Visualizing metagenomic data in R using Jetstream

Haley Leffler (hleffler@iu.edu), Department of Human Biology, Indiana University, Bloomington

Sheri A. Sanders (ss93@iu.edu), National Center for Genome Analysis Support (NCGAS), Indiana University, Bloomington

Bhavya Papudeshi (bhnala@iu.edu), National Center for Genome Analysis Support (NCGAS), Indiana University, Bloomington

ABSTRACT

Metagenomes consist of the total genome content collected from an environmental sample containing bacterial, archaeal, and viral sequences present. These datasets are complex and can be overwhelming to visualize. Using multiple visualization methods would benefit researchers by allowing them to perform exploratory analyses that could aid in downstream analysis of the data. This paper focuses on using different visualization methods including a rarefaction curve, ordination plots, alluvial plot and heatmap to represent a metagenomic dataset using Jetstream. Applying the visualization methods on a hydrocarbon seepage metagenomic dataset, we found that the samples cluster based on location, one sample was similar to both reference and seep samples, and the datasets had human contamination. These findings can now lead to potential downstream analysis questions to further assess this data. The scripts and input files used to create the different visualizations are available on GitHub (https://github.com/hleffler/Microbial-visualization).

INTRODUCTION

Metagenomics is a culture-independent technique that allows researchers to collect and study the samples that are taken directly from the environment. This technique has revealed previously unobserved richness of microorganisms in many environments [1, 2, 3], and showcased the integral role of microbes in maintaining human health and natural environments [1, 3]. In human microbiome studies, metagenomic technique has shown microbial profiles can potentially be used as biomarkers for diseases, medical treatments, disease, and other homeostatic factors [1, 2]. Researchers applying this technique generate large datasets to capture the complexity of the microbial community. In turn they face challenges during analysis, due to lack of computational training, computational power, or methods to accurately visualize these high dimensional datasets [2, 3]. While computational biologists are developing novel methods and tools to analyze these complex datasets, usability of these programs can remain challenging.

In this paper we focus on addressing these challenges, especially focusing on visualization of these complex datasets. Visualization was done using R programming language, which has been popular among biologists due to its ability to perform statistics, visualizations and data analysis. Using this language would likely address the challenge for lack of computational training. Since metagenomic datasets contain millions of bacterial, archaeal and viral sequences, visualization of these datasets in one graph can be overwhelming [1]. Visualization is an essential exploratory step that allows researchers to represent the data from different points of view, highlighting patterns and in turn perform downstream analysis accordingly [1]. In an effort to overcome these challenges, this project focuses on visualizing a metagenomic dataset in multiple ways using reusable R scripts on accessible cloud computing resource, Jetstream.

Visualization of metagenomes are commonly represented as stacked bar plots to show the taxonomic and functional differences between the different samples collected. These bar plots do help point out the differences between the datasets, however highlighting only the differences between the most abundant taxonomic or functional groups. This overlooks all the other microbes that are less abundant or rare species that could also be playing significant roles within the community. Here we present other ways to visualize metagenomes that allow for better visualization methods that represent both the abundant and rare species in these complex datasets. Generally the first question is if the dataset is a representative sample because metagenomic datasets only capture a subset of the microbial community to represent the population. To explore this idea, we can apply rarefaction curves to show species richness among the samples collected. This shape of the curve, which increases exponentially as new species are identified and begins to plateau as fewer unique species are identified, determines if the samples are representative of the larger population in the environment. The other visualization method to plot high dimensional datasets are ordination plots such as Principal Component Analysis (PCA), non-metric Multidimensional Scaling (nMDS), and Principle Coordinate Analysis (PCoA). These plots are used to visualize patterns or gradients, inferring similarity between the samples and spotting anomalies. A PCA plot will show the direction of maximal variation between the samples by condensing the data into new variables, which ensures that the data set can be represented without compromising any input variable. A non-metric Multidimensional Scaling (nMDS) is another ordination plot that visualizes the data based on the distance similarity. Like PCA and nMDS plots, PCoA plots are used to visualize patterns or gradients in the data set to infer the similar or dissimilarity of the samples. The three plots provide an initial exploratory analysis to determine any anomaly in the samples or how they can be represented in a 2D space. Recently developed Sankey or alluvial plots developed a representation of how the data changes or flows through a set of variables or metadata. These plots allow exploration of data, visualizing the structural, taxonomic and functional changes within the datasets. However, this visualization method is great to look further into subsets of the data, simplifying and looking closer into specific characteristics of the dataset. Another method is a heatmap, which is a visual representation of a table with numerical data that replaces numbers with a color scale. Heatmaps will help visualize the multiple taxa identified from these samples.

In this project, we will be applying the above described methods on metagenomic data that was downloaded from a research study on microbial response in hydrocarbon seepages. Hydrocarbon seepages are the natural release of oil or gas bubbles from the ocean floor. This study found that hydrocarbon seepages can significantly alter the microbial community structure and change their oxygen reduction system to sulfate based methane oxidation and nitrogen fixation [4]. In the paper published with this dataset, the data was represented using stacked bar charts, heatmaps, and a maximum likelihood phylogenetic tree that pointed out differences in gene abundances and compared metagenomic assembled genomes (MAGs) [4]. Majority of the reads in this dataset remain unidentified due to the novelty of the sampled environment, making it an interesting subject for visualization.

Jetstream is a fully configurable, cloud computing resource that is funded by the National Science Foundation (NSF) [5]. Jetstream provides a user interface for researchers inexperienced in computing and allows users to have access to fast computational resources interactively using virtual machines (VM) [5]. This resource is preferred for this project because users can analyze a large amount of data without using resources from their local computer.

METHODS

Materials and Data Collection

The metagenomic data set for this project was collected from BioProject number, PRJNA553005). The datasets were downloaded from the Sequence Read Archive (SRA) using sratoolkit. This data set comes from a research study that contains seven sediment metagenome samples from the Gulf of Mexico. The samples were grouped and labeled following the research study's categorizations: seep 1 (D24, D27, and D23), seep 2 (D72 and D75) , reference (D21 and D30) which are based on where the samples were collected. Sample D24 was also labeled as a transition sample in the study because it contained characteristics from the seeps and reference locations.

Computational resources

The data set was analyzed on a Ubuntu base Jetstream VM with 6 CPUs, 16 GB of memory and 60 GB of storage with an additional 100GB volume attached. Jupyter notebook (5.2.2) and R (3.4.4) was downloaded onto the VM for documentation and visualization.

Taxonomic Annotation

Using the Kraken (2.0.8) toolkit that is available at GitHub [6], each sample was aligned against the microbial database and given a taxonomic report. The reports for the samples were combined into a table with csv format using a python script from GitHub (https://github.com/npbhavya/Kraken2-output-manipulation) to format the taxonomic table to be compatible with R.

Rarefaction Curve

A rarefaction curve was made to determine if the sample was representative of the larger population of microbial communities. To plot a rarefaction curve, input data was the taxonomic report for family level classification and their corresponding abundances. R packages vegan with rarefy and rarecurve functions were used to plot these curves.

Ordination Plots

To determine how the samples correlate with each other, we used ordination plots. The ordination plots use the same input file as the rarefaction curve, but first the data is normalized using the rarefy function, which applies a variation of the min-max normalization method. This normalized data was then provided as input to PCA, nMDS and PCoA ordination plots. First the distance between the samples was calculated using vegan package's pca, mds, and betadisper functions relatively and plotted using ggbiplot and ggplot packages.

Alluvial Plot

Through an alluvial plot we visualize the structural and taxonomic changes in the datasets based on where the sample was collected. The input data for this plot included phylum level classification with their corresponding percent abundance, a column with where the sample was collected from, and other metadata such as proximity to hydrocarbons. R packages, ggalluvial and ggplot2 packages were used for this representation.

Heatmap

To look at the overall differences in the species level classification between the samples, we represented the data using a heatmap. Heatmap was plotted using the percent abundances of the taxonomic rank at species level classification in a color coded format, using the ComplexHeatmap package from Bioconductor v3.11.

RESULTS

We applied four different visualization methods to explore the metagenomes collected from the Gulf of Mexico to show that the microbial communities are different based on where the samples were collected. The microbial species present within the hydrocarbon seeps were more similar to each other compared to the reference samples. Applying the visualization methods on a hydrocarbon seepage metagenomic dataset, we found that the samples cluster based on location, one sample was similar to both reference and seep samples,and the datasets had human contamination. The input data, R scripts, and Jupyter notebook used for this project are available at https://github.com/hleffler/Microbial-visualization.

Taxonomic report from Kraken 2

The average number of reads is 336 ± 25 bp per sample, with an average length of 250bp. From the taxonomic tables generated from Kraken 2 results, the samples identified 45 phylum, 539 families, and 2,339 genera among all the seven samples. Based on the percent abundance, most samples had a low abundance for all taxa less than 9%. However, 88 to 90% of the reads remain unidentified in each sample regardless of the taxa level of classification.

Rarefaction Curve

The rarefaction plot shown in Fig. 1 represents the number of species identified relative to the sample size. The samples are color coded based on the location each sample was taken. For all seven samples, the sample size and species increases dramatically which shows new organisms being identified. The plateau shows the slow in number of identified species and the increase in the number of rare species. The rarefaction plot in Fig. 1 shows that the metagenomes collected were representative as the samples begin to plateau when 475 unique species were identified per 1.5 million sequences. shows that sample D21 (reference) does not have as many rare species identified compared to the other samples since this line stops short of the other samples. Similarly, sample D24 (seep 1) stops in between sample D21 and the rest of the samples. The other samples, D72, D75, D33, D30, and D27 (seep 1 and seep 2) show a similar pattern when plotted showing that they have a similar number in rare species. The sharp increase and plateau that is seen in all samples shows that the seven samples are a representative sample of the larger microbial community in the Gulf of Mexico.

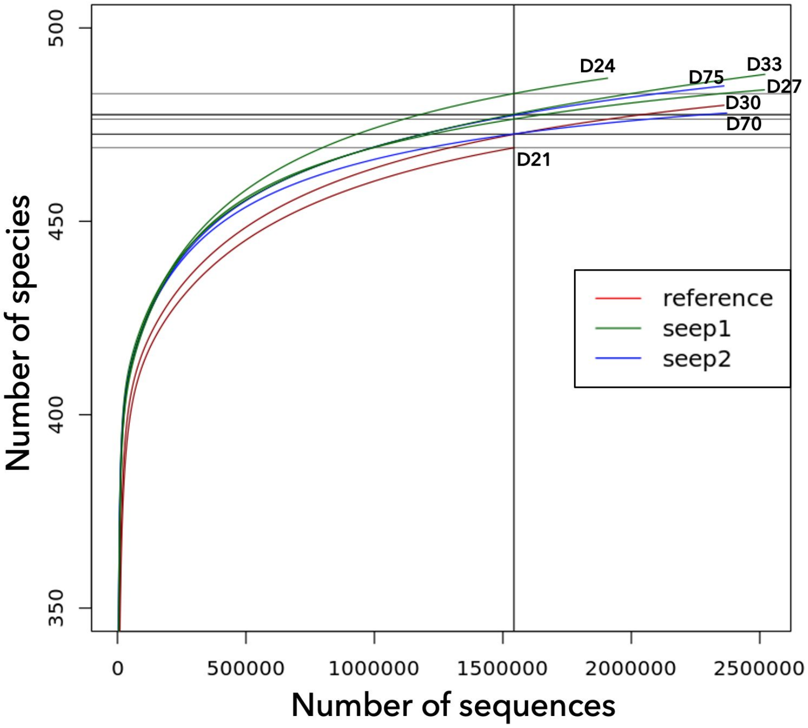

Figure 1: The rarefaction curves exponentially increase as new species are identified, then plateau as unique species are identified. Number of species is represented by the y axis and the number of sequences is represented by the x axis. Reference samples are in red, hydrocarbon seep site 1 (seep1; D24, D27, D33) is in green and hydrocarbon seep site 2 (seep2; D70, D75) is in blue.

Ordination Plots

Collectively in all three ordination plots, the seep samples and reference samples cluster together respectively, which shows higher similarity of the microbial profiles based on the source of the samples as shown in Fig 2. PCA plot shown in Fig. 2a, represents the seven samples grouped by the location the sample was taken and the correlation between them. The seep 2 samples and one seep 1 sample are plotted close together on the plot which shows there is a higher correlation between them compared to the other samples. Similarly, the reference samples are plotted closer to each other in comparison to the other samples, indicating their correlation with each other and how they are not correlated with the other locations. The taxa represented by the arrows shows that some taxa are more prevalent in some samples. This PCA plot shows that the taxa families, Hominidae, Methanosarcinaceae, and Flavobacteriaceae are more correlated with the seep samples while Pseudomonadaceae, Burkholderiaceae, Rhodobacteraceae, Bradyrhizobiaceae, and Streptomycetaceae are more correlated with the reference samples.

Looking at the nMDS plot in Fig. 2b we can determine that the two reference samples are very similar to each other as they are placed close to each other in the graph. Similarly, there are more differences within the seep samples, especially the seep 1 samples that cluster really far away from each other. In order to further gain more confidence on the representation, we can look at the goodness of fit value that determines the differences between the actual distances and predicted values. The shepard plot shows how the data scatter around the line. Large scatter around the line suggests original dissimilarities are not preserved. In this case, the data points are all on the line suggesting that the data fit the model.

Looking at the PCoA in Fig. 2c, the two reference samples are very similar to each other since they are placed closer to each other in the graph compared to the other samples. Similarly, there are more differences within the seep samples, but they cluster together.

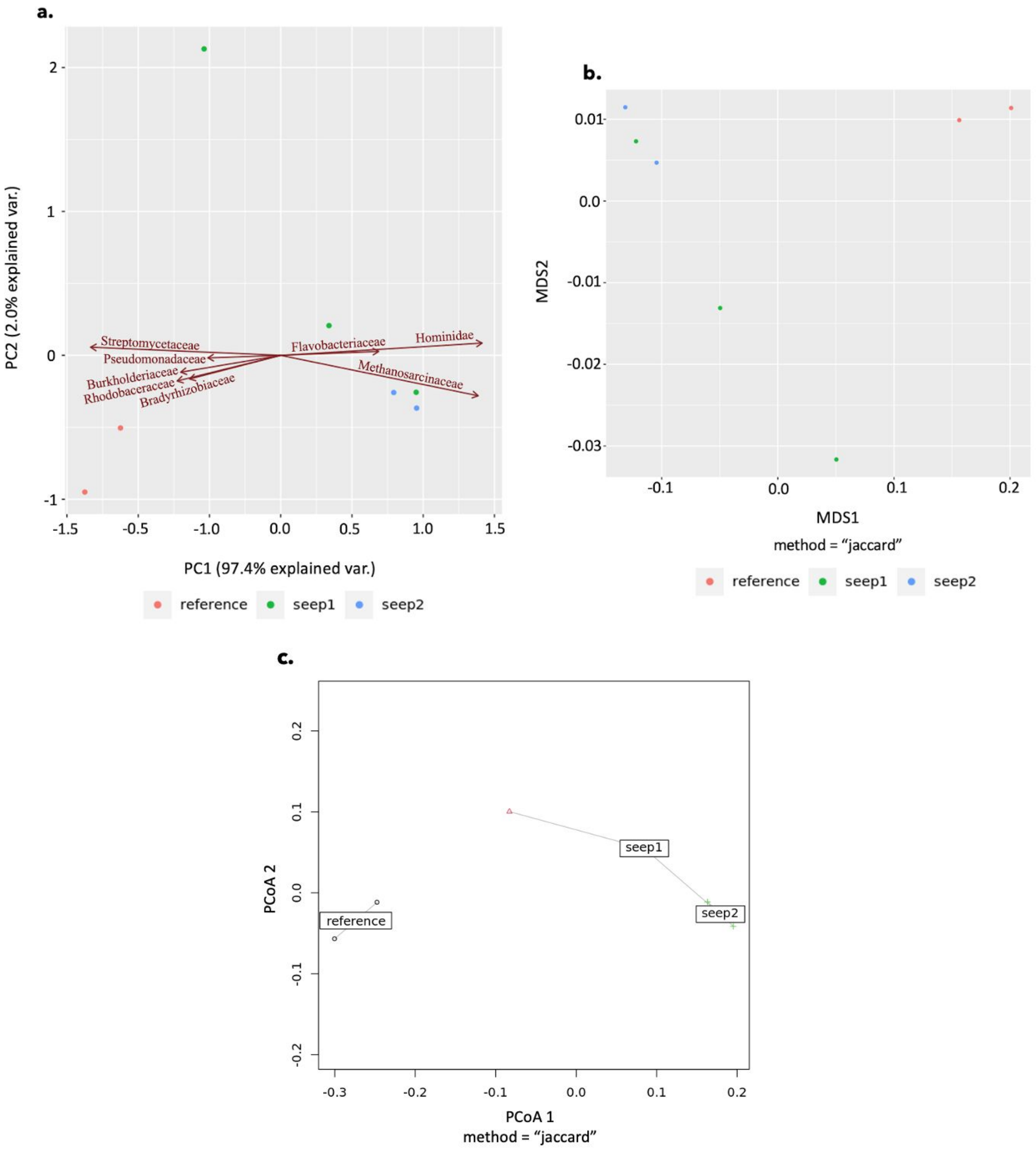

Figure 2: Ordination plots of sediment metagenome data and taxa families. Reference samples are in red, hydrocarbon seep site 1 (seep1) is in green and hydrocarbon seep site 2 (seep2) is in blue.

2a. The two axes on the PCA explain 99.4% of the variation. Vectors (arrows) indicate the direction and strength of eight most abundant taxa families to the overall distribution.

2b. nMDS of the sediment metagenomes using the Jaccard dissimilarity metric.

2c. PCoA of the sediment metagenomes using the Jaccard dissimilarity metric. Reference samples are in black, hydrocarbon seep site 1 (seep1) is in red and hydrocarbon seep site 2 (seep2) is in green.

Alluvial plot

The alluvial plot shown in Fig. 3 shows how the sediment metagenome samples flow through different variables, location, taxa abundance, and detection of hydrocarbons. This graph shows how the transition sample (D24), compares to the other sample locations. The transition sample is like the seeps since it has no hydrocarbons, and like the reference sample because there were no thermogenic gases detected.

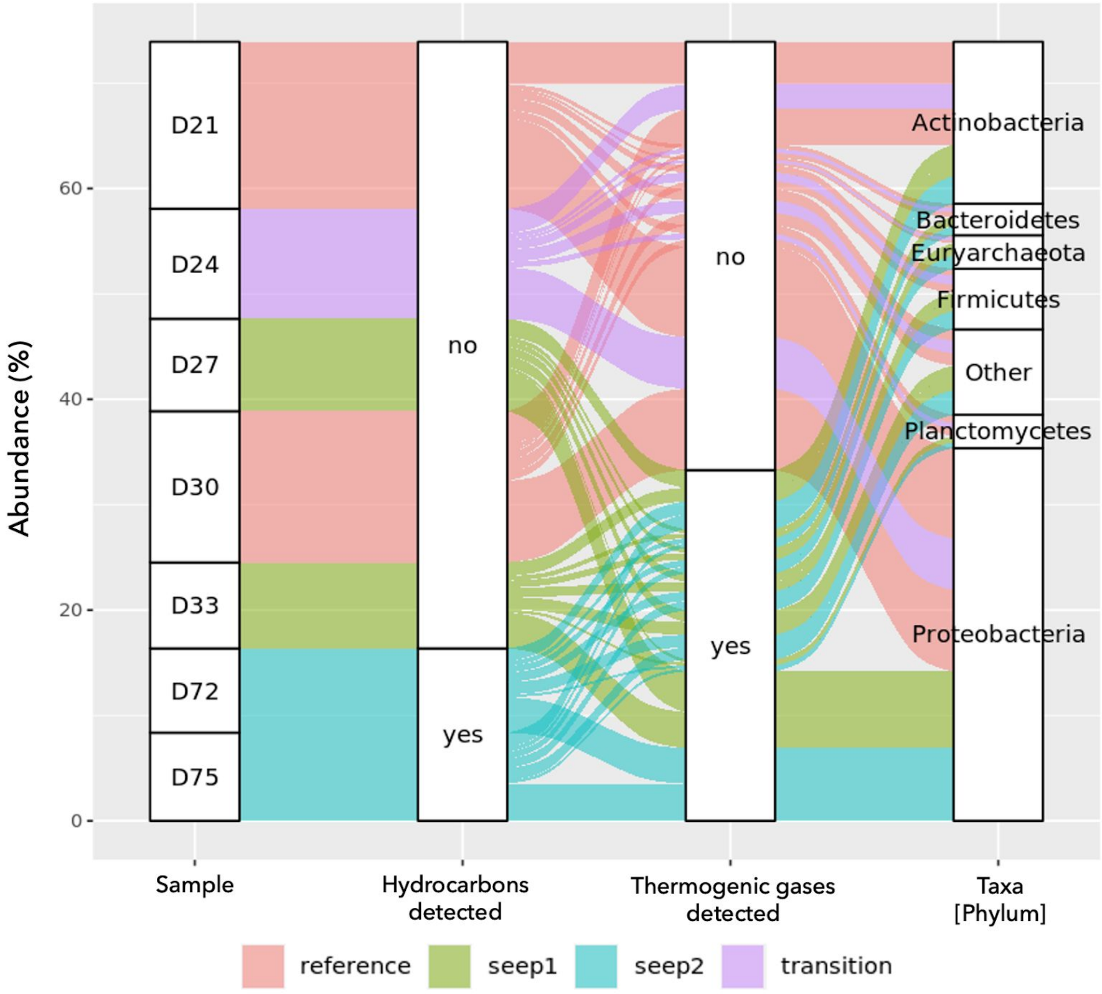

Figure 3: The alluvial plot shows differences between samples based on their metadata. Reference samples are in red, hydrocarbon seep site 1 (seep1; D27, D33) is in green, hydrocarbon seep site 2 (seep2; D70, D75) is in blue, and the transition sample from seep 1 is in purple. Detection of hydrocarbons and thermogenic gases is indicated by the two middle columns. The last column represents the percentage of the six most abundant phylum (out of 46) with the remaining binned to the category Other.

Heatmap

In order to see how the abundance of taxa at species level varies among the samples, we made a heatmap. The heatmap shows that the species, Homo sapiens, is most abundant in the samples. The other species have under 0.17% abundance values for each sample. Looking at the dendrogram for the samples, the reference samples are similar, seep2 samples are similar, and the transition sample (D24) is more similar to D27 in seep1. Additionally, the seeps are more similar to each other while the reference samples are more similar to each other.

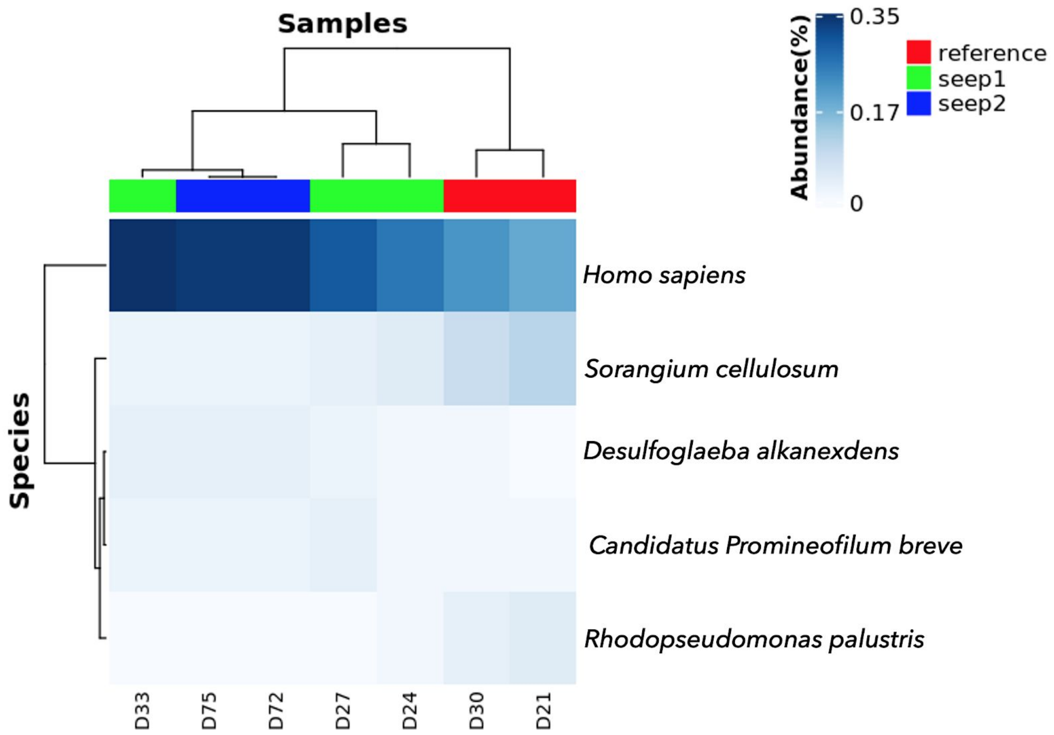

Figure 4: The heatmap shows the percentage of abundant species identified (0.1% - 0.35%) indicated by the color scale (max - dark blue, min - light blue). Samples are annotated as reference - red, seep1 - green, seep2 - blue. The dendrograms on the left and top of the heatmap indicates the relationships between the samples and species respectively.

DISCUSSION

After exploratory analysis on the hydrocarbon seepage datasets, we are able to draw conclusions that align with Zhao *et al*. and reveal more about the samples such as their representativeness and contamination. The rarefaction curve suggested that the dataset is a representative sample of the larger population, and the ordination plots revealed that the samples taken from the same seep sites are more similar to each other compared to samples from the reference sites. This trend of similarity is parallel to Zhao's findings [4]. The alluvial plot further supports Zhao *et al.,* findings by showing that the D24 sample has similarities to both locations, seep and references which suggests that the availability of nutrients in the sites may be driving change in microbial taxa and influencing their functional profiles. Lastly, the ordination plot has an eigenvector for family Hominidae and the heatmap showed *Homo sapiens* as the most abundant taxa. This suggests that the samples were contaminated with human DNA and leading us to do a quality control step prior further analysis. Additionally, on the heatmap the grouping from the dendrograms suggests that D24 (transition sample) is more similar to D27 from seep1 which contrasts other plots made. This draws out the differences between these methods and that using one method to explore the data could lead to biased conclusions. It is also good to note that while the visualization methods answer exploratory research questions, no statistical analysis was done to show if there was any significance. Limitations of visualization methods summarized in Table 1 below should be considered when planning exploratory analysis on big data.

Table 1: Summary of the strengths and limitations of the four visualization methods explored in this paper.

Plot Type	**Key Features**	**Limitations**
Rarefaction curve: Allows calculation of species richness from sample	- Determine if sample is representative of larger population - Easy to read	- Need a large sample or samples from the same community - Does not show specific species, just number - No specified number of samples to enter - No stats
Ordination plots (PCA, NMDS, PCoA): Reduces dimensions to show correlation or distance of samples based on their distance and grouping in graph	- Shows correlation or distance measure of each sample is based on groupings - Condenses data into a readable format - Optional stats	- Involves math - Correlation ≠ significance
Alluvial plot: Shows how samples flow through variables/time	- Shows flows/changes in data through variables or time - Can show many categories of data	- Specific input data type (columns) - May need to add columns for more variables - Can become overwhelming w/ a lot of samples - No stats

Heatmap: Displays visual summary of a matrix	- Color format is easy to analyze - No math	- Too many samples becomes overwhelming

Overall, this paper emphasizes the relevance of exploratory methods on complex datasets like metagenomic datasets. However these can be applied to similar datasets to help researchers learn more about their prior any further analysis. Exploring the datasets at the beginning could potentially help narrow down downstream analysis to specific variables that have an interesting pattern within the samples or point out contamination as it did in the discussed dataset.

References

1. K. Sudarikov, A. Tyakht, and D. Alexeev, “Methods for the metagenomic data visualization and analysis,” *Curr Issues Mol Biol*, vol. 24, pp. 37–58,2017.
2. F. P. Breitwieser, J. Lu, and S. L. Salzberg, “A review of methods and databases for metagenomic classification and assembly,” *Briefings in bioinformatics*, vol. 20, no. 4, pp. 1125–1136, 2019.
3. P. Hugenholtz and G. W. Tyson, “Metagenomics,” *Nature*, vol. 455, no.7212, pp. 481–483, 2008.
4. R. Zhao, Z. M. Summers, G. D. Christman, K. M. Yoshimura, andJ. F. Biddle, “Metagenomic views of microbial dynamics influenced by hydrocarbon seepage in sediments of the Gulf of Mexico,” *Scientific reports*, vol. 10, no. 1, pp. 1–13, 2020.
5. Craig A Stewart, Timothy M Cockerill, Ian Foster, David Hancock, Nirav Merchant, Edwin Skidmore, Daniel Stanzione,James Taylor, Steven Tuecke, George Turner, et al.2015. Jetstream: a self-provisioned, scalable science and engineer-ing cloud environment. *In Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*. 1–8.
6. D. E. Wood, J. Lu, and B. Langmead, “Improved metagenomic analysis with Kraken 2,” *Genome biology*, vol. 20, no. 1, p. 257, 2019.