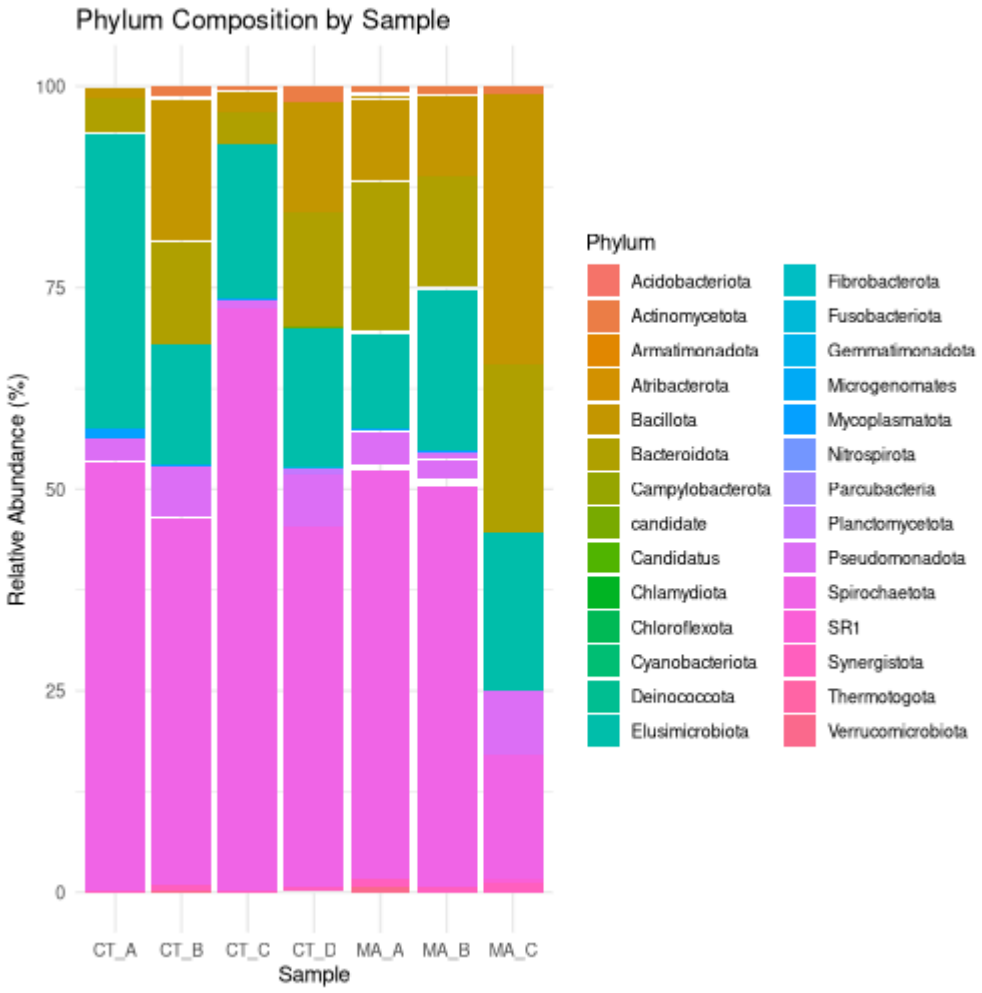
Oliver Todreas

1. **45 samples** were taken across colonies. For the caste analysis, 19 samples were taken across all 7 colonies, meaning that a total of **133 samples** were taken.
2. ***Spirochaetes*** is the most abundant phylum in the termite microbiota
3. The hindgut microbiota **did not differ significantly** between castes.
4. The `fastq` files take up **1.3066 GB** and the `xml` files take up **252 KB**.
5. The location specified by the coordinates are in **Kyrgyzstan**.
6. In this study, we want exactly two reads per individual. To ensure this is the case, we can check that each directory contains two `fastq` files where the first line of each `fastq` file is identical except for the number following the space: for one file, the character should be `1`, and the other, the character should be `2`. Only use the directories whose content match these conditions.
7. There are **7,856,118 sequences** across the `fastq` files.
8. There are an average of **227.5 bases per read** across the `fastq` files.
9. The `- - - -` in `paste` is **merging four lines** into one from the input files into a single line, returning a matrix of sorts, with four columns and the cumulative number of rows of all files subject to the `grep` divided by four.
10. The second line is cutting out all columns but the second (corresponding with the sequences) and then performing a Perl-compatible regular expression `grep`, where each line must start with `TTAGA`, be followed by a `C` or `G`, and then be followed by `AC`.
11. The “per base sequence quality” graph represents the distribution of quality scores (Phred scored) by position in the read. Typically, quality scores trend down and become more widely distributed for higher positions in reads. The “per sequence GC content” compares the average GC content across reads to a normal distribution with a mean near 50% GC.
12. The “overrepresented sequences” metric is expected to give warnings in this study because a representative sequence of an abundant microbe is expected to appear many times in the read.
13. In shell scripting, a backslash `\` allows one-line commands to **continue on a new line**.
14. A **negligible number** of reads were discarded from both the forward and reverse reads (rounds to 0.00%).
15. For `CT_A` forward reads, the trimmed paired forward reads showed almost no change. For `CT_C` reverse reads, the trimmed paired reverse reads showed a tighter sequence length distribution. For `MA_A` reverse reads, the trimmed paired reads showed a decrease in sequence duplication level, with seemingly 0% of total sequences duplicated over 50 times.

16. `pandaseq` is an assembler that assembles reads into sequences from two `fastq` files, the forward and reverse read, assembling them on their overlapping middle region.
17. The average length of merged sequences is **252.9** bases.
18. The recommended Vsearch workflow contains steps to perform paired-end read merging, quality filtering, chimera removal, and OTU clustering. We have not quality filtered or removed chimeras yet, which would be useful on our dataset at this time.
19. Dereplication drops all replicated reads. We can still access the abundance of a read because we used the `--sizeout` option with `vsearch`, saving the abundances in the `fasta` header.
20. `--relabel` relabels the output sequences. Each header (lines starting with >) now starts with the file ID, followed by the abundance rank and the abundance of the sequence.
21. **No reads were discarded**. The number of reads in the input `fastq` is identical to the sum of the read abundances in the output `fasta` file. The output has much fewer lines because it encodes abundance with a single number in the header.
22. There are **580,341 "unique" reads, together making up 3,888,919 reads** in the pooled reads file. There could in reality be fewer unique reads because a read might appear in multiple input files.
23. We need to dereplicate the pooled reads because as mentioned above, a given read could appear in multiple input files, but we want every read to appear once in the file.
24. The most frequent sequence is represented **181,549 times**.
25. There were 3,012,197; 406,046; 29,568; and 47,582 reads for samples CT_A through CT_D and 315,366; 70,644; and 7,516 reads for samples MA_A through MA_C.
26. The preclustering step drastically shortens the length of the file by clustering together similar sequences around representative sequences of a cluster, or centroids.
27. There are **48,148** `fasta` entries after preclustering.
28. There are **48,148** centroids in the `uc` file. Since each sequence in the preclustered `fasta` is a representative sequence for a cluster, it stands to reason that the number of entries and the number of centroids are equal.
29. A potential difference with running referenced-based chimera checking first is that the dataset passed to the *de novo* check is smaller than it would otherwise be, which could affect the algorithm.
30. The two `fasta` files together make up the contents of the input `fasta` file. The sequences are split based on whether they are predicted to be chimeras or not. The `uchime` file is similar to the previously generated `uc` files and contains one line for each sequence.

31. *De novo* means that the chimeras are predicted “from scratch” without any reference data, simply by an algorithm. We are not using any previously verified data, so therefore the insights gathered from this kind of chimera checking are “brand new.”
32. **33.1% of reads were classified as chimeric** during *de novo* checking.
33. The database contains **5181 sequences** with an average length of **1435 bases**.
34. —
35. **11.9% of reads were classified as chimeric** during reference based checking. See Q32 for *de novo* checking.
36. —
37. —
38. —
39. There are **3,696,289 raw reads** and the same number of dereplicated reads, although there are fewer unique reads of the latter. There are **471,809 non-chimeric preclusters**.
40. TODO: finish
41. In this case, `sed` is dropping the number following the sample ID.
42. There are **55,406 unique OTUs**, and they represent **3,696,289 sequences**, which corresponds to the number of input sequences.
43. 870,000; 674,225; 359,815; and 280,868 OTUs were identified for colonies CT_A through CT_D, and 1,028,941; 440,879; and 41,561 OTUs were identified for colonies MA_A through MA_C. These add up to the total number of 3,696,289 sequences.
44. For the entire set of OTUs, the average number of sequences per OTU is 15.7, 12.2, 6.5, and 5.1 for the CT colonies and 18.6, 8.0, and 0.8 for the MA colonies.
45. The RDB database is the database that the program uses to perform classification on input data.
46. I’m not sure how many lines it is.
47. The database can increase or decrease in size based on the maintainers’s assessment of the state of the art research since it is a manually curated database.
48. **60 cluster sequences** were not classified as bacteria and **37,284 cluster sequences** were determined at the phylum level.
49. **10,008 cluster sequences** were determined at the genus level at a cutoff of 0.8.
50. If a 0.7 cutoff was used, more cluster sequences would be considered determined at the genus level, and cutoff of 0.9 was used, fewer clusters would be considered determined at the genus level.

51.


Phylum Composition by Sample
Relative Abundance (%)
100
75
50
25
0
CT_A
CT_B
CT_C
CT_D
MA_A
MA_B
MA_C
Sample
Phylum
Acidobacteriota
Actinomycetota
Armatimonadota
Atribacterota
Bacillota
Bacteroidota
Campylobacterota
candidate
Candidatus
Chlamydiota
Chloroflexota
Cyanobacteriota
Deinococcota
Elusimicrobiota
Fibrobacterota
Fusobacteriota
Gemmatimonadota
Microgenomates
Mycoplasmatota
Nitrospirota
Parcubacteria
Planctomycetota
Pseudomonadota
Spirochaetota
SR1
Synergistota
Thermotogota
Verrucomicrobiota