Rachel Griffard

June 25, 2024

micRoclean: an R package containing decontamination pipelines for low-biomass microbiome data

Rachel Griffard, Dong Pei

1. Introduction

* Briefly introduce the research area or problem domain addressed by the R package.

*The microbiome is used as a biomarker for disease and disease progression. These studies most commonly using high-biomass samples such as stool and saliva. Compared to the high-biomass microbiome studies, however, environmental bacteria often obscure true biological signal in low-biomass microbiome studies such as skin or blood plasma (Salter et al., 2014).*

* Highlight the limitations of existing methods or tools.

*Current methods to decontaminate data from 16S-rRNA sequencing data can broadly be classified into three main method categories: blacklist, sample-based, and control-based (Hulpusch et al., 2023). Blacklist methods identify features that are previously identified as common contaminants. Sample-based methods identify contaminant features based on the relative abundance between samples. Control-based methods identify contaminant features based on included negative controls.*

* Motivate the need for your new R package.

*While there exists methods…*

2. Related Work

* Discuss existing R packages that address similar problems.

*Currently, there are many methods that aim to decontaminate 16S-rRNA sequencing samples. Most of these methods remove entire features that are identified as contaminants. A method of this type that is used extensively is decontam (Davis et al., 2018). decontam combines control- and sample-based contaminant identification and removes features from the sample that are identified as contaminants. Similarly,*

*On the other hand, some methods identify and remove partial features that are considered contamination. This partial removal appears in packages microDecon (McKnight et al., 2019) and SCRuB (Austin et al., 2023). Both of these methods are entirely control-based methods.*

* Clearly differentiate your package from existing solutions by highlighting its unique functionalities or advantages.

*We introduce micRoclean, an R package which houses multiple pipelines aimed at decontaminating low-biomass microbiome data to either: (i) characterize original composition of the sample prior to contamination or (ii) identify potential biomarkers. Furthermore, micRoclean provides insight into the impact of removal through a filtering loss statistic. This filtering loss statistic can provide insight into potential over filtering.*

3. Package Description

* Provide the name of your R package.

***See above***

* Briefly explain the core functionalities offered by the package.
* Mention the type of data the package handles (e.g., time series, text data).
* Briefly describe the key algorithms or methods implemented in the package.

4. Package Functionality

* In more detail, describe the main functionalities of the package.

5. Examples

* Showcase the functionalities of the package with practical examples.
* Use real-world or simulated data relevant to the addressed research area.
* Clearly explain the steps involved in using the package for the example tasks.

6. Evaluation (Optional)

* If applicable, include an evaluation section that compares your package to existing solutions.
* Use appropriate metrics to benchmark performance or accuracy.

7. Conclusion

* Summarize the key contributions of your R package.
* Highlight the potential applications and target audience for the package.
* Briefly mention future development plans for the package.

8. Software Availability

* Provide information on how to access and install the R package.
* Mention the repository where the package is hosted (e.g., GitHub, CRAN).

9. Appendix (Optional)

* Include additional technical details, extended code examples, or user guides as supplementary material.

Additional Tips

* Follow the specific guidelines and formatting requirements of the target journal.
* Ensure the R package is well documented with clear vignettes and help files.
* Consider including a reference list citing relevant research papers and software tools.

References

Austin, G.I., Park, H., Meydan, Y. et al. Contamination source modeling with SCRuB improves cancer phenotype prediction from microbiome data. Nat Biotechnol 41, 1820–1828 (2023). https://doi.org/10.1038/s41587-023-01696-w

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., & Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. Microbiome, 6(1), 226. https://doi.org/10.1186/s40168-018-0605-2

Hülpüsch, C., Rauer, L., Nussbaumer, T. et al. Benchmarking MicrobIEM – a user-friendly tool for decontamination of microbiome sequencing data. BMC Biol 21, 269 (2023). https://doi.org/10.1186/s12915-023-01737-5

McKnight, D. T., Huerlimann, R., Bower, D. S., Schwarzkopf, L., Alford, R. A., & Zenger, K. R. (2019). microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. Environmental DNA, 1(1), 14-25. https://doi.org/10.1002/edn3.11