Albertsen Lab

What is wrong with correlating relative abundance? Everything!

Posted on November 13, 2018 by Thomas Yssing Michaelsen

(This post is the first in a small series compiled during my visit to Segata Lab in Trento, Italy)

A provocative title I know, but as a young scientist venturing into the field of microbiology my concern about this is increasing to the point of now writing this blogpost. Sequencing data are inherently relative [Lovén et al., 2012; Gloor et al., 2017], because there is a maximum read capacity of every sequencing run. However, correlating relative data is like comparing bananas and apples – and although everyone does it – it makes little sense (In the vast majority of cases). Consider the plot from Vandeputte et al. [2017] shown below:

Recent posts

AR(10)E we there yet? September 2, 2019

EliteForsk scholarship – failure is a virtue March 4, 2019

Why is it important to remove short molecules? January 15, 2019

All I want for Christmas is a terabase of Nanopore data December 21, 2018

What is wrong with correlating relative abundance? Everything! November 13, 2018

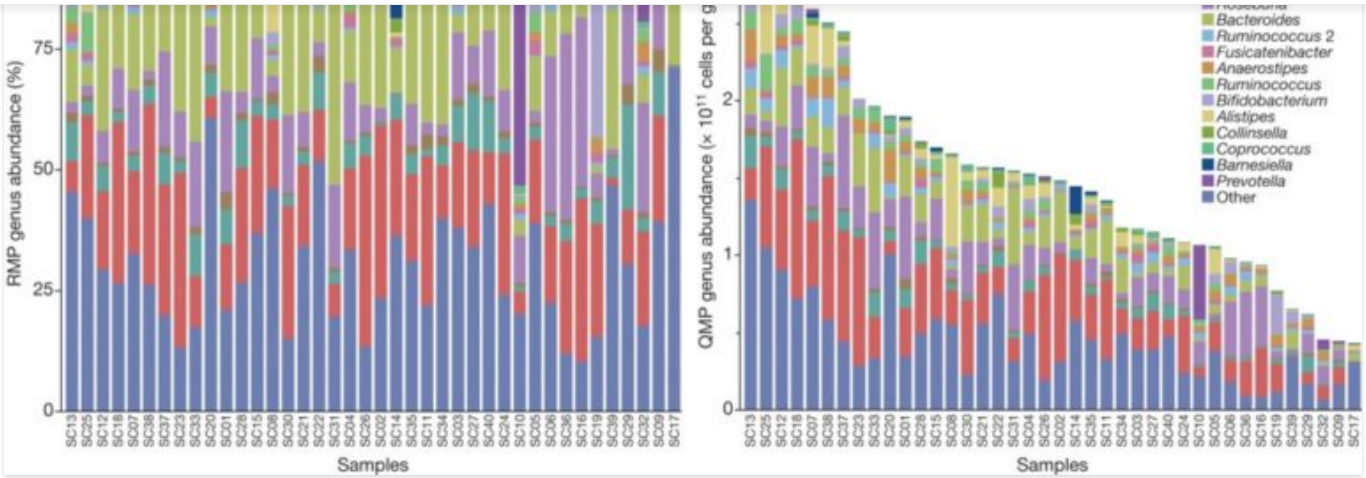

***Figure 1**: Figure 2 redrawn from Vandeputte et al. [2017]. The composition of 40 microbiomes was analysed using 16S amplicon sequencing, with colours representing different genera. Shown in (a) are the relative abundances which is quantitative output after sequencing. The authors simultanous quantified the cell numbers for each sample, which enabled them to rescale each sample according to the cell count and perform quantitative profiling as shown in (b).*

The authors did 16S sequencing and were simultaneously able to estimate the absolutte quantities of bacteria in their samples. This way they could compare the relative quantification (figure 1a) with the absolute quantification (figure 1b), which substantially influenced their results. By sequencing alone we only get the relative information – we don't know if there is a 6-fold range in the number of bacteria as in the study by Vandeputte and colleagues. Here most of the bacteria are reduced in absolute quantities from left to right (figure 1b), but the relative data shows bacteria that are constant, decreasing and increasing in all kinds of spurious relations (figure 1a).

I will give an intuitive argument about why correlation of relative data is problematic, very much inspired by this post. Let's start with the relative abundance of two bacteria measured by e.g 16S rRNA gene amplicon sequencing; B_1 and B_2. Since they are relative they have to sum to a constant:

$$B_1 + B_2 = C$$

It's easy to see here that knowing the relative abundance of one immediately tells you the relative abundance of the other. Whatever B_1 does, B_2 must go in the opposite direction with no room to vary: their correlation is a perfect -1. If we introduce a third bacteria B_3

shared between two bacteria, which means the expected correlation of any two bacteria will be -0.5. We can repeat this with any number of bacteria and find that as the number of bacteria goes up the effect of negative correlation is mitigated. Phew, we dodged the bullet!

But wait a minute…

This is not the end of it, by far. The above example is highly idealized and assumes all bacteria to be independent and identically distributed, which is not true in real data as figure 1
exemplifies. In reality, bacterial abundance exists on multiple levels, often with few dominating bacteria and a long tail of low-abundant ones. Additionally, intricate correlations between different bacteria are commonplace for biological reasons. Under these circumstances our idealized view simply does not cut it. Contrary, there seems to be a overrepresentation of negative correlations in real data [Lovell et al., 2015]. Perhaps more disturbing, the correlation of relative data tells nothing about the correlation of the underlying absolute values. This is perfectly illustrated in figure 2 from a very nice publication by Lovell et al. [2015]:

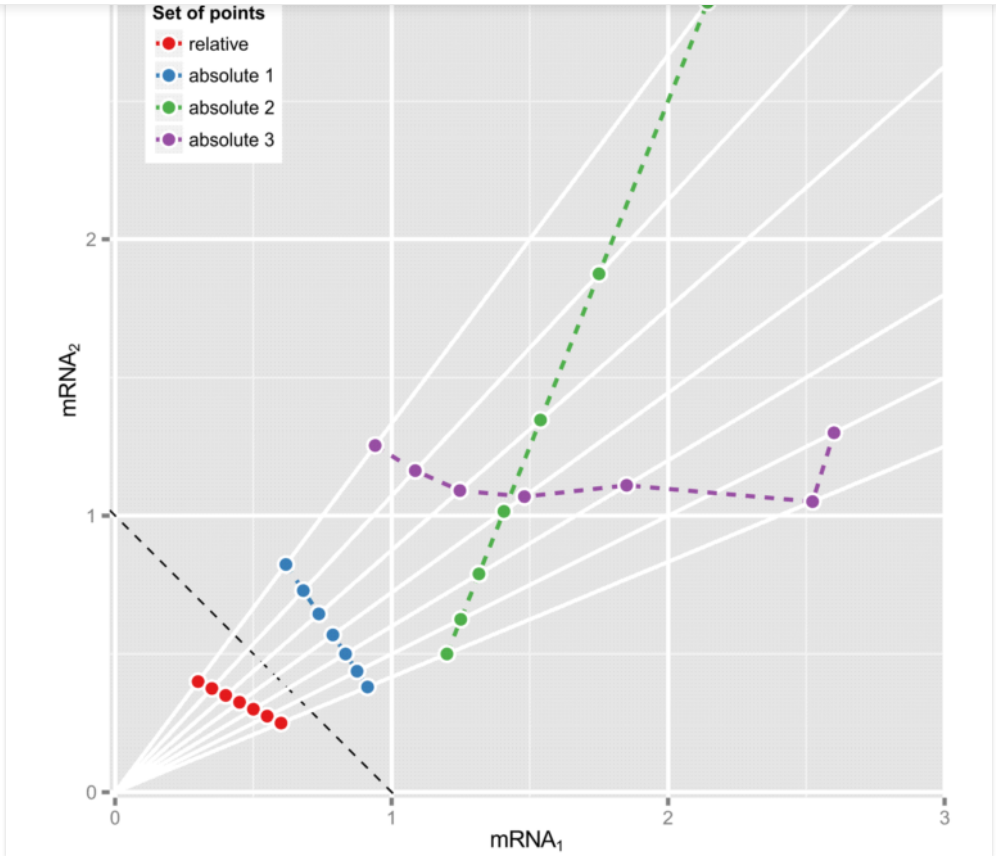

***Figure 2**: Figure 1A redrawn from Lovell et al. [2015]. The relative abundances for two hypothetical mRNAs across seven samples are shown in red in the lower left corner. The three sets of blue, green and purple points are absolute abundances that corresponds to the observed relative abundances with correlation coefficients of -1,1, and 0, respectively.*

The plot shows the relation between two mRNAs across seven samples. The lower-left corner with the red dots is what we can observe when looking at the relative values, constrained by the dotted black line because the relative values cannot sum to more than 1 (i.e. 100%). Each line extending from the origin and through a red dot is all combinations of $mRNA_1$ and $mRNA_2$ that could give rise to the observed relative abundance. For example, if we have $mRNA_1 = 5$ and

sample. If we then switch to the absolute values we get the main point: for each line we can put the real absolute value anywhere on the line we want, which is what Lovell and colleagues did for three sets of points (blue, green and purple). Thus, the same correlation of relative abundances can originate from arbitrary absolute correlations. So not only are the correlations inherently negative, they are also completely arbitrary! Check out the paper by Lovell and colleagues for all the nice details and their results when applying this on real data. In addition they also raise questions about the interpretation of differential expression analysis in the context of this problem.

But we are not all lost

I would argue that most scientist working with sequencing data know about this problem. But it is often circumvented by the ubiquitous assumption that the amount of input material is the same across all samples [Lovén et al., 2012; Lovell et al., 2015]. If that is true, the relative and absolute quantification are proportional across samples and we will not find discrepancies between the relative and absolute quantification. **This is a very, very strong assumption!** I will not go into details how this assumption can be violated, but many systems under study are constantly changing and heterogeneous. Given the complexity of microbial communities and their interactions this at least requires verification. And it is rarely done or required by the scientific community in general.

We can do other things with the data besides correlation. *Compositional data*, which is the statistical term for relative data, can be analysed by instead looking at ratios [Aitchison, 1982]. This is because ratios are invariant (i.e. unaffected) to scaling:

$$\frac{x}{y} = \frac{x/t}{y/t}$$

This is exactly what happens with compositional data, in the case of sequencing data t is the total abundance for the sample. A common starting point for analysis is to quantify how much this ratio varies across samples based on the logratio variance [Aitchison, 1982]:

$$LR_{var} = var(log(\frac{x}{y}))$$

The size of LR_{var} for x and y can be interpreted as the *proportionality* of x and y. A LR_{var} close to zero means that x and y are proportional, i.e. "synchronized" with a constant multiplication factor across samples. Contrary, large LR_{var} indicate that x and y have different multiplication factors across samples. This LR_{var} gives a great intuitive understanding of the concept, but is not directly used for calculating proportionality [Lovell et al., 2015; Quiin et al., 2017].

Implementation

The propr package [Quiin et al., 2017] implements the calculations of the ϕ-statistic. I've wrapped some of the functionality in in my own function `mt_phi` for streamlined processing, which is included in my mmtravis package for transcriptomics data analysis. This function allows you to input a $gene \times sample$ matrix of relative abundances and output a $gene \times gene$ matrix of the calculated ϕ-statistic for each gene pair. This matrix can then be used for downstream analysis, such as clustering analysis, heatmap visualization etc. Let's compare the proportionality and correlation matrices using some of my own transcriptomics data. You can find the code and data here. First we load the data:

```
library(mmtravis)
library(tidyverse)
library(magrittr)

mmt <- readRDS("mt_example.rds")
```

The data consists of some 2224 genes measured across 10 timepoints. Before we can analyse the data we need to transform it from raw counts into relative abundances. For this purpose I use transcripts per million (TPM) [Wagner et al., 2012]. Prior to this I also subset to genes with > 50 reads across all samples, to remove noisy genes.

```
mmt_sub <- mt_subset(mmt,minreads = 50,normalise = "TPM")
```

Now we can compute the ϕ-statistic and correlation matrices.

```
mmt_phi <- mt_phi(mmt_sub)
mmt_cor <- mmt_sub$mtdata %>% column_to_rownames("GeneID") %>% t()
%>% cor()
```

Let's convert the ϕ– and correlation-matrices into distance objects and do hierarchical clustering.

```
3.
4.  phi_clst <- hclust(phi_dist)
5.  cor_clst <- hclust(cor_dist)
```

Finally we can visualize the clusterings by drawing heatmaps shown in figure 3.

```
1.  par(mfrow = c(1,2),mar = c(1,1,1,1))
2.
3.  image(
4.    mmt_phi[phi_clst$order,phi_clst$order],
5.    breaks    = quantile(mmt_phi,seq(0,1,0.1)),
6.    col       = (heat.colors(10)),
7.    useRaster = TRUE,
8.    yaxt = "n",
9.    xaxt = "n",
10.   asp  = 1,
11.   bty  = "n",main = expression(phi~statistic))
12.
13. image(
14.   mmt_cor[cor_clst$order,cor_clst$order],
15.   breaks    = quantile(mmt_cor,seq(0,1,0.1)),
16.   col       = (heat.colors(10)),
17.   useRaster = TRUE,
18.   yaxt = "n",
19.   xaxt = "n",
20.   asp  = 1,
21.   bty  = "n",main = expression(Correlation~rho))
```

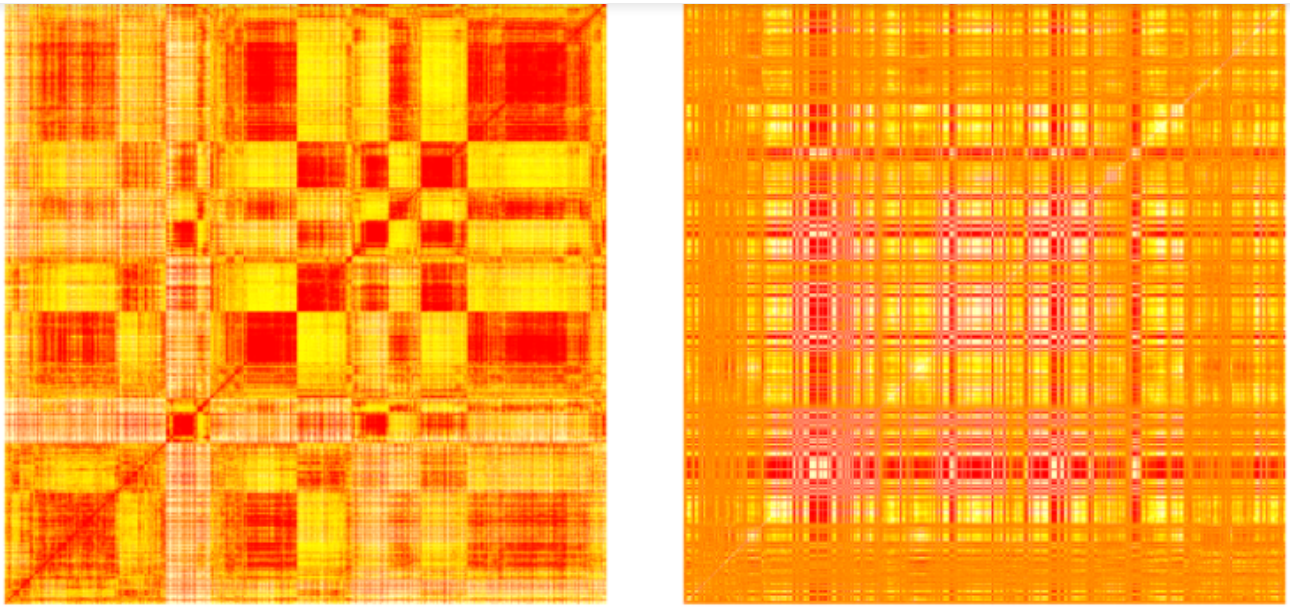

***Figure 3**: Clustered heatmap of all genes based on the proportionality calculated using the ϕ-statistic (left) and using simple correlation (right). Red indicates high proportionality or correlation.*

Note how different the two matrices in figure 3 are! The ϕ-statistic seems to yield tight clusteres of various sizes, while the correlation-based matrix is highly segmented into small clusters, with sets of genes that seem to be strongly correlated with almost all other genes, evident by all the vertical and horizontal red lines. This simple visualization is not enough for real conclusions, but the ϕ-matrix seems to have a more “biological” topology in concordance with what one may expect – cluster of genes of various sizes that are up- or downregulated together. In contrary, the clustering based on correlation seems more homogenous suggesting that noise from spurious correlations are dominating. From here the ϕ statistic could be used for further analysis, such as pathway enrichment, network analysis etc.

Conclusions

I have only scratched the surface on this topic and other approaches to handle issues with relative data exists [Pawlowsky-Glahn and Buccianti, 2011; Faust et al., 2012; Friedman and Alm, 2012; Knight et al., 2018]. The concepts

measures that in many cases are plain wrong, or just blindly trust our assumptions to hold true. Either you validate your assumptions to be true or you adapt your analysis to the constraints of the data. The former will always be problematic, therefore I believe proportionality provides a meaningful path forward analyzing sequencing data and compositional data in general.

References

Aitchison J (1982) The statistical analysis of compositional data. Chapman & Hall, Ltd.

Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, Huttenhower C. (2012). Microbial Co-occurrence Relationships in the Human Microbiome. PLoS Computational Biology 8.

Friedman J, Alm EJ (2012) Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology 8:1–11.

Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. Front. Microbiol. 8:2224.

Knight R, Vrbanac A, Taylor BC, Aksenov A, Callewaert C, Debelius J, Gonzalez A, et al. (2018) Best practices for analysing microbiomes. Nature Reviews Microbiology.

Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015) Proportionality: A Valid Alternative to Correlation for Relative Data. PLoS Computational Biology 11:1–12.

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA (2012) Revisiting global gene expression analysis. Cell 151:476–482.

Pawlowsky-Glahn V, Buccianti A (2011) Compositional Data Analysis: Theory and Applications. John Wiley & Sons.

Quinn TP, Richardson MF, Lovell D, Crowley TM (2017) propr: An R-package for identifying proportionally abundant features using compositional data analysis. Scientific reports 7(1)

Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, et al. (2017) Quantitative microbiome profiling links gut community variation to microbial load. Nature 551(7681):507-511.

Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory in Biosciences 131:281–285.

Bio Latest Posts

Thomas Yssing Michaelsen

PhD student

Wandering the greyzone between microbiology and biostatistics, where I develop novel methods to analyze “omics”-data from microbial ecosystems.

Posted in Data analysis, R.

← Article recap: method biases in…

All I want for Christmas… →

2 Comments

Jo Vandesompele
November 17, 2018 at 3:53 pm Reply

I wanted to point out that LR_{var} is equivalent to the pairwise variation V to find stable reference genes in the geNorm concept (Vandesompele et al., Genome Biology, 2002; equation 3). Indeed, stable reference genes are proportionally expressed.
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2002-3-7-research0034

Hej Thomas,
thanks for sharing your experience and creating this example. Just a quick note, that the mmtravis package requires the ampvis2 package.
Those who have not already obtained this package should add:
remotes::install_github(“MadsAlbertsen/ampvis2”)
to the next line after the installing command for the the remotes package.

Leave a Reply

Your email address will not be published. Required fields are marked *

Comment

Name *

Email *

Website

Post Comment

Center for Microbial Communities, Aalborg University, Denmark