Chang Lab / Henry Group Collaboration on Mouse Gut Microbiomes

Last Updated May 4, 2024

**Most up-to-date inputs and derivative data products for ASV to genome mapping, community model assembly, and metabolic constraints for community FBA**

**Base Folder: /MouseGutMicrobiomes\_DataProducts\_04-29-24**

*Raw data from Megan*

**Folder: /amplicon\_seq\_methods\_05-26-22**

**DFI\_methods.docx**

Written details of the amplicon sequencing methods including primer sequences, amplicon + ligated primer & barcode size, and library type.

**Folder: /asv\_data\_products\_Kennedy\_04-28-24**

**wd\_meta\_merged.csv**

16S sample metadata file. Metadata for 1029 unique sampleID entries (rows). There are 21 columns of metadata: sampleID, seqID, Mouse, Treatment, Diet, Cage, Desc, Day, Rec\_day, Rec\_day\_adj, Sac group, Aim, Cohort, reads in, reads out, nseqs, seq\_run, Abx, Post diet, Post micr, Pre diet

The columns Treatment + Rec\_day together define the groups over which abundances will be averaged for the community model parameters.

**wd\_merged\_otu.csv**

16S ASV abundance table with counts for 2770 taxa (rows) in 1030 samples (columns). Amplicon abundance table and sequences were produced by Megan with dada2 in qiime2 and include all sequencing runs from the project. Sample "HF1009" is present in the abundance table but not the metadata table.

**wd\_merged\_seqs.fasta**

16S ASV sequences. Amplicon abundance table and sequences were produced by Megan with dada2 in qiime2 and include all sequencing runs from the project.

**wd\_taxonomy.csv**

16S ASV taxonomy with seven levels (columns) for 2770 sequences. Taxonomy level names are: Kingdom, Phylum, Class, Order, Family, Genus, Species. Taxonomy assignments produced by Megan in qiime2.

**Folder: /metabolite\_data\_Kennedy\_03-29-2023**

**metabolomics\_meta.csv**

Metabolite sample metadata for 74 samples. There are 16 columns: Sample.ID, Date, Day, Rec\_day, Rec\_day\_adj, Mouse, Cage, Desc, Type, Treatment, Aim, Cohort, Sac group, Cluster, Panels, Notes. Metabolite metadata from Box folder.

**GC1315\_TMS\_Results\_Norm.csv**

Metabolite abundance table with values for 60 metabolites (columns) in 69 samples (rows). Metabolite data from Box folder. Metabolites: 12\_Ketolithocholic acid, 2\_deoxy\_Ribose, 5\_Oxoproline, Alanine, Allose, Arabinose, Asparagine, Aspartic acid, Cellobiose, Cholic/Allocholic acid, Citric acid, Cysteine, Decanoic acid, Deoxycholic acid, Dodecanoic acid, Dulcitol, Fructose, Fucose, Fumaric acid, Galactose, Glucose, Glutamic acid, Glycine, Hexanoic acid, Histidine, Indole, Isoleucine, Leucine, Linoleic acid, Lithocholic/Allolithocholic/Isolithocholic acid, Lysine, Malic acid, Melibiose, Methionine, Myristic acid, N\_Acetyl\_Glucosamine, Niacin, Octanoic acid, Ornithine, Palmitic acid, Phenylalanine, Picolinic acid, Proline, Pyruvic acid, Raffinose, Rhamnose, Ribose, Serine, Succinic acid, Sucrose, Threonine, Tyrosine, Ursodeoxycholic acid, Valeric acid, Valine, alpha\_Ketoglutaric acid, cis\_Oleic acid, meso\_Erythritol, myo\_Inositol, trans\_4\_Hydroxyproline

**scfa\_quant\_results\_20220511\_PFBBr\_GC1315\_20220516.csv**

Metabolite abundance table with values for 4 metabolites (columns) in 17 samples (rows). Metabolite data from Box folder. Metabolites: Acetate, Butyrate, Propionate, Succinate

*Processed data from Kat*

**Folder: /asv\_to\_agora\_mapping\_Beilsmith\_04-29-24**

**AGORA2\_04\_19\_24\_16S\_Compendium\_04\_26\_24.json**

Of the 7302 genomes in the AGORA2 Supplementary Table 1, only 6530 had GCF IDs in NCBI. Of these, 6529 genomes downloaded without error, 6524 were unique genomes, and 6508 had a retrievable \*.fna in the download package. This file contains a dictionary of sequences for the 16991 copies of 16S SSU rRNA found in those 6508 genomes. Each dictionary key has the following format: {genome\_id};{contig\_id};{direction};{start\_loc};{length}.

**AGORA2\_Kennedy\_over90\_best\_mappings\_05-03-24.json**

Format of each entry:

{ASV ID: [identity level of best matches, [list of best matching genomes]]}

The 16S SSU rRNA sequences in the AGORA2 compendium file were clipped with the primers in the sequencing methods to generate regions of equivalent size to the amplicons. Iterating over sequence identity levels from 100% to 90%, pairwise alignments were performed for each amplicon to each clipped AGORA2 sequence (aligment parameters match = 1, mismatch = 0, open gap = -1, extend gap = -0.5). When there was not already a matched genome and the alignment score exceeded the expected perfect score for the identity threshold, an AGORA2 genome was added to a list of best matches along with the sequence identity level of the match. Of 2770 ASVs, 1654 (59.71%) had at least one matching genome. The median number of matching genomes was 4, the mean 24, the standard deviation 145, and the range 1 to 2785. Some types of bacteria had many closely related genomes in the AGORA2 collection so there were large numbers of matches to ASVs in these groups.

**Folder: /preprocessed\_asv\_abundances\_Beilsmith\_05-03-24**

**mean\_abund\_perDay\_perTreatment\_20.csv**

**mean\_abund\_perDay\_perTreatment\_50.csv**

**mean\_abund\_perDay\_perTreatment\_100.csv**

These are modified ASV abundance tables for constraining the models. Samples are grouped by Treatment (ABX-RC, ABX-WD) and Rec\_day (-13:13) columns. The tables have a column with the averaged relative abundance for each group of the top N (20, 50, or 100) ASVs in the dataset (top N taxa determined by raw counts).