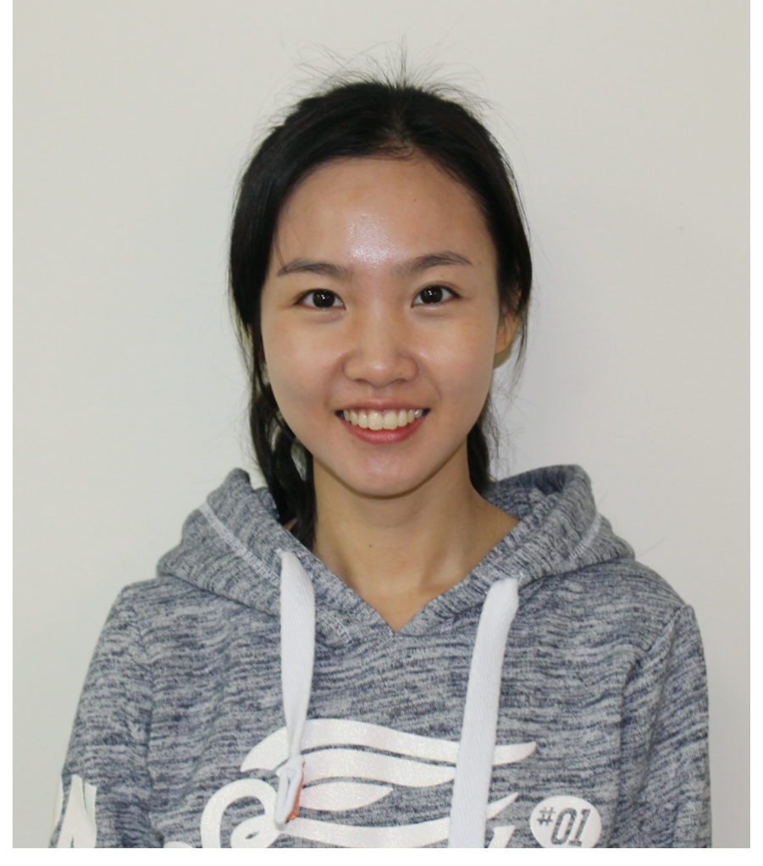


Latent Variable-Based Multivariate Methods to Correct for Batch Effects in Microbiome Data

Yiwen Wang* and Kim-Anh Lê Cao

Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne

* **email:** yiwenw5@student.unimelb.edu.au | **twitter:** @YiwenWang_Eva

Introduction

In the past years, microbial research has made enormous progress with the advent of sequencing technologies to investigate the roles of all microorganisms in different ecological habitats. However, microbiome studies based on 16S amplicon or shotgun sequencing are difficult to replicate as they may suffer from different sources of batch effects. In this context, we define "**batch effect**" as any unwanted source of variation that is unrelated to, and obscures the biological factor of interest. Such batch effects may range from biological to technical and computational factors. Traditional statistical methods developed for microarray or RNA-seq data are generally not suitable for batch effect correction in microbiome data because of the data characteristics. We propose two novel computational methods, PLSDA-batch and sPLSDA-batch based on the Partial Least Squares Discriminant Analysis regression (PLSDA) which take these characteristics into consideration. On real microbiome data, we show that our approaches are robust in removing batch variation while preserving treatment variation compared to existing correction methods. In addition, sPLSDA-batch selects discriminative variables while correcting the batch effects.

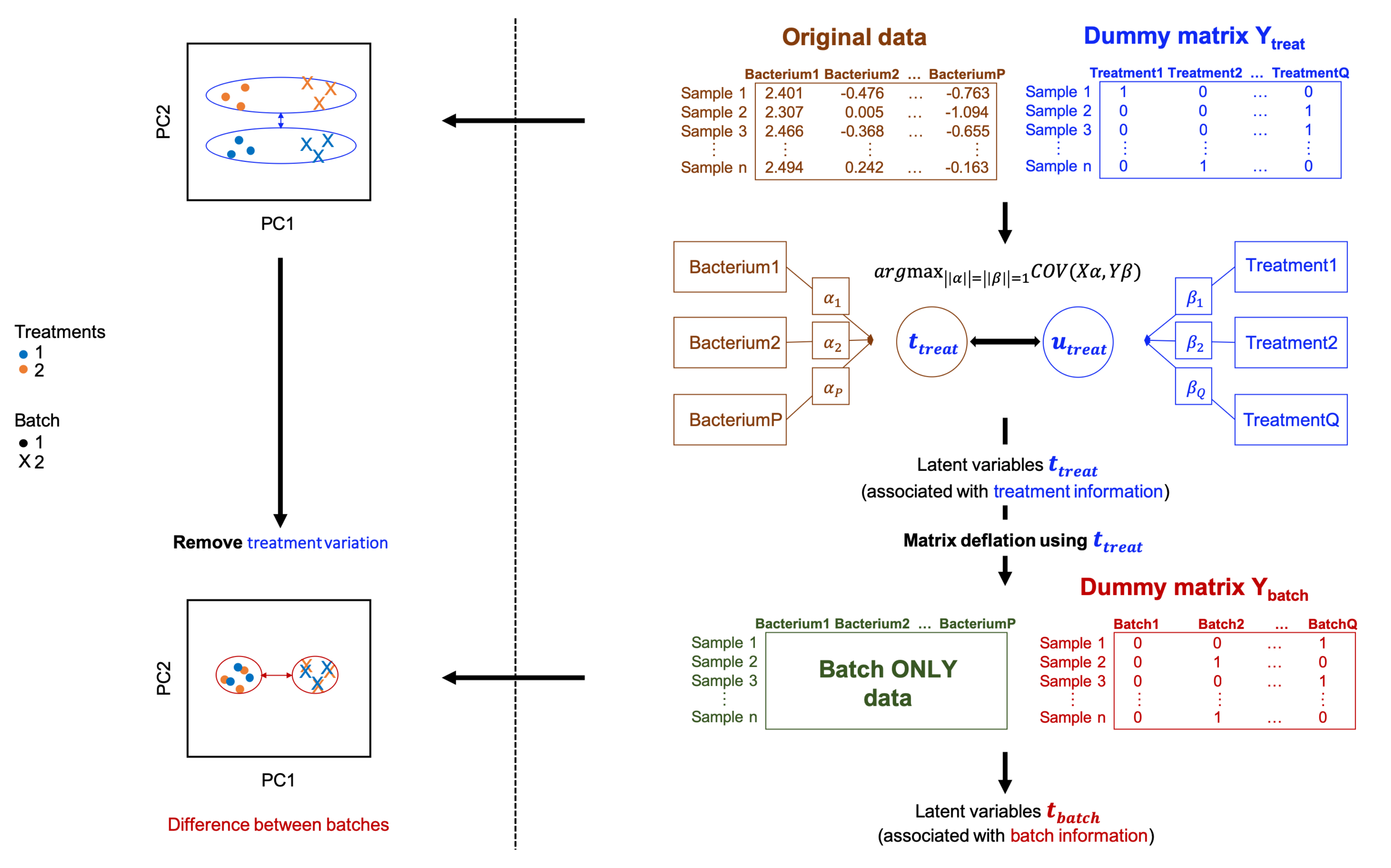
1. Data characteristics

- Sparse, overdispersed
  → skewed and non-Gaussian distribution
- Uneven library sizes
  → bias, difficult to compare samples
- Compositional structure
  → relative abundance, hence data are bounded
- Microbial variables are not independent

2. Limitations of existing methods

- ComBat (*sva* package)
  - Does not consider the correlation between microbial variables
  - Assumes Gaussian distribution
- RUVIII (*ruv* package)
  - Requires sample replicates
  - Requires negative control variables

3. PLSDA-batch


PC2
PC1
Treatments
● 1
● 2
Batch
● 1
X 2
Remove treatment variation
PC2
PC1
Difference between batches
Original data
Dummy matrix Y_{treat}
Bacterium1
Bacterium2
BacteriumP
α_1
α_2
α_P
$argmax_{||\alpha||=||\beta||=1} COV(X\alpha, Y\beta)$
t_{treat}
u_{treat}
β_1
β_2
β_Q
Treatment1
Treatment2
TreatmentQ
Latent variables t_{treat}
(associated with treatment information)
Matrix deflation using t_{treat}
Dummy matrix Y_{batch}
Batch ONLY data
Latent variables t_{batch}
(associated with batch information)

Original data

	Bacterium1	Bacterium2	...	BacteriumP
Sample 1	2.401	-0.476	...	-0.763
Sample 2	2.307	0.005	...	-1.094
Sample 3	2.466	-0.368	...	-0.655
⋮	⋮	⋮		⋮
Sample n	2.494	0.242	...	-0.163

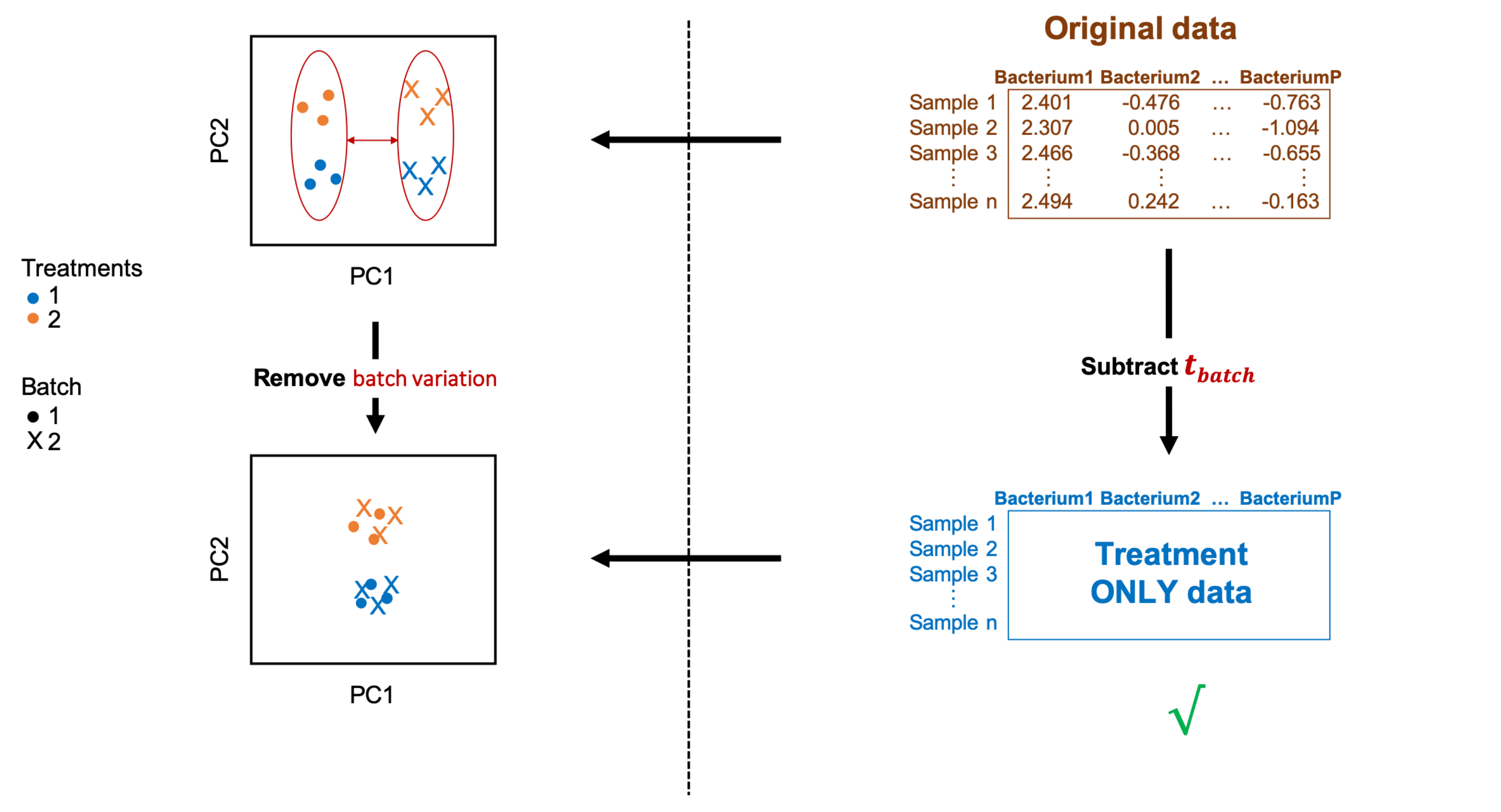
Dummy matrix Y_{treat}

	Treatment1	Treatment2	...	TreatmentQ
Sample 1	1	0	...	0
Sample 2	0	0	...	1
Sample 3	0	0	...	1
⋮	⋮	⋮		⋮
Sample n	0	1	...	0

Dummy matrix Y_{batch}

	Batch1	Batch2	...	BatchQ
Sample 1	0	0	...	1
Sample 2	0	1	...	0
Sample 3	0	0	...	1
⋮	⋮	⋮		⋮
Sample n	0	1	...	0

Figure 1: ***Step1:*** estimating the latent variables associated with batch effects



	Bacterium1	Bacterium2	...	BacteriumP
Sample 1	2.401	-0.476	...	-0.763
Sample 2	2.307	0.005	...	-1.094
Sample 3	2.466	-0.368	...	-0.655
⋮	⋮	⋮		⋮
Sample n	2.494	0.242	...	-0.163

Figure 2: ***Step2:*** removing the latent variables associated with batch effects

4. sPLSDA-batch

Adds constraint to select subsets of microbial variables:

- Only removes batch variation from selected variables that discriminate batches
  → preserves more non-batch variation compared to PLSDA-batch
- Selects variables that discriminate treatments
  → feature selection

5. Managing data characteristics

- Data are Centered Log Ratio transformed to account for uneven library sizes and compositional constraints.
- PLSDA-batch & sPLSDA-batch are non-parametric: Can handle skewed distributions caused by sparsity and overdispersion.
- Their multivariate property accounts for the data correlation structure.

6. Case study

	Anaerobic digestion
No. of microbial variables	231
No. of samples	28
Factor of interest	Phenol concentration
Batch sources	Sequencing dates

7. Results

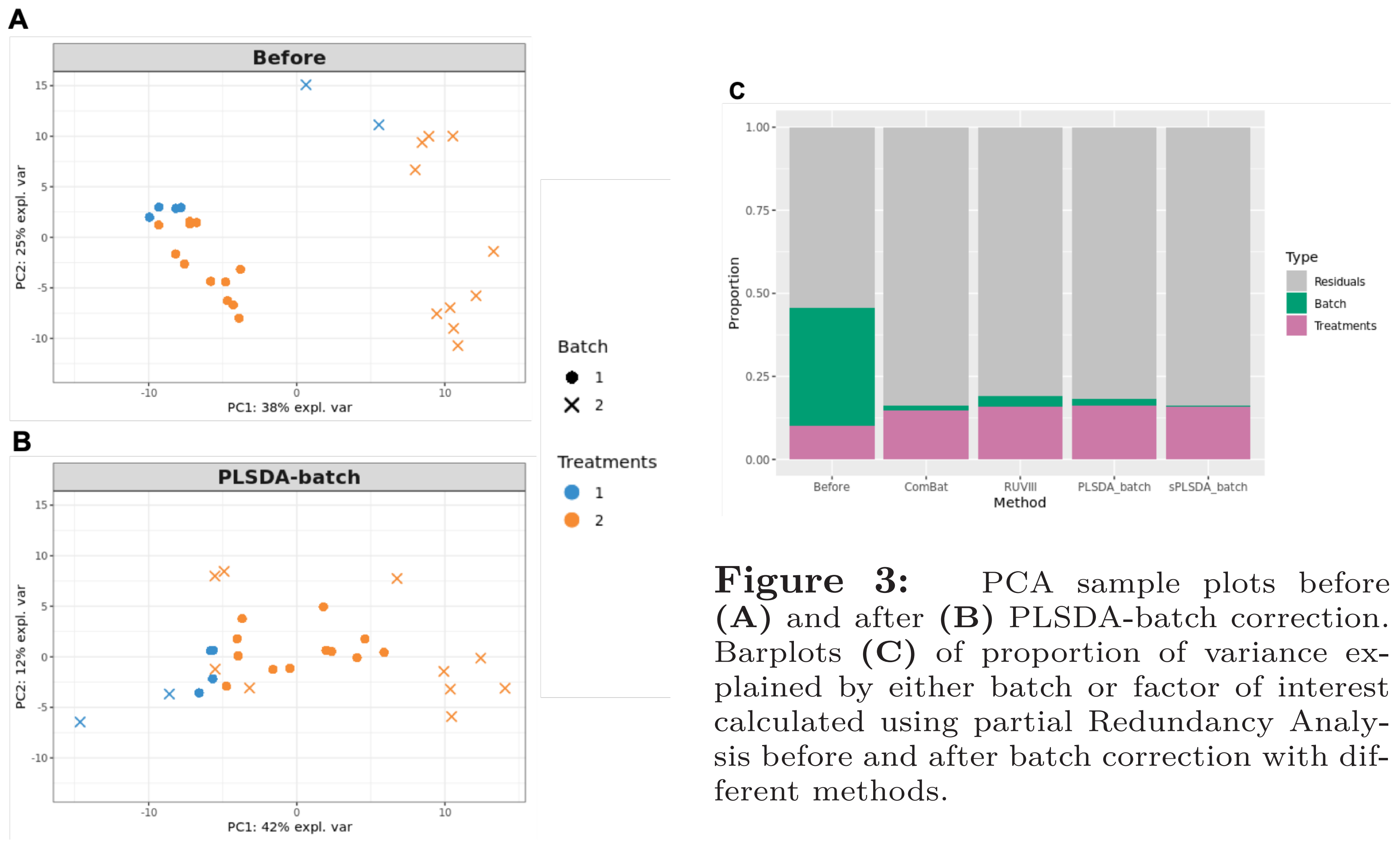


Figure 3: PCA sample plots before **(A)** and after **(B)** PLSDA-batch correction. Barplots **(C)** of proportion of variance explained by either batch or factor of interest calculated using partial Redundancy Analysis before and after batch correction with different methods.

Conclusions

- PLSDA-batch & sPLSDA-batch remove more batch variation compared to RUVIII, revealing more variation of interesting effects compared to ComBat.
- sPLSDA-batch selects discriminative variables while removing batch effects.
- Limitation: batch effects strongly confounded with treatment effects cannot be corrected for in microbiome studies.
- We have validated our methods on simulated data and will work with collaborators to interpret the microbial signature biologically.

Acknowledgements

NHMRC Career Development Fellowship CDF1 GNT1087415 (KALC) and China Scholarship Council (CSC) PhD scholarship (YW).