# Supplementary Materials

## Supplementary Methods

|  |
| --- |
| Figure 1. **1.** Initial Simulation Pipeline **A.** Simulated dataset with normally distributed random numbers **B.** Simulated class effects **C.** Simulated batch effects **D.** Introduction of 50% missing values **E.** Imputation of missing values. 3 sub-conditions for the mean-based averaging imputation method: Global mean imputation (M1), Same batch mean imputation (M2), Opposite batch mean imputation (M3) **F.** Perform batch correction. 4 types of BECAs are used: ComBat, BMC, Harman, SVA **G.** Evaluate the outcome based on the estimation of remnant batch effects and imputation accuracy  **2.** Proteomics Simulation **A.** Renal Control dataset (RCC) with only 1 true sample class is used in this proteomics simulation. Data pre-processing of RCC such that the proteomics dataset has mean=5 and standard deviation (sd) =1 (values similar to our initial simulation), while maintaining original RCC distribution **B.** No class effects simulated **C.** Amplify batch effects for the first batch **D.** Introduction of 50% missing values **E.** Imputation of missing values. 3 sub-conditions for the mean-based averaging imputation method: Global mean imputation (M1), Same batch mean imputation (M2), Opposite batch mean imputation (M3) **F.** Batch correction with ComBat. **G.** Evaluation of outcome based on the estimation of remnant batch effects and imputation accuracy  **3.** Genomics Simulation **A.** A combination of ER+ HER-2 normal breast cancer RNA dataset from 2 cohorts: GDS4056 & GDS4057 is used in this genomics simulation such that genomics dataset only comprises 1 true sample class: ER+. Data pre-processing of GDS4056/4057 such that the genomics dataset has mean=5 and standard deviation (sd) =1(values similar to our initial simulation), while maintaining original GDS4056/4057 distribution **B.** Class effects is created only when analysing power **C.** Amplify batch effects for the first batch **D.** Introduction of 50% missing values **E.** Imputation of missing values. 3 sub-conditions for the mean-based averaging imputation method: Global mean imputation (M1), Same batch mean imputation (M2), Opposite batch mean imputation (M3) **F.** Batch correction with ComBat. **G.** Evaluation of outcome based on the estimation of remnant batch effects and imputation accuracy |

## Supplementary Figures

|  |
| --- |
| A. |
| B. |

**Figure S1.** RMSE for Initial Simulations with **A.** multiplicative only and **B.** additive only batch effects have no strong differences with results with mixed batch effects (Additive + Multiplicative).

|  |
| --- |
| A. |
| B. |

**Figure S2.** gPCA delta results for Initial Simulations with **A.** multiplicative only and **B.** additive only batch effects have no strong differences with results with mixed batch effects (Additive + Multiplicative).

|  |
| --- |
| A. |
| B. |

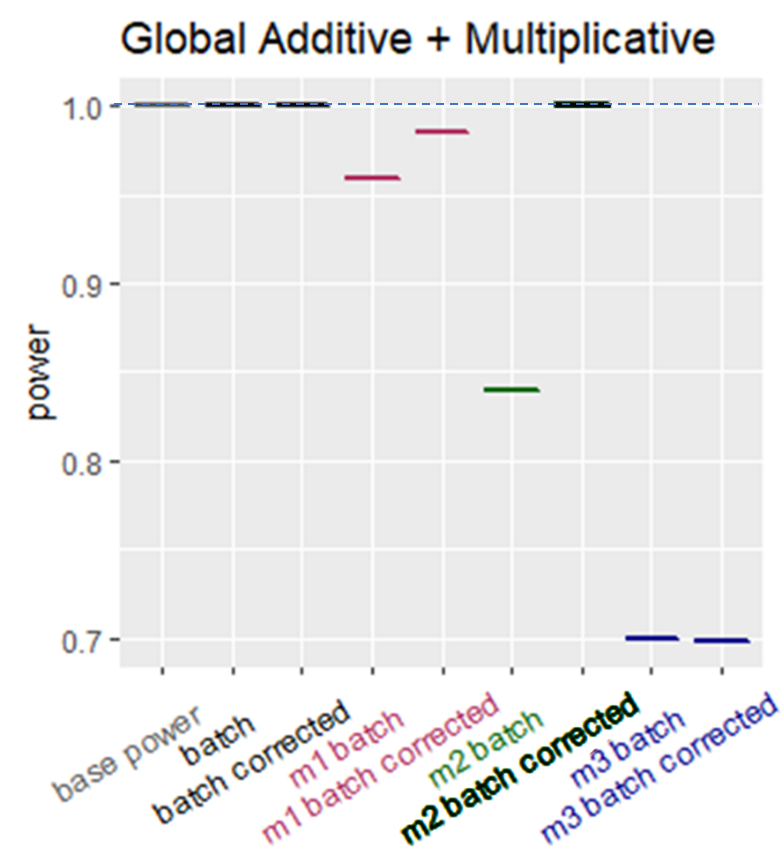
**Figure S3.** Power for Initial Simulations with **A.** multiplicative only and **B.** additive only batch effects show that BMC and SVA do not work well in these scenarios either.

|  |
| --- |
| A. |
| B. |

**Figure S4. A.** RMSE, **B.** gPCA results for Initial Simulation with 100x repeat have no strong differences with results for 10x repeat.

|  |
| --- |
| **1. Proteomics Simulation** |
| **2. Genomics Simulation** |

**Figure S5.** PCA Scatterplots (which includes both pre batch corrected and post batch corrected data) for **1.** Proteomics Simulation and **2.** Genomics Simulation showed that despite reporting higher gPCA levels for M2, samples appear well-mixed, with no apparent batch effects for all imputation strategies (M1 to M3), given the first two principal components (PC1 and PC2).

****

**Figure S6.** Only ComBat is used for evaluation of power for Genomics Simulation based on statistical feature selection. Higher values indicate better performance (higher recall of correct features).

|  |
| --- |
| A. |
| B. |
| C. |

**Figure S7.** t-statistics distribution for Initial Simulation using BECAs such as **A.** BMC, **B.** Harman, **C.** SVA also show that all imputation strategies (M1 to M3) suffer from a reduction in effect size.

|  |
| --- |
| A. Numerator |
| B. Denominator |

**Figure S8.** t-statistics split into **A.** numerator (delta means) and **B.** denominator (standard error of the mean) for Initial Simulation.

|  |  |  |
| --- | --- | --- |
| A. | B. | C. |
| D. | | |

**Figure S9. A.** RMSE, **B.** gPCA delta, **C.** Power, **D.** t-statistics results for imputation strategies (M1 to M3), including M2.1 (imputation based on same class and batch), show that while M2.1 yield t-statistics more similar to original data, it did not outperform M2 in terms of RMSE, gPCA and power.

|  |
| --- |
| A. Proteomics Simulation |
| B. Genomics Simulation |

**Figure S10.** Sample Boxplots for **A.** Proteomics Simulation and **B.** Genomics Simulation although batch effects appear to be “mitigated” in the PCA scatterplots, M1 and M3 result in increased noise (i.e. larger interquartile range) in the data.

|  |  |
| --- | --- |
| A. | B. |

**Figure S11. A.** Power and **B.** Recall for reduced genomics data (20x20 matrix) show that after reducing sample size of genomics data, M2 batch corrected no longer performs as well as batch corrected