Appendix

Data analysis code

```
git clone https://github.com/Cistron/bioc3301_data
```
Download year 2016 and year 2017 data-sets into custom folder.

```
source activate qiime1
```
Activates miniconda (Python) virtual environment.

```
validate_mapping_file.py -m map.tsv -o ./vmf
```
Ensures the mapping file is correct, faults will be highlighted in maps.tsv.html

```
split_libraries_fastq.py --barcode_type 12 -i bioc3101_2016_read1_50k.fastq.gz -m map.tsv -o ./slout -b bioc3101_2016_barcodes_50k.fastq
```
Demultiplexing and quality filtering of data according to barcode.
Only used read 1 because read 2 was of lower quality this year.

```
count_seqs.py -i ./slout/seqs.fna
```
Counts the sequences in a fna file and write results to slout.

```
pick_closed_reference_otus.py -i ./slout/seqs.fna -o ./otus
```
Picks operational taxonomic units with closed reference.

```
biom summarize-table -i ./merged_otu_table.biom
```
Produces a human readable summary of the OTU table (**Table 5**).
A total of 10453611 sequences can be observed here.

Table 5 | Summary of the OTU table generated using the `biom summarize-table` command.

Number of samples:	12
Number of observations:	15504
Total count:	10453711
Table density (fraction of non-zero values:	0.443
Counts/sample summary:	
Min:	371590.0
Max:	1427140.0
Median:	849510.000
Std. dev.:	261282.144
Sample Metadata Categories:	None provided
Observation Metadata Categories:	Taxonomy
Counts/sample detail:	
15.16.1:	1427140.0
15.16.2:	966382.0
15.16.3:	785867.0
15.16.4:	371590.0
16.17.1:	818625.0
16.17.2:	1094022.0
16.17.3:	693596.0
16.17.4:	880395.0
16.17.5:	597583.0
16.17.6:	1017801.0
16.17.7:	724145.0
16.17.8:	1076565.0

```
core_diversity_analyses.py --recover_from_failure -o cdout/ -i
merged_otu_table.biom -m map.tsv -t 97_otus.tree -e 371590 --recover_from_failure
```
Runs diversity analyses at 371590 sequences per sample.
Enables investigation of alpha (within sample) and beta (differences between habitats) diversity.
Also generates 3D principal coordinate plots, which can be subsequently viewed in EMPeror.
-e is the sampling depth, set to 371590 which is the lowest number of sequences observed in the biom table summary, else these data are excluded from the analysis.
If –e parameter is set too high, the smaller samples will be excluded.
--recover_from_failure permits analysis to be resumed should it crash.
The output of this script is an HTML file that can opened in a web browser (**Figure 10**).

Run summary data	
Master run log	log_20170324140718.txt
Previous run log	log_20170323213036.txt
Previous run log	log_20170324124136.txt
Previous run log	log_20170324125105.txt
Previous run log	log_20170324125309.txt
BIOM table statistics	biom_table_summary.txt
Filtered BIOM table (minimum sequence count: 371590)	table_mc371590.biom.gz
rarefied BIOM table (sampling depth: 371590)	table_even371590.biom.gz
Taxonomic summary results	
Taxa summary bar plots	bar_charts.html
Taxa summary area plots	area_charts.html
Beta diversity results (even sampling: 371590)	
PCoA plot (unweighted_unifrac)	index.html
Distance matrix (unweighted_unifrac)	unweighted_unifrac_dm.txt
Principal coordinate matrix (unweighted_unifrac)	unweighted_unifrac_pc.txt
PCoA plot (weighted_unifrac)	index.html
Distance matrix (weighted_unifrac)	weighted_unifrac_dm.txt
Principal coordinate matrix (weighted_unifrac)	weighted_unifrac_pc.txt
Alpha diversity results	
Alpha rarefaction plots	rarefaction_plots.html

Need help? See http://help.qiime.org.

Figure 10 | HTML result from `core_diversity_analyses.py`. This HTML file summarises and gives access to the results of the diversity analyses conducted on the given OTU table.

```
make_2d_plots.py -i unweighted_unifrac_pc.txt -m map.tsv_corrected.txt
```
Generates 2D PCoA unweighted plots which is useful for qualitative analysis, considers the presence or absence of species).

```
make_2d_plots.py -i weighted_unifrac_pc.txt -m map.tsv_corrected.txt
```
Generates 2D PCoA weighted plots which is useful for quantitative analysis, accounts for abundance of observed organisms.

```
source deactivate qiime1
```
Deactivates virtual environment.

Statistical testing and correlation testing code

The following QIIME scripts were used for performing statistical significance analyses of sample grouping using UniFrac distance matrices (http://qiime.org/scripts/compare_categories.html), and calculating the correlation between observation abundances and continuous-valued metadata (http://qiime.org/scripts/observation_metadata_correlation.html).

```
source activate qiime1
```
Activates miniconda (Python) virtual environment.

```
compare_categories.py --method adonis -i unweighted_unifrac_dm.txt -m
map.tsv_corrected.txt -c Year -o adonis_out -n 999
```
adonis is a non-parametric statistical method that takes a QIIME distance matrix file such as UniFrac distance matrix, a mapping file, and a category in the mapping file to determine the sample grouping
this command creates a new output directory named adonis_out , which will contain a single text file (`adonis_results.txt`)
R^2 value (effect size) will be computed, which shows the percentage of variation explained by the supplied mapping file category (in this case, Year)
a p-value will also be computed, which determines the statistical significance

```
compare_categories.py --method anosim -i unweighted_unifrac_dm.txt –m
map.tsv_corrected.txt -c Year -o anosim_out -n 999
```
ANOSIM (similar to adonis) tests whether 2 or more groups of samples are significantly different.
Generates the R statistic and p-value.

```
observation_metadata_correlation.py -i merged_otu_table.biom -m
map.tsv_corrected.txt -c Year -s spearman -o spearman_otu_gradient.txt
```
This script computes correlations between feature (aka. observation), abundances (relative or absolute) and numeric metadata.
Spearman's Rho was used here, which is a non-parametric measure of correlation between two sequences of numbers.
Spearman correlation is appropriate for data where the values of the observations are not necessarily accurate, but for which their relative magnitudes are.
The output generated from this script is a tab-limited text file with the following headers:
Feature ID (ID of the features being correlated – these are the observation IDs in the BIOM table)
Test stat. (test statistic value for the given test)
pval (raw p-value returned by the given test)
pval_fdr (p-value corrected by the Benjamini-Hochberg FDR procedure for multiple comparisons)
pval_bon (p-value corrected by the Bonferroni procedure for multiple comparisons)
[metadata] (this column is present only if the BIOM table contained metadata information for your features. For example, if these are OTUs, and taxonomy is present in the BIOM table, this column will contain OTU taxonomy and be named 'taxonomy')

```
source deactivate qiime1
```
Deactivates virtual environment.