

SOFTWARE

Open Access

MBECS: Microbiome Batch Effects Correction Suite



Michael Olbrich^{1,2,3*} , Axel Künstner^{1,2†} and Hauke Busch^{1*†}

[†]Axel Künstner and Hauke Busch have contributed equally to this work

*Correspondence: roland.olbrich@ku.ac.ae; hauke.busch@uksh.de

¹ Lübeck Institute for Experimental Dermatology, University of Lübeck, Lübeck, Germany

² Institute for Cardiogenetics, University of Lübeck, Lübeck, Germany

³ Center for Biotechnology, Khalifa University, Abu Dhabi, United Arab Emirates

Abstract

Despite the availability of batch effect correcting algorithms (BECA), no comprehensive tool that combines batch correction and evaluation of the results exists for microbiome datasets. This work outlines the Microbiome Batch Effects Correction Suite development that integrates several BECAs and evaluation metrics into a software package for the statistical computation framework R.

Keywords: Microbiome, Batch effects, R-package, phyloseq, Bioconductor

Introduction

The emergence of unwanted variation in next-generation sequencing applications is a well-researched challenge. A particular form of unwanted technical variation are batch effects (BE) that potentially result from any distinct grouping of samples during the processing steps. Hence, the introduced variability reflects the differences in, for example, the environmental conditions, batches of reagents, sequencing machines, or sample handling for corresponding batches [1, 2]. Consequently, unwanted variation can negatively affect the downstream statistical analyses as it represents a confounding factor that can obscure or exacerbate the biological truth in a dataset [3]. The comprehensive scientific research into causes and strategies for preventing and correcting batch effects indicates this topic's importance [4, 5]. While appropriate measures during the planning and execution of an experiment can limit the emergence and magnitude of batch effects, they are not entirely preventable and thus need to be accounted for before statistical analyses [6]. Despite the availability of batch effect correcting algorithms (BECA) and instructive guides on mitigating of BEs [4], no comprehensive tool that combines batch correction and evaluation of the results exists for microbiome datasets. This work introduces the Microbiome Batch Effects Correction Suite (MBECS), which integrates several established BECAs and evaluation metrics into a software package for the R statistical computation framework.



Features

The Microbiome Batch Effect Correction Suite is designed as a software toolbox that enables users to estimate the severity of batch effects, facilitates the utilization of different BECAs, and finally provides comparative metrics to evaluate the success of each method. To that end, the package offers a convenient 5-step workflow that produces a report to guide the user in selecting the optimal results for downstream analyses.

The software builds upon the *phyloseq* [7] package, which facilitates the intuitive import and export of existing microbiome datasets and enables the use of other count-based datasets. The packages' data object extends the *phyloseq* class with additional fields that store normalized and batch-corrected feature abundance tables. All operations are performed on this single data object that keeps track of the results, promoting tidy scripts and enabling MBECS comparative reporting.

The normalization methods implemented in MBECS are total-sum scaling (TSS) and centered log-ratio transformation (CLR) [8]. Available BECAs include established correction algorithms such as ComBat and Remove Batch Effects from the *SVA* package [9] and Remove Unwanted Variation 3 implemented in the *RUV* package [10]. Additionally, the package implements batch mean centering, Percentile Normalization, and Singular Value Decomposition as correction approaches [11].

Quantifying the variability in a dataset that can be attributed to batch effects is not trivial. A relative log expression (RLE) plot, for example, can indicate the presence of batch effects, yet it is not a suitable approach to determining whether or not they have been removed successfully by a correction algorithm [12]. Thus, the suite implements several distinct metrics to provide the user with comprehensive information to assess the severity of BEs before and after batch-correction procedures. Available methods include constructing linear models from recorded biological and batch factors to estimate the variability attributed to batch effects before and after the correction procedures. Further approaches implemented are partial redundancy analysis and principal variance components analysis [13, 14]. Finally, the silhouette coefficient is a qualitative measure of the goodness of fit of samples to their respective biological groupings [15].

The packages' native workflow depicted in Fig. 1 will create a preliminary report upon importing the dataset. This report summarizes the data concerning covariate information, distribution of samples into biological groups and known batches, heatmaps, and box plots of the most variable features concerning the batch factor and relative log-expression plots. The preliminary report also provides the metrics mentioned above to assess variability for the uncorrected data. The user can decide whether or not batch correction is required based on that account. The subsequent processing step allows the application of selected correction methods depending on the experimental design. Methods like *RUV-3* specifically require technical replicates in different batches to work; Batch mean centering is only applicable to datasets that comprise two-factor biological groupings, i.e., case-control studies [10]. Therefore, it is up to the user which methods to use, and all the correction results are stored within the data object.

The third step constructs the post-correction report. This report provides comparative analyses between uncorrected data and all the employed correction algorithms. The user can use these to evaluate the correction algorithms in terms of reduced

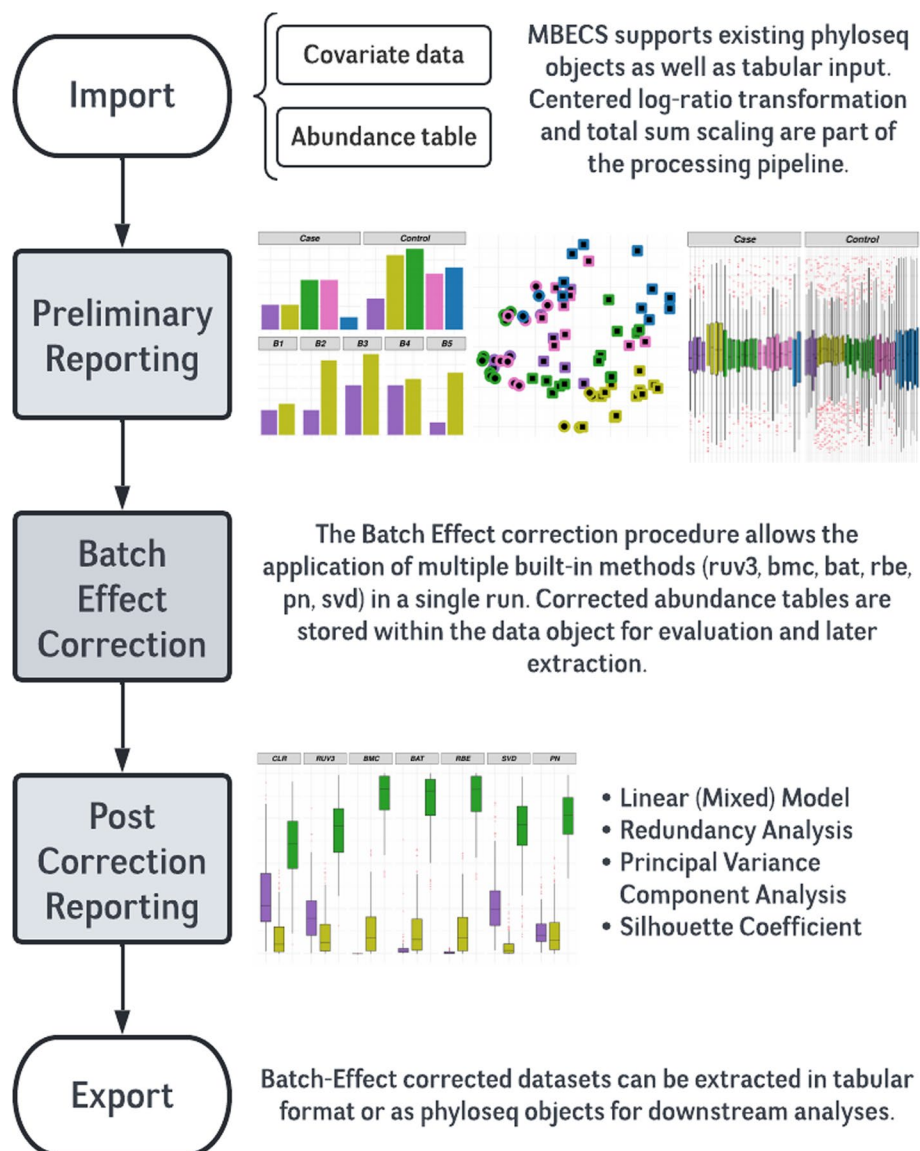


Fig. 1 The MBECS processing pipeline comprises five main steps that provide users with the means to investigate potential batch effects and mitigate them before downstream statistical analyses: The preliminary report provides an overview of a data set to assess the presence and gravity of batch effects. The correction step can apply and store the output of various BECAs. The available correction methods are Remove Unwanted Variation 3 (ruv3), Batch Mean Centering (bmc), ComBat (bat), RemoveBatchEffects (rbe), Percentile Normalization (pn), and Singular Value Decomposition (svd). Several variance assessment methods, e.g., linear modeling, principal variance component analysis, and redundancy analysis are employed in the post-processing reports to produce a comparative qualitative analysis between the uncorrected data set and the selected BECAs. The export functionality allows extraction of the transformed or corrected counts in tabular or phyloseq formats to facilitate downstream statistical analyses

unwanted variability while preserving the biological variation that is investigated with the experimental design. An instructive manual for the package and examples of preliminary and post-corrections reports are available as supplemental material accompanying the online article (Additional file 1, Additional file 2, Additional file 3).

Implementation

The Microbiome Batch Effect Correction Suite is available as a software package for the R programming framework at Bioconductor. The latest development version can be obtained from the GitHub repository.

Availability and requirements

Project name: MBECS Microbiome Batch Effect Correction Suite

Project home page: <http://www.bioconductor.org/packages/release/bioc/html/MBECS.html>

Operating system(s): Platform independent

Programming language: R (>= 4.1)

Other requirements: CRAN and Bioconductor packages (methods, magrittr, phyloseq, limma, lme4, lmerTest, pheatmap, rmarkdown, cluster, dplyr, ggplot2, gridExtra, ruv, sva, tibble, tidyr, vegan, stats, utils, Matrix)

License: Artistic-2.0

Any restrictions to use by non-academics: None

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-023-05252-w>.

Additional file 1. Vignette.

Additional file 2. Preliminary report.

Additional file 3. Post-correction report.

Acknowledgements

None.

Author contributions

Conceptualization, MO, AK, HB; Methodology, MO; Validation, MO; Formal analysis, MO; Investigation, MO; Resources, MO; Data curation, MO; Writing—original draft preparation, MO; Writing—review and editing, MO, AK, HB; Visualization, MO; Supervision, AK, HB; Project administration, AK, HB; Funding acquisition, HB. All authors read and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL. Hauke Busch acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 22167-390884018).

Availability of data and materials

The source code is freely available under Artistic-2.0 license at <https://github.com/rmolbrich/MBECS> and at <https://bioconductor.org/packages/release/bioc/html/MBECS.html>. The packages vignette and examples utilize artificial mockup data to illustrate workflow and execution. The package vignette and two exemplary reports are available as supplementary data.

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: 12 October 2022 Accepted: 20 March 2023

Published online: 03 May 2023

References

1. Chen C, et al. Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS ONE*. 2011;6:e17238.
2. Čuklina J, et al. Review of batch effects prevention, diagnostics, and correction approaches. In: Matthiesen R, editor, et al., *Mass spectrometry data analysis in proteomics, methods in molecular biology*. New York: Springer; 2020. p. 373–87.
3. Goh WWB, et al. Why batch effects matter in omics data, and how to avoid them. *Trends Biotechnol*. 2017;35:498–507.
4. Wang Y, LêCao KA. Managing batch effects in microbiome data. *Brief Bioinform*. 2020;21:1954–70.
5. Scherer A, editor. *Batch effects and noise in microarray experiments: sources and solutions*. Chichester: Wiley; 2009.
6. Zhou L, et al. Examining the practical limits of batch effect-correction algorithms: when should you care about batch effects? *J Genet Genomics*. 2019;46:433–43.
7. McMurdie PJ, Holmes S. phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE*. 2013;8:e61217.
8. Kucera M, Malmgren B. Logratio transformation of compositional data—a resolution of the constant sum constraint. *Mar Micropaleontol*. 1998;34:117–20.
9. Leek JT, et al. The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28:882–3.
10. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. *Biostatistics*. 2012;13:539–52.
11. Gibbons SM et al. Correcting for batch effects in case-control microbiome studies. 17
12. Gandolfo LC, Speed TP. RLE plots: Visualizing unwanted variation in high dimensional data. *PLoS ONE*. 2018;13:e0191629.
13. Li J, et al. Principal variance components analysis: estimating batch effects in microarray gene expression data. In: Scherer A, editor, et al., *Batch effects and noise in microarray experiments*. Chichester: Wiley; 2009. p. 141–54.
14. Liu Q. Variation partitioning by partial redundancy analysis (RDA). *Environmetrics*. 1997;8:75–85.
15. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53–65.

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

