Exploratory data analysis of cytokine files

2018-12-15

Settings

Load data

Data source: https://www.hmpdacc.org/hmp/, https://portal.hmpdacc.org/ - data portal. Samples/Studies: MOMS-PI, Files/Matrix Type: “host_cytokine” - selects 872 files. Download with `scripts/ascp-commands_biom_host_cytokine.sh`

There are NAs and values like “< OOR”. The latter are replaced by 0.

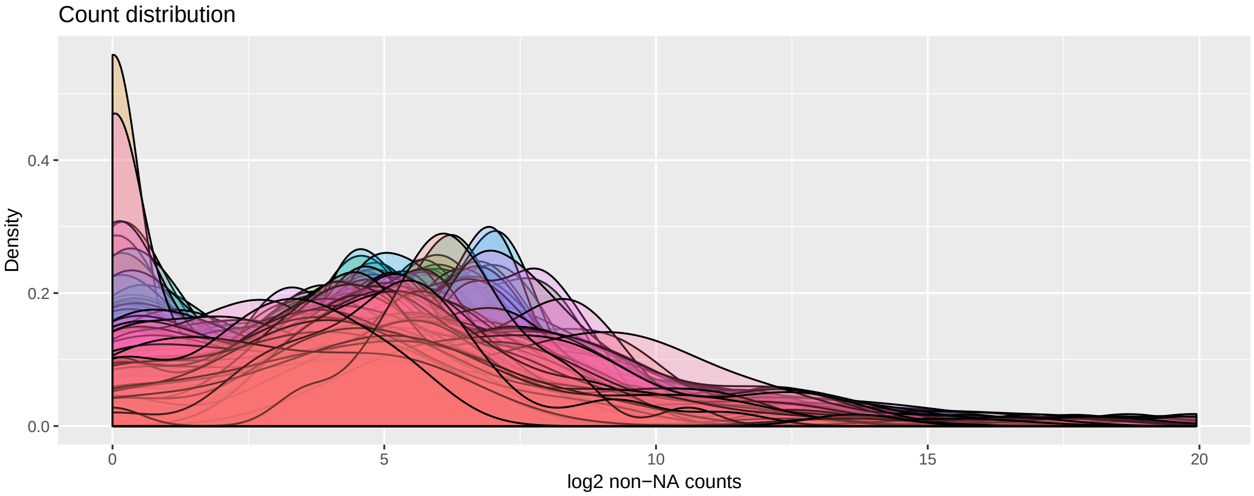
Cytokine EDA

Data dimensions: 29, 873

```
   Cytokine EP036702_K10_MVAX EP062329_K10_MP1P EP062329_K10_MVAX
1   Eotaxin             73.91             83.21             11.32
2 FGF_basic            307.07                NA                NA
3     G-CSF            430.87            324.73              0.00
4    GM-CSF           7615.00              0.00            505.35
5     IFN-g            481.79            274.65              0.00
  EP062329_K20_MVAX
1             36.42
2                NA
3            126.38
4            478.16
5            129.54
```

Count distribution for 500 randomly selected samples

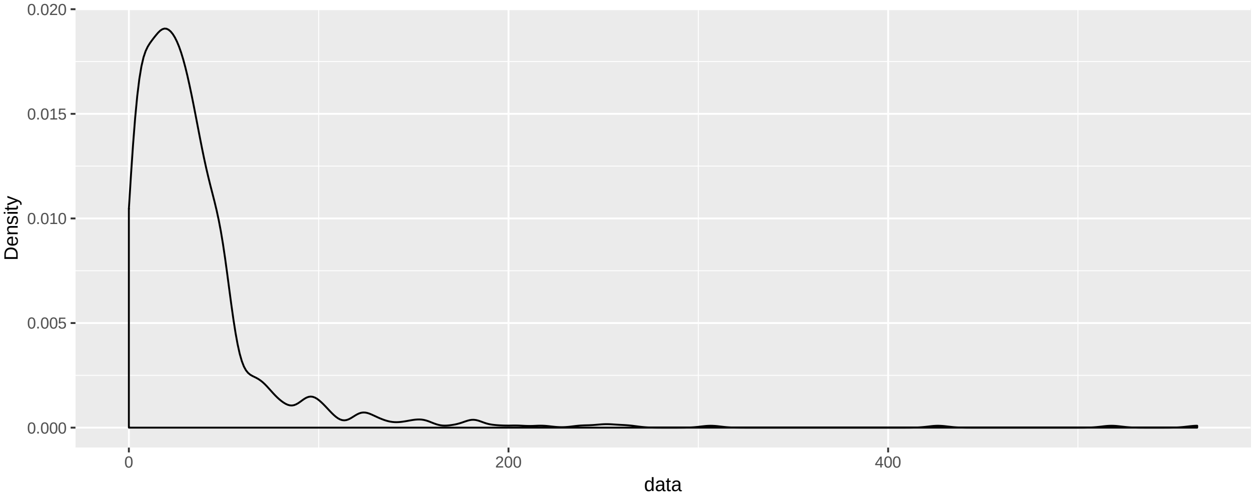
```
[1] "Summary of count data"

     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
      0.0       5.5      24.3    1363.5      83.1 1008760.0
```



Question: There are outlier counts. With median/mean counts equal to 24/1363, respectively, we have counts as high as 1008760.

Distribution of sample medians for all samples

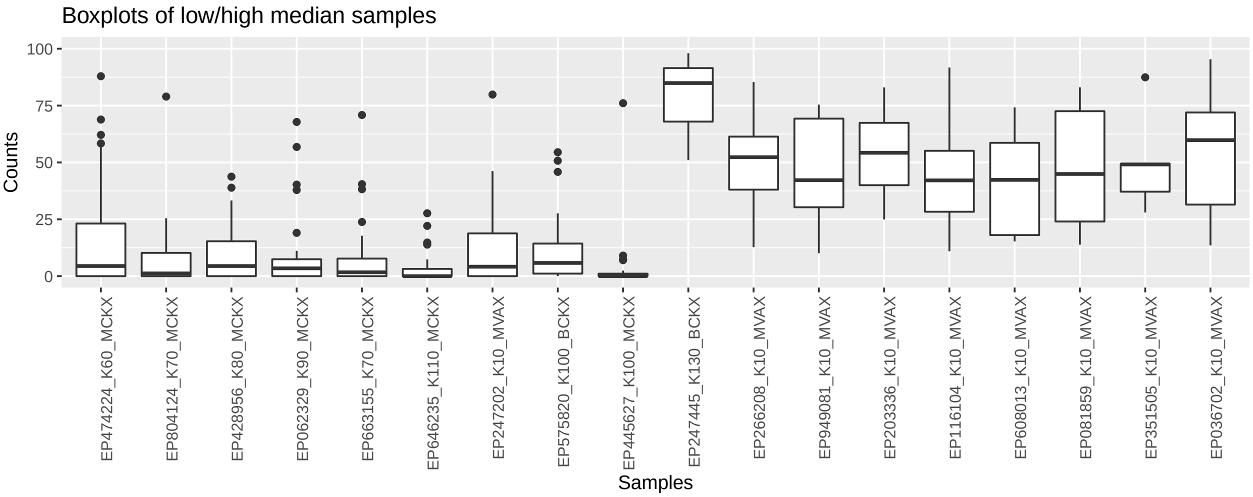
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   12.82   25.88   35.07   41.78  562.85
```



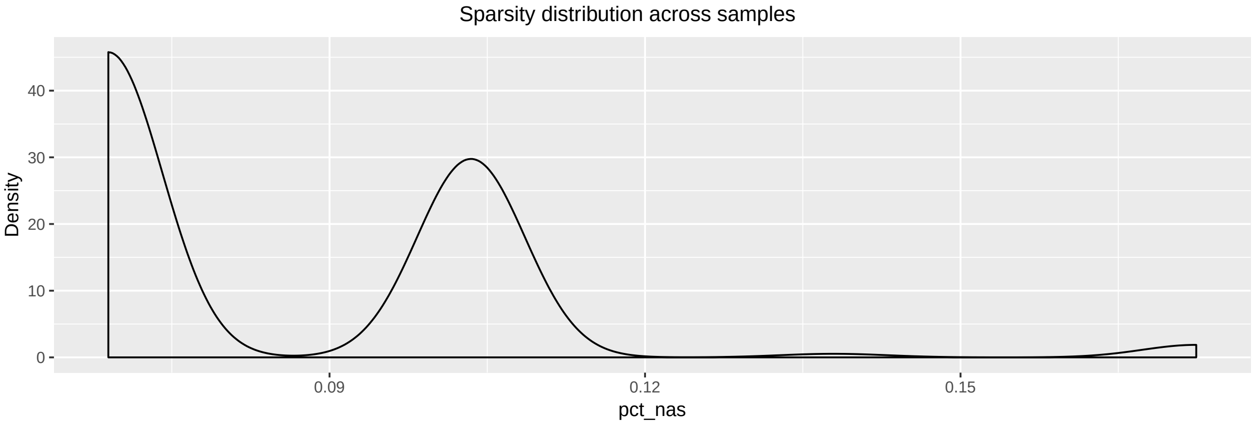
Question: The range of medians vary across samples. Most samples have median counts ~25, but some have as high as 563.

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. 
0.06897 0.06897 0.06897 0.08510 0.10345 0.17241 
```

Conclusion: Sparsity is present, but not bad

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

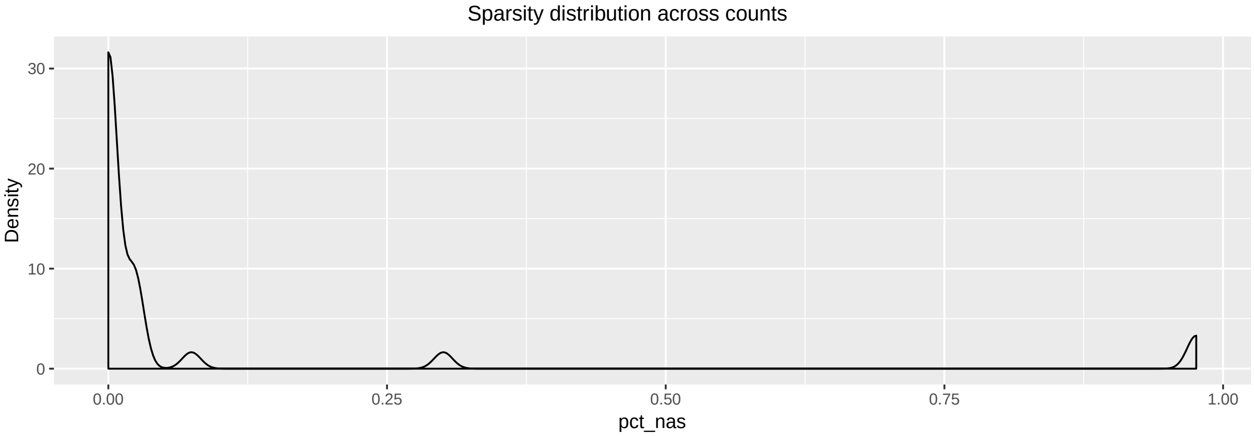
file	sample_body_site	visit_number
EP062329_K10_MVAX	vagina	1
EP062329_K30_MVAX	vagina	3
EP062329_K40_MVAX	vagina	4
EP062329_K50_MCKX	buccal mucosa	5
EP062329_K60_BCKX	buccal mucosa	6
EP062329_K80_BCKX	buccal mucosa	8
EP062329_K90_MCKX	buccal mucosa	9
EP065458_K100_BCKX	buccal mucosa	10
EP065458_K20_MVAX	vagina	2
EP065458_K80_BCKX	buccal mucosa	8

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

file	sample_body_site	visit_number
EP362253_K10_MVAX	vagina	1
EP505314_K10_MVAX	vagina	1
EP516855_K10_MVAX	vagina	1
EP523733_K10_MVAX	vagina	1
EP588271_K10_MVAX	vagina	1
EP608013_K10_MVAX	vagina	1
EP647247_K10_MVAX	vagina	1
EP663711_K10_MVAX	vagina	1
EP794231_K10_MVAX	vagina	1
EP936022_K10_MVAX	vagina	1
EP949081_K10_MVAX	vagina	1

Observation: Samples from buccal mucosa may be least sparse. Samples from vagina and first visits may be most sparse.

Sparsity of counts (across rows)



```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.00000 0.00000 0.08510 0.02408 0.97592 
```

Question: The data across counts is OK except one measure that is very sparse.

```
[1] "Top 5 least sparse counts:"
```

	x
Eotaxin	0
G-CSF	0
GM-CSF	0
IL-10	0
IL-12(p70)	0

```
[1] "Top 5 most sparse counts:"
```

	x
IL-2	0.0240826
IL-5	0.0240826
RANTES	0.0745413
IL-1ra	0.3004587
FGF__basic	0.9759174
IL-17	0.9759174

Observation: Counts for IL-5 and IL-2 cytokines are very sparse.

Metadata EDA

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

```
female 
   872 
```

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

```
unknown 
    872 
```

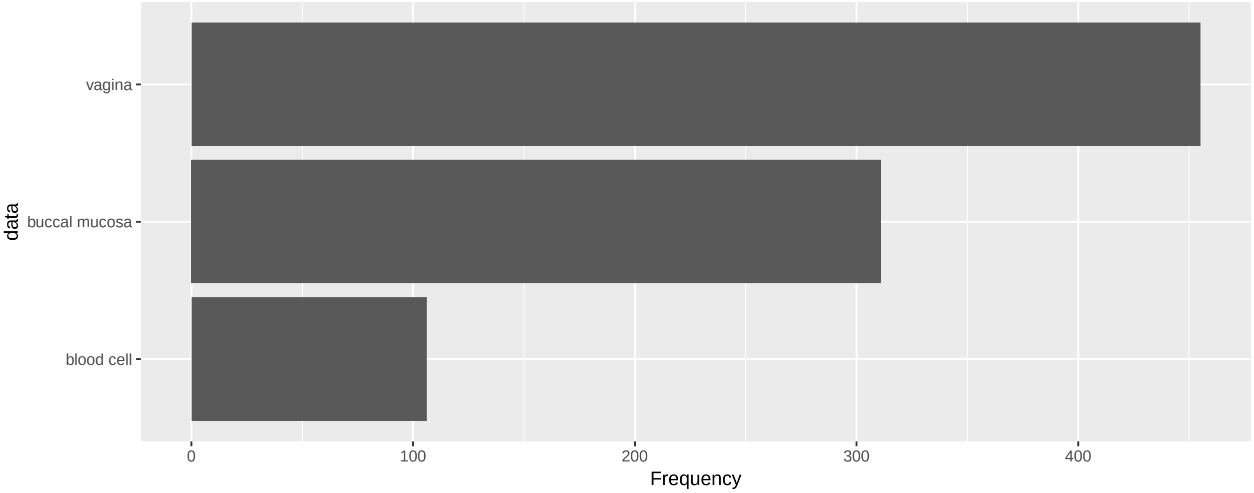
```r
table(mtx_metadata$study_full_name) %>% sort(., decreasing = TRUE)
```

```
momspi 
   872 
```

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

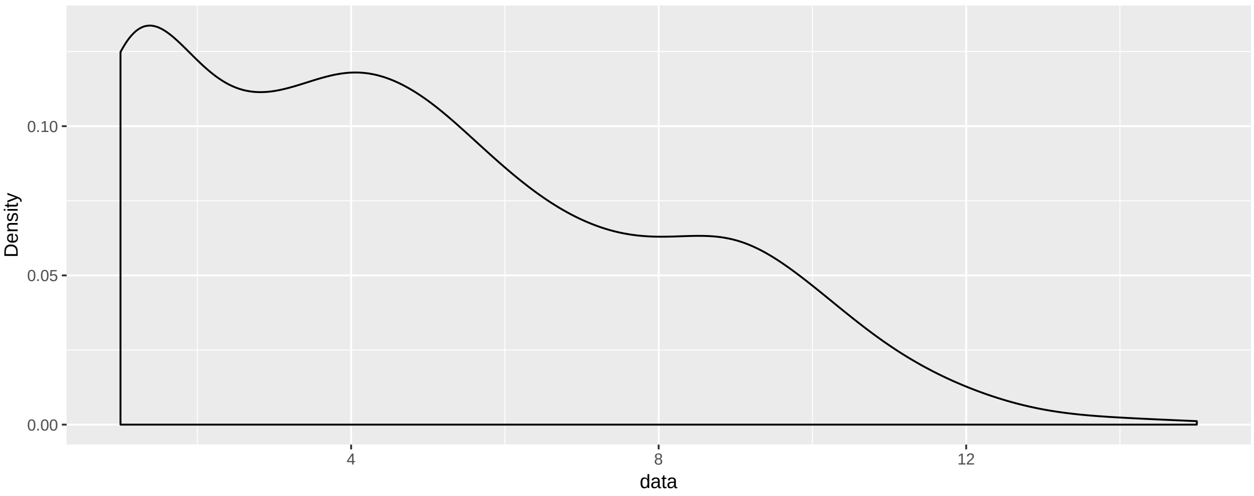
```
Integrative Human Microbiome Project 
                                 872 
```

sample__body__site



```
       vagina buccal mucosa    blood cell 
          455           311           106 
```

visit_number



```
  1   4   5   2   3   6   9   7   8  10  11  12  13  14  15
158 109  98  94  93  73  60  57  52  41  21  10   3   2   1
```

```
[1] "How many total samples: 872"
```

```
[1] "How many samples at visit 1: 158"
```

Number of subjects with vaginal samples

```
[1] 872   9
```

```
[1] 116
```

```
 1  3  4  5  6  7  8
26  8 32 29 16  4  1
```

Save the results