Your Title Here

Kory Melton

November 18th, 2019

# Introduction

Microscopic organisms are among the oldest and most prevalent lifeforms on earth. Accordingly, they play a vital role in the health of functioning of ecosystems by affecting processes like carbon dioxide and nitrogen cycling (Martiny *et al.*, 2006). As the worlds climate changes over time understanding how microbes influence the dynamics of ecosystems will be necessary to preserve the environment. Currently, the impacts of climate change on the biogeochemical cycling of microbes are not well understood (Barnard *et al.*, 2005).

# Methods

NEON collected and sequenced the samples from the UNDE site, all data can be found in data product DP1.10108.001 (<https://data.neonscience.org/data-products/DP1.10108.001>) (NEON). I analyzed the samples which were formatted and uploaded by NEON to make inferences on the microbial sequences present in each sample over time.

## Sample Collection

Each site contains ten plots that are further sampled one to three times per year. NEON randomly selects three locations for each sample event, with each sampling location is not re-sampled. Soil sampling is collected to a maximum depth of thirty centimeters. After collection, samples are stores in sterile containers, frozen and shipped to a lab for downstream analysis. In the lab, DNA is extracted and the samples are prepared for high-throughput sequence analysis using primers for 16S sequences. The sequences are delivered to NEON for quality control and acceptance. After acceptance, NEON formats and uploads the files in repositories for public use.

## Analysis

I downloaded all unique 16S files for the UNDE site from the NEON database. Then, I checked each files quality using fastqc. I subsampled each file to make the dataset computationally tractable. I trimmed each file in order to capture high quality reads and converted the fastq files to fasta files. Next, I compared each fasta file to a nucleotide database to make inferences on which microorganisms were present in the samples from the UNDE site. Finally, I used those BLAST results to BLANK!!! FILL THIS IN!!!!

### Downloading Data

I downloaded all unique files for the target gene 16S from the NEON database for the UNDE site. The script to download the data used the R library neonUtilities and the NEON API to download the data product. I used the data product ID DP1.10108.001. In addition, I downloaded the metadata for the raw data to include information on when the data was collected and where the data was collected from.

### Fastqc

I extracted the R1 files from their zipped folders. Then, I ran fastqc on each file to assess their quality and identify any problematic files (Andrews, 2010).

### Subsampling

I randomly subsampled 0.5% of the sequences in each file to make the data computationally tractable. For this, I used the srand function and the random seed 1234.

### Trimming

I trimmed the subsampled files using TrimmomaticSE (Bolger *et al.*, 2014). I used four threads, phred33 base quality encoding, a minimum quality threshold of 5 for both leading and trailing bases, and a minimum length of two-hundred base pairs. Addtionally, I implemented a sliding window with window size eight and required quality of twenty-five.

### BLAST

I used bioawk (<https://github.com/lh3/bioawk>) to convert the subsampled, trimmed fastq files to fasta files for downstream BLAST analyses. I compared each fasta file to the GenBank’s nucleotide database (Benson *et al.*, 2000). I used the BLASTN algorithm, which uses a nucleotide query sequence to search the nucleotide database (Camacho *et al.*, 2009). I used four threads, the ‘10 sscinames std’ format, and returned one match for each sequence. I excluded 2018-09-19\_environmental\_sequence.gi to reduce the number of uncultured or environmental matches.

### Analyzing Blast Data

WRITE OUT WHAT I DID AFTER I DO IT

# Results

## Subsections are ok in the results section too

# Discussion

# Sources Cited

Andrews,S. (2010) FastQC: A quality control tool for high throughput sequence data.

Barnard,R. *et al.* (2005) Global change, nitrification, and denitrification: A review. *Global biogeochemical cycles*, **19**.

Benson,D.A. *et al.* (2000) GenBank. *Nucleic acids research*, **28**, 15–18.

Bolger,A.M. *et al.* (2014) Trimmomatic: A flexible trimmer for illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Camacho,C. *et al.* (2009) BLAST+: Architecture and applications. *BMC bioinformatics*, **10**, 421.

Martiny,J.B.H. *et al.* (2006) Microbial biogeography: Putting microorganisms on the map. *Nature Reviews Microbiology*, **4**, 102.

NEON Data product dp1.10108.001, soil microbe marker gene sequences. *National Ecological Observatory Network, Battelle, Boulder, CO, USA NEON Downloaded on November 17, 2019*.