CoDa and sequencing

gg

19 July, 2016

To run this file: Rscript -e "rmarkdown::render('coda_seq.Rmd')

It is assumed that the output from a high-throughput sequencing experiment represents in some way the underlying abundance of the input DNA molecules. This is not necessarily the case as explained by the following thought experiment.

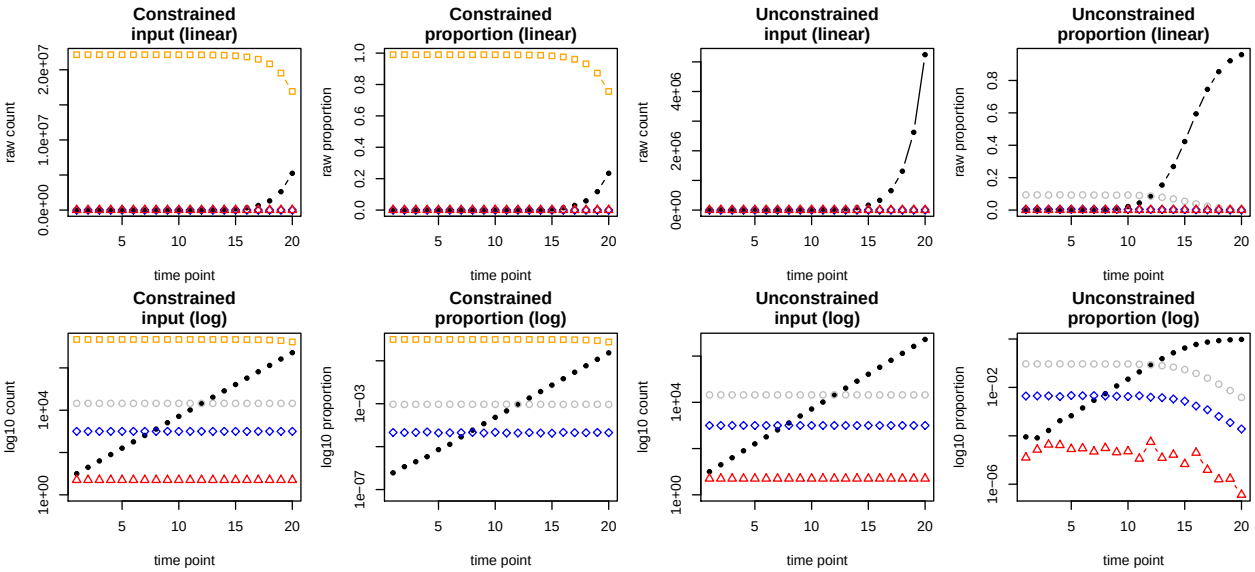

Figure 1: High-throughput sequencing affects the shape of the data differently on constrained and unconstrained data. The two left panels show the absolute number of reads in the input tube for 20 steps where the green and black OTUs are changing abundance by 2-fold each step. The gray, blue and red OTUs are held at a constant number in each step in both cases. The second column shows the output in proportions (or ppm, or FPKM) after random sampling to a constant sum, as occurs on the sequencer. The orange OTU in the constrained data set is much more abundant than any other, and is changing to maintain a constant number of input molecules. Samples in the two right columns are the same values plotted on a log scale on the Y-axis for convenience. Note how the constrained data is the same before and after sequencing while the unconstrained data is severely distorted.

Figure 1 shows two idealized experiments with four different ways of looking at the exact same data: the top row shows the data as linear points, the bottom row shows the same data after a log transform. The constrained input shows the case where the total count of all nucleic acid species in the input is constrained to a constant sum. The unconstrained input shows the case where the total sum is not constrained to a constant sum. These are modelled as a time series, but any process would produce the same results, and in practice we will likely only be comparing the first and last points (before and after some intervention). These are shown in two different ways. The first is linear counts "input", the second is proportions. The bottome row

Constrained datasets occur if the increase or decrease in any component is exactly compensated by the increase or decrease of one or more others. Here the total count remains constant across all experimental conditions. Examples of constrained datasets would include allele frequencies at a locus where the total has to be 1, and the RNA-seq where the induction of genes occurs in a steady- state cell culture. In this case, any process, such as sequencing that generates a proportion simply recapitulates the data with sampling

error. The unspoken assumption in most high throughput experimental designs is that this assumption is true— it is not!

An unconstrained dataset results if the total count is free to vary. Examples of unconstrained datasets would include ChIP-Seq, RNA-seq where we are examining two different conditions or cell populations, metagenomics, etc. Importantly, 16S rRNA gene sequencing analyses are almost always free to vary; that is, the total bacterial load is rarely constant in an environment. Thus, the unconstrained data type would be the predominant type of data that would be expected.

The relative abundance panels on the right side of Figure below shows the result of random sampling with a defined maximum value in these two types of datasets. This random sampling reflects the data that results from high throughput sequencing where the total number of reads is constrained by the instrument capacity. The data is represented as a proportion, but scales to parts per million or parts per billion without changing the shape. Here we see that the shape of the data after sequencing is very similar to the input data in the case of constrained, but is very different in the case of non-constrained data. In the unconstrained dataset, observe how the blue and red features appear to be constant over the first 10 time points, but then appear to decrease in abundance at later time points. Conversely, the black feature appears to increase linearly at early time points, but appears to become constant at late time points. Obviously, we would misinterpret what is happening if we compared early and late timepoints in the unconstrained dataset. It is also worth noting how the act of random sampling makes the proportional abundance of the rare OTU species uncertain in both the constrained and unconstrained data, but has little effect on the relative apparent effect on the relative abundance of OTUs with high counts.