
Zalewski S, Stewart CJ, Embleton ND, Berrington JE. [Brief guide to the analysis, interpretation, and presentation of microbiota data](#). *Archives of Disease in Childhood: Education and Practice Edition* (2017)

DOI link

<https://doi.org/10.1136/archdischild-2017-313838>

ePrints link

http://eprint.ncl.ac.uk/pub_details2.aspx?pub_id=244837

Date deposited

11/01/2018

Copyright

This article has been accepted for publication in *Archives of Disease in Childhood: Education and Practice Edition* following peer review. The definitive copyedited, typeset version is available online at: <https://doi.org/10.1136/archdischild-2017-313838>

A brief guide to the analysis, interpretation, and presentation of microbiota data

Dr Stefan Zalewski, Dr Christopher J. Stewart, Dr Nicholas D. Embleton, Dr Janet E. Berrington

Introduction to terminology and technologies

There has been an increase in research using new 'omic' technologies (1) (those allowing the study of a large biological dataset) designed to define and describe the microorganisms we carry, and their impact on health and disease. Omic technologies generate enormous quantities of complex data, so a major challenge is inter- and intra-study comparisons. This article provides an overview of terminology and data generation, and uses one of our study datasets to demonstrate different presentations of that data.

Bacteria are important in a range of physiological processes in humans (2): nutrient assimilation, vitamin production, modification of the nervous system (the gut-brain axis) and development of the immune system(3). Pathological changes in gut microbial communities (dysbiosis) have been associated with a wide range of diseases including skin and psychiatric disorders (4, 5), as well as diseases with high mortality in preterm infants such as infection and necrotising enterocolitis (NEC). NEC, for example, has been associated with reduced microbial diversity and increase in specific classes of bacteria, such as Gammaproteobacteria (6-8).

Identified by culture, bacteria were traditionally classified by physical characteristics into taxonomic ranks - phylum, class, order, family, genus and species. Sequencing based technologies rely on similarity in DNA sequences to determine organisms' phylogenetic relatedness to other species.

DNA Sequencing is a method for assessing microbial communities which works, either through identification of all genomes within a community (metagenomics), or using specific marker genes such as the 16S rRNA gene (9). The bacterial 16S rRNA gene can be used to group sequences by percentage similarity to each other, typically 'binning' (collecting together) all sequences with more than 97% similarity as a single operational taxonomic unit (OTU), which is then cross-referenced with databases to identify bacterial genus (or higher taxonomic levels if genus unavailable). The relatively short read length of 16S rRNA gene sequencing and homologous sequences between different genera prevent classification of OTUs to the species level. The cultivable fractions of clinical samples can be identified to species level by Sanger sequencing using the full length 16S rRNA gene. Isolates and clinical samples with single bacteria (e.g. blood infections) can also be sequenced by Sanger sequencing with amplicons as small as 320bps(10). The most comprehensive means of determining species level identification in complex clinical samples (e.g., stool) use metagenomic sequencing to computationally map random regions (i.e., not specific to the 16S rRNA gene) of DNA to known species.

Sequencing methods allow high-throughput untargeted detection of the microbial community(11) but data can be biased by methods of DNA extraction, primer selection in Polymerase chain reaction (PCR), sequencing technology, and bioinformatic processing (E.g. there are frequent annotation errors within sequencing databases such as BLAST(12)). Interpretation is further challenged when studies present results at different taxonomic levels (e.g. genus vs. phyla). Aggregation by species can miss differences in strains which could have functional importance. Associations of the relative abundance (%) of organisms also discount potentially discordant functional impacts of low abundant organisms and do not account for differences in absolute load.

Analysis and presentation of OTU data

Different numeric descriptive measures and graphical techniques can be used to illustrate key features of the microbiota that may be relevant to disease development as depicted below. We present an example of typical microbiome analyses using existing data (13-15) to show how different representations may affect interpretation. Details of the cohort are described previous work (13-15). Briefly, the dataset comprises 776 stool samples collected from 46 exclusively preterm infants (<32 weeks gestation). Of the 776 samples, 56 samples from 3 patients never received any expressed breast milk (EBM).

Numeric descriptors:

Diversity: Diversity relates to the number of different OTUs and their relative abundance, and is often a primary outcome in microbiome studies, although variably presented. Higher diversity is often associated with better health outcomes but not always (16). The number of distinct OTUs corresponds to richness and the relative abundance reads for each OTU allows a measure of diversity. Diversity is dictated by richness and evenness, where evenness considers if a single OTU dominates the sample (lower diversity) or no OTU is dominant (higher diversity). For example, two communities that contain only taxa A and B, one with 9 OTUs of taxa A and 1 of taxa B and the other with 5 of each, both have equal richness but the latter is more even. Diversity indices offer a combination measure of richness and evenness. The most commonly reported in neonatal studies is the Shannon diversity index (H) which describes the relative uncertainty of correctly predicting the next taxa, such that where one OTU dominates H approaches zero (reflecting low diversity), even if multiple other rare taxa are present. Shannon diversity typically ranges from 0 to 2 in neonatal studies (7, 17). Alpha diversity describes diversity in one site or sample i.e the total number of different OTUs and how abundant each of them are in relation to one another within one sample (Figure 1A and 1B). Beta diversity is a measure of how different sites or samples are (in terms of OTUs composition) compared to each other (Figure 1C) (18).

Graphical techniques:

Ordination

Ordination, for instance principal coordinate analysis (PCoA), places samples in ordination space based on their total microbiota profiles. If all OTUs from all samples were shared evenly between samples, these samples would sit at ordinates 0,0. Thus samples further from this point are more different from 'average' than those near to this point, and samples placed closer together are more similar to each other than those further away from each other. This is used to compare samples with different variables of interest, (e.g. exposure or health outcome). Figure 1C demonstrates differences in stool microbiota between breast-fed and formula-fed infants. The data can be combined to produce centroids (hypothetical 'average' sample points) for each group. Ordination axes show proportion of total variability accounted for by the ordination space components, and thus residual unexplained differences.

As well as studying the community as a whole, many studies seek to identify and compare specific taxa in association with health and disease states.

Heat Maps

Heat maps (Figure 1D) cluster samples and OTUs based on similarity within the dataset. Each column represents a different patient/sample with each row representing single taxa, allowing visualisation of samples that are similar in one or more key taxa, termed clusters, which can then be related to

exposure or outcome variables of interest. Image 1D shows the 10 most abundant genera in this dataset.

Stacked Bar Charts

Stacked bar charts show relative abundance of each taxa within a sample, and can be drawn at different taxonomic levels. They can identify change in relative abundance over time within an individual or between groups. In samples with high diversity, the charts often show family or order taxonomic level (anything lower is visually uninterpretable). Fig 1E shows our data at order and 1F at genus level for comparison. As shown, the data can be presented split by intervention or outcome measure to aid interpretation.

Associating patient data and taxonomic data

Ultimately analysis will seek to associate variables of clinical relevance, such as the development or protection from disease, with changes in bacterial taxa. Because many OTUs are detected, and thus many comparisons will be made for each OTU and each clinical variable, Statistics must be applied cautiously and it is important to adjust P values for multiple comparisons. For instance, if comparing 100 OTUs and using a P value <0.05 to define significance, simply by chance 5 OTUs will be reported as significant. Likewise, if analysing multiple samples from the same person then the statistics should account for repeated measures to avoid inflation of P values. Despite these statistical difficulties, appropriately analyses datasets can point to changes in key OTUs associated with disease or treatments.

Summary

Paediatric research increasingly focuses on, or includes data from, microbial analysis, but presentation is highly variable and can be challenging for clinicians. Understanding the key descriptive terms and visual representation measures is important in data interpretation and any potential clinical inferences.

References

1. Hasin Y, Seldin M, Lusi A. Multi-omics approaches to disease. *Genome Biology*. 2017;18(1):83.
2. Amon P, Sanderson I. What is the microbiome? *Archives of disease in childhood - Education & practice edition*. 2017.
3. Laukens D, Brinkman BM, Raes J, De Vos M, Vandenabeele P. Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *FEMS Microbiology Reviews*. 2016;40(1):117-32.
4. Rieder R, Wisniewski PJ, Alderman BL, Campbell SC. Microbes and mental health: A review. *Brain, Behavior, and Immunity*. 2017.
5. Penders J, Thijs C, van den Brandt PA, Kummeling I, Snijders B, Stelma F, et al. Gut microbiota composition and development of atopic manifestations in infancy: the KOALA Birth Cohort Study. *Gut*. 2007;56(5):661-7.
6. Stewart CJ, Nelson A, Scribbins D, Marrs ECL, Lanyon C, Perry JD, et al. Bacterial and fungal viability in the preterm gut: NEC and sepsis. *Archives of Disease in Childhood - Fetal and Neonatal Edition*. 2013;98(4):F298-F303.
7. Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Nelson A, Abdulkadir B, et al. Temporal bacterial and metabolic development of the preterm gut reveals specific signatures in health and disease. *Microbiome*. 2016;4(1):67.

8. Warner BB, Deych E, Zhou Y, Hall-Moore C, Weinstock GM, Sodergren E, et al. Gut bacteria dysbiosis and necrotising enterocolitis in very low birthweight infants: a prospective case-control study. *The Lancet*. 387(10031):1928-36.
9. Hilton SK, Castro-Nallar E, Pérez-Losada M, Toma I, McCaffrey TA, Hoffman EP, et al. Metataxonomic and Metagenomic Approaches vs. Culture-Based Techniques for Clinical Pathology. *Frontiers in Microbiology*. 2016;7:484.
10. Harris KA, Hartley JC. Development of broad-range 16S rDNA PCR for use in the routine diagnostic clinical microbiology service. *Journal of Medical Microbiology*. 2003;52(8):685-91.
11. Patel A, Harris KA, Fitzgerald F. What is broad-range 16S rDNA PCR? *Archives of disease in childhood - Education & practice edition*. 2017.
12. Klimke W, O'Donovan C, White O, Brister JR, Clark K, Fedorov B, et al. Solving the Problem: Genome Annotation Standards before the Data Deluge. *Standards in Genomic Sciences*. 2011;5(1):168-93.
13. Abdulkadir B, Nelson A, Skeath T, Marrs ECL, Perry JD, Cummings SP, et al. Routine Use of Probiotics in Preterm Infants: Longitudinal Impact on the Microbiome and Metabolome. *Neonatology*. 2016;109(4):239-47.
14. Stewart CJ, Embleton ND, Clements E, Luna PN, Smith DP, Fofanova TY, et al. Cesarean or Vaginal Birth Does Not Impact the Longitudinal Development of the Gut Microbiome in a Cohort of Exclusively Preterm Infants. *Frontiers in Microbiology*. 2017;8:1008.
15. Stewart CJ, Embleton ND, Marrs ECL, Smith DP, Fofanova T, Nelson A, et al. Longitudinal development of the gut microbiome and metabolome in preterm neonates with late onset sepsis and healthy controls. *Microbiome*. 2017;5(1):75.
16. Johnson KVA, Burnet PWJ. Microbiome: Should we diversify from diversity? *Gut Microbes*. 2016;7(6):455-8.
17. Morris EK, Caruso T, Buscot F, Fischer M, Hancock C, Maier TS, et al. Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. *Ecology and Evolution*. 2014;4(18):3514-24.
18. Kuczynski J, Lauber CL, Walters WA, Parfrey LW, Clemente JC, Gevers D, et al. Experimental and analytical tools for studying the human microbiome. *Nat Rev Genet*. 2012;13(1):47-58.