# lake-microbiome

SY, RA, EG, CSL

June 2023

## 1 Description of 16-sample data

The presented count dataset in this study comprises 16 samples collected over a span of several months, encompassing different dates and depths. With an abundance matrix capturing the count of 431 genomes in each sample, this dataset offers a comprehensive view of the microbial community's dynamics. Analyzing this dataset allows for the exploration of temporal and spatial variations, identification of significant abundance changes, and assessment of diversity patterns. The findings from this study will contribute to a deeper understanding of the intricate interactions between the microbial community and environmental factors, shedding light on the ecological dynamics within this specific ecosystem. Before normalization, a boxplot was created to visualize the relative abundance of genomes per sample in the presented dataset (Figure 1). This boxplot provides insights into the distribution and variability of the relative abundance values across the 16 samples. After normalization, a boxplot was generated to visualize the distribution of the relative abundance of genomes per sample (Figure 2). This plot provides a clear view of the central tendency, variability, and potential outliers in the relative abundance values across the 16 samples collected from different dates and depths. By examining this plot, we can identify the range of relative abundance, the median value representing the typical abundance level, and any potential outliers that deviate significantly from the overall distribution. Notably, upon closer examination, it is observed that there are more outliers in the middle seasons compared to the other seasons. These outliers represent samples with relative abundance values that deviate substantially from the general trend observed in the dataset. By identifying these outliers, researchers can gain insights into the composition of the microbial community within each sample and detect any notable variations in abundance, particularly during the middle seasons.

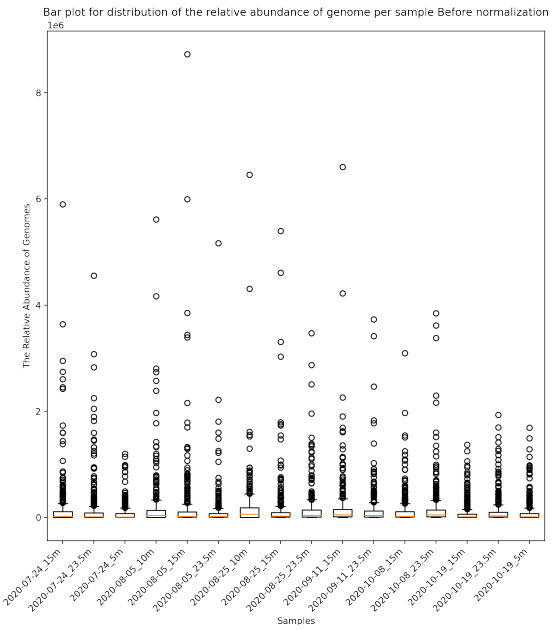


Figure 1: Boxplot for distribution of the relative abundance of genome per sample Before normalization

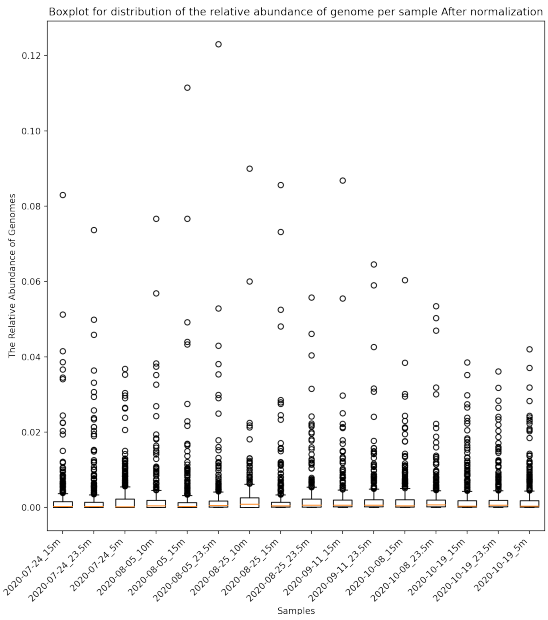


Figure 2: Boxplot for distribution of the relative abundance of genome per sample After normalization

The count dataset was subjected to meticulous analysis, including categorizing the data into different months: July (3 samples), August (6 samples), September (2 samples), and October (5 samples). This categorization facilitated an in-depth exploration of temporal dynamics within the microbial community. By summing up genome counts based on their names, a comprehensive understanding of abundance variations across the months was obtained. A boxplot was generated to visualize these abundance distributions (Figure 3). This boxplot provides insights into the abundance distributions for each month. Notably, in August, the month with the highest number of samples, more outliers are observed compared to the other months. It is important to note that the presence of outliers might be expected due to the larger sample size in August. These outliers represent samples with relative abundance values that deviate significantly from the general trend observed within the specific month. Moreover, a binary heatmap was created specifically for the top ten most popular genomes, displaying their appearance for each month (Figure 4). This visualization allows for a clear depiction of the temporal patterns and dominance of these top genomes. The findings contribute to an enhanced comprehension of the ecological dynamics and temporal fluctuations within the studied ecosystem, providing valuable insights for environmental monitoring and microbiome research.

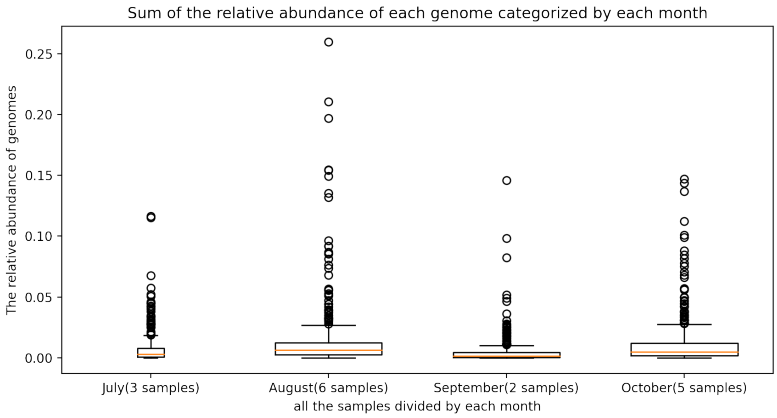


Figure 3: Sum of the relative abundance of each genome categorized by each month based on the normalized data set (row sum method)

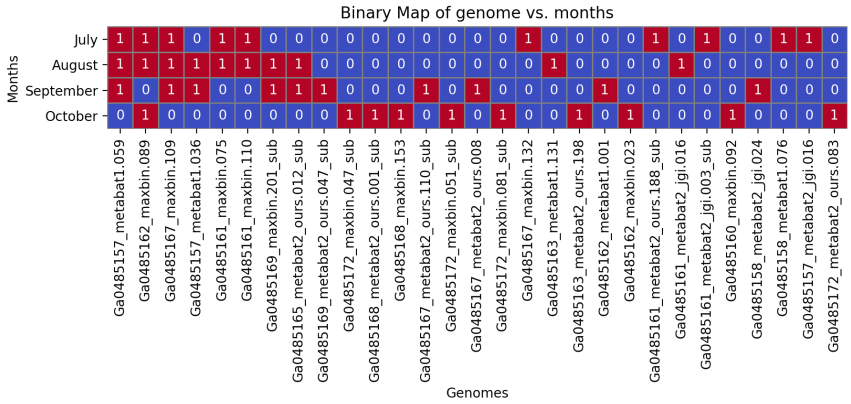


Figure 4: Binary heat map for the occurrence of the top ten most popular genomes for each months

In addition, a comprehensive analysis was conducted on the count dataset, focusing on the diverse sampling depths: 5 meters (2 samples), 10 meters (2 samples), 15 meters (6 samples), and 23.5 meters (6 samples). This approach enabled a detailed exploration of the microbial community dynamics specific to each depth. Through the aggregation of genome abundance by their names, a nuanced understanding of abundance variations within each depth was obtained. To visually represent these variations, a boxplot was generated, offering insights into potential patterns or distinctions in genome abundance across

depths(Figure 5). Notably, as the depth becomes deeper, more outliers are observed. It is important to note that since there is a lower number of samples for depths 5 and 10 meters (2 samples each), it is expected to have a smaller count in those depths. The boxplot analysis enables the identification of central tendency, variability, and potential outliers, thereby revealing the abundance dynamics within each depth. Furthermore, a distinct binary heatmap was constructed, specifically highlighting the occurrence of the top ten most prevalent genomes for each depth (Figure 6). The insights gained from this analysis are valuable for environmental monitoring and microbiome research, shedding light on the intricate relationships between microbial communities and specific environmental features within the ecosystem.

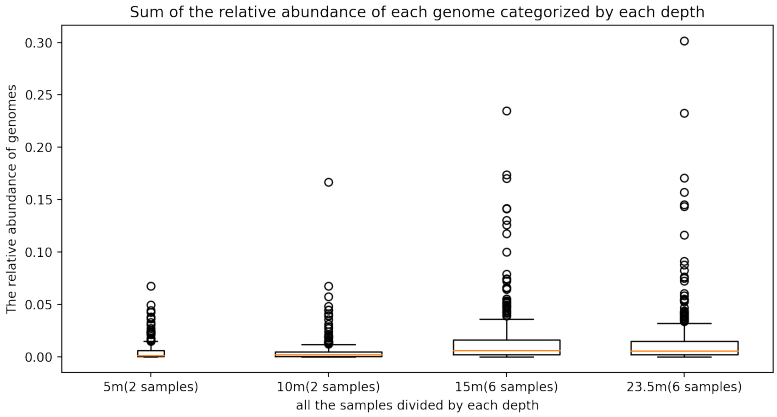


Figure 5: Sum of relative abundance of each genome categorized by each depth based on the normalized data set (row sum method)

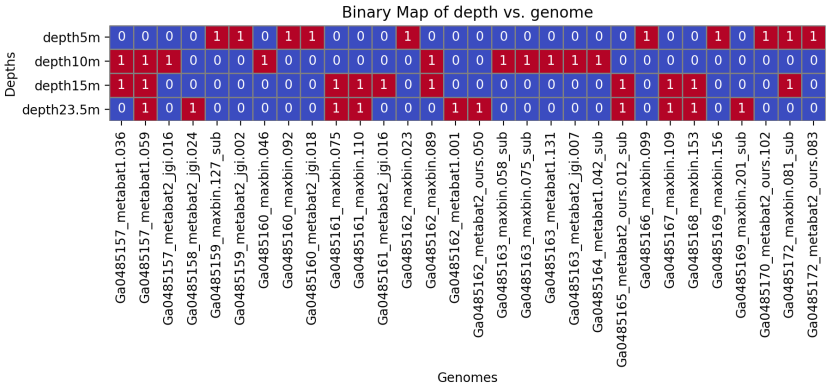


Figure 6: Binary heat map for the occurrence of the top ten most popular genomes for each depths

In order to gain further insights into the microbial community dynamics within the studied ecosystem, a comparative analysis was performed on the relative abundance of specific genomes across different time periods (July, August, September, and October) and depths (5 meters, 10 meters, 15 meters, and 23.5 meters). The relative abundance represents the ratio of the appearance frequency of each genome that has been classified as the top ten most prevalent genomes. To visualize this data, a bar plot was created (Figure 7). The x-axis represents the sum of the ratio of each genome, and the y-axis represents the genomes. Each bar or cell in the plot represents the cumulative appearance ratio of specific genomes for a given time period (month) and depth. By examining the bar plot, patterns and variations in the relative abundance of genomes across months and depths can be identified. This analysis provides a comprehensive view of how the microbial community composition varies both temporally and spatially within the ecosystem. It allows for the identification of the shared presence or exclusivity of certain key genomes across different dimensions. These comparative analyses pave the way for further investigations into the complex interactions between the microbial community and environmental variables, facilitating a deeper understanding of the interconnectedness between months and depths in shaping the microbial community dynamics.

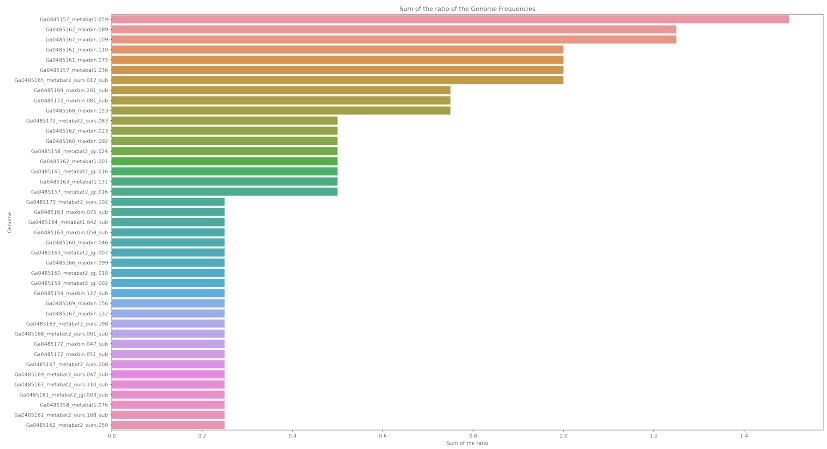


Figure 7: Sum of the ratio of the appearance of the Genome Frequencies

To assess the interrelationships and co-occurrence patterns of the 16 samples, correlation analysis was conducted using two distinct methods: Pearson's correlation and Spearman's correlation. Interestingly, both methods produced similar results in terms of correlation values, indicating a consistent pattern of associations among the samples. Pearson's correlation revealed linear relationships between the samples, providing insights into potential linear dependencies and associations (Figure 8). 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. Similarly, Spearman's correlation, which captures both linear and non-linear dependencies, confirmed the presence of monotonic relationships among the samples (Figure 9). By examining the figure, we can conclude that the coefficient tend to increase(or perhaps decrease) which shows 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. This convergence in results strengthens the robustness of the observed correlations and enhances our confidence in the findings. The correlation matrices derived from these methods provided valuable information on the similarities and dissimilarities between the samples, elucidating potential shared microbial composition. These correlation analyses contribute to a deeper comprehension of the complex interactions and co-occurrence patterns within the microbial community, offering essential insights for understanding the ecological dynamics and functional relationships in

the studied ecosystem.

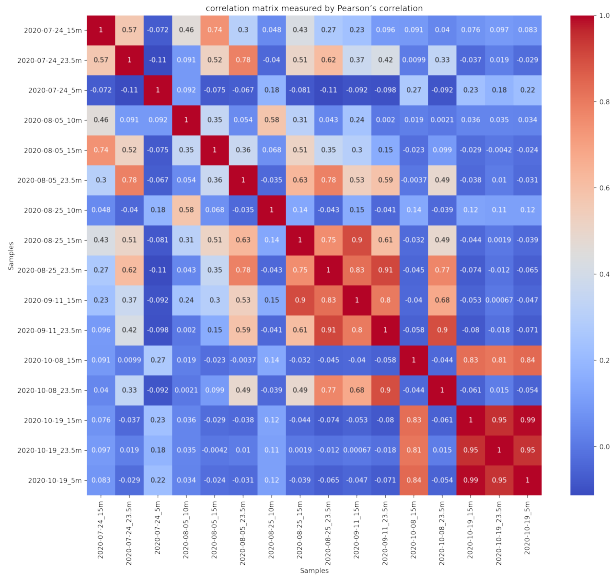


Figure 8: correlation matrix measured by Pearson's correlation

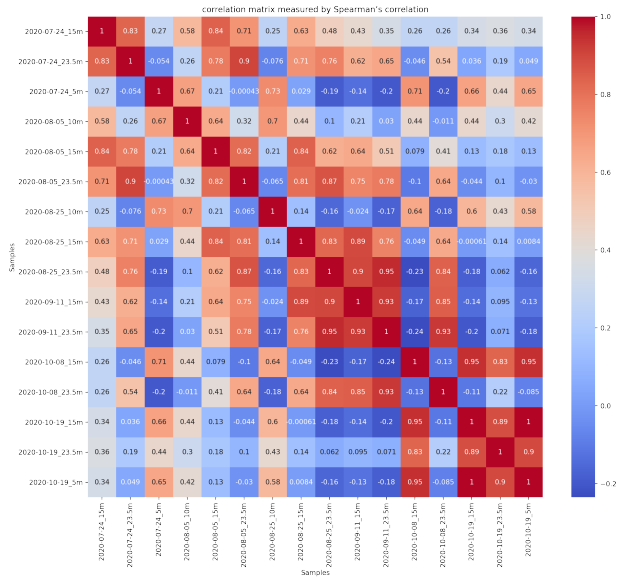


Figure 9: correlation matrix measured by Spearman's correlation