Visualizing metagenomic data in R using Jetstream

Haley Leffler[1], Bhavya Papudeshi[2], Sheri A. Sanders[2]
[1]Department of Human Biology, Indiana University, Bloomington, IN,
[2]National Center for Genome Analysis Support, Pervasive Institute of Technology, Indiana University, Bloomington, IN

Introduction

- After the 2010 BP oil spill, it was estimated that half of the oil was removed from the ocean by physical and chemical methods and the other half was unaccounted for at the time.
- Metagenomic techniques allow researchers to study microbial communities, revealed their ability to degrade oil or hydrocarbons.
- Data for this project was downloaded from a research study focusing on the microbial response to hydrocarbon seepages (the natural release of oil or gas bubbles from the ocean floor shown in Fig 1). This study found that hydrocarbon seepages can significantly alter the microbial community.
- Researchers applying metagenomics techniques generate large datasets to capture the complexity of the microbial community and face challenges during visualization. Using several visualization methods to display the data from multiple samples allows for exploratory analysis and represents the data from different points of view.
- **This project performs an exploratory analysis of metagenomes using Jetstream, a cloud-based infrastructure that aids analysis of large datasets.**

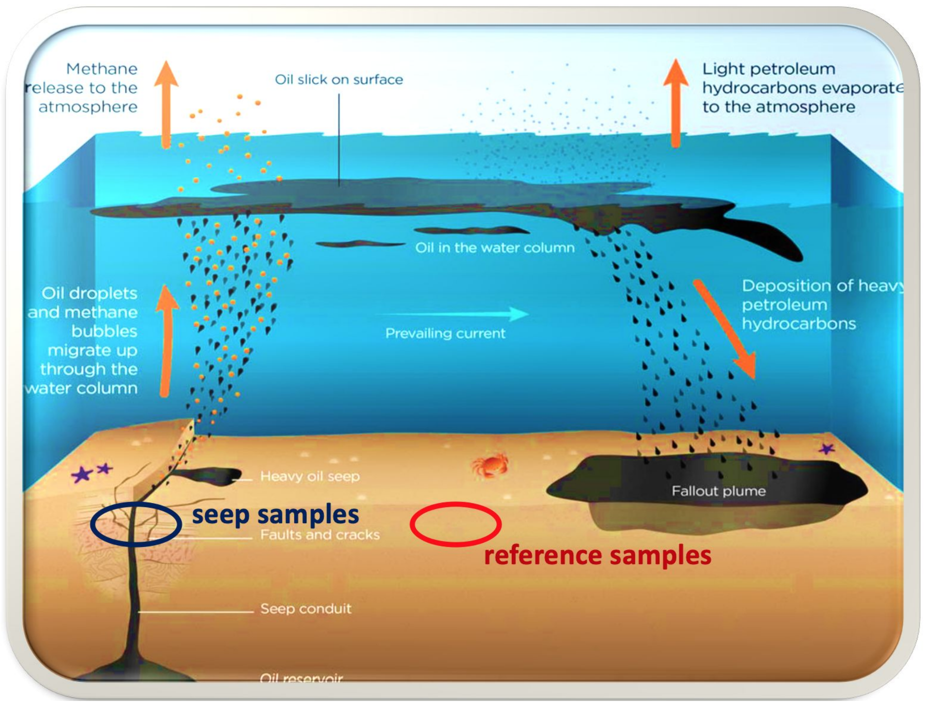

Figure 1: Representation of the hydrocarbon seep environment where the seep and reference samples were likely collected for this study.

Methods

Materials and Data Collection

- The metagenomic dataset (BioProject number PRJNA553005) was downloaded from the Sequence Read Archive (SRA) using sratoolkit.
- The dataset was analyzed on an Ubuntu-based Jetstream virtual machine (CPU: 6, Memory: 16 GB, Disk: 60 GB) with Jupyter notebook (5.2.2) and R (4.0.2) installed.

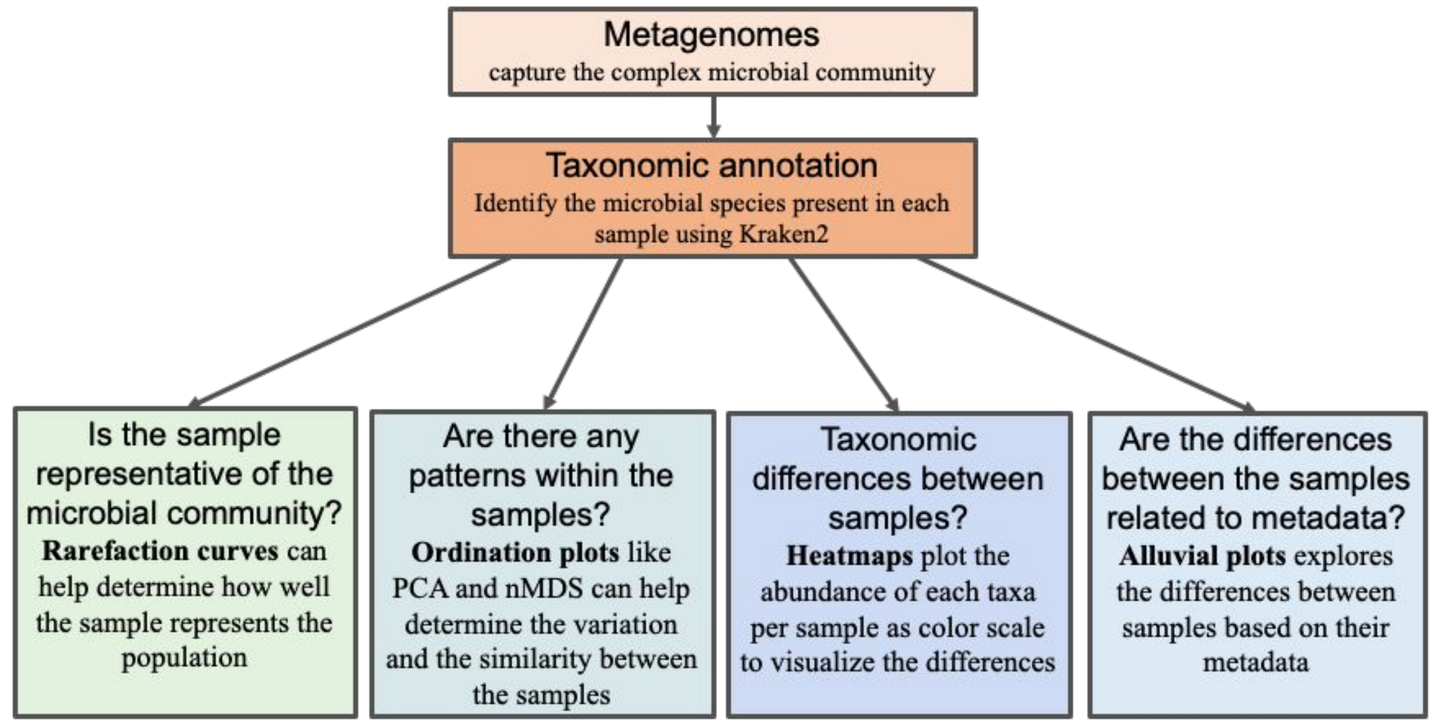

Figure 2: Metagenomes were first annotated to identify the taxa per sample. Taxonomic differences were then explored using four different visualization methods

Results

- Samples were collected from three sites: site 1 (seep1 - D24, D27, and D23), site 2 (seep2 - D72 and D75), and reference (D21 and D30). D24 contained characteristics from all three sites and was labeled as a transition sample.
- Kraken taxonomic reports revealed that 88-90% of the sequences remain unidentified per sample. Regardless of the level of classification, identified taxa have low (<91%) abundance in total.
- From the rarefaction curves (Fig 3) we determined that the metagenomes collected were representative. The samples begin to plateau when 475 unique species were identified per 1.5 million sequences.
- In the two ordination plots (Fig 4) the seep samples and reference samples cluster together respectively, showing higher correlations of the microbial profiles between them.
- The heatmap (Fig 5) showed that *Homo sapiens* sequences are most abundant, likely due to human contamination, and there are other species-level differences between seep and reference sites.
- Looking further into these taxonomic changes in the alluvial plot (Fig 6), the transition sample (D24) has a similar taxonomic profile to both the reference and the seep samples.

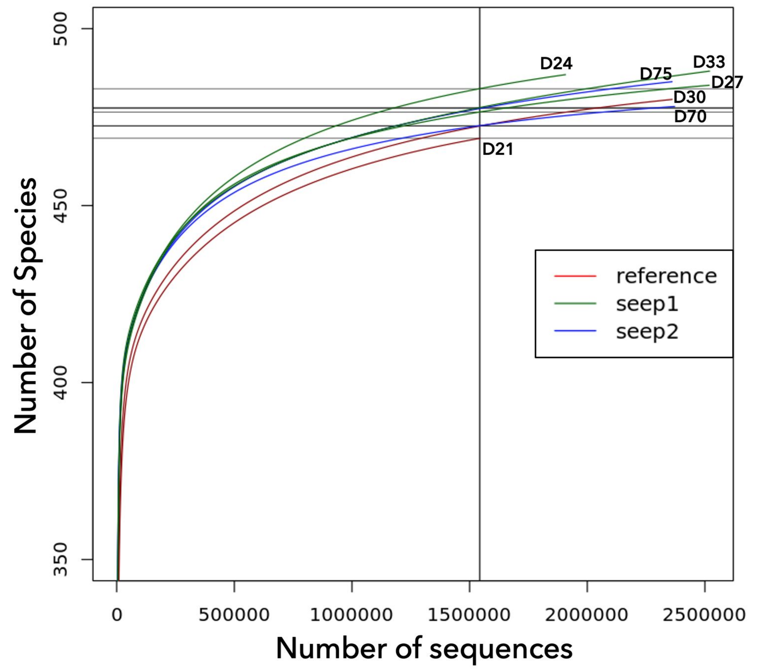

Figure 3: The rarefaction curves exponentially increase as new species are identified, then plateau as no new unique species are identified. Number of species is represented by the y axis and number of sequences is represented by the x axis.

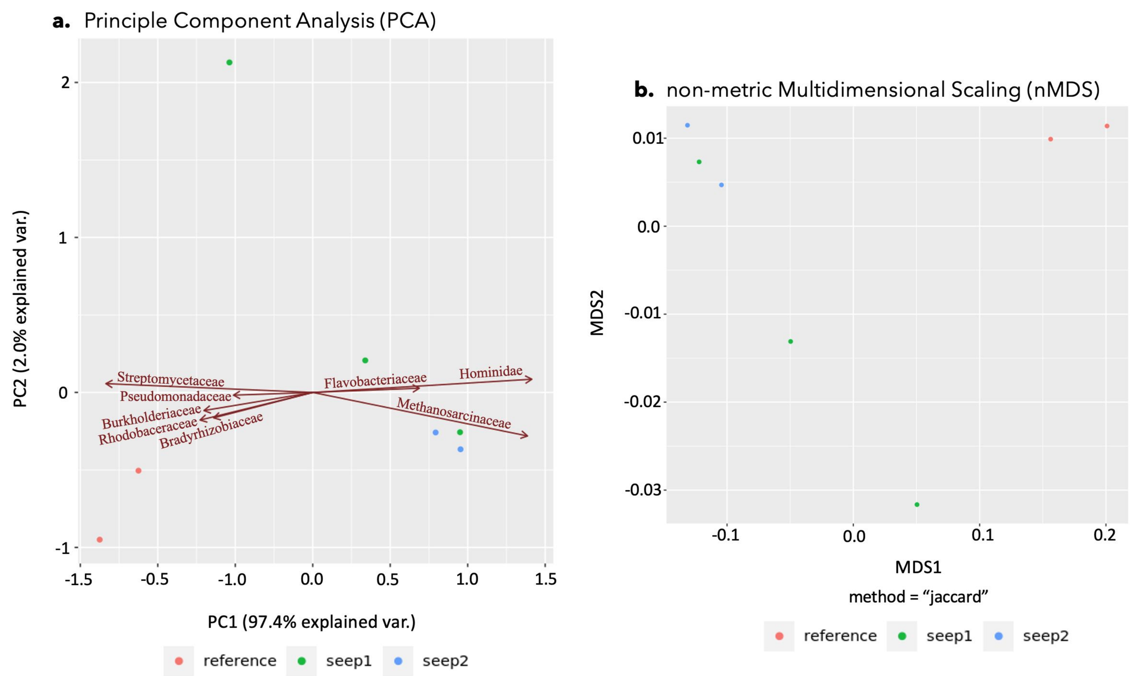

Figure 4: Ordination plots of sediment metagenome data and taxa families. (a) The two axis on the PCA explain 99.4% of the variation. Vectors (arrows) indicate the direction and strength of eight most abundant taxa families to the overall distribution. (b) nMDS of the sediment metagenomes using jaccard dissimilarity metric.

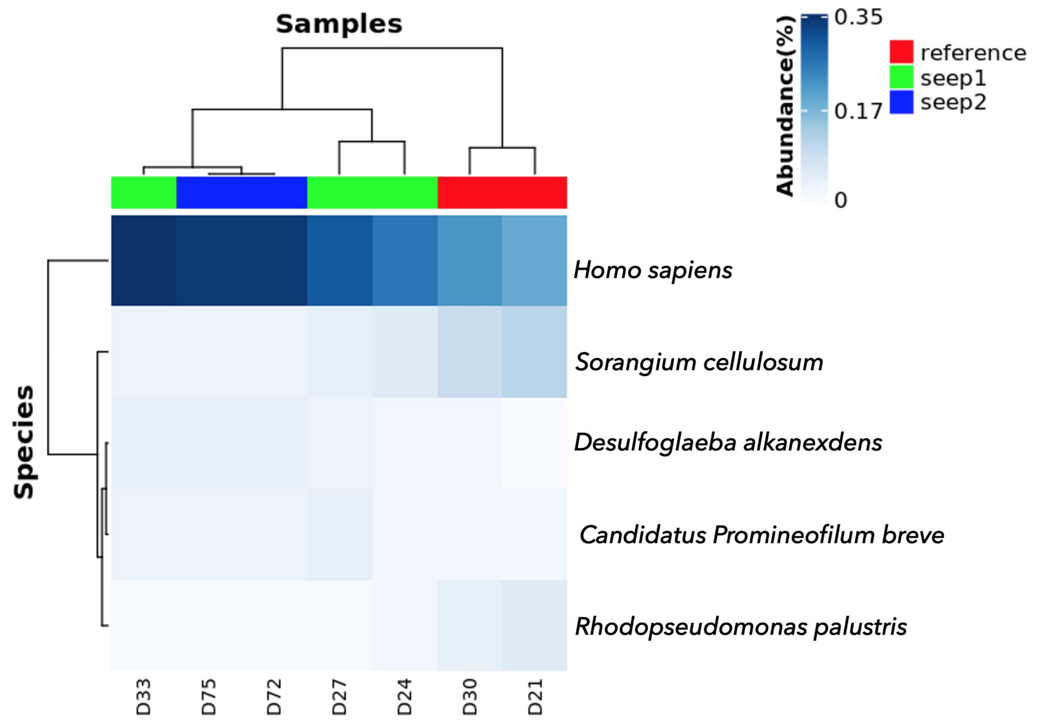

Figure 5: The heatmap shows the percentage of abundant species identified (0.1% - 0.35%) indicated by the color scale (max - dark blue, min - light blue). The dendrograms on the left and top of the heatmap indicates the relationships between the samples and species respectively.

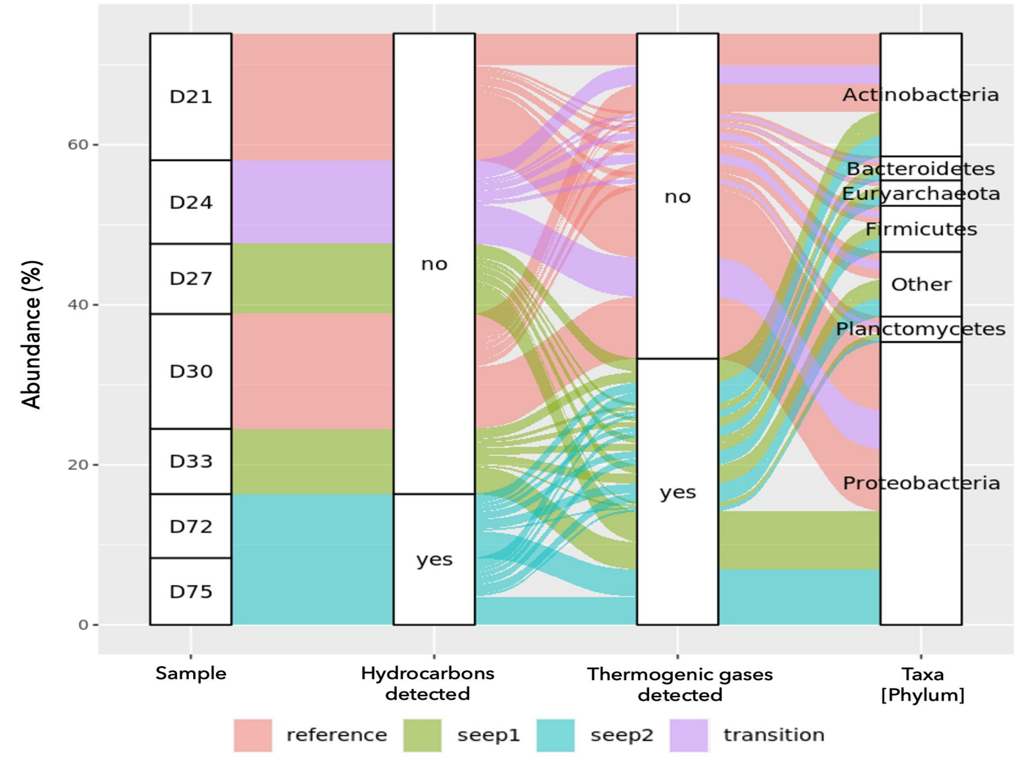

Figure 6: The alluvial plot shows differences between samples based on their metadata. Presence/absence of hydrocarbons and thermogenic gases is shown by the two middle columns. The last column represents the percentage of the six most abundant phylum (out of 46) with the remaining binned to the category Other.

Discussion

- After exploratory analysis of the hydrocarbon seepage datasets, our conclusions align with Zhao *et al.* 2020, and reveal more about the samples' representativeness and contamination.
- While the visualization methods perform exploratory analysis exploring different questions, each method has their strengths and limitations that should be considered. This is summarized in Table 1 below.
- The scripts written to make the plots and input data are available on GitHub in a Jupyter notebook for public use: https://github.com/hleffler/Microbial-visualization.
- Exploration of the structure and function of sediments metagenomes from hydrocarbon seepages can aid in future environmental disasters

Table 1: Strengths and limitation of the four visualization methods explored in this study.

Plot Type	Key Features	Limitations
Rarefaction curve: Allows calculation of species richness from sample	- Determine if sample is representative of a larger population - Easy to read	- Need a large sample or samples from the same community - Does not show specific species, just # - No specified # of samples to enter - No stats
Ordination plots (PCA and NMDS): 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, 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 - No stats

Future Directions

- Perform a quality control step to remove the human DNA contamination before further analysis
- Explore network plots to show sample interconnectedness of taxa between samples
- Apply other metagenomic pipelines to identify functional roles of these taxa within the samples
- Apply statistics to test the significance of the differences observed in the visualizations

Acknowledgements
This material is based upon work supported by the National Science Foundation under Grant No. 1445604 (Jetstream) and 1759906 (NCGAS). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Special thanks to Seth Adams and Dakila Ledesma's neural network resources, Jefferson Davis, Carrie Ganote and Laura Huber's for help setting up and debugging our neural networks, the Jetstream system administration team (Steve Bird, Mike Lowe, and George Turner), and Bhavya Papudeshi for guidance with Jetstream.