MOMS-PI

2018-11-20

Exploratory data analysis

Data source: http://vmc.vcu.edu/resources/momspi. Download with `scripts/download_moms-pi.sh`

MOMS-PI_16S_STIRRUPS_profiles_POP1.txt

```
[1] 4943    6
```

SampleID	Taxa	Threshold.Status	No_of_Reads	Percentage_Abundance	Avg._Score
EP288825_K10_MV1D	Lactobacillus_crispatus_cluster	AT	26762	39.82	99.52
EP288825_K10_MV1D	Lactobacillus_jensenii	AT	9715	14.45	98.71
EP288825_K10_MV1D	Prevotella_cluster2	AT	8448	12.57	99.58
EP288825_K10_MV1D	Prevotella_disiens	AT	3690	5.49	99.34
EP288825_K10_MV1D	Prevotella_bivia	AT	2123	3.159	99.52
EP288825_K10_MV1D	Streptococcus_anginosus	AT	1847	2.748	99.67

Questions

- What are the “Threshold.Status”, “No_of_Reads”, “Percentage_Abundance”, “Avg._Score" columns?

Manual EDA

```
[1] "How many unique samples?"
```

```
[1] 69
```

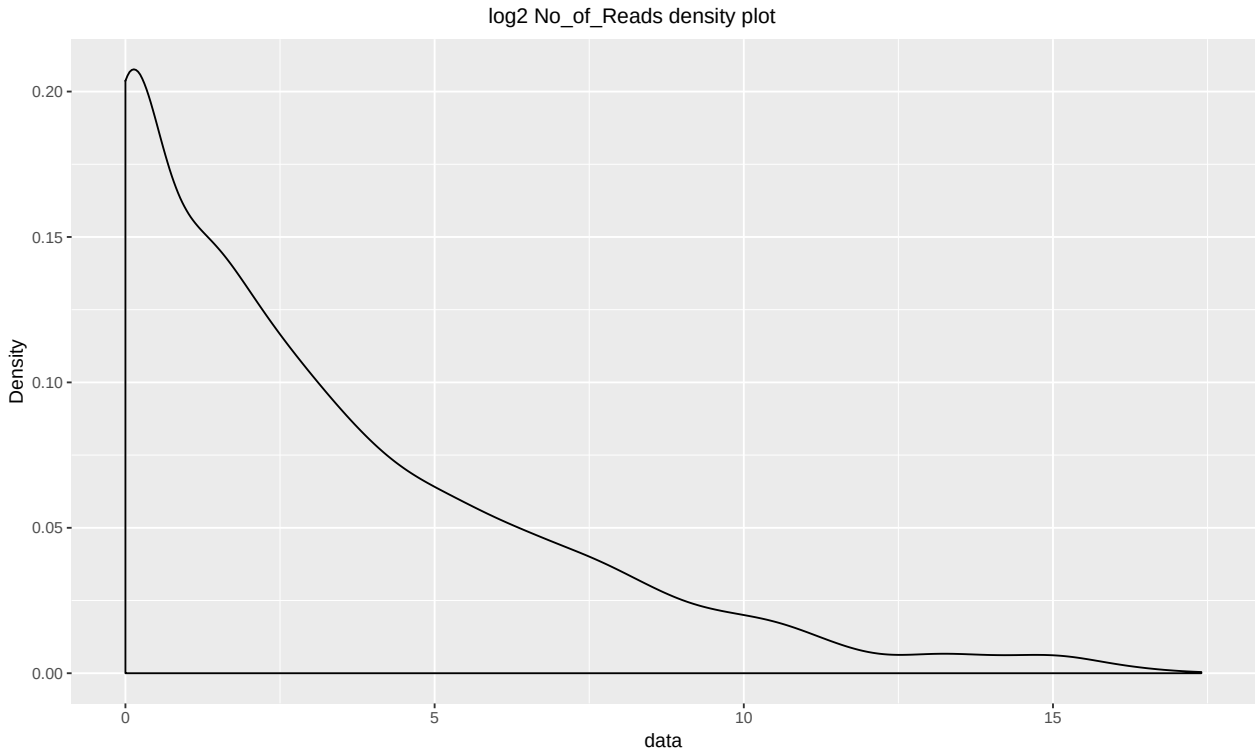
```
[1] "How many unique taxa?"
```

```
[1] 299
```

```
[1] "How many unique Threshold.Status?"
```

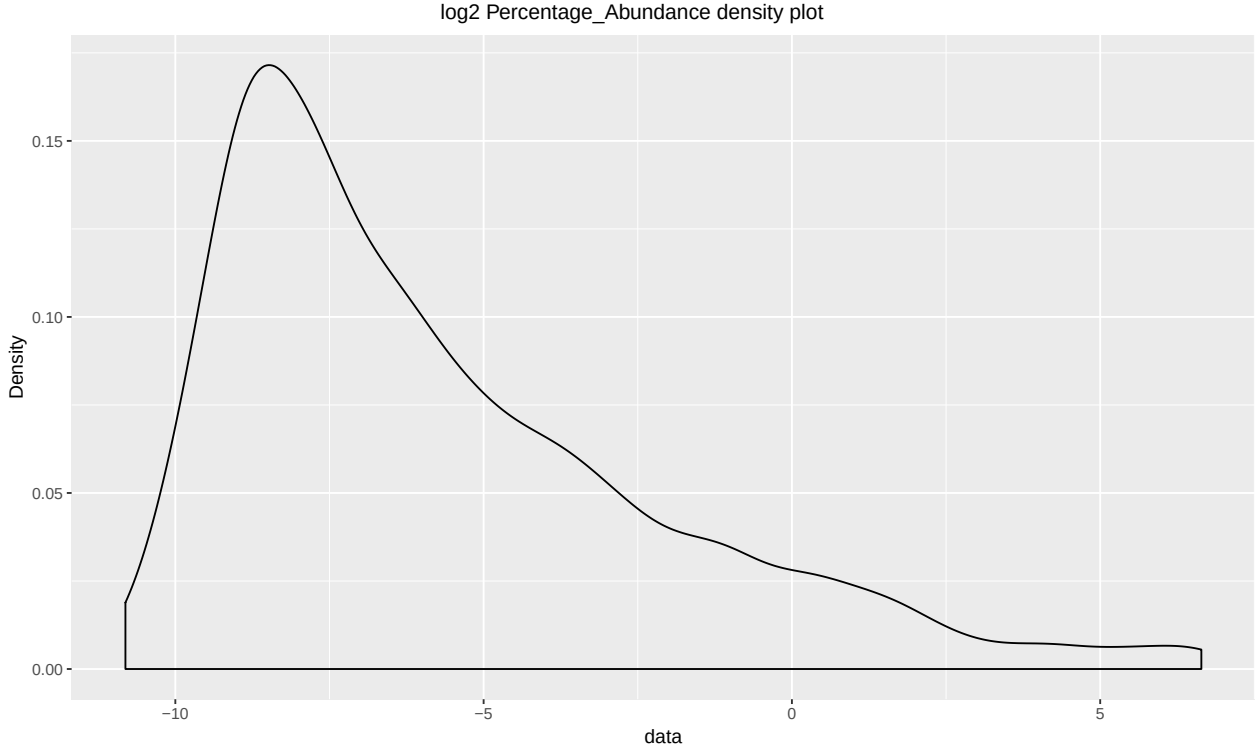
```
  AT   BT
2616 2327
```

```
[1] "What is the distribution of the No_of_Reads?"
```



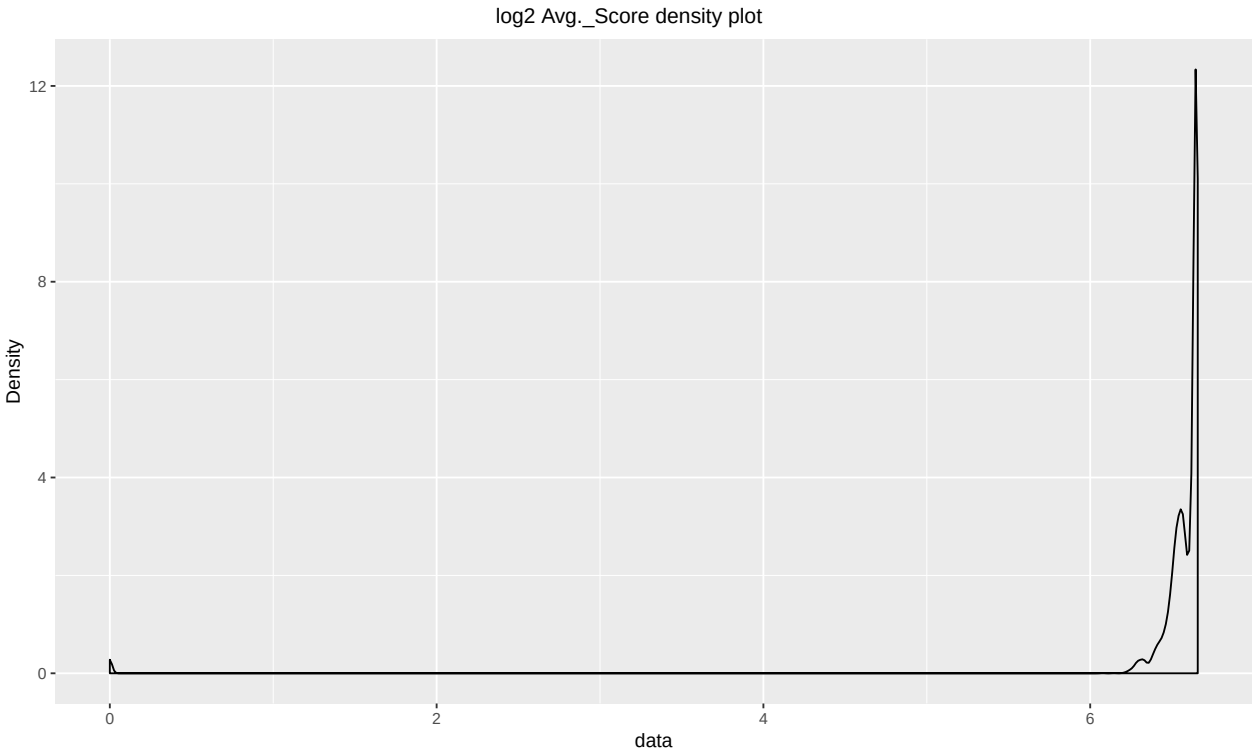

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      1       1       5     766      33  172858 
```

```
[1] "What is the distribution of the Percentage_Abundance?"
```




```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
 0.00056  0.00281  0.00956  1.39591  0.06871 99.75557 
```

```
[1] "What is the distribution of the Avg._Score?"
```




```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   91.97   98.05   94.28   99.40  100.00
```

Questions

- No_of_Reads, Percentage_Abundance, Avg._Score have outlies - is that normal?

Convert to wide format

```
[1]  69 300
```

SampleID	Acidaminococcus_fermentans	Acinetobacter_baumannii	Actinobaculum_massiliense	Actinobaculum_schaalii
EP036702_K10_MV1D	0	0	0	0
EP081373_K10_MV1D	0	0	0	0
EP081859_K10_MV1D	0	0	0	0
EP086560_K10_MV1D	0	0	0	0
EP087695_K10_MV1D	0	0	0.03774	0.1887
EP101628_K10_MV1D	0	0	0	0
EP116104_K10_MV1D	0	0	0	0
EP123585_K10_MV1D	0	0	0	0
EP154266_K10_MV1D	0	0	0	0
EP175456_K10_MV1D	0	0	0	0
EP203336_K10_MV1D	0	0.00388	0	0
EP222037_K10_MV1D	0	0.01756	0	0
EP223167_K10_MV1D	0	0	0	0
EP237361_K10_MV1D	0	0	0.01022	0
EP247202_K10_MV1D	0	0	0	0

Questions

- Many zeros (missing values?) - is that normal?

What are the most abundant strains?

```
            Lactobacillus_iners Lactobacillus_crispatus_cluster
                         1724.3418                       1408.9113
  Lactobacillus_gasseri_cluster          Gardnerella_vaginalis
                          639.6009                        619.6503
         Lachnospiraceae_BVAB1              Atopobium_vaginae
                          619.0304                        378.2886
```

MOMS-PI_CytokineProfiles_POP1.txt

```
[1] 1625    5
```

SampleID	Cytokine	Cytokine.Concentration.pg.ml.	Protein.Concentration.mg.ml.	Normalized.Protein.Concentration.pg.cytokine.mg.prote
EP642999_K10_MVAX	EotaxinMVAX	12.77	0.226	56.5
EP376326_K10_MVAX	EotaxinMVAX	10.18	0.365	27.89
EP718673_K10_MVAX	EotaxinMVAX	11.92	0.113	105.5
EP575820_K10_MVAX	EotaxinMVAX	7.39	0.168	43.99
EP036702_K10_MVAX	EotaxinMVAX	13.6	0.184	73.91
EP203336_K10_MVAX	EotaxinMVAX	8.19	0.073	112.2

MOMS-PI_LipidomicsProfiles_POP1.txt

```
[1] 4880    4
```

SampleID	LipidType	Lipid	Expression
EP438739_K10_MVAL	Eicosanoid	12-HETE	0.0
EP438739_K10_MVAL	Eicosanoid	6ketoPGF1alpha	0.145994065282
EP438739_K10_MVAL	Eicosanoid	8isoprostane	0.139534883721
EP438739_K10_MVAL	Eicosanoid	DHA	0.66976744186
EP438739_K10_MVAL	Eicosanoid	EPA	0.0918699186992
EP438739_K10_MVAL	Eicosanoid	PGD2	7.6397515528

MOMS-PI_subject_kit_sample_mapping_POP1.txt

```
[1] 203   7
```

SubjectID	KitID	KitType	SampleID	SampleType	BodySite	VisitNumber
EP203336	EP203336_K10	New OB	EP203336_K10_MV1DMV1D		Vagina	1
EP642999	EP642999_K10	New OB	EP642999_K10_MV1DMV1D		Vagina	1
EP154266	EP154266_K10	New OB	EP154266_K10_MV1DMV1D		Vagina	1
EP718673	EP718673_K10	New OB	EP718673_K10_MV1DMV1D		Vagina	1
EP222037	EP222037_K10	New OB	EP222037_K10_MV1DMV1D		Vagina	1
EP824738	EP824738_K10	New OB	EP824738_K10_MV1DMV1D		Vagina	1

Manual EDA

```
[1] "How many unique SubjectID?"
```

```
[1] 74
```

```
[1] "How many unique SampleType?"
```

```
MV1D MVAL MVAX
  69   69   65
```

```
[1] "How many BodySite?"
```

```
Vagina
   203
```

```
[1] "How many VisitNumber?"
```

```
  1   2
201   2
```

Questions

- 74 unique subjects, but should be 69?
- What is "mtx4$SampleType"?