Exploring Pre-processing of data using phyloseq

Ben Temperton

1 Introduction

A recent paper was published on arXiv that called into question the use of rarefying amplicon sequence data down to the smallest dataset prior to analysis[1]. The argument against doing it is two-fold:

1.1 Loss of Statistical Power and increase of Type II errors

Imagine you have 16S rRNA amplicon libraries from a summer and a winter surface water community at BATS. You sequence them both on the same multiplexed 454 run. For the summer community you get 100 sequences back, and for the winter community you get 1000 sequences back. Now, imagine that these two communities contain two OTUs, OTU1 and OTU2 at different ratios in the summer and winter:

	Summer	Winter	Rarefied Winter
OTU1	62	500	50
OTU2	38	500	50

Table 1: Original Abundance and Rarefied abundance

Now, if we use a Chi-squared test to see if summer and winter are statistically different, for the original data we get:

	P-value
Original	0.029
Rarefied	0.117

Table 2: Statistical test of Sample difference - Original vs. Rarefied

So, by rarefying the data, the P-value has increased from 0.029 to 0.1171, and would thus not be considered statistically significant at a P-value cutoff of 0.05.

1.2 Overdispersion in rarefied datasets increases Type I errors

Count data, like that used in community analyses have interesting properties that are worth remembering:

- They are bounded by 0 (no negative counts)
- As the mean increases, the variance increases (a difference of 3 is big when the mean is 1, but not when the mean is 1000)
- Errors are not normally distributed

It is fairly common for this type of data to be *overdispersed*, meaning that the variance is larger than the mean. This can be due to missing model terms such as interactions, covariates etc. Remember that count data is typically modeled by a Poisson distribution, with a single parameter - the average density or arrival rate ($\lambda > 0$). Both the mean and the variance of a Poisson distribution are equal to $\lambda = rt$, where r is the density per sampling effort (how likely are you to find OTU1 if you have a single sequence), and t is the sampling effort.

This has two effects: Firstly, it not allow the variance to be adjusted independently of the mean (as both equal rt). Secondly, even when the population density is constant, you can change the poisson distribution of counts by sampling more extensively[2] - i.e. by collecting more 16S rRNA sequences for a given sample.

Overdispersion can also be caused by an abundance of 0 counts in the data, clustering of observations or correlation between observations[3].

2 Solving the problem with phyloseq

```
suppressMessages(library(phyloseq))
data(GlobalPatterns)
```

Global Patterns is a phyloseq object that contains the following:

- An OTU table (containing 19,216 taxa and 26 samples)
- Sample data (26 samples with 7 sample variables)
- A taxonomy table (19,216 taxa with 7 taxonomic ranks)
- A phylogenetic tree (19,216 tips)

References

[1] P. J. McMurdie and S. Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data is Inadmissible. *ArXiv e-prints*, October 2013.

[2] Benjamin M Bolker. *Ecological Models and Data in R.* Princeton University Press, Princeton and Oxford, 2007.

[3] Alain Zuur, Elena N. Ieno, Neil Walker, Anatoly A. Saveiliev, and Graham M. Smith. *Mixed Effects Models and Extensions in Ecology with R.* Springer, New York, 2009. ISBN 978-0-387-87457-9.