BIOM

2018-11-18

Exploratory data analysis

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

Load data

Read .biom data

```
[1] TRUE
```

Biom EDA

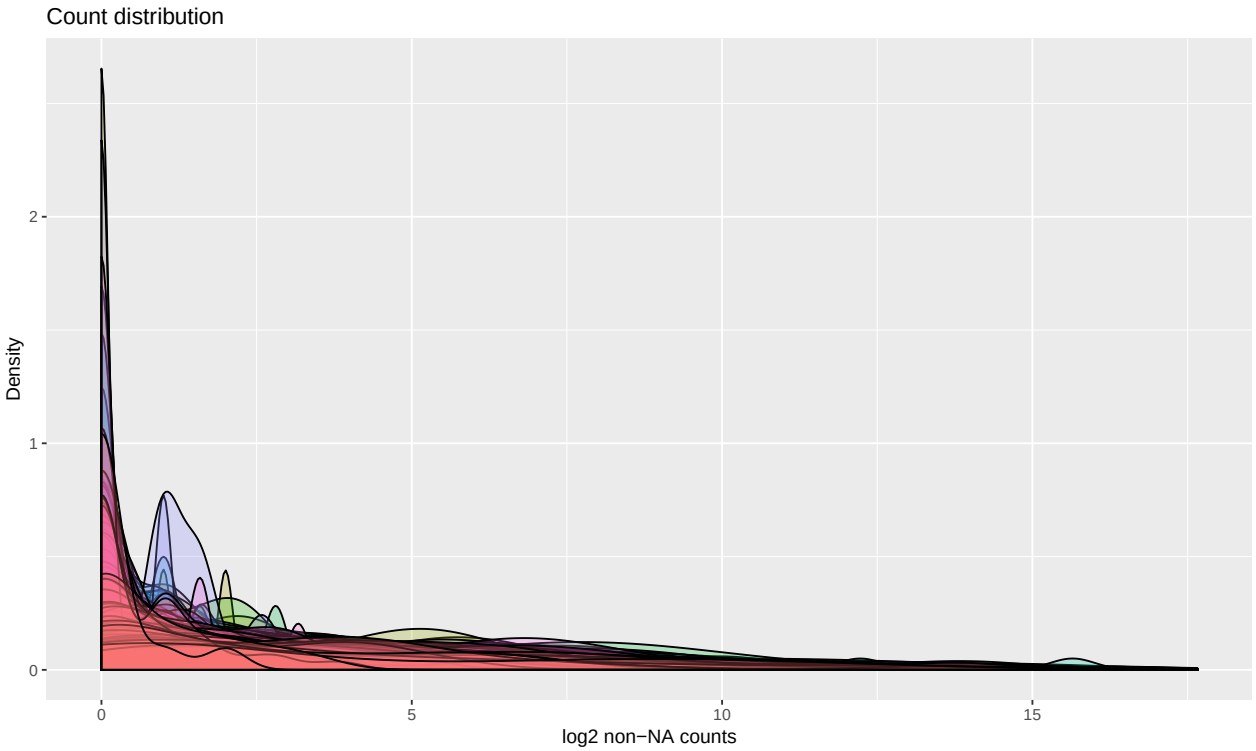
500 files are processed. Data dimensions: 3728, 501

```
      ID EP003595_K10_MV1D.otu_table.biom
1 807795                                20
2 134726                                 1
3 215097                                 2
4 542066                                 1
5 851634                                 1
  EP003595_K100_BRCD.otu_table.biom EP003595_K90_BCKD.otu_table.biom
1                                NA                               NA
2                                NA                               NA
3                                NA                               NA
4                                NA                               NA
5                                NA                               NA
  EP003595_K90_BRCD.otu_table.biom
1                               NA
2                               NA
3                               NA
4                               NA
5                               NA
```

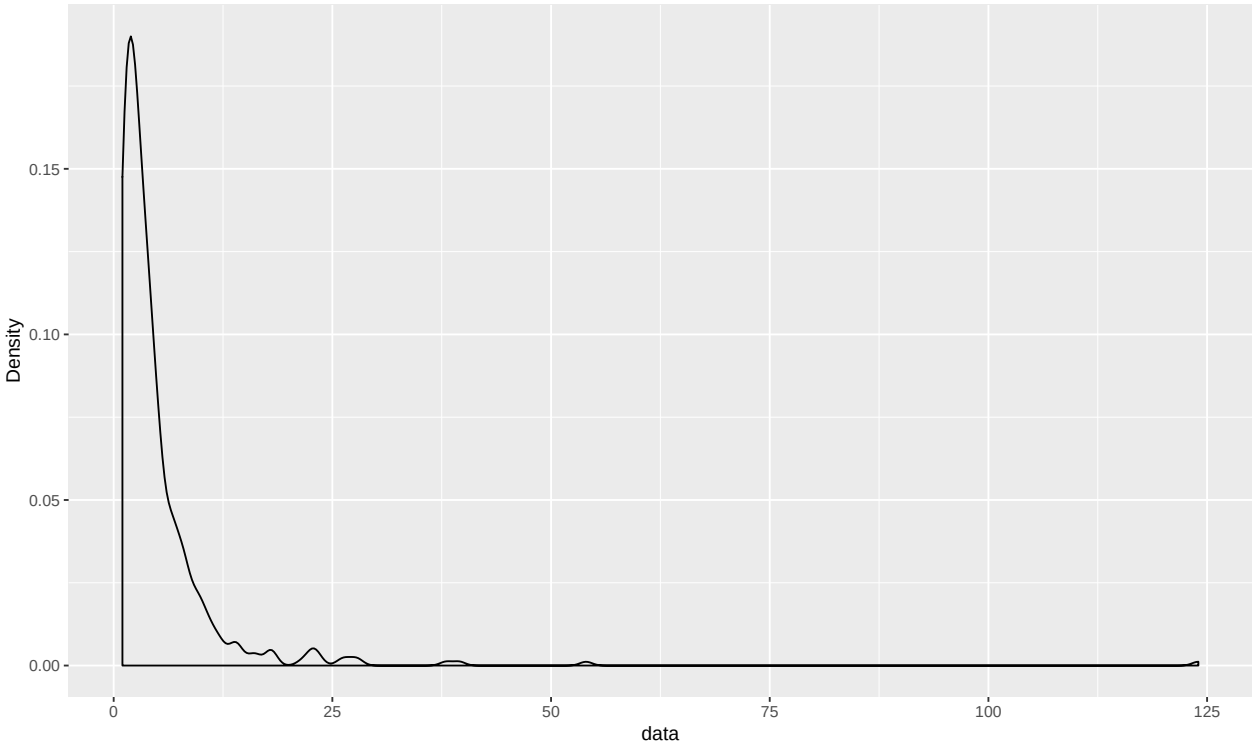
Count distribution

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

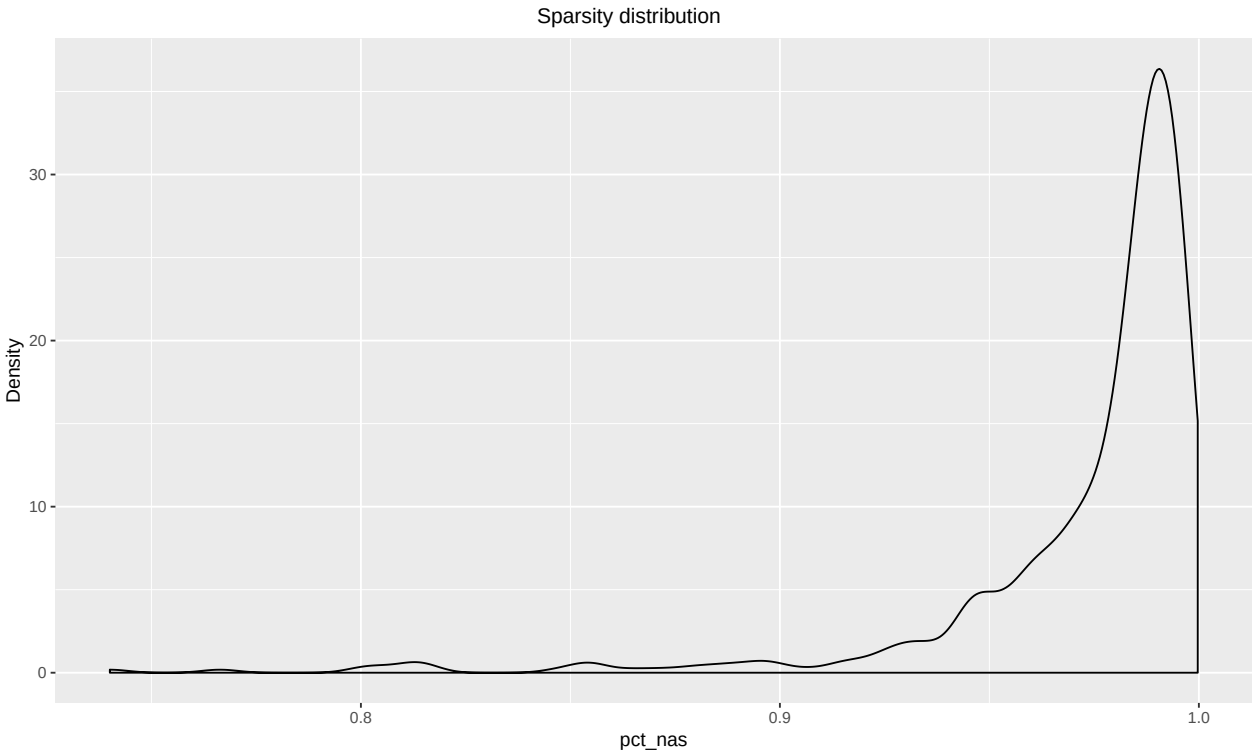
```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
     1.0      1.0      4.0    306.2     28.0 206705.0
```



Distribution of sample medians



Sparsity

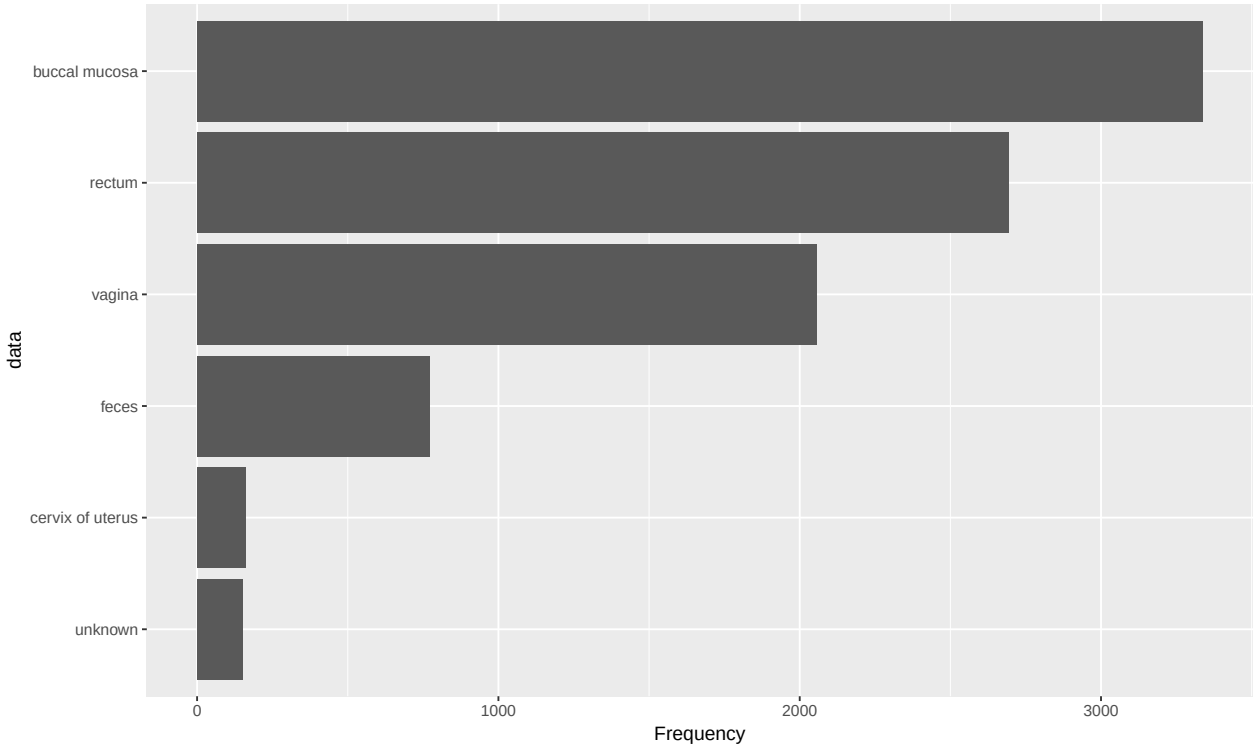



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.7401  0.9694  0.9861  0.9741  0.9920  0.9997
```

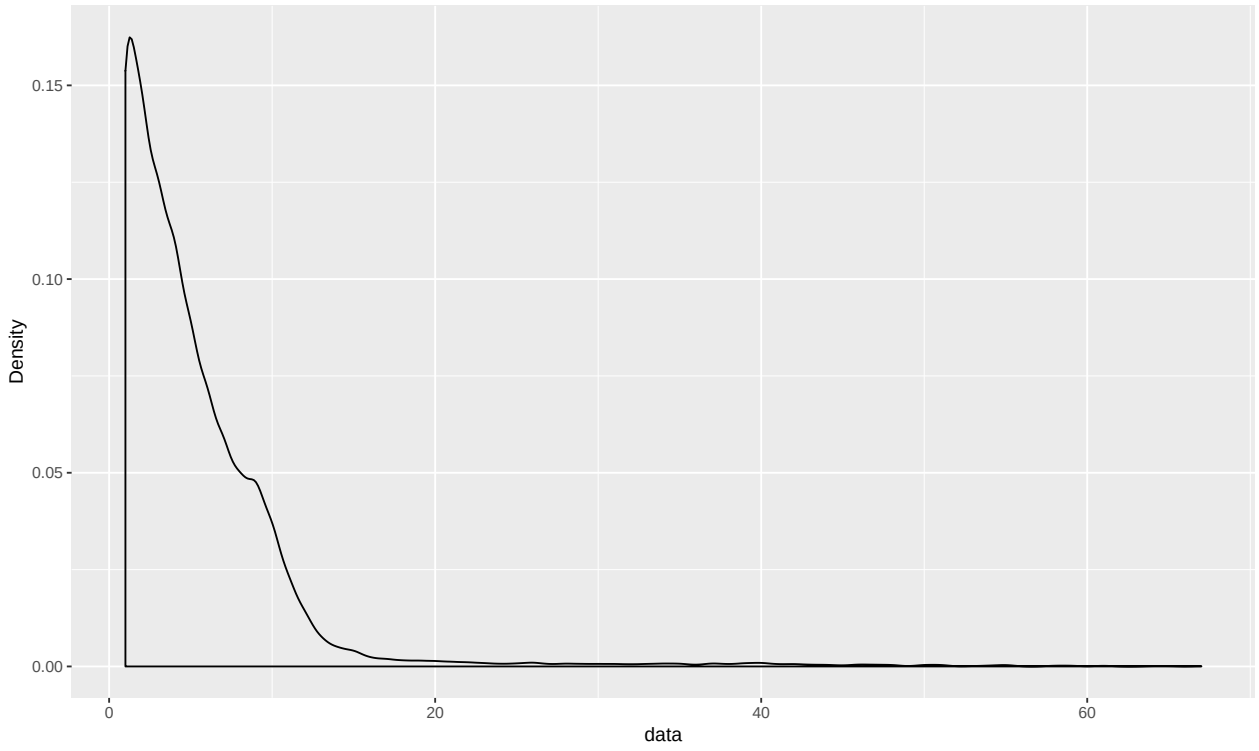
Questions

- What are the row IDs?
- Most sample counts are very small, median = 2. Is it expected?
- There are outliers, like a sample with median counts = 124, maximum count = 206705. Shall we remove such samples?
- The data is very sparse (~97% NAs) - is it expected?

Metadata EDA




```
   buccal mucosa           rectum          vagina           feces
            3338             2692            2055             771
cervix of uterus          unknown
             162              152
```




```
   1    2    3    4    5    6    7    9    8   10   11   12   13   14   15
1688 1325 1143 1019  805  652  540  448  446  341  209  129   65   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   32   38   41   46   43   44   47
   7    6    6    6    6    6    6    5    5    5    5    5    4    4    4
  48   50   51   55   36   45   54   58   59   61   53   64   65   67
   4    4    4    4    3    2    2    2    2    2    1    1    1    1
```

```
[1] "How many total samples: 9170"
```

```
[1] "How many unique sample IDs: 9170"
```

```
[1] "How many unique subject IDs: 596"
```

```
[1] "How many samples at visit 1: 1688"
```

```
# Quick EDA
table(mtx2$subject_gender) %>% sort(., decreasing = TRUE)
```

```
female
  9170
```

```
table(mtx2$subject_race) %>% sort(., decreasing = TRUE)
```

```
unknown
   9170
```

```
table(mtx2$study_full_name) %>% sort(., decreasing = TRUE)
```

```
momspi
  9170
```

```r
table(mtx2$project_name) %>% sort(., decreasing = TRUE)
```

```
Integrative Human Microbiome Project
                                9170
```