Exploratory data analysis of .biom files

2019-01-01

Settings

Load data

Data source: https://www.hmpdacc.org/hmp/, https://portal.hmpdacc.org/ - data portal. Files to download. Download with `scripts/ascp-commands.sh`

Biom EDA

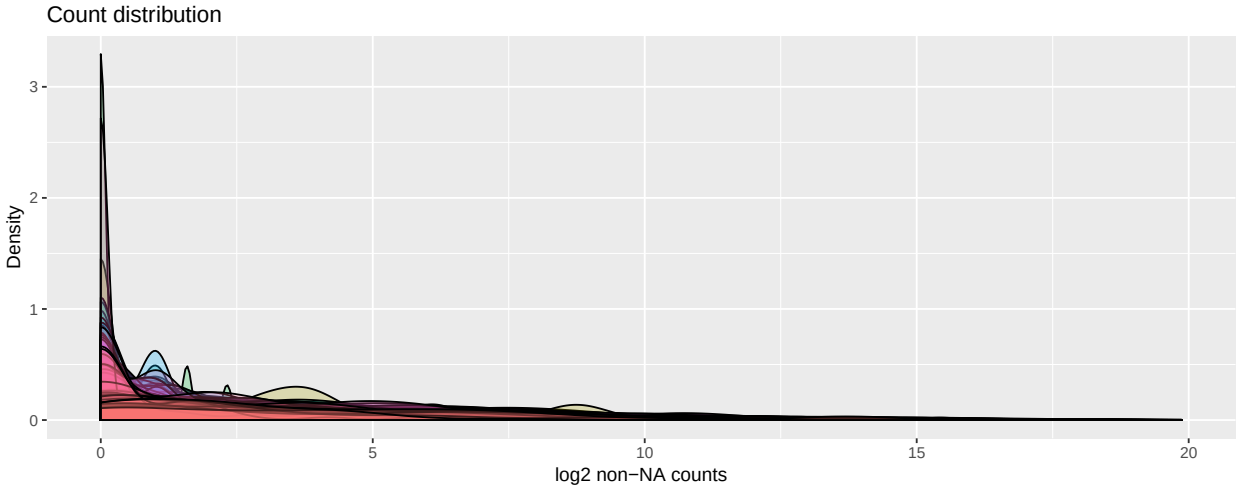
Biom data dimensions: 7665, 9108

```
       ID EP003595_K10_MV1D EP003595_K100_BRCD EP003595_K90_BCKD
1  100039                NA                 NA                NA
2 1000547                NA                 NA                NA
3 1005406                NA                 NA                NA
4 1005533                NA                 NA                NA
5 1007399                NA                 NA                NA
  EP003595_K90_BRCD
1                NA
2                NA
3                NA
4                NA
5                NA
```

Count distribution for 500 randomly selected samples

```
[1] "Summary of count data"
```

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
    1.0      1.0      4.0    342.7     28.0 956642.0
```



Question: There are outlier counts. With median/mean counts equal to 4/300, respectively, we have counts as high as 956642.

Library sizes (column sums)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      2    1912   19050   34412   45663 1501486
```

Taxa counts (row sums)

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0       0       1    2245      14 3222950
```

Library variability (column SD)

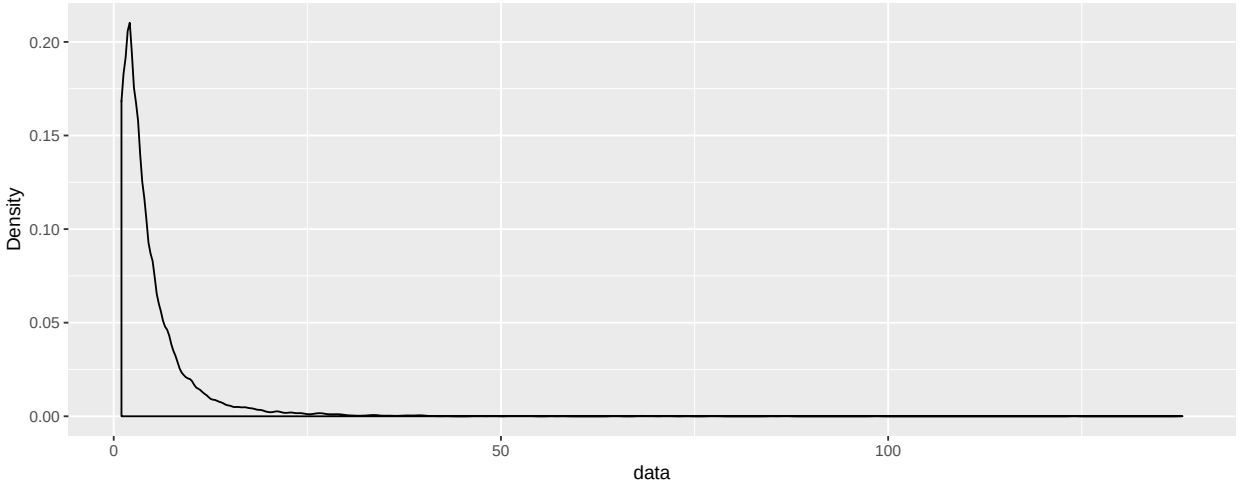
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    0.0   110.3   679.3  2431.8  2311.1 71345.8
```

Taxa variability (row SD)

```
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.     NA's
   0.00     0.92     4.95   175.73    35.34 46841.49     4791
```

Distribution of sample medians for all samples

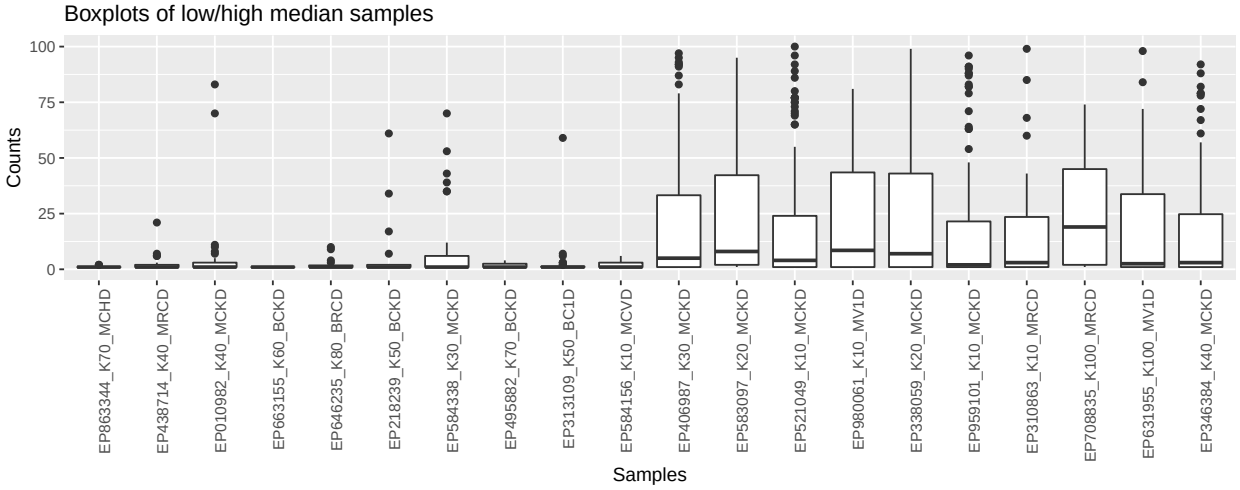
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.854   6.000 138.000
```



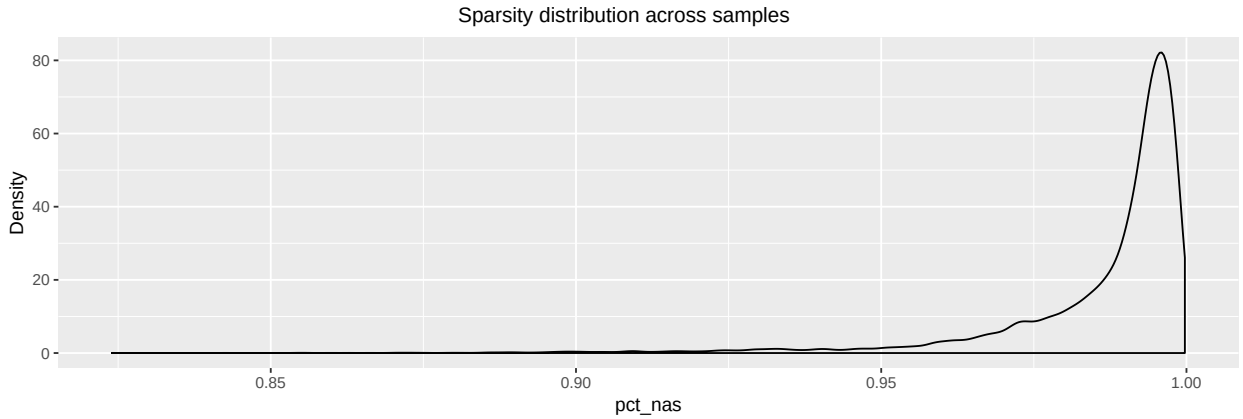
Question: The range of medians vary across samples. Most samples have median counts ~3-4, but some have as high as 138.

Boxplots of low/high median samples

We see how different are count distributions between low/high median samples



Sparsity of samples (across columns)




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.8240  0.9842  0.9928  0.9868  0.9961  0.9997
```

Question: The data across samples is very sparse (~97% NAs) - is it expected? There are samples with >99% sparsity.

```
[1] "Top 10 least sparse samples:"
```

file	sample_body_site	visit_number
EP070043_K20_MRCD	rectum	2
EP488403_K30_MRCD	rectum	3
EP580086_K30_MRCD	rectum	3
EP631608_K10_MRCD	rectum	1
EP631608_K20_MRCD	rectum	2
EP646001_K10_MRCD	rectum	1
EP751500_K10_MRCD	rectum	1
EP751500_K20_MRCD	rectum	2
EP936464_K100_MRCD	rectum	10
EP949081_K40_MRCD	rectum	4

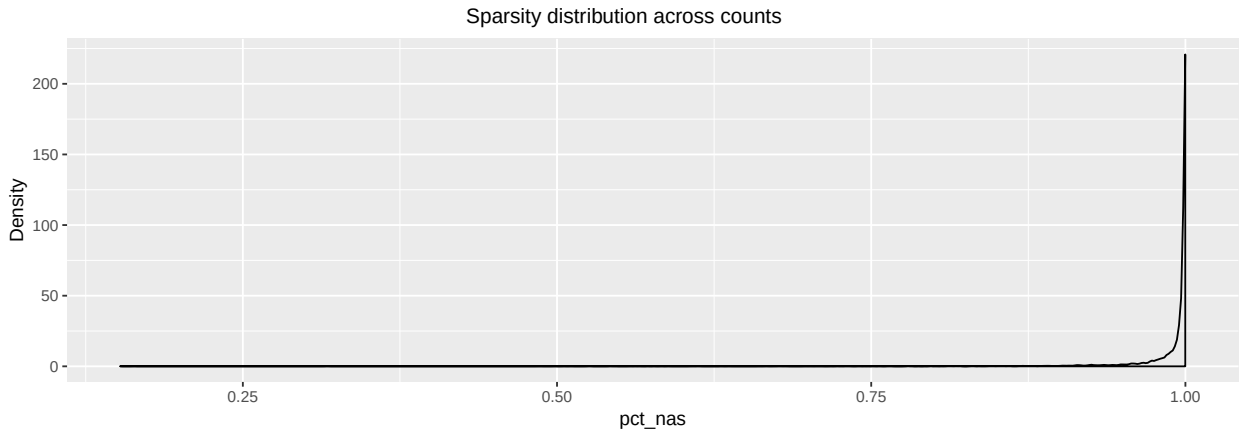
```
[1] "Top 10 most sparse samples:"
```

file	sample_body_site	visit_number
EP338059_K70_BRCD	rectum	7
EP463151_K70_BS1D	feces	7
EP550087_K10_MCKD	buccal mucosa	1
EP553021_K100_MCKD	buccal mucosa	10
EP641566_K340_BCKD	buccal mucosa	34
EP673818_K30_MCKD	buccal mucosa	3
EP771695_K150_BRCD	rectum	15
EP785033_K90_MCHD	unknown	9
EP828421_K40_BRCD	rectum	4
EP853831_K70_MCKD	buccal mucosa	7

file	sample_body_site	visit_number
EP897729_K110_BCKD	buccal mucosa	11

Observation: Samples from rectum and first visits may be least sparse. Samples from buccal mucosa/rectum and later visits may be most sparse.

Sparsity of counts (across rows)




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1525  0.9924  0.9988  0.9868  0.9998  0.9999
```

Question: The data across counts is very sparse (~99% NAs) - is it expected? There are counts with almost 100% sparsity.

```
[1] "Top 10 least sparse counts:"
```

ID	Taxonomy
130468	k___Bacteria; p___Firmicutes; c___Bacilli; o___Lactobacillales; f___Lactobacillaceae; g**Lactobacillus; s**
332718	k___Bacteria; p___Firmicutes; c___Bacilli; o___Lactobacillales; f___Lactobacillaceae; g**Lactobacillus; s**
134265	k___Bacteria; p___Bacteroidetes; c___Bacteroidia; o___Bacteroidales; f___Prevotellaceae; g**Prevotella; s**
137183	k___Bacteria; p___Actinobacteria; c___Actinobacteria; o___Bifidobacteriales; f___Bifidobacteriaceae; g**Gardnere**
469663	k___Bacteria; p___Actinobacteria; c___Coriobacteriia; o___Coriobacteriales; f___Coriobacteriaceae; g___Atopobiu
581782	k___Bacteria; p___Proteobacteria; c___Gammaproteobacteria; o___Enterobacteriales; f___Enterobacteriaceae; g**;**
137580	k___Bacteria; p___Firmicutes; c___Bacilli; o___Lactobacillales; f___Lactobacillaceae; g**Lactobacillus; s**
529233	k___Bacteria; p___Firmicutes; c___Bacilli; o___Lactobacillales; f___Streptococcaceae; g**Streptococcus; s**
580629	k___Bacteria; p___Bacteroidetes; c___Bacteroidia; o___Bacteroidales; f___Bacteroidaceae; g**Bacteroides; s**
1109247	k___Bacteria; p___Proteobacteria; c___Gammaproteobacteria; o___Enterobacteriales; f___Enterobacteriaceae; g**;**

```
[1] "Top 10 most sparse counts:"
```

ID	Taxonomy
982877	k___Bacteria; p___Actinobacteria; c___Actinobacteria; o___Actinomycetales; f___Actinomycetaceae; g**Varibaculu**
982363	k___Bacteria; p___Actinobacteria; c___Actinobacteria; o___Actinomycetales; f___Corynebacteriaceae; g**Coryneba**
956702	k___Bacteria; p___Proteobacteria; c___Gammaproteobacteria; o___Pasteurellales; f___Pasteurellaceae; g___Haemo
995822	k___Bacteria; p___Firmicutes; c___Bacilli; o___Gemellales; f___Gemellaceae; g**; s**
965048	k___Bacteria; p___Proteobacteria; c___Betaproteobacteria; o___Neisseriales; f___Neisseriaceae; g**; s**
968363	k___Bacteria; p___Firmicutes; c___Clostridia; o**Clostridiales; f**[Tissierellaceae]; g**Anaerococcus; s**
982266	k___Bacteria; p___Proteobacteria; c___Gammaproteobacteria; o**Alteromonadales; f**[Chromatiaceae]; g**Rheinhe**
99517	k___Bacteria; p___Bacteroidetes; c___Bacteroidia; o___Bacteroidales; f___Porphyromonadaceae; g**Dysgonomona**
955102	k___Bacteria; p___Actinobacteria; c___Actinobacteria; o___Actinomycetales; f___Actinomycetaceae; g**Actinomyc**
953463	k___Bacteria; p___Proteobacteria; c___Gammaproteobacteria; o___Pasteurellales; f___Pasteurellaceae; g**Aggrega**

ID	Taxonomy
96894	k___Bacteria; p___Actinobacteria; c___Actinobacteria; o___Actinomycetales; f___Microbacteriaceae; g**; s**

Observation: Counts for "Lacto*" bacteria seem to be least sparse.

Metadata EDA

```
# Quick EDA
table(mtx_metadata$subject_gender) %>% sort(., decreasing = TRUE)
```

```
female
  9107
```

```
table(mtx_metadata$subject_race) %>% sort(., decreasing = TRUE)
```

```
unknown
   9107
```

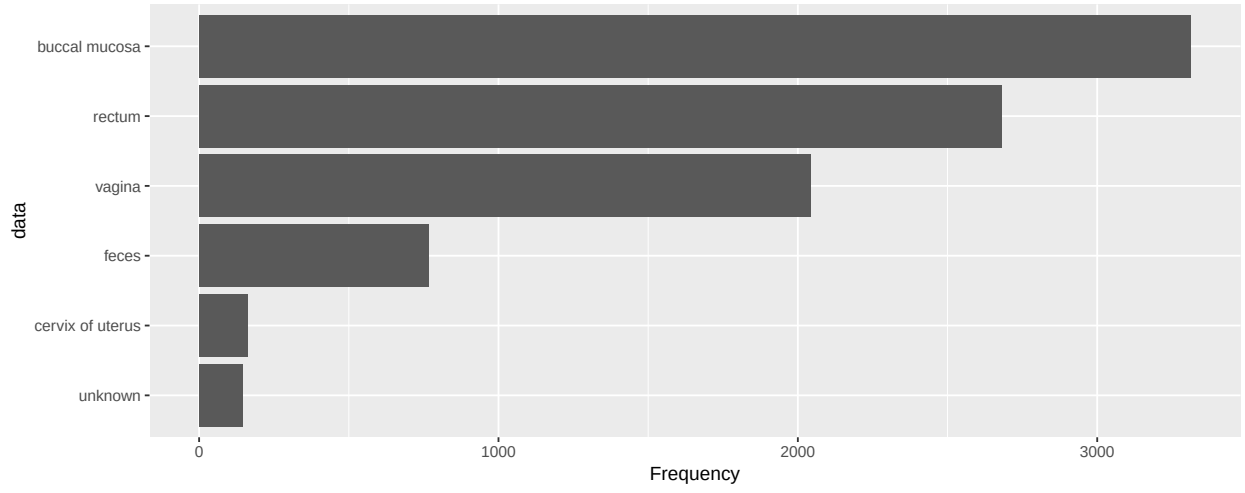
```
table(mtx_metadata$study_full_name) %>% sort(., decreasing = TRUE)
```

```
momspi
  9107
```

```
table(mtx_metadata$project_name) %>% sort(., decreasing = TRUE)
```

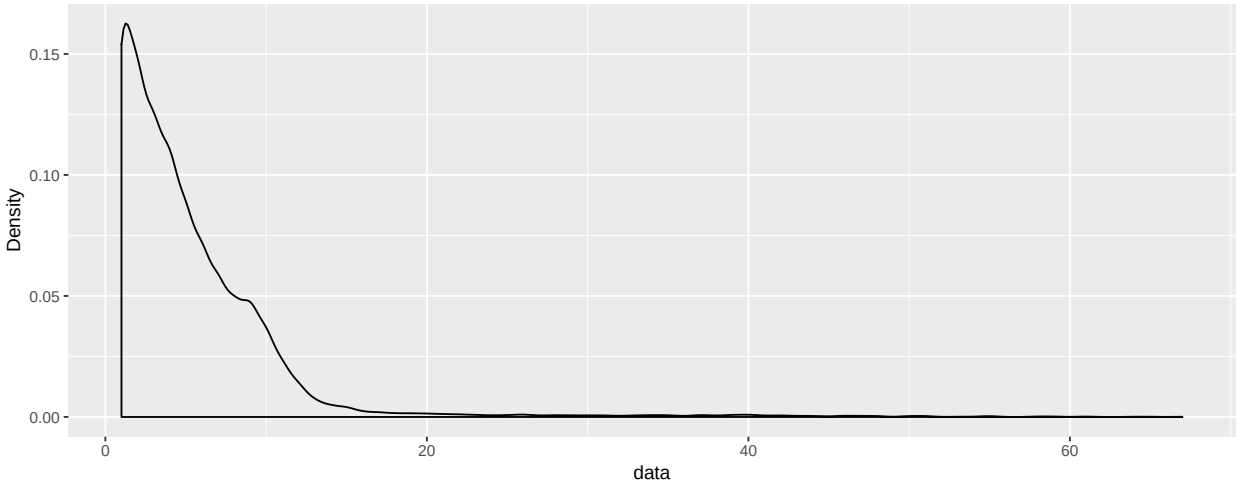
```
Integrative Human Microbiome Project
                                9107
```

sample_body_site




```
   buccal mucosa          rectum          vagina           feces
            3313            2679            2042             765
cervix of uterus         unknown
             162             146
```

visit_number




```
   1    2    3    4    5    6    7    9    8   10   11   12   13   14   15
1680 1314 1134 1017  797  648  534  445  441  338  209  128   64   44   39
  16   17   18   19   20   21   22   26   40   23   37   39   25   28   34
  20   18   14   14   13   11   10   10    9    8    8    8    7    7    7
  35   24   29   30   31   33   42   27   38   41   46   32   43   44   47
   7    6    6    6    6    6    6    5    5    5    5    4    4    4    4
  48   50   51   55   36   45   58   59   61   53   54   64   65   67
   4    4    4    4    3    2    2    2    2    1    1    1    1    1
```

```
[1] "How many total samples: 9107"
```

```
[1] "How many samples at visit 1: 1680"
```

Number of subjects with vaginal samples

```
[1] 9107    9
```

```
[1] 596
```

```
  1   2   3   4   5   6   7   8   9
 82  94 104 147  82  53  17   2   1
```

Save the results