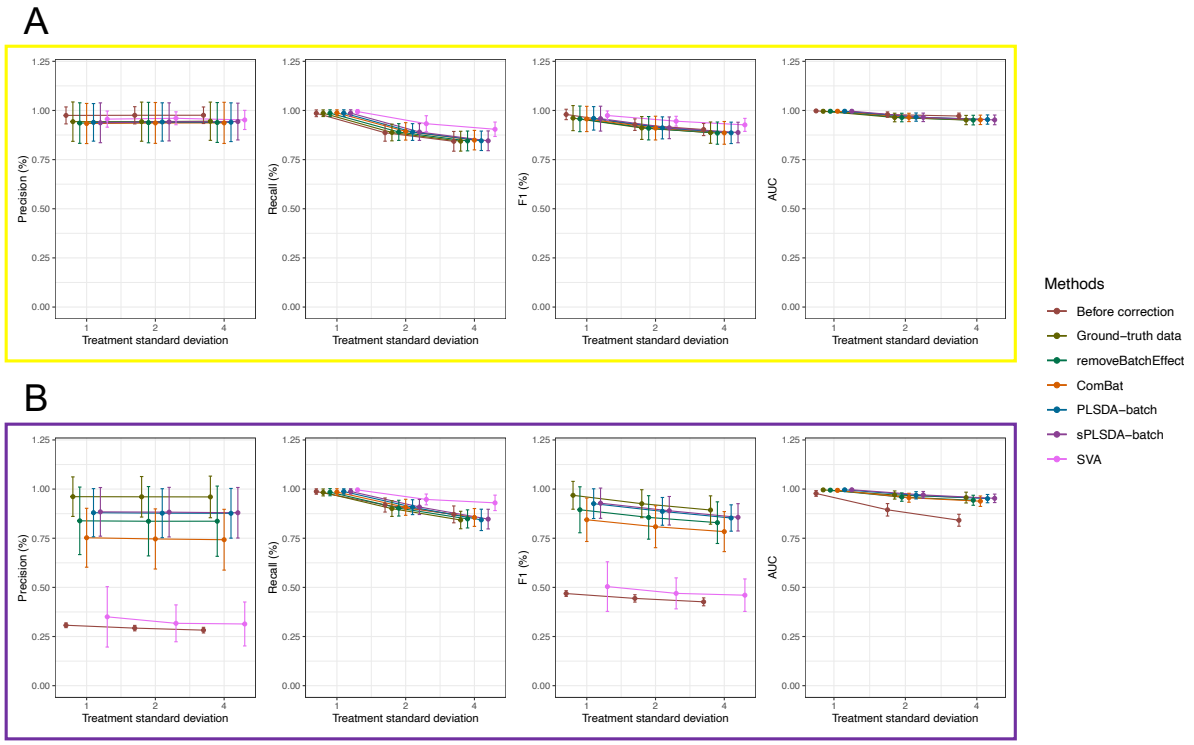

Figure S3. Simulation 3 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different treatment effect variability among variables (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Treatment effects were generated with three choices of variability $\sigma_{(trt)}$ among variables (x-axis). The description of these plots is detailed in Figure S1. In both designs, the increase of $\sigma_{(trt)}$ did not affect the precision, but decreased the recall, F1 score and AUC of different methods corrected datasets. In the balanced design, SVA controlled data had higher accuracy measurements compared to the other datasets including the ground-truth data. In the unbalanced design, PLSDA-batch and sPLSDA-batch corrected data had higher precision, while SVA controlled data had higher recall, but overall, PLSDA-batch and sPLSDA-batch had higher F1 score compared to the others. The reason is SVA selected more than twice number of variables with treatment effects compared to the selection from the other batch effect corrected datasets. The AUC highlighted the importance of removing batch effects.

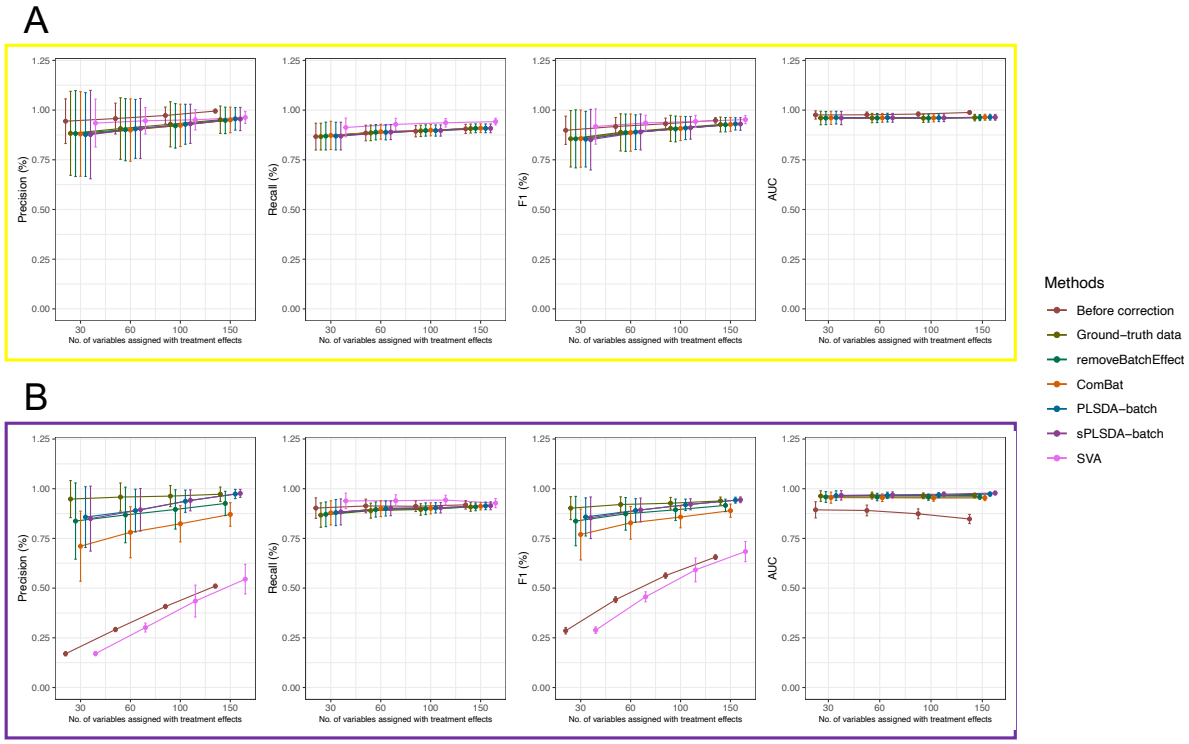

Figure S4. Simulation 4 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different numbers of variables with a true treatment effect (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with four choices of numbers of treatment associated variables $M^{(trt)}$ (x-axis). The description of these plots is detailed in Figure S1. In the balanced design, there was a rise in the accuracy measurements from one-way ANOVA of different datasets before and after batch effect correction as $M^{(trt)}$ increased. When $M^{(trt)} = 150$, the precision of different methods corrected data was similar as the ground-truth data, while in the other scenarios, different measurements of SVA were higher than the other datasets except the original data. In the unbalanced design, the precision and F1 score of PLSDA-batch and sPLSDA-batch were higher than the other methods corrected data, but lower than the ground-truth data. These two measurements became similar as the ground-truth data when $M^{(trt)} = 150$. SVA controlled data (balanced & unbalanced designs) and the original data (unbalanced design) had slightly higher recall than the others. The AUC was similar across different $M^{(trt)}$ choices.

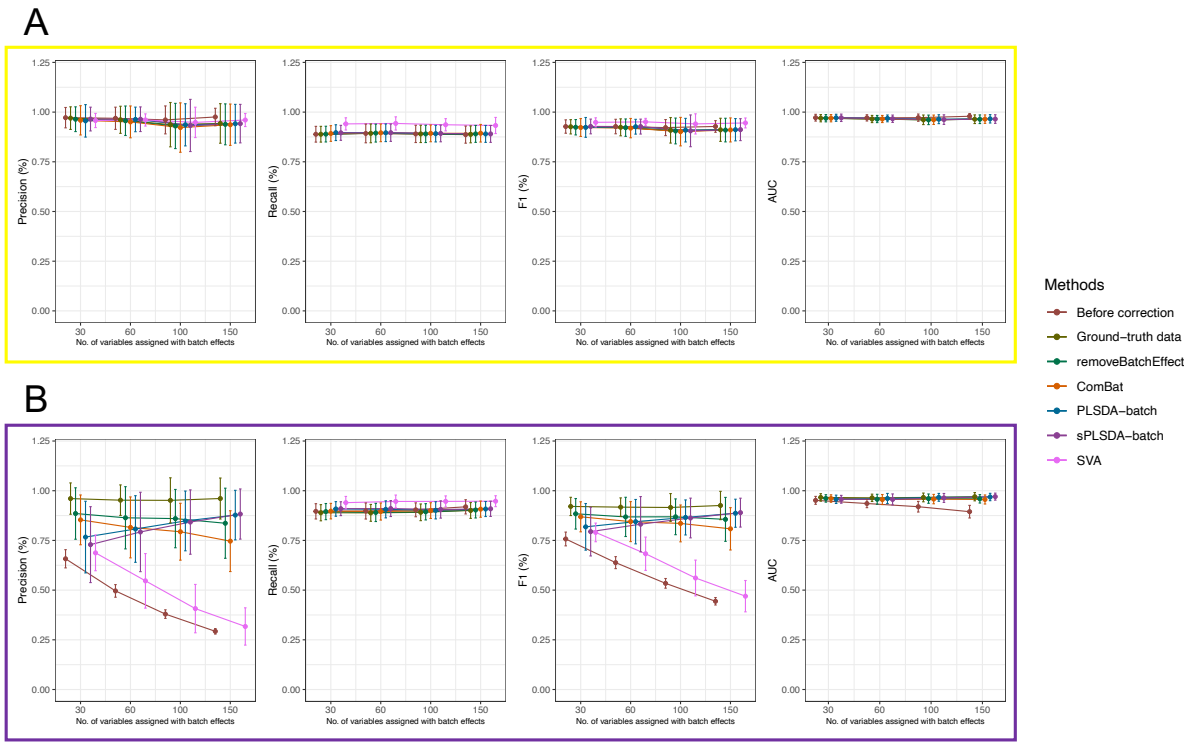

Figure S5. Simulation 5 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different numbers of variables with a true batch effect (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with four choices of numbers of batch associated variables $M^{(batch)}$ (x-axis). The description of these plots is detailed in Figure S1. In the balanced design, the increase of $M^{(batch)}$ did not largely affect the performance of different batch effect correction methods. SVA controlled data had slightly higher recall and F1 score compared to the other corrected data and the ground-truth data, while the other methods performed similarly. In the unbalanced design, as $M^{(batch)}$ increased, the precision and F1 score of PLSDA-batch and sPLSDA-batch corrected data increased while removeBatchEffect and ComBat corrected data, SVA controlled data and the original data decreased. SVA controlled data had slightly higher recall but much lower precision and F1 score compared to the other datasets, as SVA selected far more variables with treatment effects than the selection from the other datasets. The AUC highlighted the importance of removing batch effects.

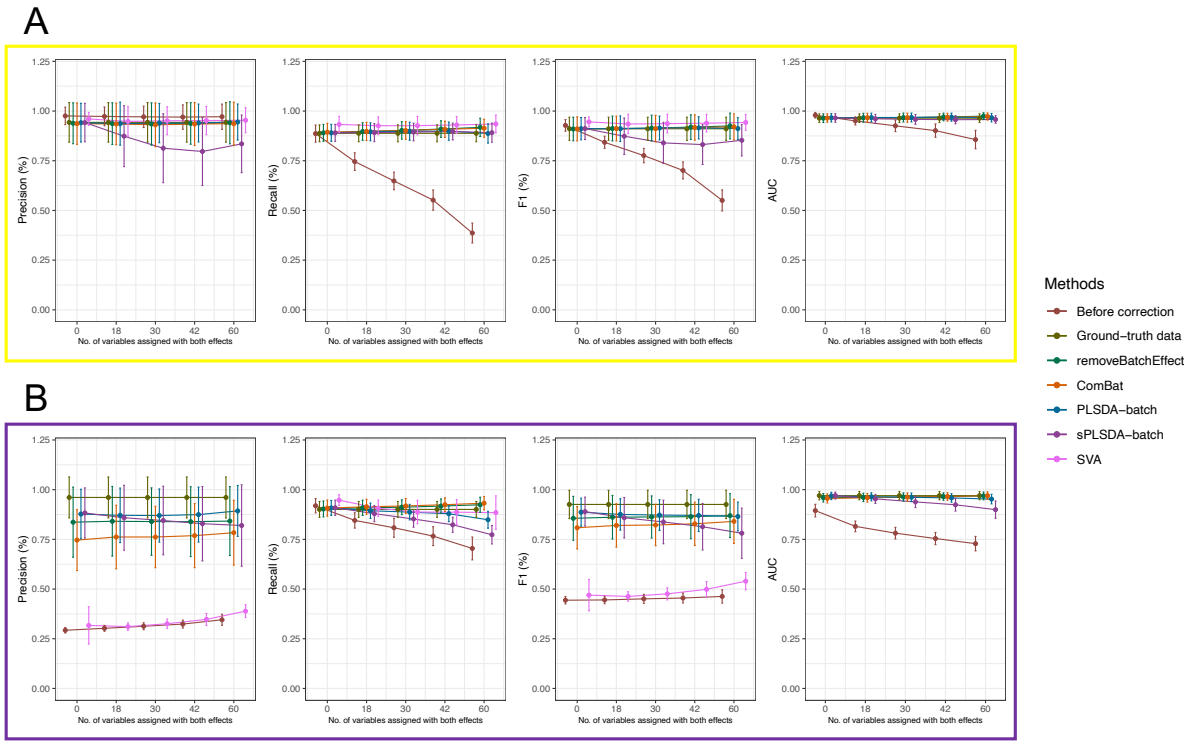

Figure S6. Simulation 6 (two batch groups): summary of accuracy measurements before and after batch effect correction for the data simulated with different numbers of variables with both treatment and batch effects (see Table 1) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with five choices of numbers of relevant variables with both treatment and batch effects $M^{(trt\ \&\ batch)}$ (x-axis). The description of these plots is detailed in Figure S1. In the balanced design, the increase of $M^{(trt\ \&\ batch)}$ did not change the precision of different datasets, except sPLSDA-batch corrected data in which the precision decreased. The recall, F1 score and AUC of the original data decreased dramatically as $M^{(trt\ \&\ batch)}$ increased. SVA performed slightly better than the other methods based on recall and F1 score. In the unbalanced design, all accuracy measurements of sPLSDA-batch corrected data decreased gradually, while the other methods were comparatively stable as $M^{(trt\ \&\ batch)}$ increased. SVA performed the worst among all the methods based on the precision and F1 score. The AUC of both designs highlighted the importance of removing batch effects.

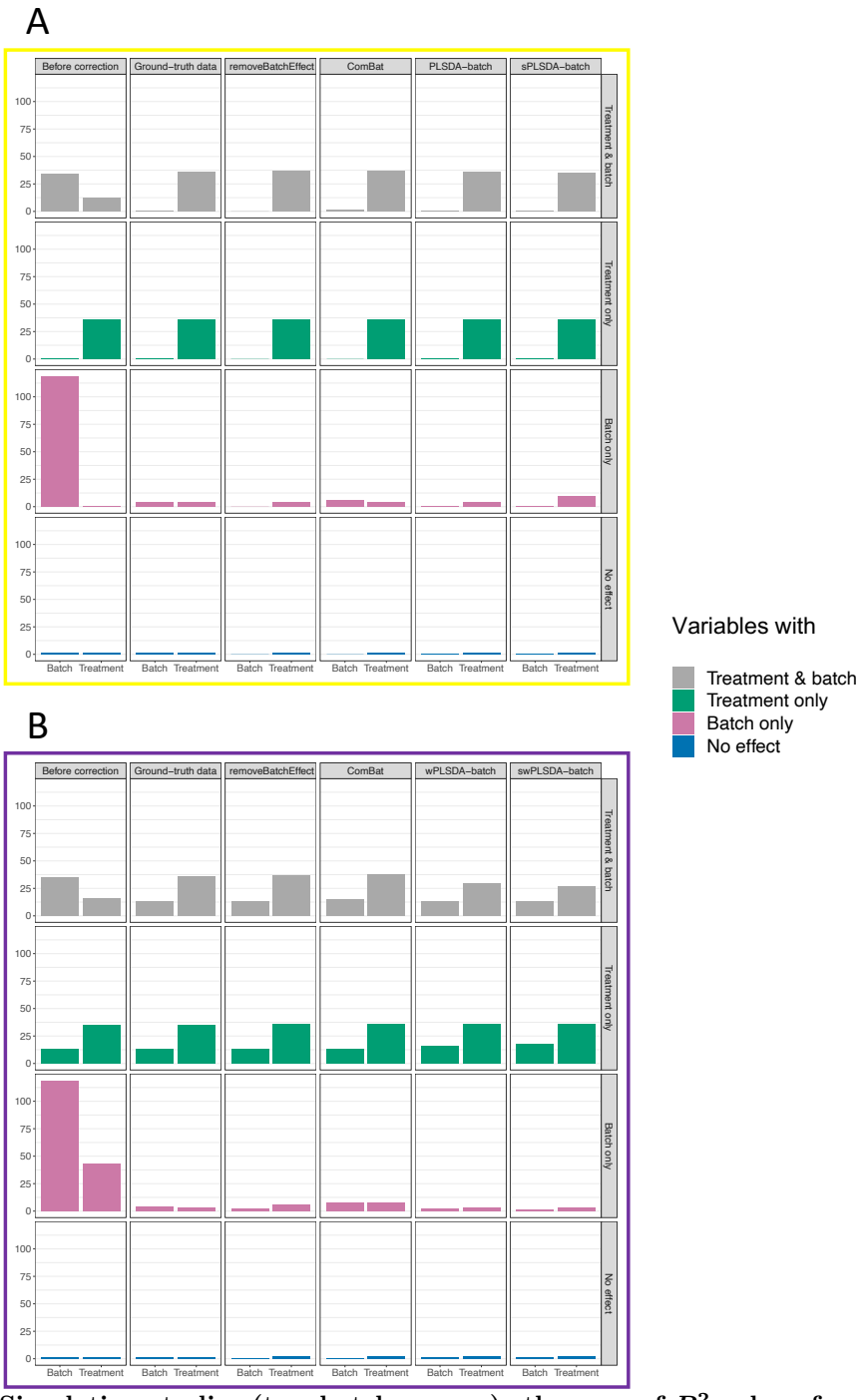

Figure S7. Simulation studies (two batch groups): the sum of R^2 values for each microbial variable before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. Each bar represents the sum of R^2 values for variables simulated with the associated effects (batch or/and treatment effects). Each R^2 value was fitted for each variable from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the effects assigned to each variable. In both designs, ComBat did not remove enough batch variation. For the balanced design, sPLSDA-batch generated slightly spurious treatment variation for the variables with batch effects only. For the unbalanced design, wPLSDA-batch and swPLSDA-batch generated data with less treatment variation for the variables with both treatment and batch effects compared to the ground-truth data.

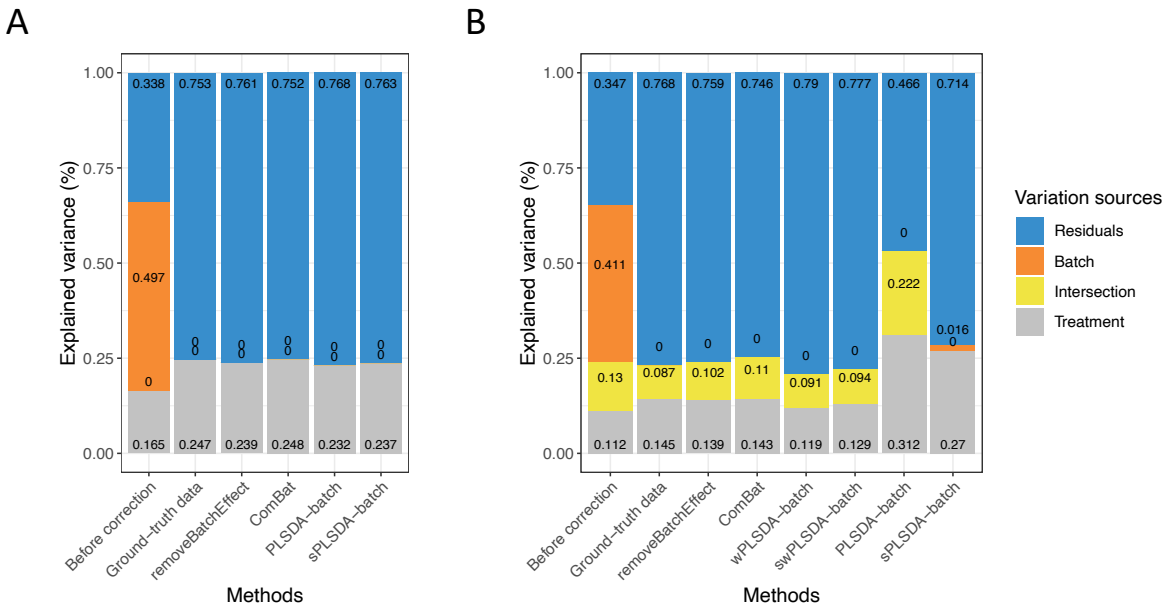

Figure S8. Simulation studies (three batch groups): comparison of explained variance before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. The method pRDA estimated the proportion of variance explained by (from top to bottom) residuals, batch effects, intersection of batch and treatment effects, and treatment effects. All methods performed equally well in removing batch variance for a balanced design, while in an unbalanced design, our weighted variants wPLSDA-batch and swPLSDA-batch performed better than their unweighted counterparts.

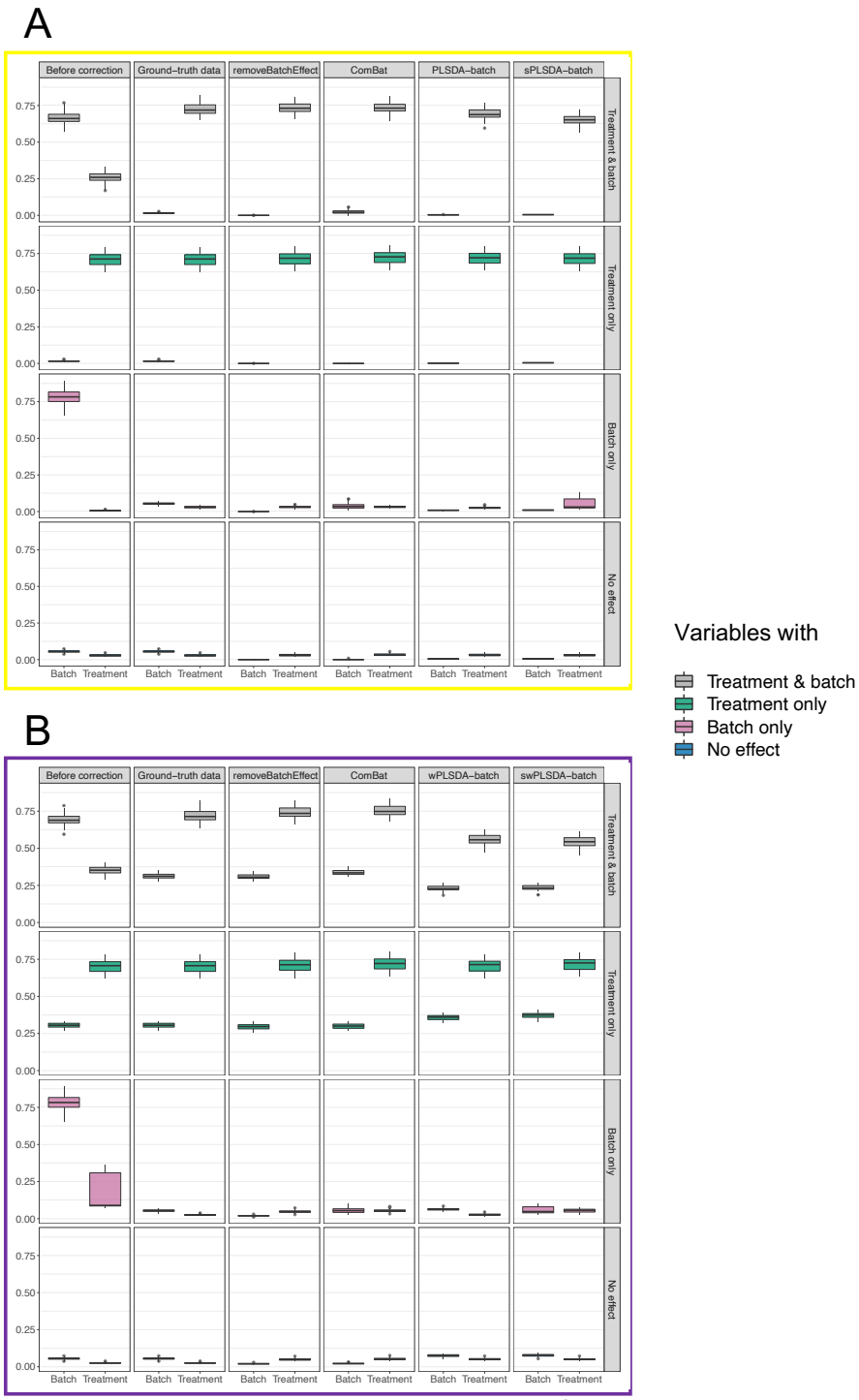

Figure S9. Simulation studies (three batch groups): R^2 values for each microbial variable before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. Each box represents a summary of R^2 values for variables simulated with the associated effects (batch or/and treatment effects). Each R^2 value was fitted for each variable from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the effects assigned to each variable. For the balanced design, ComBat did not remove enough batch variation. sPLSDA-batch generated slightly spurious treatment variation for the variables with batch effects only. For the unbalanced design, wPLSDA-batch and swPLSDA-batch generated data with less treatment variation for the variables with both treatment and batch effects compared to the ground-truth data.

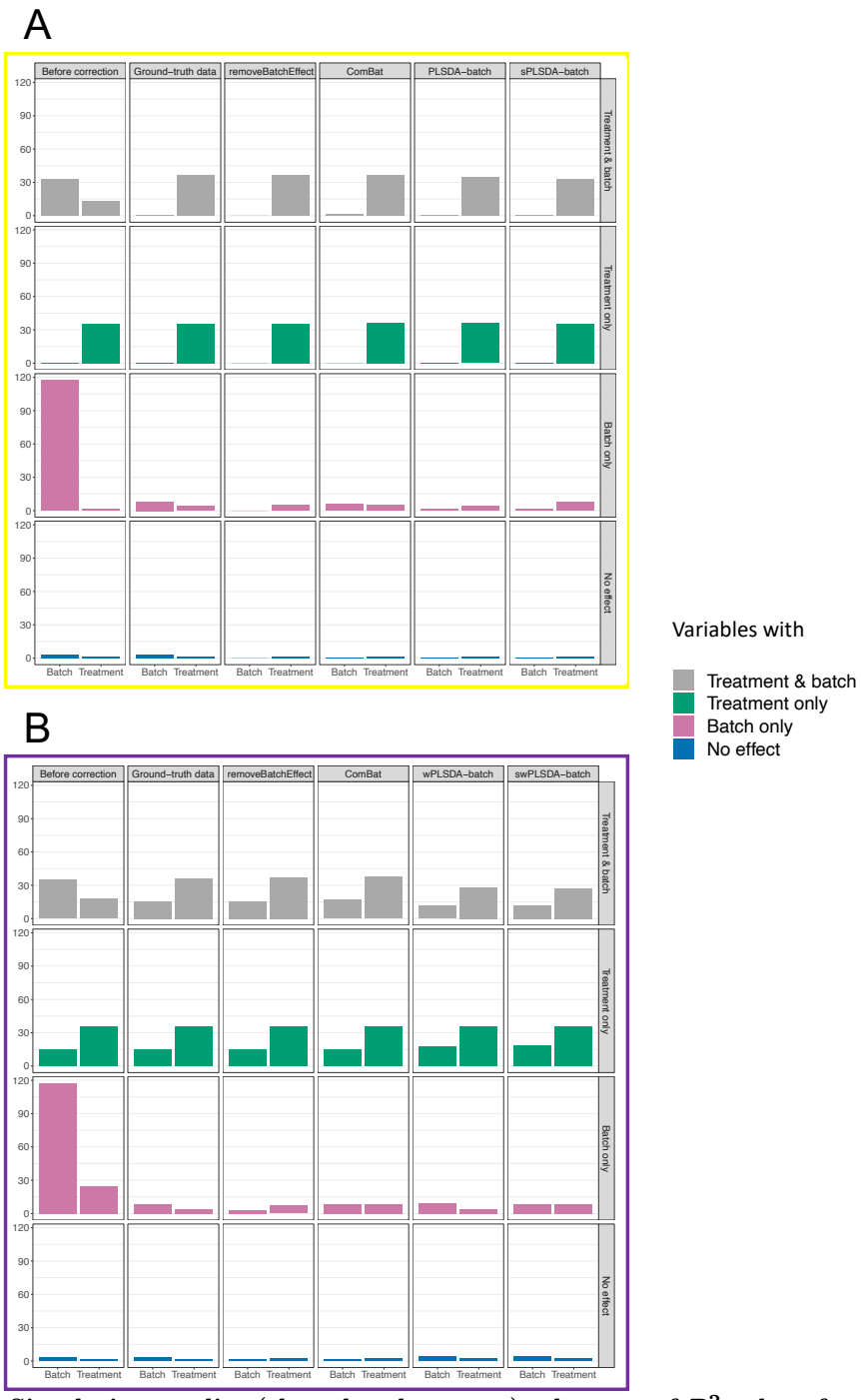

Figure S10. Simulation studies (three batch groups): the sum of R^2 values for each microbial variable before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. Each bar represents the sum of R^2 values for variables simulated with the associated effects (batch or/and treatment effects). Each R^2 value was fitted for each variable from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the effects assigned to each variable. For the balanced design, ComBat did not remove enough batch variation. For the unbalanced design, wPLSDA-batch and swPLSDA-batch generated data with less treatment variation for the variables with both treatment and batch effects compared to the ground-truth data.

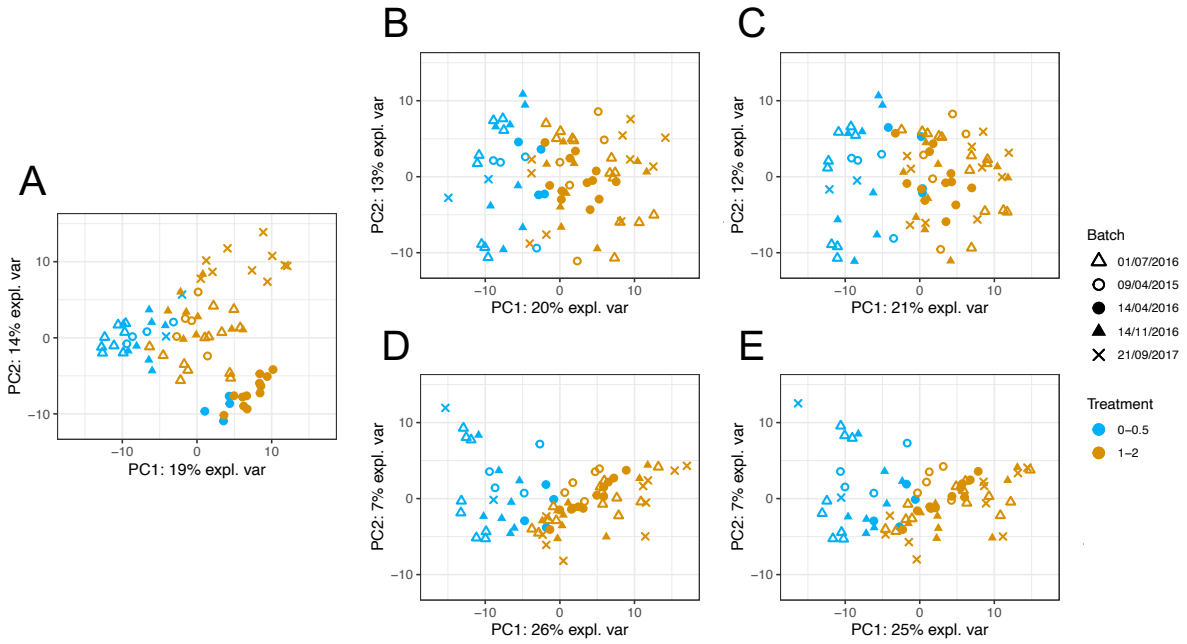

Figure S11. PCA sample plots of the AD data (A) before or after batch effect correction using **(B)** removeBatchEffect, **(C)** ComBat, **(D)** PLSDA-batch and **(E)** sPLSDA-batch. Colours represent the effect of interest (treatment types), and shapes the batch types. The variance explained by the first principal component that separated the different treatment groups was increased in all of the corrected data, with PLSDA-batch resulting in the highest proportion of variance.

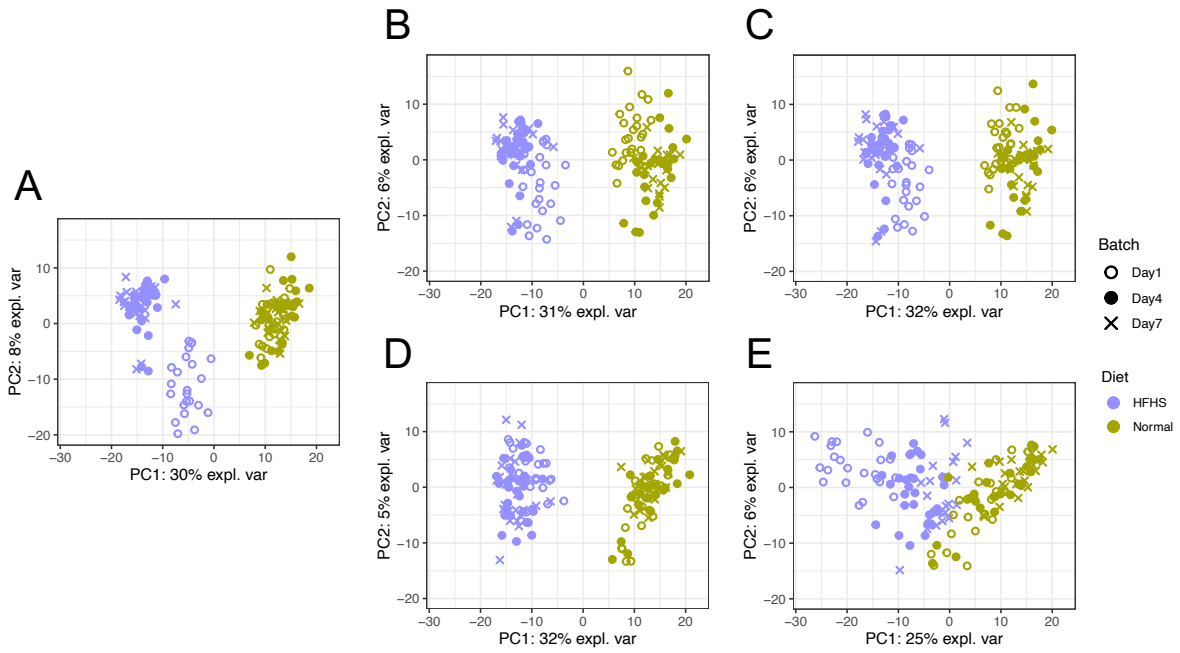

Figure S12. PCA sample plots of the HFHS data (A) before or after batch effect correction using **(B)** removeBatchEffect, **(C)** ComBat, **(D)** PLSDA-batch and **(E)** sPLSDA-batch. Colours represent the effect of interest (diet types), and shapes the batch types. The batch variation between Day 1 and the other days is only shown in the HFHS group before correction. The proportion of variance explained by the first principal component (related to diet effects) before batch effect correction and after was almost the same except after sPLSDA-batch correction, indicating a good preservation of treatment variation.

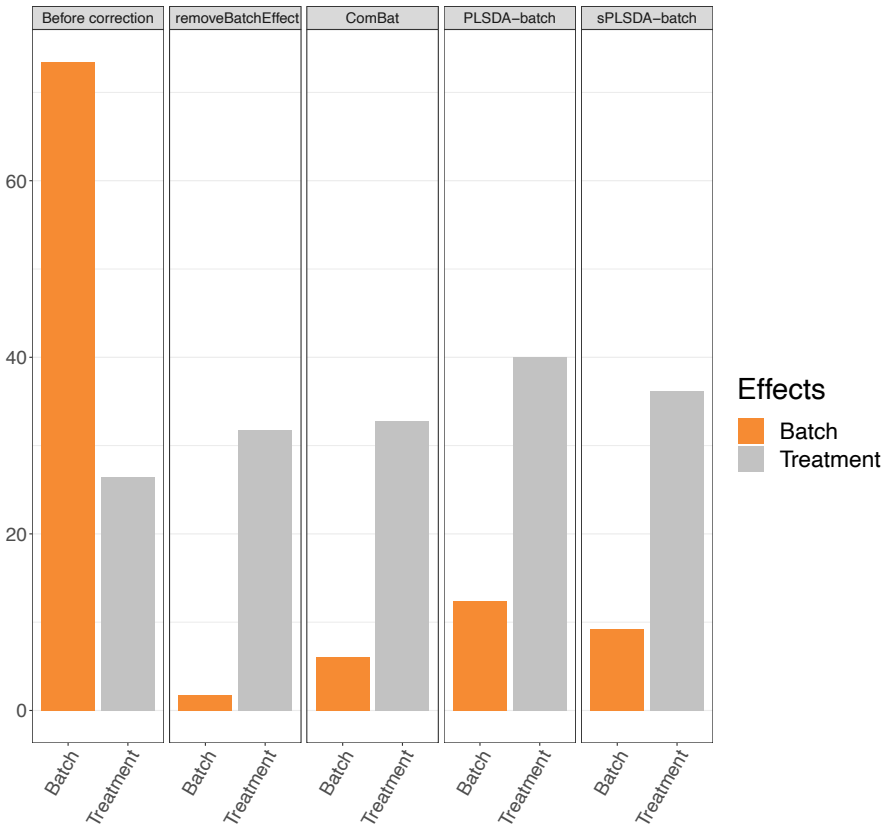

Figure S13. AD study: the sum of R^2 values for each microbial variable before and after batch effect correction. Each bar represents the sum of R^2 values fitted for variables from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the fitted effects in ANOVA. removeBatchEffect and ComBat removed slightly more batch variance, but preserved less treatment variance than our proposed PLSDA-batch and sPLSDA-batch.

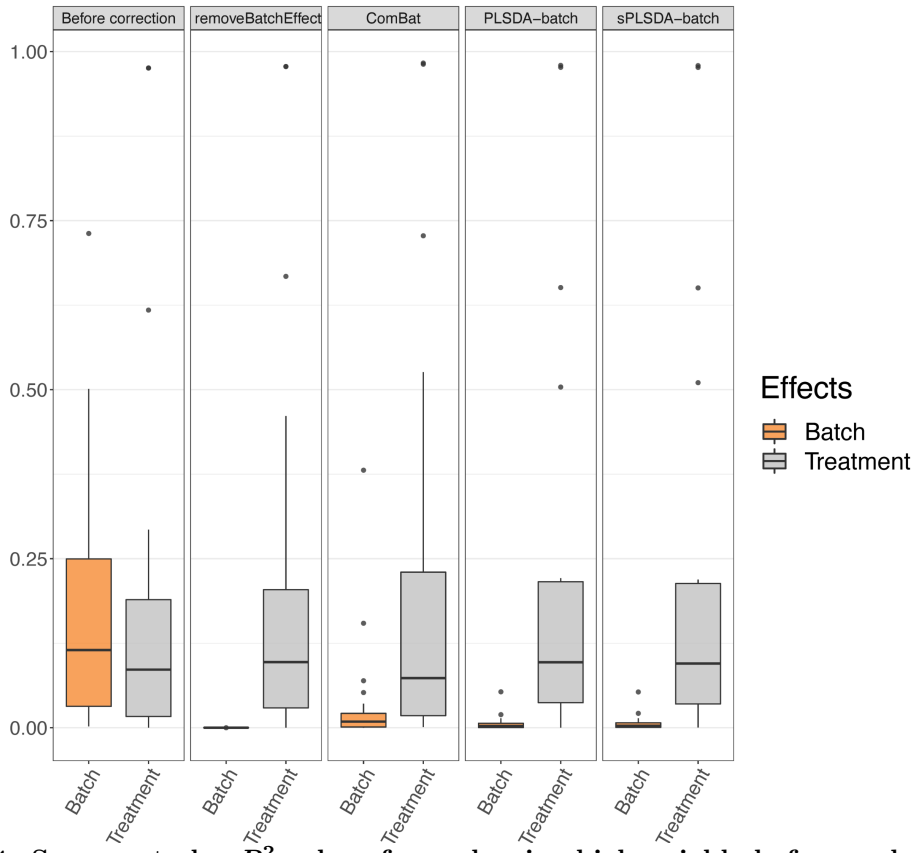

Figure S14. Sponge study: R^2 values for each microbial variable before and after batch effect correction. Each box represents a summary of R^2 values fitted for variables from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the fitted effects in ANOVA. Combat corrected data included variables with a larger proportion of batch variance than the other methods.

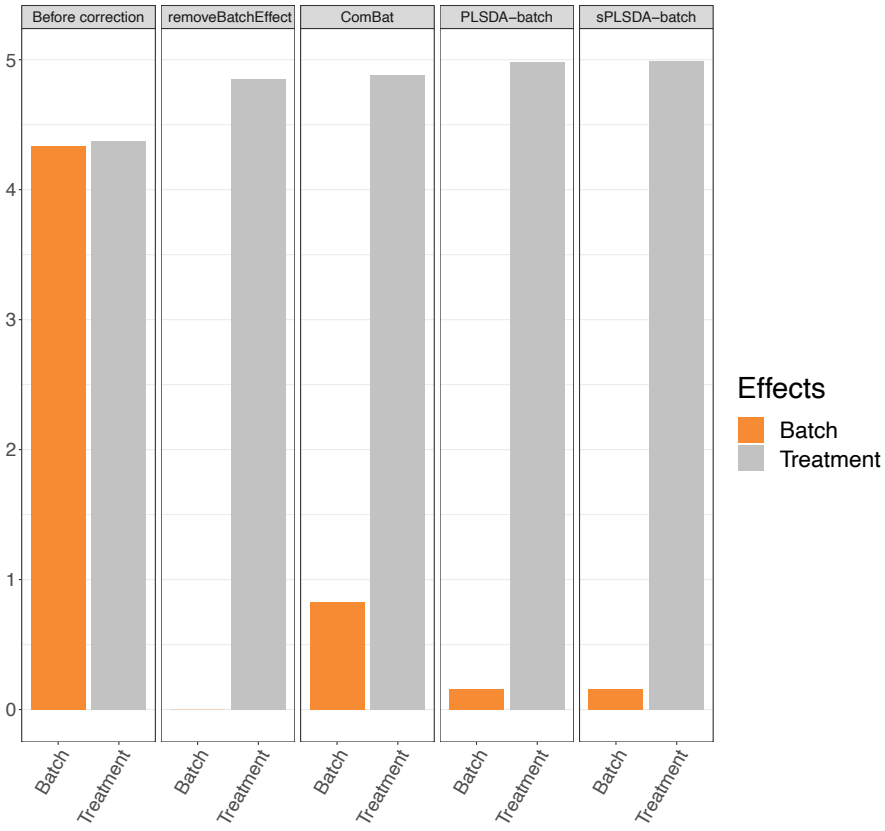

Figure S15. Sponge study: the sum of R^2 values for each microbial variable before and after batch effect correction. Each bar represents the sum of R^2 values fitted for variables from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the fitted effects in ANOVA. ComBat did not remove enough batch variation. removeBatchEffect removed slightly more batch variance, but preserved less treatment variance than our proposed PLSDA-batch and sPLSDA-batch.

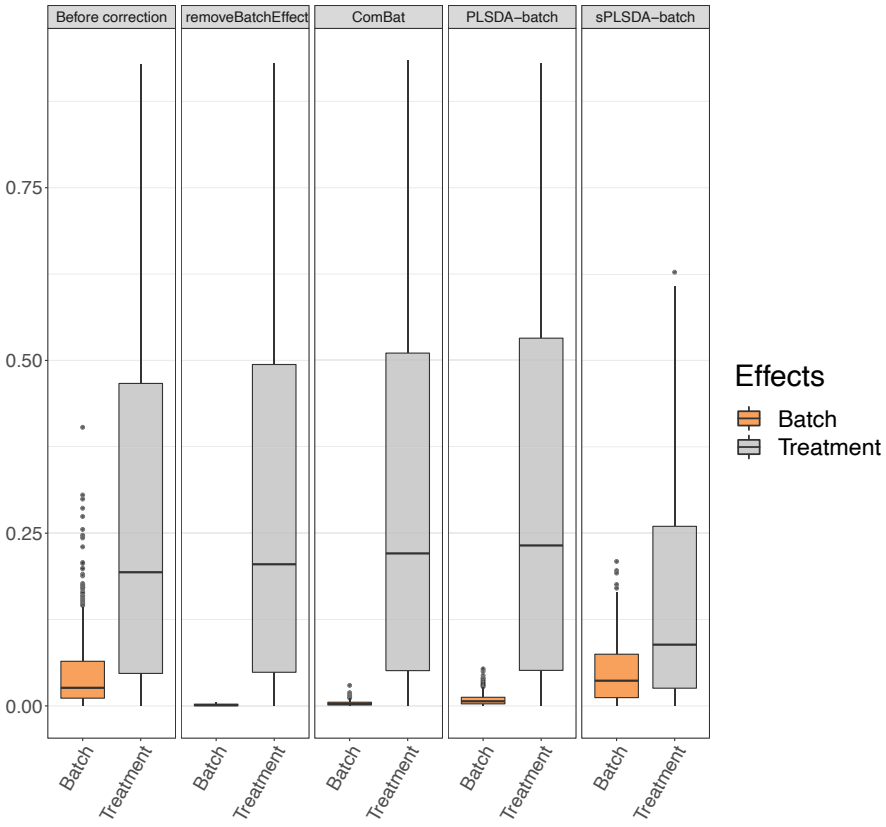

Figure S16. HFHS study: R^2 values for each microbial variable before and after batch effect correction. Each box represents a summary of R^2 values fitted for variables from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the fitted effects in ANOVA. PLSDA-batch is more appropriate than sPLSDA-batch in the situation with weak batch effects.

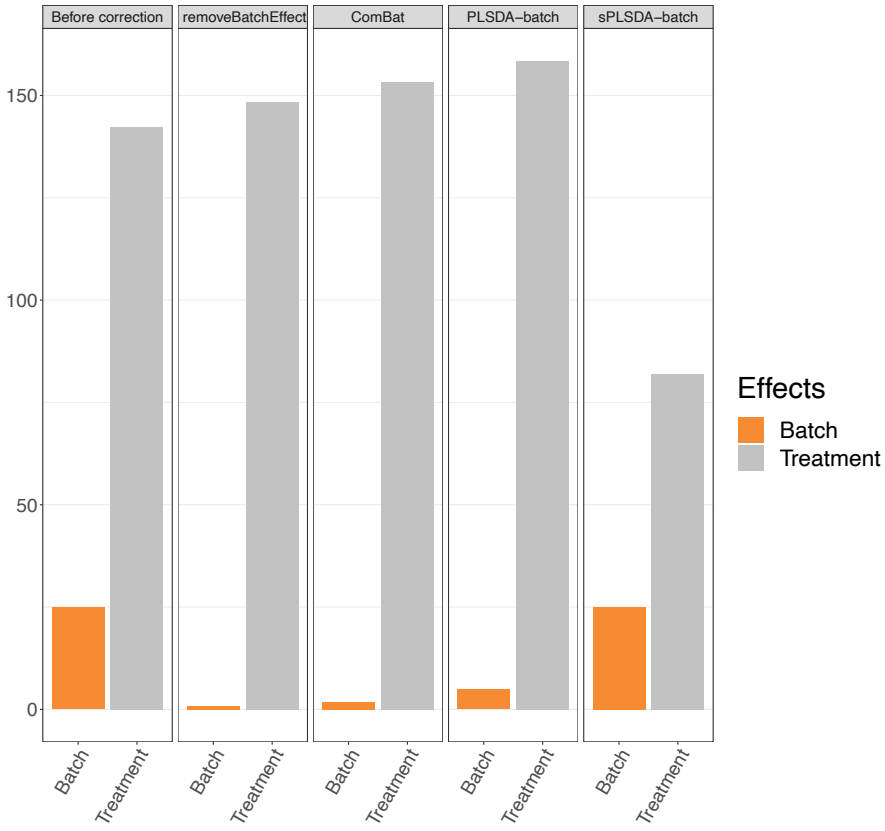

Figure S17. HFHS study: the sum of R^2 values for each microbial variable before and after batch effect correction. Each bar represents the sum of R^2 values fitted for variables from a one-way ANOVA with a treatment effect or batch effect as covariate (x-axis). Colours indicate the fitted effects in ANOVA. PLSDA-batch is more appropriate than sPLSDA-batch in the situation with weak batch effects.

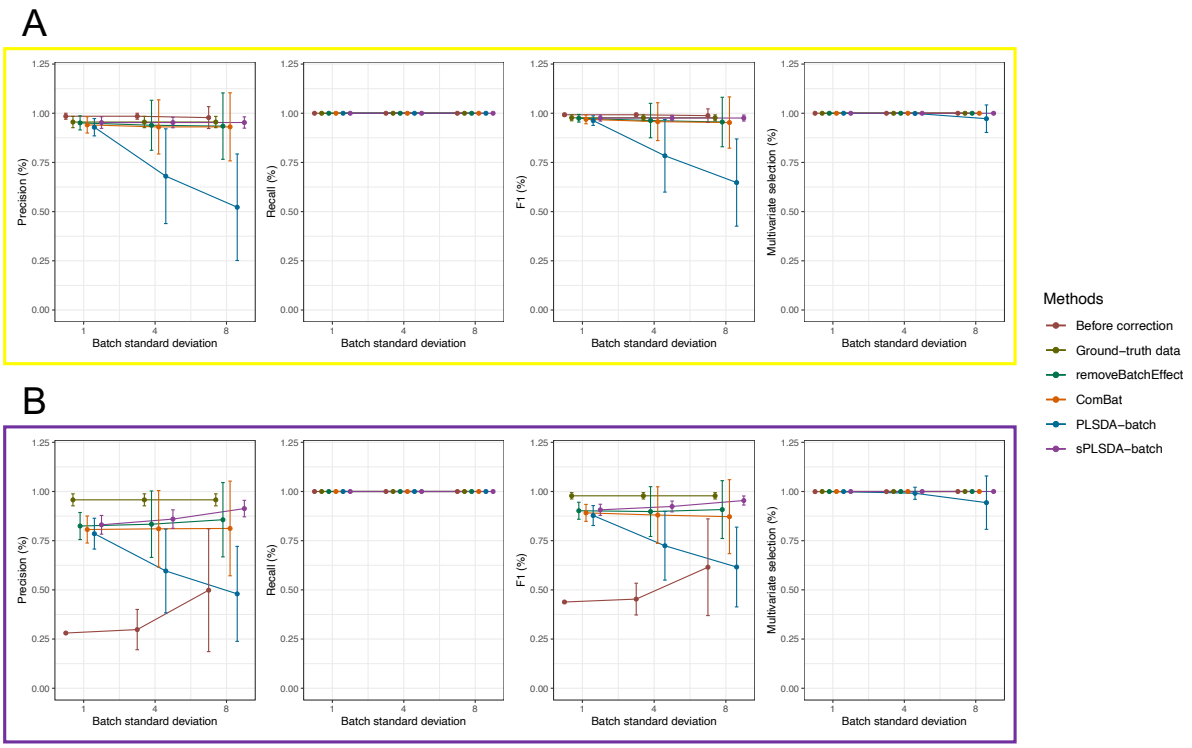

Figure S18. Simulation 1 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different batch effect variability (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Batch effects were generated with three choices of variability among samples $\sigma'_{(batch)}$ (x-axis). The proportion of correctly identified microbial variables with a true treatment effect was assessed with Precision, Recall, F1 score and Multivariate selection score using one-way ANOVA or sPLSDA. Each point was averaged over 50 repeatedly simulated data, with error bars indicating estimated sample standard deviations. As $\sigma'_{(batch)}$ increased, the precision of corrected data from PLSDA-batch dramatically decreased while with sPLSDA-batch slightly increased in both cases of balanced and unbalanced designs. The standard deviation of precision calculated from removeBatchEffect and ComBat corrected data increased with $\sigma'_{(batch)}$. sPLSDA-batch corrected data slightly outperformed the other corrected data with a higher precision or/and a smaller standard deviation of the precision in both designs. The resulting recall and multivariate selection score were similar among different data. F1 score calculated from the precision and recall therefore displayed the same information as the precision.

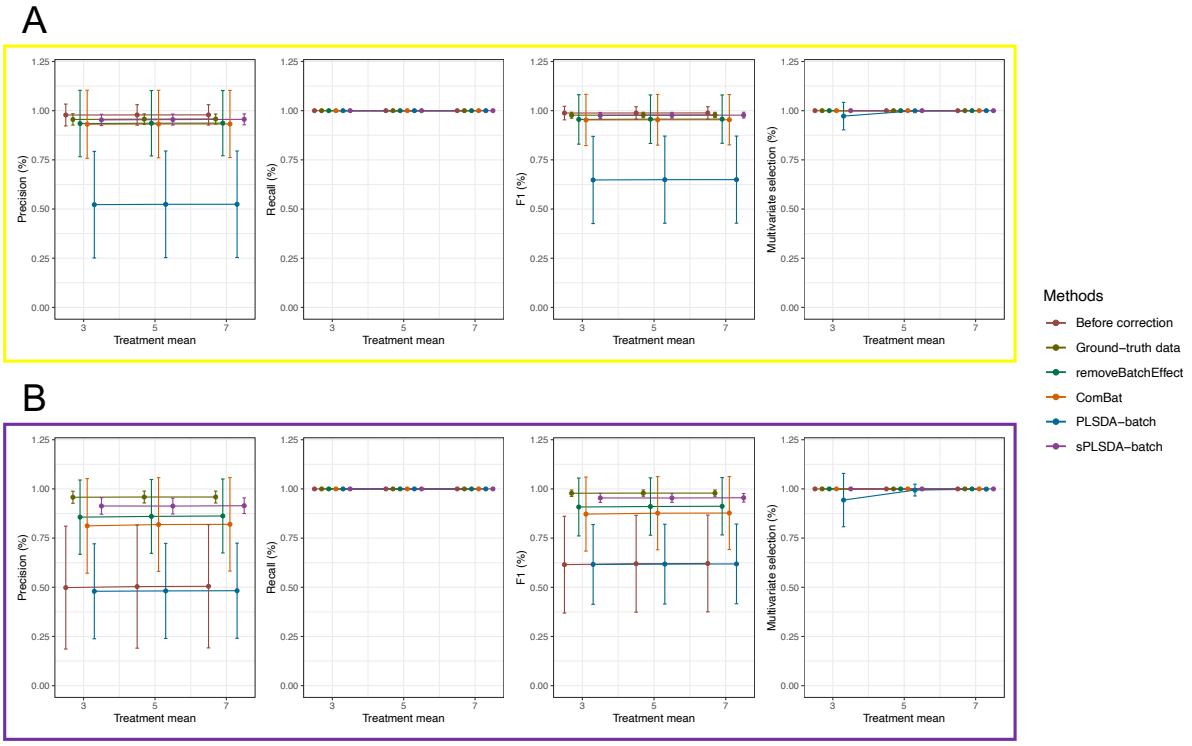

Figure S19. Simulation 2 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different sizes of treatment effects (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Treatment effects were generated with three choices of sizes $\mu_{(trt)}$ (x-axis). The description of these plots is detailed in Figure S18. The change of $\mu_{(trt)}$ did not affect the performance of different batch effect correction methods.

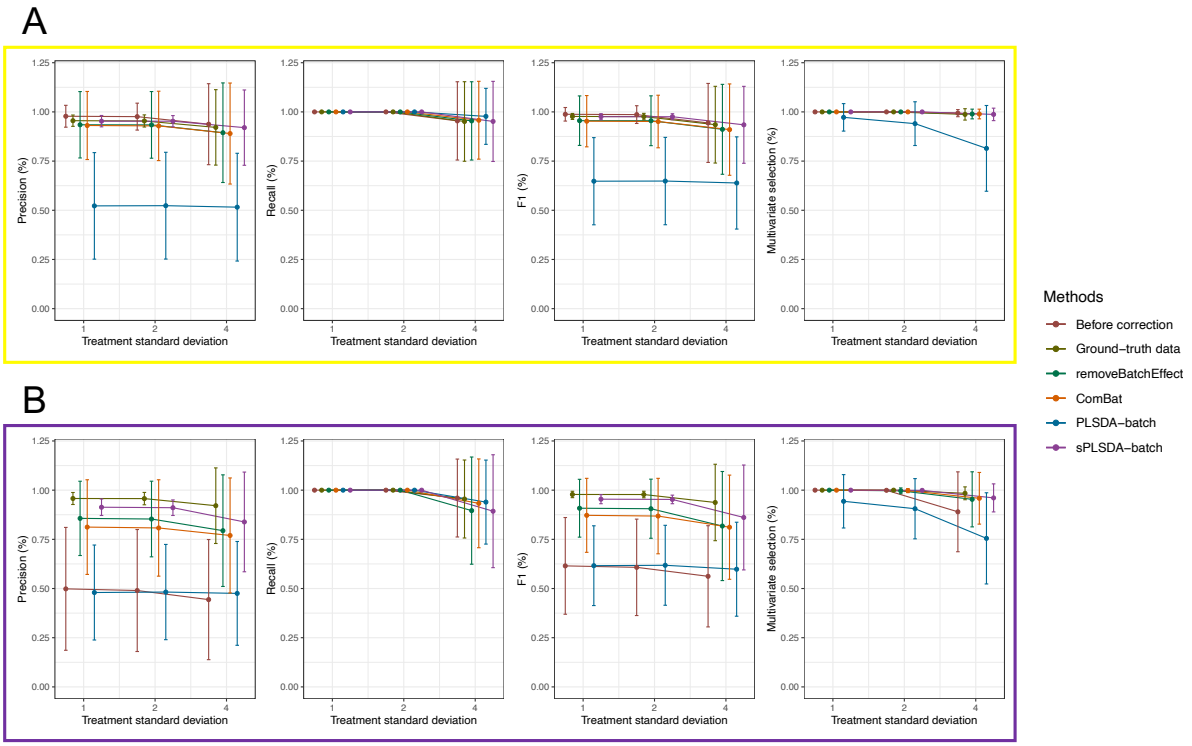

Figure S20. Simulation 3 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different treatment effect variability (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Treatment effects were generated with three choices of variability among samples $\sigma'_{(trt)}$ (x-axis). The description of these plots is detailed in Figure S18. All accuracy measurements of different batch effect corrected data slightly decreased and their standard deviations increased when the $\sigma'_{(trt)}$ is larger than 2.

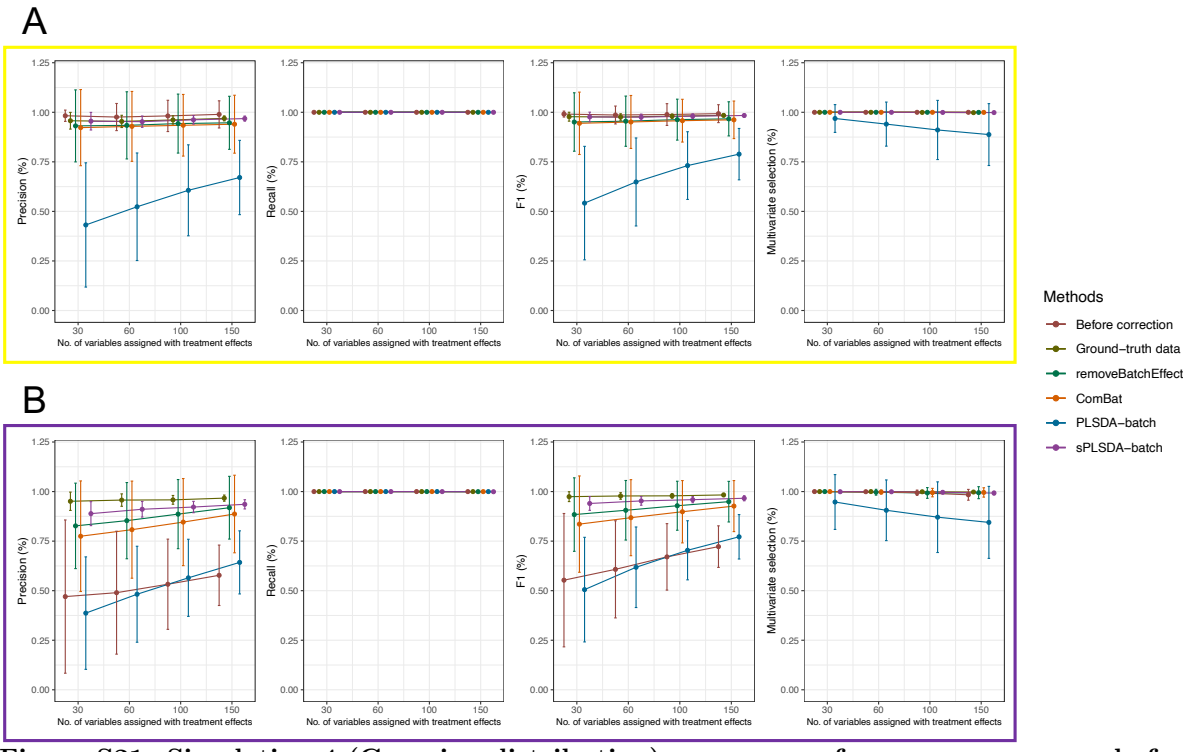

Figure S21. Simulation 4 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different numbers of variables with a true treatment effect (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with four choices of numbers of treatment associated variables $M^{(trt)}$ (x-axis). The description of these plots is detailed in Figure S18. The precision of corrected data from different methods slightly increased because of the increase of $M^{(trt)}$ for the unbalanced design, while similar among different $M^{(trt)}$ for the balanced deign with an exception of PLSDA-batch corrected data. The multivariate selection scores of different corrected data were similar, except PLSDA-batch corrected data whose multivariate selection score decreased.

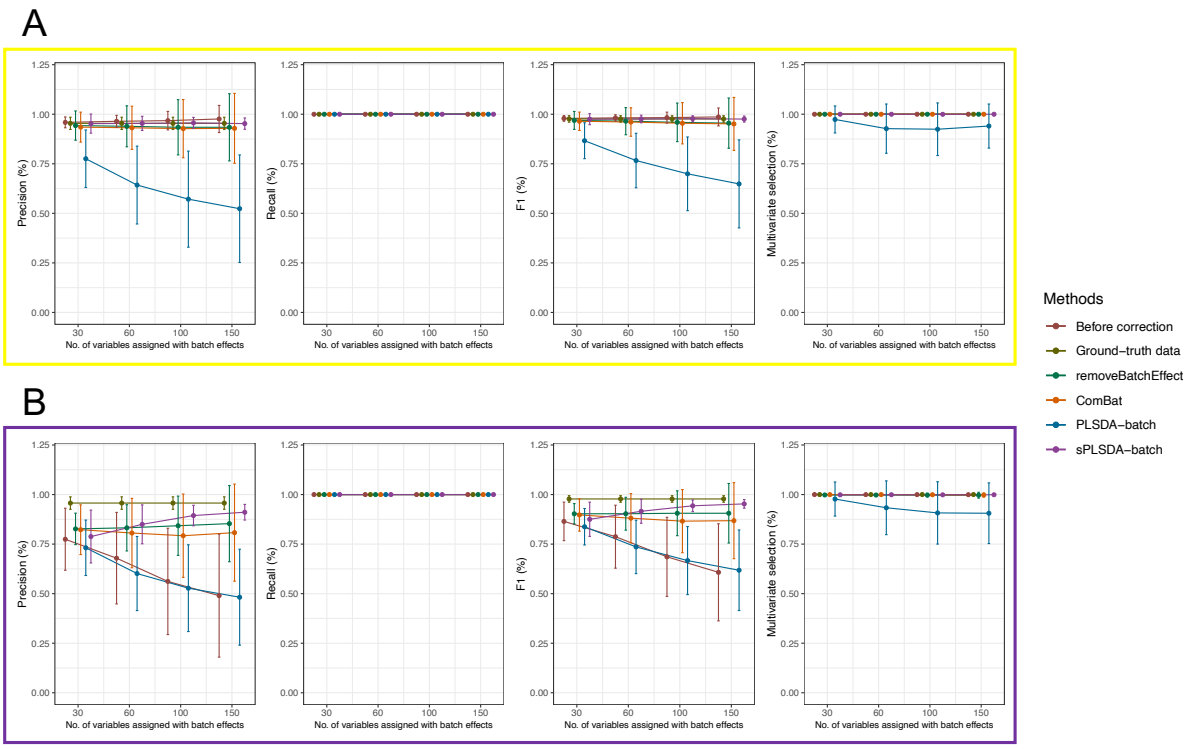

Figure S22. Simulation 5 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different numbers of variables with a true batch effect (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with four choices of numbers of batch associated variables $M^{(batch)}$ (x-axis). The description of these plots is detailed in Figure S18. The increase of $M^{(batch)}$ resulted in an increase of the precision of data corrected with removeBatchEffect, ComBat and sPLSDA-batch, while a decrease with PLSDA-batch for the unbalanced design. The precision of all corrected data and with different $M^{(batch)}$ were similar for the balanced deign except PLSDA-batch corrected data.

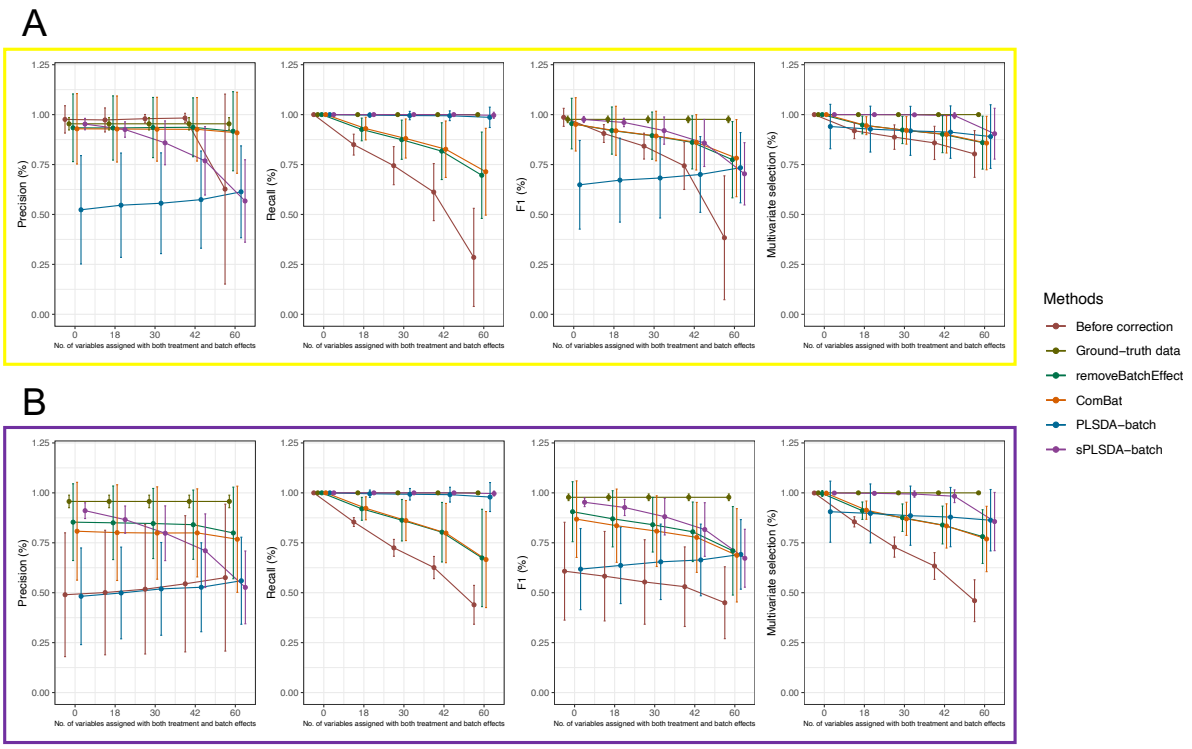

Figure S23. Simulation 6 (Gaussian distribution): summary of accuracy measures before and after batch effect correction for the data simulated with different numbers of variables with both treatment and batch effects (see Table S2) with **(A)** balanced and **(B)** unbalanced batch × treatment designs. Simulated data were generated with five choices of numbers of relevant variables with both treatment and batch effects $M^{(trt\ \&\ batch)}$ (x-axis). The description of these plots is detailed in Figure S18. When $M^{(trt\ \&\ batch)}$ was larger than 30 (a half of $M^{(trt)}$), the precision of data corrected with sPLSDA-batch was lower compared to removeBatchEffect and ComBat, but the recall and multivariate selection score were higher regardless of different $M^{(trt\ \&\ batch)}$.

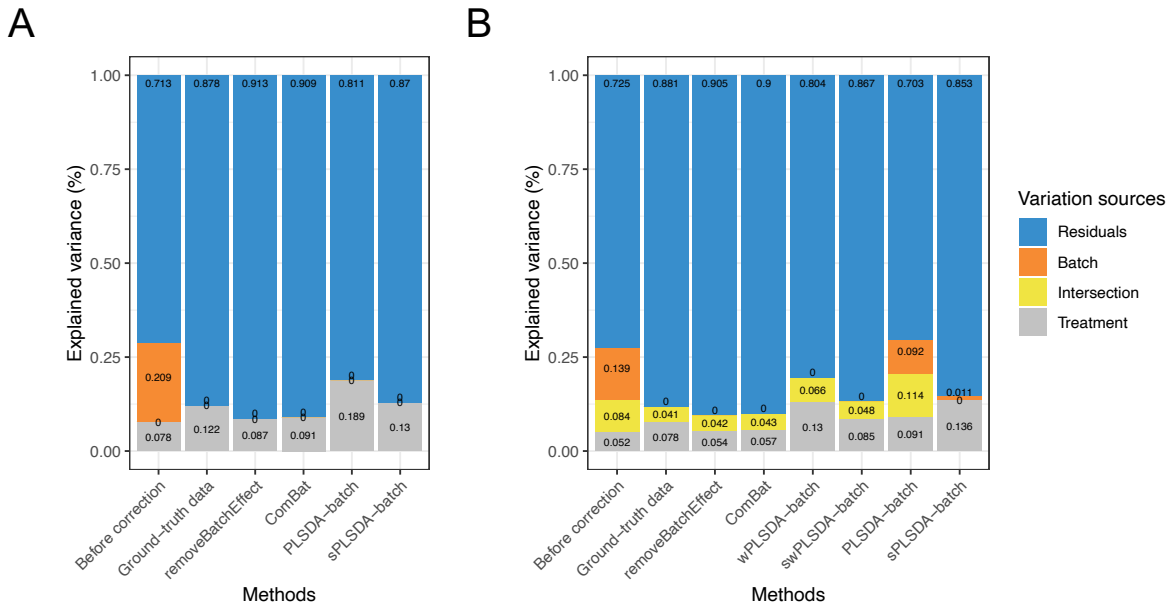

Figure S24. Simulation studies (Gaussian distribution): comparison of explained variance before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. The partitioned variance explained by (from top to bottom) residuals, batch effects, intersection of batch and treatment effects, and treatment effects was estimated with pRDA. sPLSDA-batch and swPLSDA-batch performed best in correcting for batch effects as the explained variance was most similar to the ground-truth data that included no batch effect.

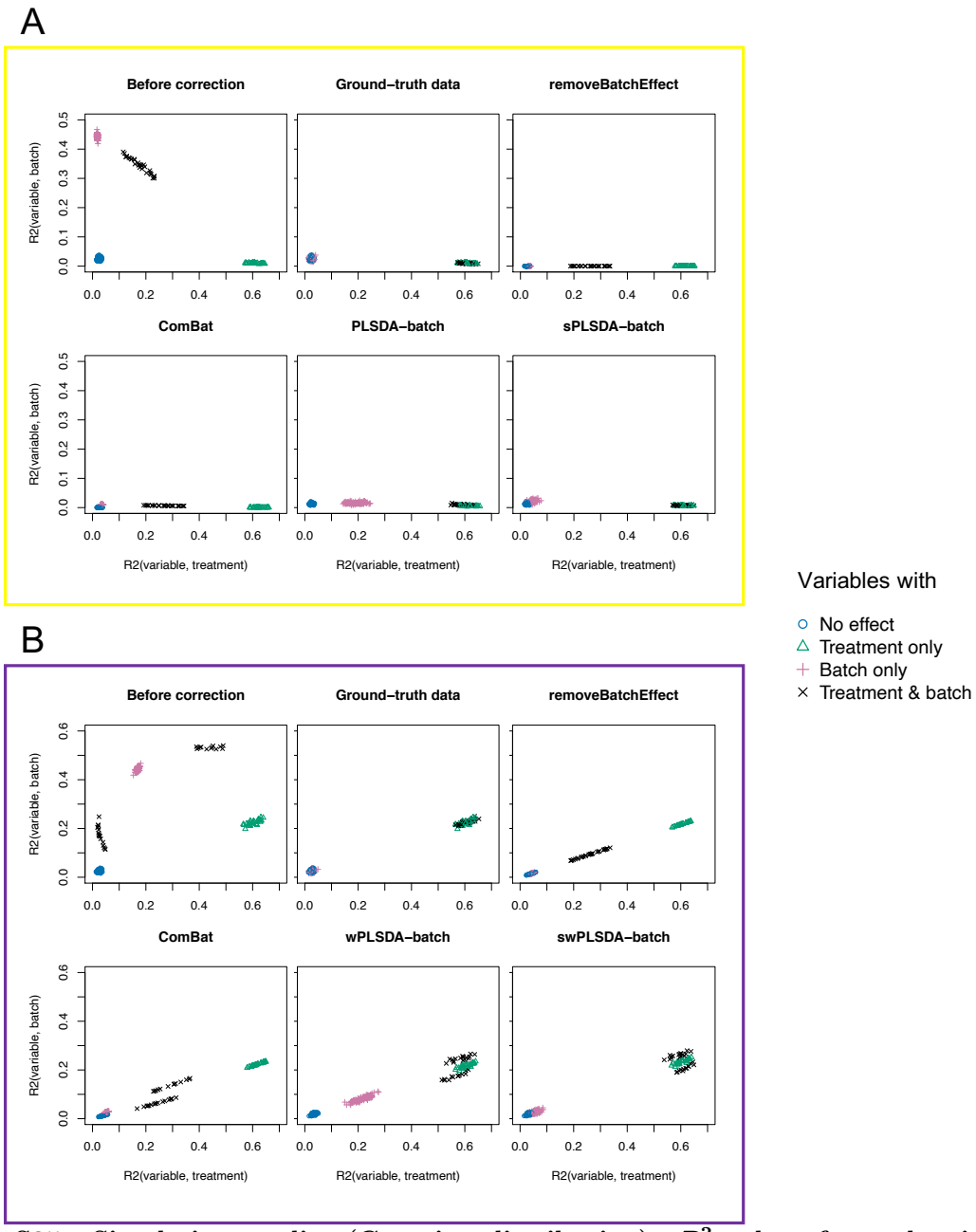

Figure S25. Simulation studies (Gaussian distribution): R^2 values for each microbial variable before and after batch effect correction for **(A)** balanced and **(B)** unbalanced batch × treatment designs. Each point represents one variable with respect to its fitted R^2 from a one-way ANOVA with a treatment effect (x-axis) or batch effect (y-axis) as covariate. Colours and shapes indicate the associated effects (batch or/and treatment effects) for each variable. RemoveBatchEffect and ComBat did not preserve enough treatment variation for variables with both treatment and batch effects, while PLSDA-batch and wPLSDA-batch generated spurious treatment variation for variables with batch effect only. sPLSDA-batch and swPLSDA-batch corrected data are the most similar to the ground-truth data that include no batch effects.

References

1. Ritchie ME, Phipson B, Wu D, *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* 2015;**43**:e47–e47.

2. Wu J, Peters BA, Dominianni C, *et al.* Cigarette smoking and the oral microbiome in a large study of American adults. *The ISME journal* 2016;**10**:2435–2446.

3. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007;**8**:118–127.

4. Hong By, Paulson JN, Stine OC, *et al.* Meta-analysis of the lung microbiota in pulmonary tuberculosis. *tuberculosis* 2018;**109**:102–108.

5. Kubinski R, Djamen-Kepaou JY, Zhanabaev T, *et al.* Benchmark of data processing methods and machine learning models for gut microbiome-based diagnosis of inflammatory bowel disease. *bioRxiv* 2021;.

6. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS genetics* 2007;**3**:e161.

7. Ho EXP, Cheung CMG, Sim S, *et al.* Human pharyngeal microbiota in age-related macular degeneration. *PloS One* 2018;**13**:e0201768.

8. Thompson KJ, Ingle JN, Tang X, *et al.* A comprehensive analysis of breast cancer microbiota and host gene expression. *PloS One* 2017;**12**:e0188873.

9. Holmes S, Huber W. *Modern statistics for modern biology.* Cambridge University Press, 2018.

10. Singh A, Shannon CP, Gautier B, *et al.* DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics* 2019;**35**:3055–3062.

11. Weiss S, Xu ZZ, Peddada S, *et al.* Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 2017;**5**:1–18.