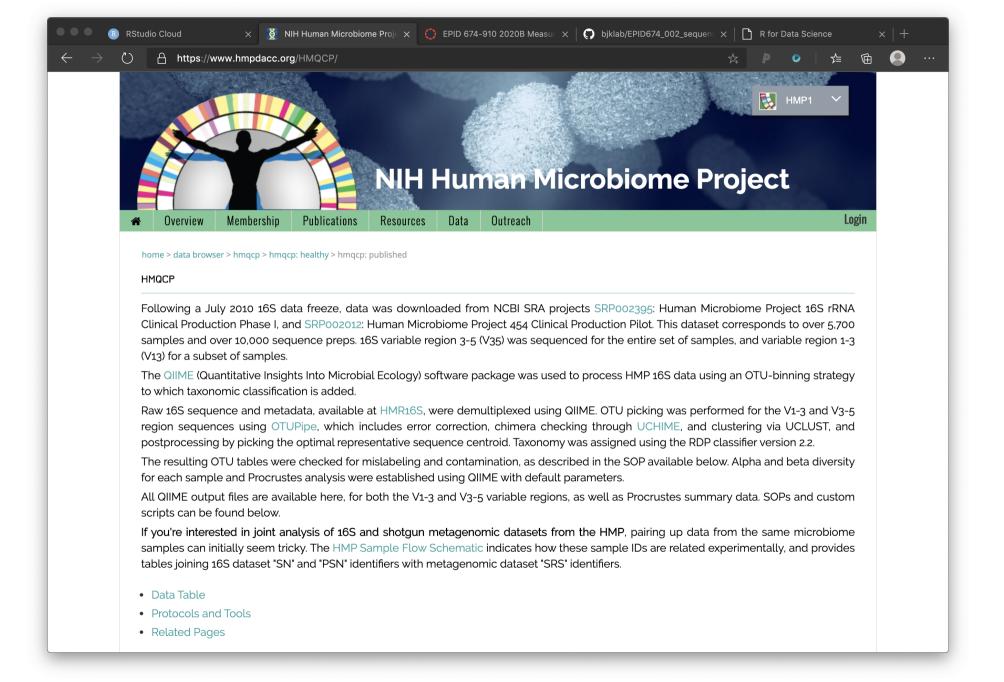
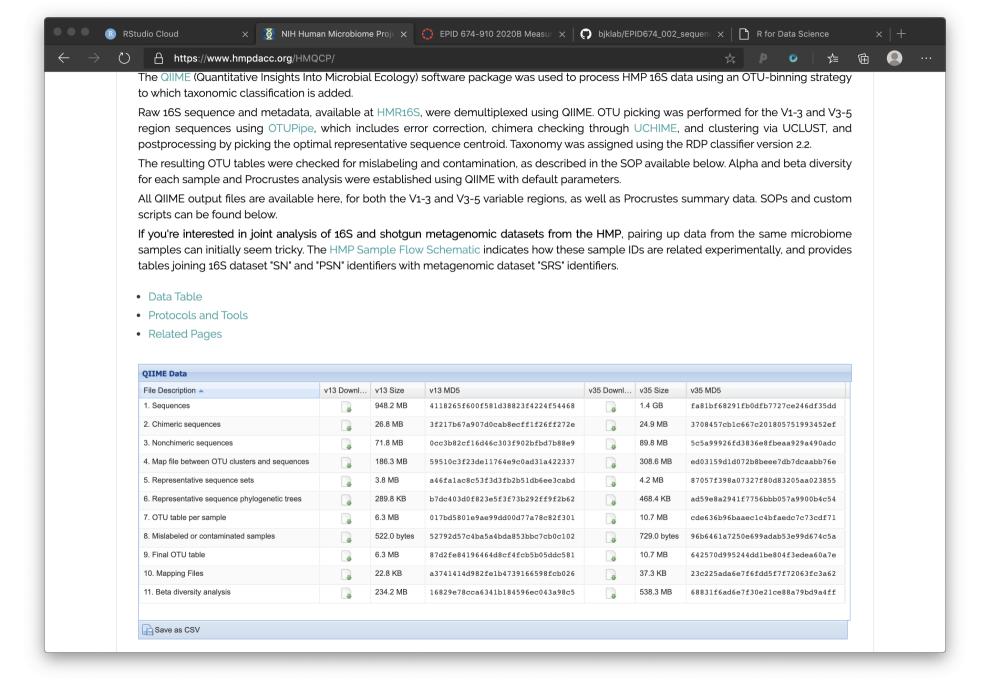
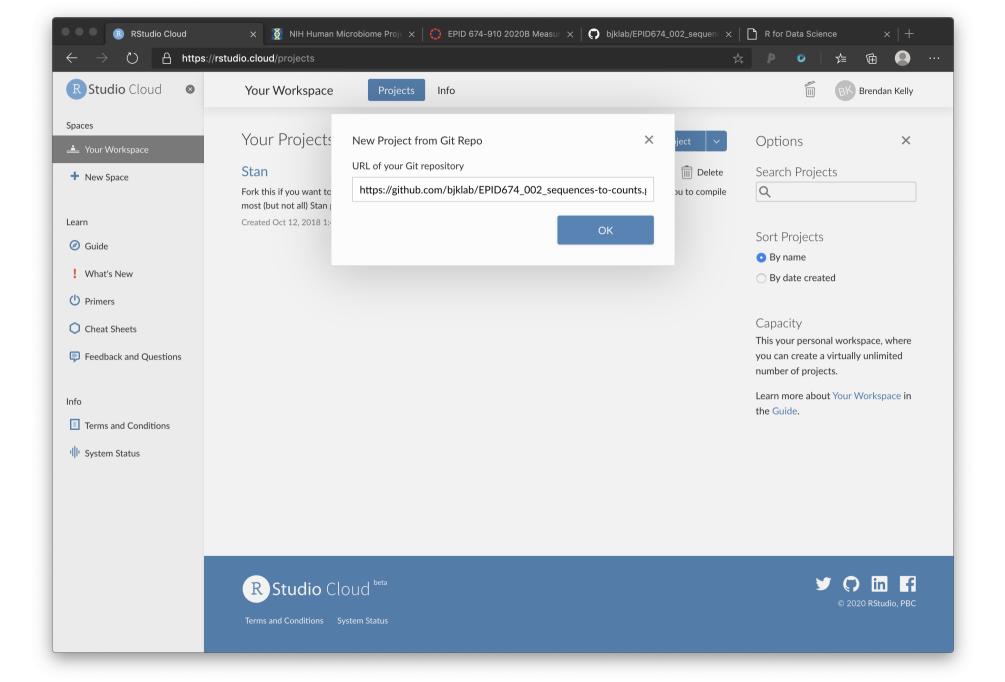


Summarise Alpha Diversity in R

- 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
- Assigment:
 - Summarize by specimen: unique OTUs, sequencing depth
 - Summarize by specimen: alpha diversity







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</pre>
```

```
## # A tibble: 1,380,480 x 4
      HMPbodysubsite specimen_id otu_id
                                             read count
      <chr>>
                          <dbl> <chr>
                                                  <fdb>
   1 Anterior nares 700014445 OTU 97.1
                                                      0
   2 Anterior nares 700014445 OTU 97.10
   3 Anterior nares 700014445 OTU 97.100
   4 Anterior nares 700014445 OTU 97.1000
                     700014445 OTU 97.10000
   5 Anterior nares
                      700014445 OTU 97.10001
   6 Anterior nares
                      700014445 OTU_97.10002
   7 Anterior nares
   8 Anterior nares
                    700014445 OTU 97.10003
   9 Anterior nares
                      700014445 OTU_97.10004
                                                      0
## 10 Anterior nares
                      700014445 OTU 97.10005
                                                      0
## # ... with 1,380,470 more rows
```

Try mutate()

```
## # A tibble: 1,380,480 x 5
      HMPbodysubsite specimen_id otu_id
                                              read_count log_reads
      <chr>>
                           <dbl> <chr>
                                                   <dbl>
                                                              <dbl>
                      700014445 OTU_97.1
                                                               -Inf
   1 Anterior nares
   2 Anterior nares
                      700014445 OTU 97.10
                                                               -Inf
                       700014445 OTU_97.100
                                                               -Inf
   3 Anterior nares
   4 Anterior nares
                       700014445 OTU 97.1000
                                                               -Inf
   5 Anterior nares
                       700014445 OTU_97.10000
                                                               -Inf
                       700014445 OTU 97.10001
                                                               -Inf
   6 Anterior nares
## 7 Anterior nares
                       700014445 OTU_97.10002
                                                               -Inf
   8 Anterior nares
                       700014445 OTU_97.10003
                                                               -Inf
   9 Anterior nares
                       700014445 OTU 97.10004
                                                               -Inf
## 10 Anterior nares
                       700014445 OTU_97.10005
                                                               -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	${\sf mean_reads}$	sum_reads
##		<chr></chr>	<dbl></dbl>	<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

ии и A ваньта, 10 у о

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

```
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")
```

```
## # A tibble: 32 x 3
      specimen_id HMPbodysubsite
                                               rare shannon
            <dbl> <chr>
                                                      <dbl>
       700013549 Stool
                                                       4.58
## 2 700014386 Stool
                                                       5.05
       700014403 Saliva
                                                       5.17
       700014409 Tongue_dorsum
                                                       5.30
       700014412 Hard_palate
                                                       4.70
       700014415 Buccal mucosa
                                                       4.97
       700014418 Attached_Keratinized_gingiva
                                                       4.57
       700014421 Palatine Tonsils
                                                       5.06
       700014424 Throat
                                                       4.81
       700014427 Supragingival_plaque
                                                       4.97
## # ... with 22 more rows
```

##	# /	A tibble: 16 x 2	
##		HMPbodysubsite	median_shannon
##		<chr></chr>	<dbl></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

ии и и и на Бъта и и о

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?



Thank you!

Slides available: github.com/bjklab

brendank@pennmedicine.upenn.edu