Processing Stool Sample Whole Genome Sequencing Data

Ankan Choudhury

2025-03-05

Introduction

This code includes the bash code for performing Whole-genome Sequence analysis on the Stool samples using Trimmomatic, Bowtie, Kraken, Bracken, Kaiju, Metaphlan, and Humann.The paired sequences ("raw") were first trimmed using Trimmomatic v0.39, which trimmed the sequences and separated out into paired ("trimmed") and unpaired, sequences who do not have a matching read pairing with it. The trimmed sequences were then filtered for host reads (*Homo sapiens*) using Bowtie2 resultant sequences ("trimmed-filtered") were used for taxonomic classification. Taxonomic classification was done at all three levels of sequence refinement were done using Kraken2, Kaiju, and Metaphlan. Kraken2 classification used standard Kraken database and was normalized by Bracken using standard Bracken database. Kaiju classification was done using the Kaiju RefSeq database. Metaphlan 4.1.1 used its own ChocoPhlan database (updated October 2024). The trimmed-filtered sequences were also assembled into contigs using Megahit, and the assembled sequences ("trimmed-filtered-contigs") were also classified using Kraken2-Bracken algorithm. The raw, trimmed, trimmed-filtered sequences were also functionally annotated using Humann 3.9. The resultant taxonomic classification data and functional annotation data were transformed and compiled into phyloseq objects in R.

Prerequisites

1. Install qiime2-metagenome-2024.10
2. Install Trimmomatic 0.39
3. Install Metaphlan 4.4.1
4. Install Humann 3.9
5. Install R 4.3.1 or higher

1. Getting the sequences

All the code chunks given below should be made into a .sh files and executed by batch submission in kodiak server as they usually take long time to run.

```bash
ZIP_FILE="061824KGmetagenome.zip"  # The zip file containing the sequences
EXTRACT_DIR="extracted_files"
DEST_DIR="demultiplexed/raw-paired"
mkdir -p "$EXTRACT_DIR" "$DEST_DIR"

unzip -o "$ZIP_FILE" -d "$EXTRACT_DIR"

find "$EXTRACT_DIR" -type f -name "*.fastq.gz" -exec mv {} "$DEST_DIR" \;
```

For subsequent operations, please use a custom temporary folder when using kodiak server as if the temporary files generated during any operation exceeds the given space, it will shut the operation down. Add this piece of code to your bash script every time

```
TEMP_DIR="tmp"
mkdir -p "$TEMP_DIR"
export TMPDIR="$TEMP_DIR"
export TMP="$TEMP_DIR"
export TEMP="$TEMP_DIR"
```

2. Trimming the sequences using Trimmomatic v0.39

Install Trimmomatic v0.39 prior to this. Include the path where trimmomatic-0.39.jar executable file is stored.

```
cd demultiplexed

mkdir -p trimmed-paired # to store trimmed and pair matched sequences
mkdir -p trimmed-unpaired # to store unmatched sequences

for f in $(find raw-paired -maxdepth 1 -name "*R1_001.fastq.gz" -printf "%f\n"
→   | sed 's/1_001.fastq.gz//' | sort -u)
do
java -jar /path/to/Trimmomatic-0.39/trimmomatic-0.39.jar PE -phred33
→   ${f}1_001.fastq.gz ${f}2_001.fastq.gz trimmed-paired/${f}1_001.fastq.gz
→   trimmed-unpaired/${f}1_001.fastq.gz trimmed-paired/${f}2_001.fastq.gz
→   trimmed-unpaired/${f}2_001.fastq.gz LEADING:3 TRAILING:3
→   SLIDINGWINDOW:4:15 MINLEN:35;
echo Processed: ${f}1_001.fastq.gz & ${f}2_001.fastq.gz
done
```

Storing the trimmed sequences as a qiime2 artifact (.qza)

```
source activate qiime2-metagenome-2024.10 #activating qiime2 metagenome
→   distribution

qiime tools import \
  --type 'SampleData[PairedEndSequencesWithQuality]' \
  --input-path demultiplexed/trimmed-paired \
  --input-format CasavaOneEightSingleLanePerSampleDirFmt \
  --output-path demux-paired/trimmed.qza

qiime demux summarize \
  --i-data demux-paired/trimmed.qza \
  --o-visualization demux-paired/trimmed.qzv  #summary visualization of the
  →   sequences
```

3. Filtering the sequences with Bowtie2

Bowtie2 requires a Bowtie database built from genomic data provided by user. Currently, the database used here was built from Human Genome Data gathered from NCBI. If you are using Bowtie2 standalone installation then you

can download it from the Bowtie indices for H. sapiens, GRCh38 no-alt analysis set from here. Here Bowtie2 was used from the qiime2 metagenome distribution platform, so the database has to be in a qimme2 artifact (.qza) format. For that, we can download the FASTA version of H. sapiens genome GRCh38 in NCBI from here, and build it into Bowtie2 datbase format using the following bash script.

```
cd /path/to/where/it/is/stored/GRCh38_FASTA

source activate qiime2-metagenome-2024.10

qiime tools import \                      #Importing the genome stored in .fna to
↪  .qza format
  --input-path GRCh38.fna \
  --output-path GRCh38.qza \
  --type 'FeatureData[Sequence]'

qiime quality-control bowtie2-build \    #Building Bowtie2 Index from it
    --i-sequences GRCh38.qza \
    --o-database Hsapiens_BowTie2Index.qza
```

After BowTie2 Indices are built, we can use that to filter human reads from the sequences using the following bash script.

```
source activate qiime2-metagenome-2024.10

qiime quality-control filter-reads \
    --i-demultiplexed-sequences demux-paired/trimmed.qza \             #Input
    ↪  trimmed sequences
    --i-database /path/to/where/it/is/stored/Hsapiens_BowTie2Index.qza \
    ↪  #Bowtie2 Index built from Human genome
    --o-filtered-sequences demux-paired/trimmed-filtered.qza          #Output
    ↪  trimmed & filtered sequences
```

4. Assembling the sequences into contigs using Megahit

The assemble contigs can be used for taxonomic classification using Kraken2 and functional annotation using eggNOG.

```
mkdir -p contigs

source activate qiime2-metagenome-2024.10

qiime assembly assemble-megahit \
    --i-seqs demux-paired/trimmed-filtered.qza \         #Input trimmed and
    ↪  filtered sequences
    --p-presets "meta-sensitive" \        #Preset which looks for more rarer
    ↪  species in the metagenome
    --p-num-cpu-threads 24 \        #number of cpus to be used
    --p-num-partitions 1 \
    --o-contigs contigs/trimmed-filtered-contigs.qza \
    --verbose
```

5. Taxonomic classification using Kraken2

We can use Kraken2 to classify both sequences and contigs using the same basic script. Kraken2 is a taxonomic classification system using exact k-mer matches to achieve high accuracy and fast classification speeds. This classifier matches each k-mer within a query sequence to the lowest common ancestor (LCA) of all genomes containing the given k-mer. You can install Kraken2 as a stand-alone, but here we have used kraken2 through the qiime2 metagenome platform's moshpit distribution package. Before that we will have to build databses for Kraken2 and Bracken using this script. We will be storing the database in the qiime cache. A cache will store the "data" of a dataset in a separate data folder with a random generated name, which is stored in a separate "key" file named after the dataset in the keys folder. In a qiime artifact (.qza), they are all bundled up together into one thing that has to zipped and unzipped everytime you use it.

```
source activate qiime2-metagenome-2024.10

qiime tools cache-create --cache cache        #creating the cache

qiime moshpit build-kraken-db \
    --p-collection standard \        #Using the standard Kraken2 and Bracken
    ↪   database
    --o-kraken2-database cache:kraken_standard \
    --o-bracken-database cache:bracken_standard \
    --verbose
```

After the databases are built we can go ahead and classify the data using Kraken2.

```
source activate qiime2-metagenome-2024.10

qiime moshpit classify-kraken2 \
    --i-seqs contigs/trimmed-filtered.qza \        #Input sequences/contigs
    --i-kraken2-db cache:kraken_standard \
    --p-threads 40 \        #Number of threads to use in computation
    --p-confidence 0.1 \        #Confidence level (0 to 1) of the matches that
    ↪   will be made, higher the value more stringent the algorithm
    --p-minimum-base-quality 0 \        #Minimum base quality at which Kraken2
    ↪   classifies a read
    --o-hits cache:kraken_hits-trimmed-filtered \        #Kraken2 hits stored
    ↪   in cache format
    --o-reports cache:kraken_reports-trimmed-filtered \        #Kraken2 reports
    ↪   stored in cache format
    --use-cache cache \
    --verbose \
  --p-memory-mapping False
```

Kraken outputs conatin a feature table (.biom), taxnomy file (.txt) and kraken reports (.txt). Kraken reports contain taxon information on each line: Percentage of fragments covered by the clade root, number of fragments covered by clade root, Number of fragments assigned directly to this taxon, a rank code: indicating (U)nclassified, (R)oot, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies, NCBI taxonomic ID number, and taxonomic annotation.

```
##   26.51    1453057 1453057 U   0    unclassified
##   73.49    4028832 338 R   1    root
##   73.49    4028400 1671    R1  131567     cellular organisms
##   73.44    4025812 61371   D   2        Bacteria
```

4

```
## 55.30    3031565 15   D1   1783270          FCB group
## 55.30    3031550 47   D2   68336              Bacteroidota/Chlorobiota group
## 55.30    3031503 10014 P   976                  Bacteroidota
## 55.12    3021465 124 C    200643                  Bacteroidia
## 55.11    3021340 251536 O   171549                  Bacteroidales
## 49.56    2716871 99066 F    815                      Bacteroidaceae
## 29.01    1590028 202688 G   909656                      Phocaeicola
## 25.00    1370421 1370421 S   821                        Phocaeicola vulgatus
##  0.28    15136   14805 S    357276                      Phocaeicola dorei
##  0.01    331 331 S1   997877                          Phocaeicola dorei CL03T12C01
##  0.03    1782    1782  S    387090                        Phocaeicola coprophilus
##  0.00    1   0   S    376805                        Phocaeicola salanitronis
##  0.00    1   1   S1   667015                          Phocaeicola salanitronis DSM 18170
## 18.75    1027777 505940 G   816                        Bacteroides
##  5.42    297365  294611 S   820                        Bacteroides uniformis
##  0.05    2754    2754  S1   997890                          Bacteroides uniformis CL03T12
##  3.27    179117  179117 S   28116                        Bacteroides ovatus
##  0.36    19970   19970 S    818                        Bacteroides thetaiotaomicron
##  0.10    5697    652 G1   2646097                      unclassified Bacteroides
##  0.03    1600    1600  S    2528203                        Bacteroides sp. A1C1
##  0.02    1100    1100  S    2785531                        Bacteroides sp. HF-162
```

Renormalization with Bracken

Bracken (Bayesian Reestimation of Abundance with KrakEN) is a highly accurate statistical method that computes the abundance of species in DNA sequences from a metagenomics sample. Bracken is a companion program to Kraken 2. While Kraken classifies reads to multiple levels in the taxonomic tree, Bracken allows estimation of abundance at a single level using those classifications (e.g. Bracken can estimate abundance of species within a sample). Bracken uses a Bracken database, the length of your reads and the kraken reports to give you a feature frequency table that renormalizes the data by dropping the unclassified reads and according to the Bayesian probability of the kraken hits being correct.

```
source activate qiime2-metagenome-2024.10

mkdir kraken-outputs

name="trimmed-filtered"

qiime moshpit estimate-bracken \
    --i-bracken-db cache:bracken_standard \
    --p-read-len 150 \
    --i-kraken-reports cache:kraken_reports-$name \
    --o-reports kraken-outputs/bracken-reports-$name.qza \
    --o-taxonomy kraken-outputs/taxonomy-bracken-$name.qza \
    --o-table kraken-outputs/table-bracken-$name.qza
```

The Bracken reports are very similar to Kraken2 reports except for having no unclassified reads

```
## 100.00   4028748 0   R   1    root
## 100.00   4028645 0   R1   131567    cellular organisms
## 99.95    4026906 0   D   2        Bacteria
## 76.64    3087472 0   D1   1783270      FCB group
## 76.64    3087472 0   D2   68336          Bacteroidota/Chlorobiota group
```

```
## 76.64   3087472 0   P   976             Bacteroidota
## 76.64   3087439 0   C   200643              Bacteroidia
## 76.64   3087438 0   O   171549               Bacteroidales
## 75.25   3031475 0   F   815              Bacteroidaceae
## 45.53   1834358 0   G   909656                 Phocaeicola
## 44.98   1812114 1812114 S   821                 Phocaeicola vulgatus
## 0.50 20141    20141   S   357276               Phocaeicola dorei
## 0.05 2101     2101    S   387090               Phocaeicola coprophilus
## 0.00 1   1   S   376805           Phocaeicola salanitronis
## 29.71    1197116 0   G   816               Bacteroides
## 13.24    533218  533218  S   820                  Bacteroides uniformis
## 12.94    521396  521396  S   28116                 Bacteroides ovatus
## 0.74 29918    0   G1  2646097            unclassified Bacteroides
## 0.25 10241    10241   S   2528203                Bacteroides sp. A1C1
## 0.24 9713     9713    S   2785531                Bacteroides sp. HF-162
```

The taxonomy file contains the NCBI taxa ID number and the string of taxonomic names leading to the species or the final taxonomic ranks for that ID.

```
## Feature ID   Taxon
## 821  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bactero:
## 357276   d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 387090   d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 376805   d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 820  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bactero:
## 28116    d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2528203  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2785531  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2763022  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2847299  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 556259   d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2755405  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 3117552  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2709390  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 818  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bactero:
## 817  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bactero:
## 371601   d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2792859  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
## 2715212  d__Bacteria;k__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Bacteroidales;f__Bact
```

The feature table (.biom) is a BIOM format matrix table with Feature IDs as the row number and the numbe rof reads for that feature in each samples as column data.

```
##        1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 1002546            0                          0
## 1002689            0                          0
## 1005039            0                          0
## 1005665            0                          0
## 1006               0                          0
## 1006155            0                          0
##        14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 1002546                               0                          0
## 1002689                               0                          0
```

```
## 1005039                                                      0                       0
## 1005665                                                      0                       0
## 1006                                                         0                       0
## 1006155                                                      0                       0
##         17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A
## 1002546                0                0                         0
## 1002689                0                0                         0
## 1005039                0                0                         0
## 1005665                0                0                         0
## 1006                   0                0                         0
## 1006155                0                0                         0
##         23-MDcol033-1A 25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A
## 1002546              0              0              0         0     0
## 1002689              0              0              0         0     0
## 1005039              0              0              0         0     0
## 1005665              0              0              0         0     0
## 1006                 0              0              0         0     0
## 1006155              0              0              0         0     0
##         3-MDcol002-5-1 30-8B 31-8B 33-6-2 36-40-1B 38-S-Swanson
## 1002546              0     0     0      0        0            0
## 1002689              0     0     0      1        0            0
## 1005039              0     0     1      0        0            0
## 1005665              0     0     0      0        0            0
## 1006                 0     0     0      0        0           59
## 1006155              0     0     0      6        0            0
##         40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B 46-MDCOL015-1a 48-MDCOL039-1a
## 1002546                0        0              0              3              0
## 1002689                0        0              0              0              0
## 1005039                0        0              0              0              0
## 1005665                0        0              1              0              0
## 1006                   0        0              0              0              0
## 1006155                0        0              0              0              0
##         5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2 50-MDCOL027-1a 52-MDCOL021-1b
## 1002546                                        0              0              0
## 1002689                                        0              0              0
## 1005039                                        0              0              0
## 1005665                                        0              0              0
## 1006                                           0              0              0
## 1006155                                        0              0              0
##         54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 1002546              1                0               0
## 1002689              0                0               0
## 1005039              0                0               0
## 1005665              1                0               0
## 1006                 0                0               0
## 1006155              0                0               0
##         65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 1002546                                           0              0
## 1002689                                           0              0
## 1005039                                           0              0
## 1005665                                           0              0
## 1006                                              0              0
## 1006155                                           0              0
##         68-MDCOL003-3-1-Before-Chemo-2-3A
```

```
## 1002546                                           0
## 1002689                                           0
## 1005039                                           0
## 1005665                                           0
## 1006                                               0
## 1006155                                           0
##           7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3 70-MDCOL004-3-2-Post-Chemo-A
## 1002546                                           0                           0
## 1002689                                           0                           0
## 1005039                                           0                           0
## 1005665                                           0                           0
## 1006                                               0                           0
## 1006155                                           0                           0
##           72-MDCOL003-4-1-4A-Before-Chemo-6 75-MDCOL053-1-1-1-Unknown
## 1002546                                           0                   0
## 1002689                                           0                   0
## 1005039                                           0                   0
## 1005665                                           0                   0
## 1006                                               0                   0
## 1006155                                           0                   0
##           9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1002546                                           0
## 1002689                                           0
## 1005039                                           0
## 1005665                                           0
## 1006                                               0
## 1006155                                           0
```

Creating phylogenetic trees from the bracken data

We can generate phylogenetic trees from the brcaken data using a package called gracken, which can be installed by `pip install gracken` or from this link (https://github.com/jonasoh/gracken). But gracken uses scientific names as phylogenetic tree tip labels whereas here the feature table contains the NCBI taxa IDs, so you can replace the executable python file for gracken in the miniconda path "miniconda/lib/python3.9/site-packages/gracken/" with this file gracken.py, which contains an updated code. Once you have it installed, this code will create the phylogentic tree for each dataset ("raw", "trimmed","filtered"... etc) and an OTU table , but we can ignore that.

```
mkdir -p bracken-outputs

source activate qiime2-metagenome-2024.10

for f in $(find kraken-outputs -maxdepth 2 -name "table-bracken*")
↪   #Converting the bracken feature tables from a .qza to folder format
do
filename=$(basename "$f" .qza)
qiime tools export \
    --input-path $f \
    --output-path bracken-outputs/$filename
echo $filename
done


for f in $(find kraken-outputs -maxdepth 2 -name "taxonomy-bracken*")
↪   #Converting the bracken taxonomy from a .qza to folder format
```

8

```
do
filename=$(basename "$f" .qza)
qiime tools export \
    --input-path $f \
    --output-path bracken-outputs/$filename
echo $filename
done

for f in $(find kraken-outputs -maxdepth 2 -name "bracken-reports*")
↪   #Converting the bracken reports from a .qza to folder format
do
filename=$(basename "$f" .qza)
qiime tools export \
    --input-path $f \
    --output-path bracken-outputs/$filename
echo $filename
done

find bracken-outputs -type d -name "bracken-reports-*" | while read dir; do
↪   #Create a phylogenetic tree for each dataset using the bracken reports
     gracken --input_dir $dir --out_prefix $dir/output --mode bracken
done
```

6. Taxonomic Classification using Kaiju

Kaiju is another taxonomic classification tool that assigns each sequencing read to a taxon in the NCBI taxonomy by comparing it to a reference protein database containing microbial and viral protein sequences. By using protein-level classification, Kaiju achieves a higher sensitivity compared with methods based on nucleotide comparison. You can use any of the several reference protein databases, such as complete genomes from NCBI RefSeq or the microbial subset of the NCBI BLAST non-redundant protein database nr, optionally also including fungi and microbial eukaryotes. We have used the NCBI RefSeq and have run the kaiju classifation through the qiime2 metagenome platform's moshpit distribution package. The databses can be built in the following manner and stored as a cache artifact.

```
source activate qiime2-metagenome-2024.10
qiime moshpit fetch-kaiju-db \
    --p-database-type 'refseq' \
    --o-database cache:KaijuDB_refseq \
    --verbose
```

After the databasehas been buit we can classify the datasets by Kaiju to any taxonomic level of choice. Here we have classiifed it to the species level, hence any reads that could not be assigned to any species level with enough confidence but can be assigned to other taxonomic ranks will be put in the "cannot be assigned to a (non-viral) species" category, which is different than the "unclassified" category, which contains reads that cannot be assigned to any taxonomic category at all. To classify using Kaiju, use the following code.

```
source activate qiime2-metagenome-2024.10
qiime moshpit classify-kaiju \
    --i-seqs /data/choudhurya/COLON_MD/contigs/filtered-contigs.qza \
    --i-db cache:KaijuDB_refseq \
    --p-z 40 \
    --p-a 'greedy' \
    --p-c 0 \
```

```
  --o-abundances cache:kaiju_abun-trimmed-filtered \      #Stores the
  ↪ abundance table
  --o-taxonomy cache:kaiju_taxonomy-trimmed-filtered \      #Stores the
  ↪ taxonomy file
  --use-cache cache \
  --verbose
```

The output from Kaiju is a BIOM formated feature table (.biom) and a taxonmy file.

Feature Table from Kaiju:

```
##          1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 0                107510                          128875
## 100886            1062                            1096
## 102148               0                               0
## 10239              297                             343
## 1042156            753                            1060
## 105841             836                            1093
##          14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 0                                           92791                    160866
## 100886                                        568                       856
## 102148                                          0                       733
## 10239                                        2204                     84395
## 1042156                                      2081                      2218
## 105841                                       1655                       937
##          17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A
## 0                  66562            69952                    166779
## 100886                 0                0                      1156
## 102148                 0                0                         0
## 10239                203              382                       520
## 1042156                0                0                         0
## 105841               455              861                      3894
##          23-MDcol033-1A 25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A
## 0                74759         103900         117685    118836 78654
## 100886               0            872            843       710   477
## 102148               0              0              0         0     0
## 10239            34575            514          17332       790   737
## 1042156            621           1146           1993       902   487
## 105841            3625           6632           1159       908   504
##          3-MDcol002-5-1 30-8B   31-8B 33-6-2 36-40-1B 38-S-Swanson
## 0                154682 42309 129571 103313   171345       234863
## 100886              828   294    982    635     1406         2094
## 102148                0     0    546    395        0            0
## 10239               612   448    569    260      513         2846
## 1042156            1068   396   1662   1230     1234         3697
## 105841             1192   387   1981   1360     1506         2760
##          40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B 46-MDCOL015-1a 48-MDCOL039-1a
## 0                  189502   115185         118140         134839         135318
## 100886               1769      600              0            572            867
## 102148                  0        0              0            851              0
## 10239                 388      816           8264            317           4951
## 1042156              1597      827           1787           1930           2703
## 105841               1913     1098           2959           6296           1430
```

```
##         5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2 50-MDCOL027-1a 52-MDCOL021-1b
## 0                                          60205        164777         56860
## 100886                                       449          7126           464
## 102148                                         0           794          1648
## 10239                                        215          5659         58741
## 1042156                                      462          2630          1060
## 105841                                       438          1403         21535
##         54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 0                85864           101022           37472
## 100886               0                0               0
## 102148             809                0               0
## 10239             1693              242           13305
## 1042156              0             1153               0
## 105841            1179             5277               0
##         65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 0                                            75672          10878
## 100886                                         807             51
## 102148                                           0             82
## 10239                                         1695           3027
## 1042156                                       2534            104
## 105841                                        1480             76
##         68-MDCOL003-3-1-Before-Chemo-2-3A
## 0                                   63818
## 100886                                425
## 102148                                  0
## 10239                                3053
## 1042156                              1489
## 105841                               1098
##         7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3 70-MDCOL004-3-2-Post-Chemo-A
## 0                                          47845                         2676
## 100886                                       329                            0
## 102148                                         0                            0
## 10239                                        152                          725
## 1042156                                      396                           24
## 105841                                       385                           21
##         72-MDCOL003-4-1-4A-Before-Chemo-6 75-MDCOL053-1-1-1-Unknown
## 0                                   48410                     88527
## 100886                                377                       846
## 102148                                  0                         0
## 10239                                 406                      1170
## 1042156                               737                      1557
## 105841                                864                       746
##         9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 0                                         110766
## 100886                                       998
## 102148                                         0
## 10239                                        768
## 1042156                                     2019
## 105841                                      1065
```

Taxonomy from Kaiju:

```
##   Feature.ID
## 1          0
```

```
## 2      100886
## 3      102148
## 4       10239
## 5     1042156
## 6      105841
##
## 1                                                                  d__belong to a (non-
## 2 d__Bacteria;p__Bacillota;c__Erysipelotrichia;o__Erysipelotrichales;f__Coprobacillaceae
## 3     d__Bacteria;p__Bacillota;c__Erysipelotrichia;o__Erysipelotrichales;f__Erysipelotri
## 4
## 5                      d__Bacteria;p__Bacillota;c__Clostridia;o__Eubacteriales;f__Clostri
## 6                      d__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachno
```

7. Taxonomic Classification using Metaphlan

MetaPhlAn 4 relies on unique clade-specific marker genes identified from ~1M microbial genomes for taxonomic assignment of metgenome sequence reads. Metaphlan 4.1.1 was used as a standalone installation as a miniconda environment, whose installation information can be gathered from here (https://github.com/biobakery/MetaPhlAn/wiki/MetaPhlAn-4.1). Metaphlan automatically downloads its latest database "ChocoPhlan" when you use it for the first time, but you can also do it manually with this `$ metaphlan --install --index <latest database index> --bowtie2db <database folder>`. The latest database indices can be found here (http://cmprod1.cibio.unitn.it/biobakery4/metaphlan_databases/). Once everything is installed, the following code was used to perform classification using Metaphlan for our datasets.

```
name="trimmed-filtered"          #dataset name

source activate qiime2-metagenome-2024.10

qiime tools export \          #Fetching the datset from .qza format to a folder
↪  format
    --input-path demux-paired/$name \
    --output-path demultiplexed/$name

conda activate mpa           #Activating the miniconda environment, here named as
↪  "mpa"

mkdir -p MetaphlanAnalysis/BowTieOutput/$name MetaphlanAnalysis/Output/$name
↪  #folders for storing Metaphlan output and the intermediated Bowtie
↪  alignment ouput

folder="demultiplexed/$name"          #folder storing our sequences

# Loop over sample names
for f in $(find ${folder} -maxdepth 1 -name "*R1_001.fastq.gz" | sed 's|.*/||'
↪  | sed 's/R1_001.fastq.gz//' | sort -u)
do
    # Define output filename without _S[number]_L[number]_R
    clean_f=$(echo ${f} | sed -E 's/_S[0-9]+_L[0-9]+_//')

    # Run MetaPhlAn
     metaphlan ${folder}/${f}R1_001.fastq.gz,${folder}/${f}R2_001.fastq.gz
        ↪  --bowtie2db <database folder> --bowtie2out
        ↪  MetaphlanAnalysis/BowTieOutput/$name/${clean_f}.bowtie2.bz2
        ↪  --input_type fastq --nproc 28 -o
        ↪  MetaphlanAnalysis/Output/$name/${clean_f}.output.txt
```

```
      echo ${clean_f} is done
done
```

A typical Metaphlan output contains taxonomic relative abundances

```
## [1]   #mpa_vOct22_CHOCOPhlAnSGB_202403
## [2]   #/data/choudhurya/miniconda/envs/mpa/bin/metaphlan demultiplexed/trimmed-filtered-p
## [3]   #10709265 reads processed
## [4]   #SampleID   Metaphlan_Analysis
## [5]   #clade_name NCBI_tax_id relative_abundance  additional_species
## [6]   k__Bacteria 2   100.0
## [7]   k__Bacteria|p__Firmicutes   2|1239  51.68535
## [8]   k__Bacteria|p__Bacteroidetes    2|976   46.48729
## [9]   k__Bacteria|p__Actinobacteria   2|201174    1.22089
## [10]  k__Bacteria|p__Proteobacteria   2|1224  0.60648
## [11]  k__Bacteria|p__Bacteroidetes|c__Bacteroidia    2|976|200643    46.42399
## [12]  k__Bacteria|p__Firmicutes|c__Clostridia    2|1239|186801   29.73363
## [13]  k__Bacteria|p__Firmicutes|c__CFGB3002   2|1239| 15.59072
## [14]  k__Bacteria|p__Firmicutes|c__CFGB3012   2|1239| 2.14771
## [15]  k__Bacteria|p__Firmicutes|c__Negativicutes 2|1239|909932   1.73986
## [16]  k__Bacteria|p__Firmicutes|c__CFGB3005   2|1239| 1.38625
## [17]  k__Bacteria|p__Actinobacteria|c__Coriobacteriia    2|201174|84998  1.22089
## [18]  k__Bacteria|p__Proteobacteria|c__Betaproteobacteria    2|1224|28216    0.46794
## [19]  k__Bacteria|p__Firmicutes|c__CFGB3054   2|1239| 0.25919
## [20]  k__Bacteria|p__Firmicutes|c__Firmicutes_unclassified   2|1239| 0.24734
```

The file has a five line header (lines beginning with #). We've added [i]s (line numbers) to the header text above to help call out their roles more clearly.

1. Indicates the version of the marker gene database that MetaPhlAn used in this run. There are currently ~1.1M unique clade-specific marker genes identified from ~100k reference genomes (~99,500 bacterial and archaeal and ~500 eukaryotic) included in the database.
2. Provides a copy of the command that was run to produce this profile. This includes the path to the MetaPhlAn executable, the input file that was analyzed, and any custom parameters that were configured.
3. Indicates the number of sample reads processed.
4. Lists your sample name.
5. Provides column headers from the profile data that follows. More specifically, [5] includes four headers to be familiar with:

clade_name: The taxonomic lineage of the taxon reported on this line, ranging from Kingdom (e.g. Bacteria/Archaea) to species genome bin (SGB). Taxon names are prefixed to help indicate their rank: Kingdom: k__, Phylum: p, **Class: c**, Order: o, **Family: f**, Genus: g, **Species: s**, and SGB: t__.

NCBI_tax_id: The NCBI-equivalent taxon IDs of the named taxa from clade_name.

relative_abundance: The taxon's relative abundance in %. Since typical shotgun-sequencing-based taxonomic profile is relative (i.e. it does not provide absolute cell counts), clades are hierarchically summed. Each taxonomic level will sum to 100%. That is, the sum of all kingdom-level clades is 100%, the sum of all phylum-level clades is 100%, and so forth.

additional_species: Additional species names for cases where the metagenome profile contains clades that represent multiple species. The species listed in column 1 is the representative species in such cases.

8. Functional annotation using Humann 3.9

HUMAnN is a pipeline for efficiently and accurately profiling the presence/absence and abundance of microbial pathways in a community from metagenomic or metatranscriptomic sequencing data (typically millions of short DNA/RNA reads). This process, referred to as functional profiling, aims to describe the metabolic potential of a microbial community and its members. You can install Humann 3.9 as a stand-alone through here (https://github.com/biobakery/humann). Once installed, it automatically downloads the ChocoPhlan database (if not already installed) to do taxonomic analysis using Metaphlan on its first run, but you will have to download the translated search database for anotation. You can do it using this code `humann_databases --download uniref uniref90_diamond $INSTALL_LOCATION`, which will download the full UniRef90 database used in our analysis. After that, run this code to annotate the sequences using HUMAnN. (P.S: Unlike MetaPhlan, HUMAnN cannpot parse paired read sequences, so we will have to first concatenate both reads from each sample into a single .fastq file and then send it off for functional annotation)

```
name="trimmed-filtered"        #Dataset name

mkdir -p HumannAnalysis/HumannOutput/$name        #where the output is to be
 ↪  stored

folder="demultiplexed/$name"        #Dataset folder

cd $folder
forward=$(echo *R1*.fastq.gz)
reverse=$(echo *R2*.fastq.gz)

mkdir -p concatenated

for f in $forward
do
    for r in $reverse
    do
        clean_f=$(echo ${f} | sed -E
         ↪  's/_S[0-9]+_L[0-9]+_R[0-9]_[0-9]+.fastq.gz//')
        clean_r=$(echo ${r} | sed -E
         ↪  's/_S[0-9]+_L[0-9]+_R[0-9]_[0-9]+.fastq.gz//')

        if [ $clean_f = $clean_r ]; then
            cat $f $r > concatenated/$clean_f.fastq.gz        #concatenating
             ↪  the forward and reverse reads of the same sample
            humann --input concatenated/$clean_f.fastq.gz --threads 16
             ↪  --output HumannAnalysis/HumannOutput/$name
            fi
    done
done
rm -rf concatenated        #deleting the folder storing the conactenated reads
```

HUMAnN output consists of three files: Gene families file (`$SAMPLENAME_genefamilies.tsv`), Path Coverage file (`$SAMPLENAME_pathcoverage.tsv`), and Path Abundance file (`$SAMPLENAME_pathabundance.tsv`).

The **gene families file** contains the the abundance of each gene family in the community. Gene families are groups of evolutionarily-related protein-coding sequences that often perform similar functions. Gene family abundance at the community level is stratified to show the contributions from known and unknown species. Individual species' abundance contributions sum to the community total abundance. Gene family abundance is reported in RPK (reads per kilobase) units to normalize for gene length; RPK units reflect relative gene (or transcript) copy number in the community.

```
##                                                      Gene-Family
## 1                                                        UNMAPPED
## 2                                                UniRef90_A0A174NA57
## 3 UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## 4                                                UniRef90_A0A395VRV3
## 5    UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## 6  UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##    1-MDcol002-5-2_Abundance-RPKs
## 1                 2364890.000
## 2                   55247.258
## 3                   55247.258
## 4                   49394.382
## 5                   25032.358
## 6                    8720.539
```

The **path coverage file** provides information on the completeness of metabolic pathways in a given sample. It contains two main columns: Pathway, which lists metabolic pathway identifiers, and Coverage, which represents the proportion of the pathway detected in the sample as a value between 0 and 1. A coverage value of 0.00 indicates that the pathway is absent, 0.50 suggests that half of its reactions are present, and 1.00 means the pathway is fully reconstructed. These pathway identifiers are derived from the MetaCyc database.

```
##                                                      Pathway
## 1                                                      UNMAPPED
## 2                                                  UNINTEGRATED
## 3 UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## 4          UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## 5    UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## 6          UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##    1-MDcol002-5-2_Coverage
## 1                        1
## 2                        1
## 3                        1
## 4                        1
## 5                        1
## 6                        1
```

The **path abundance file** details the abundance of each pathway in the community as a function of the abundances of the pathway's component reactions, with each reaction's abundance computed as the sum over abundances of genes catalyzing the reaction. Pathway abundance is proportional to the number of complete "copies" of the pathway in the community but unlike gene abundance, a pathway's community-level abundance is not necessarily the sum of its stratified abundance values. It consists of two main columns: Pathway, which contains metabolic pathway identifiers from MetaCyc, and Abundance, which represents the cumulative abundance of reactions contributing to the pathway

```
##                                                Pathway 1-MDcol002-5-2_Abundance
## 1                                               UNMAPPED                778748.2
## 2                                           UNINTEGRATED               2481983.1
## 3  UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                917463.5
## 4 UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                464620.0
## 5    UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                354114.6
## 6                                 UNINTEGRATED|unclassified                147854.1
```

After gathering all the data, we will move to performing furtehr data transfromation in R in steps 9 onwards.

Required packages for the following R code

```r
library(biomformat)
library(tidyverse)
library(data.table)
library(fs)
library(phyloseq)
library(writexl)
library(phylostratr)
library(ape)
library(taxize)
```

9. Transforming Kaiju Data

The following code is to read the Kaiju feature table (.biom) and Kaiju taxonomy, merge them together to create an abundance table that has both the taxonomy and the reads in each sample for each taxa, provided all the biom files and taxonomy files are kept in a folder named "Kaiju Taxonomy Abundance".

```r
Kaijubioms<-dir("Kaiju Taxonomy Abundance",pattern="*biom")     #Listing all
↪ biom files from the kaiju folder
Kaijutaxa<-dir("Kaiju Taxonomy Abundance",pattern="kaiju_taxonomy*")  #Listing
↪ all taxonomy files from the kaiju folder
for (biom in Kaijubioms){
  for(taxa in Kaijutaxa){
    if(substring(biom,21,nchar(biom)-5)==substring(taxa,16,nchar(taxa)-4))
      name=substring(biom,21,nchar(biom)-5)
    OTUtable<-as.data.frame(as.matrix(biom_data(read_biom(paste("Kaiju
↪ Taxonomy Abundance",biom,sep="/")))))%>% #Reading biom files
      rownames_to_column("Feature.ID")
    taxatable<-read.table(paste("Kaiju Taxonomy
↪ Abundance",taxa,sep="/"),sep='\t',header = TRUE)       #Reading the
↪ taxonomy file
    Abundance<-inner_join(OTUtable,taxatable)%>%       #merging them together
↪ and separating the taxonomy string into separate columns for each
↪ taxonomic ranks
      select(-"Feature.ID")%>%
      relocate(Taxon)%>%
      mutate(Taxon=str_replace_all(Taxon,".___",""))%>%

      ↪  separate(Taxon,c("Domain","Phylum","Class","Order","Family","Genus","Species"),se
      mutate(Phylum=if_else(is.na(Phylum),Domain,Phylum),
             Class=if_else(is.na(Class),Phylum,Class),
             Order=if_else(is.na(Order),Class,Order),
             Family=if_else(is.na(Family),Order,Family),
             Genus=if_else(is.na(Genus),Family,Genus),
             Species=if_else(is.na(Species),Genus,Species))
    write_xlsx(Abundance,paste("Kaiju Taxonomy
    ↪ Abundance/kaiju_abun-",name,".xlsx",sep=""))   #Writing the table as
    ↪ an excel file
  }
}
```

Sample Kaiju feature table

```
##    Feature.ID 1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 1          0         107567                       128946
## 2     100886           1063                         1096
## 3     102148              0                            0
## 4      10239            297                          344
## 5    1042156            754                         1061
## 6     105841            837                         1093
##    14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 1                                      92867                   164303
## 2                                        568                      856
## 3                                          0                      734
## 4                                       2207                    84670
## 5                                       2083                     2220
## 6                                       1655                      938
##    17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A 23-MDcol033-1A
## 1            66642            70074                    166933          74818
## 2                0                0                      1157              0
## 3                0                0                         0              0
## 4              204              383                       520          34634
## 5                0                0                         0            622
## 6              455              861                      3895           3628
##    25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A 3-MDcol002-5-1 30-8B   31-8B
## 1          104008         117792    118937 78699         154777 42339  129679
## 2             872            844       710   478            828   294     983
## 3               0              0         0     0              0     0     546
## 4             515          17373       791   737            612   448     572
## 5            1147           1995       903   487           1068   396    1662
## 6            6639           1165       908   504           1193   387    1983
##    33-6-2 36-40-1B 38-S-Swanson 40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B
## 1 103400   171450       235039           189626   115278         118226
## 2    635     1406         2095             1769      600              0
## 3    395        0            0                0        0              0
## 4    261      513         2850              388      818           8271
## 5   1231     1234         3699             1598      828           1787
## 6   1360     1506         2764             1913     1098           2961
##    46-MDCOL015-1a 48-MDCOL039-1a 5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2
## 1         134994         135414                                    60250
## 2            572            867                                      449
## 3            853              0                                        0
## 4            319           4953                                      215
## 5           1931           2704                                      462
## 6           6302           1430                                      438
##    50-MDCOL027-1a 52-MDCOL021-1b 54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 1         164912          56900          86452           101341           37508
## 2           7133            466              0                0               0
## 3            794           1650            810                0               0
## 4           5663          58791           1695              331           13311
## 5           2632           1061              0             1153               0
## 6           1403          21548           1187             5279               0
##    65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 1                                        77198          33754
## 2                                          807              0
```

```
## 3                                                                         0                 0
## 4                                                                      1801              5560
## 5                                                                      2536                 0
## 6                                                                      1480                 0
##   68-MDCOL003-3-1-Before-Chemo-2-3A 7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3
## 1                               63548                                            47689
## 2                                 425                                              331
## 3                                   0                                                0
## 4                                3058                                              152
## 5                                1490                                              397
## 6                                1099                                              385
##   70-MDCOL004-3-2-Post-Chemo-A 72-MDCOL003-4-1-4A-Before-Chemo-6
## 1                         13449                             48576
## 2                             0                               377
## 3                             0                                 0
## 4                          3811                               416
## 5                             0                               737
## 6                             0                               865
##   75-MDCOL053-1-1-1-Unknown 9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                     88590                                    110828
## 2                       847                                       999
## 3                         0                                         0
## 4                      1170                                       768
## 5                      1560                                      2020
## 6                       747                                      1068
```

Sample Kaiju taxonomy file

```
##   Feature.ID
## 1          0
## 2     100886
## 3     102148
## 4      10239
## 5    1042156
## 6     105841
##
## 1                                                                       d__belong to a (non-
## 2 d__Bacteria;p__Bacillota;c__Erysipelotrichia;o__Erysipelotrichales;f__Coprobacillaceae
## 3     d__Bacteria;p__Bacillota;c__Erysipelotrichia;o__Erysipelotrichales;f__Erysipelotri
## 4
## 5                  d__Bacteria;p__Bacillota;c__Clostridia;o__Eubacteriales;f__Clostri
## 6                  d__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachno
```

Sample Kaiju combined abundance table

```
##                                                                       Domain
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                                     Bacteria
## 3                                                                     Bacteria
## 4                                                                      Viruses
## 5                                                                     Bacteria
## 6                                                                     Bacteria
##                                                                       Phylum
```

```
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                              Bacillota
## 3                                                              Bacillota
## 4                                                               Viruses
## 5                                                              Bacillota
## 6                                                              Bacillota
##                                                                  Class
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                       Erysipelotrichia
## 3                                                       Erysipelotrichia
## 4                                                               Viruses
## 5                                                             Clostridia
## 6                                                             Clostridia
##                                                                  Order
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                      Erysipelotrichales
## 3                                                      Erysipelotrichales
## 4                                                               Viruses
## 5                                                           Eubacteriales
## 6                                                          Lachnospirales
##                                                                 Family
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                      Coprobacillaceae
## 3                                                     Erysipelotrichaceae
## 4                                                               Viruses
## 5                                                          Clostridiaceae
## 6                                                         Lachnospiraceae
##                                                                  Genus
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                        Catenibacterium
## 3                                                          Solobacterium
## 4                                                               Viruses
## 5                                                             Clostridium
## 6                                                            Anaerostipes
##                                                                Species
## 1 belong to a (non-viral) species with less than 0.01% of all reads
## 2                                                 Catenibacterium mitsuokai
## 3                                                   Solobacterium moorei
## 4                                                               Viruses
## 5                                                 Clostridium sp. SY8519
## 6                                                    Anaerostipes caccae
##   1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 1         107567                         128946
## 2           1063                           1096
## 3              0                              0
## 4            297                            344
## 5            754                           1061
## 6            837                           1093
##   14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 1                                     92867                   164303
## 2                                       568                      856
## 3                                         0                      734
## 4                                      2207                    84670
## 5                                      2083                     2220
```

19

```
## 6                                            1655                       938
##    17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A 23-MDcol033-1A
## 1            66642            70074                    166933          74818
## 2                0                0                      1157              0
## 3                0                0                         0              0
## 4              204              383                       520          34634
## 5                0                0                         0            622
## 6              455              861                      3895           3628
##    25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A 3-MDcol002-5-1 30-8B  31-8B
## 1         104008         117792    118937 78699         154777 42339 129679
## 2            872            844       710   478            828   294    983
## 3              0              0         0     0              0     0    546
## 4            515          17373       791   737            612   448    572
## 5           1147           1995       903   487           1068   396   1662
## 6           6639           1165       908   504           1193   387   1983
##    33-6-2 36-40-1B 38-S-Swanson 40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B
## 1 103400   171450       235039           189626   115278         118226
## 2    635     1406         2095             1769      600              0
## 3    395        0            0                0        0              0
## 4    261      513         2850              388      818           8271
## 5   1231     1234         3699             1598      828           1787
## 6   1360     1506         2764             1913     1098           2961
##    46-MDCOL015-1a 48-MDCOL039-1a 5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2
## 1         134994         135414                                    60250
## 2            572            867                                      449
## 3            853              0                                        0
## 4            319           4953                                      215
## 5           1931           2704                                      462
## 6           6302           1430                                      438
##    50-MDCOL027-1a 52-MDCOL021-1b 54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 1         164912          56900          86452           101341           37508
## 2           7133            466              0                0               0
## 3            794           1650            810                0               0
## 4           5663          58791           1695              331           13311
## 5           2632           1061              0             1153               0
## 6           1403          21548           1187             5279               0
##    65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 1                                        77198          33754
## 2                                          807              0
## 3                                            0              0
## 4                                         1801           5560
## 5                                         2536              0
## 6                                         1480              0
##    68-MDCOL003-3-1-Before-Chemo-2-3A 7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3
## 1                             63548                                     47689
## 2                               425                                       331
## 3                                 0                                         0
## 4                              3058                                       152
## 5                              1490                                       397
## 6                              1099                                       385
##    70-MDCOL004-3-2-Post-Chemo-A 72-MDCOL003-4-1-4A-Before-Chemo-6
## 1                        13449                            48576
## 2                            0                              377
## 3                            0                                0
```

20

```
## 4                                     3811                                       416
## 5                                        0                                       737
## 6                                        0                                       865
##    75-MDCOL053-1-1-1-Unknown 9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                     88590                                       110828
## 2                       847                                          999
## 3                         0                                            0
## 4                      1170                                          768
## 5                      1560                                         2020
## 6                       747                                         1068
```

Creating Phyloseq Objects for the Kaiju data

The following code packages the feature table and the taxonomy file into a single phyloseq object for the datasets
("raw", "trimmed-filtered" etc). A Phyloseq object usually contains,

- a **sample metadata**, which contains the sample IDs and metadata pertaining to it.
- a **feature table**, which is a matrix where the rows are the taxa/feature and columns are the sample IDs with read
  count for each taxa/feature.
- a **taxonomy file**, which is a matrix where the rows are the taxa/feature and columns are the taxonomy path for
  each taxa/feature.
- a **phylogenetic tree**, a phylogenetic tree with the taxa/feature as the tip labels.

The phyloseq object can be made with one or more of these data file.

```r
Kaijubioms<-dir("Kaiju Taxonomy Abundance",pattern="*biom")    #Listing all
↪    biom files from the kaiju folder
Kaijutaxa<-dir("Kaiju Taxonomy Abundance",pattern="kaiju_taxonomy*")   #Listing
↪    all taxonomy files from the kaiju folder
dir.create("PhyloseqOutputs/KaijuPhyloseq")
for (biom in Kaijubioms){
  for(taxa in Kaijutaxa){
    if(substring(biom,21,nchar(biom)-5)==substring(taxa,16,nchar(taxa)-4))
      name=substring(biom,21,nchar(biom)-5)
    OTUtable<-as.matrix(biom_data(read_biom(paste("Kaiju Taxonomy
↪    Abundance",biom,sep="/"))))  #Reading biom files
    taxatable<-read.table(paste("Kaiju Taxonomy
↪    Abundance",taxa,sep="/"),sep='\t',header = TRUE)%>%       #Reading the
↪    taxonomy file
      mutate(Taxon=str_replace_all(Taxon,".__","")))%>%

       ↪    separate(Taxon,c("Domain","Phylum","Class","Order","Family","Genus","Species"),se
      mutate(Phylum=if_else(is.na(Phylum),Domain,Phylum),
             Class=if_else(is.na(Class),Phylum,Class),
             Order=if_else(is.na(Order),Class,Order),
             Family=if_else(is.na(Family),Order,Family),
             Genus=if_else(is.na(Genus),Family,Genus),
             Species=if_else(is.na(Species),Genus,Species))%>%
      column_to_rownames("Feature.ID")%>%
      as.matrix()
    physeq <- phyloseq(otu_table(OTUtable, taxa_are_rows = TRUE),
↪    tax_table(taxatable))    #Compiling the phyloseq object

    ↪    saveRDS(physeq,paste("PhyloseqOutputs/KaijuPhyloseq/phyloseq_kaiju-",name,".rds",se
    ↪    #Saving it as an .RDS file
```

```
    }
}
```

Summary of the sample phyloseq object

```
##   Length   Class    Mode
##       1 phyloseq     S4


## phyloseq-class experiment-level object
## otu_table()   OTU Table:       [ 621 taxa and 37 samples ]
## tax_table()   Taxonomy Table:  [ 621 taxa by 7 taxonomic ranks ]
```

10. Transforming Kraken2-Bracken Data

The following code is to read the Bracken feature table (.biom) and Bracken taxonomy, merge them together to create an abundance table that has both the taxonomy and the reads in each sample for each taxa.

```
brackentables<-dir(path="BrackenOutputs",pattern="table-bracken*")    #Listing
↪  all the folders containing the bracken feature table
brackentaxonomy<-dir(path="BrackenOutputs",pattern="taxonomy-bracken*")
↪  #Listing all the folders containing the bracken taxonomy

dir.create("Kraken-Bracken Taxonomy Abundance")
for(each in brackentables){
  name<-substring(each,15)
  for(each2 in brackentaxonomy){
    name2<-substring(each2,18)
    if(name==name2)

↪  {table<-as.data.frame(as.matrix(biom_data(read_biom(paste("BrackenOutputs",each,"featur
↪  #Reading biom file
      rownames_to_column("Feature.ID")

↪  taxonomy<-read.table(paste("BrackenOutputs",each2,"taxonomy.tsv",sep='/'),sep="\t",hea
↪  = TRUE)%>%  #Reading taxonomy file
      mutate(Feature.ID=as.character(Feature.ID))
    taxatable<-inner_join(table,taxonomy)%>%  #Joining into an abundance table
      select(-"Feature.ID")%>%
      relocate(Taxon)%>%
      mutate(Taxon=str_replace_all(Taxon,".__",""))%>%

      ↪  separate(Taxon,c("Domain","Kingdom","Phylum","Class","Order","Family","Genus","Sp
    write_xlsx(taxatable,paste("Kraken-Bracken Taxonomy
      ↪  Abundance/Kraken-Bracken_abun-",name,".xlsx",sep="")) #Writing the
      ↪  abundance table as excel files
    }
  }
}
```

Sample Bracken feature table, taxonomy and combined abundance table

```
##    Feature.ID 1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 1      100886              5                               8
## 2      102148              0                               0
## 3      102684              0                               0
## 4     1042156              0                               0
## 5      104609              0                               0
## 6     1051631              0                               0
##    14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 1                                          0                        1
## 2                                          0                        2
## 3                                          0                        0
## 4                                          0                        3
## 5                                          0                        0
## 6                                          1                        0
##    17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A 23-MDcol033-1A
## 1                9                5                        12              5
## 2                0                0                         0              0
## 3                0                0                         0              0
## 4                0                0                         1              1
## 5                0                0                         0              0
## 6                0                0                         0              0
##    25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A 3-MDcol002-5-1 30-8B 31-8B
## 1               1              2        11     7              3     4    19
## 2               0              0         0     0              0     0     0
## 3               0              0         0     0              0     0     0
## 4               0              3         0     0              0     0     1
## 5               0              0         0     0              0     0     0
## 6               0              0         0     0              0     0     0
##    33-6-2 36-40-1B 38-S-Swanson 40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B
## 1      7        8           10                2        2              0
## 2      0        0            0                0        0              0
## 3      0        0            0                0        0              0
## 4      0        0            0                3        0              0
## 5      0        0            0                0        0              0
## 6      0        0            0                1        0              2
##    46-MDCOL015-1a 48-MDCOL039-1a 5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2
## 1               4              8                                        4
## 2               0              0                                        0
## 3               0              1                                        0
## 4               3              8                                        0
## 5               0              0                                        0
## 6               0              0                                        0
##    50-MDCOL027-1a 52-MDCOL021-1b 54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 1               55              2              0                4               0
## 2                1             40              0                0               0
## 3                0              0              0                0               0
## 4                5             10              0                0               0
## 5                0              0              0                0               0
## 6                0              0              0                0               0
##    65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 1                                            2              0
## 2                                            0              1
## 3                                            0              0
## 4                                            7              0
```

```
## 5                                              1              0
## 6                                              5              0
##   68-MDCOL003-3-1-Before-Chemo-2-3A 7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3
## 1                                 1                                        4
## 2                                 0                                        1
## 3                                 0                                        0
## 4                                16                                        0
## 5                                 1                                        0
## 6                                 0                                        0
##   70-MDCOL004-3-2-Post-Chemo-A 72-MDCOL003-4-1-4A-Before-Chemo-6
## 1                            0                                 0
## 2                            0                                 0
## 3                            0                                 0
## 4                            0                                19
## 5                            0                                 0
## 6                            0                                 0
##   75-MDCOL053-1-1-1-Unknown 9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                        24                                       18
## 2                         0                                        0
## 3                         0                                        0
## 4                        11                                        4
## 5                         0                                        0
## 6                         0                                        0


##   Feature.ID
## 1      40520
## 2     418240
## 3    2479767
## 4      89014
## 5      33035
## 6       1532
##
## 1      d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnosp
## 2    d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnospira
## 3 d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnospiracea
## 4       d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnos
## 5    d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnospira
## 6   d__Bacteria;k__Bacteria;p__Bacillota;c__Clostridia;o__Lachnospirales;f__Lachnospirac


##    Domain        Kingdom        Phylum              Class
## 1 Bacteria      Bacteria      Bacillota   Erysipelotrichia
## 2 Bacteria      Bacteria      Bacillota   Erysipelotrichia
## 3 Bacteria      Bacteria      Bacillota            Bacilli
## 4 Bacteria      Bacteria      Bacillota          Clostridia
## 5 Bacteria      Bacteria Pseudomonadota Gammaproteobacteria
## 6  Viruses Heunggongvirae   Uroviricota      Caudoviricetes
##                        Order             Family
## 1           Erysipelotrichales   Coprobacillaceae
## 2           Erysipelotrichales Erysipelotrichaceae
## 3             Lactobacillales   Streptococcaceae
## 4              Eubacteriales     Clostridiaceae
## 5                 Vibrionales       Vibrionaceae
## 6 containing Aliceevansviridae   Aliceevansviridae
##                                                   Genus             Species
```

```
## 1                                          Catenibacterium    Catenibacterium mitsuokai
## 2                                          Solobacterium          Solobacterium moorei
## 3                                          Streptococcus    Streptococcus infantarius
## 4                                             Clostridium        Clostridium sp. SY8519
## 5                                                  Vibrio           Vibrio penaeicida
## 6 containing Streptococcus phage YMC-2011 Streptococcus phage YMC-2011
##   1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## 1              5                                8
## 2              0                                0
## 3              0                                0
## 4              0                                0
## 5              0                                0
## 6              0                                0
##   14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4 15-MDcol004-3-1-pretreat
## 1                                         0                        1
## 2                                         0                        2
## 3                                         0                        0
## 4                                         0                        3
## 5                                         0                        0
## 6                                         1                        0
##   17-MDCOL006-3-1A 19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A 23-MDcol033-1A
## 1                9                5                        12              5
## 2                0                0                         0              0
## 3                0                0                         0              0
## 4                0                0                         1              1
## 5                0                0                         0              0
## 6                0                0                         0              0
##   25-MDcol036-1A 27-MDcol025-1B 28-7A-1-1 29-7A 3-MDcol002-5-1 30-8B 31-8B
## 1              1              2        11     7              3     4    19
## 2              0              0         0     0              0     0     0
## 3              0              0         0     0              0     0     0
## 4              0              3         0     0              0     0     1
## 5              0              0         0     0              0     0     0
## 6              0              0         0     0              0     0     0
##   33-6-2 36-40-1B 38-S-Swanson 40-MDCol002-7-43 42-51-1B 45-MDCOL005-1B
## 1      7        8           10                2        2              0
## 2      0        0            0                0        0              0
## 3      0        0            0                0        0              0
## 4      0        0            0                3        0              0
## 5      0        0            0                0        0              0
## 6      0        0            0                1        0              2
##   46-MDCOL015-1a 48-MDCOL039-1a 5-MDCOL055-Stool-1A-Baseline-BSB-2-4-1-2
## 1              4              8                                        4
## 2              0              0                                        0
## 3              0              1                                        0
## 4              3              8                                        0
## 5              0              0                                        0
## 6              0              0                                        0
##   50-MDCOL027-1a 52-MDCOL021-1b 54-MDCOL028-1b 60-MDcol006-4-1A 63-MDcol003-3-2
## 1             55              2              0                4               0
## 2              1             40              0                0               0
## 3              0              0              0                0               0
## 4              5             10              0                0               0
## 5              0              0              0                0               0
```

25

```
## 6                0               0               0               0               0
##    65-MDCOL003-Post-Chemo-Collection-Cycle-1-A 66-MDCOL004-1A
## 1                                            2              0
## 2                                            0              1
## 3                                            0              0
## 4                                            7              0
## 5                                            1              0
## 6                                            5              0
##    68-MDCOL003-3-1-Before-Chemo-2-3A 7-MDCOL058-Stool-1A-Baseline-BSB-2-4-1-3
## 1                                  1                                        4
## 2                                  0                                        1
## 3                                  0                                        0
## 4                                 16                                        0
## 5                                  1                                        0
## 6                                  0                                        0
##    70-MDCOL004-3-2-Post-Chemo-A 72-MDCOL003-4-1-4A-Before-Chemo-6
## 1                             0                                 0
## 2                             0                                 0
## 3                             0                                 0
## 4                             0                                19
## 5                             0                                 0
## 6                             0                                 0
##    75-MDCOL053-1-1-1-Unknown 9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                         24                                       18
## 2                          0                                        0
## 3                          0                                        0
## 4                         11                                        4
## 5                          0                                        0
## 6                          0                                        0
```

Creating Phyloseq Objects for the Kraken2-Bracken data

The following code packages the feature table, taxonomy file, and the phylogeentic tree (obtained from gracken, see Step 5) into a single phyloseq object for the datasets ("raw", "trimmed-filtered" etc).

```r
brackentables<-dir(path="BrackenOutputs",pattern="table-bracken*")    #Listing
↪   all the folders containing the bracken feature table
brackentaxonomy<-dir(path="BrackenOutputs",pattern="taxonomy-bracken*")
↪   #Listing all the folders containing the bracken taxonomy
brackentree<-dir(path="BrackenOutputs",pattern="bracken-reports*")   #Listing
↪   all the folders containing the bracken reports and tree

dir.create("PhyloseqOutputs/BrackenPhyloseq")
counter=0
for(each in brackentables){
  name<-substring(each,15)

↪   table<-as.matrix(biom_data(read_biom(paste("BrackenOutputs",each,"feature-table.biom",s
↪   #Reading biom file

↪   taxonomy<-read.table(paste("BrackenOutputs/taxonomy-bracken-",name,"/taxonomy.tsv",sep=
↪   = TRUE)%>%   #Reading taxonomy file
    mutate(Taxon=str_replace_all(Taxon,".__",""))%>%
```

26

```
  ⤷  separate(Taxon,c("Domain","Phylum","Class","Order","Family","Genus","Species"),sep=
    mutate(Phylum=if_else(is.na(Phylum),Domain,Phylum),
           Class=if_else(is.na(Class),Phylum,Class),
           Order=if_else(is.na(Order),Class,Order),
           Family=if_else(is.na(Family),Order,Family),
           Genus=if_else(is.na(Genus),Family,Genus),
           Species=if_else(is.na(Species),Genus,Species))%>%
    column_to_rownames("Feature.ID")%>%
    as.matrix()

 ⤷  tree<-read_tree(paste("BrackenOutputs/bracken-reports-",name,"/output.tree",sep=''))
 ⤷  #Reading the phylogenetic tree for that dataset
  counter=counter+1
  physeqBracken<-phyloseq(otu_table(table,taxa_are_rows =
 ⤷  TRUE),tax_table(taxonomy))  #creating the phyloseq object
  physeqBracken<-merge_phyloseq(physeqBracken,tree)  #Adding the phylogenetic
 ⤷  tree to the phyloseq object

  ⤷  saveRDS(physeqBracken,paste("PhyloseqOutputs/BrackenPhyloseq/phyloseq_bracken-",name,
  ⤷  #Saving it as an .RDS file

}
```

Summary of the phyloseq object generated from the Kraken2-Bracken data

```
##     Length    Class     Mode
##         1 phyloseq       S4


## phyloseq-class experiment-level object
## otu_table()   OTU Table:        [ 812 taxa and 37 samples ]
## tax_table()   Taxonomy Table:   [ 812 taxa by 7 taxonomic ranks ]
## phy_tree()    Phylogenetic Tree: [ 812 tips and 234 internal nodes ]
```

10. Transforming Metaphlan Data and creating phyloesq objects from it

Since, MetaPhlan data does not have the regular feature table and taxonomy format as Kraken-2 or Kaiju, we will have to tarnsform the data to make it fit that format. Also, we will use this tree generated from all the species level features in ChocoPhlan found here.

```
MetaphlanOutputs<-dir("MetaphlanOutput")    #Folder where the MetaPhlan
 ⤷  outputs are stored
dir.create("PhyloseqOutputs/MetaphlanPhyloseq")
MetaphlanTree<-read_tree("20201013_mpav30_speciesMod.nwk")    #The Tree made
 ⤷  from ChocoPhlan database
new_names <- MetaphlanTree$tip.label
new_names<-new_names%>%
  str_replace_all("s__","")
MetaphlanTree$tip.label<-new_names
for(folder in MetaphlanOutputs){    #Compiling all the outputs from different
 ⤷  sample into one feature abundance table and one taxonomy file
  Alltaxonomy<-NULL
```

```
  AllOTU<-NULL
  for(file in list.files(paste("MetaphlanOutput",folder,sep="/"))){
    content<-readLines(paste("MetaphlanOutput",folder,file,sep="/"))[-c(1:4)]
↪   #Parsing the MetaPhlan files
    id<-substring(basename(file),1,nchar(basename(file))-11)    #Stores the
↪   Sample ID
    totalreads<-readLines(paste("MetaphlanOutput",folder,file,sep="/"))[3]
↪   #Parsing the total reads for that sample
    totalreads<-as.numeric(gsub("^#(\\d+).*", "\\1", totalreads))
    table<-read.table(text=content,header=FALSE,sep="\t")%>%    #Transforming
↪   the Metaphlan file into feature table with taxonomy
      select(-4)%>%
      mutate(V1=str_replace_all(V1,".__",""),
             V1 = str_replace_all(V1, "\\|", ";"),
             reads=round(V3/100*totalreads)) %>%

      ↪  separate(V1,c("Kingdom","Phylum","Class","Order","Family","Genus","Species"),sep=
      unique()%>%
      filter(!(is.na(Species)))
    taxonomy<-table[,c(1:7)]    #Storing the taxonomy
    OTU<-table[,c(7,10)]%>%      #Storing the feature table
      unique()%>%
      mutate(Sample=id)
    Alltaxonomy<-rbind(Alltaxonomy,taxonomy)    #Adding taxonomy extracted
↪   from that sample with the rest in that dataset
    AllOTU<-rbind(AllOTU,OTU)    #Adding feature table extracted from that
↪   sample with the rest in that dataset
}

  taxa_table<-Alltaxonomy%>%    #Prepping taxatable for phyloseq
    arrange("Species")%>%
    unique()
  rownames(taxa_table)<-taxa_table$Species
  otu_table<-AllOTU%>%        #Prepping OTU/feature table for phyloseq
    group_by(Sample,Species)%>%
    summarize(reads=sum(reads))%>%
    pivot_wider(id_cols = "Species",values_from = "reads",names_from =
    ↪  "Sample",values_fill = 0)%>%
    column_to_rownames("Species")
  physeqMetaphlan<-phyloseq(otu_table(as.matrix(otu_table),taxa_are_rows =
↪   TRUE),tax_table(as.matrix(taxa_table)))    #Creating the phyloseq object
  physeqMetaphlan<-merge_phyloseq(physeqMetaphlan,phy_tree(MetaphlanTree))
↪   #Adding the tree to the phyloseq object

  ↪  saveRDS(physeqMetaphlan,paste("PhyloseqOutputs/MetaphlanPhyloseq/phyloseq_Metaphlan-
  ↪  #Saving the phyloseq object as an .RDS file
}
```

Sample Metaphlan output for a sample from our datasets

```
## [1]  #mpa_vOct22_CHOCOPhlAnSGB_202403
## [2]  #/data/choudhurya/miniconda/envs/mpa/bin/metaphlan demultiplexed/trimmed-filtered-
## [3]  #10709265 reads processed
## [4]  #SampleID    Metaphlan_Analysis
```

```
## [5]   #clade_name NCBI_tax_id relative_abundance  additional_species
## [6]   k__Bacteria 2    100.0
## [7]   k__Bacteria|p__Firmicutes   2|1239  51.68535
## [8]   k__Bacteria|p__Bacteroidetes   2|976   46.48729
## [9]   k__Bacteria|p__Actinobacteria   2|201174    1.22089
## [10]  k__Bacteria|p__Proteobacteria  2|1224  0.60648
## [11]  k__Bacteria|p__Bacteroidetes|c__Bacteroidia   2|976|200643    46.42399
## [12]  k__Bacteria|p__Firmicutes|c__Clostridia   2|1239|186801   29.73363
## [13]  k__Bacteria|p__Firmicutes|c__CFGB3002   2|1239| 15.59072
## [14]  k__Bacteria|p__Firmicutes|c__CFGB3012   2|1239| 2.14771
## [15]  k__Bacteria|p__Firmicutes|c__Negativicutes 2|1239|909932   1.73986
## [16]  k__Bacteria|p__Firmicutes|c__CFGB3005   2|1239| 1.38625
## [17]  k__Bacteria|p__Actinobacteria|c__Coriobacteriia   2|201174|84998  1.22089
## [18]  k__Bacteria|p__Proteobacteria|c__Betaproteobacteria   2|1224|28216   0.46794
## [19]  k__Bacteria|p__Firmicutes|c__CFGB3054   2|1239| 0.25919
## [20]  k__Bacteria|p__Firmicutes|c__Firmicutes_unclassified   2|1239| 0.24734
```

Sample taxonomy and feature table extracted from the MetaPhlan outputs

```
##                           Kingdom       Phylum        Class         Order
## Phocaeicola_vulgatus  Bacteria Bacteroidetes Bacteroidia Bacteroidales
## GGB9464_SGB14857        Bacteria     Firmicutes     CFGB3002       OFGB3002
## Bacteroides_uniformis Bacteria Bacteroidetes Bacteroidia Bacteroidales
## Blautia_wexlerae        Bacteria     Firmicutes   Clostridia Eubacteriales
## Bacteroides_ovatus    Bacteria Bacteroidetes Bacteroidia Bacteroidales
## Vescimonas_coprocola  Bacteria     Firmicutes   Clostridia Eubacteriales
##                           Family         Genus                 Species
## Phocaeicola_vulgatus    Bacteroidaceae Phocaeicola  Phocaeicola_vulgatus
## GGB9464_SGB14857              FGB3002     GGB9464       GGB9464_SGB14857
## Bacteroides_uniformis   Bacteroidaceae Bacteroides Bacteroides_uniformis
## Blautia_wexlerae        Lachnospiraceae     Blautia       Blautia_wexlerae
## Bacteroides_ovatus      Bacteroidaceae Bacteroides    Bacteroides_ovatus
## Vescimonas_coprocola  Oscillospiraceae  Vescimonas  Vescimonas_coprocola


##                                 1-MDcol002-5-2 11-MDCol002-label-12-on-bottle
## Adlercreutzia_equolifaciens              74804                         146604
## Agathobaculum_butyriciproducens            814                              0
## Alistipes_communis                       24059                              0
## Amedibacillus_dolichus                   11373                          12777
## Anaerobutyricum_soehngenii               68025                          85720
## Anaerostipes_hadrus                      31240                          14025
##                                 14-MDCOL003-Stool-1B-Baseline-BSB-2-4-1-4
## Adlercreutzia_equolifaciens                                             0
## Agathobaculum_butyriciproducens                                         0
## Alistipes_communis                                                      0
## Amedibacillus_dolichus                                                  0
## Anaerobutyricum_soehngenii                                              0
## Anaerostipes_hadrus                                                     0
##                                 15-MDcol004-3-1-pretreat 17-MDCOL006-3-1A
## Adlercreutzia_equolifaciens                        50268           288766
## Agathobaculum_butyriciproducens                     3402                0
## Alistipes_communis                                 35327                0
## Amedibacillus_dolichus                                 0                0
```

```
## Anaerobutyricum_soehngenii                                 36663                        0
## Anaerostipes_hadrus                                         64514                        0
##                                  19-MDCOL006-3-2A 21-MDCOL006-2-pre-chemo-A
## Adlercreutzia_equolifaciens                 668696                    492852
## Agathobaculum_butyriciproducens                  0                         0
## Alistipes_communis                               0                         0
## Amedibacillus_dolichus                           0                         0
## Anaerobutyricum_soehngenii                       0                         0
## Anaerostipes_hadrus                           3078                         0
##                                  23-MDcol033-1A 25-MDcol036-1A 27-MDcol025-1B
## Adlercreutzia_equolifaciens                 0              0          32991
## Agathobaculum_butyriciproducens             0              0          58289
## Alistipes_communis                          0              0              0
## Amedibacillus_dolichus                      0              0              0
## Anaerobutyricum_soehngenii                  0              0           2979
## Anaerostipes_hadrus                         0              0         323302
```

Summary of the phyloseq object generated from the MetaPhlan data

```
##    Length    Class      Mode
##       1 phyloseq        S4


## phyloseq-class experiment-level object
## otu_table()    OTU Table:        [ 207 taxa and 37 samples ]
## tax_table()    Taxonomy Table:   [ 207 taxa by 7 taxonomic ranks ]
## phy_tree()     Phylogenetic Tree: [ 207 tips and 206 internal nodes ]
```

11. Creating Phyloseq Objects with the HUMAnN data

The following code parses the genefemilies.tv, pathcoverage.tsv, and pathabundance.tv files and combines them from individual sample files into three phyloseq objects for each dataset.

```
HumannOutput <- dir("HumannOutput")  #Folder containing all the HUMAnN outputs
dir.create("PhyloseqOutputs/HumannPhyloseq")
for (folder in HumannOutput) {
    AllCoverage <- data.frame()
    AllAbundance <- data.frame()
    counter = 1
    for (file in list.files(paste("HumannOutput", folder,
        sep = "/"), pattern = "*_pathabundance.tsv")) {
        id <- substring(basename(file), 1, nchar(basename(file)) -
            18)  #Parsing the Sample ID
        abundance <- read.delim(paste("HumannOutput",
            folder, file, sep = "/"), sep = "\t", header = TRUE,
            comment.char = "")  #Reading the pathway abundance file for a
            ↪ sample
        df <- colnames(abundance) %>%
            str_replace_all("X\\.\\.", "") %>%
            str_replace_all("X", "") %>%
            str_replace_all("\\.", "_") %>%
            str_replace_all("_Abundance", "")
        colnames(abundance) <- df
```

```r
        coverage <- read.delim(paste("HumannOutput/",
            folder, "/", id, "_pathcoverage.tsv", sep = ""),
            sep = "\t", header = TRUE, comment.char = "")  #Reading the
            ↪  pathway coverage file for a sample
        df <- colnames(coverage) %>%
            str_replace_all("X\\.\\.", "") %>%
            str_replace_all("X", "") %>%
            str_replace_all("\\.", "-") %>%
            str_replace_all("_Coverage", "")
        colnames(coverage) <- df

        # Combining the pathway abundances and
        # pathway coverage data with those from
        # the previous samples
        if (counter == 1) {
            AllAbundance <- abundance
        } else {
            AllAbundance <- left_join(AllAbundance,
                abundance)
        }
        if (counter == 1) {
            AllCoverage <- coverage
        } else {
            AllCoverage <- left_join(AllCoverage, coverage)
        }
        counter = counter + 1
    }

    AllAbundance <- AllAbundance %>%
        column_to_rownames("Pathway") %>%
        as.matrix()
    AllCoverage <- AllCoverage %>%
        column_to_rownames("Pathway") %>%
        as.matrix()
    physeqcoverage <- phyloseq(otu_table(AllCoverage,
        taxa_are_rows = TRUE))  #Creating phyloseq object for the pathway
        ↪  abundance data
    physeqabundance <- phyloseq(otu_table(AllAbundance,
        taxa_are_rows = TRUE))  #Creating phyloseq object for the pathway
        ↪  coverage data
    saveRDS(physeqabundance,
    ↪  paste("PhyloseqOutputs/HumannPhyloseq/phyloseq_humann-",
        folder, "_pathwayabundance.rds", sep = ""))
    saveRDS(physeqcoverage,
    ↪  paste("PhyloseqOutputs/HumannPhyloseq/phyloseq_humann-",
        folder, "_pathwaycoverage.rds", sep = ""))
}

for (folder in HumannOutput) {
    AllRPK <- data.frame()
    counter = 1
    for (file in list.files(paste("HumannOutput", folder,
        sep = "/"), pattern = "*_genefamilies.tsv")) {
        id <- substring(basename(file), 1, nchar(basename(file)) -
```

```
            17)
        # Reading the gene family abundance for a
        # sample
        RPK <- read.delim(paste("HumannOutput", folder,
            file, sep = "/"), sep = "\t", header = TRUE,
            comment.char = "")
        df <- colnames(RPK) %>%
            str_replace_all("X\\.\\.", "") %>%
            str_replace_all("X", "") %>%
            str_replace_all("\\.", "-") %>%
            str_replace_all("_Abundance-RPKs", "")
        colnames(RPK) <- df
        # Joining the gene family abundance data
        # of the current sample to the rest of
        # the samples
        if (counter == 1) {
            AllRPK <- RPK
        } else {
            AllRPK <- left_join(AllRPK, RPK)
        }
        counter = counter + 1
    }

    AllRPK <- AllRPK %>%
        column_to_rownames("Gene-Family") %>%
        as.matrix()
    physeqRPK <- phyloseq(otu_table(AllRPK, taxa_are_rows = TRUE))  #Creating
↪  phyloseq object for the gene family abundance data
    saveRDS(physeqRPK,
     ↪  paste("PhyloseqOutputs/HumannPhyloseq/phyloseq_humann-",
        folder, "_genefamilies.rds", sep = ""))
}
```

Sample pathway abundance file

```
##                                                   Pathway
## 1                                                 UNMAPPED
## 2                                             UNINTEGRATED
## 3    UNINTEGRATED|g__Prevotella.s__Prevotella_copri
## 4                       UNINTEGRATED|unclassified
## 5 UNINTEGRATED|g__Alistipes.s__Alistipes_finegoldii
## 6          UNINTEGRATED|g__Blautia.s__Blautia_obeum
##   9_MDCOL062_Stool_1A_Baseline_BSB_2_4_1_3
## 1                               706374.82
## 2                              1990573.07
## 3                               280605.01
## 4                               211052.96
## 5                               161814.73
## 6                                77940.21
```

Combined pathway abundances of all samples in a dataset

```
##                                                          1_MDcol002_5_2
```

```
## UNMAPPED                                                        778748.2
## UNINTEGRATED                                                   2481983.1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus            917463.5
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis           464620.0
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              354114.6
## UNINTEGRATED|unclassified                                      147854.1
##                                              11_MDCol002_label_12_on_bottle
## UNMAPPED                                                          812515.6
## UNINTEGRATED                                                     2338562.7
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus              740989.9
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis            205282.2
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus               248232.9
## UNINTEGRATED|unclassified                                       229298.6
##                                          14_MDCOL003_Stool_1B_Baseline_BSB_2
## UNMAPPED                                                              72
## UNINTEGRATED                                                          317
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
## UNINTEGRATED|unclassified                                            30
##                                              15_MDcol004_3_1_pretreat
## UNMAPPED                                                         853278.4
## UNINTEGRATED                                                    2607555.3
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus             187785.6
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis           193568.8
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus               27733.6
## UNINTEGRATED|unclassified                                      406532.7
##                                              17_MDCOL006_3_1A
## UNMAPPED                                                       640092.1
## UNINTEGRATED                                                  2365706.0
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis         255956.6
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus            486901.1
## UNINTEGRATED|unclassified                                    176732.7
##                                              19_MDCOL006_3_2A
## UNMAPPED                                                      728047.75
## UNINTEGRATED                                                 2412350.76
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          84115.85
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             78634.97
## UNINTEGRATED|unclassified                                    178803.49
##                                              21_MDCOL006_2_pre_chemo_A
## UNMAPPED                                                       1170228.6
## UNINTEGRATED                                                   4808488.1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                  NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis           240857.1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              828564.4
## UNINTEGRATED|unclassified                                      532576.9
##                                              23_MDcol033_1A
## UNMAPPED                                                       691496.1
## UNINTEGRATED                                                  2425535.6
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          159044.4
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             178257.6
```

```
## UNINTEGRATED|unclassified                                      141054.7
##                                                         25_MDcol036_1A
## UNMAPPED                                                    1068527.81
## UNINTEGRATED                                                4715990.30
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus          665352.47
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus            85896.49
## UNINTEGRATED|unclassified                                    245014.75
##                                                 27_MDcol025_1B 28_7A_1_1
## UNMAPPED                                            945989.87   902644.2
## UNINTEGRATED                                       3843549.24 2664917.9
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 313605.33   665346.6
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis 271299.80   716802.5
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus    59771.61   354245.5
## UNINTEGRATED|unclassified                           191873.16   216273.1
##                                                   29_7A 3_MDcol002_5_1
## UNMAPPED                                         630687.3    1054449.8
## UNINTEGRATED                                    2027313.2    2771711.1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 617776.9    502390.7
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis 642820.4    658787.2
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus   274152.5    399946.3
## UNINTEGRATED|unclassified                          150555.6    308745.2
##                                                      30_8B       31_8B
## UNMAPPED                                         364662.41   860531.46
## UNINTEGRATED                                   1176143.90 1898317.92
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 265896.80    62848.75
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis 357642.20   259350.24
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus   154126.96   113944.27
## UNINTEGRATED|unclassified                           65569.98   186109.70
##                                                      33_6_2   36_40_1B
## UNMAPPED                                         683345.04 1168856.3
## UNINTEGRATED                                   1265336.05 3293921.9
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus  50452.27   896384.5
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis 136902.02   369106.3
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus    64743.90   644273.4
## UNINTEGRATED|unclassified                          150448.35   296277.0
##                                                 38_S_Swanson
## UNMAPPED                                          1317047.4
## UNINTEGRATED                                      2714109.3
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 125167.7
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis        NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus          NA
## UNINTEGRATED|unclassified                          524893.7
##                                             40_MDCol002_7_43   42_51_1B
## UNMAPPED                                         1234938.85   795759.2
## UNINTEGRATED                                     3188195.55 2269103.2
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 1000793.98   961751.5
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis         NA   331637.7
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus    77151.05   257447.8
## UNINTEGRATED|unclassified                          336153.66   229043.3
##                                                 45_MDCOL005_1B
## UNMAPPED                                          1246331.44
## UNINTEGRATED                                      6680423.14
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus 597264.54
```

```
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          68584.03
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                    NA
## UNINTEGRATED|unclassified                                     243317.20
##                                                      46_MDCOL015_1a
## UNMAPPED                                                       911520.5
## UNINTEGRATED                                                  2033420.0
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                   NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                  NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     NA
## UNINTEGRATED|unclassified                                      300806.1
##                                                      48_MDCOL039_1a
## UNMAPPED                                                     994944.84
## UNINTEGRATED                                                 3228306.39
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus           305187.68
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          405561.98
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              56772.45
## UNINTEGRATED|unclassified                                     243334.71
##                                         5_MDCOL055_Stool_1A_Baseline_BSB_2_
## UNMAPPED                                                           4334
## UNINTEGRATED                                                       14781
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                 3570
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                2953
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                   4328
## UNINTEGRATED|unclassified                                            810
##                                                      50_MDCOL027_1a
## UNMAPPED                                                     904651.65
## UNINTEGRATED                                                 2385396.13
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus            40831.38
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis           26875.49
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              12807.89
## UNINTEGRATED|unclassified                                     324132.36
##                                                      52_MDCOL021_1b
## UNMAPPED                                                    608032.696
## UNINTEGRATED                                                2086633.441
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus            6378.159
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis           6021.116
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             14722.860
## UNINTEGRATED|unclassified                                    173575.249
##                                                      54_MDCOL028_1b
## UNMAPPED                                                     940612.29
## UNINTEGRATED                                                 5641354.19
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                   NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                  NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     NA
## UNINTEGRATED|unclassified                                      96868.25
##                                                      60_MDcol006_4_1A
## UNMAPPED                                                       709619.7
## UNINTEGRATED                                                  2663982.1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                   NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          1151220.5
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              371673.6
## UNINTEGRATED|unclassified                                      208778.9
##                                                      63_MDcol003_3_2
## UNMAPPED                                                       552487.6
```

```
## UNINTEGRATED                                                                3076573.7
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                              NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                             NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                                NA
## UNINTEGRATED|unclassified                                                   199483.9
##                                                          65_MDCOL003_Post_Chemo_Collection_C
## UNMAPPED
## UNINTEGRATED                                                                              1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
## UNINTEGRATED|unclassified
##                                                          66_MDCOL004_1A
## UNMAPPED                                                      102994.388
## UNINTEGRATED                                                 211972.628
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus           24242.876
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          18036.559
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              6011.369
## UNINTEGRATED|unclassified                                     31441.128
##                                                          68_MDCOL003_3_1_Before_Chemo_2_3A
## UNMAPPED                                                               575372.3
## UNINTEGRATED                                                          2622809.6
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                          NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                         NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                            NA
## UNINTEGRATED|unclassified                                               221928.5
##                                                          7_MDCOL058_Stool_1A_Baseline_BSB_2_
## UNMAPPED                                                                    3288
## UNINTEGRATED                                                                6894
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                         1468
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                         479
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                           1102
## UNINTEGRATED|unclassified                                                    912
##                                                          70_MDCOL004_3_2_Post_Chemo_A
## UNMAPPED                                                       34068.790
## UNINTEGRATED                                                   39832.621
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus            6643.549
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             4221.045
## UNINTEGRATED|unclassified                                      8543.140
##                                                          72_MDCOL003_4_1_4A_Before_Chemo_6
## UNMAPPED                                                      461003.24
## UNINTEGRATED                                                 1808251.90
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                  NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          186720.43
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             26477.84
## UNINTEGRATED|unclassified                                    104787.82
##                                                          75_MDCOL053_1_1_1_Unknown
## UNMAPPED                                                      595072.92
## UNINTEGRATED                                                 1816052.92
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus           49070.33
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis          11643.02
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             26855.97
## UNINTEGRATED|unclassified                                    223668.29
```

```
##                                                              9_MDCOL062_Stool_1A_Baseline_BSB_2_
## UNMAPPED                                                                           7063
## UNINTEGRATED                                                                       19905
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_vulgatus                                 519
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_uniformis                                281
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                                   220
## UNINTEGRATED|unclassified                                                          2110
```

Sample pathway coverage file

```
##                                                              Pathway
## 1                                                           UNMAPPED
## 2                                                       UNINTEGRATED
## 3 UNINTEGRATED|g__Agathobaculum.s__Agathobaculum_butyriciproducens
## 4          UNINTEGRATED|g__Akkermansia.s__Akkermansia_muciniphila
## 5              UNINTEGRATED|g__Alistipes.s__Alistipes_finegoldii
## 6              UNINTEGRATED|g__Alistipes.s__Alistipes_onderdonkii
##   9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                                         1
## 2                                         1
## 3                                         1
## 4                                         1
## 5                                         1
## 6                                         1
```

Combined pathway coverage of all samples in a dataset

```
##                                                              1-MDcol002-5-2
## UNMAPPED                                                                  1
## UNINTEGRATED                                                              1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens              1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                       1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus               1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                         1
##                                                              11-MDCol002-label-12-on-bot
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                              14-MDCOL003-Stool-1B-Baseli
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                              15-MDcol004-3-1-pretreat
## UNMAPPED                                                                  1
## UNINTEGRATED                                                              1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens              1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                       1
```

```
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                 1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                          1
##                                                        17-MDCOL006-3-1A
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens          1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                  NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus            1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1
##                                                        19-MDCOL006-3-2A
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens          1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                  NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus            1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1
##                                                        21-MDCOL006-2-pre-chemo-A
## UNMAPPED                                                                   1
## UNINTEGRATED                                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens                   1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                          NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                    1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                             1
##                                                        23-MDcol033-1A
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens         NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                  NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus           NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1
##                                                        25-MDcol036-1A
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens         NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                  NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus           NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1
##                                                        27-MDcol025-1B
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens          1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                   1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus            1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1
##                                                        28-7A-1-1 29-7A
## UNMAPPED                                                          1     1
## UNINTEGRATED                                                      1     1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens          1     1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                   1     1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus            1     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                     1     1
##                                                        3-MDcol002-5-1
## UNMAPPED                                                          1
## UNINTEGRATED                                                      1
```

```
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens           1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                    1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus             1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                      1
##                                                         30-8B 31-8B 33-6-2
## UNMAPPED                                                   1     1     1
## UNINTEGRATED                                               1     1     1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens   1     1     1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus            1     1     1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus     1     1     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              1     1     1
##                                                         36-40-1B
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens   1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus            1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              1
##                                                         38-S-Swanson
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens   1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus            1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             NA
##                                                         40-MDCol002-7-43
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens   1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus            1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              1
##                                                         42-51-1B
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens  NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus           NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus    NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus              1
##                                                         45-MDCOL005-1B
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens   1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus           NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus     1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             NA
##                                                         46-MDCOL015-1a
## UNMAPPED                                                   1
## UNINTEGRATED                                               1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens  NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus            1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus    NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus             NA
##                                                         48-MDCOL039-1a
```

```
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens                1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                         1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                  1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                           1
##                                               5-MDCOL055-Stool-1A-Baselir
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                               50-MDCOL027-1a
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens                1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                         1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                  1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                           1
##                                               52-MDCOL021-1b
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens               NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                        NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                           1
##                                               54-MDCOL028-1b
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens               NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                        NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                          NA
##                                               60-MDcol006-4-1A
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens                1
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                         1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                  1
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                           1
##                                               63-MDcol003-3-2
## UNMAPPED                                                                    1
## UNINTEGRATED                                                                1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens               NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                        NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                 NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                          NA
##                                               65-MDCOL003-Post-Chemo-Col:
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
```

```
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                               66-MDCOL004-1A
## UNMAPPED                                                                  1
## UNINTEGRATED                                                              1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens            NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                     NA
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus              NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                        1
##                                                               68-MDCOL003-3-1-Before-Cher
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                               7-MDCOL058-Stool-1A-Baselin
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                               70-MDCOL004-3-2-Post-Chemo-
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens                  N
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                           N
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus                    N
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                               72-MDCOL003-4-1-4A-Before-C
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
##                                                               75-MDCOL053-1-1-1-Unknown
## UNMAPPED                                                                  1
## UNINTEGRATED                                                              1
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens            NA
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus                      1
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus              NA
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus                        1
##                                                               9-MDCOL062-Stool-1A-Baselir
## UNMAPPED
## UNINTEGRATED
## UNINTEGRATED|g__Adlercreutzia.s__Adlercreutzia_equolifaciens
## UNINTEGRATED|g__Anaerostipes.s__Anaerostipes_hadrus
## UNINTEGRATED|g__Asaccharobacter.s__Asaccharobacter_celatus
## UNINTEGRATED|g__Bacteroides.s__Bacteroides_ovatus
```

Sample gene family abundance file

```
##                                                    Gene-Family
```

```
## 1                                               UNMAPPED
## 2                                       UniRef90_A5ZYV3
## 3          UniRef90_A5ZYV3|g__Blautia.s__Blautia_obeum
## 4                                       UniRef90_U3JBQ2
## 5 UniRef90_U3JBQ2|g__Eggerthella.s__Eggerthella_lenta
## 6                                    UniRef90_A0A174PA28
##   9-MDCOL062-Stool-1A-Baseline-BSB-2-4-1-3
## 1                              1987609.00
## 2                                19139.59
## 3                                19139.59
## 4                                15087.70
## 5                                15087.70
## 6                                13552.05
```

Combined gene family abundances of all samples in a dataset

```
##                                                              1-MDcol002-5-2
## UNMAPPED                                                       2364890.000
## UniRef90_A0A174NA57                                             55247.258
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis     55247.258
## UniRef90_A0A395VRV3                                             49394.382
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus        25032.358
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus       8720.539
##                                                              11-MDCol002-label-12-on-bott
## UNMAPPED                                                               2374486.0
## UniRef90_A0A174NA57                                                      43582.0
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis             43582.0
## UniRef90_A0A395VRV3                                                      31559.4
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                  7825.1
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus               5414.0
##                                                              14-MDCOL003-Stool-1B-Baselin
## UNMAPPED
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                              15-MDcol004-3-1-pretreat
## UNMAPPED                                                            2462301.0000
## UniRef90_A0A174NA57                                                   12391.8543
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis           12391.8543
## UniRef90_A0A395VRV3                                                    6052.5065
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus              2456.2799
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus             823.7582
##                                                              17-MDCOL006-3-1A
## UNMAPPED                                                       1570111.000
## UniRef90_A0A174NA57                                                     NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis             NA
## UniRef90_A0A395VRV3                                             5697.658
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus              NA
##                                                              19-MDCOL006-3-2A
## UNMAPPED                                                       1791726.000
## UniRef90_A0A174NA57                                                     NA
```

```
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis               NA
## UniRef90_A0A395VRV3                                             629.729
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                   NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus                 NA
##                                                           21-MDCOL006-2-pre-chemo-A
## UNMAPPED                                                    2.798869e+06
## UniRef90_A0A174NA57                                                      NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis               NA
## UniRef90_A0A395VRV3                                           1.730327e+04
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus       7.697973e+00
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus               NA
##                                                           23-MDcol033-1A
## UNMAPPED                                                    1.816065e+06
## UniRef90_A0A174NA57                                           1.139728e+01
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis               NA
## UniRef90_A0A395VRV3                                           1.754432e+04
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus       8.162051e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus               NA
##                                                           25-MDcol036-1A
## UNMAPPED                                                    2.876898e+06
## UniRef90_A0A174NA57                                           3.411955e+02
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis               NA
## UniRef90_A0A395VRV3                                           1.683431e+04
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus       1.683046e+04
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus     4.744231e-01
##                                                           27-MDcol025-1B
## UNMAPPED                                                    2.535351e+06
## UniRef90_A0A174NA57                                           9.803922e+00
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis               NA
## UniRef90_A0A395VRV3                                           6.165756e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus       6.164478e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus     6.428122e-01
##                                                           28-7A-1-1
## UNMAPPED                                                   2736361.000
## UniRef90_A0A174NA57                                          74789.702
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis  74789.702
## UniRef90_A0A395VRV3                                          34024.889
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus     17735.630
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus    5423.871
##                                                           29-7A
## UNMAPPED                                                   1980954.000
## UniRef90_A0A174NA57                                          46810.359
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis  46810.359
## UniRef90_A0A395VRV3                                          29657.473
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus     15926.570
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus    4748.206
##                                                           3-MDcol002-5-1
## UNMAPPED                                                   3126697.000
## UniRef90_A0A174NA57                                          60087.977
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis  60087.977
## UniRef90_A0A395VRV3                                          38175.660
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus      9597.700
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus    6431.994
##                                                           30-8B
```

```
## UNMAPPED                                                         1105105.000
## UniRef90_A0A174NA57                                                 25387.917
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis         25387.917
## UniRef90_A0A395VRV3                                                 21139.826
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus            10124.156
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           2733.621
##                                                                         31-8B
## UNMAPPED                                                         2514157.000
## UniRef90_A0A174NA57                                                 21054.698
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis         21054.698
## UniRef90_A0A395VRV3                                                 14503.126
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             6558.171
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           2147.991
##                                                                        33-6-2
## UNMAPPED                                                         1998210.000
## UniRef90_A0A174NA57                                                 14323.268
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis         14323.268
## UniRef90_A0A395VRV3                                                  6942.182
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             3888.373
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           1008.496
##                                                                       36-40-1B
## UNMAPPED                                                          3434720.00
## UniRef90_A0A174NA57                                                  78682.87
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          78682.87
## UniRef90_A0A395VRV3                                                  56618.00
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             13359.20
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           10429.88
##                                                                   38-S-Swanson
## UNMAPPED                                                          3.951244e+06
## UniRef90_A0A174NA57                                                4.437334e+01
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis                 NA
## UniRef90_A0A395VRV3                                                3.114029e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                    NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus         1.426303e+03
##                                                              40-MDCol002-7-43
## UNMAPPED                                                         3609851.0000
## UniRef90_A0A174NA57                                                  478.7466
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis                 NA
## UniRef90_A0A395VRV3                                                41862.4500
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus           22803.9210
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus          9008.1536
##                                                                       42-51-1B
## UNMAPPED                                                         2458377.000
## UniRef90_A0A174NA57                                                  39347.805
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          39347.805
## UniRef90_A0A395VRV3                                                  45337.774
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             24467.937
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           6390.123
##                                                                  45-MDCOL005-1B
## UNMAPPED                                                          2.619226e+06
## UniRef90_A0A174NA57                                                         NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis                 NA
## UniRef90_A0A395VRV3                                                1.742780e+04
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                    NA
```

```
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus   1.037072e+01
##                                                            46-MDCOL015-1a
## UNMAPPED                                                         2508626
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                                    NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           NA
##                                                            48-MDCOL039-1a
## UNMAPPED                                                      2.862189e+06
## UniRef90_A0A174NA57                                          1.139754e+00
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                          2.077782e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus    1.541726e+03
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus  1.107623e+02
##                                                            5-MDCOL055-Stool-1A-Baseline
## UNMAPPED
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                            50-MDCOL027-1a
## UNMAPPED                                                      2.569471e+06
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                          8.392341e+02
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus    8.490451e+01
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus  1.747656e+02
##                                                            52-MDCOL021-1b
## UNMAPPED                                                      1332322.0000
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                              811.9280
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus      259.6574
##                                                            54-MDCOL028-1b
## UNMAPPED                                                      1.875517e+06
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                          1.006431e+01
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           NA
##                                                            60-MDcol006-4-1A
## UNMAPPED                                                      2136531.000
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
## UniRef90_A0A395VRV3                                              7466.442
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus             NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           NA
##                                                            63-MDcol003-3-2
## UNMAPPED                                                      1172962.000
## UniRef90_A0A174NA57                                                    NA
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis          NA
```

```
## UniRef90_A0A395VRV3                                                             1801.322
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus                              NA
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus                            NA
##                                                            65-MDCOL003-Post-Chemo-Colle
## UNMAPPED
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                            66-MDCOL004-1A
## UNMAPPED                                                      274343.0000
## UniRef90_A0A174NA57                                             2498.5665
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis     2498.5665
## UniRef90_A0A395VRV3                                             1859.1191
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus         909.3233
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus       236.6618
##                                                            68-MDCOL003-3-1-Before-Chemo
## UNMAPPED                                                                       12
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                            7-MDCOL058-Stool-1A-Baseline
## UNMAPPED
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                            70-MDCOL004-3-2-Post-Chemo-A
## UNMAPPED                                                          97034.00000
## UniRef90_A0A174NA57                                                1436.54015
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis        1436.54015
## UniRef90_A0A395VRV3                                                 225.24666
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus            102.72230
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus           49.84632
##                                                            72-MDCOL003-4-1-4A-Before-Ch
## UNMAPPED                                                             117068
## UniRef90_A0A174NA57                                                     188
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis             188
## UniRef90_A0A395VRV3                                                     409
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
##                                                            75-MDCOL053-1-1-1-Unknown
## UNMAPPED                                                        1578382.0000
## UniRef90_A0A174NA57                                                1906.2450
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis        1906.2450
## UniRef90_A0A395VRV3                                                1064.6096
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus            171.7574
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus          255.1828
##                                                            9-MDCOL062-Stool-1A-Baseline
## UNMAPPED
```

```
## UniRef90_A0A174NA57
## UniRef90_A0A174NA57|g__Bacteroides.s__Bacteroides_uniformis
## UniRef90_A0A395VRV3
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_ovatus
## UniRef90_A0A395VRV3|g__Bacteroides.s__Bacteroides_vulgatus
```