

## Microbiome Scoping Review Data extraction form

*Adapted from an original Qualtrics survey*

1. Who extracted information from this article? *(select one)*
- YC
  - JD
  - ZL
  - YS

### ARTICLE INFORMATION

2. What is the confidence citation number? *(free text)*
3. What is the article DOI? *(free text)*
4. What is the article title? *(free text)*
5. Who were the authors of the article? *(free text)*
6. What was the primary purpose of the combined analysis? (Why did they combine data) *(select all)*
- Replication or validation of targeted findings from a single cohort
  - To describe the core microbiome of an environment. (The core microbiome is a shared set of bacteria found across all samples)
  - To compare synthesize across multiple studies to describe a population
  - The study described didn't actually involve combining multiple data sets from multiple prior publications or sequencing data sets
  - Other *(free text)*
  - Not described **[End Survey]**
7. What was the primary goal of this meta-analysis (what was the hypothesis they were testing)? *(select all)*
- To characterize a microbiome-outcome association
  - To characterize a microbiome-exposure association
  - To characterize assembly, development, or temporal variation in the microbial community
  - To characterize a unique microbiome (microbiome environment, population)
  - Other *(free text)*
  - Not described

### SYSTEMATIC APPROACH TO DATA COLLECTION

8. Was there a clear systematic approach for gathering reference data sets? (select one)

- No [Skip to 9-11]
- Yes [Answer 9-11]

[If 8 is Yes]

9. Was the search strategy for this meta-analysis described? (select one)

- No
- Yes

[If 8 is Yes]

10. Was a PRISMA diagram describing how papers were selected included in the paper? (select one)

- No
- Yes

[If 8 is Yes]

11. Was there a clear set of inclusion/exclusion criteria for the final studies? (select one)

- No
- Yes
- Unclear

**DATA SOURCES**

12. How many studies were analyzed in the primary meta analysis, after the data was processed? (If this is not listed, leave it blank) *(free text)*

13. How many samples were analyzed in the primary meta-analysis, after the data was processed? (If not listed, leave blank) *(free text)*

14. What are the sources for sequencing and metadata for data included in the primary meta-analysis? *(select all)*

- The authors of the meta-analysis or consortium
- SRA/GenBank
- ENA
- Request to the authors of the original data set
- MG-RAST
- QIIME DB or Qiita
- Other
- Not described

*(free text)*

15. Did the study include any independent validation cohorts? (not included in the original analysis) *(select one)*

- No
- Yes

**[Skip 16-18]**

**[Answer 16-18]**

**[If 15 is Yes]**

16. How many studies were analyzed for validation (final number after processing)? (If not known, leave blank) *(free text)*

**[If 15 is Yes]**

17. How many samples were analyzed for validation (final number after processing)? (if not listed, leave blank) *(free text)*

**[If 15 is Yes]**

18. What were the sources for sequences and metadata for the validation analysis? *(select all)*

- The authors of the meta-analysis or consortium
- SRA/GenBank
- ENA
- Request to the authors of the original data set
- MG-RAST
- QIIME DB or Qiita
- Other
- Not described

*(free text)*

**SAMPLING ENVIRONMENT**

19. Which environments were included? *(select all)*

- Built environment (i.e. house walls, office floor, etc)
- Host associated (non-human animals)
- Human-associated
- Other
- Not described

**[Answer 20, 21, 22]**

**[Answer 20]**

*(free text)*

**[If “Host associated (non-human animals)” or “Human-associated” are selected in 19]**

20. If human or non-human animals were included, which body sites were analyzed? *(select all)*

- Gut (Feces, rectal swab, biopsy)
- Oral
- Urogenital
- Skin
- Airway
- Other

*(free text)*

**[If “Host associated (non-human animals)” is selected in 19]**

21. What non-human animal species were included? *(free text)*

**[If “Host associated (non-human animals)” is selected in 19]**

22. If non-human animals were included, was this analysis a comparison between host species?  
*(select one)*

- No
- Maybe
- Ye

**STUDY DESCRIPTION**

23. Which of the following are reported for the studies that were combined? (select all)

- Population description
- Experimental design (randomization, variable matching)
- Sampling method(s)/collection kit(s)
- Collection kit(s) used
- Extraction kit(s) used
- Hypervariable region(s)
- Sequencing platform

[Answer 25]

[Answer 24]

[If “sequencing platform” was selected in 23]

24. Which sequencing platforms were used? (select all)

- 454 pyrosequencing
- Illumina (MiSeq, HiSeq, NovoSeq)
- Ion Torrent
- Pac Bio
- Oxford Nanopore
- Other

(free text)

[If “hypervariable region(s) was selected in 23]

25. Which hypervariable regions were included (select all)

- V12
- V13
- V2
- V23
- V3
- V34
- V35
- V4
- V45
- V46
- V68
- V69
- Ion torrent or multiple region kit (TSLR)
- Other

(free text)

**TABLE CONSTRUCTION**

26. Were all the data processed using a similar pipeline? (select one)
- Yes **[Continue to 27]**
  - No **[Skip to 33]**
  - Not described **[Skip to 33]**

**[If 26 is Yes]**

27. Did the authors perform taxonomic profiling and analyze the data? (i.e. denoise to ASVs, OTU clustering)? (select one)
- No
  - Yes

**[If 26 is Yes]**

28. Did the authors perform functional profiling? (i.e. tools like PICRUSt, Tax4Fun, mention of KEGGs in results) (select one)
- No
  - Yes

*If the authors performed both taxonomic profiling and functional profiling, please answer these questions [29 - 33] focused on the taxonomy profiling*

**[If 26 is Yes]**

29. Were sequences denoised? (This might be indicated by the used of DADA2, Deblur, Unoise3, or the description of features as ASVs)? (select one)
- No
  - Yes **[Skip 32; Answer 33]**
  - Unclear/not described

**[If 26 is Yes]**

30. Were the sequences clustered into operational taxonomic units (OTUs)? (This might be indicated by the use of algorithms like mothur, usearch, or vsearch or the mention of clustering) (select one)
- No
  - Yes **[skip 32; Answer 31]**
  - Unclear/not described

**[If 30 is Yes]**

31. What type of OTU clustering was performed and used for primary analysis (if multiple types of clustering were performed, or validation clustering was perform)? (select all)
- de novo clustering **[Answer 33]**
  - Open reference clustering **[Answer 33]**
  - Closed reference clustering
  - Unclear
  - Not described

**[If 29 is No and 30 is No]**

32. If the data was not denoised or clustered into OTUs, was the data collapsed to a taxonomic level without denoising or clustering? (select one)

- No
- Yes
- Unclear/not described

**[If 29 is Yes or 31 is “de novo clustering” or “open reference clustering”]**

33. If the authors performed denoising without clustering, open reference clustering or de novo clustering, how did they describe their phylogenetic tree construction? (select one)

- They did not construct a phylogenetic tree (no UniFrac or Faith's diversity)
- They used MAFFT, Fast Tree, or another de novo approach
- They used fragment insertion (SEPP, epa-ng)
- Unclear/not described

**[If 26 is Yes]**

34. Which tools were used to generate the feature table? (select all)

- DADA2 in R or QIIME 2
- Deblur
- Unoise
- Usearch
- Uclust
- Mothur
- CD-Hit
- SortMeRNA
- Sumacrust
- QIIME 1
- QIIME 2
- QIIME DB or Qiita
- vsearch
- naive bayesian classifier
- cutadapt
- clawback
- PICRUSt 1
- PICRUSt 2
- PICRUSt (version not listed)
- Tax4Fun
- Kraken
- Other

(free text)

35. Are there any other notes you'd like to add for the feature table generation? (free text)

**STATISTICAL ANALYSIS**

36. Which aspects of the microbiome were analyzed?

*(select all)*

- Descriptive analysis including taxonomy plots [Answer 37]
- Alpha diversity [Answer 39, 40]
- Beta diversity [Answer 41, 42]
- Differential abundance [Answer 43-46]
- Sample classification (i.e. random forest) [Answer 47-49]
- Core microbiome analysis (looking for a set of features common across all the samples)
- Co-occurrence networking
- Other *(free text)*
- Not described

**[If “Descriptive analysis” was selected in 36]**

37. Which taxonomic levels were used for descriptive analysis?

*(select all)*

- OTU/ASV
- Species
- Genus
- Family
- Order
- Class
- Phylum
- Not described

**[If “Alpha Diversity” was selected in 36]**

38. At which taxonomic levels was alpha diversity analyzed? (Faith's PD must be analyzed at the OTU/ASV level)

*(select all)*

- OTU/ASV
- Species
- Genus
- Family
- Order
- Class
- Phylum
- Not described



**[If “Alpha Diversity” was selected in 36]**

39. How did the authors handle differences between studies in their alpha diversity analyzes?  
(select all)

- Pooled all samples within a study
- Not adjustment for study effects (i.e. kruskal wallis testing)
- Adjusted for study as a fixed term (linear regression)
- Linear mixed effects model with study as a random effect
- Meta-analysis framework (i.e. forest plot, effect pooling)
- Not described
- Other

(free text)

**[If “Beta Diversity” was selected in 36]**

40. At which taxonomic levels was beta diversity analyzed? (UniFrac distance must be analyzed at the OTU/ASV level)  
(select all)

- OTU/ASV
- Species
- Genus
- Family
- Order
- Class
- Phylum
- Not described

**[If “Beta Diversity” was selected in 36]**

41. How did beta diversity analysis handle study effects?  
(select all)

- Descriptive PCoA that did not show study effects
- PCoA showing study effect
- Adonis permanova showing or adjusting for study effect size
- Other test adjusted for study effect
- Not adjustment or acknowledgment of study effect
- Other
- Not described

(free text)

**[If “Differential abundance” was selected in 36]**

42. At which taxonomic levels was differential abundance analyzed?  
(select all)

- OTU/ASV
- Species
- Genus
- Family
- Order
- Class
- Phylum
- Not described

**[If “Differential abundance” was selected in 36]**

43. What types of microbiome differential abundance testing was performed? (select all)

- Differential abundance of a targeted taxa (IE is F. nucleatum higher in all the studies)
- Untargeted differential abundance

**[If “Differential abundance” was selected in 36]**

44. How was the data filtered before differential abundance? (select all)

- The features were not filtered
- The filter was applied individually in each study
- The filter was applied across all studies
- the features were filtered based on prevalence
- The features were filtered based on abundance
- The features were filtered based on taxonomic assignment or other information

**[If “Differential abundance” was selected in 36]**

45. How were multiple studies handled in differential abundance testing? (select all)

- The studies were pooled with no effect adjustment
- Pooled analysis adjusted for study or technical effect (fixed effect)
- Pooled analysis with study as a random effect
- A comparison of results across individual cohorts (i.e. forest plot, effect pooling
- Other (free text)
- Not described

**[If “Differential abundance” was selected in 36]**

46. Which differential abundance algorithm(s) were used? (select all)

- t-test or ordinary least squares regression on rarefied data
- kruskal wallis test on rarefied data
- LefSe
- DeSeq2
- CornCob
- edgeR
- MaAslin
- metagenomSeq
- limma voom
- kruskal wallis, t-test, or other parametric model with on CLR-transformed data
- ALDEx2
- ANCOM I or ANCOM II
- ANCOM-BC
- PhILR or PhyloFactor
- Gneiss
- Differential ranking (songbird, Birdman, Bayesian DR)
- Other (free text)
- Not described

**[If “Sample classification (i.e. random forest)” was selected in 36]**47. What taxonomic level(s) were used to build a sample classifier? *(select all)*

- OTU/ASV
- Species
- Genus
- Family
- Order
- Class
- Phylum
- Not described

**[If “Sample classification (i.e. random forest)” was selected in 36]**48. How was the classifier trained? *(select one)*

- It was trained on a single study
- It was trained on multiple studies with no consideration of study effects
- Leave a study out or cross validation training
- Other
- Not described

*(free text)***[If “Sample classification (i.e. random forest)” was selected in 36]**49. Did the authors validate their classifier on other cohorts not originally used to train the classifier? *(select one)*

- No
- Yes
- Not describe

**OTHER NOTES**

50. Is there anything else you'd like to say about the study?

*(free text)*