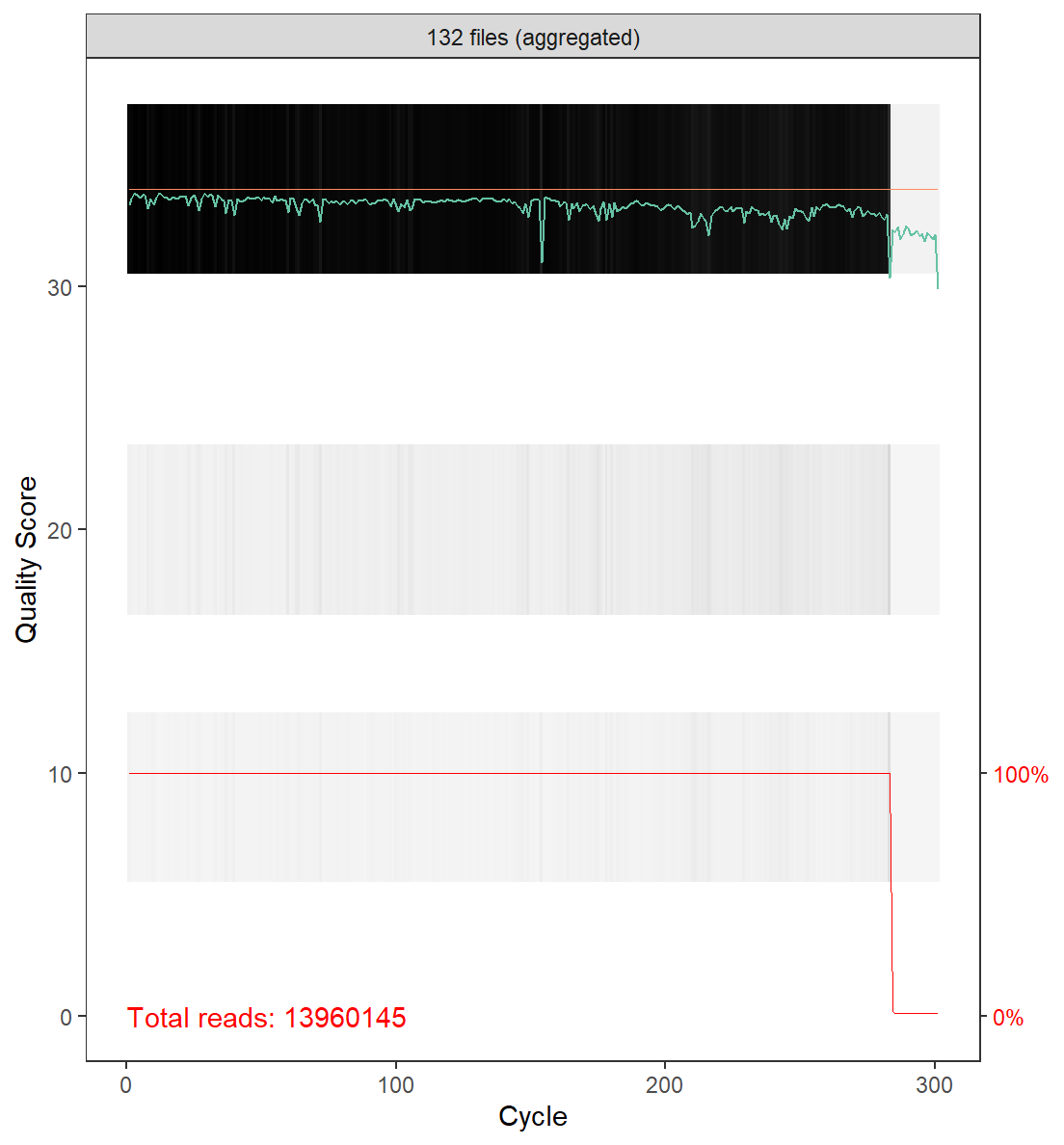
This document describes attempts at improving the DADA2/phyloseq pipeline with the microbiota 17 dataset:

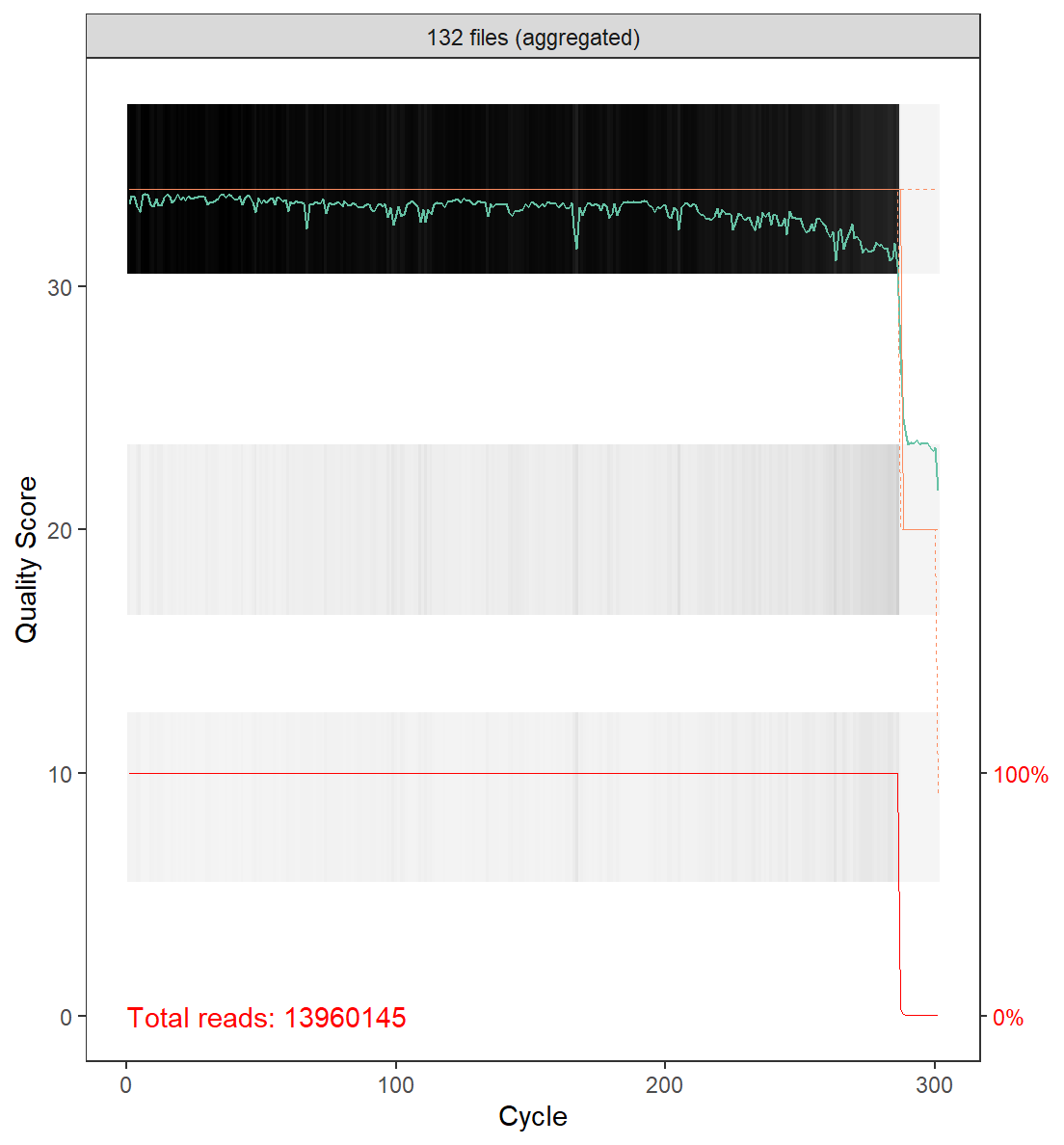
New addition: using the sink function to save the output R console text on a text file

Plot quality profiles: I tried to put aggregate = TRUE to print a single quality profile

It worked out:

For R1 files:

Quality starts dipping at around 283 nts



I should trimmed shorter on R2

A black text on a white background

Description automatically generated

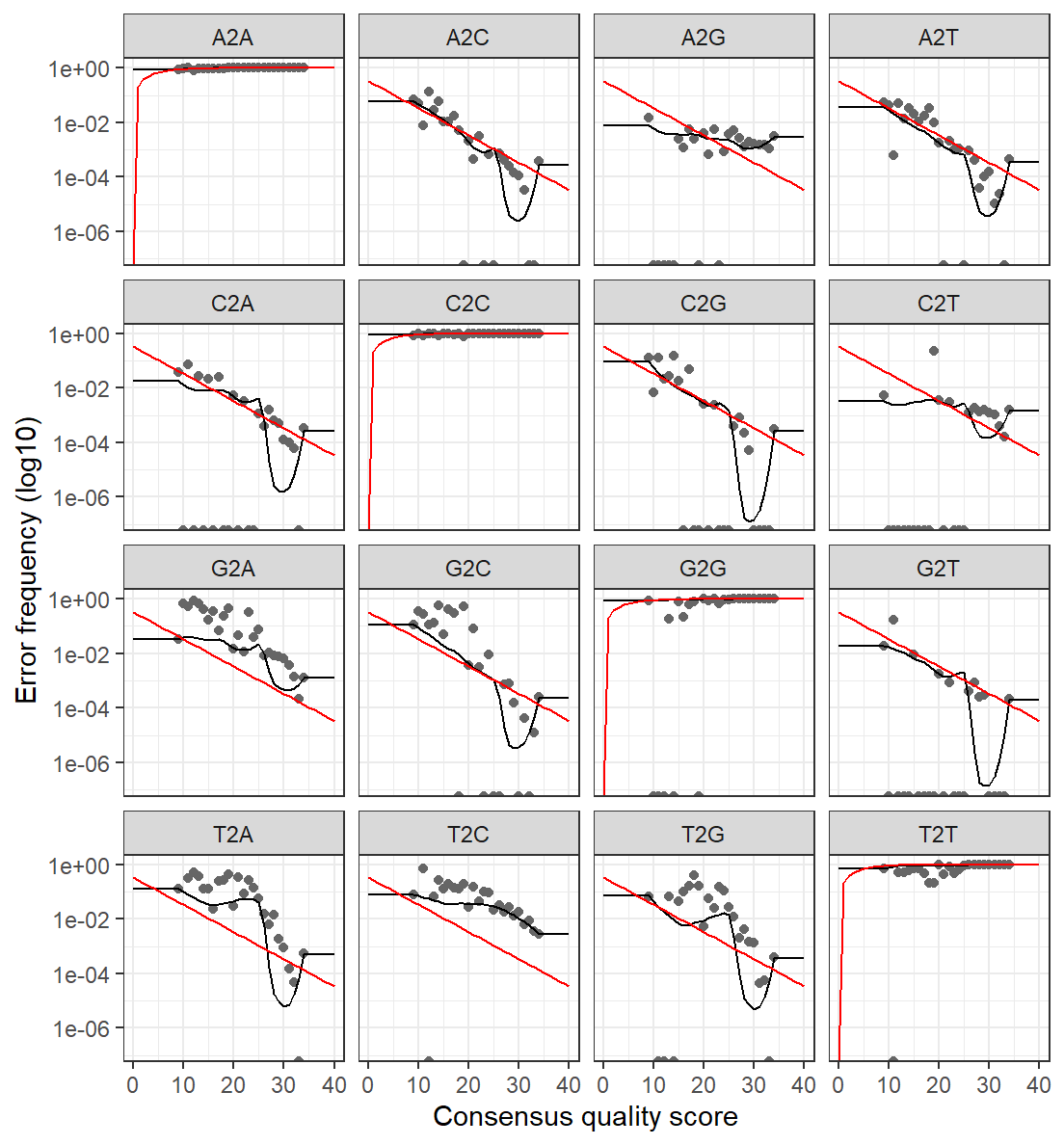
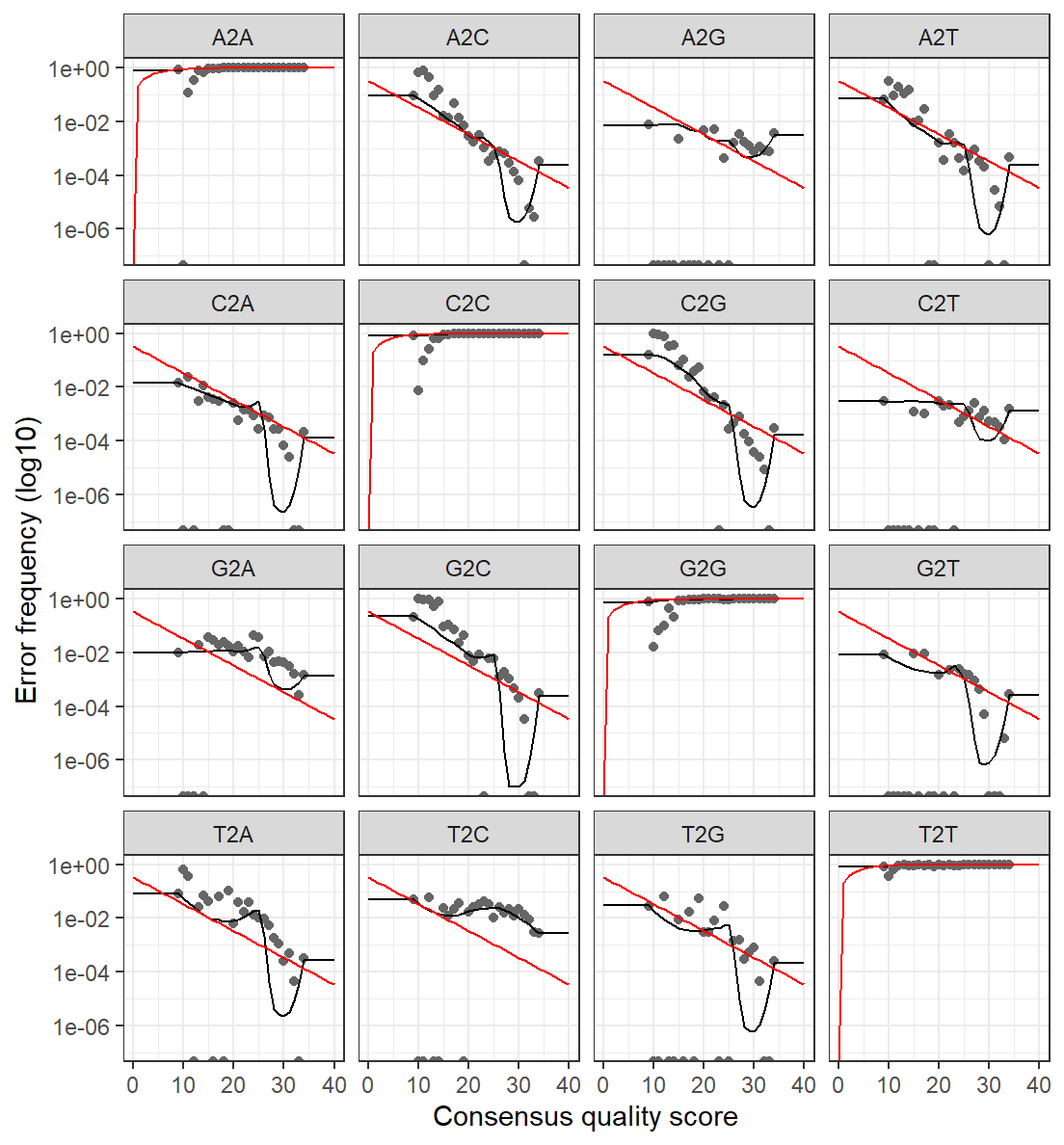
V5-V6 region was amplified, which accounts for an expected amplicon size of 300-350 (source ChatGPT 💀)

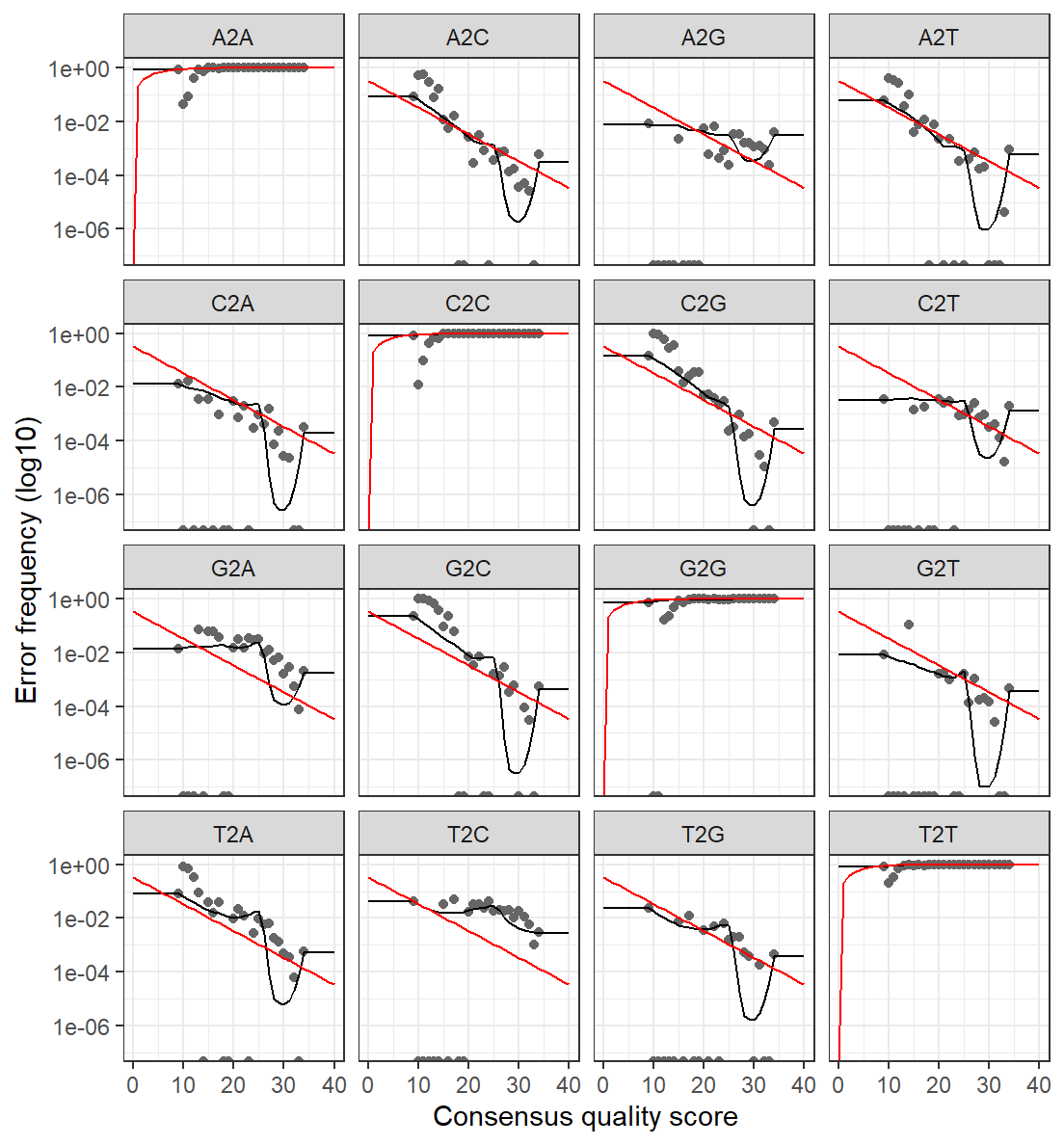
Filter and trim: Choose trimming parameters based on the quality profile  
-I chose (275,270)  
-Results are pretty good, with average number of reads that passed of 85%

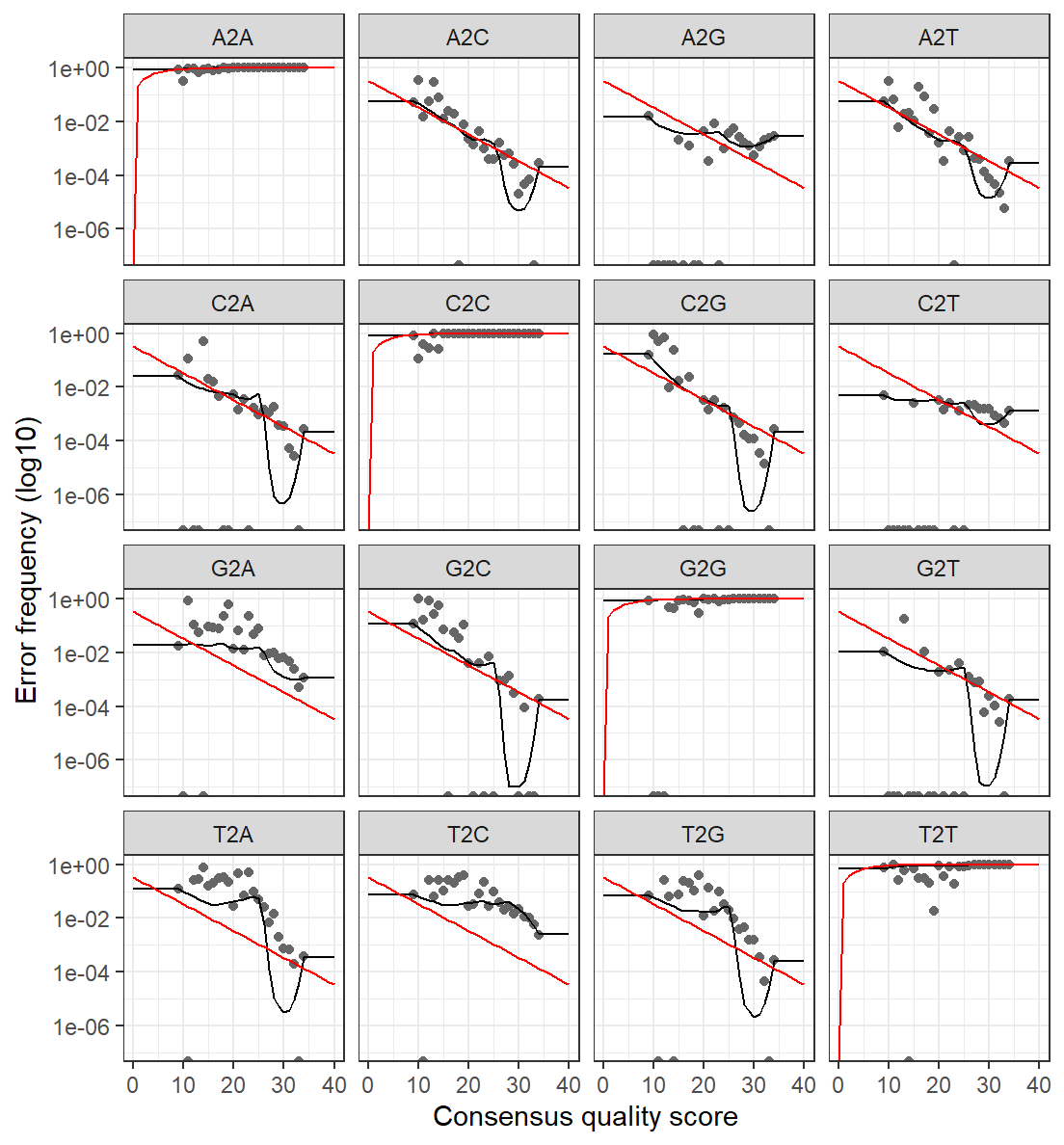
Derep: At this point we can reduce the dataset size by using the derep function

Learning error rates: You can learn the error rates with derep class objects

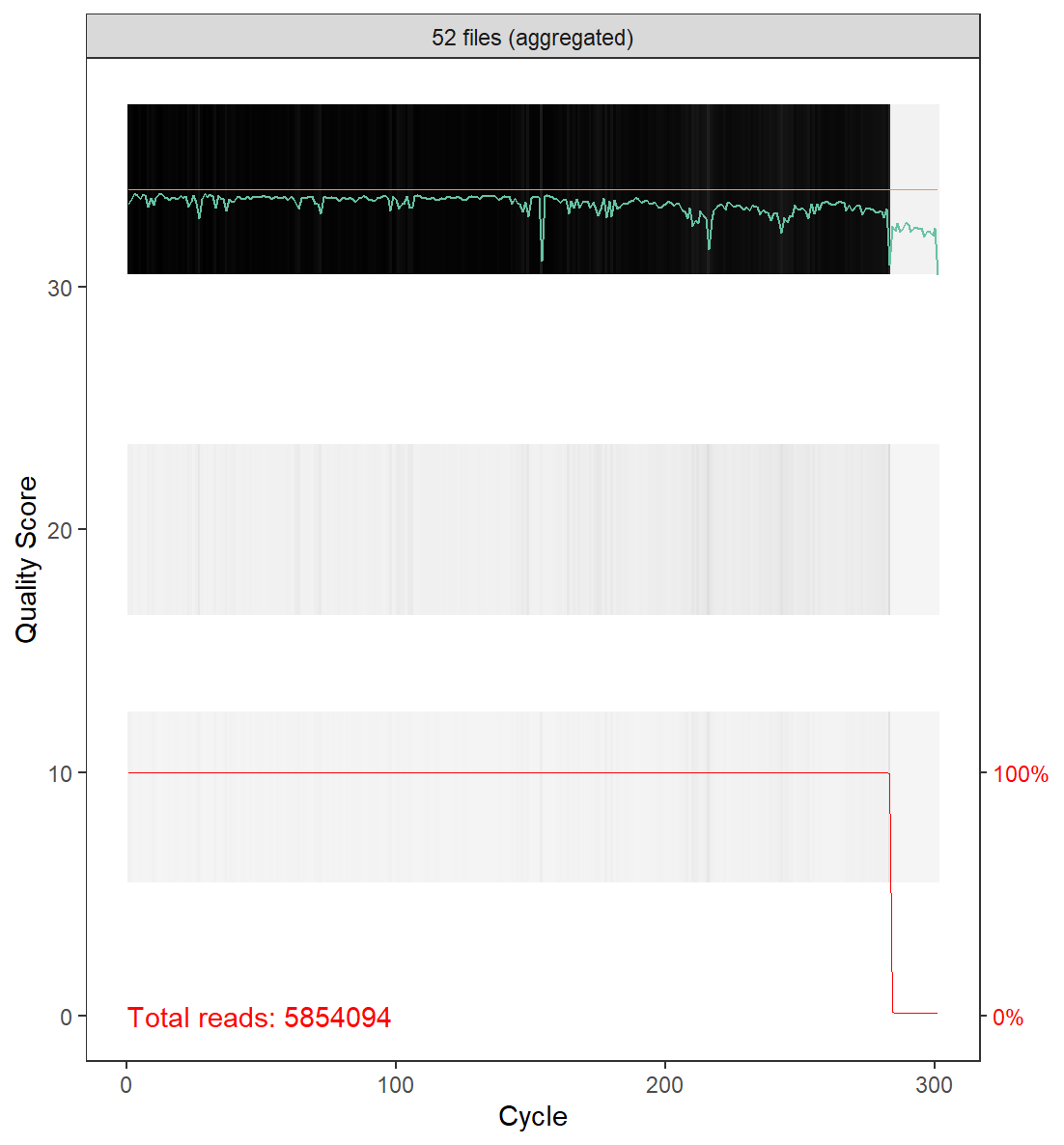
But doing the derep step before the learning error rate stuff gives that:

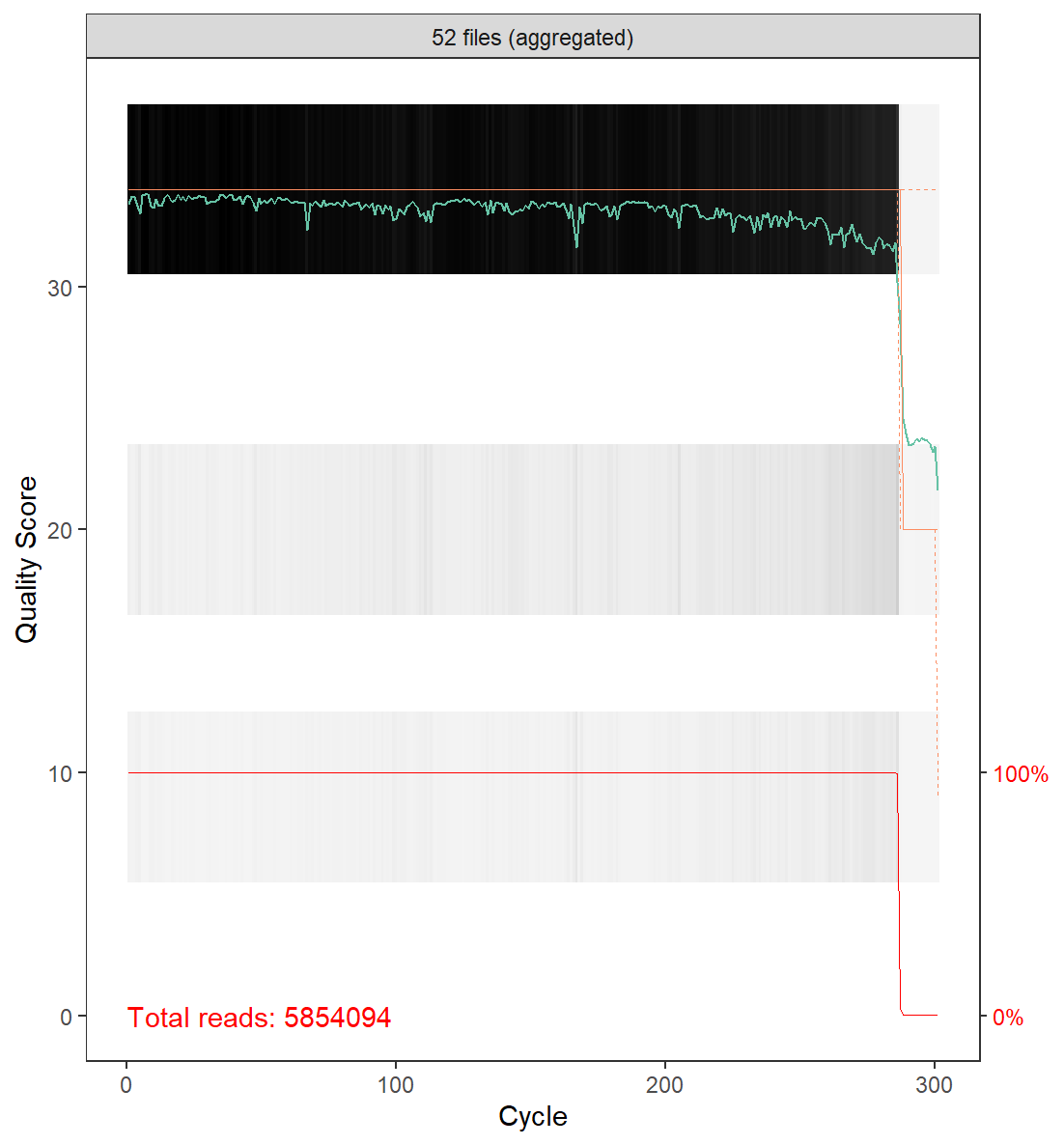
R1 R2

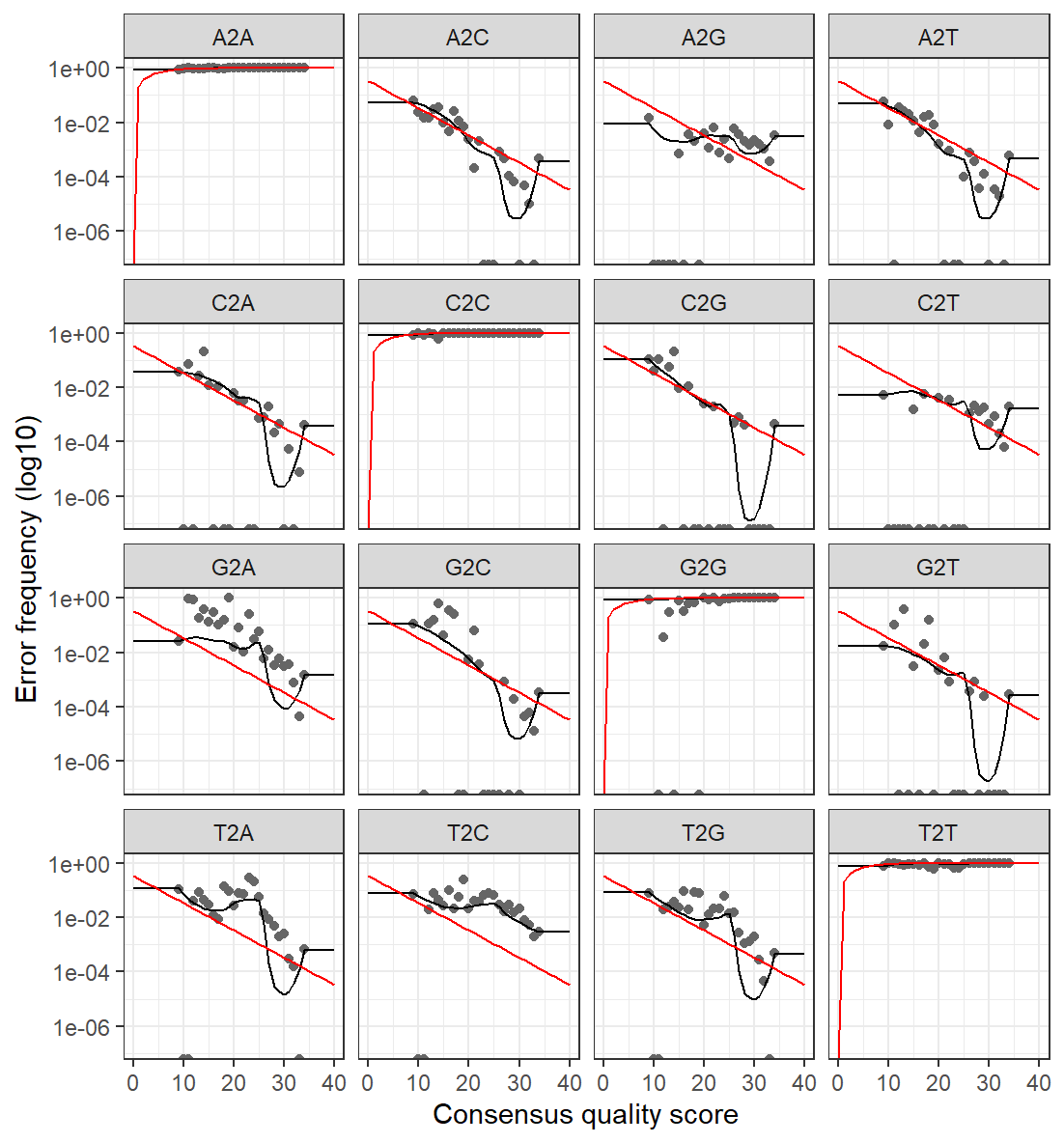
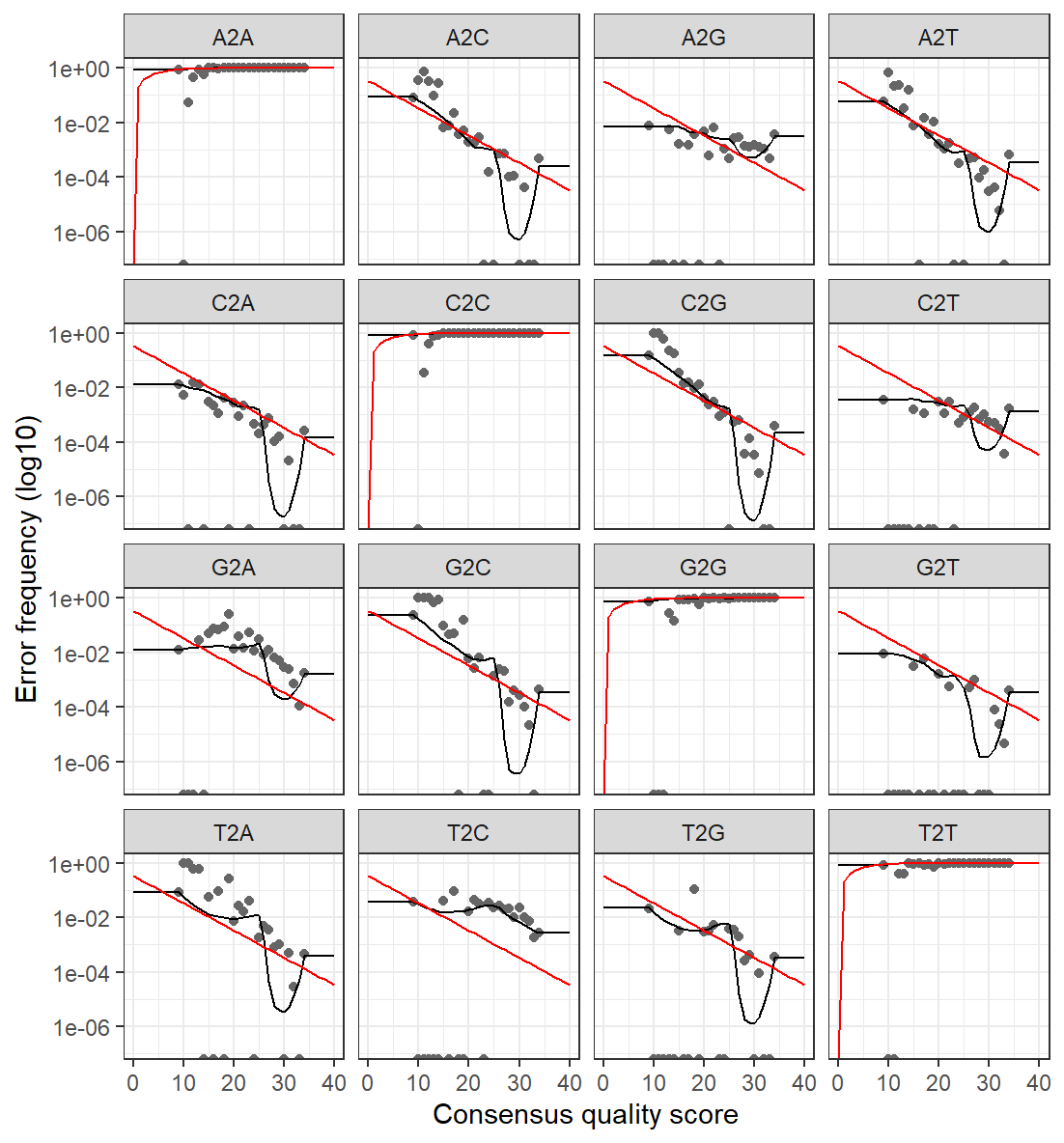
So that does not seem really good, let’s perform the learning error rate before applying the derep function



We have the same issue, so probably that we should think about separating both datasets (different genetic backgrounds, and completely different experiments)

We start with Samuel dataset:



We are facing similar issues, need to figure out why

Apparently the red line is the mean, and the black the model, as long as the points do correspond to the model and that error rates seem to be decreasing as the quality score goes up, things should be fine for what comes after

DADA function:

Thinking about pooling: strategy for denoising and identifying the unique sequence variants  
-Full pooling: good to identify rare sequences variants (data from all samples is combined) = good for low read depth BUT computationally intensive  
-Pseudo-pooling: still more computationally efficient than full pooling, it’s pretty good when computational resources are limited and that you’re still looking forward to identifying rare sequence variants  
-No pooling: each sample processed independently

Considering for the microbiota17 dataset we are at 90k reads, that’s a moderate depth, so pseudo-pooling seems appropriate.

* Combining these things + installing ubuntu on “the beast” greatly improved the time to generate ASVs as well as taxonomies

Phyloseq: what metrics should we focus on and what kind of graphs should we generate

Alpha diversity: observed richness (number of taxa) or evenness relative abundances of those taxa  
Beta-diversity: variability in community composition among samples within a habitat

While alpha diversity is **a measure of microbiome diversity applicable to a single sample**, beta diversity is a measure of the similarity or dissimilarity of two communities. As for alpha diversity, many indices exist, each reflecting different aspects of community heterogeneity.

 **Alpha diversity** looks at how diverse one sample is by itself.

 **Beta diversity** looks at how two samples compare in terms of diversity.

For alpha diversity there is:

-Shannon: measures both richness and evenness  
-Simpson: measures probability that two individuals randomly selected from sample belong to same species => focuses on species dominance   
-Chao1: estimates total species richness (including rare species) in a sample  
-OTUs count: straightforward measure of richness = **does it work with ASVs?**

For beta diversity there is:

Bray-Curtis: compares difference in species composition between two samples  
good for comparing two microbiota communities   
Jaccard Index: similarity between samples based ONLY on presence or absence of species  
UniFrac: phylogenetic distance between microbial communities  
-Unweighted: focus on presence or absence of phylogenetic lineages   
-Weighted: considers both presence/absence and relative abundance of phylogenetic lineages

Now knowing all of that, let’s take Thibault’s antibiotics paper as an example and check what kind of analyzes he performed:

Alpha: Shannon and chao1  
Beta: weighted UniFrac and associated PCoA = levels of differences between experimental groups

Statistical significance of weighted UniFrac distance between groups was performed using Adonis function from vegan R package (version 2.5)

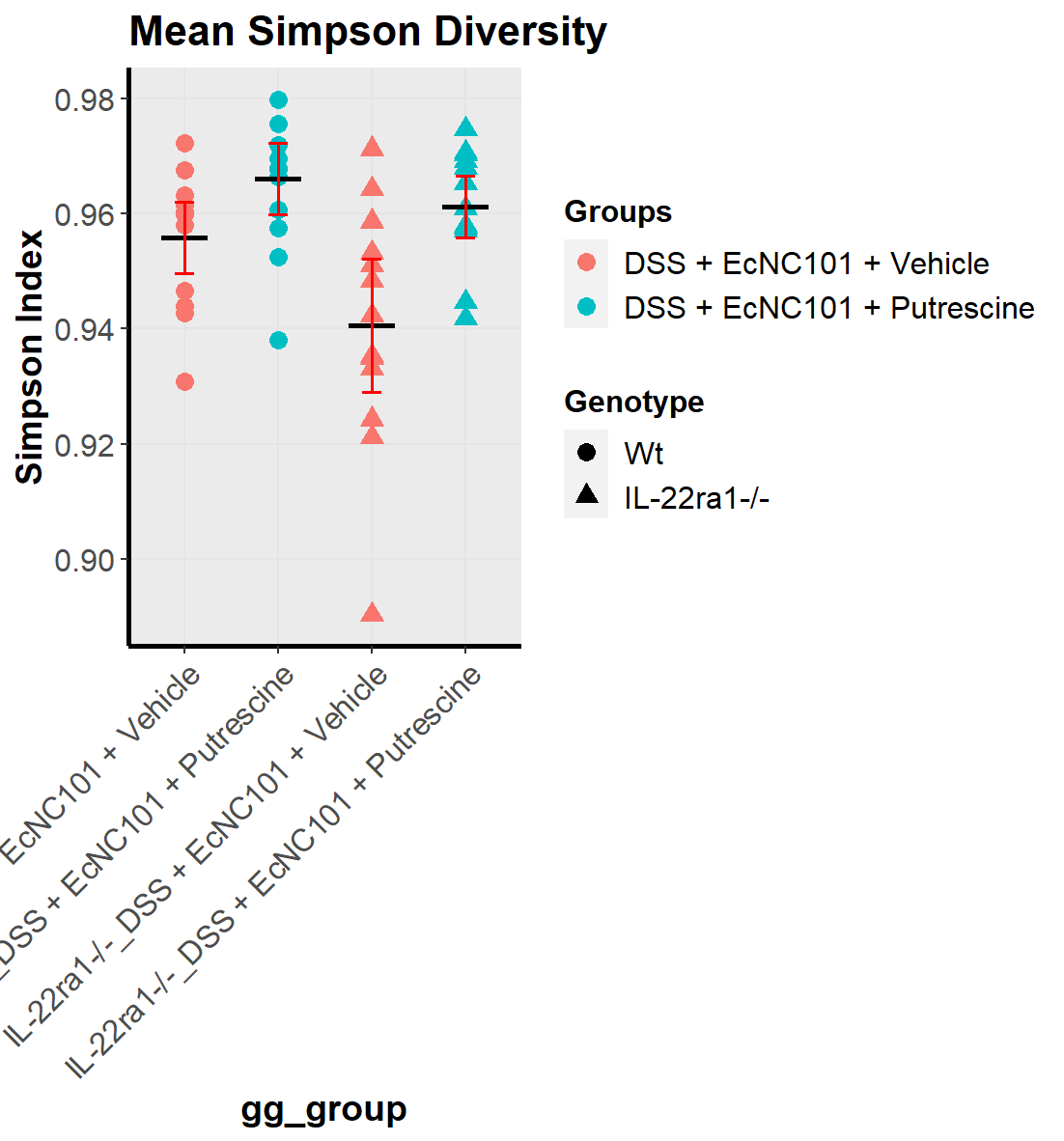
Bacterial relative abundance at the phylum and family levels

“To detect differentially abundant bacterial genera and species induced by the iron-enriched diet, we applied Generalized Additive Models for Location, Scale and Shape (GAMLSS) with a zero-inflated beta (BEZI) family (GAMLSS-BEZI).” = **focus on this cause it’s gonna suck ass**

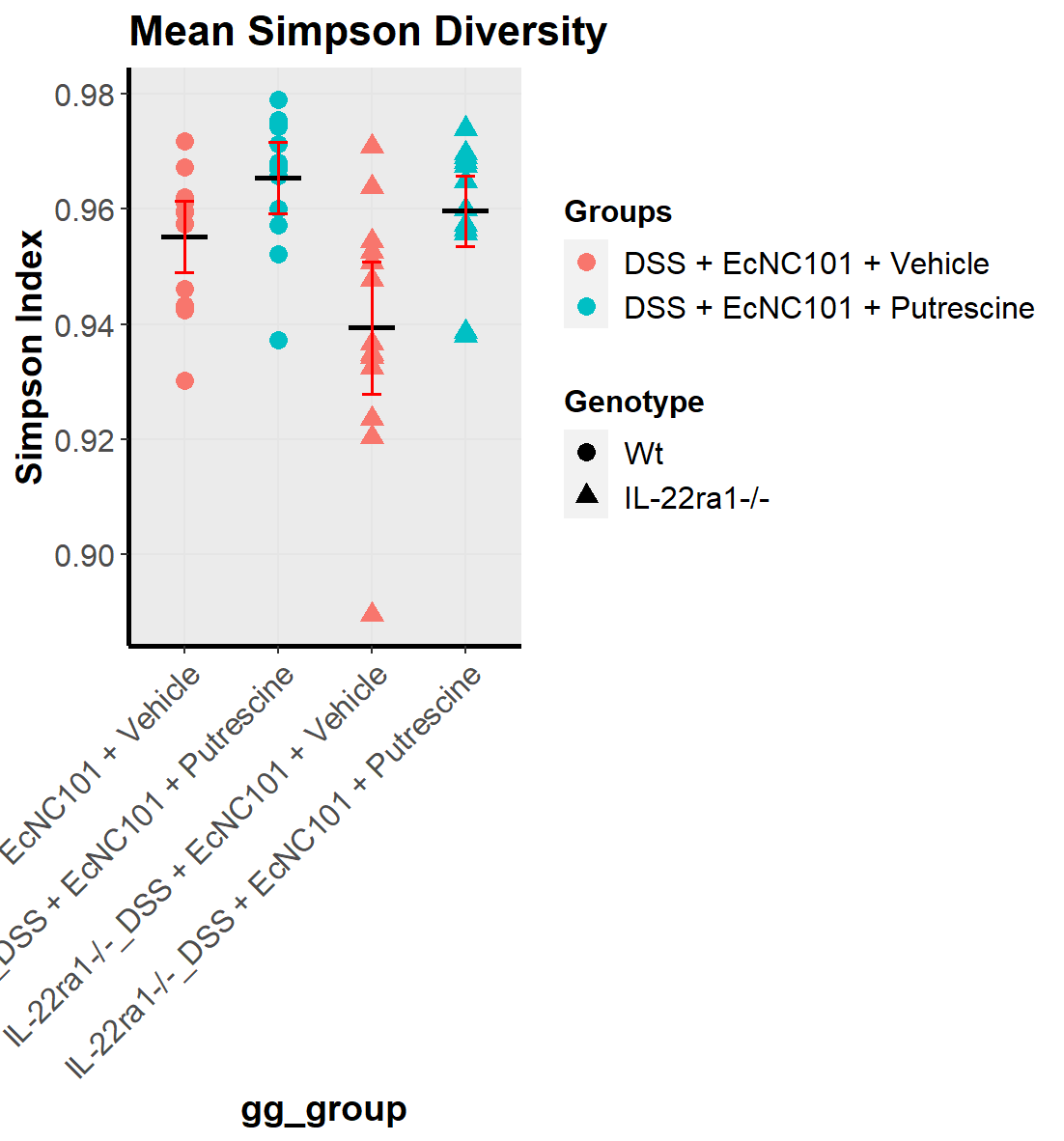
Apparently it’s an approach of statistical modelling = semiparametric regression = supervised machine learning

https://pubmed.ncbi.nlm.nih.gov/30991942/

So apparently it’s better to remove low frequency ASVs as it reduces the noise?  
*Edit* I added a bit of code to that: starting with a 10% minimum threshold



Mean Simpson diversity before doing the filtering



There is some slight differences

But code is working and is fairly easy to use so 👌

Now testing metamicrobiomeR

DESeq2 is the shit to go for statistics and stuff but we need to understand its functioning.

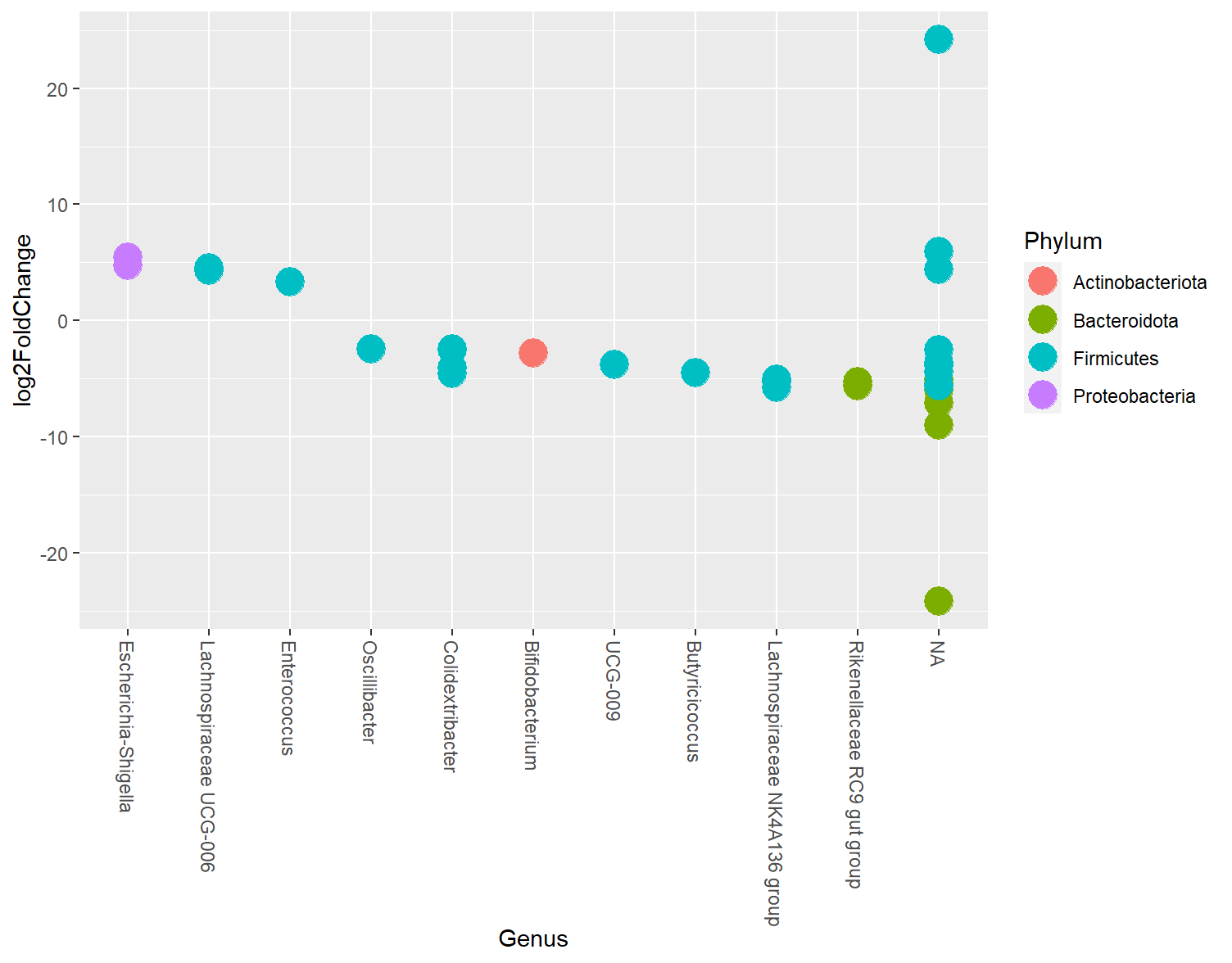
Its description says: **Differential gene expression analysis based on the negative binomial distribution**

Estimate variance-mean dependence in count data from high-throughput sequencing assays and test for differential expression based on a model using the negative binomial distribution.

It’s supposed to correct for depth of sequencing (i.e. number of reads per sample)

<https://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0550-8>

Like other techniques (such ANCOM, ALDEX etc), it is meant for differential abundance analysis which aims to find the differences in the abundance of each taxa between two classes of subjects or samples, assigning a significance value to each comparison.



I have no idea what that shit means, we have work to do…