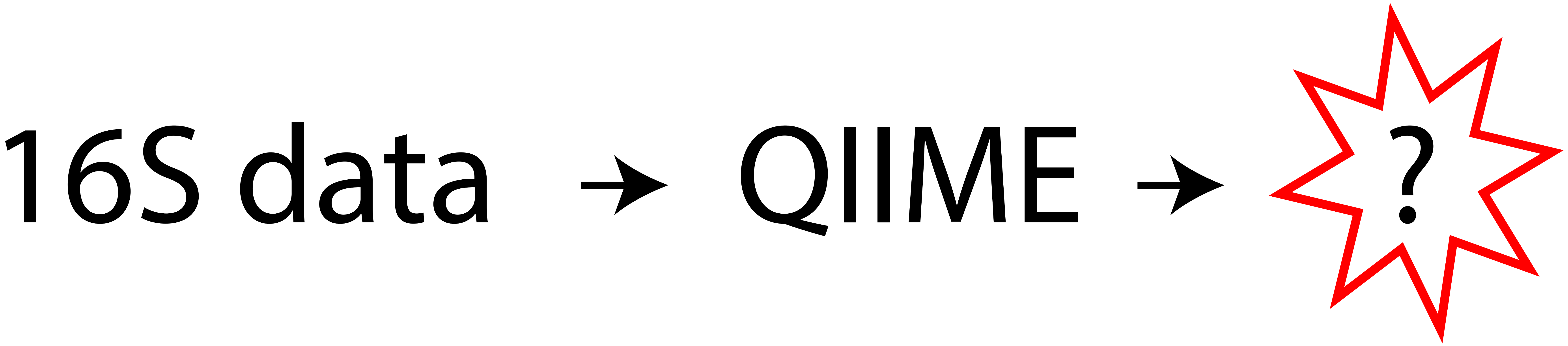
(R –Phyloseq)



16S data → QIIME → ?

So far...

1. NeCTAR NeCTAR website tutorials

2. UNIX command line basics google it

3. Installing QIIME QIIME website instructions

4. Using QIIME QIIME tutorials as templates

Mostly A-to-Z processes = **easy** ...when you know how

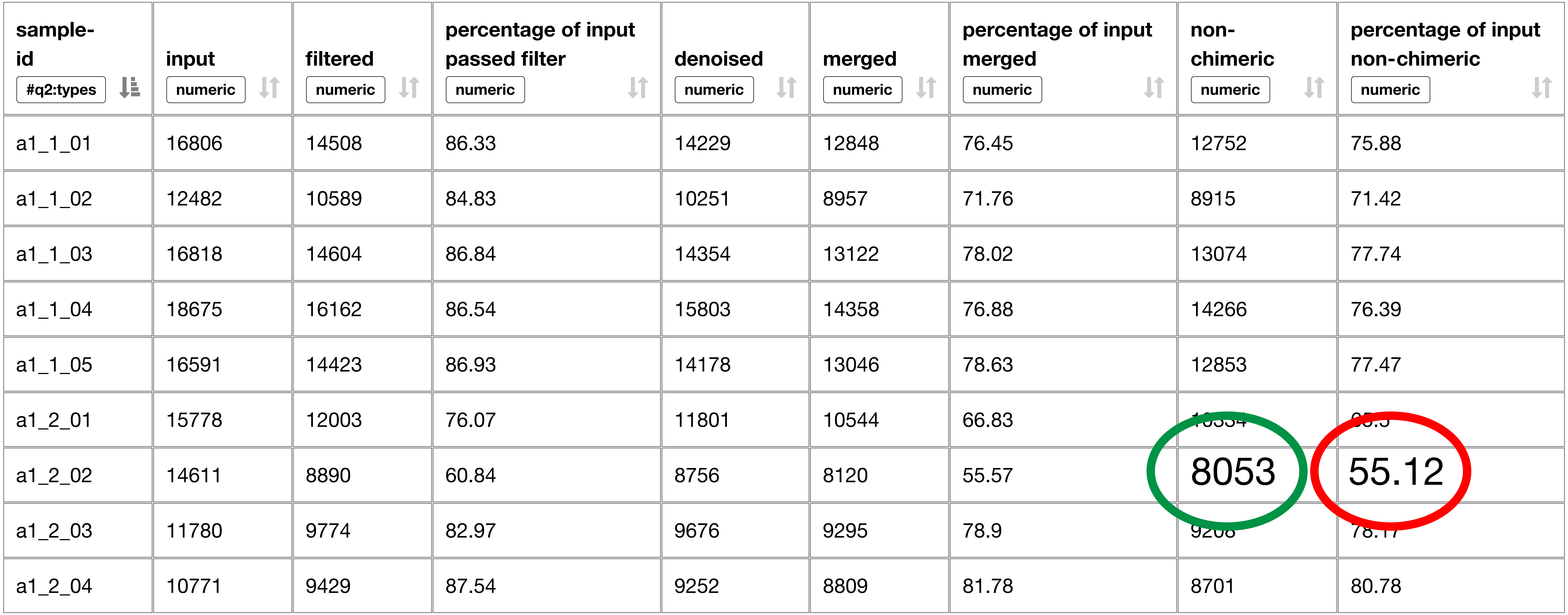
Following on from the last session...

Know your data

- QIIME’s .qzv files
  - <u>reads in</u> vs <u>reads out</u>
  - <u>outliers</u>: ASVs, samples
  - general relationship between samples

reads in vs reads out – denoising stats.qzv

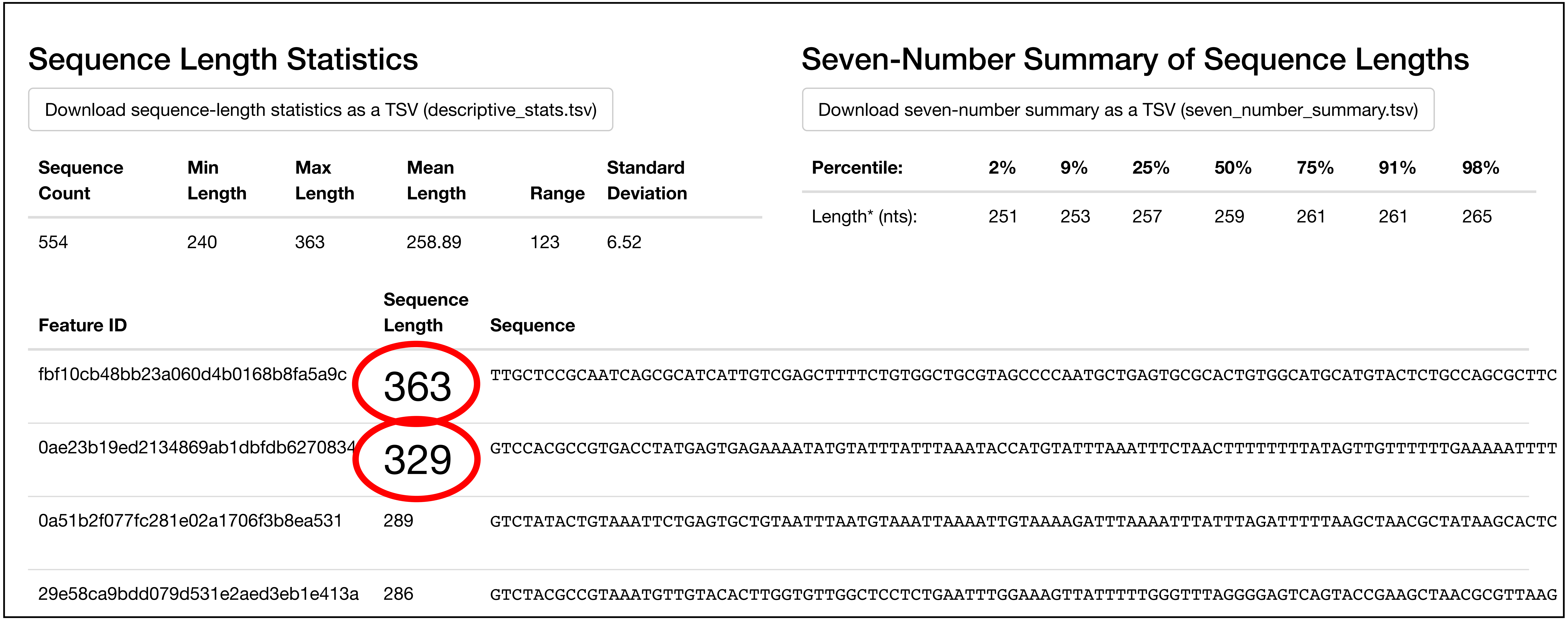
The highlighted sample retained only 55% of reads after processing, but the final read count is still good.



sample-id	input	filtered	percentage of input passed filter	denoised	merged	percentage of input merged	non-chimeric	percentage of input non-chimeric
#q2:types	numeric	numeric	numeric	numeric	numeric	numeric	numeric	numeric
a1_1_01	16806	14508	86.33	14229	12848	76.45	12752	75.88
a1_1_02	12482	10589	84.83	10251	8957	71.76	8915	71.42
a1_1_03	16818	14604	86.84	14354	13122	78.02	13074	77.74
a1_1_04	18675	16162	86.54	15803	14358	76.88	14266	76.39
a1_1_05	16591	14423	86.93	14178	13046	78.63	12853	77.47
a1_2_01	15778	12003	76.07	11801	10544	66.83	10334	65.5
a1_2_02	14611	8890	60.84	8756	8120	55.57	8053	55.12
a1_2_03	11780	9774	82.97	9676	9295	78.9	9208	78.17
a1_2_04	10771	9429	87.54	9252	8809	81.78	8701	80.78

outliers ASVs, samples – representative_seqs.qzv

Our target region was ~277 bp. These ASVs might be PCR artefacts. Check read count and taxonomy.



Sequence Length Statistics

Download sequence-length statistics as a TSV (descriptive_stats.tsv)

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
554	240	363	258.89	123	6.52

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV (seven_number_summary.tsv)

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	251	253	257	259	261	261	265

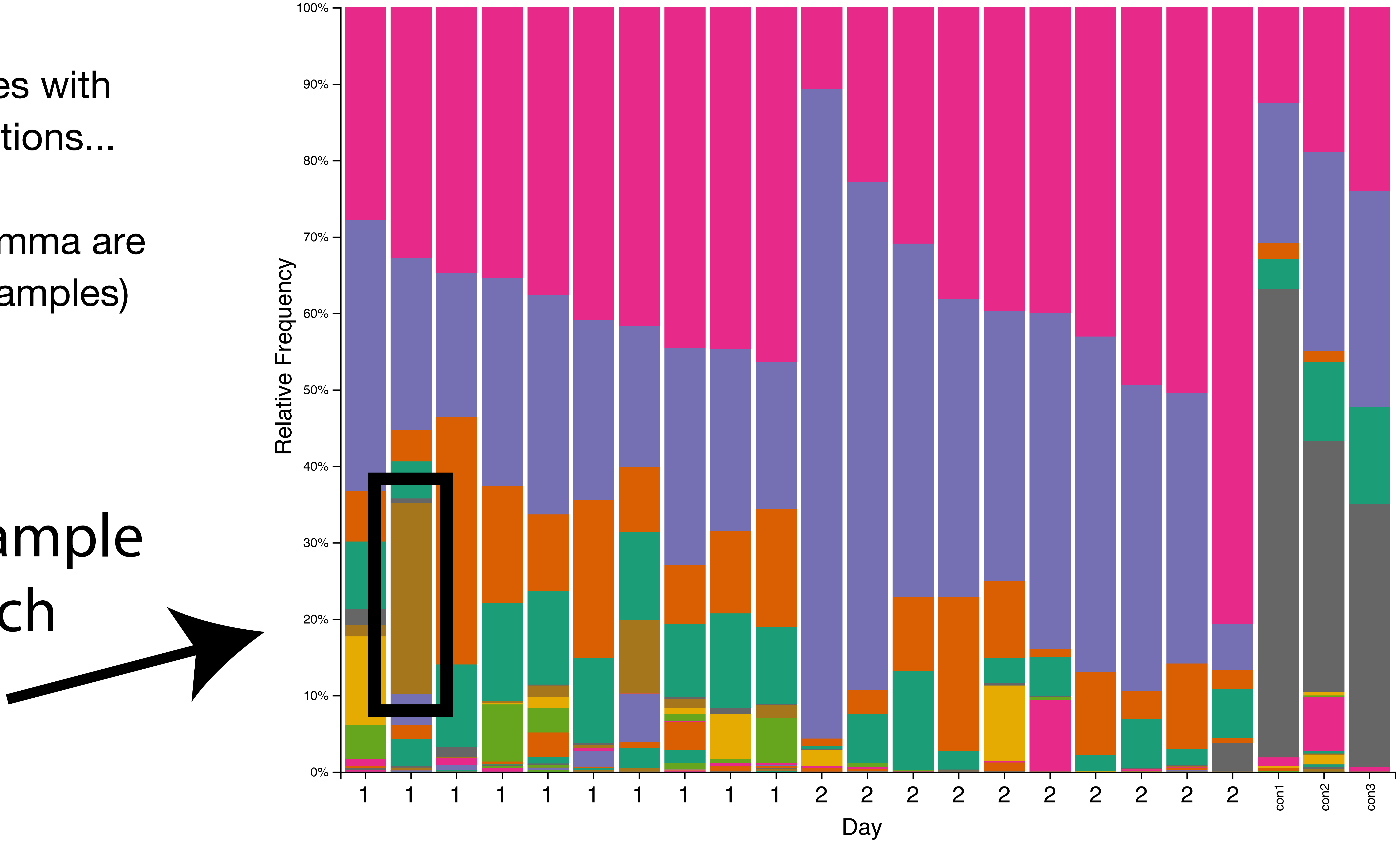
Feature ID	Sequence Length	Sequence
fbf10cb48bb23a060d4b0168b8fa5a9c	363	TTGCTCCGCAATCAGCGCATCATTGTCGAGCTTTTCTGTGGCTGCGTAGCCCCAATGCTGAGTGCGCACTGTGGCATGCATGTACTCTGCCAGCGCTTC
0ae23b19ed2134869ab1dbfdb6270834	329	GTCCACGCCGTGACCTATGAGTGAGAAAATATGTATTTATTTAAATACCATGTATTTAAATTTCTAACTTTTTTTTATAGTTGTTTTTTGAAAAATTTT
0a51b2f077fc281e02a1706f3b8ea531	289	GTCTATACTGTAAATTCTGAGTGCTGTAATTTAATGTAAATTAAAATTGTAAAAGATTTAAAATTTATTTAGATTTTTAAGCTAACGCTATAAGCACTC
29e58ca9bdd079d531e2aed3eb1e413a	286	GTCTACGCCGTAAATGTTGTACACTTGGTGTTGGCTCCTCTGAATTTGGAAAGTTATTTTTGGGTTTAGGGGAGTCAGTACCGAAGCTAACGCGTTAAG

outliers ASVs, samples – barchart.qzv

Check for samples with unusual compositions...

...and trends (Gamma are higher in Day 2 samples)

25%
Only one sample with so much of this ASV



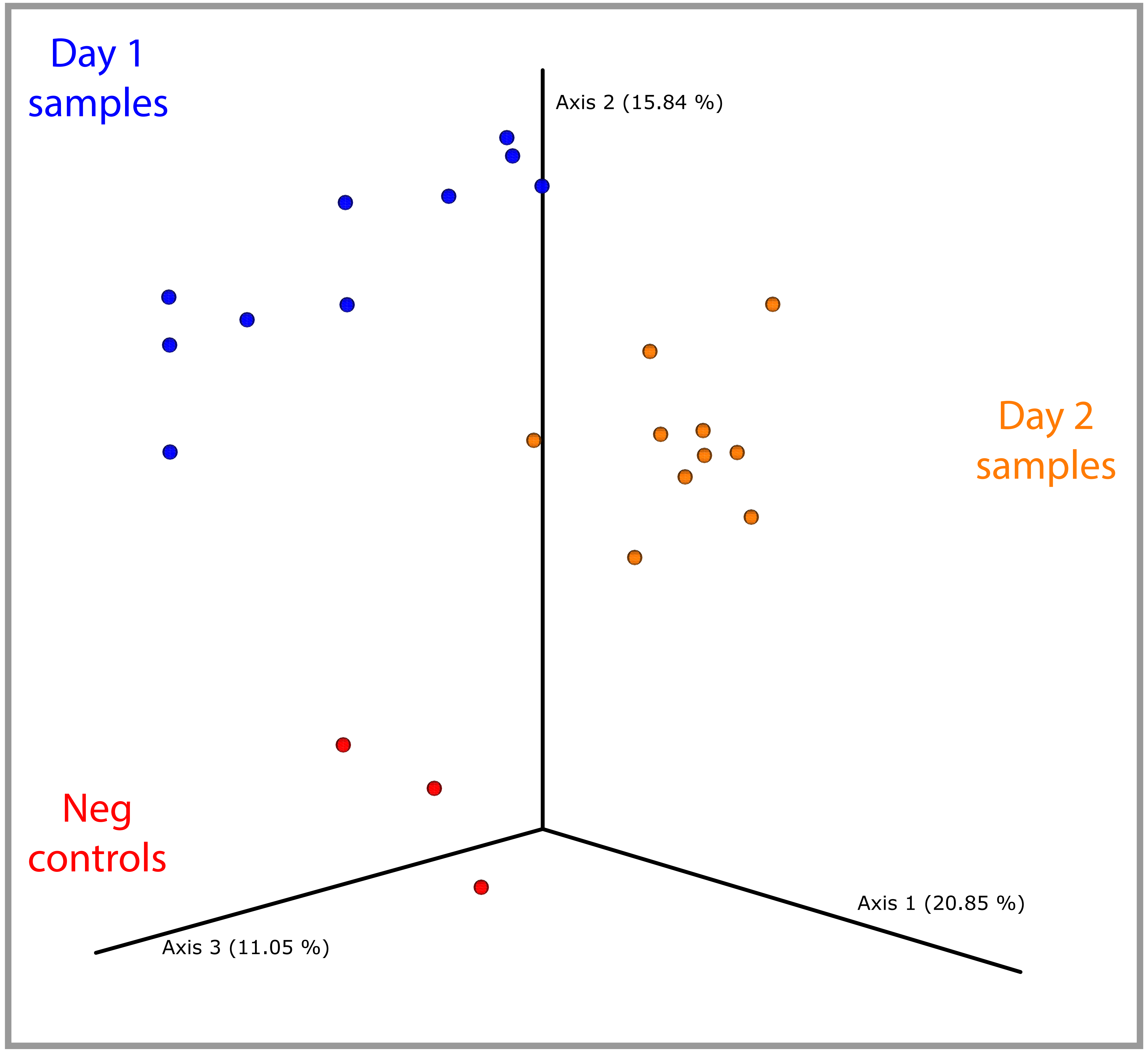
outliers ASVs, samples – emperor.qzv

A useful exploratory tool

- Neg controls are distinct from samples

- We see some separation of samples by day, but...

...this image is a 2-D depiction of a 3-D ordination = dodgy!



Today's session...

QIIME ➔ R ➔ basic visualisations + stats

- analysis of our 'toy' data set
- this will be a demonstration...

  but if you can run the scripts, great!

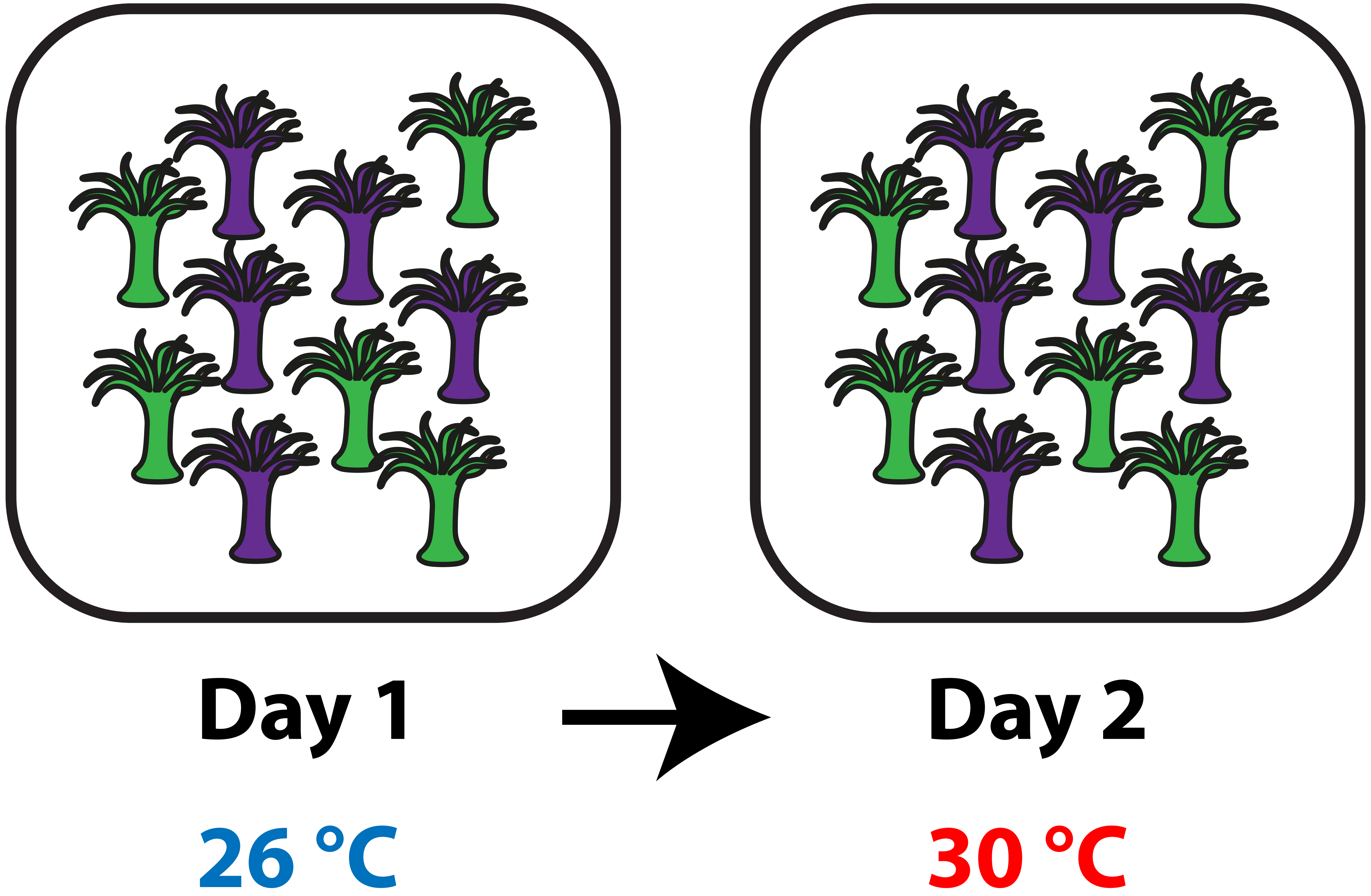
- this will not be an R 'tutorial' – 100-1000's already exist

16S 'toy data'

- 2 x sea anemone genotypes (a1, a4)

- Exposed to temp. increase (26 - 30 °C) over 24 hr

- 2 x sampling time-points:
  Day 1, 26 °C
  Day 2, 30 °C



Day 1 → Day 2

26 °C

30 °C

Q: How did the bacterial communities differ/change?

QIIME output (naming is arbitrary):

tax.tsv	tab-separated taxonomic information links ASV to tax info
table.tsv	tab-separated ASV read count data ASV ‘abundance’
tree.nwk	phylogenetic info in Newick format reqd for some distance matrices e.g. unifrac
metadata	user defined

Following on from the last session...

QIIME2:

- see today’s demo script (30 min runtime: 64 Gb RAM, 16 cores)

- Pro tip: if files have odd names, use a manifest file

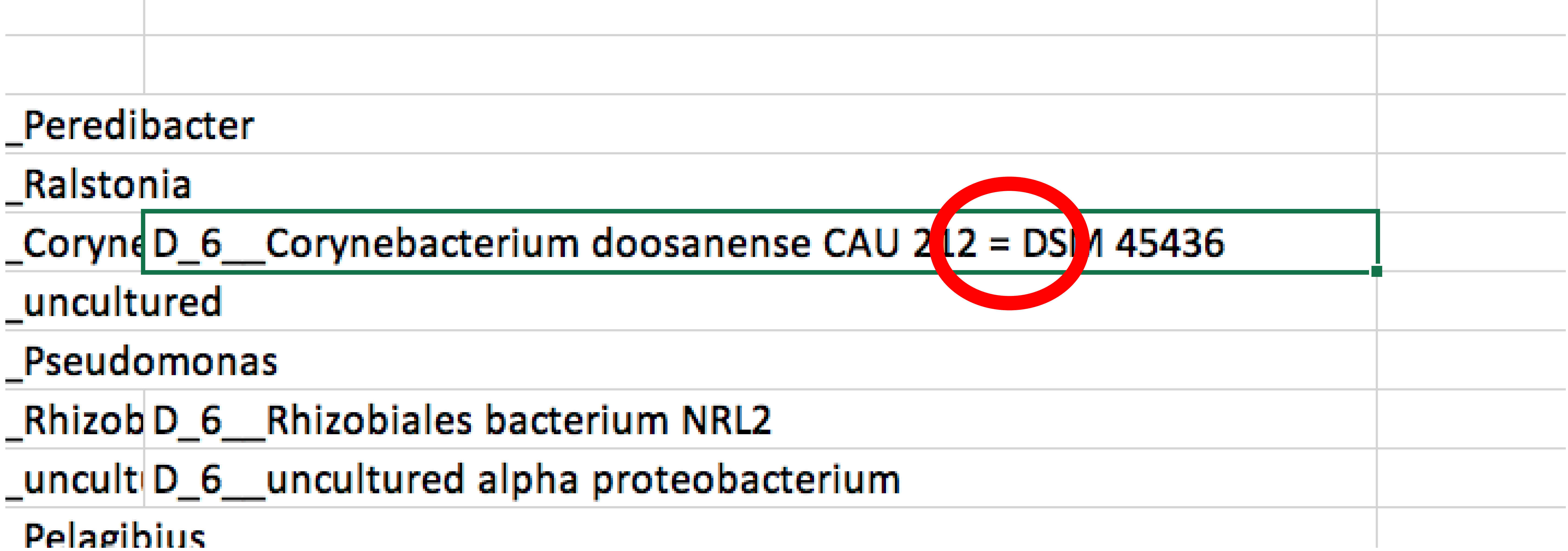
→ **reproducible** + **publishable**

Getting QIIME output in to R:

- we can tidy up files programmatically

```
# Remove header from otu table
sed -i "1d" ~/output/table.tsv
```

- but Excel is easiest for trouble-shooting



_Peredibacter	
_Ralstonia	
_Coryne	D_6__Corynebacterium doosanense CAU 212 = DSM 45436
_uncultured	
_Pseudomonas	
_Rhizob	D_6__Rhizobiales bacterium NRL2
_uncult	D_6__uncultured alpha proteobacterium
_Pelagibius	

a_import_data.R

```
emri_otu <- read.table(
  "table.tsv",
  header = TRUE,
  sep = "\t",
  row.names = 1)

phy <- phyloseq(otu_table(emri_otu_mat,
                taxa_are_rows = T),
                tax_table(emri_tax_mat),
                sample_data(emri_met))
```

Phyloseq (R package)

- combines all files:

  OTU table, taxonomy, tree, metadata

- many useful functions for microbiome analysis

OK, what now?

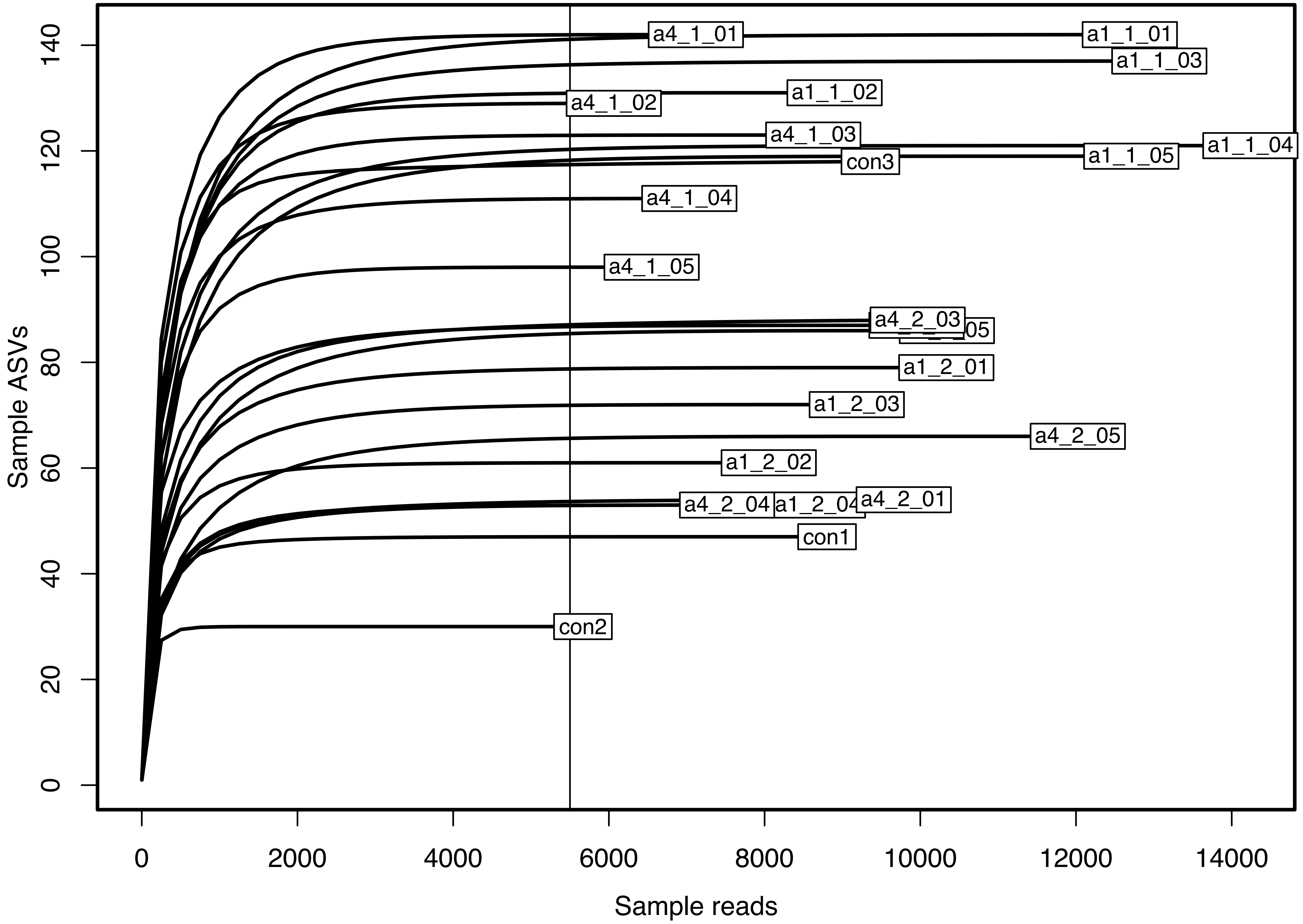
- essential checks
    - QC (seq depth, contamination)

- some basic analyses:
    - α diversity – metrics + stats
    - β diversity – ordinations + stats

b_rare_curves.R

Curves plateau

= seq depth OK



c_decontam.R

Davis, N.M.; Proctor, D.M.; Holmes, S.P.; Relman, D.A.; Callahan, B.J. Simple **statistical identification and removal of contaminant sequences** in marker-gene and metagenomics data. Microbiome 2018, 6, doi:10.1186/s40168-018-0605-2.

...the prevalence of contaminants will be higher in negative controls than in true samples...

Using the 'prevalence' method, only ASVs that are present in the negative controls will be considered potential contaminants.

c_decontam.R

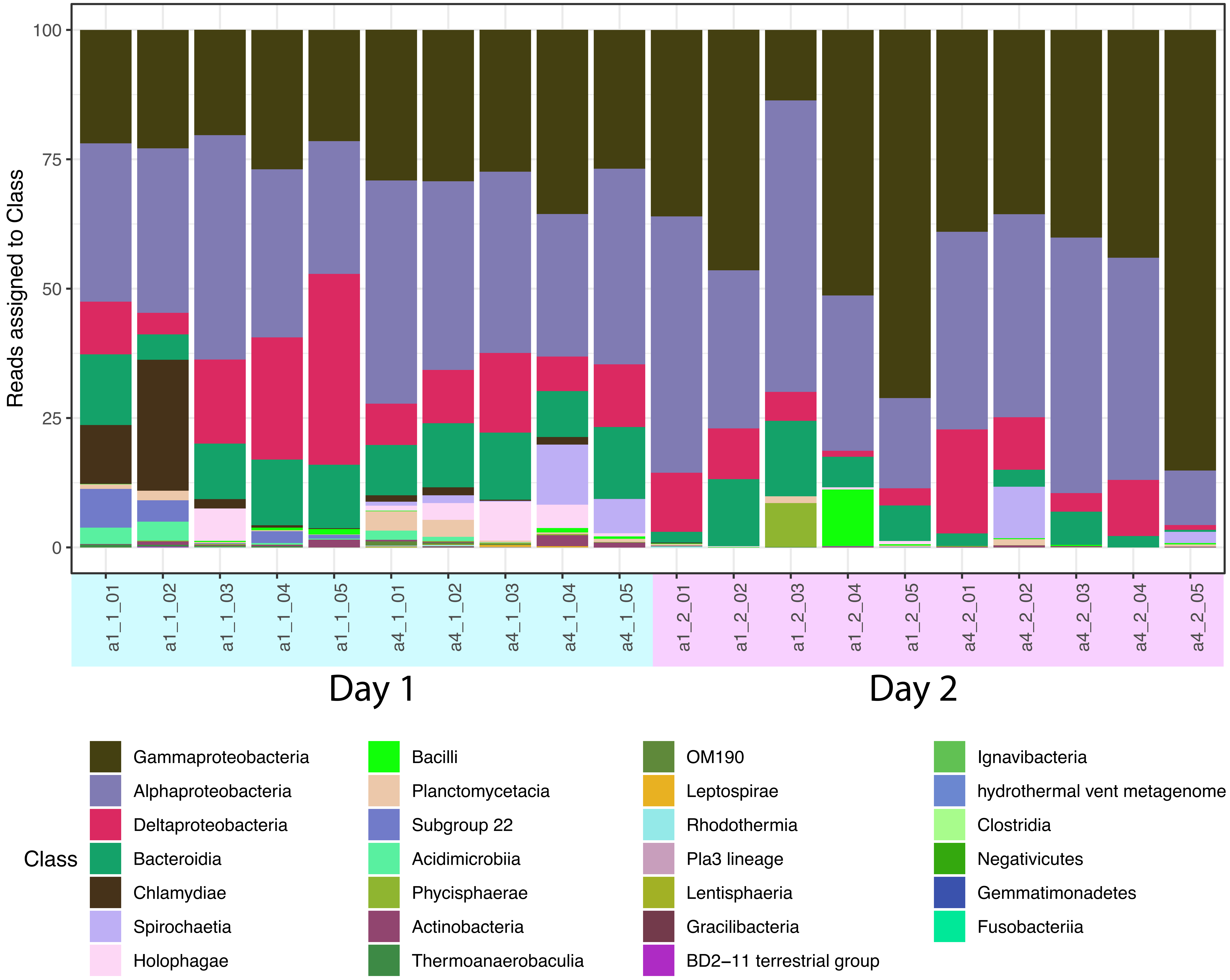
```
isContaminant(seqtab = phy, neg = "neg",
                        method = "prevalence")
```

	Phylum	Class	Order	Family	Genus
0.469	Proteobacteria	Gammaproteobacteria	Oceanospirillales	Litoricolaceae	Litoricola
0.136	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Hydrotalea
0.015	Bacteroidetes	Bacteroidia	Chitinophagales	Chitinophagaceae	Sediminibacterium
0.009	Actinobacteria	Actinobacteria	Micrococcales	Micrococcaceae	Micrococcus
0.004	Firmicutes	Bacilli	Lactobacillales	Carnobacteriaceae	Granulicatella
0.007	Firmicutes	Bacilli	Lactobacillales	Lactobacillaceae	Lactobacillus
0.372	Proteobacteria	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseibacterium
0.049	Proteobacteria	Gammaproteobacteria	Betaproteobacteriales	Burkholderiaceae	Pelomonas
1.062					

e_barchart.R

Select the deepest level of taxonomy that show a trends in the data.

If not very informative, leave it out or put it in the supplementary / appendix.



f_alpha_metrics.R

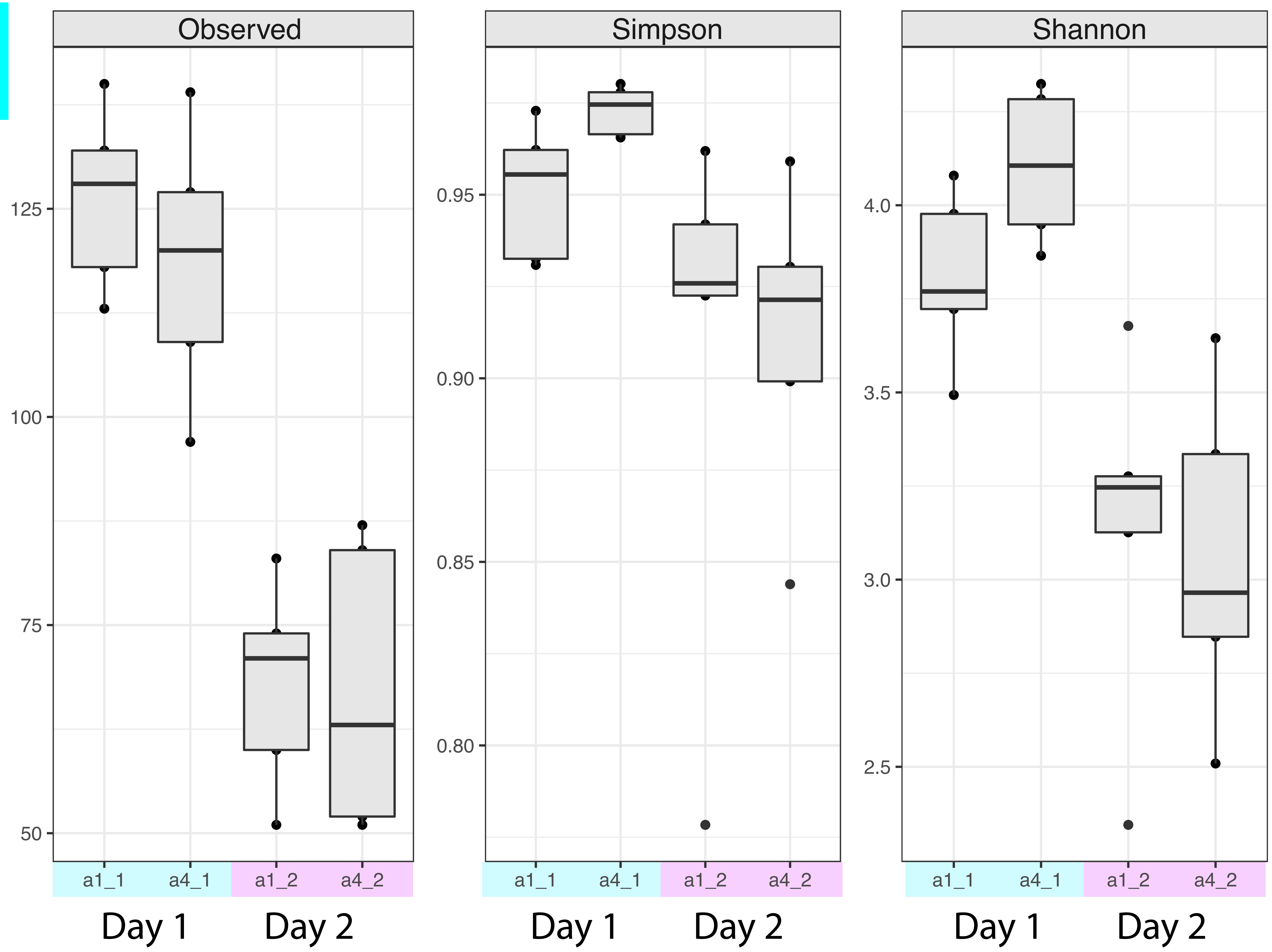
Info about intra-sample diversity:

- How many ASVs?

- Dominance?

- Overall alpha diversity?

"Alpha diversity decreased from Day 1 to Day 2 in both genotypes"



g_alpha_stats.R

Was the decrease in alpha diversity (Shannon index) significant? (p < .05)

- Check whether data meet assumptions for ANOVA
  - normality: shapiro.test
  - homogeneity of variance: leveneTest

- If assumptions not met, look at non-parametric alternatives
  - Kruskal-Wallis etc

- Likewise if data have special features e.g. irregular time-series
  - GLS model etc

g_alpha_stats.R

Was the decrease in alpha diversity (Shannon index) significant? (p < .05)

a1_1: shapiro.test((diversityMetrics$Observed)[1:5])

p-value = 0.7842

Does not differ sig from normality :)

all samples: leveneTest(Observed ~ grouping, data = diversityMetrics)

p-value = 0.7351

Variance is not sig dif :)

g_alpha_stats.R

Was the decrease in alpha diversity (Shannon index) significant? (p < .05): Yes!

```
summary(aov(Shannon ~ genotype * samplingDay, data = diversityMetrics))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
genotype	1	0.074	0.074	0.522	0.480568	
samplingDay	1	3.636	3.636	25.549	0.000117	***
genotype:samplingDay	1	0.167	0.167	1.176	0.294253	

“There was no sig dif in Shannon diversity based on genotype (p = 0.48), however both genotypes’ Shannon diversity was lower on Day 2 compared to Day 1 (p < 0.05)”

h_beta_nMDS.R

What is the relationship between the sample types?

- Ordinations are great for helping us assess the relationships between samples based on their bacterial community compositions

- Choose a distance matrix (see Ashley's QIIME2 notes)
  - Jaccard – presence-absence
  - Bray-Curtis – presence-absence + relative abundance
  - Unifrac – incorporates phylogenetic relatedness

  ...many more

h_beta_nMDS.R

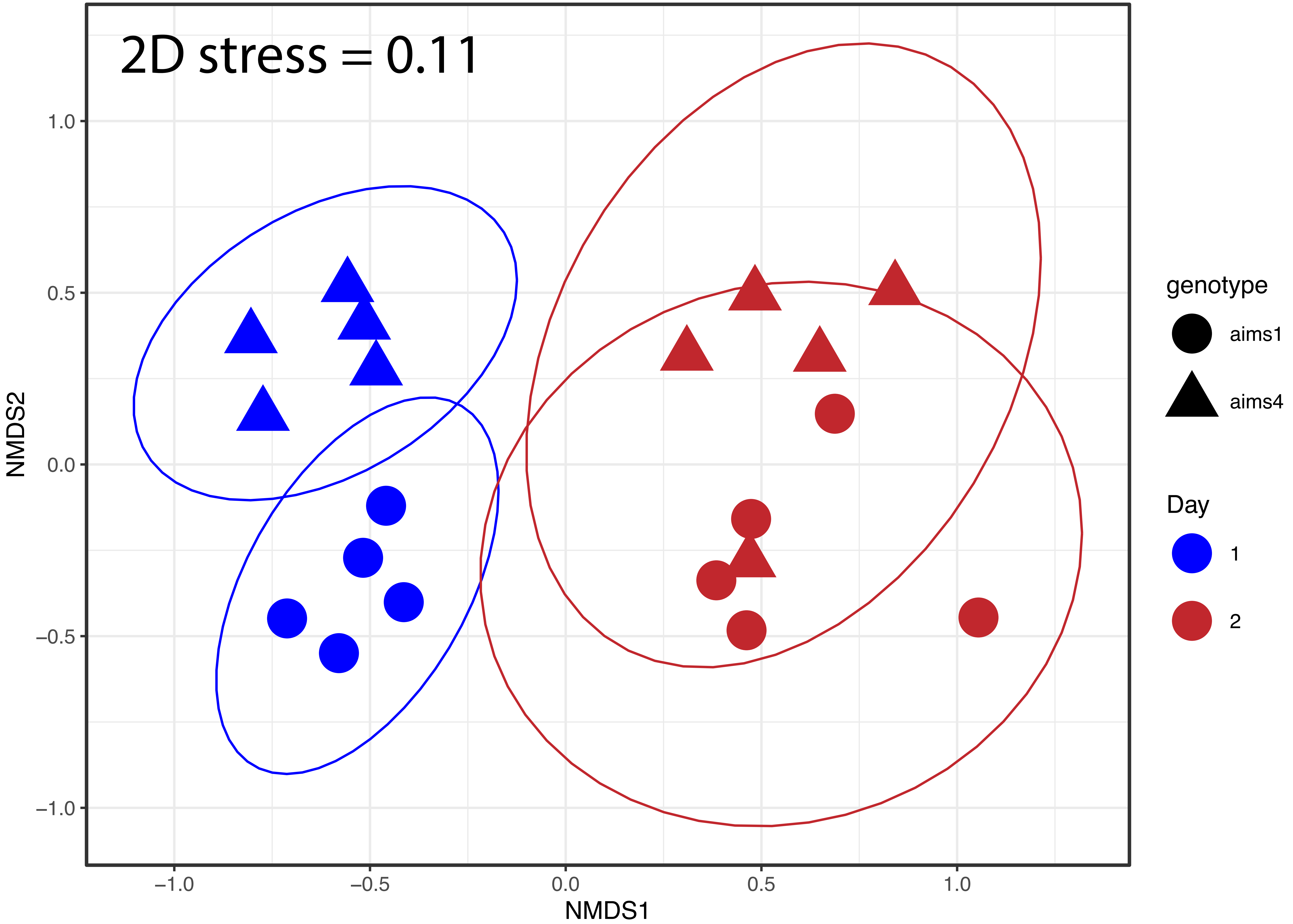
What is the relationship between the sample types?

- nMDS based on Bray-Curtis distance

- The bacterial communities of both genotypes differ from from Day 1 to Day 2.

- The bacterial communities of the genotypes differ from each other at Day 1, but not Day 2.

- Were these differences significant?



h_beta_mvabund.R

What is the relationship between the sample types?

- Analysis of count data is problematic – zeroes, non-normality...
- permANOVA and ANOSIM overcome this through permutation = loss of power
- So, we will use a method that accomodates the nature of the data: mvabund (GLM)

	Res.Df	Df.diff	Dev	Pr(>Dev)
genotype	18	1	270.7	0.017 *
samplingDay	17	1	680.4	0.001 ***
genotype : samplingDay	16	1	133.6	0.026 *

- The data differ significantly based on genotype & sampling-day.
- However, there is also a significant interaction. We need more p-values! :-P

h_beta_mvabund.R

What is the relationship between the sample types?

- Day 1

	Res.Df	Df.diff	Dev	Pr(>Dev)
Day1$genotype	8	1	330.2	0.024 *

"The compositions of the genotypes' bacterial communities differed significantly on Day 1 (p = 0.024) ..."

- Day 2

	Res.Df	Df.diff	Dev	Pr(>Dev)
Day2$genotype	8	1	111.5	0.269

"...but by Day 2 they were no longer significantly different (p = 0.269)."