BIOL 432

Stafford Maracle 20005033
Shannon Edie 20008148
Logan Lafreniere 20017084
Yukthi Wickramarachchi
Katy Waters 10155094

Bio ‘R’ Us

Final Project Presentation

Outline

Overview

Data Description

Data Analysis

Dataset Diversity

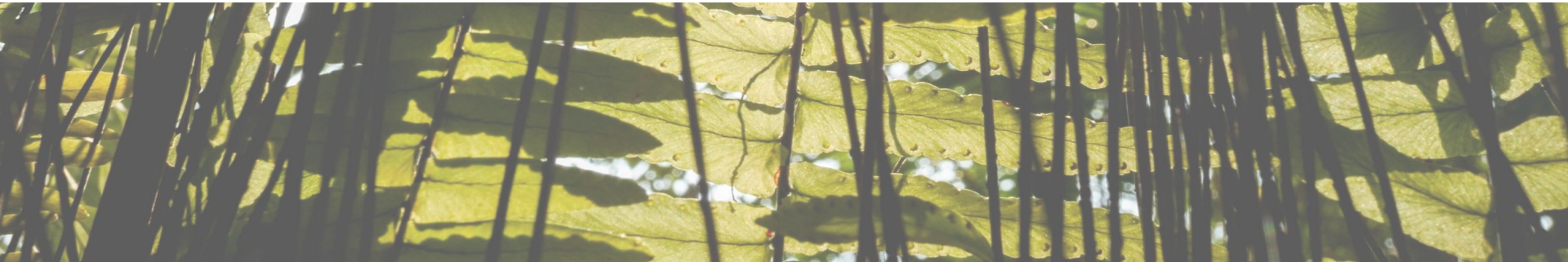
Comparing Datasets

Conclusion

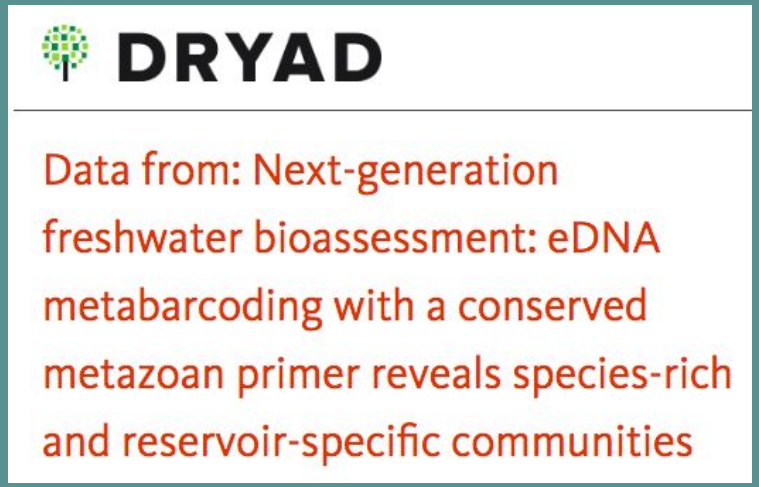
Overview

Environmental DNA (eDNA), or the sequencing of DNA fragments from environmental samples such as from lakes and rivers, has recently become a popular topic in ecology. eDNA has applications in environmental assessment, conservation and monitoring of wildlife, discovery of novel niche species, and more.

Using the techniques we have learned in BIOL 432, we hope to explore an eDNA dataset to identify species, investigate species richness, and compare diversity amongst sampled sites.



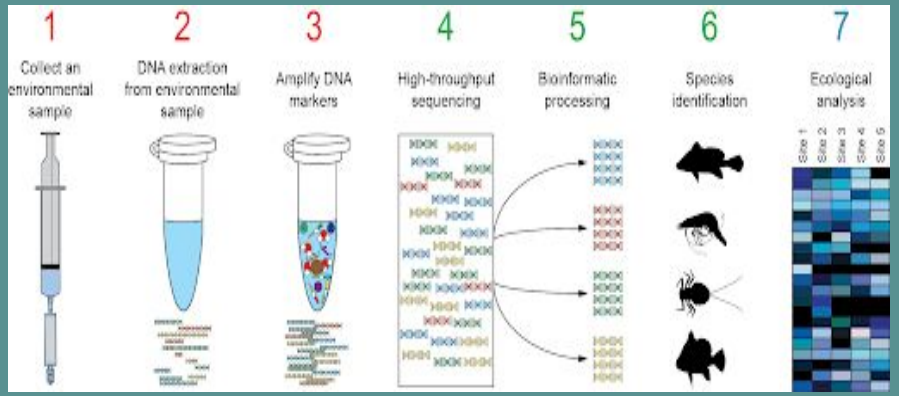
Data Description



- Our data utilizes environmental DNA (eDNA) to analyze species diversity in freshwater ecosystems
- Using data from Lim (2016) obtained from DryadDatasets, our project will consist of water-based environmental DNA (eDNA) samples with universal metazoan primers applied
- 42 samples were collected from each of two freshwater reservoirs.

Data Description

- Environmental DNA (dna) is a cutting edge technique primarily used to assess freshwater ecosystems diversity and health.
- This technique involves isolating and sequencing DNA segments within water samples using DNA primers created to target specific species or our case to target any metazoan species.
- DNA segments tend to below quality due to the nature of eDNA largely composing of detritus or excrement. This leaves us with a data set of 84 fragmented DNA strands in FASTA format (a string of nucleic acid bases) from unknown metazoan species.



Focal Questions

1

What species are we working with?

Working with eDNA and universal primers our nucleotide sequences may belong to any metazoan species. Narrowing down each sample to the family or genus level, with hopes to the species level, will allow us to broadly determine the diversity of species among our samples. We will do this using BLAST and Rentrez in R to compare unknown sequences with NCBI GenBank database.

2

How diverse is our dataset?

This can be done by creating a Shannon index to determine species richness and relative abundance among the sample population. This method may have limitations due to some sequences only being identified at the family level, and with some of our sequences being identified to a species level

3

What is the weighted phylogenetic distribution of our sample population?

Adding to our understanding of the sample populations diversity, we will then question the relatedness of the species in these reservoirs. By creating a weighed phylogenetic tree with the raw FASTA data we can determine the variation in the sequences collected by using eDNA. Another tree may be created using the results from question one to sort samples according to determined taxonomy.

4

What is the distribution, density, and diversity of our eDNA samples found between two reservoirs?

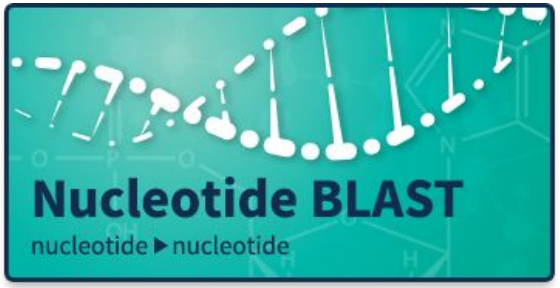
Frequently, eDNA is used for environmental assessment and monitoring. As such, comparison amongst sites is an apt exploration of the powers of eDNA.

Data Analysis

SPECIES IDENTIFICATION

Data Analysis:

Read in and BLAST sequences

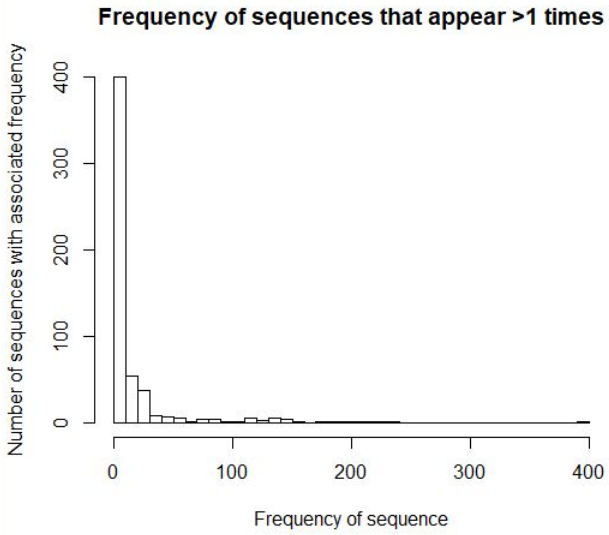


- Our dataset contained 2,155 unique sequences
- All of the sequences were pre-processed (by the authors of the data) to only include Metazoan COI sequences
- We tried to BLAST all sequences, but were unable to due to time constraints
    - Attempted R interface, BLAST+ command line, and BOLD identification
- Instead, we decided to base most metrics for dataset diversity on the sequence count data (using metagenomics approaches)
- We still chose to BLAST the 10 most frequent sequences

Data Analysis:

Analyse BLAST hits

Of our 2155 sequences, 1602 only appeared once. We then looked at the sequences that appeared the most times, and ran the top 10 through BLASTn using the R interface. From these results, we inferred the species based on the read, locational probability, and accuracy of alignment.



Hit_num	Hit_id	Hit_def	Hit_accession	Hit_len
1	gi|331136926|gb|JF844023.1|	Lepidoptera sp. BOLD:AAP9232 cytochrome oxidase sub...	JF844023	658
2	gi|296789142|gb|GU686956.1|	Hypena obesalis voucher BC ZSM Lep 22024 cytochrome ...	GU686956	658
3	gi|630053950|gb|KJ379296.1|	Hypena atomaria voucher UASM7089 cytochrome oxidas...	KJ379296	625
4	gi|630052572|gb|KJ378607.1|	Hypena atomaria voucher CNC LEP00052416 cytochrome ...	KJ378607	658
5	gi|558477265|gb|KF491803.1|	Hypena stygiana voucher AYK-06-7273 cytochrome oxida...	KF491803	658
6	gi|331157750|gb|JF854433.1|	Hypena obesalis voucher MM19072 cytochrome oxidase ...	JF854433	658
7	gi|331157748|gb|JF854432.1|	Hypena obesalis voucher MM19071 cytochrome oxidase ...	JF854432	658
8	gi|326369998|gb|JF415790.1|	Hypena obesalis voucher BC ZSM Lep 21834 cytochrome ...	JF415790	658
9	gi|213984498|gb|FJ412699.1|	Hypena abalienalis voucher UBC-2007-0133 cytochrome ...	FJ412699	658
10	gi|630049692|gb|KJ377167.1|	Hypena atomaria voucher BL456 cytochrome oxidase sub...	KJ377167	658

Data Analysis:

Identify Sequences

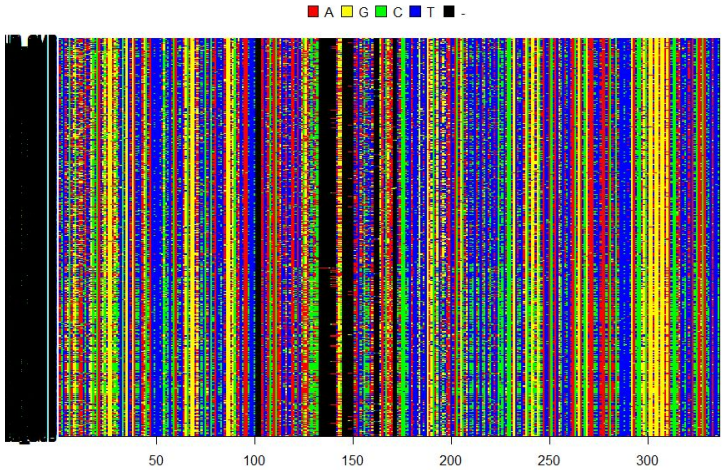
Sequence	Count	Suspected species from BLAST results
1	395	Hypena obesalis
2	232	Plationus patulus macracanthus
3	227	Polyarthra remata
4	222	Bachelotia antillarum
5	211	Monosiga brevicollis
6	209	Amathia vidovici
7	195	Euchlanis dilatata
8	183	Bachelotia antillarum
9	174	Brachionus rubens
10	157	Polyarthra remata

Several sequences appeared to be *Bachelotia* genus, which is an algae. This is interesting as the authors of the paper said their data was filtered to focus on Metazoan sequences only. We do not know if they mistakenly included such sequences, or if our BLAST identification could have been tweaked to be improved.

Data Analysis:

Align data and construct phylogeny

- We aligned the data using MUSCLE and from this alignment constructed a phylogenetic tree of all of our samples using neighbor-joining
- We considered selecting representative sequences from higher-level monophyletic groups to sequence and see if we could get a better image of species diversity at the phylum level but unfortunately, the construction of the “phylo” object in R does not appear to easily permit tree-cropping

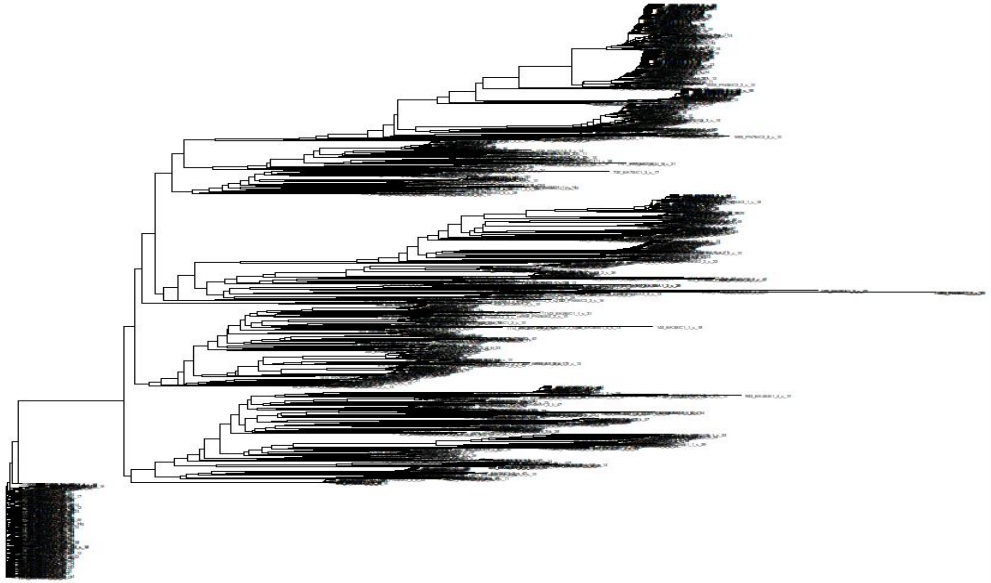


Snapshot of MUSCLE alignment of sequences.

Data Analysis:

Align data and construct phylogeny

- Phylogenetic tree was constructed from the MUSCLE alignment using neighbor-joining based on a DNA distance matrix.
- Suspicion that the two apparent “clades” may represent the Metazoan and non-Metazoan sequences (unfortunately unconfirmed due to insufficient BLAST data)



Constructed phylogeny of 2,155 unique sequences found at all of the collective sites.

Dataset Diversity

Data Diversity

In order to analyze the species diversity of our dataset at each collection site, the first approach that we took was to run a BLAST analysis on all of our sequences. Unfortunately, due to the size of our dataset and computational limitations this was not possible, thus we took a new approach to attempting to estimate the diversity of our dataset. We filtered our dataset for unique sequence reads which had been sequenced more than 5 times, this attempted to account for differences in gapping that created “one-off” sequence reads. This approach allowed us to assume that each unique read was a distinct species of Metazoa, and resulted in a total of 258 unique reads. We then calculated the Shannon-Index value of diversity for each collection site. To compare this to our full dataset we also calculated the Shannon-Index value of diversity using our full dataset to see whether our method of filtering out less abundant sequences had an impact.

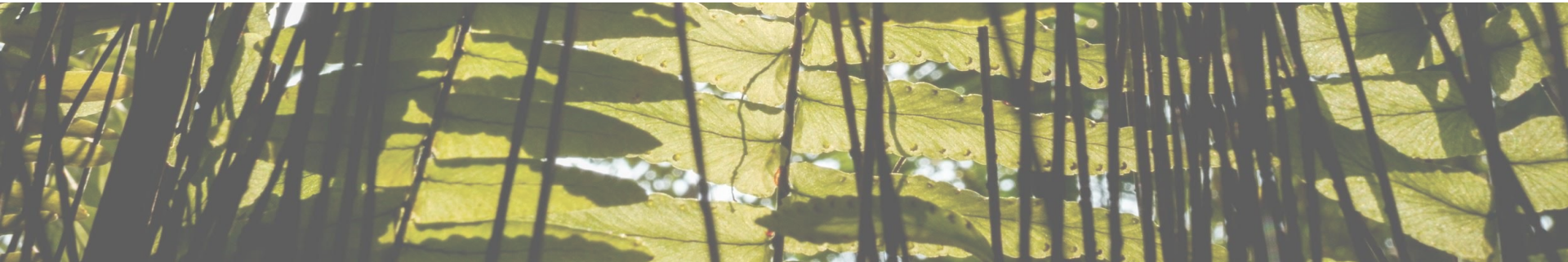
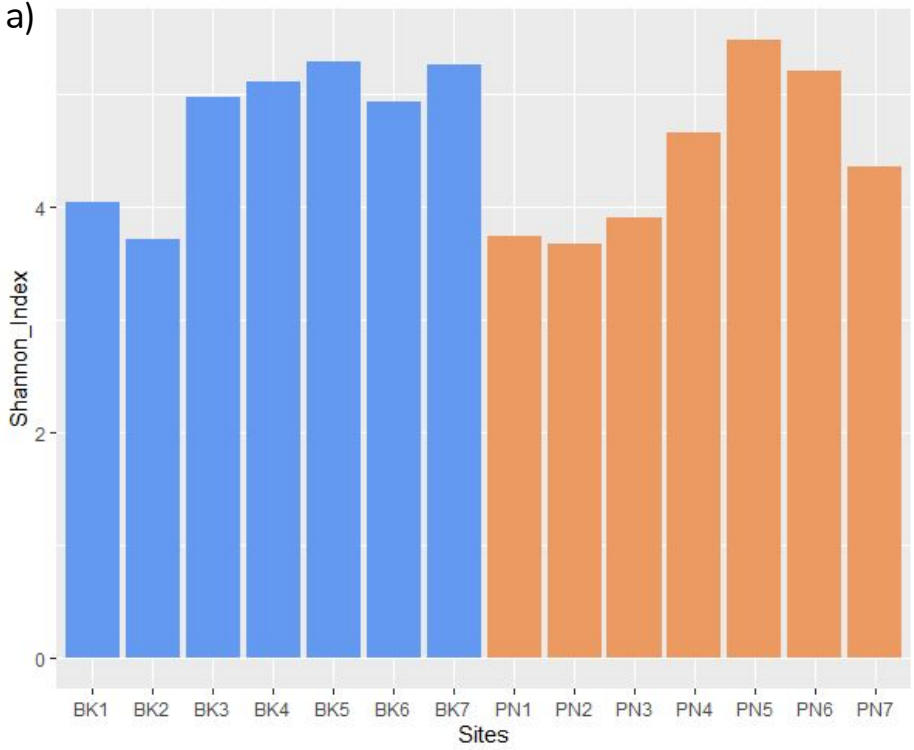
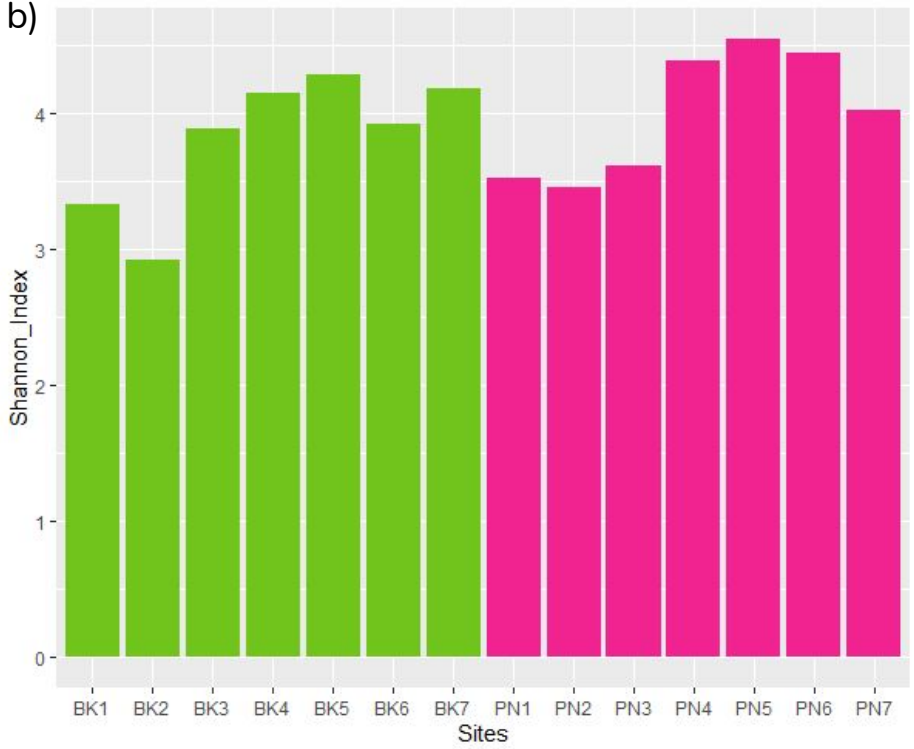


Figure 1. Shannon index calculated on raw sequence data (a) and filtered common sequence data (b)



Shannon-Index Values for Full Data Set



Shannon-Index Values for Filtered Data Set

Diversity Discussion and Limitations

After calculating the Shannon-Index values for each collection site, it can be observed that the species diversity in each site is relatively similar, with the BK sites having an average Shannon-Index value of 3.81 and PN sites having an average value of 4.00. Interestingly though, when using the full dataset, the BK sites have an average Shannon-Index value of 4.76 while the PN sites have an average value of 4.43. This might suggest that the BK site collections may have had more uncommon species which were only sequenced a few times contributing to its higher diversity score in the full dataset in comparison to the filtered one.

This difference in relative H score between the two datasets brings us to the largest limitation with our approach to analyzing the species diversity of our dataset. As mentioned, due to our inability to BLAST our dataset, we had no way of confirming how many different species were truly present in the eDNA dataset that was provided by Lim et al. (2016). Our approach of using each unique DNA sequence with more than 5 reads may have been effective at filtering out most of the repeated sequences, it most certainly wasn't perfect and may have also left out the sequences of species which may have been more uncommon (which would affect our diversity calculations as well.

Comparing Datasets

Comparing Datasets

- Use metagenomics techniques to compare the reservoirs
- We created a counts table of the numbers of each sequence per sampling location and depth. In total, there were 2 reservoirs with 7 sampling sites each, and at each site, samples were taken at 2 different depths. Thus, we had 28 sampling locations to compare
- From the counts data, we generated an NMDS
- It appeared that the reservoir explained most of the variation between samples, and site number explained some of the variation between some samples. Depth did not appear to have a large impact on species variation between sites.

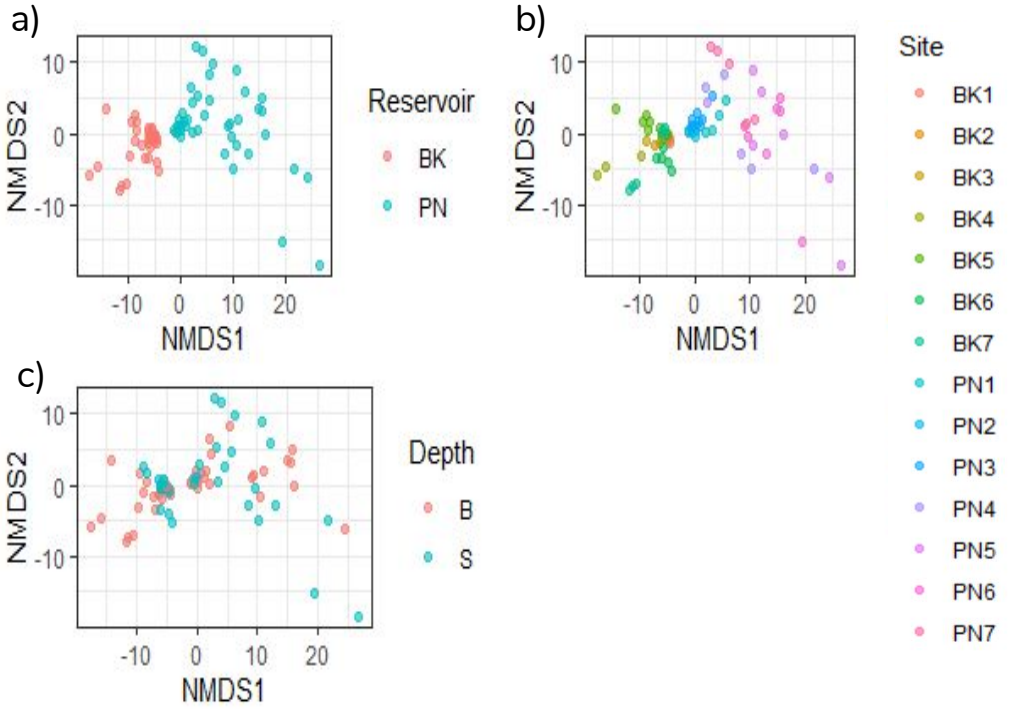


Figure 3. Non-metric multidimensional scaling (NMDS) plots grouped by: (a) Reservoir, Bedok (BK) and Pandan (PN) depicted by colour. (b) Site, 7 from each reservoir depicted by colour. and (c) Depth, benthic (B) and surface (S) depicted by colour.

Metagenomic Tree

Using a Hierarchical cluster analysis we were able to create a metagenomic tree showing the relative relatedness of reads found in each sample site (Figure 2). The tree shows sequences from Bedok clustered together and similarly did Pandan. Clustering is seen within each sample site with slight dispersal within each reservoir and little to none between reservoirs.

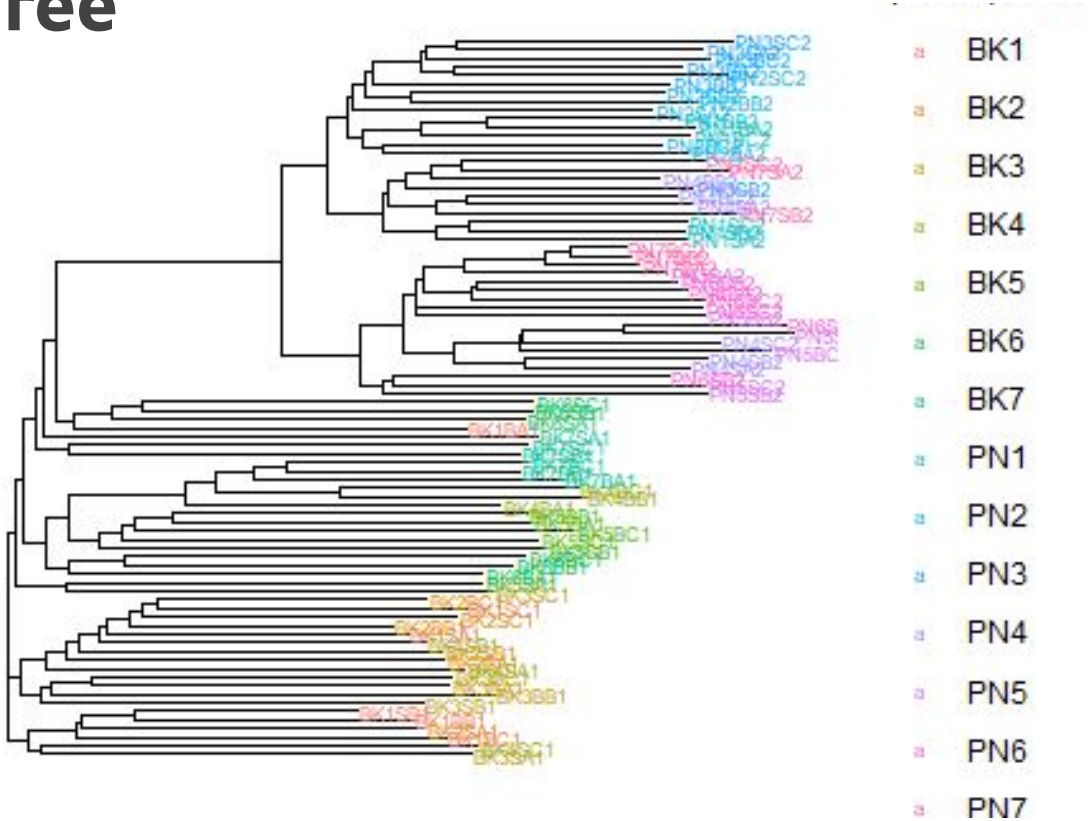


Figure 2. Metagenomic breakdown of samples from reservoirs Bedok (BK) and Pandan (PN). Tree depicts relatedness of reads found within each sample. Each of the 14 sites (7 from each reservoir) were grouped by colour.

Discussion

Discussion/Conclusion

1

What species are we working with?

Using Blastn we were able to run each sequence against the NCBI database. We were able to identify ## of unique species from ## sequences. With the use of private or local databases more accurate results may have been obtained,

2

How diverse is our dataset?

Using a Shannon-index for both raw and filtered data we were able to roughly determine the how diverse the reads were for each reservoir (Figure 1). Through this analysis we found that samples in Bedok reservoir had more uncommon reads and thus a more species rich reservoir. Again, approaching this analysis with identified species rather than with raw sequence, stronger results may have been obtained.

3

What is the weighted phylogenetic distribution of our sample population?

Using a Hierarchical cluster analysis we were able to create a metagenomic tree showing the relative relatedness of reads found in each sample (Figure 2). From this analysis we were able to determine common reads found in each reservoir. Further analysis could be done to show relatedness at the depth level with less clean result. With the use of identified species reads, stronger results may have been obtained

4

What is the distribution, density, and diversity of our eDNA samples found between two reservoirs?

According to the NMDS breakdown of the metagenomic data, the most significant variation between samples was due to the reservoir the sequence was sourced from. The site within each reservoir also explained some of the variation in the sequences, but to a lesser extent. The depth of the sample did not appear to explain much of the

Further Directions

1

BLAST all sequences (or possibly run with BOLD)

Due to time constraints, we were not able to BLAST all of the unique sequences in our dataset. Being able to BLAST all sequences could give us a better understanding of the diversity present, including a broader overview of the various phyla represented.

2

Increase replicate count per each sampling location

The data tables we were working with were rather sparse, which could be augmented either by taking more samples (there were three samples per sampling location/depth), or perhaps by using a different DNA extraction technique. Note that this further direction is not on the informatics side of study design.

3

Further investigate the possible presence of plant sequences in the metazoan dataset

When evaluating our BLAST results, we found the presence of several species of seaweed; theoretically, however, all of the sequences in our dataset were supposedly only Metazoan. Additionally, our phylogeny indicated a significant difference between two groups of our sequences. It may be valuable to further investigate the integrity of the data.

4

Incorporate more geographic data into the metagenomic analysis

Our current comparisons of site, depth, and reservoir metagenomics did not take into account the geographic distance between each site (particularly within the same reservoirs). Having this auxiliary information could help us conduct more thorough investigations into the similarity between sites, and whether geographically close sites

Thank you!

- Bio ‘R’ Us