RESEARCH

A sample article title

Jane E Doe[??????], John RS Smith[??,??]

[??]Correspondence:
[??]Department of Zoology, Cambridge, Waterloo Road, London, UK
Full list of author information is available at the end of the article
[??]Equal contributor

Abstract

First part title: Text for this section.

Second part title: Text for this section.

Keywords: sample; article; author

Introduction

Taxonomic profiling using high throughput sequencing

Limitations of culturing techniques have prevented scientists from investigating the dynamics of highly complex microbial communities, especially human associated microbiomes. Advances in high-throughput sequencing have enabled the culture-free analysis of such communities, however sequencing data comes with additional statistical challenges.

One major difficulty of analyzing microbiome data is that it is strictly compositional [**?**]. This is because each sample has a different library size, induced through the PCR procedure embedded in short-read sequecning technologies. However, unlike RNAseq or even scRNAseq, microbiome data does not have "consistent features", such as UMIs or housekeeping genes, that can be used to estimate "size factors", allowing RNAseq-type data to break open the composition [**?**]. Even though methods have emerged in the scRNAseq domain addressing estimation of effective library size independent of persistent variables, the assumptions of these approaches have not been verified in the context of the microbiome. As such, microbiome data primarily exists in the form of relative abundances, where the principles of compositional data analysis (CoDA) applies [**?**].

Variable aggregation with microbiome data

However, microbiome data is also high dimensional. A common way to reduce this burden is to aggregate taxa, most naturally to higher Linnean taxonomic levels. This would reduce the number of hypotheses being tested, as well as improving interpretation. Currently, most microbiome studies have performed taxonomic aggregation through element-wise summation of the count vectors for all taxa assigned to the taxonomic rank of interest.

Prior to any downstream statistical analysis, these aggregated counts are then transformed back to compositional form. As such, we can define the sum-based taxonomic aggregation processs as simply the element-wise summation of the relative abundances. Let P_i be the relative abundance of higher taxonomic (HT) rank $\mathcal{P}$ in

sample i with raw counts x_{ij} where j is the column index of the lower taxonomic (LT) proportions. Let $\boldsymbol{P}$ be the set of column indices that belong to the HT rank $\mathcal{P}$ of interest. As such we have:

$$\mathrm{P}_i = \frac{\sum_{j \in \boldsymbol{P}} x_{ij}}{\sum_j x_{ij}} = \sum_{j \in \boldsymbol{P}} \frac{x_{ij}}{\sum_j x_{ij}} = \sum_{j \in \boldsymbol{P}} c_{ij} \quad (1)$$

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section . . .

Acknowledgements
Text for this section . . .

Figures

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Tables

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.