



Database Tools

met v1: Expanding on old estimations of biodiversity from eDNA with a new database framework

David C. Molik^{1,2,*}

¹Navari Family Center for Digital Scholarship, University of Notre Dame, Notre Dame, 46556, United States of America

²Department of Biological Sciences, University of Notre Dame, Notre Dame, 46556, United States of America

Associate Editor: XXXXXXXX

Received on XXXXXX; revised on XXXXXX; accepted on XXXXXX

Abstract

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 framework 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 metabarcode datasets. 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, doi.org/10.17605/OSF.IO/SPB8V. Project pages includes explanation of query used to gather data and software framework code.

1 Introduction

We are approaching the ten-year anniversary of "Conservation in a Cup of Water" (Lodge *et al.*, 2012), a journal article in Molecular Ecology and something of a landmark in Environmental DNA (eDNA, a subtype of metabarcode data) 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 cliché to say that we have seen explosive growth in the number of available environmental DNA datasets for analysis, however this deluge of data requires new methods to analyze it. eDNA analysis, as with much bioinformatics analysis, has not kept up in a way that allows for the comparison of thousands, or tens of thousands of samples. Similarly, computational and methodological technology in the field of ecology has been trying to compare samples across large swaths of area and environment (Thompson *et al.*, 2017; Pawlowski *et al.*, 2018). However, the goal of true meta-analysis, loosely defined as combining data from different experiments, has as of yet been

out of reach, or at the very least extremely time-consuming (Yates *et al.*, 2019). The framework presented here, *met*, attempts to make a first pass at achieving Big Data eDNA sample computation as well as showing the benefit to ecological research 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 an arbitrarily large number of sample comparisons possible. Principally, *met* stores eDNA data, and allows for thousands of pairwise comparisons of samples, or the search of a specific gene through thousands of samples.

eDNA relies on metabarcoding. Like gene barcoding, metabarcoding selects for a gene, but instead the selection is across species (Deiner *et al.*, 2017). The metabarcoding in question should be conserved enough to be in an entire taxonomic group of interest, but different enough in all relevant taxa to tell them apart (Deiner *et al.*, 2017). In effect, this means that a "single cup of water" can determine the diversity of species in an area. Being a relatively low cost method of sampling diversity, a not unexpected

use of the technology has been to determine the total amount of diversity of organisms on our planet (examples of large sampling projects: (Rusch *et al.*, 2007; Turnbaugh *et al.*, 2007; Gilbert *et al.*, 2014). More often, eDNA is used to determine the representative diversity of a given sample of an environment (examples of such projects: (Crits-Christoph *et al.*, 2013; Armitage, 2017). There have been a few efforts to do this, and perhaps most notably has been Knight *et al.* 2012's sampling of the English Channel, which claimed sixty percent representative diversity of the Atlantic Ocean in a single sampling of the channel (Caporaso *et al.*, 2012).

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: require some standardization through data format requirements, while utilizing alignment methods which allow for some effects caused from differences in data analysis methods (Molik *et al.*, 2020). To address the challenges of cross-dataset comparison and to increase the speed of analysis, we created *met* as a framework around which to build other 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. 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 multi-threaded web API layer. *met* can compare thousands of samples from different experiments in a single analysis.

eDNA presents a unique set of challenges in Computational Biology. The fact that eDNA relies on a single gene means that modern alignment algorithms can be bemusingly eschewed for older DNA alignment strategies, which in specific cases, may be faster than modern alignment. Specifically, this means that Levenshtein distance can be used (Levenshtein, 1966; Buschmann and Bystrykh, 2013). Those familiar with alignment strategies will at once notice the similarities between Levenshtein distance and usual alignment strategies—both have cost functions to differentiate between strings, here referring to both text and DNA, and both are used to compare similar but slightly different strings. Genome level alignment necessarily requires comparing many wholly different strings and generally comes with a storage strategy that makes strings easier to compare, for instance suffix trees (i.e., (Weiner, 1973; Delcher *et al.*, 2002)), or de buijn graphs (Compeau *et al.*, 2011). These data structures are not a cheap computational operation to initialize, and generally require expensive computational operations to update the data structure given new strings. If, however, comparing many very similar strings, especially around kingdom metabarcoding cutoffs (e.g. a 450 base pair reference sequence might be considered the same species at ninety-seven percent identity, or 14 base pair differences), searching for a similar sequence would be faster than an alignment if a cut-off was used in Levenshtein (i.e. after so many differences move on to the next string). This assumes that the sequences are in the same orientation. Since *met* is making comparisons against similar sequences Levenshtein can be used and would be faster than a bag of words comparison of k-mers, as both operations would require the complete comparison of all sequences for the detection of small differences. Considering that the *met* use case is to find the most similar sequences, the property of stopping comparison after too many differences in Levenshtein is more desirable.

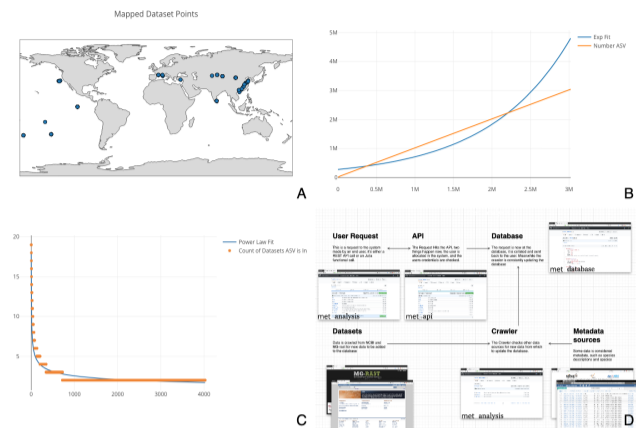


Fig. 1. (A) Map of the 515 samples with latitude and longitude data. Samples tended to tightly cluster around locations, correlating with particular biodiversity assay experiments. (B) Number of sequences found per ASV, sorted by the number of ASVs found. If each ASV was counted across all datasets, 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.

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). 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.

met is written in Perl, Julia, and PostgreSQL PL/pgSQL (PostgreSQL Procedure Language SQL [Structured Query Language]). *met-db* is written as an optimized PostgreSQL schema restoring external datasets. A decreased emphasis on database views and an increased emphasis on efficient database functions written in PL/pgSQL means that the data storage backend is compressed due to the benefits of a database. Writing in this layered approach ensures that *met* components (e.g., Data Storage in PostgreSQL, API as a pass-through layer, and analysis in the API client) are organized as separate entities. This organization method ensures not only the sequestration of code, but that computational resources are

easily partitioned and allocated. The upshot of this structure is that an organization could host a `met-db` and `met-api` install, and utilize grid computing for `met-analysis`. The implementation of `met` for this project was deployed on Amazon Web Services (AWS) Relational Database Service (RDS) on a `db.r4.2xlarge` instance. The component `met-api`, written in Perl using the Dancer framework, was deployed via docker containers to a `t3.large` instance. The component `met-analysis`, written in Julia, was run on the Notre Dame Center for Research Computing (CRC) servers using minimal memory.

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). Furthermore, multiple genes can be utilized in the same of instance of `met` meaning that non-specific shotgun metagenome approaches could be utilized with `met`. `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 `texttmet` is a powerful tool for metabarcoding researchers going forward.

Acknowledgements

The authors would like to thank Prof. Micheal E. Pfrender, Natalie Meyers, and Dr. Matthew Sisk at the University of Notre Dame as well as Daniel A. Molik for his implementation advice

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, and partially by Oracle for Research Grant CPQ-214706.

References

- Armitage, D. W. (2017). Linking the development and functioning of a carnivorous pitcher plant's microbial digestive community. *The ISME journal*, **11**(11), 2439.
- Buschmann, T. and Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed dna sequencing. *BMC bioinformatics*, **14**(1), 1–10.
- Caporaso, J. G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J. A. (2012). The western english channel contains a persistent microbial seed bank. *The ISME journal*, **6**(6), 1089–1093.
- Compeau, P. E., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, **29**(11), 987–991.
- Crits-Christoph, A., Robinson, C. K., Barnum, T., Fricke, W., Davila, A. F., Jedynek, B., McKay, C. P., and DiRuggiero, J. (2013). Colonization patterns of soil microbial communities in the Atacama Desert. *Microbiome*, **1**(1), 28.
- Deiner, K., Bik, H. M., Mächler, E., Seymour, M., Lacoursière-Roussel, A., Altermatt, F., Creer, S., Bista, I., Lodge, D. M., De Vere, N., *et al.* (2017). Environmental dna metabarcoding: Transforming how we survey animal and plant communities. *Molecular ecology*, **26**(21), 5872–5895.
- Delcher, A. L., Phillippy, A., Carlton, J., and Salzberg, S. L. (2002). Fast algorithms for large-scale genome alignment and comparison. *Nucleic acids research*, **30**(11), 2478–2483.
- Gilbert, J. A., Jansson, J. K., and Knight, R. (2014). The earth microbiome project: successes and aspirations. *BMC biology*, **12**(1), 1–4.
- Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.
- 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.
- Molik, D. C., Pfrender, M. E., and Emrich, S. J. (2020). Uncovering effects from the structure of metabarcode sequences for metagenetic and microbiome analysis. *Methods and protocols*, **3**(1), 22.
- 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.
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yoosheph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., and Others (2007). The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS biology*, **5**(3), e77.
- 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.
- Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., and Gordon, J. I. (2007). The human microbiome project. *Nature*, **449**(7164), 804–810.
- Weiner, P. (1973). Linear pattern matching algorithms. In *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, pages 1–11. IEEE.
- 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.