lake-microbiome

Sophia Yang

Datasets

Folder_Name	Filename	Descritpion I	Descritpion II
Fw_ lake mendota data	coverm_431_MAGS_metagenomes_reads_count.csv	raw count data of the MAG(metagenomes-assembled-genomes of Bacteria and Archea) abundance	431 columns × 16 rows
	Samples-mendota.csv	Metadata for each sample(only 16)	10 columns × 16 rows
	MAG_taxonomy.tsv	Taxonomy for each of the MAG	431 rows × 8 columns
	MAG_abundance_matrix_rel_abund.tsv	Reletive adundance of MAG	431 columns not including the unmapped× 16 rows
	MAG_RPKM_normalized.tsv	MAG normalized matrix(divided each RPKM value to an internal standard they used during sequencing already)	431 columns × 16 rows
	phages_abundance_matrix_metagenomes.tsv	Abundance matrix of Bacteriophages	2246 columns not including the unmapped × 16 rows
	phages_rpkm_normalized_matrix.tsv	Bacteriophages normalized matrix(divided each RPKM value to an internal standard they used during sequencing already)	2246 columns not including the unmapped × 16 rows
2020 Profile Data Lake Mendota(Environmental)	EnvironmentalData2020-Mendota.xlsx	The description of Environmental Data Column Names	2 columns × 12 rows
	ME_profile_071620.csv	Environmental information that has been measure on 2020/07/16, including different depths, water temperature and sample time etc.	12 columns × 21 rows
	ME_profile_072420.csv with the other 11 environmental sample datasets….	…… the same info as the previous cell, but has an extra sample info at depth 23.5.	12 columns × 22 rows

Row sum...(is not equal to1)

- Fw_ lake mendota data
  - MAG_abundance_matrix_rel_abund.tsv

  - MAG_RPKM_normalized.tsv

  - phages_abundance_matrix_metagenomes.tsv

  - phages_rpkm_normalized_matrix.tsv

```
1  row_sums_MAG_abundance= MAG_abundance.sum(axis=1)
2  row_sums_MAG_abundance
```

```
2020-07-24_15m       51.388456
2020-07-24_23.5m     41.211519
2020-07-24_5m        33.847230
2020-08-05_10m       55.278206
2020-08-05_15m       54.719929
2020-08-05_23.5m     39.912086
2020-08-25_10m       45.643490
2020-08-25_15m       53.954127
2020-08-25_23.5m     43.067711
2020-09-11_15m       51.559964
2020-09-11_23.5m     43.448078
2020-10-08_15m       38.991339
2020-10-08_23.5m     39.313672
2020-10-19_15m       29.948284
2020-10-19_23.5m     38.422739
2020-10-19_5m        30.432611
dtype: float64
```

```
1  row_sums_MAG_RPKM_normalized= MAG_RPKM_normalized.sum(axis=1)
2  row_sums_MAG_RPKM_normalized
```

```
2020-07-24_5m        0.000166
2020-07-24_15m       0.000125
2020-07-24_23.5m     0.000204
2020-08-05_10m       0.000183
2020-08-05_15m       0.000118
2020-08-05_23.5m     0.000178
2020-08-25_10m       0.000955
2020-08-25_15m       0.000080
2020-08-25_23.5m     0.000108
2020-09-11_15m       0.000110
2020-09-11_23.5m     0.000094
2020-10-08_15m       0.000608
2020-10-08_23.5m     0.000082
2020-10-19_5m        0.000977
2020-10-19_15m       0.001004
2020-10-19_23.5m     0.000793
```

```
1  row_sums_phages_abundance_matrix_metagenomes= phages_abundance_matrix_metagenomes.sum(axis=1)
2  row_sums_phages_abundance_matrix_metagenomes
```

```
2020-10-19_23.5m     0.655490
2020-10-08_15m       0.653947
2020-10-19_15m       0.777846
2020-09-11_15m       0.504348
2020-08-05_15m       0.577252
2020-08-25_10m       0.468715
2020-07-24_5m        2.384052
2020-07-24_23.5m     2.011944
2020-07-24_15m       0.536729
2020-08-25_15m       0.560599
2020-10-08_23.5m     0.709240
2020-10-19_5m        0.826225
2020-09-11_23.5m     0.932316
2020-08-25_23.5m     1.253326
2020-08-05_23.5m     1.858483
2020-08-05_10m       0.525313
```

```
1  row_sums_phages_rpkm_normalized_matrix= phages_rpkm_normalized_matrix.sum(axis=1)
2  row_sums_phages_rpkm_normalized_matrix
```

```
2020-07-24_15m       0.016504
2020-07-24_23.5m     0.040392
2020-07-24_5m        0.009160
2020-08-05_10m       0.024225
2020-08-05_15m       0.022192
2020-08-05_23.5m     0.057709
2020-08-25_10m       0.106863
2020-08-25_15m       0.015531
2020-08-25_23.5m     0.026158
2020-09-11_15m       0.021904
2020-09-11_23.5m     0.021430
2020-10-08_15m       0.030538
2020-10-08_23.5m     0.015622
2020-10-19_15m       0.057807
2020-10-19_23.5m     0.047207
2020-10-19_5m        0.051827
```

Q: Row sum does not sum to one??? Even for the normalized datasets
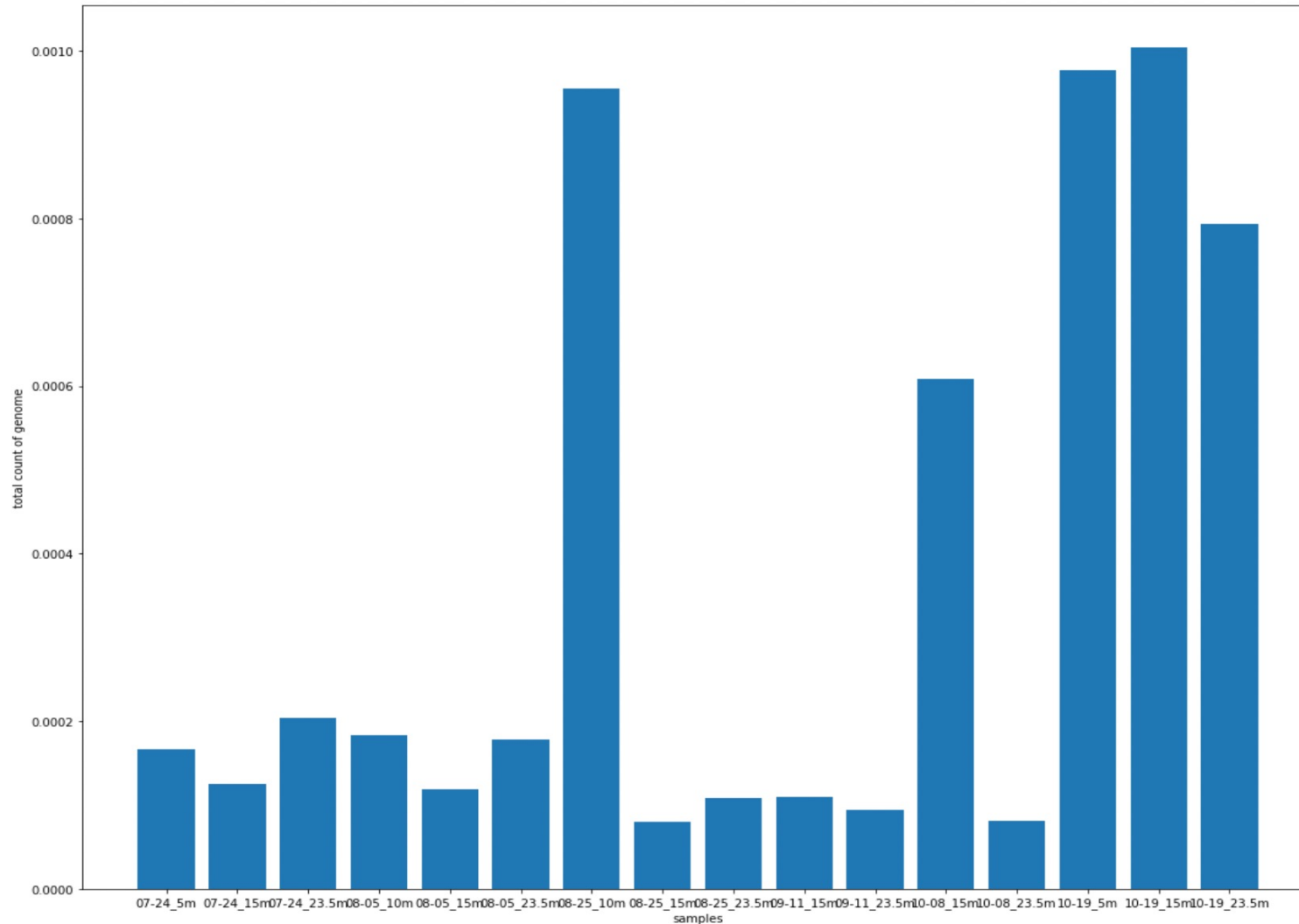**A: since different normalization method is being applied, this might be the reason that the row sum is not equal to 1, and some of the datasets have a vast majority of zeros.**

Let's take a look at of the row sum of each of the datasets(they do not sum up to 1...)
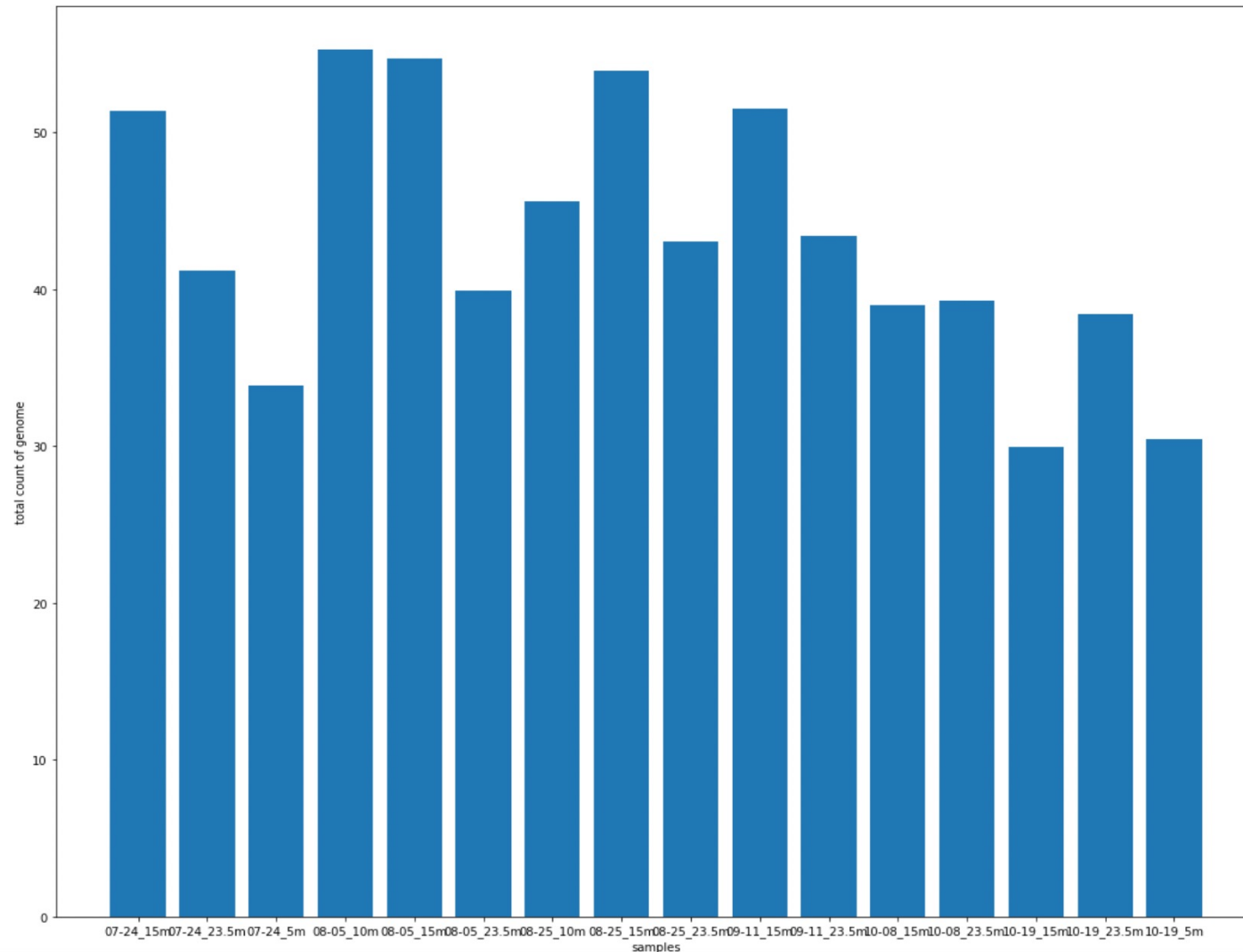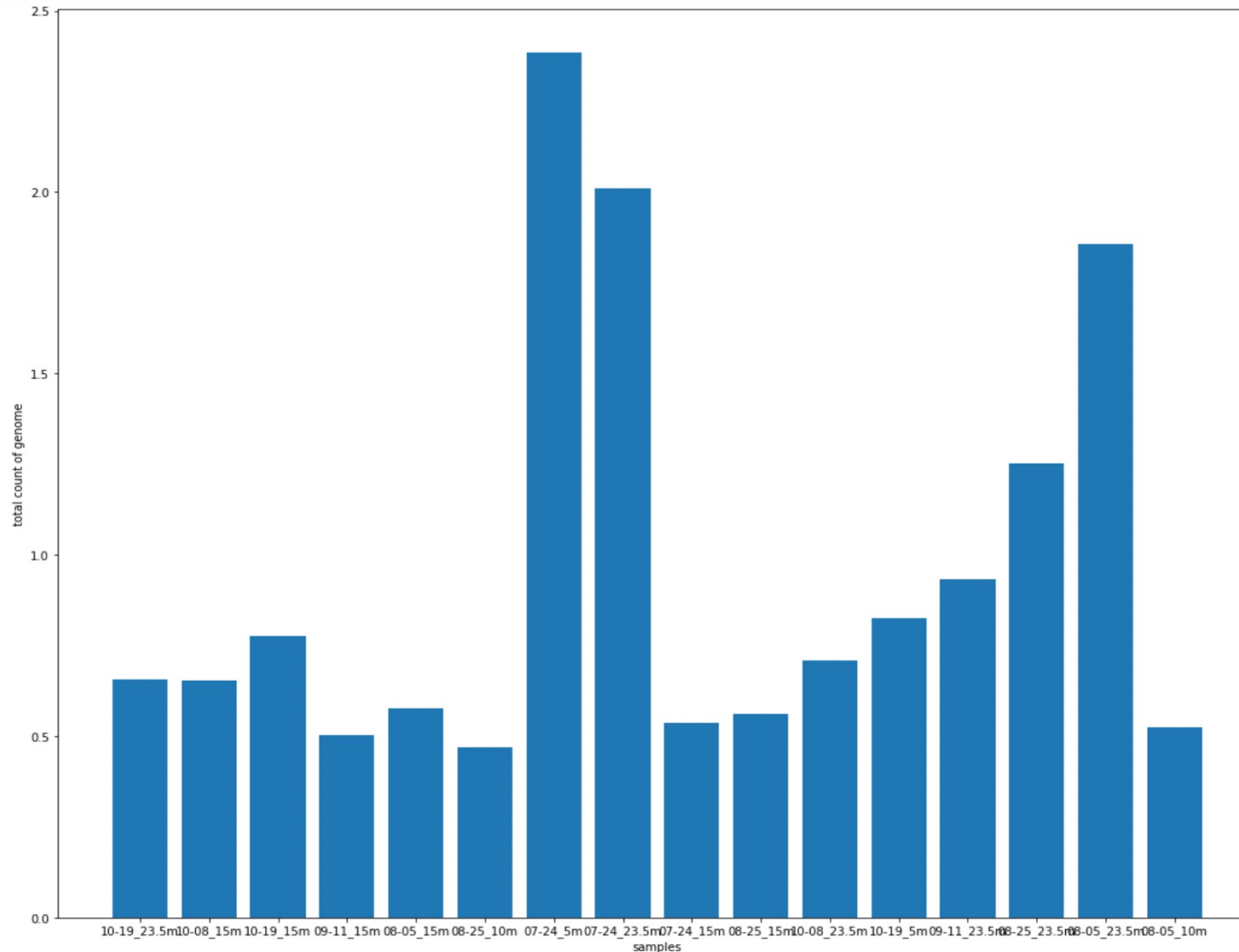

MAG_abundance_matrix_rel_abund.tsv

Let's take a look at of the row sum of each of the datasets(they do not sum up to 1...)
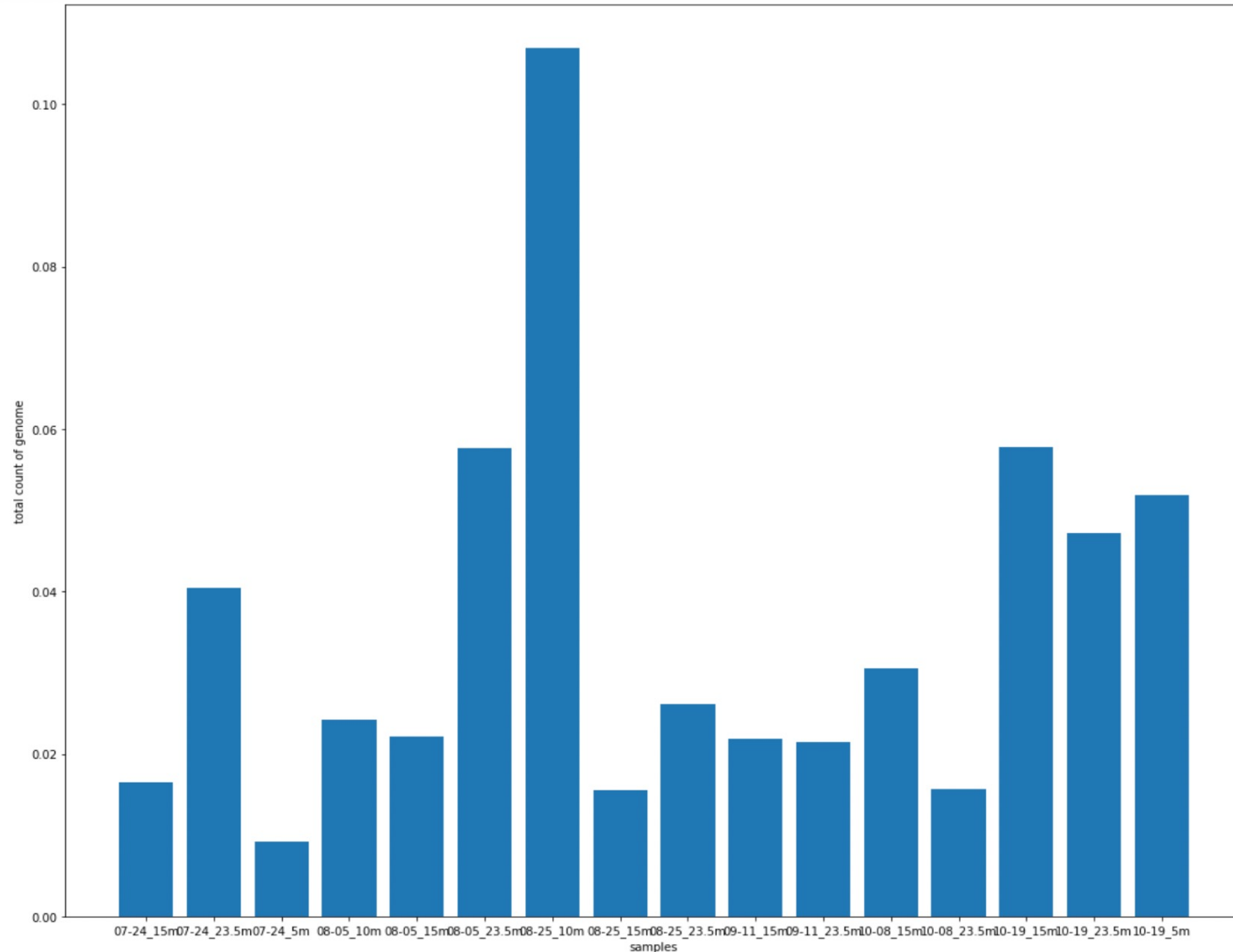

MAG_RPKM_normalized.tsv

Let's take a look at of the row sum of each of the datasets(they do not sum up to 1...)

phages_abundance_matrix_metagenomes.tsv

Let's take a look at of the row sum of each of the datasets(they do not sum up to 1...)

phages_rpkm_normalized_matrix.tsv

of zeros

- No zeros for the count dataset

- 1384 zeros for *MAG_abundance_matrix_rel_abund.tsv*

  ```
  the number of zeros: 1384
  Percentage of zeros: 20.07%
  ```

- 5893 zeros for *MAG_RPKM_normalized.tsv*

  ```
  the number of zeros: 5893
  Percentage of zeros: 85.46%
  ```

- 13125 zeros for *phages_abundance_matrix_metagenomes.tsv*

  ```
  the number of zeros: 13125
  Percentage of zeros: 36.52%
  ```

- 34983 zeros for phages_rpkm_normalized_matrix.tsv

  ```
  the number of zeros: 34983
  Percentage of zeros: 97.35%
  ```
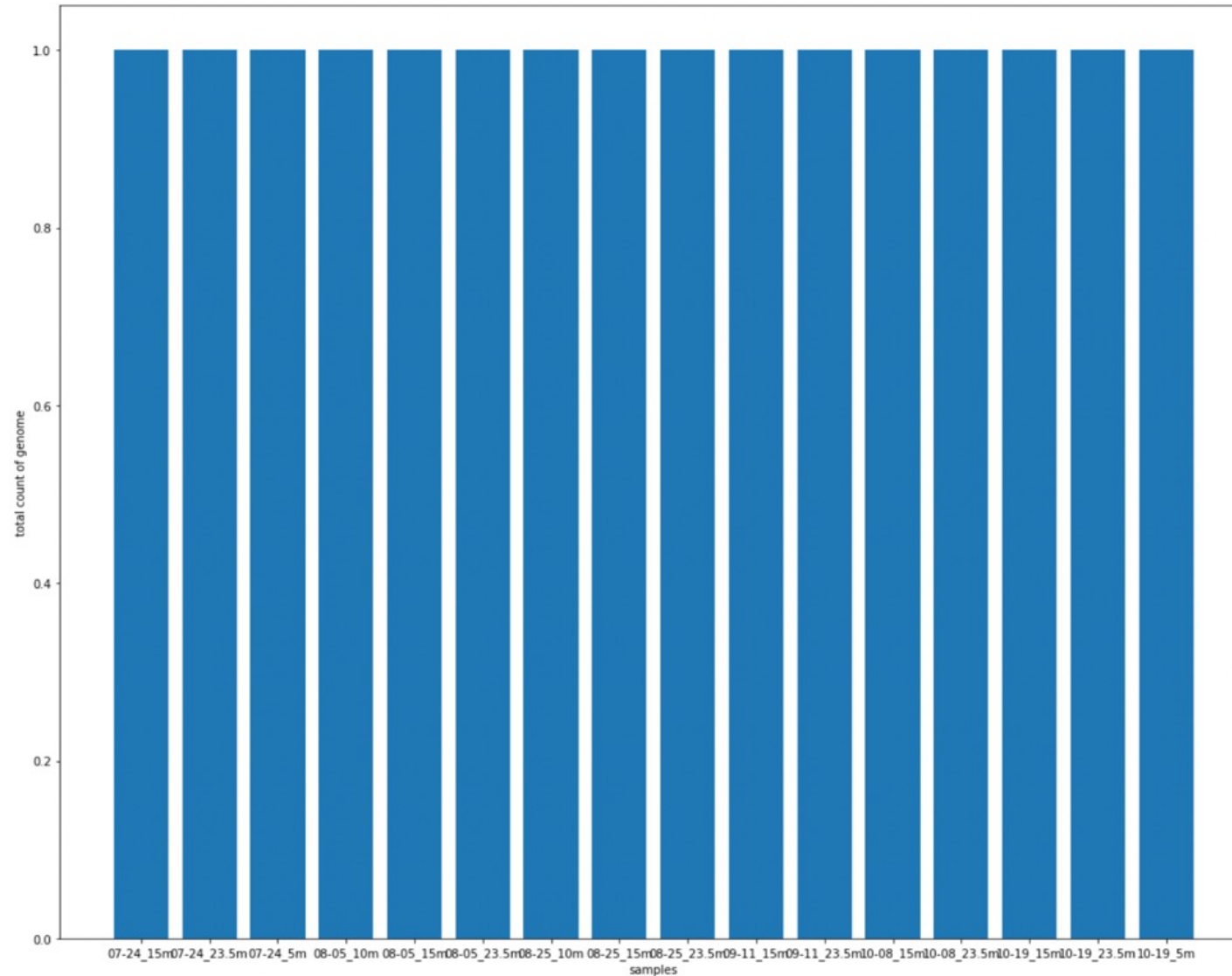
Take a look at of the count dataset…
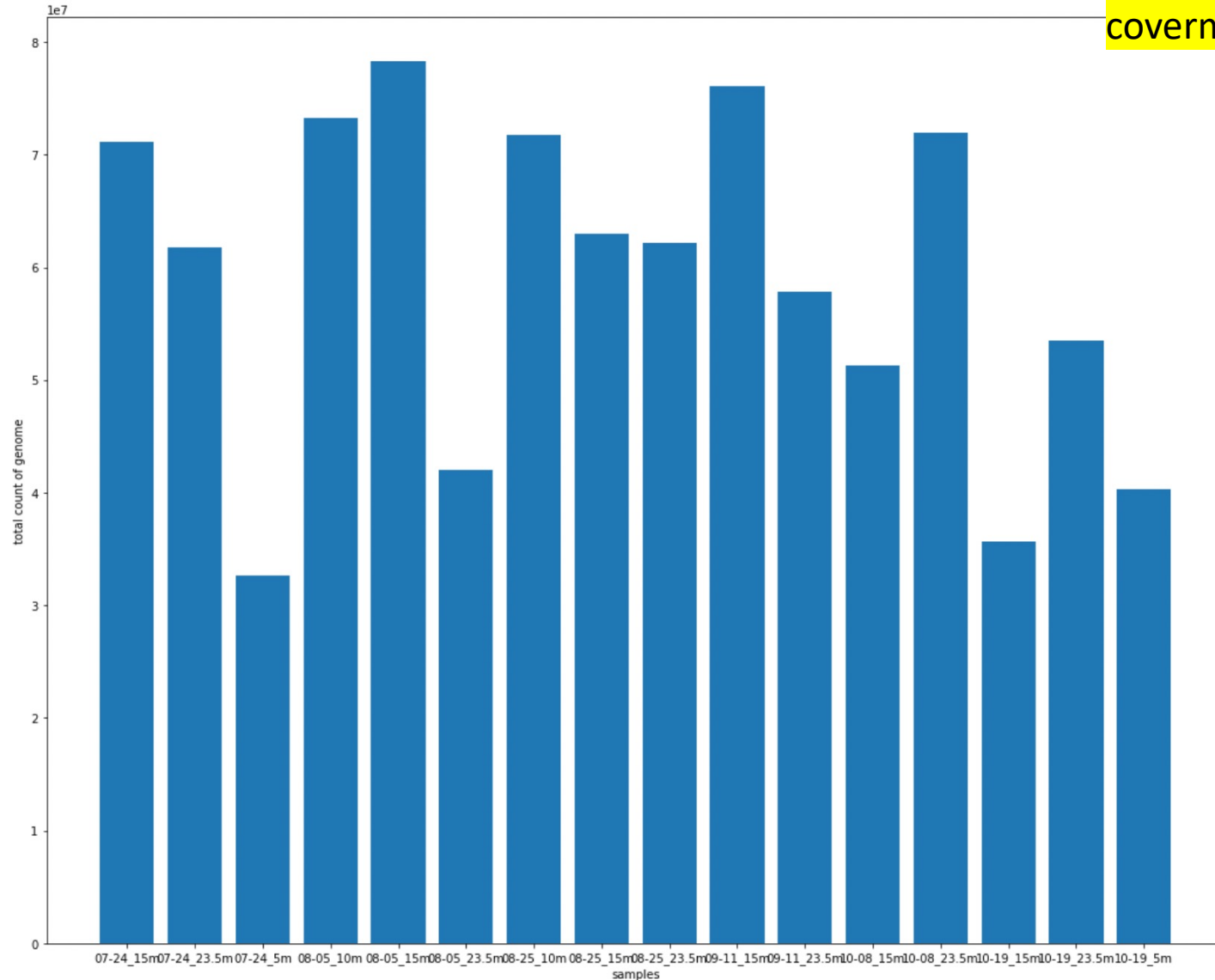
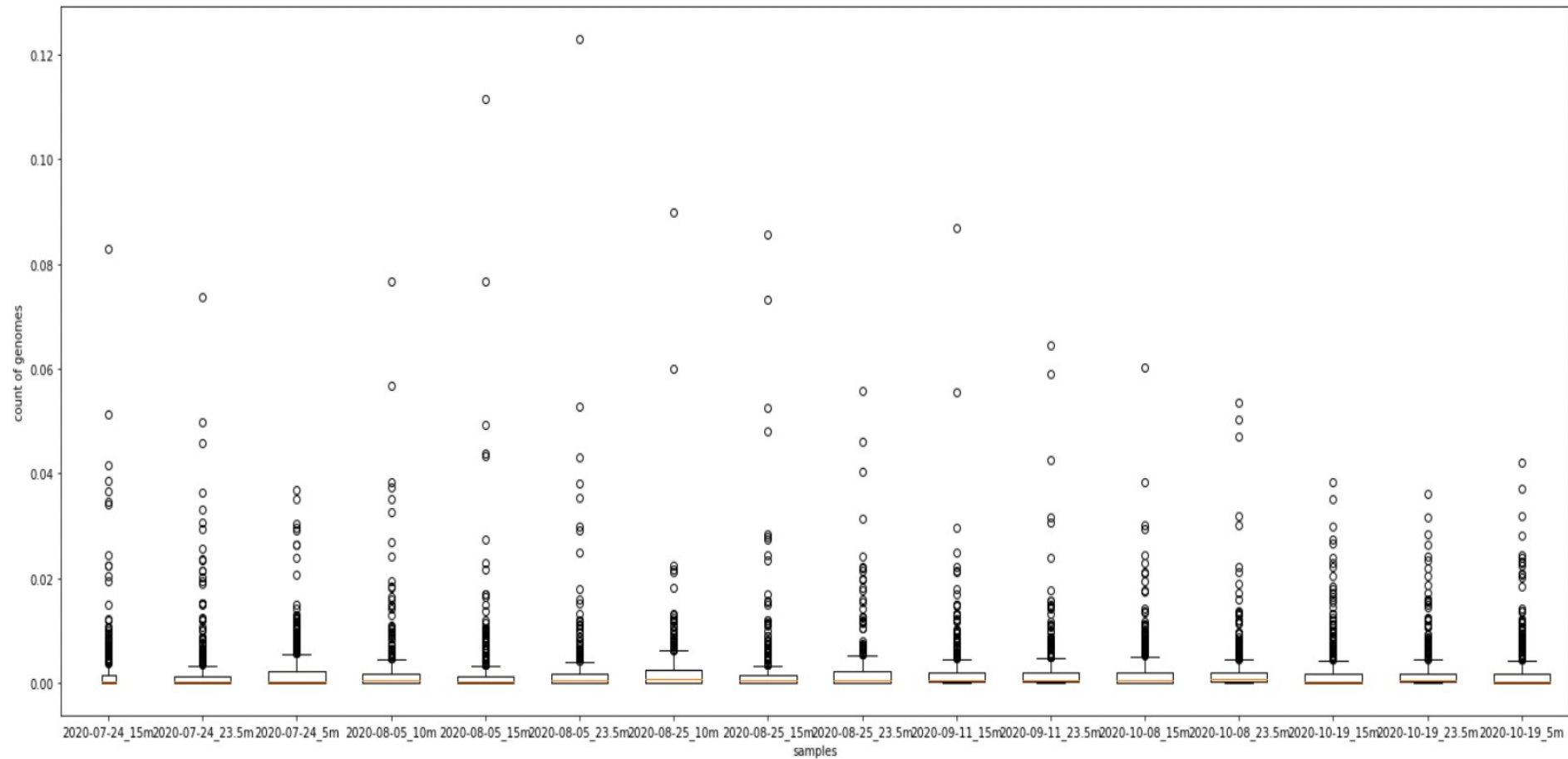Total count of genome per sample
Before normalization

Shallow depth less genome

Take a look at of the count dataset...

This is the count for all the genome per sample after normalization (sanity check, should sum to 1)
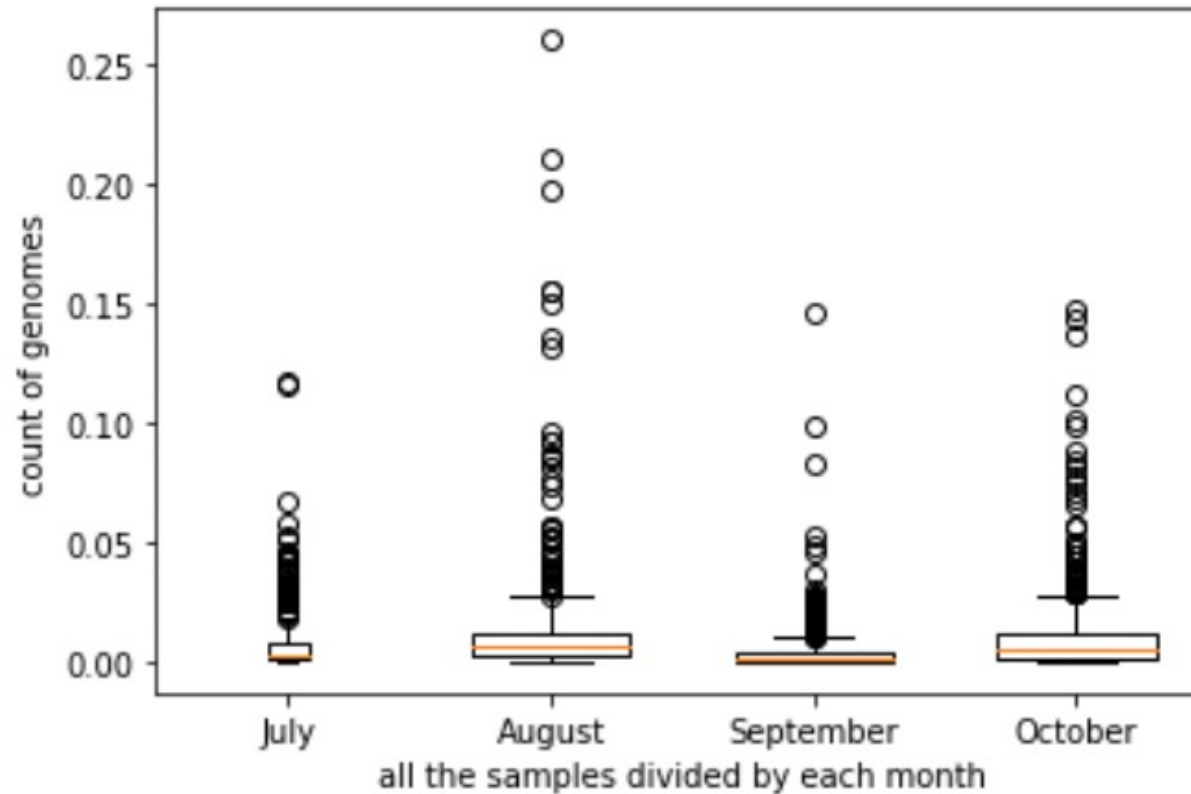
Take a look at of the count dataset...

Boxplot for dictribution of count of genome per sample

More outliers in the middle seasons

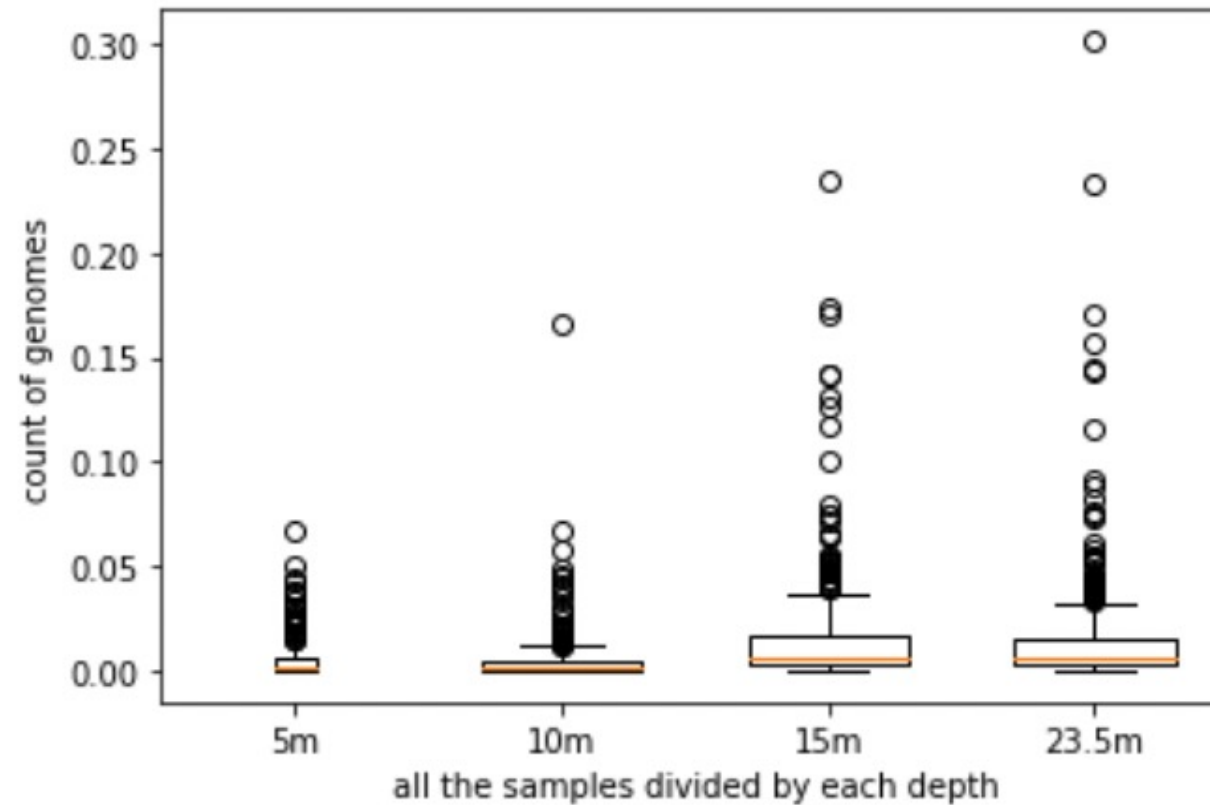# Continued(count dataset)...

Sum of count of each genome categorized by each month

Figures are generated based on normalized data set (row sum method)

August has more outliers than all the other months. However more samples in August…. Outlier might be expected.

Continued(count dataset)…

Sum of count of each genome categorized by each depth
Figures are generated based on normalized data set (row sum method)

As the depth became deeper, there are more outliers. However, since we have less sample data on depth 5 and 10m(2 for each), less count is expected.
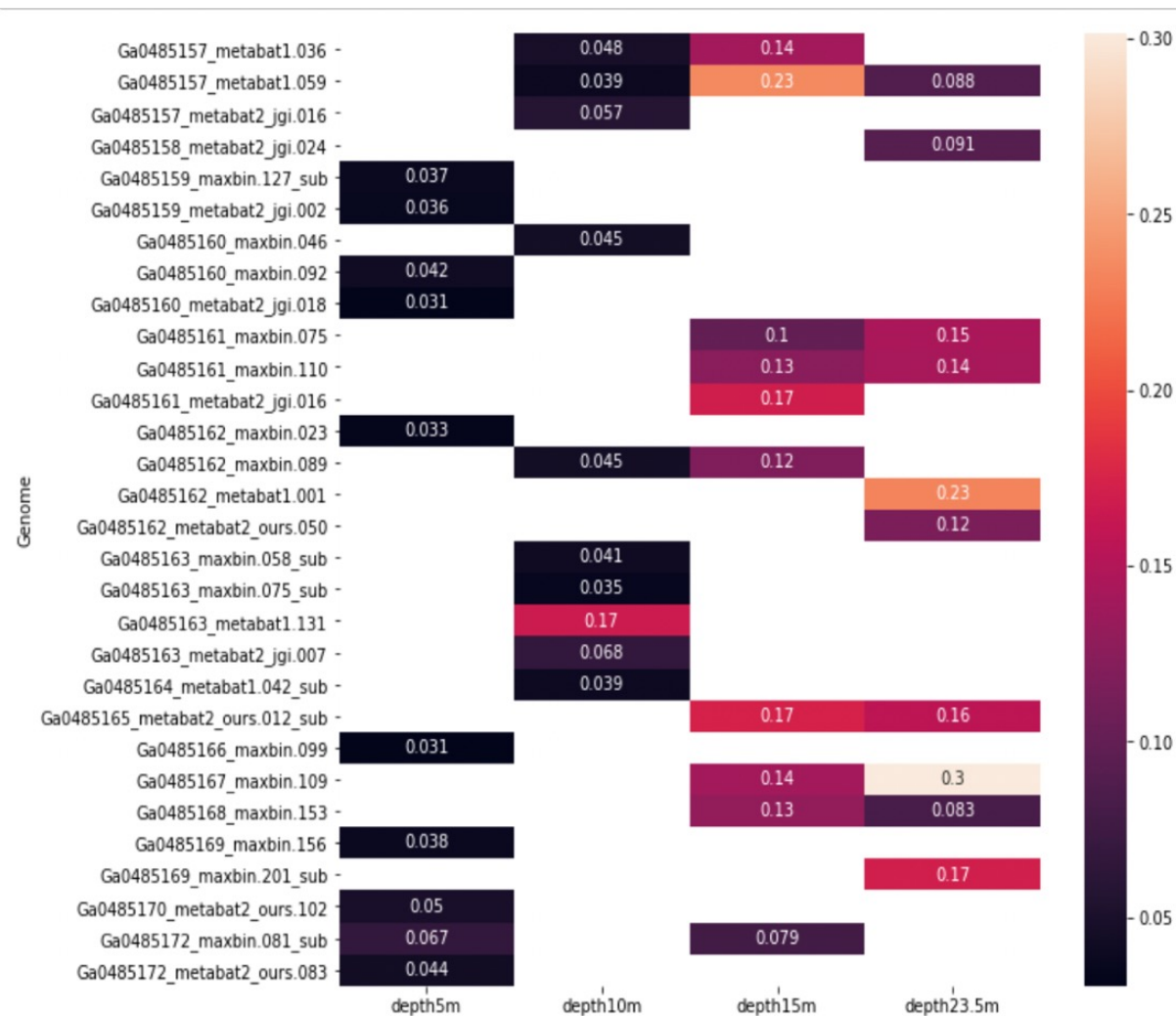
Still the count datset....

{'Ga0485157_metabat1.059': 3,
 'Ga0485157_metabat1.036': 2,
 'Ga0485162_maxbin.089': 2,
 'Ga0485165_metabat2_ours.012_sub': 2,
 'Ga0485167_maxbin.109': 2,
 'Ga0485161_maxbin.110': 2,
 'Ga0485168_maxbin.153': 2,
 'Ga0485161_maxbin.075': 2,
 'Ga0485172_maxbin.081_sub': 1,
 'Ga0485172_metabat2_ours.083': 1,
 'Ga0485170_metabat2_ours.102': 1,
 'Ga0485160_maxbin.092': 1,
 'Ga0485169_maxbin.156': 1,
 'Ga0485162_maxbin.023': 1,
 'Ga0485159_metabat2_jgi.002': 1,
 'Ga0485159_maxbin.127_sub': 1,
 'Ga0485166_maxbin.099': 1,
 'Ga0485172_maxbin.051_sub': 1,
 'Ga0485163_metabat1.131': 1,
 'Ga0485163_metabat2_jgi.007': 1,
 'Ga0485157_metabat2_jgi.016': 1,
 'Ga0485160_maxbin.046': 1,
 'Ga0485163_maxbin.058_sub': 1,
 'Ga0485164_metabat1.042_sub': 1,
 'Ga0485163_maxbin.075_sub': 1,
 'Ga0485161_metabat2_jgi.016': 1,
 'Ga0485167_maxbin.132': 1,
 'Ga0485162_metabat1.001': 1,
 'Ga0485169_maxbin.201_sub': 1,
 'Ga0485162_metabat2_ours.050': 1,
 'Ga0485158_metabat2_jgi.024': 1}

Here is a dictionary that is the count for how many times that the genomes appear in the top10(normalization)

This time I am only focsuing on the most popular(count) top ten genomes that have appeared in all the depths.

Ga0485157_metabat1.059 exists three times across all the depths(3/4).

Count for depths(top10)

plotting heatmap for all of the genome vs. depth. However, we are only focusing on the top 10 this time.... this would actually be all of the top10 genomes that appears in all of the depths. Generally speaking, shallower depths have less percentage, which means less count of the genome.

Count for depths(top10)--Binary Map

Binary Map of genome vs. depth
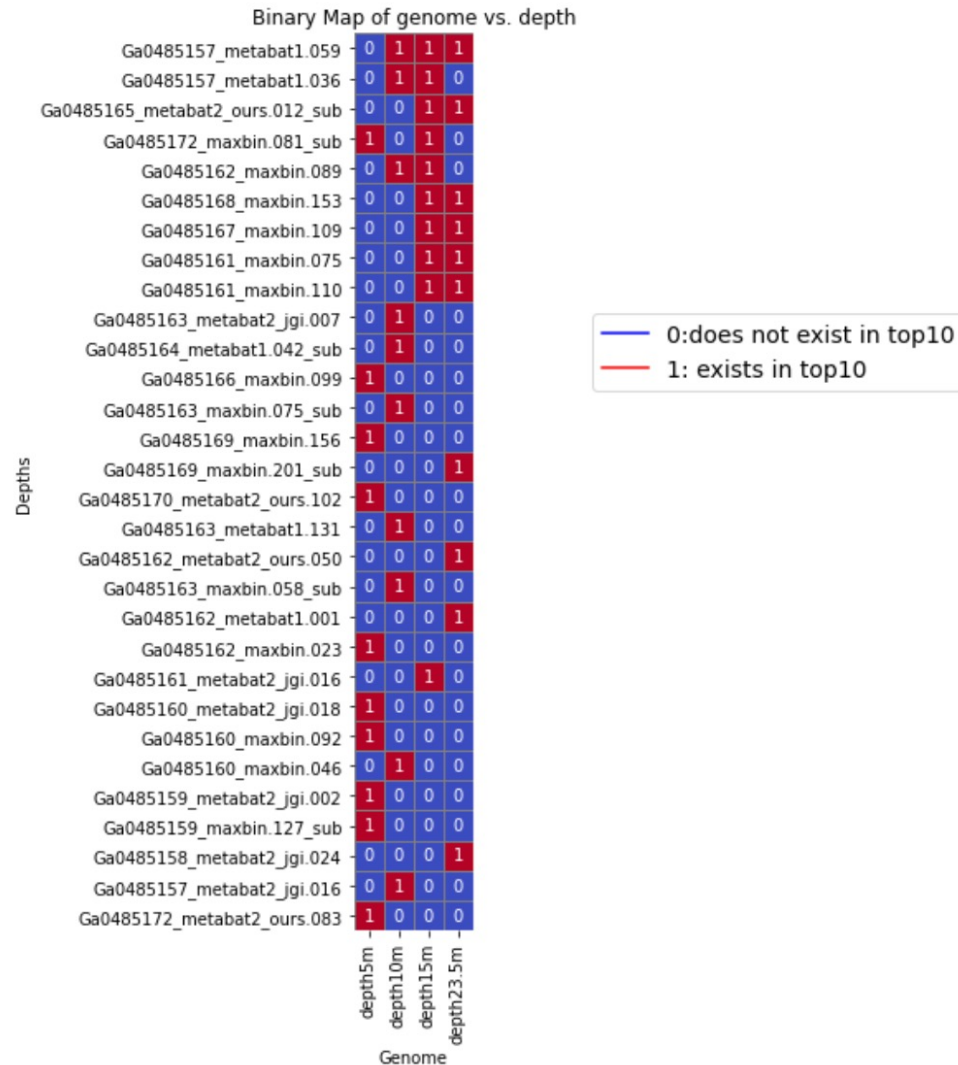
plotting heatmap for all of the genome vs. depth. This time we are treating the Nan values as 0 and others as 1, then generated the plot which shows that this specific genome does exist in the top10 that has been classifies across all the depths.

Still the count datset….

```
{'Ga0485157_metabat1.059': 3,
 'Ga0485167_maxbin.109': 3,
 'Ga0485162_maxbin.089': 3,
 'Ga0485161_maxbin.110': 2,
 'Ga0485161_maxbin.075': 2,
 'Ga0485157_metabat1.036': 2,
 'Ga0485165_metabat2_ours.012_sub': 2,
 'Ga0485169_maxbin.201_sub': 2,
 'Ga0485161_metabat2_ours.188_sub': 1,
 'Ga0485157_metabat2_jgi.016': 1,
 'Ga0485158_metabat1.076': 1,
 'Ga0485167_maxbin.132': 1,
 'Ga0485161_metabat2_jgi.003_sub': 1,
 'Ga0485163_metabat1.131': 1,
 'Ga0485161_metabat2_jgi.016': 1,
 'Ga0485162_metabat1.001': 1,
 'Ga0485167_metabat2_ours.110_sub': 1,
 'Ga0485169_metabat2_ours.047_sub': 1,
 'Ga0485167_metabat2_ours.008': 1,
 'Ga0485158_metabat2_jgi.024': 1,
 'Ga0485172_metabat2_ours.083': 1,
 'Ga0485172_maxbin.081_sub': 1,
 'Ga0485168_maxbin.153': 1,
 'Ga0485160_maxbin.092': 1,
 'Ga0485162_maxbin.023': 1,
 'Ga0485172_maxbin.051_sub': 1,
 'Ga0485172_maxbin.047_sub': 1,
 'Ga0485168_metabat2_ours.001_sub': 1,
 'Ga0485163_metabat2_ours.198': 1}
```
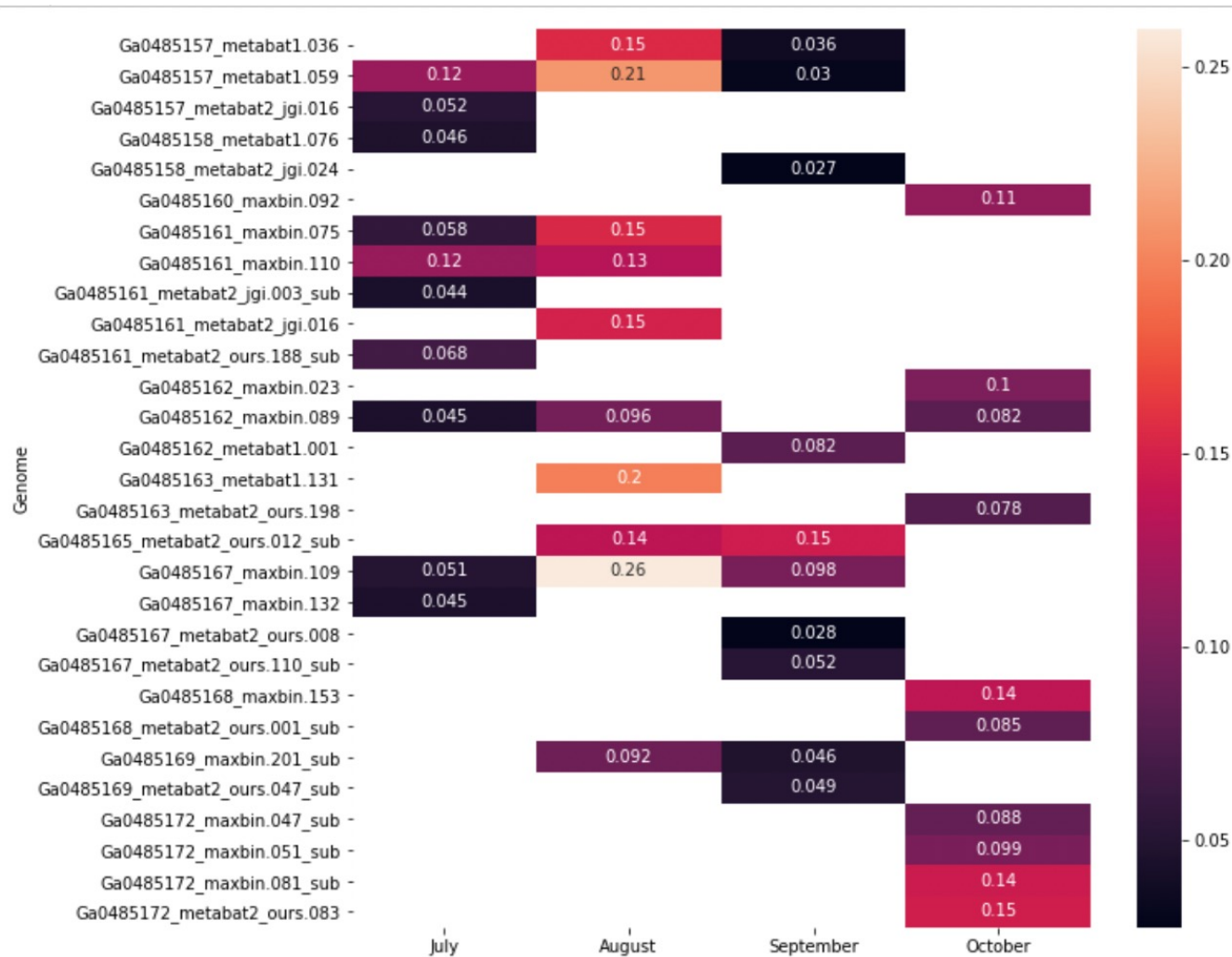
This would be a dictionary that represent the <u>most popular(count) top ten genomes</u> that have appeared in all the months.

I am only focsuing on the <u>most popular(count) top ten genomes</u> that have appeared in all the months.
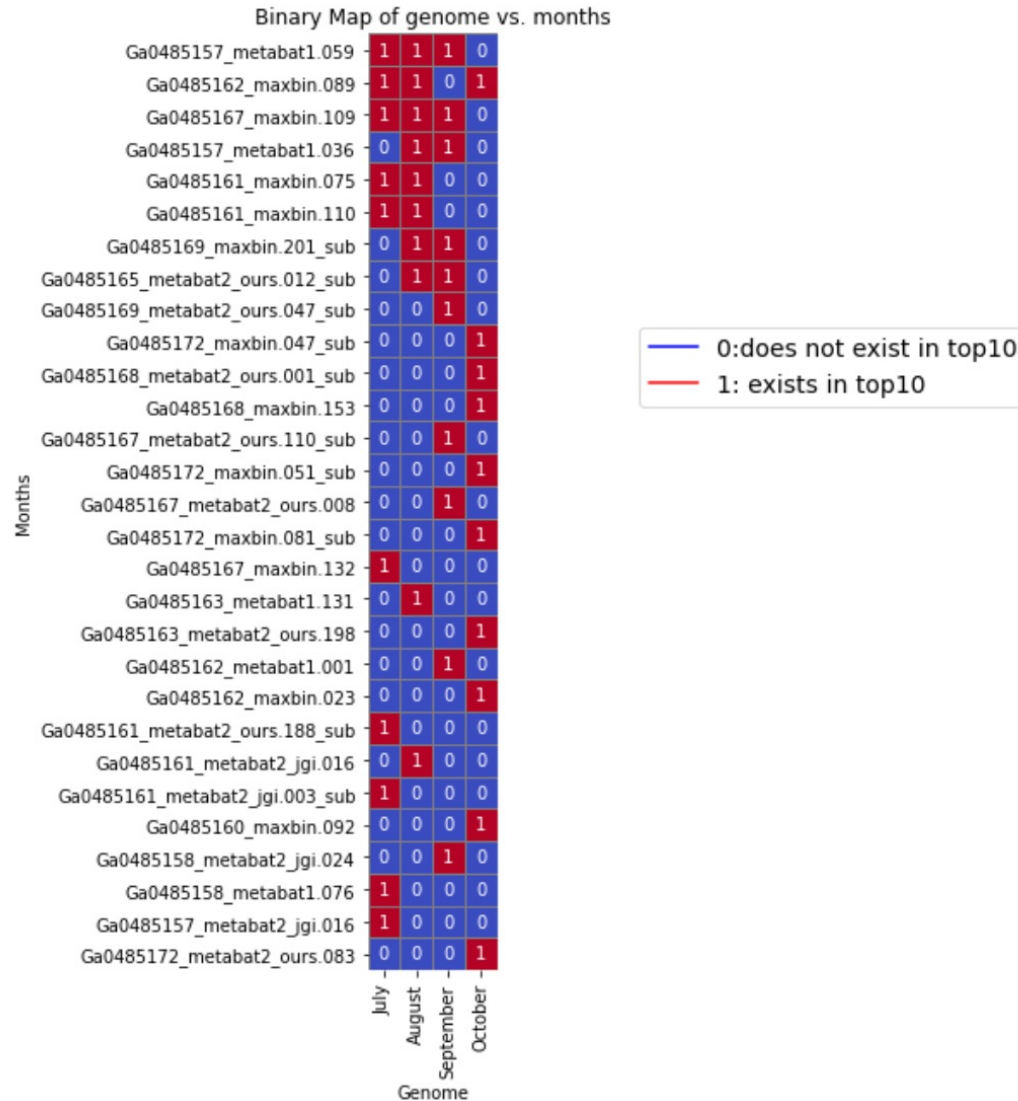
'Ga0485157_metabat1.059', 'Ga0485167_maxbin.109' 'Ga0485162_maxbin.089' exist 3 out of 4 times across all the months.

Count for months(top10)

Plotting heatmap for all of the genome vs. month. However, we are only focusing on the top 10 this time....(since more samples in August, the percentage is expected to be larger).
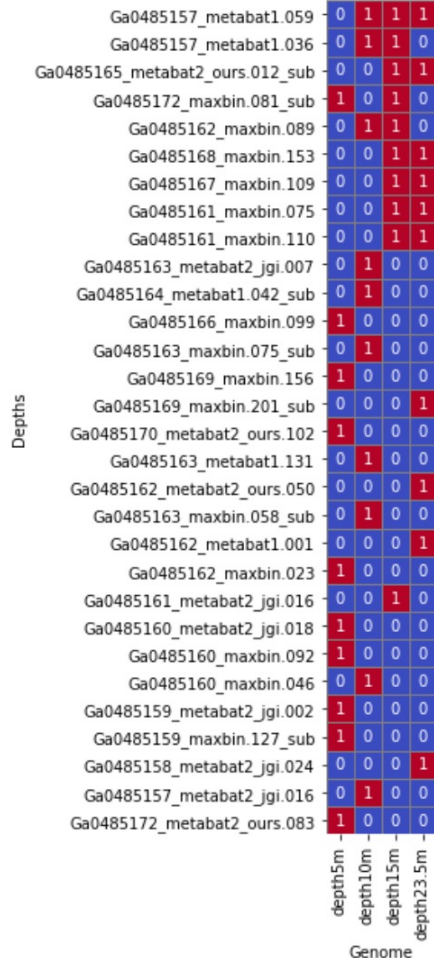
Count for depths(top10)--Binary Map
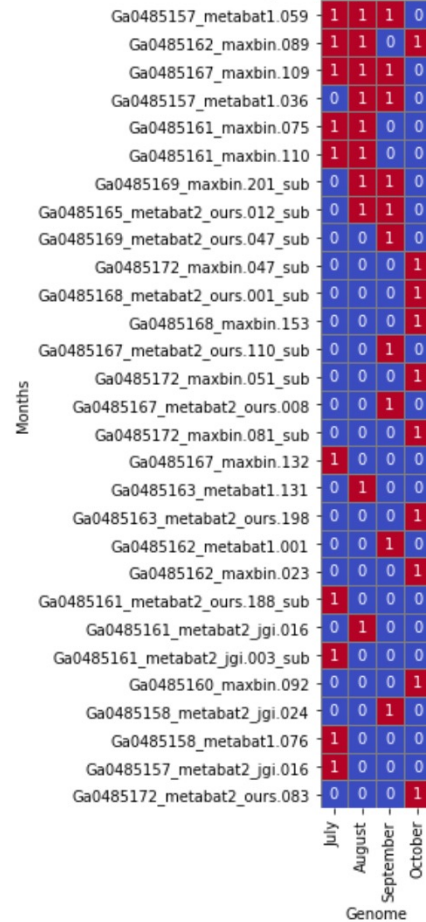

Binary Map of genome vs. months

plotting heatmap for all of the genome vs. months. This time we are treating the Nan values as 0 and others as 1, then generated the plot which show that this specific genome does exist in the top10 that has been classifies across all the months.

Compare two binary plot

Binary Map of genome vs. depth

Binary Map of genome vs. months

0:does not exist in top10
1: exists in top10

{'Ga0485157_metabat1.059': 2,
 'Ga0485167_maxbin.109': 2,
 'Ga0485162_maxbin.089': 2,
 'Ga0485161_maxbin.110': 2,
 'Ga0485161_maxbin.075': 2,
 'Ga0485157_metabat1.036': 2,
 'Ga0485165_metabat2_ours.012_sub': 2,
 'Ga0485169_maxbin.201_sub': 2,
 'Ga0485157_metabat2_jgi.016': 2,
 'Ga0485163_metabat1.131': 2,
 'Ga0485161_metabat2_jgi.016': 2,
 'Ga0485162_metabat1.001': 2,
 'Ga0485158_metabat2_jgi.024': 2,
 'Ga0485172_metabat2_ours.083': 2,
 'Ga0485172_maxbin.081_sub': 2,
 'Ga0485168_maxbin.153': 2,
 'Ga0485160_maxbin.092': 2,
 'Ga0485162_maxbin.023': 2,
 'Ga0485161_metabat2_ours.188_sub': 1,
 'Ga0485158_metabat1.076': 1,
 'Ga0485167_maxbin.132': 1,
 'Ga0485161_metabat2_jgi.003_sub': 1,
 'Ga0485167_metabat2_ours.110_sub': 1,
 'Ga0485169_metabat2_ours.047_sub': 1,
 'Ga0485167_metabat2_ours.008': 1,
 'Ga0485172_maxbin.051_sub': 1,
 'Ga0485172_maxbin.047_sub': 1,
 'Ga0485168_metabat2_ours.001_sub': 1,
 'Ga0485163_metabat2_ours.198': 1,
 'Ga0485170_metabat2_ours.102': 1,
 'Ga0485169_maxbin.156': 1,
 'Ga0485159_maxbin.127_sub': 1,
 'Ga0485159_metabat2_jgi.002': 1,
 'Ga0485160_metabat2_jgi.018': 1,
 'Ga0485166_maxbin.099': 1,
 'Ga0485163_metabat2_jgi.007': 1,
 'Ga0485160_maxbin.046': 1,
 'Ga0485163_maxbin.058_sub': 1,
 'Ga0485164_metabat1.042_sub': 1,
 'Ga0485163_maxbin.075_sub': 1,
 'Ga0485162_metabat2_ours.050': 1}

These genomes have appeared in both the top10 depths as well as the top10 months..

Correlation matrix of the count dataset

#measured by Pearson's correlation
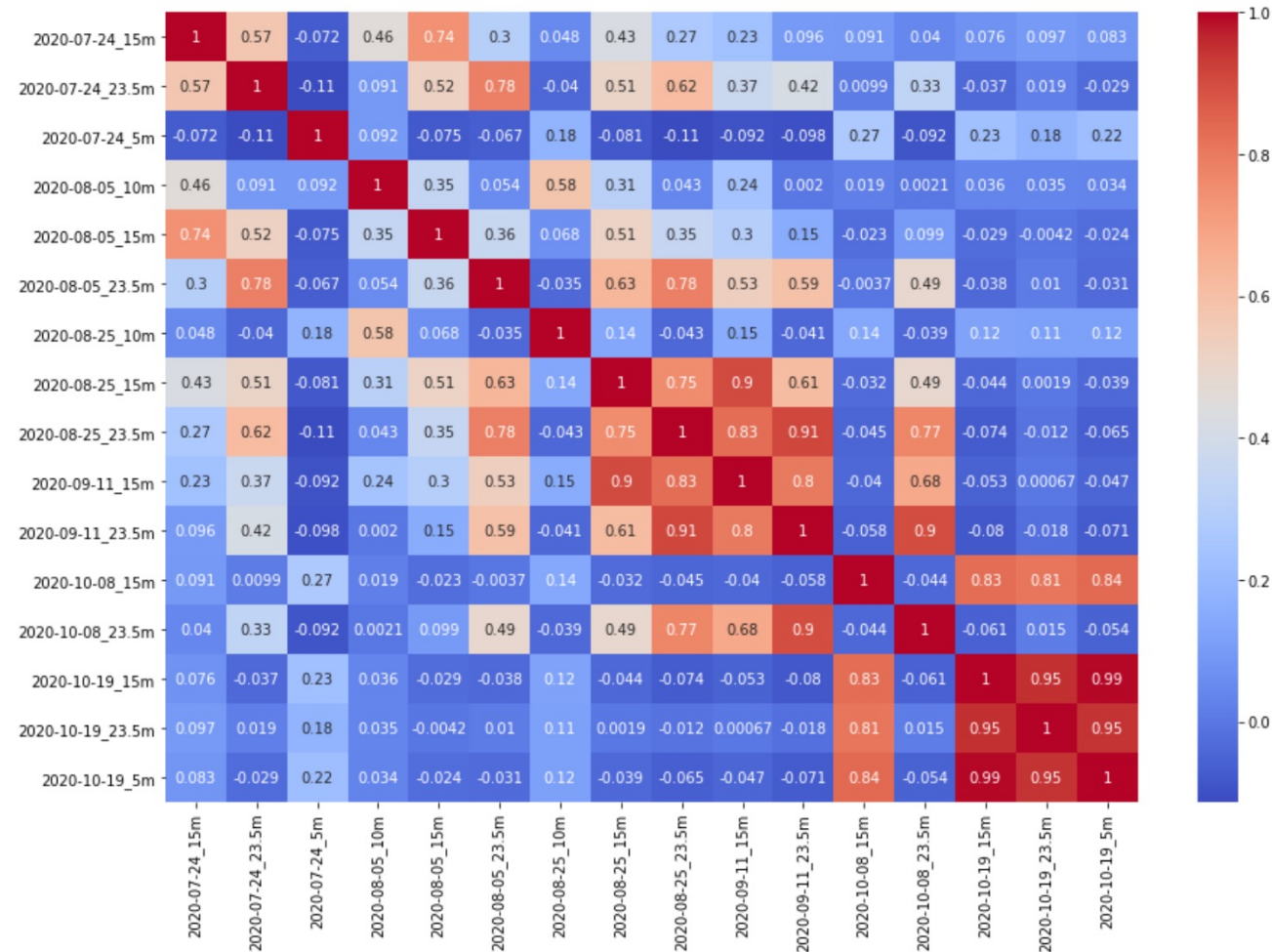
We concluded that sample that are measured at 23.5m on 2020-10-19 is highly correlated with sample that are measured at 15m on 2020-10-19. Sample at 5m on 2020-10-19 is highly correlated with sample that are measured at 15m on 2020-10-19. Sample at 5m on 2020-10-19 is highly correlated with Sample at 23.5m on 2020-10-19.

Correlation matrix of the count dataset
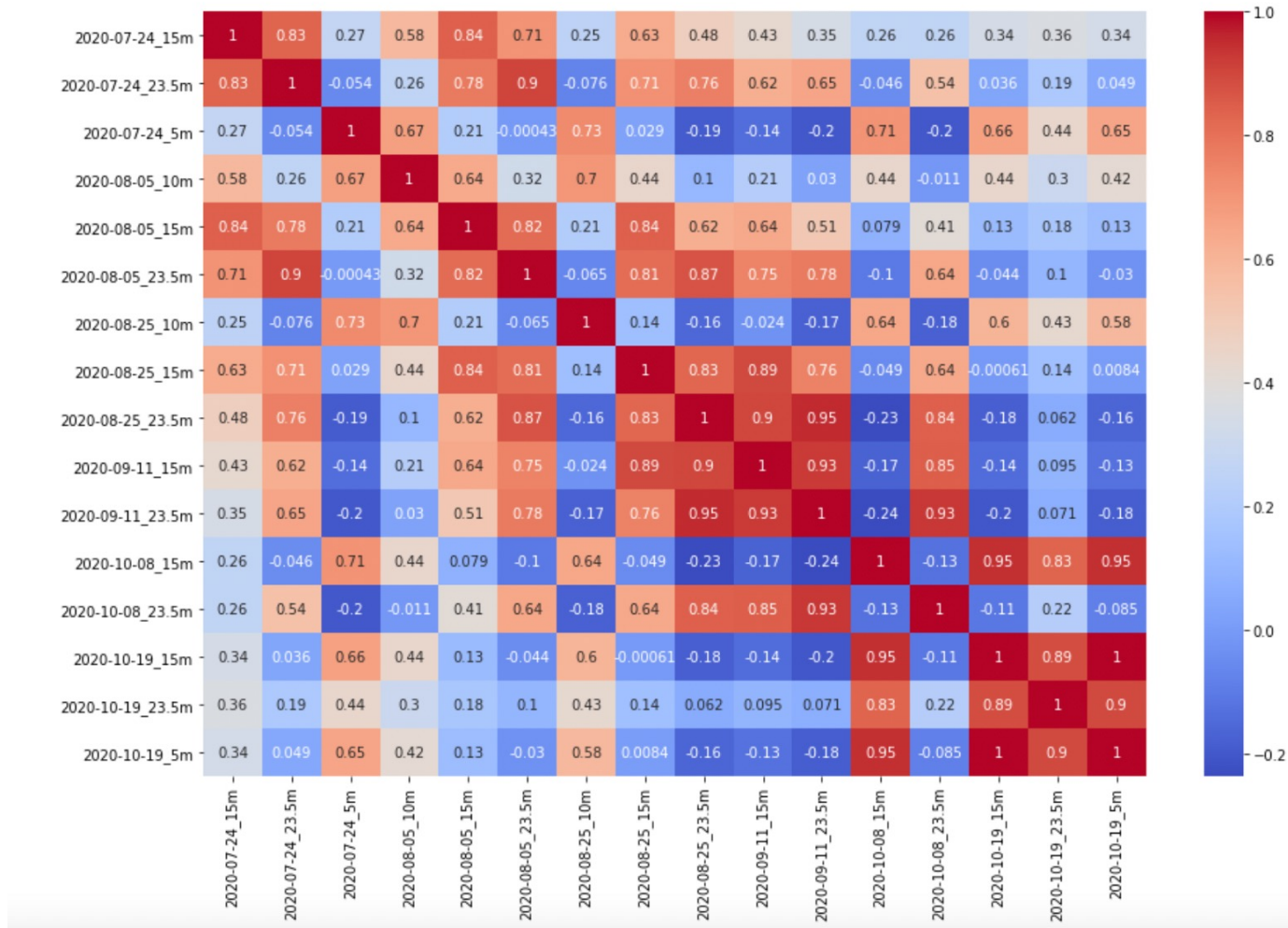
Since this heatmap is generated by using Spearman correlation which measures the monotonic relationship between two variable. we can conclude that the coefficient tend to increase(or perhaps decrease) shows show that the two variables are more related. The last samples that have been mentioned in the conclusion of Pearson's correlation are still very correlated. Besides that, Sample at 15m on 2020-10-08 is highly correlated with all of the samples that have been measured on 2020-10-19.