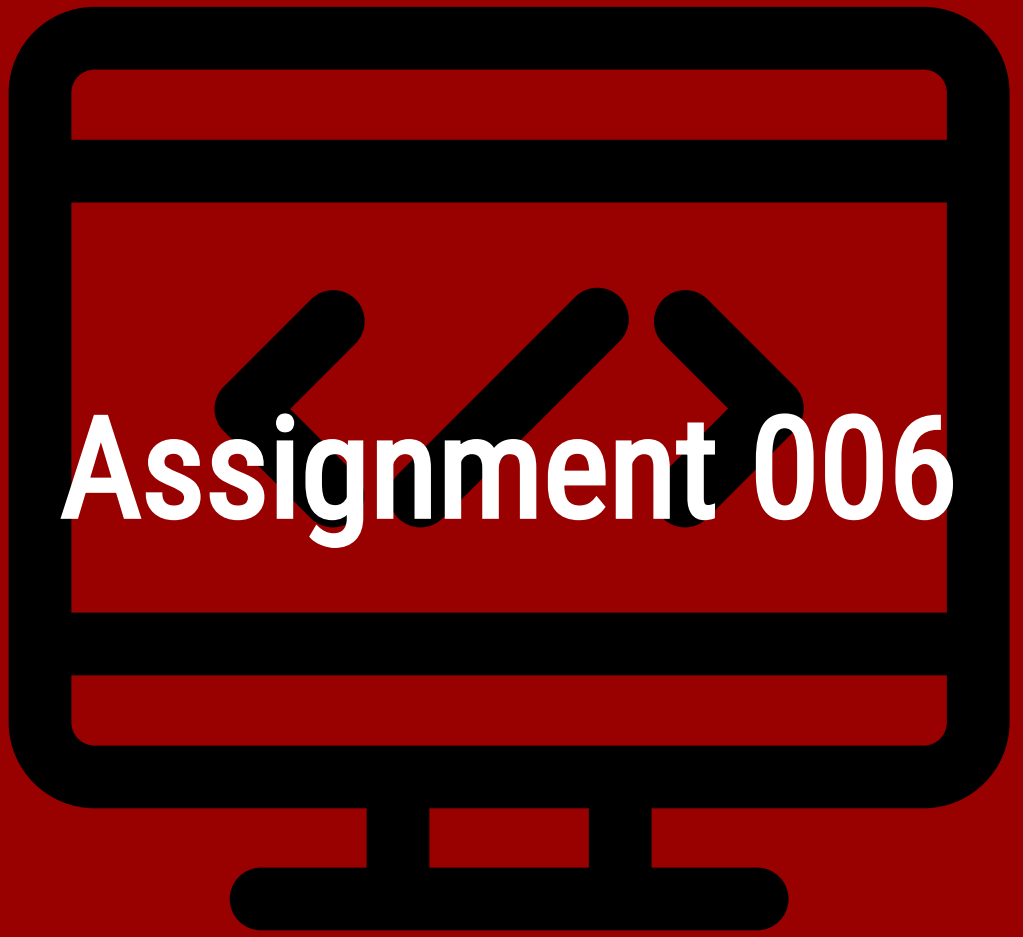
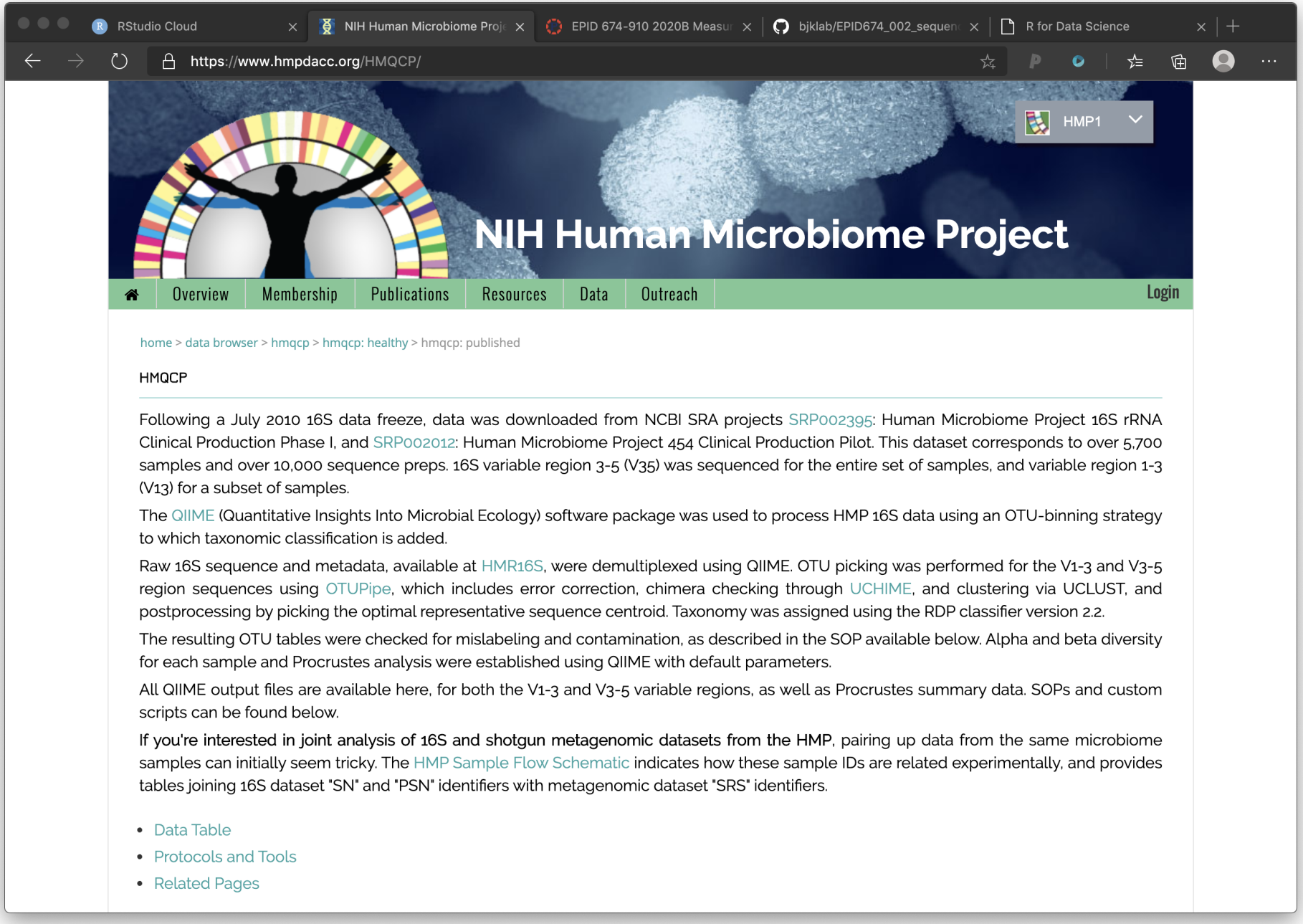

Assignment 006

Beta Diversity in R

- Clone the sixth class repository: https://github.com/bjklab/EPID674_006_beta-diversity.git
- Install & load `tidyverse`, `vegan`, & `ape` packages
- Examine OTUs, distances, and principal coordinates
- Assigment:
  - Calculate pairwise distances
  - Plot principal coordinates


RStudio Cloud
NIH Human Microbiome Proj
EPID 674-910 2020B Measur
bjklab/EPID674_002_sequenc
R for Data Science
https://www.hmpdacc.org/HMQCP/
HMP1
NIH Human Microbiome Project
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

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.

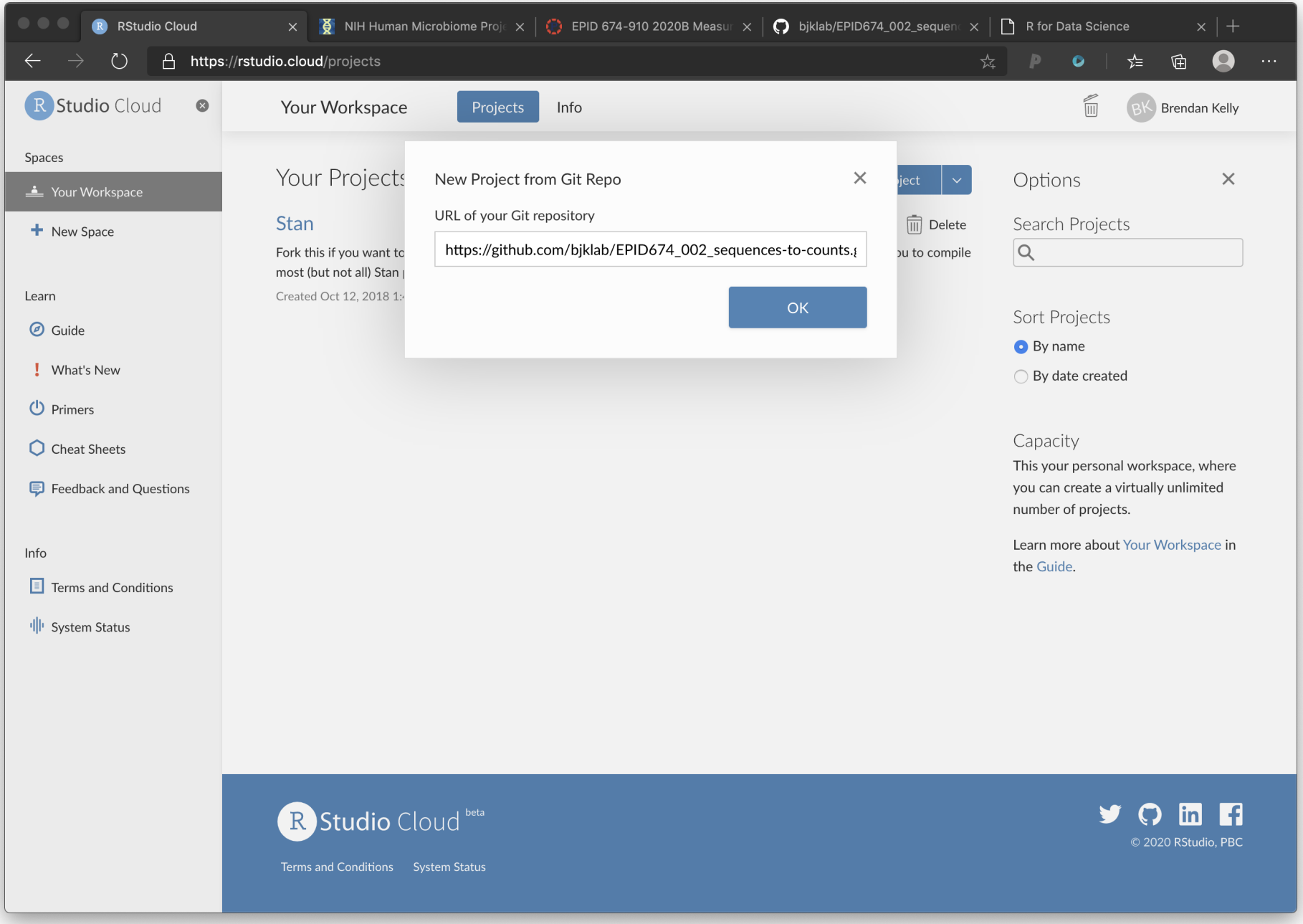
If you're interested in joint analysis of 16S and shotgun metagenomic datasets from the HMP, pairing up data from the same microbiome samples can initially seem tricky. The HMP Sample Flow Schematic indicates how these sample IDs are related experimentally, and provides tables joining 16S dataset "SN" 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	0cc3b82cf16d46c303f902bfbd7b88e9		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. Mislabeled 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
NIH Human Microbiome Proj
EPID 674-910 2020B Measur
bjklab/EPID674_002_sequenc
R for Data Science
https://rstudio.cloud/projects
Studio Cloud
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
most (but not all) Stan
Created Oct 12, 2018 1:
New Project from Git Repo
URL of your Git repository
https://github.com/bjklab/EPID674_002_sequences-to-counts.g
OK
Delete
u to compile
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.
Studio Cloud beta
Terms and Conditions
System Status
© 2020 RStudio, PBC

Using the rstudio.cloud console

```
# make sure packages installed
# install.packages("tidyverse")
# install.packages("vegan")
# install.packages("ape")

# load packages
library(tidyverse)
library(vegan)
library(ape)

# load data
otu_long <- read_csv(
file =
"./data/HMP_OTU_table_longformat_stool_nares.csv
)

otu_long # show what you've read
```

```
## # A tibble: 431,400 x 4
##    otu_id       specimen_id read_count HMPbodysubsite
##    <chr>              <dbl>      <dbl> <chr>
##  1 OTU_97.1       700014718          0 Stool
##  2 OTU_97.10      700014718          0 Stool
##  3 OTU_97.100     700014718          0 Stool
##  4 OTU_97.1000    700014718          0 Stool
##  5 OTU_97.10000   700014718          0 Stool
##  6 OTU_97.10001   700014718          0 Stool
##  7 OTU_97.10002   700014718          0 Stool
##  8 OTU_97.10003   700014718          0 Stool
##  9 OTU_97.10004   700014718          0 Stool
## 10 OTU_97.10005   700014718          0 Stool
## # … with 431,390 more rows
```

Using the rstudio.cloud console

```
# load data
otu_matrix <- read_rds(
path =
"./data/HMP_OTU_table_matrix_stool_nares.rds"
)

# microbiome format
otu_matrix %>%
  str(vec.len = 2)

# ecology format
otu_matrix %>%
  t() %>% # TRANSPOSE
  str(vec.len = 2)
```

```
##  num [1:43140, 1:10] 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:43140] "OTU_97.1" "OTU_97.10" ...
##   ..$ : chr [1:10] "700014718" "700014767" ...

##  num [1:10, 1:43140] 0 0 0 0 0 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:10] "700014718" "700014767" ...
##   ..$ : chr [1:43140] "OTU_97.1" "OTU_97.10" ...
```

Using the rstudio.cloud console

```
# load data
hmp_dm <- read_csv(
file =
"./data/HMP_DM_weighted_unifrac_stool_nares.csv
)

hmp_pc <- read_csv(
file =
"./data/HMP_PC_weighted_unifrac_stool_nares.csv
)

# long-format distance matrix
hmp_dm %>%
  print(n = 4)

# long-format principal coordinates
hmp_pc %>%
  print(n = 4)
```

```
## # A tibble: 106,276 x 5
##   specimen_1 specimen_2 weighted_unifrac site_1
##        <dbl>      <dbl>            <dbl> <chr>
## 1  700014386  700014386            0     Stool
## 2  700113013  700014386            0.217 Stool
## 3  700111523  700014386            0.721 Anterior_nares
## 4  700105685  700014386            0.232 Stool
## # … with 106,272 more rows

## # A tibble: 326 x 4
##   specimen_id    PC1    PC2 HMPbodysubsite
##         <dbl>  <dbl>  <dbl> <chr>
## 1   700014386 0.0893  0.144 Stool
## 2   700113013 0.0857  0.127 Stool
## 3   700111523 0.231  -0.138 Anterior_nares
## 4   700105685 0.110   0.156 Stool
## # … with 322 more rows
```

vegan::vegdist()

```
otu_matrix %>%
  t() %>% #TRANSPOSE
  vegdist(x = .,
          method = "jaccard",
          binary = TRUE) %>%
  as.matrix() %>%
  str(vec.len = 2)
```

```
##  num [1:10, 1:10] 0 1 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:10] "700014718" "700014767" ...
##   ..$ : chr [1:10] "700014718" "700014767" ...
```

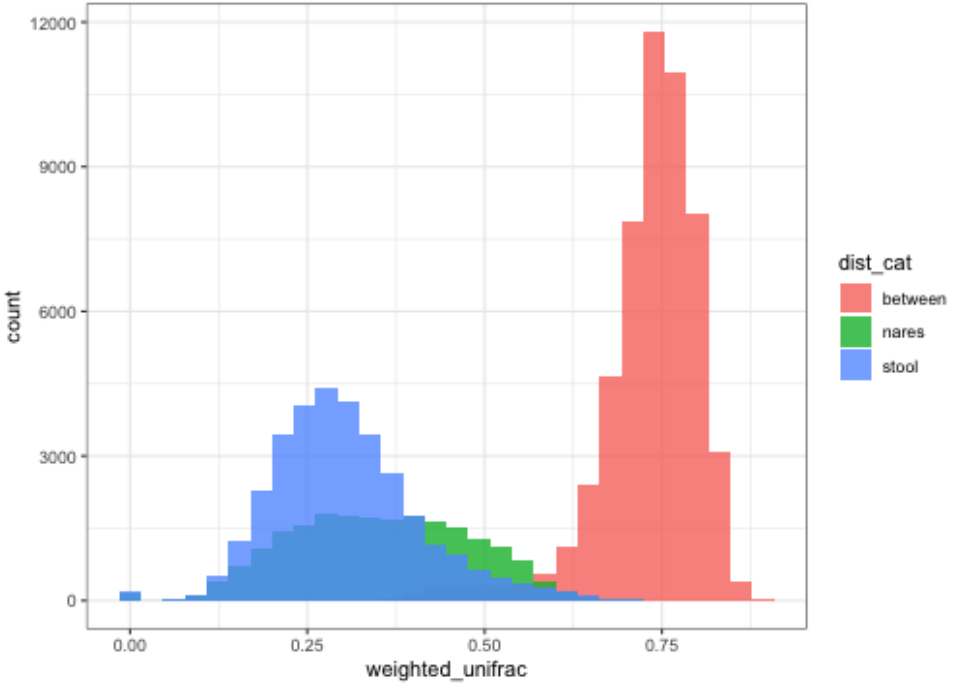
ape::pcoa()

```
otu_matrix %>%
  t() %>% #TRANSPOSE
  vegdist(x = .,
          method = "jaccard",
          binary = TRUE) %>%
  pcoa() %>%
  str(vec.len = 2)
```

```
## List of 5
##  $ correction: chr [1:2] "none" "1"
##  $ note      : chr "There were no negative eigenvalues.
##  $ values    :'data.frame':    9 obs. of  5 variables:
##   ..$ Eigenvalues   : num [1:9] 0.863 0.562 ...
##   ..$ Relative_eig  : num [1:9] 0.21 0.137 ...
##   ..$ Broken_stick  : num [1:9] 0.314 0.203 ...
##   ..$ Cumul_eig     : num [1:9] 0.21 0.347 ...
##   ..$ Cumul_br_stick: num [1:9] 0.314 0.518 ...
##  $ vectors   : num [1:10, 1:9] 0.336 -0.27 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : chr [1:10] "700014718" "700014767" ...
##   .. ..$ : chr [1:9] "Axis.1" "Axis.2" ...
##  $ trace     : num 4.11
##  - attr(*, "class")= chr "pcoa"
```

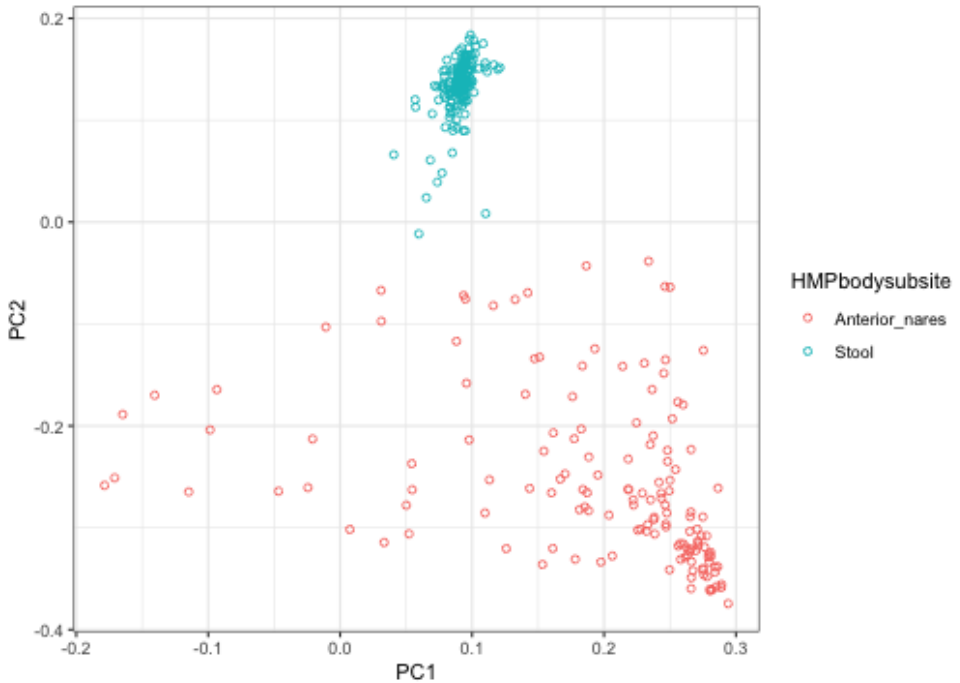
Plot Distances

```
hmp_dm %>%
  mutate(dist_cat =
           ifelse(site_1 != site_2,
                  "between",
                  ifelse(site_1 == "Stool",
                         "stool",
                         "nares"))) %>%
  ggplot(data = .) +
  geom_histogram(aes(
    x = weighted_unifrac,
    fill = dist_cat
  ), alpha = 0.8,
  position = "identity") +
  theme_bw()
```



Plot Principal Coordinates

```
hmp_pc %>%
  ggplot(data = .) +
  geom_point(
    aes(
    x = PC1,
    y = PC2,
    color = HMPbodysubsite
  ),
  shape = 21) +
  theme_bw()
```



Questions

- What is the relationship between PERMANOVA testing and the distance matrix (look at slide #11)?
- Does PCoA show the relationship between specimens (communities) or between OTUs?

Post questions to the discussion board!

Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu