met: A cup of water revisited: expanding on old estimations of biodiversity and population abundance from environmental DNA with a new software framework

David Molik[1,2*], Caroline DeVoto[3‡], Daniel A. Molik[4‡],

1 Navari Family Center for Digital Scholarship, University of Notre Dame, Notre Dame, IN, United States of America
2 Department of Biological Sciences, University of Notre Dame, Notre Dame, IN, United States of America
3 Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN, United States of America
4 Independent Researcher, Buffalo, NY, United Sates of America

‡These authors also contributed equally to this work.
* dmolik@nd.edu

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 datasets.

Author summary

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. The capabilities of met allow the researcher to think about what could be achieved in data reuse and data utility, while the coding of met is especially a practice in solving the research problems that hold back that kind of analysis.

1 met

We are approaching the ten-year anniversary of Conservation in a Cup of Water [1], something of a landmark in Environmental DNA (eDNA, a subtype of metabarcode data, for further explanation see Text Box 1: 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 [2–4] 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 [5]. 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 14 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).

Text Box One: What is eDNA?
eDNA relies on metabacoding. Like gene barcoding, metabarcoding selects for a gene, but instead the selection is across species [6]. The metabarcode gene 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 [6]. 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: [7–9]). More often, eDNA is used to determine the representative diversity of a given sample of an environment (examples of such projects: [10, 11]). 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 [12].

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 [13, 14], the field could change acceptable reporting standards for metadata [15], or as met does, strike a balance between the two. Met allows for a data framework that could enforce some standards and allow for user designed data correction, assuming expert users will have specific data correction analysis methods in mind. This is why met is a software framework and not a traditional web application—it allows for customization by the end user specific to their requirements.

To address the challenges of cross-dataset comparison (for more information on eDNA comparison challenges, see Text box 2: "Computational Problems with the Analysis of eDNA") and to increase the speed of analysis, we created met, designed 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, met-db (https://github.com/molikd/met-db/), met-api (https://github.com/molikd/met-api), and met-analysis (https://github.com/molikd/met-analysis), the framework is designed to be portable to different compute scenarios. All three components are scalable and continuously integrated as docker containers [16], and scalable (for more information on containers see Text Box 3: "Why Containers?") As a result of met's design, it can effectively and 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.

Text Box Three: Why Containers?
Containers for met are stored in the container registry, meaning that while only one version of the met code is "live" or public to the general user base, multiple versions of the code can be stored. This enables a process that when buggy code is accidentally released, the code can be "rolled-back" to earlier versions [17]. Containers have a secondary effect of increasing accessibility of code by making the code base easier to install. In met, there are three major components: a centralized database, a centralized api, and a de-centralized analysis pack. It is easy to see how a user may want to run all three components on their own machine for testing or workstations setups. Through containerization, that can be achieved by installing all three components via containers.

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)(see [18] for common eDNA metabarcode genes). 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. Using the query:

```
"ecological metagenomes"[Organism] AND (CO1 OR COI OR COX1) AND
(cluster\_public[prop] AND "biomol dna"[Properties]) AND
("filetype fastq"[Properties])
```

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. 2) and the cumulative increase of ASVs across samples (see: Fig. 3). We also mapped the 515 samples that had latitude and longitude information (see: Fig. 1). Using met, the data retrieval and functions to generate these plots took only a matter of seconds. CSV.jl [19] and DataFrames.jl [20] were used to munge (import and conform) the data.

2 Design Philosophy

met is written in Perl, Julia, and PostgreSQL PL/pgSQL (PostgreSQL Procedure Language SQL [Structured Query Language]). met-db is written as an optimized 64 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 70 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.

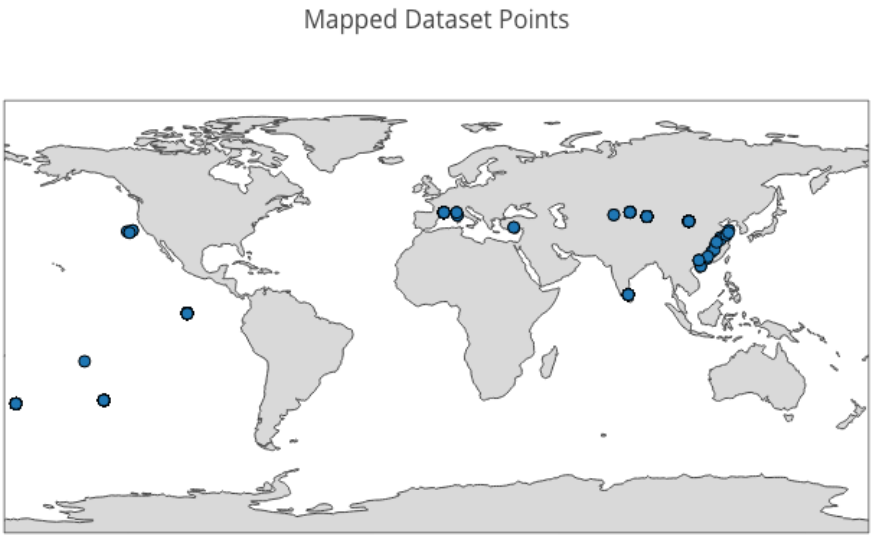

Fig 1. Map of the 515 samples with latitude and longitude data. Samples tended to tightly cluster around locations, correlating with particular biodiversity assay experiments. Plotted with PlotlyJS.jl.

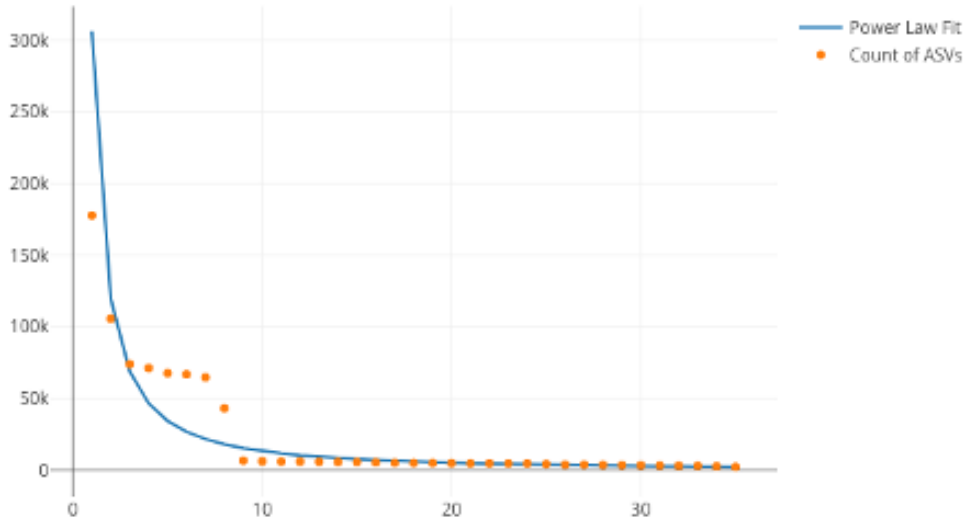

Fig 2. Number of sequences found per ASV, sorted by the number of ASVs found. If each ASV was counted across all datasets, it would necessitate a n^2 operation of all sequences compared to all other sequences. Most analysis software have some solution to this all-on-all problem. met overcomes this difficulty 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. ASV diversity plots were constructed with PlotlyJS.jl and curve fits were done with CurveFit.jl

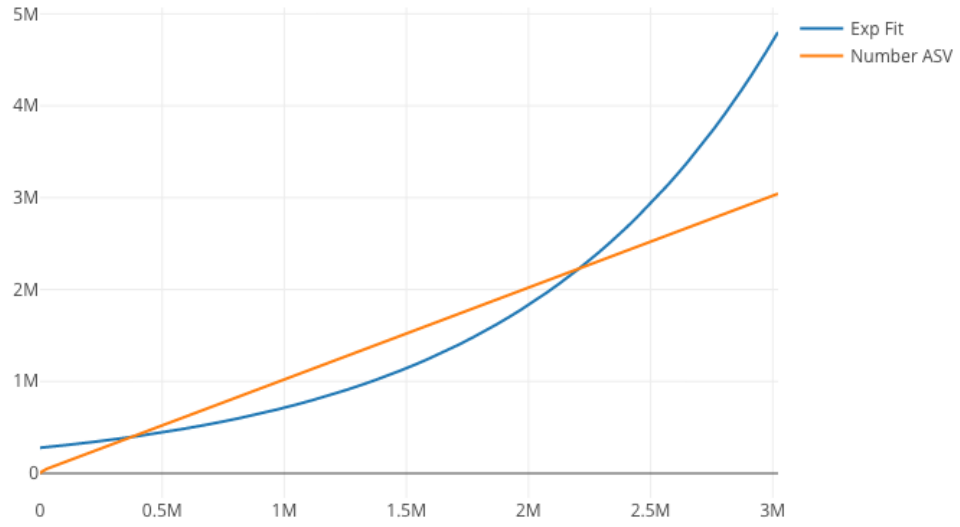

Fig 3. Cumulative plot of any particular ASV found across samples. The plot is reverse sorted by count of samples in which the ASV is found. Although it may not look like it to the eye, no single sequence was found in over 20 datasets. Plotted with plotlyJS.jl, Curve fits done with curveFit.jl

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. 4).

3 Comparisons

Other technologies have started to touch on cross-dataset capabilities, the closest of which is Qiita [14], an open source project which stores sample data from microbiome assays. Qiita uses a plugin system designed to integrate with a compute environment. While Qiita is open source design, it mostly operates within its main web service at qiita.ucsd.edu. What is particularly interesting about the design of Qiita is the availability of cross-study analysis. Within this software framework, datasets of different studies can be combined in new analysis, however, this feature would be difficult and slow if the entirety of a sample type were to be studied (e.g., all 16S V3 samples available through Qiita). While not providing the simplified analysis approach of Qiita, met's main design philosophy is large scale comparison of samples, and easy deployment as a framework for different use-cases.

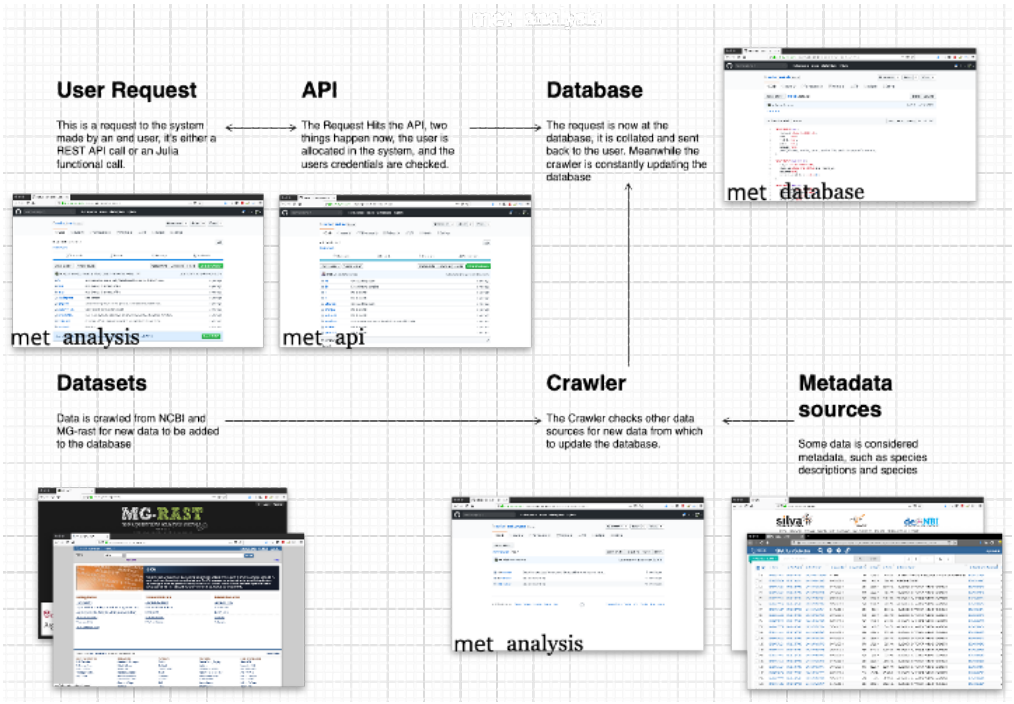

Fig 4. A diagram of met's different pieces: met-api is composed of three major components: met-analysis, met-api, and met-db. met-analysis is the main point of entry for the framework. Data gathered by crawlers would be inserted via met-analysis, and data for further downstream computation would come 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.

4 Final Thoughts

This new software package, met, is a tool that allows the massive comparison of different metabarcode experiments. While the specific results from our example generating AVS 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.

References

1. Lodge DM, Turner CR, Jerde CL, Barnes MA, Chadderton L, Egan SP, et al. Conservation in a cup of water: estimating biodiversity and population abundance from environmental DNA. Molecular ecology. 2012;21(11):2555–2558.

2. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. Nature. 2017;551(7681):457–463.

3. McDonald D, Hyde E, Debelius JW, Morton JT, Gonzalez A, Ackermann G, et al. American gut: an open platform for citizen science microbiome research. Msystems. 2018;3(3).

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

5. Yates MC, Fraser DJ, Derry AM. Meta-analysis supports further refinement of eDNA for monitoring aquatic species-specific abundance in nature. Environmental DNA. 2019;1(1):5–13.

6. Deiner K, Bik HM, Mächler E, Seymour M, Lacoursière-Roussel A, Altermatt F, et al. Environmental DNA metabarcoding: transforming how we survey animal and plant communities. Molecular ecology. 2017;26(21):5872–5895.

7. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS biology. 2007;5(3):e77.

8. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The human microbiome project. Nature. 2007;449(7164):804–810.

9. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC biology. 2014;12(1):1–4.

10. Crits-Christoph A, Robinson CK, Barnum T, Fricke W, Davila AF, Jedynak B, et al. Colonization patterns of soil microbial communities in the Atacama Desert. Microbiome. 2013;1(1):28. doi:10.1186/2049-2618-1-28.

11. Armitage DW. Linking the development and functioning of a carnivorous pitcher plant's microbial digestive community. The ISME journal. 2017;11(11):2439.

12. Caporaso JG, Paszkiewicz K, Field D, Knight R, Gilbert JA. The Western English Channel contains a persistent microbial seed bank. The ISME journal. 2012;6(6):1089–1093.

13. Oliveira RR, Silva RL, Nunes GL, Oliveira G. PIMBA: a PIpeline for MetaBarcoding Analysis. bioRxiv. 2021;.

14. Gonzalez A, Navas-Molina JA, Kosciolek T, McDonald D, Vázquez-Baeza Y, Ackermann G, et al. Qiita: rapid, web-enabled microbiome meta-analysis. Nature methods. 2018;15(10):796–798.

15. Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. Nature biotechnology. 2011;29(5):415–420.

16. Di Tommaso P, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. The impact of Docker containers on the performance of genomic pipelines. PeerJ. 2015;3:e1273.

17. Machado GS, Daitx FF, da Costa Cordeiro WL, Both CB, Gaspary LP, Granville LZ, et al. Enabling rollback support in IT change management systems. In: NOMS 2008-2008 IEEE Network Operations and Management Symposium. IEEE; 2008. p. 347–354.

18. Evans KM, Wortley AH, Mann DG. An assessment of potential diatom "barcode" genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). Protist. 2007;158(3):349–364.

19. JuliaData. JuliaData/CSV.jl;. Available from: `https://github.com/JuliaData/CSV.jl`.

20. JuliaData. JuliaData/DataFrames.jl;. Available from: `https://github.com/JuliaData/DataFrames.jl`.