Summary of results of the analysis of gut microbiome data

J. Ignacio Lucas Lledó

25/4/2020

Pre-processing

Sequence data from the gut microbiome came in two batches. In all there samples from 24 isolines, and two time points considered *early* and *late* in life ages. There are between 2 and 4 replicates of each isoline × age combination. Early and late samples are balanced between the two batches. Most isolines got all their samples sequenced in one of the two batches, except isolines 22, 23, and 24.

The sequencing center had already merged two forward and reverse reads from every sample, with overall statistics showing good quality data. The only issue with the data was an acute imbalance of sequencing effort among samples in the second sequencing batch. As a result, early samples from isolines 27 and 35 had a relatively small number of sequences (see figure 1).

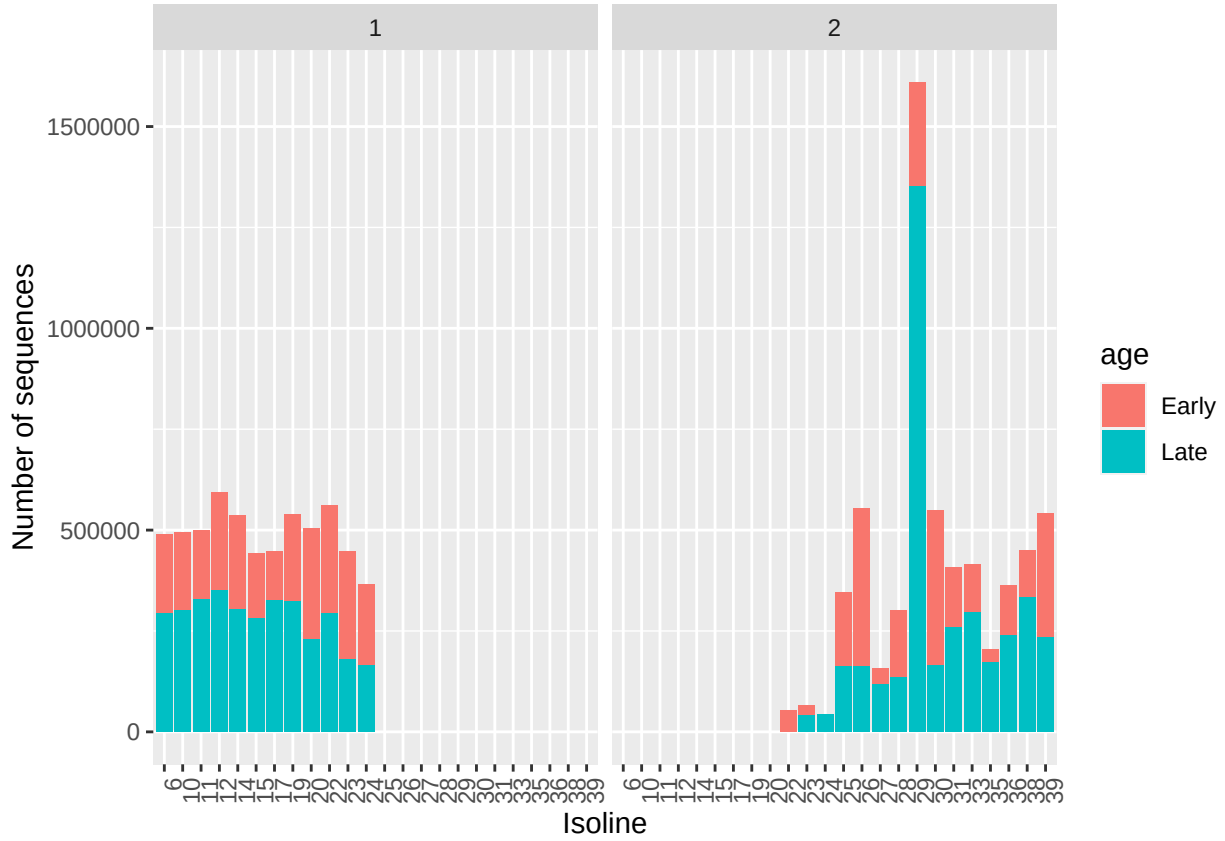

Figure 1: Sequencing effort per isoline and sequencing batch

I used the package `dada2` (Callahan et al. 2016) for *denoising*. This package implements algorithmns to

infer the exact sequences of different amplicons present in a sample, to single nucleotide precision, despite sequencing errors. Thus, there is no need to cluster similar sequences together in *operational taxonomic units* of doubtful functional homogeneity. Instead, the unit of microbiome composition is the *amplicon sequence variant* (ASV).

The length distribution of ASVs was quite wide, most of its width occupied by noise coming from spurious hybridization of sequencing primers with nuclear DNA. After some manual BLAST searches, I retained only ASVs of lengths between 432 and 464. In all, there were 2683 different ASVs left.

See full report here

Taxonomic attribution

The same package `dada2` allows for taxonomic attribution, based on public databases of RNA sequences with known taxonomy. I removed a few additional ASVs with missing high level taxonomic labels, or with spurious `Eurkaryotic` attribution (16S rRNA does not exist in Eukaryotes). Only 16 (0.006 %) of ASVs got the species assigned. But 0.96 got their genus assigned.

I did not check the quality of the taxonomic attribution. The whole report is available here.

Exploratory analysis

I run some exploratory analyses, including ordination plots, comparisons of diversity levels between early and late samples, or relationship between average proportion of an ASV in a sample and its prevalence among samples. These results can be checked here, and here.

I was lucky to find a copy of *Numerical ecology* (Legendre and Legendre 2012) avaiable online, and I took the chance to learn about ordination methods and multivariate statistics. I highly recommend the book. Know that it's the first time I use these methods, and I apologize if there are mistakes.

My conclusions on the exploration of data are the following:

- **There is a batch effect**. Some ordination methods sparate clearly samples sequenced in the first batch from those sequenced in the second. After talking with Zahida, I understand the sequencing run itself may not be the only source of variation, because samples sequenced later had also been processed later, with potentiall different room temperature, or whatever. Nothing to worry about, I think, but it's good to keep in mind.
- **Microbiome composition changes with age**. In later analyses (see below), it becomes clear that some measures of diversity (not all) are lower in late microbiomes.

Differential abbundance

To test for a difference in abundance between early and late samples requires specific packages, because we need to take into account the biological variation among replicates of the same isoline. There are thousands of ASVs and few replicates. The package I used is `DESeq2` (Love, Huber, and Anders 2014), originaly designed for RNA-seq data.

Keeping false discovery rate below 0.001, I find 219 ASVs with significantly different abundances between early and late samples. Figure 2 shows their fold change, and how they are distributed taxonomically.

Remarkably, ASVs from four different phyla and at least 15 genera reduce their abundances in late relative to early samples, while amplicons from only two genera increase their abundances. Some of these amplicons

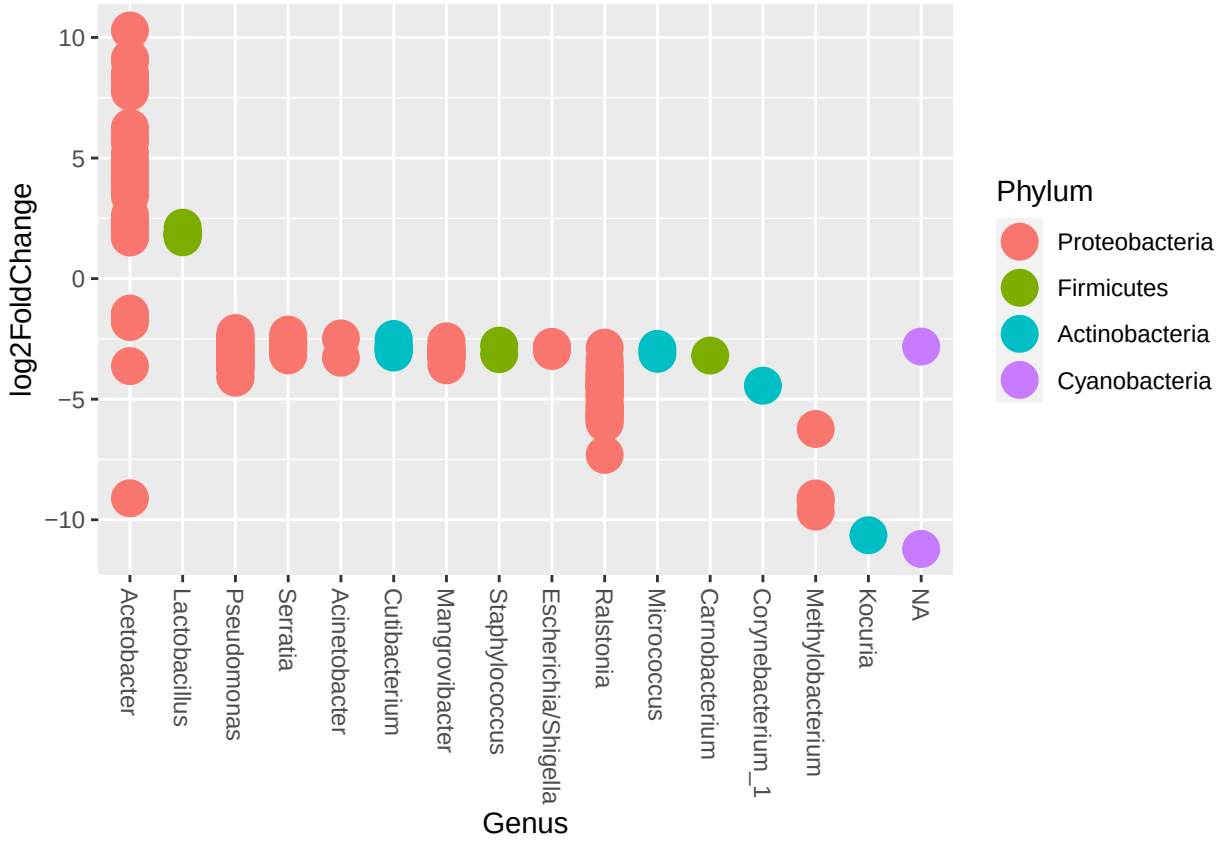

Figure 2: Fold change and taxonomic distribution of the amplicons with most differentiated abundances between early and late samples.

(maybe the Cyanobacteria and the Acetobacter with negative fold changes) could be false positives. But the plot suggests a reduction in diversity with age.

Redundancy analysis

The redundancy analysis was motivated by Pau Carazo's joint analysis of life history data and abundance summary data. The summary data included diversity measures and ordination vectors from a multidimensional scaling (MDS). The MDS was based on a binary (presence/absence) distance measure among samples. That distance was selected because it separated well early from late samples, as well as samples from first and second batches. Even though there were some interesting correlations, it was impossible to interpret the ordination axes in terms of ASVs, because only a distance matrix and not the original abundances entered the algorithm.

Redundancy analysis (RDA) seemed a good choice, because it is a multivariate method to explain the variance in a set of variables with another set of variables. But it has an important limitation: the number of explanatory variables cannot be larger than the number of observations. The high dimensionality problem is typical, and one way to deal with it is to run a PCA before the redundancy analysis (Song et al. 2016). Reducing the dimensionality of abundance data by selecting the first few principal components would make interpretation difficult, but not impossible.

The main issue with the RDA is that we are bound to find false positives if we do not test the global fit before believing the correlations that the RDA finds (Blanchet, Legendre, and Borcard 2008). My first attempts were naive: the abundance data, early and late in life, have a lot of variation that we do not expect to be correlated with life-history traits variation. The RDA looks for the axes of variation that maximize correlation, and it should not surprise us that it always finds something.

In the most recent attempts, I tried to use more meaningful predictable variables. Zahida told me to use diversity measures, which I had forgotten before. I also summarized the amplicon abundances by genus, and even by class. I focused on the amplicons (and genera) with significantly different abundances between early and late. I used abundances only early, or only late... Nothing that I tried produced a significant global fit. The lowest p values I've seen are >0.11. You can see here the full report of the latest attempt.

References

Blanchet, F.G., P. Legendre, and D. Borcard. 2008. "Forward Selection of Explanatory Variables." *Ecology* 89 (9): 2623–32. https://doi.org/10.1890/07-0986.1.

Callahan, B.J., P.G. McMurdie, M.J. Rosen, A.W. Han, A.J.A. Johnson, and S.P. Holmes. 2016. "DADA2: High-Resolution Sample Inference from Illumina Amplicon Data." *Nature Methods* 13: 581–83. https://doi.org/10.1038/nmeth.3869.

Legendre, Pierre, and Louis Legendre. 2012. *Numerical Ecology.* Amsterdam: Elsevier.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15: 550. https://doi.org/https://doi.org/10.1186/s13059-014-0550-8.

Song, Y., P.J. Schreier, D. Ramírez, and T. Hasija. 2016. "Canonical Correlation Analysis of High-Dimensional Data with Very Small Sample Support." *Signal Processing* 128: 449–58. https://doi.org/10.1016/j.sigpro.2016.05.020.