STUDENT FINAL REPORT TO THE UNIVERSITY OF HAWAI'I AT HILO

MARINE OPTION PROGRAM

Statistics Workflows for Environmental DNA Observational Data

Grant Sanderson

Marine Science Department

University of Hawai'i at Hilo

MOP ADVISOR

Lisa Parr

Marine Science Department

University of Hawai'i at Hilo

PROPOSAL DATE

March 5, 2021

ABSTRACT

Environmental DNA analysis is becoming an increasingly useful tool for ecological monitoring, due to its ability to detect rare species, and determine community composition without taking a direct census. However, the tools and workflows used to analyze eDNA are still in their infancy, due to the novelty of the technique. In particular, consistent statistical analysis of eDNA diversity results is cumbersome, and difficult to grasp for those who do not have extensive knowledge of ecological computing. Therefore, it may be worthwhile to develop a consistent workflow for eDNA statistical analysis, in order to expedite the research process of future studies. As part of an internship with the National Oceanic and Atmospheric Administration (NOAA), a set of code was created using the R statistics platform that easily allows eDNA diversity data to be imported into the platform, performs basic statistical analyses, and provides a framework to conducive to more specialized analyses, should the need arise. The code was created to assist with NOAA's ongoing mission to categorize and predict cyanobacteria blooms in the North American Great Lakes.

Table of Contents

INTRODUCTION

Environmental DNA (eDNA) is a term used to refer to unincorporated genetic material in an environment. This loose genetic material is shed by all organisms, and can be found in most environments. Through the process of collecting and amplifying eDNA, it is possible to identify the organism from which the material originated (Ficetola et al. 2008). During the past decade, it has been found that eDNA can be used to gain deep insight into the ecological state of the environment it was collected from (Deiner et al. 2016). Environmental DNA analysis has been conducted for both terrestrial and aquatic environments, but aquatic environments are typically favored, due to the possibility for eDNA to drift through the water column, away from the source organism (Thomson et al. 2012). As a result, eDNA can be used to detect organisms in aquatic environments from a single water sample, without the need for invasive survey techniques (Uthicke et al. 2018). eDNA analysis has also been shown to be sensitive to rare or secretive species, which may be missed using traditional surveying techniques (Ficetola et al. 2008). eDNA analysis has been shown to be a very accurate method of detecting species in an environment, so eDNA has been a topic of intense study in recent years (Bálint et al. 2018).

Environmental DNA analysis is a very promising tool for marine ecology, both in terms of accuracy of results and in cost-effectiveness. However, the first use of eDNA for an ecological study was only little more than a decade ago (Ficetola et al. 2008). As a result, eDNA analysis is still a very new field of study, with many potential pitfalls (Berney et al. 2004). Previous studies such as a study by Coissac et al. in 2012 have shown that the conclusions made by a study incorporating eDNA may be influenced by variations in the methods and tools used, so it is important to have standard, reusable toolsets for eDNA analysis (Coissac et al. 2012). Additionally, the tools used for eDNA analysis are often cumbersome and difficult to use for those without an extensive knowledge of ecological computing (Larson et al. 2020).

In recent years, advances in amplicon sequencing and ribosomal RNA analysis have allowed entire communities to be studied, rather than single species (Caporaso et al. 2012). Through this technique, it is now possible to measure the species diversity of aquatic habitats with only simple water samples, which is a major breakthrough for ecological monitoring. Using such analyses, it is possible to examine both the alpha diversity (The variety and abundance of species in a habitat) and the beta diversity (The variety and abundance of species between different habitats) from simple water samples, which is a major breakthrough for ecological monitoring (Li et al. 2018). As a result, organizations such as the National Oceanic and Atmospheric Administration (NOAA) have begun to implement eDNA into their ecological research platforms in recent years (Liu et al. 2019). As a result, the demand is high for tools to aid in the process of eDNA analysis, in order to remove some of the difficulty.

One of the most popular tools for eDNA analysis is Qiime2 (Bolyen et al. 2019). This platform uses the DADA2 pipeline to infer complete amplicon sequences from short amplified fragments, create Amplicon Sequence Variants (ASVs) from those full sequences, and give taxonomic assignments to the ASVs. This tool is widely used for processing amplicon data, because it can detect more variants than other methods, and infer differences as little as one

nucleotide in width (Callahan et al. 2016). This tool is widely used within eDNA workflows at NOAA and elsewhere (Goodwin et al. 2020). However, there are no standard workflows used for statistical analysis of the output produced by Qiime2, which is a gap that must be filled by each new study.

eDNA analysis has been successfully performed in order to detect invasive species in the North American Great Lakes, showing that this method is useful for studying the ecology of the Great Lakes (Klymus 2017). Currently, NOAA and other research institutions are using eDNA to determine the causes of harmful Cyanobacteria blooms in the Great Lakes. These blooms have become a topic of concern for researchers, due to their increasing frequency and severity (Steffen et al. 2012). The algal blooms in the Great Lates are known to be caused by the phosphorus contained within excess agricultural runoff, but the exact mechanisms by which these blooms are occurring are still largely unknown (Smith et al. 2015). As a result, a great deal of research has been focused on the Great Lakes in recent years in order to determine the causes of these blooms. Due to the sensitivity of the technique, it can be used to detect bloom-causing bacteria before the bloom starts, and determine where the bloom will start. Using this information, the stressors that cause the blooms to develop can be identified, and potentially removed (Liu et al. 2020).

The main objectives of this project were to complete a research internship at the NOAA Atlantic Oceanographic and Meteorological Laboratory (AOML), to become familiar with the tools and workflows that are currently used to process and analyze eDNA samples at NOAA, and to create a block of R code providing a standard set of statistical tests and visualizations, including analysis of Sequence length parameters, comparison of Alpha Diversity metrics, and rapid visualization of taxonomic assemblage. The goal in designing this code was to simplify the methods NOAA is using in an ongoing eDNA analysis regarding Cyanobacteria blooms in North America's Lake Erie. The framework that this code provides will hopefully also make the process of designing and performing future eDNA studies easier. Additionally, this framework could potentially allow more ecological studies to be conducted using environmental DNA, thereby reducing the number of invasive ecological studies performed in marine ecosystems.

METHODS

Site Location

The internship was conducted as part of NOAA's Ernest F. Hollings Scholarship program, and was chosen using the Student Scholarship Internship Opportunities online system provided by the Scholarship staff during the spring of 2020. The Internship was originally scheduled to be conducted at the NOAA Atlantic Oceanographic and Meteorological Laboratory in Miami, Florida, 33149. This laboratory is an ideal place to study eDNA, because it has a large genomics team that is dedicated to testing new eDNA analyses and workflows. This team often collaborates with the Monterey Bay Aquarium Research Institute (MBARI) and Scripps Institution of Oceanography for eDNA research (Goodwin et al. 2020).

Due to the COVID-19 global pandemic occurring during the summer of 2020, the decision was made to perform the entirety of the internship remotely. In lieu of physical meetings, correspondence with mentors and NOAA staff was conducted online, and no travel to the internship site was done at any point. This shift to a remote internship was possible due to the fact that the nature of the internship was largely digital to begin with, and that most of the work planned involved datasets that had already been collected and processed. However, planned activities involving methods of eDNA collection and processing were canceled to accommodate the new internship format. As a substitute, these methods were instead discussed during weekly correspondence meetings with NOAA staff members.

Working with NOAA Staff

The internship was performed under the mentorship of several experienced NOAA staff members at the Atlantic Oceanographic and Meteorological Laboratory. Due to the remote nature of the internship, meetings were conducted online, using the Google Hangouts virtual conferencing platform. During the internship period, meetings were held every Monday with the internship mentors to discuss strategy and present work that had been completed. Additionally, there were meetings between all members of the genomics group at AOML on Wednesdays, to discuss current and future work by individual members of the group.

Every two weeks, hours and activities performed during the internship were submitted to the NOAA internship program staff, and stipends were disbursed accordingly. The internship encompassed 40 hours per week for the duration of the 10-week internship

Writing Code in R

The code for this project was written using the R programming language v. 4.0.1 (R, Berkely, California). Amplicon Sequence processing had already been conducted prior to the beginning of the internship, so coding in R could begin as soon as the internship began. The data were processed using the command line tools Qiime2 and DADA2, which are some of the most popular tools currently used for eDNA analysis.

The data output from Qiime2 are bundled into Qiime2 Artifacts, which are the default output from Qiime2. Therefore, the Qiime2R package for R was used to import the data into R sessions (Bisanz 2018). The code was written to import Qiime2 artifacts using automatically repeating “for loops” in R. These functions repeat for each entry in a list, which in this case, is a list of the folders containing sample data. As a result, as long as the file structure for the data stays consistent, the code is able to import as many data factors as is required for statistical analysis. This also serves to make the code re-usable for different ecological studies, as long as the amplicon sequence data are processed using a Qiime2/DADA2 workflow. Representative Sequence analysis was conducted using R’s built-in statistical tools, while the diversity analysis made use of the Phyloseq package for R. This package has been used widely for microbiome and eDNA analysis, as it provides many useful tools for genetic analysis in R (McMurdie and Holmes 2013).

Testing R Code

The code created during the course of this internship was not designed to be used with any particular data set, but the code was tested using 16S rRNA amplicon data collected from the western basin of Lake Erie in 2018 as part of NOAA's goal to understand Cyanobacteria blooms in the Great Lakes. The data had already been processed using Qiime2 and DADA2 prior to the beginning of the internship. The data were formatted as an Amplicon Sequence Variant (ASV) Table, which contains each individual sequence after they had been inferred by the DADA2 pipeline. The versatility of the code was tested by comparing the overall alpha diversity, number of ASVs assigned to specific taxa, and mean sequence length across several factors changed during processing in Qiime2 and DADA2. The code was found to perform satisfactorily, producing results that were similar to those of analyses performed using other tools and workflows in place at NOAA.

DISCUSSION

Broader Impacts

Environmental DNA has been used for ecological studies for little more than a decade, so most of the tools and workflows used for eDNA analysis in current studies are largely experimental (Deiner et al. 2014). Due to their novelty, these tools are often difficult to understand for those without an extensive knowledge of ecological computing, and they are cumbersome to implement in a study for researchers of any knowledge level. As a result, tools for eDNA analysis that are more streamlined will allow more studies to implement this technique in their workflows. Increased use of eDNA in scientific studies will likely reduce the number of invasive studies being performed in marine environments, and increase the accuracy and sensitivity of new studies (Deiner et al. 2016).

Future Directions

NOAA's research into the harmful cyanobacteria blooms in the North American Great Lakes is still ongoing, and eDNA will continue to be used for this purpose. As the field of metagenomic research continues to evolve, the tools used to analyze eDNA data for ecological purposes will also continue to evolve. As a result, new eDNA analysis workflows will continue to be developed. The code that was written for this project will hopefully be used as a framework for some elements of these future workflows.

REFERENCES

Bálint M, Nowak C, Márton O, Paulis SU, Wittwer C, Aramayo BJL, Schulze A, Chambert T, Cocchiararo B, Jansen M (2018) Accuracy, limitations and cost-efficiency of eDNA-based community survey in tropical frogs. Mol Ecol Resour 18:1415-1426.

Berney C, Fahrni J, Pawlowski J (2004) How many novel eukaryotic 'kingdoms'? Pitfalls and limitations of environmental DNA surveys. BMC Biol 2:13.

Bisanz JE (2018) Qiime2R: Importing Qiime2 artifacts and associated data into R sessions. Version 0.99, 13.

Bolyen E et al. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37:852-857.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R (2012) Ultra-high throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J 6:1621-1624.

Coissac E, Riaz T, Puillandre N (2012) Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21:1834-1847.

Callahan BJ, McMurdia PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP (2016) DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods 13:581-583.

Deiner K, Walser J, Mächler E, Altermatt F (2014) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. Biol Conserv 183:53-63.

Deiner K, Bik HM, Mächler E, Seymour M, Lacoursiere-Roussel A, Altermatt F, Creer S, Bista I, Lodge DM, de Vere N, Pfrender ME, Bernatchez L (2016) Environmental DNA metabarcoding: Transforming how we survey animal and plant communities. Mol Ecol 26:5872-5895.

Ficetola GF, Miaud C, Pompanon F, Taberlet P (2008) Species detection using environmental DNA from water samples. Biol lett 4:423-425.

Goodwin K, Certner R, Strom M, Arzayus F, Bohan M, Busch S, Canonico G, Cross S, Davis J, Egan K, Grieg T, Kearns E, Koss J, Larsen K, Layton D, Nichols K, O'Neil J, Parks D, Poussard L, Werner C (2020) NOAA 'Omics White Paper: Informing the NOAA 'Omics Strategy and Implementation Plan.

Klymus KE, Marshall NT, Stepien CA (2017) Environmental DNA (edna) metabarcoding assays to detect invasive invertebrate species in the Great Lakes. PLoS One 12:e0177643.

Larson ER, Graham BM, Achury R, Coon JJ, Daniels MK, Gambrell DK, Jonasen KL, King GD, LaRacuente N, Perrin-Stowe Tolulope IN, Reed EM, Rice CJ, Ruzi SA,Thiaru MW, Wilson JC, Suarez AV (2020) From eDNA to citizen science: emerging tools for the early detection of invasive species. Ecol Environ 18:194-202.

Li Y, Evans NT, Renshaw MA, Jerde CL, Olds BP, Shogren AJ, Deiner K, Lodge DM, Lamberti GA, Pfrender ME (2018) Estimating fish alpha- and beta-diversity along a small stream with environmental DNA metabarcoding. Metabarcoding Metagenom 2:1-11.

Liu Q, Zhang Y, Wu H, Liu F, Peng W, Zhang X, Chang F, Xie P, Zhang H (2020) A Review and Perspective of eDNA Application to Eutrophication and HAB Control in Freshwater and Marine Ecosystems. Microorganisms 8:1-15

Liu Y, Wikfors GH, Rose JM, McBride RS, Milke LM, Mercaldo-Allen R (2019) Application of Environmental DNA Metabarcoding to Spatiotemporal Finfish Community Assessment in a Temperate Embayment. Front Mar Sci 6:674.

McMurdie PJ and Holmes S (2013) Phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLoS One e:61217.

Smith DR, Wing KW, Williams MR (2015) What is causing the harmful algal blooms in Lake Erie? J Soil Water Conserv 70:27-29.

Steffen MM, Li Z, Effler TC, Hauser LJ, Boyer GL, Wilhelm SW (2012) Comparative Metagenomics of Toxic Freshwater Cyanobacteria Bloom Communities on Two Continents. PLos One 7:e44002.

Thomson PF, Kielgast J, Iversen LL, Møller PR, Rasmussen M, Willerslev E (2012) Detection of a Diverse Marine Fish Fauna Using Environmental DNA from Seawater Samples. PLoS One 7:e41732.

Uthicke S, Lamare M, Doyle JR (2018) eDNA detection of corallivorous seastar (*Acanthaster cf. solaris*) outbreaks on the great barrier reef using digital droplet PCR. Coral Reefs 37:1229-1239.