Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Applications Note

OXFORD

Databases and ontologies

met: Expanding on old estimations of biodiversity from eDNA with a new software framework

David C. Molik [1,*], Caroline DeVoto [2,‡] and Daniel Molik [3,‡]

[1]Navari Family Center for Digital Scholarship, University of Notre Dame, Notre Dame, 46556, United States of America
[2]Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, 46556, United States of America
[3]Independent Researcher, Buffalo, 14226, United States of America

*To whom correspondence should be addressed.
‡These authors also contributed equally to this work.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: A long standing problem in Environmental DNA has been the inability to compute across large number of datasets. Here we introduce an Open Source software frame work that can store a large number of Environmental DNA datasets, as well as provide a platform for analysis, in an easily customizable way. We show the utility of such an approach by analyzing over 1400 arthropod datasets.

Results: This article introduces a new software framework, met, which utilizes large numbers of metabarcode datasets to draw conclusions about patterns of diversity at large spatial scales. Given more accurate estimations on the distribution of variance in metabarcode datasets, this software framework could facilitate novel analyses that are outside the scope of currently available similar platforms.

Availability: All code are published under the Mozilla Public License ver 2.0 on the met project page: doi.org/10.17605/OSF.IO/SPB8V

Contact: dmolik@nd.edu

Supplementary information: Supplementary data are available at met project page online.

1 Introduction

We are approaching the ten-year anniversary of Conservation in a Cup of Water (Lodge *et al.* (2012)), something of a landmark in Environmental DNA (eDNA, a subtype of metabarcode data, for further explanation see Supplemental Wiki Page: What is eDNA?)) describing the use of a fairly new technology at the time, eDNA, which the paper showed could be used to determine biodiversity at a relatively low cost. It is now a cliche to say that we have seen explosive growth in the number of available environmental DNA datasets. Now, computational and methodological technology has been trying to compare samples across large swaths of area and environment (Thompson *et al.* (2017); Pawlowski *et al.* (2018)) and the field is approaching that target. However, the goal of true meta-analysis, loosely defined as combining data from different experiments, has as yet been out of reach, or at the very least extremely time-consuming (Yates *et al.* (2019)). This work attempts to make a first pass at achieving numerous eDNA sample computation as well as showing the ecological benefit of doing so. In order to achieve this target, we introduce "met," an acronym for metabarcode, metagenomic, metagenetic enrichment toolkit. The "met" in met stands in for three words starting in "met", with the e and t standing for enrichment and toolkit, respectively. met is a software framework, utilizing databasing, web frameworks, and just in time compiling, which starts to make a large number of sample comparisons possible. Principally, met stores eDNA data (DOI: 10.17605/OSF.IO/SPB8V).

Meta-analysis in eDNA is difficult due to the lack of standardization across experiments. Differences in preparation of samples and in sequencing can cause slight changes in comparisons of data between different experiments. There are a few ways to tackle this problem: either the field or application of eDNA could enforce more stringent controls on data production (Oliveira *et al.* (2021)), the field could change acceptable reporting standards for metadata (Yilmaz *et al.* (2011)), or as met does, strike a balance between the two. To address the challenges of cross-dataset comparison (for more on eDNA comparison challenges, see Supplemental Wiki Page: "Computational Problems with the Analysis of eDNA") and to increase the speed of analysis, we created met as a framework around which to build analysis solutions. Consisting of three main software repositories, all published Open Source under the Mozilla Public License Version

2.0 (see supplemental project page), the framework is designed to be portable to different compute scenarios. All three components are scalable and continuously integrated as docker containers (for more on containers see Supplemental Wiki Page: "Why Containers?") As a result of met's design, it can simultaneously compare numerous metabarcoded datasets. met achieves this capability through database compression, reorganized database schema, scaling, and a multithreaded web API layer. met can compare thousands of samples from different experiments in a single analysis.

2 Methods

To demonstrate some of the notable features of met, we explore Cytochrome C Oxidase I (COX1) arthropod eDNA samples accessible through the National Center for Biotechnology Information's (NCBI) Sequence Read Archive (SRA). The SRA is part of the International Nucleotide Sequence Database Collaboration (INSDC) that includes data from the European Bioinformatics Institute (EBI) and DNA Data Bank of Japan (DDBJ). We downloaded relevant data sets en masse to determine global arthropod Amplicon Sequence Variant (ASV) diversity (for more on the query used see Supplemental Wiki Page: "The Query"). We loaded 1405 datasets into met to calculate world-wide aquatic COX1 diversity. ASVs are composed of each unique barcode variant found in a sample. This set was pared down manually from an initial 5900 COX1 samples by filtering for only aquatic arthropod samples. To demonstrate the utility of met, we compared all samples by calculating the total diversity of ASVs (see: Fig. 1, panel C) and the cumulative increase of ASVs across samples (see: Fig. 1, panel B). We also mapped the 515 samples that had latitude and longitude information (see: Fig. 1, panel A). Using met, the data retrieval and functions to generate these plots took only a matter of seconds. For more on how met was written, see Supplemental Wiki Page: "The How of met."

3 Conclusions

met is designed to allow for comprehensive analysis of metabarcoded datasets, either in pair-wise comparison of datasets or for the search of specific taxa. This functionality allows for the location of any unique sequence in all previously published metabarcode data. met is adaptable for commonly used microbiome barcodes (i.e.: 16S, 18S) and eDNA barcodes (i.e.: ITS, COX1, ND2). met's scaling ability is achieved through a scaling web server pool, as well as possible database sharding. Met works via met-analysis interacting with met-api and in turn, met-api interacts with met-db (see: Fig. 1, Panel D).

While the specific results from our example generating ASV abundance curves from geographically disparate locations are largely confirmatory, met itself has proven to be an efficient tool for analysis. When the "Conservation in a Cup of Water" paper was first published, the authors were thinking about how biodiversity could be determined in a particular spot, at a relatively low cost. The next logical extension is to take advantage of the power gained by combining data from multiple experiments in this rapidly expanding field in new and interesting ways to increase data utility. This analysis is a way to increase data utility and combine metabarcode experiments. In met we have a way to computationally process large number of samples and we can compare them quickly and come back with useful output, demonstrating that met is a powerful tool for metabarcoding researchers going forward.

Acknowledgements

The authors would like to thank Micheal E. Pfrender and Natalie Mevers

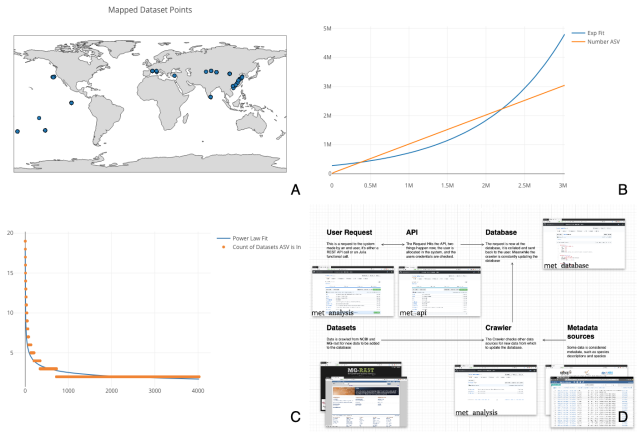

Fig. 1. (A) Map of the 515 samples with latitude and longitude data. Samples tended to tightly cluster around locations, correlating with particular biodiversity assay experiments. (B) Number of sequences found per ASV, sorted by the number of ASVs found. If each ASV was counted across all datasets, it would necessitate a n^2 operation of all sequences compared to all other sequences. Most analysis software have some solution to this all-on-all problem. met overcomes this difficulty by storing ASVs in a separate table so that this operation becomes a 'n' operation of grouping and counting the ASV's associated datasets. The inferred ASV diversity followed an exponential function, with a substantially long tail. (C) Cumulative plot of any particular ASV found across samples. The plot is reverse sorted by count of samples in which the ASV is found. Although it may not look like it to the eye, no single sequence was found in over 20 datasets. (D) A diagram of met's different pieces: met-api is composed of three major components: met-analysis, met-api, and met-db. met-analysis is the main point of entry for the framework. Data gathered by crawlers would be inserted via met-analysis, and data for further downstream computation would come out of met-analysis. met-api is the only entry point for met-db, and met-db contains all information an analysis project may be interested in.

Funding

This research was supported in part by the University of Notre Dame Navari Family Center for Digital Scholarship (https://cds.library.nd.edu/) (NFCDS), the University of Notre Dame Professional Development Zahm Research Travel Grant Fund, and by an AWS Cloud Credits for Research Grant.

References

Lodge, D. M., Turner, C. R., Jerde, C. L., Barnes, M. A., Chadderton, L., Egan, S. P., Feder, J. L., Mahon, A. R., and Pfrender, M. E. (2012). Conservation in a cup of water: estimating biodiversity and population abundance from environmental dna. *Molecular ecology*, **21**(11), 2555–2558.

Oliveira, R. R., Silva, R. L., Nunes, G. L., and Oliveira, G. (2021). Pimba: a pipeline for metabarcoding analysis. *bioRxiv*.

Pawlowski, J., Kelly-Quinn, M., Altermatt, F., Apothéloz-Perret-Gentil, L., Beja, P., Boggero, A., Borja, A., Bouchez, A., Cordier, T., Domaizon, I., *et al.* (2018). The future of biotic indices in the ecogenomic era: Integrating (e) dna metabarcoding in biological assessment of aquatic ecosystems. *Science of the Total Environment*, **637**, 1295–1310.

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., Prill, R. J., Tripathi, A., Gibbons, S. M., Ackermann, G., *et al.* (2017). A communal catalogue reveals earth's multiscale microbial diversity. *Nature*, **551**(7681), 457–463.

Yates, M. C., Fraser, D. J., and Derry, A. M. (2019). Meta-analysis supports further refinement of edna for monitoring aquatic species-specific abundance in nature. *Environmental DNA*, **1**(1), 5–13.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., *et al.* (2011). Minimum information about a marker gene sequence (mimarks) and minimum information about any (x) sequence (mixs) specifications. *Nature biotechnology*, **29**(5), 415–420.