An Automated Pipeline for Nitrogen-cycling Gene Abundance Analysis

Briana Hackos[1,2], Zoey Werbin[3], Jennifer Bhatnagar[3]
[1]The University of Colorado Boulder, [2]Boston University Bioinformatics BRITE REU program, Summer 2019, [3] Department of Biology, Boston University.

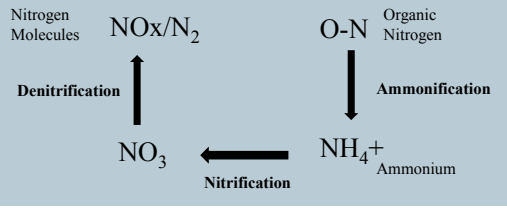
Abstract

Nitrogen (N) cycling is a key ecosystem service carried out by microbes in the soil[1]. Metagenomes, which represent genetic material from an entire microbial community, have been used to infer the rates of processes like nitrification and mineralization, but past studies have been limited by small geographical scales and/or lack of ecosystem diversity[1]. We built a fast and simple pipeline to download and analyze bioinformatic and ecological data from the National Ecological Observatory Network (NEON), which collects yearly data from 81 U.S. sites[5]. We use the R statistical environment, command line tools, and the Sunbeam extensible data pipeline to automate the steps in our metagenomic sequence analysis. We built software extensions to quantify a variety of genes related to the N cycle so that correlations between N-cycle genes and other environmental phenomena can be made. Our pipeline has produced a valuable, accessible dataset that will allow for the identification of continental-scale N-cycle patterns related to agricultural runoff, root exudates, biogeochemical fluxes and more.

Background

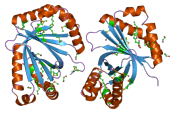
- Humans are adding immense amounts of nitrogen (N) into terrestrial ecosystems through industrial agriculture and fossil fuel combustion[4].
- N pollution poses a threat to public health so it is important to understand how N-cycling microbes - key players in the N cycle – are affected by other ecological phenomena
- N-cycling bacteria contain genes that indicate the ability to transform N, but links between specific N-cycling genes and other environmental processes have not been solidified across ecosystems.
- We developed a data analysis pipeline to explore relationships between N-cycling genes and available NEON data.

Nitrogen Transformations



Questions:

- Can our pipeline accurately profile nitrogen cycling genes within metagenomic data?
- Do N-cycling genes vary across different environments?



Methods

Data

- This study used metagenomic data from soil cores from 5 different NEON sites (highlighted in red)
- Metagenomic data can be coupled to NEON ecological data allowing for analysis on a large spatial scale and a wide temporal scale.
- In the future, NEON will release 2017 metagenomic data from 1000+ soil samples across the United States

Figure 1. Visual representation of NEON sites across the United States. Sites used in this study are in red.

Metagenome Bioinformatics Pipeline

NEON provides plant, animal, soil, nutrient, and freshwater and atmospheric environmental data for samples

↓

Ecological data is accessed through NEON provided R packages

↓

NEON sequences microbes from samples to produce shotgun metagenomic **reads** (gene fragments)

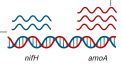
↓

Quality control: Low-quality reads are trimmed and sequencing artefacts such as primers or adapters are removed. Done with Trimmomatic and CutAdapt

Assembly: Reads are assembled into larger fragments, called **contigs**, using overlapping sections of each read. Done with MEGAHIT

↓

Gene prediction: Databases are searched for sequences of known organisms to predict what genes the sequences came from. Done with Prodigal



↓

Gene annotation: Genes are categorized according to their role in the N-cycle using NCycDB

↓

Mapping: Reads are mapped onto genes to quantify the number of genes in samples. Done with Bowtie2 and anvi'o

Results and Conclusions

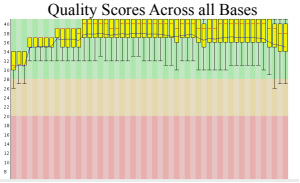


Figure 2. Per base sequence quality of SCBI site raw sequences.

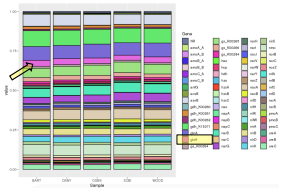


Figure 3. Normalized gene abundance plotted by sample site. 68 genes were annotated by the NCyc gene profiler.

- A Sunbeam report showed NEON fastq files had high per base sequence quality, normal sequence and GC content, and few duplicated and overrepresented sequences
- Accurate gene annotation is attainable with high-quality raw sequences from NEON data
- Normalized relative gene abundances show that N-cycling genes vary slightly among different ecosystems
- Specific genes like glsA (highlighted) present more variation

Future Directions

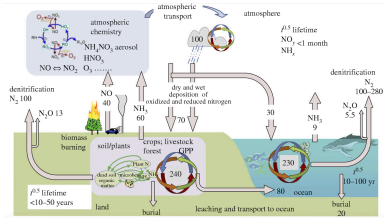


Figure 4. [4]Nitrogen processes in terrestrial and marine ecosystems and in the atmosphere.

- Download 2017 NEON data with nitrification and N-mineralization rates for samples. Analyze relationships between these rates and N-cycling gene presence in data samples.

References and Acknowledgments

[1]Dorthe Groth Peterson, S. J. (2012). Abundance of microbial genes associated with nitrogen cycling as indices of biogeochemical process rates across a vegetation gradient in Alaska. *Environmental Microbiology, Volume 12*, 993-1008.

[2]Magdalena A. Dytczak, K. L. (2008). Activated sludge operational regime has significant impact on the type of nitrifying community and its nitrification rates. *Water Research, Volume 42, Issues 8-9*, 2320-2328.

[3]Qichao Tu, L. L. (2019). NCycDB: a curated integrative database for fast and accurate metagenomic profiling of nitrogen cycling genes. *Bioinformatics, Volume 35, Issue 6*, 1040-1048.

[4]Fowler, David R. et al. "The global nitrogen cycle in the twenty-first century." *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*

[5]National Ecological Observatory Network. 2016. Provisional data downloaded from http://data.neonscience.org 2019. Battelle, Boulder, CO, USA

This work was funded, in part, by NSF grant DBI-1559829, awarded to the Boston University Bioinformatics BRITE REU program.