https://doi.org/10.1038/s41564-025-01944-6

Planning and describing a microbiome data analysis

Amy D. Willis & David S. Clausen

We provide guidance on the planning, execution and description of statistical analyses in microbiome studies.

Scientific advances in microbial ecology and microbiome science rely on both high-quality data and its high-quality analysis. To achieve this, the findings must be both reproducible (the analysis must be repeatable on the same data) and replicable (the findings must hold up in a future study that collects analogous data)[1]. To ensure reproducibility, information is needed on sample collection and laboratory processing, such as sample storage temperature, DNA extraction kit and protocol, PCR primers and number of PCR cycles performed, as well as post-processing (bioinformatics) details, including software used for contig assembly or sequence variant construction. Unfortunately, when it comes to statistical analysis, essential details and justification are frequently missing from the methods.

For studies that focus on a small number of targets, a single paragraph is often sufficient to describe the statistical analysis performed. However, modern genomics fields, including microbiome science, now frequently collect quantitative information on thousands of biological units, including strains, genes and metabolic pathways, which can be analysed and summarized in myriad ways. Many modern microbiome investigations span within-ecosystem summary statistics such as alpha diversity, across-ecosystem community overlap or beta diversity, differential abundance, phylogenetic analysis and more. As with laboratory and raw data processing steps, each of these methods requires a careful description of the approach, justification and quality control assurances.

Describing the statistical analysis plan in the level of detail described above demands more of an author, but there are a number of advantages. The process of determining outcome and predictor variables (discussed in more detail below) may assist authors in selecting an analysis plan that is consistent with the goals of the research. As a result, peer review is likely to be streamlined as reviewers and readers can unambiguously discern what analyses were performed. Readers are more likely to correctly interpret results, be confident findings are robust and to believe that significant results are attributable to biology. Finally, because authors' assumptions are clearly stated, it will accelerate the development of statistical methods that are consistent with authors' needs.

To assist authors in planning, executing and describing their statistical analyses, we provide general guidance for microbiome researchers and a checklist (Box 1) to confirm the transparency and completeness of a described analysis.

Creating a statistical analysis plan

Drafting the statistical analysis before obtaining data can improve the quality of research. It can align experiments with the target scientific questions, ensure that the experimental design can be accounted for in the analysis and highlight relevant data for quality control or validation. For example, despite widespread enthusiasm for longitudinal microbiome studies, many studies are, in fact, interested in cross-sectional comparisons (for example, between populations who differ in an environmental exposure), for which longitudinal designs may not be resource-optimal. Similarly, if the abundance of a specific microbial community member is a scientific target, absolute quantitation data could be collected in addition to community survey data. Below we provide a workflow for drafting and refining the statistical analysis section of a microbiome study with examples.

List your planned analyses

First, make a list of all analyses that you plan to perform. This should be done for each biological unit, even if distinct biological units will be constructed from the same raw data. For example, your list might include:

- Taxonomic analysis: alpha diversity (richness, Shannon diversity), beta diversity (Bray–Curtis, UniFrac), differential abundance (proportions, fold changes), differential presence/absence (odds of being observed).
- Gene analysis: alpha diversity (richness), differential abundance (differences in proportions, fold changes), differential presence/absence (odds of being observed).
- Sequence analysis: phylogenetic analysis (species tree for metagenome-assembled genomes).

Each item in the list above is a different biological unit that can be constructed from the same raw data, such as shotgun sequencing data. You should include all analyses in your list, not just those you plan to discuss in your manuscript or those that produced statistically significant hypothesis tests.

Describe the model

For each analysis that you plan to perform, state the model to be fit. In many cases, a minimum full description of the model includes:

- The outcome variable, for example, Chao1 (richness); sample Shannon diversity; or whether more than five reads were detected (presence–absence).
- The set of predictors, for example, some combination of age (continuously measured); BMI (continuously measured); binarized sex (female as baseline); and an indicator for statin medication use.
- The model used to connect the outcome and predictor variables. This could be a simple model (for example, a linear regression) or a more complex model that accounts for the experimental design (for example, generalized estimating equations with an exchangeable correlation structure within co-housed animals).

BOX 1

Checklist for describing a statistical analysis

The below checklist is intended to help authors preparing manuscripts communicate transparency of both scientific process and statistical analysis. This information could be contained in the methods, results or supplemental material. Additionally, well-documented code and all relevant data are essential for reproducibility and provide a backstop for resolving ambiguities about the analysis.

1. Were all analyses that were performed reported and described? Can a list of analyses that were investigated but not included in the main text be found in the methods? What assurances are provided that analyses were specified in advance, rather than mining the data for significant results?
2. Are exploratory analyses (for example, sequencing depth and quality control reporting, and visualizations of low-dimensional data summaries) and inferential (statements about statistical parameters via hypothesis testing) analyses distinguished from each other? Confirm that analyses described as exploratory rely on visualizations and descriptive statistics, and not statistical inference.
3. Is each analysis fully described? For all results that arise from regression models, was the approach for choosing the adjustment variables detailed? Were the outcome and each predictor variable clearly described? Is the model used to link the outcome to the predictors detailed?
4. For all statistical inference and *P* values reported, is the null hypothesis unambiguous? Are the quantities tested clearly stated, along with their units?
5. For any hypothesis tests that arose from a regression model, is it clear if the model is 'adjusted' or 'unadjusted' for additional predictors? If adjusted, is the adjustment set clear and its justification provided?
6. If statistical inference is used to screen many null hypotheses, is it clear whether the reported significance values are corrected for multiple comparisons? If corrected, is the method of correction (for example, Bonferroni or Benjamini–Hochberg) clear?
7. If point or interval estimates for a parameter obtained via a regression model are reported, is a clear interpretation of the parameter provided? Does this interpretation take into account all variables adjusted for in the regression model?
8. Is it clear whether the primary scientific question is causal or associative? Does the analysis align with this objective? If the scientific question is causal, are the causal assumptions clearly stated? Is there a discussion of both their justification and validity? Is confounding bias discussed and are reasonable steps taken to address it?
9. Is it clear whether any data manipulation was performed before analysis? For example, were zero counts replaced by a pseudocount? If so, what pseudocount was chosen? Are data transformations described (for example, centred log-ratio)? Was aggregation across taxonomy or phylogeny performed? If so, was aggregation performed before or after data transformation? If any samples or sequences were discarded, was the procedure for doing so described and justification given?
10. For the most important results from the paper, what sensitivity analyses were performed? For example:
    a. For any methods that are sensitive to the inclusion of infrequently detected units (for example, estimating and/or comparing species richness and Shannon diversity), how did excluding increasingly rare units impact the findings?
    b. If any Bayesian methods were used, how were priors chosen? How was convergence to the posterior distribution assessed?
    c. For any method that requires tuning or other hyperparameters, was the approach to select these parameters clear and justified within the study setting? If software defaults were selected, are the chosen parameters reasonable for the sample size and data structures considered in the current study?
11. Is all data required to repeat the analysis available, including both sample data and processing information (for example, extraction and sequencing batches)? Is protected personally identifiable information (including human host genome sequences) removed?
12. Is a code appendix provided? Is a 'readme' file describing the workflow available? Is the code well-commented? Were all computing tools (for example, R and GraphPad) and software packages (for example, ggplot and data.table) cited?

In other cases, such as phylogeny estimation, there is no set of predictors, and describing the model used to estimate the outcome variable becomes more important (for example, general time reversible model with unequal mutation rates and unequal nucleotide frequencies[2]).

Justify your choice of predictors

Next, justify the choice of predictors in each model. Most researchers intuitively know that observations that are alike in all ways except for the primary exposure or intervention (for example, treatment) are most appropriate for comparison. However, a formal framework for classifying predictors facilitates scientific discourse about the implications of their omission or inclusion and can clarify hypotheses about mechanism. We provide some guidelines for choosing predictors below.

Predictor of interest. The predictor of interest is the main predictor variable. This variable should always be included because you would like to study how the distribution of the outcome (or a summary of the outcome's distribution, such as the mean, odds or hazard) varies with changes in the predictor variable. In analyses where a microbial feature is the outcome variable, and clinical or environmental characteristics are the predictor variables, common predictors of interest include disease, treatment and amendment.

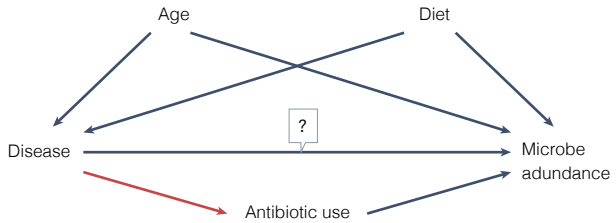

Fig. 1 | Selecting adjustment sets by identifying causal assumptions. Causal assumptions should drive the choice of variables to adjust for (adjustment sets) because different causal assumptions imply different adjustment sets. A generic observational cross-sectional study (no longitudinal sampling) is used as an example, where we are interested in the total effect of disease (the predictor of interest) on microbial abundances (the outcome variable). First, a causal diagram should be constructed based on context, scientific knowledge and/or published literature (see ref. 6 for guidance). Then, the causal diagram can be queried to obtain adjustment sets (for example, using ref. 13). Given the causal assumptions shown here, both age and diet are in the adjustment set and should therefore be included as predictors in the model. In the absence of the red causal assumption, we could include antibiotic use as a precision variable. However, in the presence of the red causal assumption, we should not adjust for antibiotic use in our model, as antibiotic use mediates the relationship between microbial abundances and disease.

Precision variables. Precision variables are correlated with the outcome but uncorrelated with the predictor of interest. Precision variables are so named because they can increase the power to reject false null hypotheses[3], often by reducing the uncertainty in the parameter associated with the predictor of interest. Common precision variables in studies where microbial abundances are the outcome include batch and sequencing variables (sequencing batch may be correlated with observed counts but not with microbial abundances). There are typically a large number of both unmeasured and measured-but-omitted precision variables in any analysis. Omitted precision variables do not bias parameter estimates in linear regression but can induce bias and impact results interpretation in other regression models[3].

Confounders. In any study that aims to understand causes, a sufficient set of confounders should be adjusted for to prevent confounding bias. In simple settings, confounders are variables that are (1) causally upstream of the outcome, (2) causally upstream of the predictor and (3) not in the causal pathway of interest[4]. To determine if a variable is a confounder, a set of causal assumptions must be stated. This is typically done through the construction of a causal diagram (Fig. 1). After a causal diagram has been constructed, it can be queried to obtain adjustment sets that control for confounding. Determining adjustment sets from causal diagrams ensures that confounders are adjusted for, and variables that would induce bias in parameter estimates, such as colliders[5], are correctly omitted (see ref. 6 for further reading). Importantly, selecting adjustment sets using the data carries substantial risks, including overstating statistical significance, understating uncertainty and compromising prediction[7]. For this reason, we strongly advocate for adjustment set selection via causal reasoning.

Selecting adjustment variables. A particular measurement's 'type' depends on the study. For example, variables associated with diet (for example, fibre intake) could be the predictor of interest, a precision variable, a confounder, a collider or a completely irrelevant variable, depending on the specific research question. For this reason, in studies where a measure of microbial abundance is the outcome, it is not possible to give universally applicable advice on which variables should be adjusted for. This advice requires causal assumptions, which is context dependent. That said, we provide an example set of causal assumptions and the resulting adjustment sets for a cross-sectional study in Fig. 1.

We encourage authors to clearly state if the goals of the study are causal or associative. For example, answering 'do adults who have colorectal cancer have a higher abundance of *Fusobacterium nucleatum* than adults who do not?' is an associative goal. By contrast, answering 'does having colorectal cancer increase the abundance of *F. nucleatum*?' is a causal goal. If confounding is a concern, then the goal of the study must be causal[8]. This is true even though the potential for unmeasured confounding means estimates cannot be causally interpreted without making untestable assumptions. Nevertheless, because causal and associative questions require different analyses, the study's goal must be stated clearly[8].

Adjusting for variables in a regression model means that comparisons are made across groups that are alike with respect to the adjustment variable. For example, if you are interested in comparing bacterial species diversity in the human gut across populations that differ in their use of a medication, comparing populations that are similar in age, sex and adiposity informally 'matches' subjects who differ in their medication use, but are otherwise similar. If the goal of the study is to answer the causal question 'does the medication reduce species diversity in the human gut', adjusting for these characteristics reduces the risk of entangling differences in diversity due to the medication with differences in diversity due to other variables. Of course, even with adjustment, the estimate may be biased owing to an unmeasured variable that may confound the relationship between medication and diversity.

Statistical inference

If statistical inference was performed, clearly state the null hypothesis that was tested. For example, 'we tested the null hypothesis that the average observed Shannon diversity was equal across groups who differed in statin medication use, when comparing individuals of the same age, BMI and sex'. The test that was performed (for example, a robust score test using a normal likelihood or a Wald test assuming homoscedasticity) should be clearly stated. Citing relevant software can clarify the test statistic and inference procedure that was performed, but it cannot clarify the null hypothesis that was tested. Note that the null hypothesis is impacted by the choice of variables that are adjusted for in the analysis. A common error that we have seen in microbiome studies is interpretations of hypotheses as either adjusted or unadjusted when the opposite is true, creating confusion about the comparison being made and ambiguity about the treatment of adjustment variables. For example, the null hypothesis that 'the average observed Shannon diversity is equal across groups who differ in statin medication use' is critically different from the aforementioned null hypothesis.

Assumptions

Key assumptions of each method should be stated alongside approaches used to validate them. For example, many methods that estimate differences in proportions of microbial abundances assume that all microbes are equally well detected (for example, refs. 9–11), which could be validated using sequencing controls. Similarly, independence of observational units (conditional on covariates) is a widespread and critical assumption of most statistical tests. Independence should be discussed in the context of the experiment's design, especially when

the study design is longitudinal, involves repeated measurements or batch processing, or otherwise involves sampling correlated observations. Phrasing assumptions as 'limitations' of the analysis highlights the rigour of the investigation.

Sensitivity analyses

A description of any sensitivity analyses performed should be included. For example, the results of any Bayesian analysis depend on the choice of prior parameters, and either justifying the choice of priors on scientific grounds or illustrating the robustness of results to the prior choice strengthens quantitative arguments. Similarly, many approaches used to analyse microbial abundance replace the empirical abundance of unobserved taxa with a small non-zero abundance. Investigating sensitivity to the choice of this 'pseudocount' gives assurance that differential abundance findings are not artefactual. Likewise, the choice to remove rare species from an analysis can substantially alter estimates of diversity, but results can easily be reanalysed with different thresholds for prevalence filtering[12]. Focusing the sensitivity analysis on the most important findings of the manuscript is more feasible and rigorous than a shallow reinvestigation of all results.

Conclusion

We advocate for a detailed and organized 'statistical analysis' section to be included in all microbiome research papers, including descriptions and justification for all analyses. We recommend that key assumptions be stated as limitations, and sensitivity analyses be performed on key results. Although most authors often have excellent justification for their choices, transparency for readers and reviewers will improve the reproducibility and replicability of research; clarify interpretations of results; and enhance the rigour of our field.

Our guidelines are more involved than current expectations for describing microbiome data analyses, and we acknowledge that scientists may not know the answers to questions we pose about statistical assumptions and parameter interpretation. In an ideal world, every research team that deals with quantitative data would have access to statistical expertise. In the absence of such expertise in their teams, we encourage scientists to consider connecting with their local statistics department when designing, executing and describing statistical analyses.

In addition, we encourage users of statistical software to request clarification from tool developers and to critically consider limitations of a methodology in the scientific context. We believe that interdisciplinary dialogue and a culture of transparency will benefit microbiome scientists by supporting more efficient use of experimental resources and providing a stronger quantitative framework for knowledge accumulation.

Amy D. Willis ✉ **& David S. Clausen**
Department of Biostatistics, University of Washington, Seattle, WA, USA.
✉e-mail: adwillis@uw.edu

Published online: 21 February 2025

References

1. National Academies of Sciences, Engineering, and Medicine. *Reproducibility and Replicability in Science* 46–46 (National Academies, 2019).
2. Tavaré, S. in *Lectures on Mathematics in the Life Sciences* Vol. 17 (ed. Miura, R. M.) 57–86 (1986).
3. Robinson, L. D. & Jewell, N. P. *Int. Stat. Rev.* **59**, 227–240 (1991).
4. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* Ch. 7 (Chapman and Hall, 2024).
5. Hernán, M. A. & Monge, S. *BMJ* **381**, 1135 (2023).
6. Barrett, M., McGowan, L. D. & Gerke, T. *Causal Inference in R* Ch. 5 (GitHub, 2024).
7. Greenland, S. & Pearce, N. *Annu. Rev. Public Health* **36**, 89–108 (2015).
8. Hernán, M. A. *Am. J. Public Health* **108**, 616–619 (2018).
9. Martin, B. D., Witten, D. & Willis, A. D. *Ann. Appl. Stat.* **14**, 94–115 (2020).
10. Paulson, J. N., Stine, O. C., Bravo, H. C. & Pop, M. *Nat. Methods* **10**, 1200–1202 (2013).
11. Segata, N. et al. *Genome Biol.* **12**, R60 (2011).
12. Callahan, B. J., Sankaran, K., Fukuyama, J. A., McMurdie, P. J. & Holmes, S. P. *F1000Res.* **5**, 1492 (2016).
13. Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M. & Ellison, G. T. *Int. J. Epidemiol.* **45**, 1887–1894 (2017).

Acknowledgements

We gratefully acknowledge the very constructive feedback of S. Gibbons, T. Ye, A. M. Eren and three anonymous referees. A.D.W. was supported by the National Institute for General Medical Sciences award GM133420 and a Whiteley Center Scholarship.

Competing interests

The authors declare no competing interests.