

# Exploratory data analysis of .biom files

2018-12-03

## 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

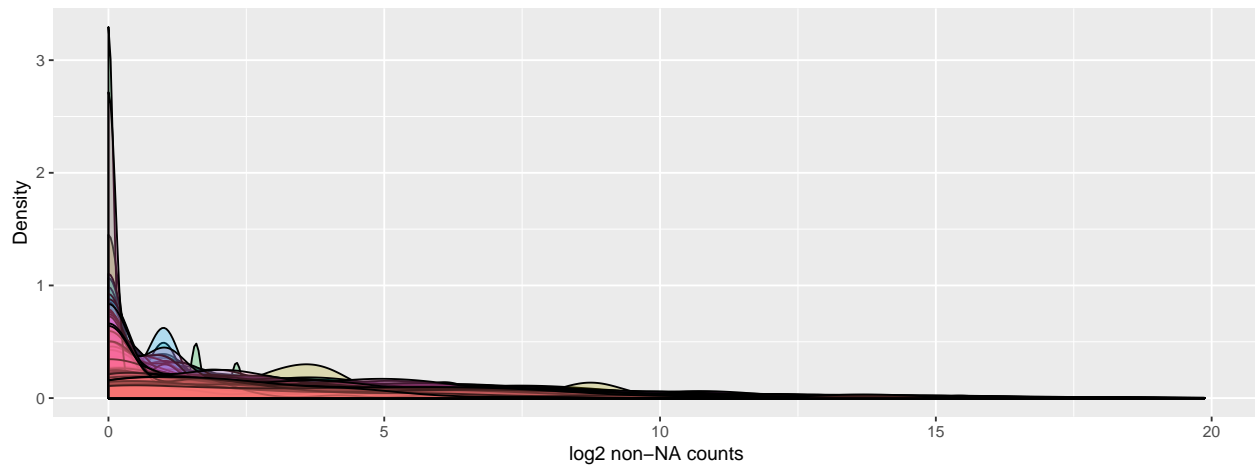
	EP003595_K90_BRCD
1	NA
2	NA
3	NA
4	NA
5	NA

## Count distribution for 500 randomly selected samples

[1] "Summary of count data"

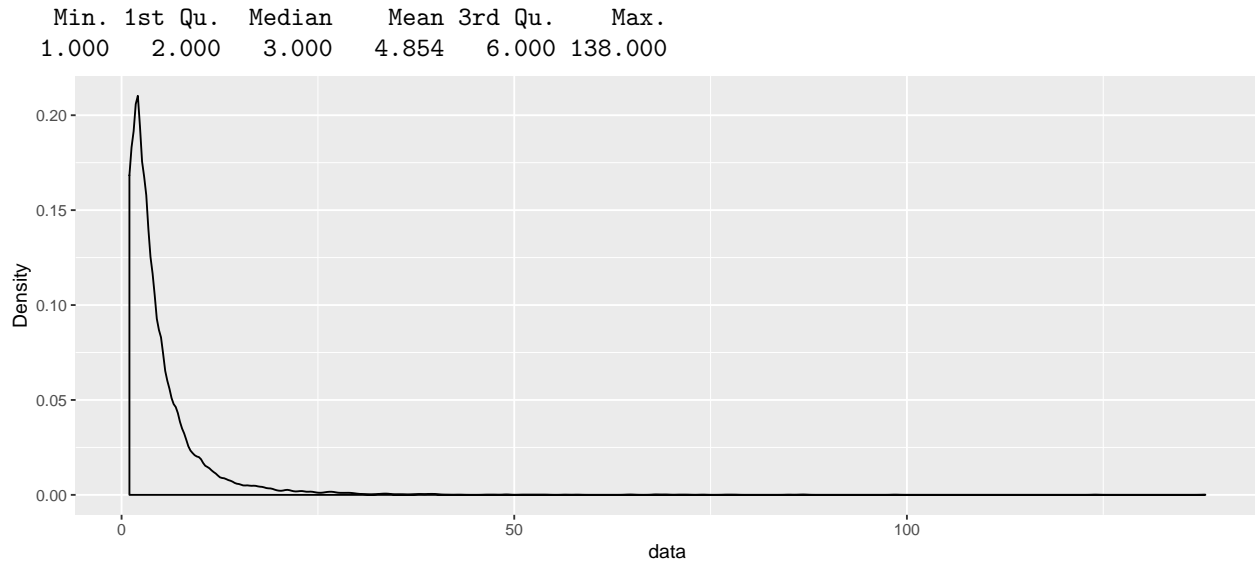
Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1.0	1.0	4.0	342.7	28.0	956642.0

Count distribution



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

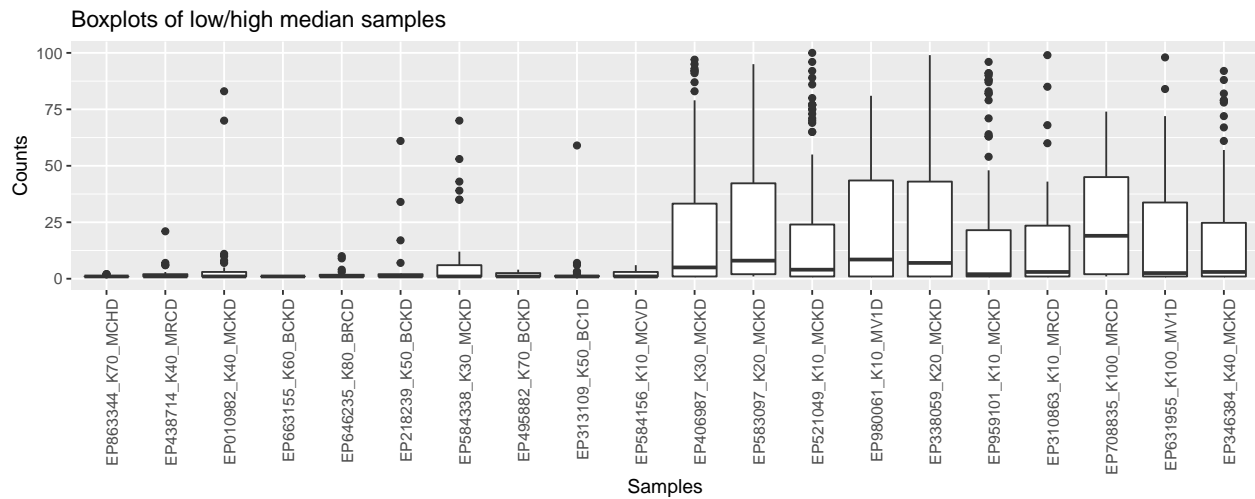
## Distribution of sample medians for all samples



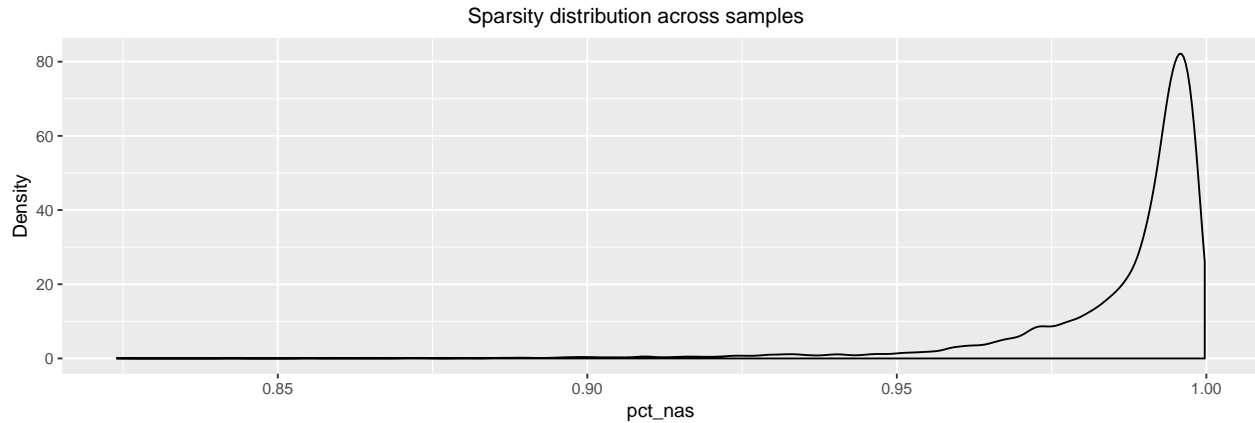
**Question:** The range of medians vary across samples. Most samples have median counts ~3-4, but some have as high as 138.

## 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:"

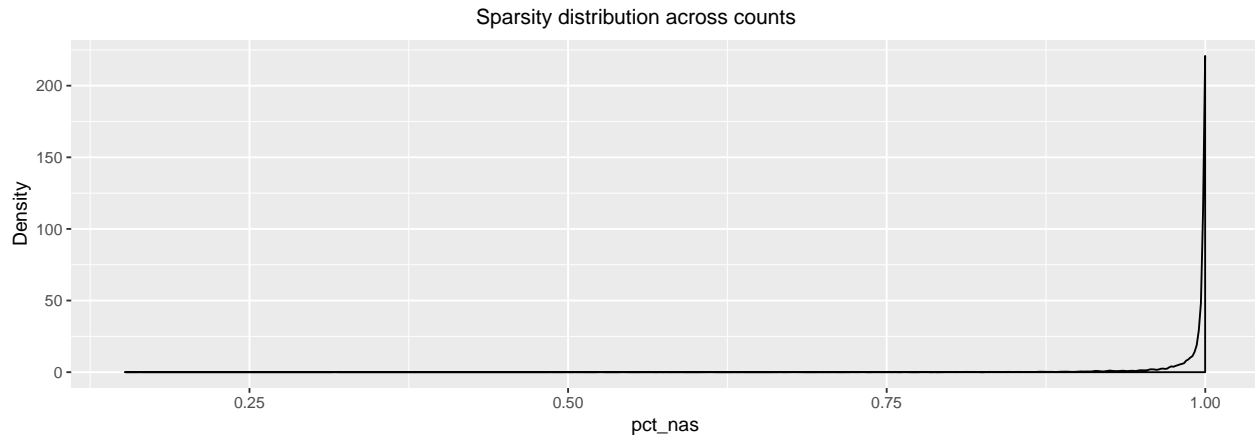
file	sample_body_site	visit_number
EP070043_K20_MRCD	rectum	2
EP488403_K30_MRCD	rectum	3
EP580086_K30_MRCD	rectum	3
EP631608_K10_MRCD	rectum	1
EP631608_K20_MRCD	rectum	2
EP646001_K10_MRCD	rectum	1
EP751500_K10_MRCD	rectum	1
EP751500_K20_MRCD	rectum	2
EP936464_K100_MRCD	rectum	10
EP949081_K40_MRCD	rectum	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
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 <b>Lactobacillus</b> ; s
332718	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g <b>Lactobacillus</b> ; s
134265	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Prevotellaceae; g <b>Prevotella</b> ; s
137183	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Bifidobacteriales; f__Bifidobacteriaceae; g <b>Gardnerella</b> ; s
469663	k__Bacteria; p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Atopobium; s
581782	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g; s
137580	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Lactobacillaceae; g <b>Lactobacillus</b> ; s
529233	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g <b>Streptococcus</b> ; s
580629	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g <b>Bacteroides</b> ; s
1109247	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Enterobacteriales; f__Enterobacteriaceae; g; s

[1] "Top 10 most sparse counts:"

ID	Taxonomy
982877	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Actinomycetaceae; g <b>Varibaculum</b> ; s
982363	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Corynebacteriaceae; g <b>Corynebacterium</b> ; s
956702	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g__Haemophilus; s
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 <b>Clostridiales</b> ; f[Tissierellaceae]; g <b>Anaerococcus</b> ; s
982266	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o <b>Alteromonadales</b> ; f[Chromatiaceae]; g <b>Rheinheimera</b> ; s
99517	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Porphyromonadaceae; g <b>Dysgonomonas</b> ; s
955102	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Actinomycetaceae; g <b>Actinomyces</b> ; s
953463	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae; g <b>Aggregatibacter</b> ; s
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

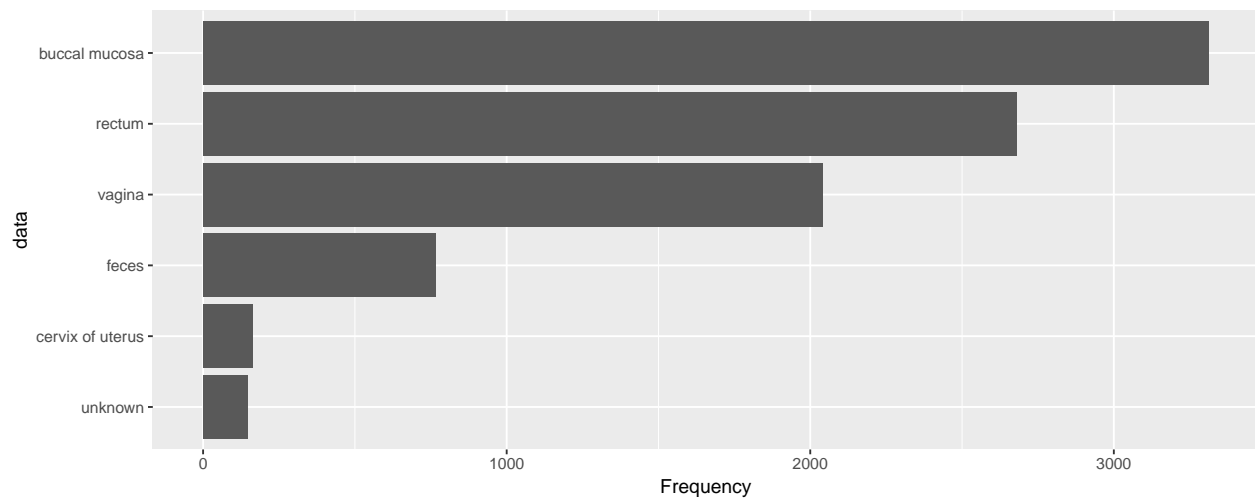
9107

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

Integrative Human Microbiome Project

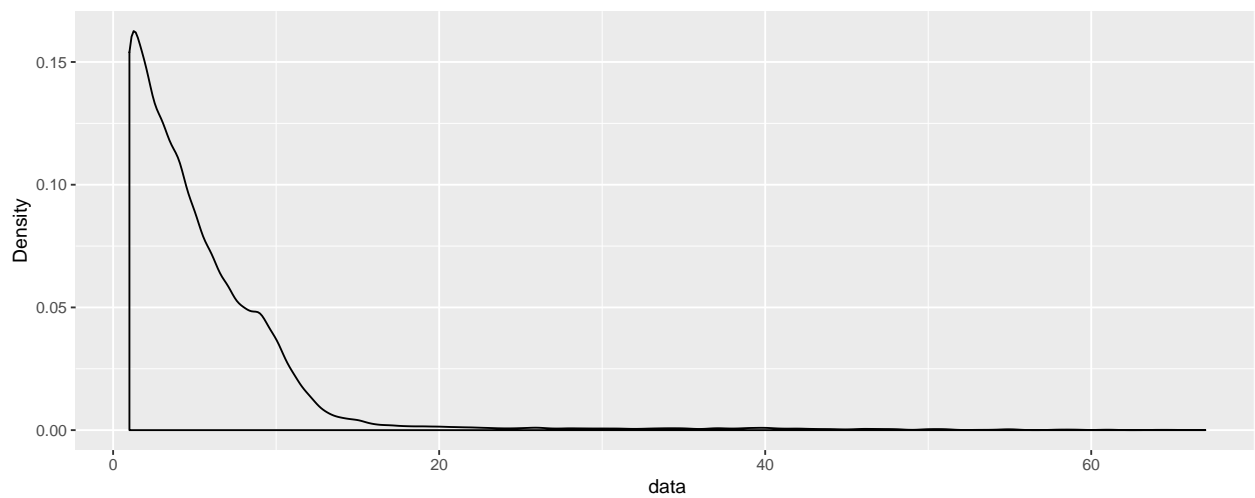
9107

## 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