Rachel Griffard

Low-Biomass Microbiome Decontamination Pipeline

Background: The microbiome is used as a biomarker for disease and disease progression, most commonly using high-biomass samples such as stool and saliva. Cell-free microbial DNA (cfmDNA) in blood plasma is an attractive alternative sample as it may be less invasive and is often more readily available cfmDNA is often bacterial in origin and has previously shown differences between healthy and diseased patients and among prognoses (Pietrzak et al., 2023). Compared to the high-biomass microbiome studies, however, environmental bacteria often obscure true biological signal in these low-biomass microbiome studies (Salter et al., 2014).

Gap: R package and optional dashboard whose primary function is to provide a structured decontamination pipeline leveraging multiple established methods for low-biomass microbiome data from 16S-rRNA sequencing.

Goal: Develop and release R package containing a low-biomass microbiome decontamination pipeline.

Data: Below are some potential data options for both comparison of methods and vignette of future package

* Zozaya-Valdes et al., 2023 – Low-biomass blood plasma
* Karsets et al., 2019 – Even mock community
* Hülpüsch et al., 2023 two staggered mock community datasets, one skin low-biomass dataset

Pipeline:

1. Remove features with different abundance in different batches
   1. ANCOMBC
   2. Compare based on batch-dependent variables
   3. **Sample-based approach (Hulspusch et al., 2023) pg. 2**
      1. Similar to frequency approach in Decontam
2. Remove features higher in negative control and lower in samples
   1. Decontam prevalence test
   2. **Control-based approach (Hulspusch et al., 2023)**
      1. Similar to microbIEM and Decontam prevalence method
3. Remove features with different abundance in different batches for technical replicates
   1. Interrater reliability (Zozaya et al., 2021)
   2. “Spearman’s ρ was calculated using centred log-ratio-transformed microbial relative abundances. Centred log-ratio transformations and Spearman’s ρ were calculated using the clr function of the compositions package78 and cor.test function in R” (Tan et al., 2023)
4. Remove previously known contaminant
   1. Previous published “blocklist” (Eisenhofer et al., 2019)
   2. Possible contaminant species and genera list from Piro et al., 2023
   3. **Blocklist approach (Hulpusch et al., 2023)**

Potential pipeline:

* Liu, Y., Elworth, R.A.L., Jochum, M.D. et al. De novo identification of microbial contaminants in low microbial biomass microbiomes with Squeegee. Nat Commun 13, 6799 (2022). <https://doi.org/10.1038/s41467-022-34409-z>
  + For use when no negative controls
* Cross-contamination v. contaminant DNA (Eisenhofer et al., 2019)
  + See three types of negative controls encouraged in study (sampling blank controls, DNA extraction blank controls, no-template amplification controls)

Methods to compare to:

* MicrobIEM (Hülpüsch et al., 2023)
  + R Shiny
* GRIMER (Piro et al., 2023)
  + HTML
  + Leverages Decontam?
* Decontam (Davis et al., 2018)
  + R
  + Primarily for high-biomass
  + Both prevalence and frequency filters
* Frequency filter
  + Remove all below abundance threshold
* Presence filter
  + Removes all sequences appearing in negative control
* SourceTracker (Knights et al., 2011)
  + R
  + Bayesian

Package:

* decontaminate(object, contaminant = 2, prevalence\_threshold = 0.5, blocklist = blocklist, venn = FALSE, dashboard = FALSE)
  + **Goal: raw object in, returned filtered object and list of removed features with information on which steps removed (along with optional Venn diagram)**
  + object
    - raw phyloseq or summarized experiment object
  + prevalence\_threshold = 0.5
    - threshold used for decontam *isContaminant()* method
  + blocklist = blocklist
    - blocklist object with genus level for known contaminants
    - default is blocklist from Eisenhofer et al., 2019
  + venn = FALSE
    - indicates if user would like to return Venn diagram comparing overlap between steps of features identified as contaminants
  + dashboard = FALSE
    - indicates if user would like to return HTML interactive Shiny dashboard to explore the data

Next steps:

* How does user indicate which samples in phyloseq/summarized experiment object are negative controls
* Consideration of impact if change from ANCOM-BC to equivalence test
  + Is this possible being that we are working with compositional data?
  + Is this beneficial?
  + If we choose not to, how does this impact our interpretation? (see notes from SOWG meeting)
    - **ANCOMBC2 does not leverage equivalence test**
* How can we maintain a “blocklist” that is relevant to the time users run the pipeline?
  + Concern about only being at genus level
  + Can we maintain this
* How do we “compare” these methods?
  + See Hülpüsch et al., 2023 for further ideas
* Main differences between microbIEM and GRIMER methods **(see below)**
  + Both methods are focused on creating an interactive dashboard to make data exploration more accessible to non-coding users
  + GRIMER visualizes results from Decontam method
  + microbIEM uses only control-based approach, compares negative controls to environmental samples (similar to Decontam)
* **How do we account for the use of multiple methods for decontamination?**
  + Is there an impact from this? Thinking multiple testing for example

References

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

Eisenhofer, R., Minich, J. J., Marotz, C., Cooper, A., Knight, R., & Weyrich, L. S. (2019). Contamination in Low Microbial Biomass Microbiome Studies: Issues and Recommendations. Trends in microbiology, 27(2), 105–117. <https://doi.org/10.1016/j.tim.2018.11.003>

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>

Karstens, L., Asquith, M., Davin, S., Fair, D., Gregory, W. T., Wolfe, A. J., Braun, J., & McWeeney, S. (2019). Controlling for Contaminants in Low-Biomass 16S rRNA Gene Sequencing Experiments. mSystems, 4(4), e00290-19. https://doi.org/10.1128/mSystems.00290-19

Knights, D., Kuczynski, J., Charlson, E. et al. Bayesian community-wide culture-independent microbial source tracking. Nat Methods 8, 761–763 (2011). https://doi.org/10.1038/nmeth.1650

Pietrzak, B., Kawacka, I., Olejnik-Schmidt, A., & Schmidt, M. (2023). Circulating Microbial Cell-Free DNA in Health and Disease. International journal of molecular sciences, 24(3), 3051. <https://doi.org/10.3390/ijms24033051>

Pollock, J., Glendinning, L., Wisedchanwet, T., & Watson, M. (2018). The Madness of Microbiome: Attempting To Find Consensus "Best Practice" for 16S Microbiome Studies. Applied and environmental microbiology, 84(7), e02627-17. https://doi.org/10.1128/AEM.02627-17

Salter, S.J., Cox, M.J., Turek, E.M. et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol 12, 87 (2014). <https://doi.org/10.1186/s12915-014-0087-z>

Piro, V. C., & Renard, B. Y. (2022). Contamination detection and microbiome exploration with GRIMER. GigaScience, 12, giad017. https://doi.org/10.1093/gigascience/giad017

Tan, C.C.S., Ko, K.K.K., Chen, H. et al. No evidence for a common blood microbiome based on a population study of 9,770 healthy humans. Nat Microbiol 8, 973–985 (2023). <https://doi.org/10.1038/s41564-023-01350-w>

Zozaya-Valdés, E., Wong, S.Q., Raleigh, J. et al. Detection of cell-free microbial DNA using a contaminant-controlled analysis framework. Genome Biol 22, 187 (2021). <https://doi.org/10.1186/s13059-021-02401-3>

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Zozaya et al. | GRIMER | microbIEM | DECONTAM |
| **Goal** | * Decontaminate low-biomass microbiome pipelines | * Create interactive report/dashboard from DECONTAM output * Make accessible for non-coding researchers | * ID contaminants based on relative abundance in negative controls * Provide GUI for non-coding researchers | * Provide R package with statistical classification to ID contaminants based on:  1. Contaminants at higher frequencies (sample) 2. Contaminants often found more in negative controls (control) |
| **Method** | Sample-based approach, control-based approach, blacklist-based approach | Control-based approach, sample-based approach (both from decontam) | Control-based approach | Control-based approach, sample-based approach |
| **What is the difference?** | * Not wrapped in R package * Used limma * Not streamlined | * Uses prior method, just puts into dashboard | * Focuses on one method | * Leverages two methods, not uses blacklist approach * Lacks dashboard approach? |
| **What can we pull from this?** | Overall pipeline layout/order | Dashboard/HTML output makes more accessible and sharable for dissemination of information? | Mock communities to test effectiveness  Youden’s index | Prevalence method (step 2)  ~ Should we consider the frequency method? |

Split Pipeline Method

Pipeline 1: Multiple batches, negative controls

* See above pipeline
* Implement toggle between microDecon and decontam

Pipeline 2: Negative controls, well information

With well information, users would be well off to use the SCRuB decontamination as it uses this spatial information

1. SCRuB
2. Measure filtering loss (PERfect)