

Assignment 005

Summarise Alpha Diversity in R

- Clone the fift class repository:
https://github.com/bjklab/EPID674_005_alpha-diversity.git
- Read long-form Human Microbiome Data (included in repo)
- Examine the dataset
- Install & load **tidyverse** & **vegan** packages
- Assignment:
 - Summarize by specimen: unique OTUs, sequencing depth
 - Summarize by specimen: alpha diversity

RStudio Cloud


NIH Human Microbiome Project

EPID 674-910 2020B Measur

bjklab/EPID674_002_sequenc

R for Data Science

← → ↺ 🔒 https://www.hmpdacc.org/HMQCP/ ☆ P 🔍 ⚙️ 👤 ⋮



HMP1 ▾

NIH Human Microbiome Project

[Home](#) [Overview](#) [Membership](#) [Publications](#) [Resources](#) [Data](#) [Outreach](#) [Login](#)

[home](#) > [data browser](#) > [hmqcp](#) > [hmqcp: healthy](#) > [hmqcp: published](#)

HMQCP

Following a July 2010 16S data freeze, data was downloaded from NCBI SRA projects [SRP002395](#): Human Microbiome Project 16S rRNA Clinical Production Phase I, and [SRP002012](#): Human Microbiome Project 454 Clinical Production Pilot. This dataset corresponds to over 5,700 samples and over 10,000 sequence preps. 16S variable region 3-5 (V35) was sequenced for the entire set of samples, and variable region 1-3 (V13) for a subset of samples.

The [QIIME](#) (Quantitative Insights Into Microbial Ecology) software package was used to process HMP 16S data using an OTU-binning strategy to which taxonomic classification is added.

Raw 16S sequence and metadata, available at [HMR16S](#), were demultiplexed using QIIME. OTU picking was performed for the V1-3 and V3-5 region sequences using [OTUPipe](#), which includes error correction, chimera checking through [UCHIME](#), and clustering via UCLUST, and postprocessing by picking the optimal representative sequence centroid. Taxonomy was assigned using the RDP classifier version 2.2.

The resulting OTU tables were checked for mislabeling and contamination, as described in the SOP available below. Alpha and beta diversity for each sample and Procrustes analysis were established using QIIME with default parameters.

All QIIME output files are available here, for both the V1-3 and V3-5 variable regions, as well as Procrustes summary data. SOPs and custom scripts can be found below.

If you're interested in joint analysis of 16S and shotgun metagenomic datasets from the HMP, pairing up data from the same microbiome samples can initially seem tricky. The [HMP Sample Flow Schematic](#) indicates how these sample IDs are related experimentally, and provides tables joining 16S dataset "SN" and "PSN" identifiers with metagenomic dataset "SRS" identifiers.

- [Data Table](#)
- [Protocols and Tools](#)
- [Related Pages](#)

3 / 15

RStudio Cloud
NIH Human Microbiome Proj
EPID 674-910 2020B Measu
bjklab/EPID674_002_sequeni
R for Data Science

https://www.hmpdacc.org/HMQCP/

The [QIIME](#) (Quantitative Insights Into Microbial Ecology) software package was used to process HMP 16S data using an OTU-binning strategy to which taxonomic classification is added.

Raw 16S sequence and metadata, available at [HMR16S](#), were demultiplexed using QIIME. OTU picking was performed for the V1-3 and V3-5 region sequences using [OTUPipe](#), which includes error correction, chimera checking through [UCHIME](#), and clustering via UCLUST, and postprocessing by picking the optimal representative sequence centroid. Taxonomy was assigned using the RDP classifier version 2.2.

The resulting OTU tables were checked for mislabeling and contamination, as described in the SOP available below. Alpha and beta diversity for each sample and Procrustes analysis were established using QIIME with default parameters.

All QIIME output files are available here, for both the V1-3 and V3-5 variable regions, as well as Procrustes summary data. SOPs and custom scripts can be found below.

If you're interested in joint analysis of 16S and shotgun metagenomic datasets from the HMP, pairing up data from the same microbiome samples can initially seem tricky. The [HMP Sample Flow Schematic](#) indicates how these sample IDs are related experimentally, and provides tables joining 16S dataset "SN" and "PSN" identifiers with metagenomic dataset "SRS" identifiers.

- [Data Table](#)
- [Protocols and Tools](#)
- [Related Pages](#)

QIIME Data						
File Description	v13 Downl...	v13 Size	v13 MD5	v35 Downl...	v35 Size	v35 MD5
1. Sequences		948.2 MB	4118265f600f581d38823f4224f54468		1.4 GB	fa81bf68291fb0dfb7727ce246df35dd
2. Chimeric sequences		26.8 MB	3f217b67a907d0cab8ecff1f26ff272e		24.9 MB	3708457cb1c667c201805751993452ef
3. Nonchimeric sequences		71.8 MB	0cc3b82cf16d46c303f902bfb7b88e9		89.8 MB	5c5a99926fd3836e8fbeaa929a490adc
4. Map file between OTU clusters and sequences		186.3 MB	59510c3f23de11764e9c0ad31a422337		308.6 MB	ed03159d1d072b8beee7db7dcaabb76e
5. Representative sequence sets		3.8 MB	a46fa1ac8c53f3d3fb2b51db6ee3cabd		4.2 MB	87057f398a07327f80d83205aa023855
6. Representative sequence phylogenetic trees		289.8 KB	b7dc403d0f823e5f3f73b292ff9f2b62		468.4 KB	ad59e8a2941f7756bbb057a9900b4c54
7. OTU table per sample		6.3 MB	017bd5801e9ae99dd00d77a78c82f301		10.7 MB	cde636b96baaec1c4bfaedc7c73cdf71
8. Mislabelled or contaminated samples		522.0 bytes	52792d57c4ba5a4bda853bbc7cb0c102		729.0 bytes	96b6461a7250e699adab53e99d674c5a
9. Final OTU table		6.3 MB	87d2fe84196464d8cf4fcb5b05ddc581		10.7 MB	642570d995244dd1be804f3edea60a7e
10. Mapping Files		22.8 KB	a3741414d982fe1b4739166598fcb026		37.3 KB	23c225ada6e7f6fdd5f7f72063fc3a62
11. Beta diversity analysis		234.2 MB	16829e78cca6341b184596ec043a98c5		538.3 MB	68831f6ad6e7f30e21ce88a79bd9a4ff

Save as CSV

RStudio Cloud

https://rstudio.cloud/projects

Your Workspace Projects Info

Brendan Kelly

Spaces

Your Workspace

New Space

Learn

- Guide
- What's New
- Primers
- Cheat Sheets
- Feedback and Questions

Info

- Terms and Conditions
- System Status

Your Projects

Stan

Fork this if you want to...

Created Oct 12, 2018 1:...

New Project from Git Repo

URL of your Git repository

https://github.com/bjklab/EPID674_002_sequences-to-counts.js

OK

Options

Search Projects

Sort Projects

- By name
- By date created

Capacity

This your personal workspace, where you can create a virtually unlimited number of projects.

Learn more about [Your Workspace](#) in the [Guide](#).

RStudio Cloud ^{beta}

Terms and Conditions System Status

© 2020 RStudio, PBC

Using the **rstudio.cloud** console

```
# make sure tidyverse loaded
library(tidyverse)

# load (trimmed) HMP V1-V3 OTU table
otu <- read_csv(
  file =
    "./data/HMP_OTU_table_longformat.csv.gz",
)

otu # show what you've read
```

```
## # A tibble: 1,380,480 x 4
##   HMPbodysubsite specimen_id otu_id      read_count
##   <chr>           <dbl> <chr>          <dbl>
## 1 Anterior_nares  700014445 OTU_97.1         0
## 2 Anterior_nares  700014445 OTU_97.10        0
## 3 Anterior_nares  700014445 OTU_97.100       0
## 4 Anterior_nares  700014445 OTU_97.1000      0
## 5 Anterior_nares  700014445 OTU_97.10000     0
## 6 Anterior_nares  700014445 OTU_97.10001     0
## 7 Anterior_nares  700014445 OTU_97.10002     0
## 8 Anterior_nares  700014445 OTU_97.10003     0
## 9 Anterior_nares  700014445 OTU_97.10004     0
## 10 Anterior_nares 700014445 OTU_97.10005     0
## # ... with 1,380,470 more rows
```

Try mutate()

```
otu %>%  
  mutate(  
    HMPbodysubsite =  
      gsub("_", " ", HMPbodysubsite),  
    log_reads = log10(read_count)  
  )
```

```
## # A tibble: 1,380,480 x 5  
##   HMPbodysubsite specimen_id otu_id      read_count log_reads  
##   <chr>          <dbl> <chr>          <dbl>      <dbl>  
## 1 Anterior nares 700014445 OTU_97.1          0      -Inf  
## 2 Anterior nares 700014445 OTU_97.10         0      -Inf  
## 3 Anterior nares 700014445 OTU_97.100        0      -Inf  
## 4 Anterior nares 700014445 OTU_97.1000       0      -Inf  
## 5 Anterior nares 700014445 OTU_97.10000      0      -Inf  
## 6 Anterior nares 700014445 OTU_97.10001      0      -Inf  
## 7 Anterior nares 700014445 OTU_97.10002      0      -Inf  
## 8 Anterior nares 700014445 OTU_97.10003      0      -Inf  
## 9 Anterior nares 700014445 OTU_97.10004      0      -Inf  
## 10 Anterior nares 700014445 OTU_97.10005     0      -Inf  
## # ... with 1,380,470 more rows
```

Try summarise()

```
otu %>%  
  group_by(HMPbodysubsite) %>%  
  summarise(  
    mean_reads = mean(read_count, na.rm = TRUE),  
    sum_reads = sum(read_count, na.rm = TRUE)  
  ) %>%  
  ungroup()
```

```
## # A tibble: 16 x 3  
##   HMPbodysubsite      mean_reads sum_reads  
##   <chr>          <dbl>     <dbl>  
## 1 Anterior_nares      0.0958      8262  
## 2 Attached_Keratinized_gingiva 0.121     10458  
## 3 Buccal_mucosa       0.172     14831  
## 4 Hard_palate         0.116      9977  
## 5 Left_Retroauricular_crease 0.0403      3481  
## 6 Mid_vagina          0.267     23043  
## 7 Palatine_Tonsils    0.133     11445  
## 8 Posterior_fornix    0.257     22189  
## 9 Right_Retroauricular_crease 0.0835      7207  
## 10 Saliva             0.213     18336  
## 11 Stool              0.221     19072  
## 12 Subgingival_plaque 0.262     22573  
## 13 Supragingival_plaque 0.260     22404  
## 14 Throat            0.117     10061  
## 15 Tongue_dorsum      0.202     17431  
## 16 Vaginal_introitus  0.247     21354
```


summarise() + vegan

```
install.packages("vegan")
library(vegan)

otu %>%
  group_by(specimen_id) %>%
  summarise(
    specimen_shannon = diversity(x = read_count,
                                index = "shannon",
                                base = exp(1))

  ) %>%
  ungroup()
```

```
## # A tibble: 32 x 2
##   specimen_id specimen_shannon
##   <dbl>         <dbl>
## 1 700013549         4.80
## 2 700014386         5.41
## 3 700014403         5.63
## 4 700014409         5.70
## 5 700014412         5.02
## 6 700014415         5.32
## 7 700014418         4.83
## 8 700014421         5.52
## 9 700014424         5.02
## 10 700014427         5.33
## # ... with 22 more rows
```

summarise() + vegan

```
otu %>%
  group_by(specimen_id, HMPbodysubsite) %>%
  summarise(
    specimen_shannon = diversity(x = read_count,
                                index = "shannon",
                                base = exp(1))
  ) %>%
  ungroup() %>%
  qplot(data = ., x = HMPbodysubsite,
        y = specimen_shannon,
        fill = HMPbodysubsite,
        geom = "boxplot") +
  theme_bw() +
  theme(axis.text.x = element_text(angle = 270,
                                    hjust = 0,
                                    vjust = 0.5),
        legend.position = "none")
```

summarise() + vegan

```
otu %>%
  group_by(specimen_id, HMPbodysubsite) %>%
  summarise(
    read_count = read_count,
    rare_count = as.vector(
      rrarefy(read_count, sample = 1000)
    ) %>%
  summarise(
    #shannon = ...,
    rare_shannon = diversity(x = rare_count,
                           index = "shannon",
                           base = exp(1))

  ) %>%
  ungroup()
```

```
## # A tibble: 32 x 3
##   specimen_id HMPbodysubsite rare_shannon
##   <dbl> <chr> <dbl>
## 1 700013549 Stool 4.58
## 2 700014386 Stool 5.05
## 3 700014403 Saliva 5.17
## 4 700014409 Tongue_dorsum 5.30
## 5 700014412 Hard_palate 4.70
## 6 700014415 Buccal_mucosa 4.97
## 7 700014418 Attached_Keratinized_gingiva 4.57
## 8 700014421 Palatine_Tonsils 5.06
## 9 700014424 Throat 4.81
## 10 700014427 Supragingival_plaque 4.97
## # ... with 22 more rows
```

summarise() + vegan

```
otu %>%
  group_by(specimen_id, HMPbodysubsite) %>%
  summarise(
    specimen_shannon = diversity(x = read_count,
                                index = "shannon",
                                base = exp(1))
  ) %>%
  ungroup() %>%
  group_by(HMPbodysubsite) %>%
  summarise(
    median_shannon = median(specimen_shannon)
  ) %>%
  ungroup()
```

```
## # A tibble: 16 x 2
##   HMPbodysubsite      median_shannon
##   <chr>          <dbl>
## 1 Anterior_nares      4.54
## 2 Attached_Keratinized_gingiva 4.62
## 3 Buccal_mucosa       5.37
## 4 Hard_palate         5.47
## 5 Left_Retroauricular_crease 5.31
## 6 Mid_vagina          2.91
## 7 Palatine_Tonsils    5.33
## 8 Posterior_fornix     2.83
## 9 Right_Retroauricular_crease 5.37
## 10 Saliva              6.01
## 11 Stool               5.10
## 12 Subgingival_plaque  5.84
## 13 Supragingival_plaque 5.82
## 14 Throat              5.31
## 15 Tongue_dorsum       5.80
## 16 Vaginal_introitus   3.01
```

Questions

- What is the median Shannon diversity of the Stool specimens?
- What's the median sequencing depth of the Stool specimens? (use `sum()`)
- Compare the Shannon diversity of Stool samples with and without rarefaction to 1000 reads per specimen. How do the diversity estimates differ?



Post questions to the discussion board!

Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu

