The Journal of Open Source Software

EUKulele: Taxonomic annotation of the unsung eukaryotic microbes

Arianna I. Krinos[1,2], **Sarah K. Hu**[3,4], **Natalie R. Cohen**[3], **and Harriet Alexander**[*1]

1 Biology Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA **2** MIT-WHOI Joint Program in Oceanography, Cambridge and Woods Hole, MA, USA **3** Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA, USA **4** Center for Dark Energy Biosphere Investigations, University of Southern California, Los Angeles, CA, USA

DOI: 10.21105/joss.02817

Software

- Review
- Repository
- Archive

Editor: William Rowe

Reviewers:

- @johanneswerner
- @jcmcnch

Submitted: 29 October 2020
Published: 08 January 2021

Summary

The assessment of microbial species biodiversity is essential in ecology and evolutionary biology (Reaka-Kudla et al., 1996), but especially challenging for communities of microorganisms found in the environment (Das et al., 2006; Hillebrand et al., 2018). Beyond providing a census of organisms in the ocean, assessing marine microbial biodiversity can reveal how microbes respond to environmental change (Salazar & Sunagawa, 2017), clarify the ecological roles of community members (Hehemann et al., 2016), and lead to biotechnology discoveries (Das et al., 2006). Computational approaches to characterize taxonomic diversity and phylogeny based on the quality of available data for environmental sequence datasets is fundamental for advancing our understanding of the role of these organisms in the environment. Even more pressing is the need for comprehensive and consistent methods to assign taxonomy to environmentally-relevant microbial eukaryotes. Here, we present `EUKulele`, an open-source software tool designed to assign taxonomy to microeukaryotes detected in meta-omic samples, and complement analysis approaches in other domains by accommodating assembly output and providing concrete metrics reporting the taxonomic completeness of each sample.

`EUKulele` is motivated by ongoing efforts in our community to create and curate databases of genetic and genomic information (Allen, 2015; Caron et al., 2017; Richter et al., 2020; UGA, 2020). For decades, it has been recognized that genetic and genomic techniques are key to understanding microbial diversity (Fell et al., 1992). Genetic approaches are particularly useful in poorly-understood or difficult-to-access environmental systems, which may have a high degree of species diversity (Das et al., 2006; Mock et al., 2016). The most common approach for censusing microbial diversity is genetic barcoding, which targets the hyper-variable regions of highly conserved genes such as 16S or 18S rRNA (Leray & Knowlton, 2016). Computational approaches to assess the origin of these barcode-based studies (or tag-sequencing) have been well established (Bolyen et al., 2018; Schloss et al., 2009), and enable biologists to compare microbial communities and estimate sequence phylogeny. The recent collation of reference databases, e.g. PR2 and EukRef, for ribosomal RNA in eukaryotes have enabled more accurate taxonomic assessment (Del Campo et al., 2018; Guillou et al., 2012). However, barcoding approaches that focus on single marker genes or variable regions limit the field of view of microbes–especially protists, which have complex and highly variable genomes (Campo et al., 2014)–potentially limiting the organisms recovered and leaving the "true" diversity poorly constrained (Caron & Hu, 2019; Piganeau et al., 2011).

*Corresponding author

Shotgun sequencing approaches (e.g., metagenomics and metatranscriptomics) have become increasingly tractable, emerging as a viable, untargeted means to simultaneously assess community diversity and function. Large-scale meta-omic surveys, such as the Tara Oceans project (Zhang & Ning, 2015), have presented opportunities to assemble and annotate full "genomes" from environmental metagenomic samples (Delmont et al., 2018; Tully et al., 2018) and assemble massive eukaryotic gene catalogs from environmental metatranscriptomic samples (Carradec et al., 2018). The interpretation of these meta-omic surveys hinges upon curated, culture-based reference material. Several curated databases that contain predicted proteins from a mixture of genomic and transcriptomic references from eukaryotes, as well as bacteria and archaea have been created (e.g., Allen, 2015; Liu & Hu, 2020; Richter et al., 2020; UGA, 2020). Building upon the creation of high-quality reference databases, we sought to create a tool similar to MEGAN (Beier et al., 2017), CCMetagen (Marcelino et al., 2020), and MG-RAST (Keegan et al., 2016), but independent of NCBI databases and useful for both metagenomes and metatranscriptomes, as well as the study of environmental eukaryotes. Further, we sought to create a tool with a single function to download and format databases, which is necessary for computational tools to remain relevant and usable as reference databases grow.

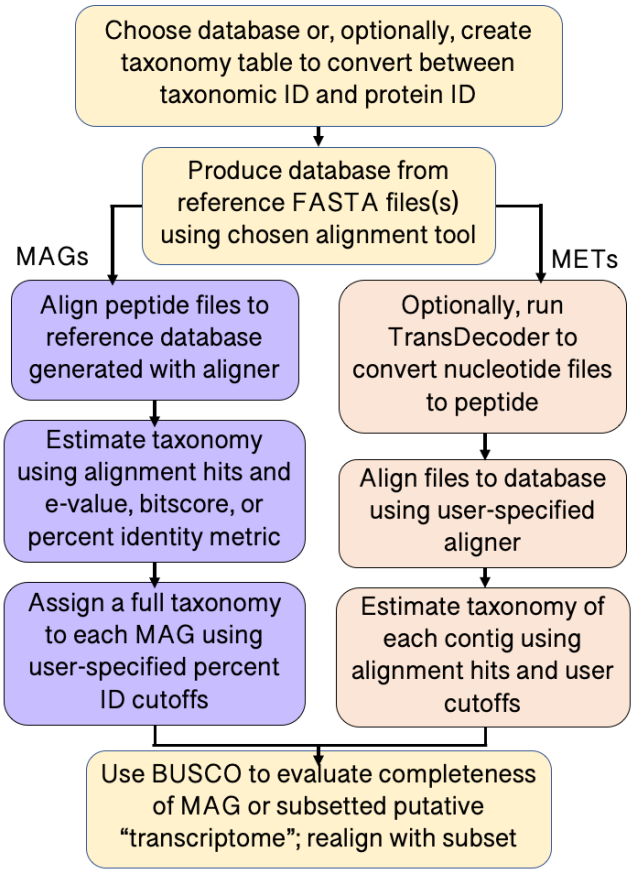

Figure 1: A flowchart describing the general workflow of the software as it relates to metatranscriptomes (METs) and metagenomes (MAGs).

Implementation

We built a tool with default databases MMETSP (Caron et al., 2017), PhyloDB (Allen, 2015), EUKZoo (Liu & Hu, 2020), MarineRefII (UGA, 2020), and EukProt (Richter et al., 2020), for optimum compatibility with environmental eukaryotic sequences. In particular, the Marine

Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) database, which contains over 650 fully-assembled reference transcriptomes (Johnson et al., 2019; Keeling et al., 2014), is among the largest single projects to create a unified reference. These databases are an invaluable resource yet, to our knowledge, no single integrated software tool currently exists to enable an end-user to harness databases in a consistent and reproducible manner.

`EUKulele` (Krinos et al., 2020) (Figure 1) is an open-source `Python`-based package designed to simplify taxonomic identification of marine eukaryotes in meta-omic samples. The package is written in `Python`, but may be installed as a `Python` module via PyPI, as a standalone tool via `conda`, or through download of the `EUKulele` tarball through `GitHub`. User-provided meta-transcriptomic or metagenomic samples are aligned against a database of the user's choosing, using a user-chosen aligner (`BLAST` (Kent, 2002) or `DIAMOND` (Buchfink et al., 2015)). The "blastx" utility is used by default if metatranscriptomic samples are only provided in nucleotide format, while the "blastp" utility is used for samples available as translated protein sequences. Any consistently-formatted database may be used, but five microbial eukaryotic database options are provided by default: MMETSP (Caron et al., 2017; Johnson et al., 2019; Keeling et al., 2014), PhyloDB (Allen, 2015), EukProt (Richter et al., 2020), EukZoo (Liu & Hu, 2020), and a combination of the MMETSP and MarRef (Caron et al., 2017; Johnson et al., 2019; Keeling et al., 2014; Klemetsen et al., 2018) (referred to as MarRef-MMETSP). This final database is the default database option, and allows the eukaryotic sequences to be compared against the expansive and high-quality MMETSP, while also distinguishing prokaryotic sequences that may be present in the sample. The package returns comma-separated files containing all of the contig matches from the metatranscriptome or metagenome, as well as the total number of transcripts that matched, at each taxonomic level, from domain or supergroup to species. If a quantification tool has been used to estimate the number of counts associated with each transcript ID, counts may also be returned.

After a desired database is either specified by the user from a previous install, or downloaded by the program (with the MarRef-MMETSP database downloaded by default), the user-selected alignment tool will create a database from the reads in the database peptide file. That database is aligned against the sample metatranscriptomic or metagenomic reads, resulting in a transcript ID, a percentage identity, e-value, and bitscore, all of which are common metrics in bioinformatics for assessing the quality of an alignment comparison between sequences. The user can specify which metric should be used for filtering out low-quality matches.

The alignment output is compared to an accompanying phylogenetic reference specific to the database (which can be generated via a script included in the package). Taxonomy is estimated at eight levels of taxonomic resolution, labeled as they are defined in the MMETSP (Keeling et al., 2014) and MarRef (Klemetsen et al., 2018) from "species" to "domain"/"supergroup." Additionally, the software returns barplots displaying the relative composition of each sample at each taxonomic level, according to the number of transcripts or number of estimated counts if provided from `Salmon` (an external transcript quantification tool (Patro et al., 2017)), which enable users to get a quick sense of the diversity in their metagenomic or metatranscriptomic sample. For metagenomic samples, a consensus taxonomic annotation is assigned based on the majority assignment of the contigs in the metagenome-assembled genome (MAG). For the metatranscriptomic option, only the taxonomic breakdown of the mixed community detected in the assembly will be returned.

`EUKulele` will assess the relative "completeness" of a given taxonomic group by taking a user-inputted list of names at some taxonomic level to determine BUSCO completeness and redundancy (Simao et al., 2015). For example, if the user was interested whether there was a set of relatively complete contigs available for genus *Phaeocystis* within their metagenomic sample, they could pass *Phaeocystis*, along with its taxonomic level, "genus," to `EUKulele`. By default, `EUKulele` will assess the BUSCO completeness of the most commonly encountered classifications at each taxonomic level. To do this, `BUSCO` (Simao et al., 2015) is used to identify the core eukaryotic genes present in each sample. Using the list of genes identified as "core," a secondary taxonomic estimation step (and consensus assignment step, for MAGs) is

performed to compare the taxonomic assignment predicted using all of the genes in comparison to the assignment made using only the genes that would be expected to be found in most reference transcriptomes. This approach is particularly useful for MAGs, and offers a method for avoiding conflicting or spurious matches made due to strain-level inconsistencies. For metatranscriptome samples, BUSCO completeness can be used to estimate the completeness of taxonomic groups to better inform their downstream interpretation.

Statement of Need

A growing number of databases have been created to catalog eukaryotic and bacterial diversity, but even when the same database is used, taxonomic assessment is not always consistent and fully documented (Menzel et al., 2016; Rasheed & Rangwala, 2012). Databases often contain distinct compilations of organisms and custom databases are commonly compiled for only a particular study (Geisen et al., 2015; Kranzler et al., 2019; Obiol et al., 2020). Database variability might influence interpretation by splitting taxonomic annotations between groups, and often impacts the proportion of contigs that are annotated (Price & Bhattacharya, 2017). A software tool can bridge the gap between database availability and efficient taxonomic assessment, making environmental meta-omic analyses more reproducible. Further, such a tool can control and assess the quality of the annotation, enable inference for specific organisms or taxonomic groups, and provide more conservative annotation in the case of organisms with exceptional amounts of inter-strain variability. We have designed the `EUKulele` (Krinos et al., 2020) package to enable efficient and consistent taxonomic annotation of metagenomes and metatranscriptomes, in particular for eukaryote-dominated samples.

Future Outlook

As single species isolates continue to be sequenced, databases are growing and becoming more reliable for assigning taxonomy in diverse environmental communities. `EUKulele` provides a platform to enable the repeated and consistent linkage of these databases to metagenomic and metatranscriptomic analyses. Taxonomic annotation is not the only desired outcome of meta-omic datasets against available databases, hence we envision eventually integrating functional annotation into the `EUKulele` package.

Acknowledgements

This software was developed with support from the Computational Science Graduate Fellowship (DOE; DE-SC0020347) awarded to AIK and from the Woods Hole Oceanographic Independent Research & Development grant awarded to HA. NRC was supported by grant 544236 from the Simons Foundation. The Center for Dark Energy Biosphere Investigations (C-DEBI; OCE-0939564) supported the participation of SKH through a C-DEBI Postdoctoral Fellowship. The High-Performance Computing cluster at Woods Hole Oceanographic Institution (Poseidon) was used to generate assemblies and run `EUKulele`.

Author Contributions

HA and SKH conceived the original idea for the tool. HA wrote the initial code for the tool. HA and AIK refined ideas for `EUKulele` related to designing it as an installable package and adding the feature of classification via core gene taxonomy for metagenomic applications. AIK developed the Python package code, wrote the conda package, implemented multiple alignment tools, and the `BUSCO` integration. AIK, NC, SKH, and HA wrote tests and documentation. HA and AIK wrote the paper.

References

Allen, A. (2015). PhyloDB. In *GitHub repository*. GitHub. https://github.com/allenlab/PhyloDB

Beier, S., Tappu, R., & Huson, D. H. (2017). Functional analysis in metagenomics using MEGAN 6. In *Functional metagenomics: Tools and applications* (pp. 65–74). Springer. https://doi.org/10.1007/978-3-319-61510-3_4

Bolyen, E., Rideout, J. R., Dillon, M. R., Bokulich, N. A., Abnet, C., Al-Ghalith, G. A., Alexander, H., Alm, E. J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J. E., Bittinger, K., Brejnrod, A., Brislawn, C. J., Brown, C. T., Callahan, B. J., Caraballo-Rodriguez, A. M., Chase, J., … Caporaso, J. G. (2018). *QIIME 2: Reproducible, interactive, scalable, and extensible microbiome data science*. https://doi.org/10.7287/peerj.preprints.27295v1

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60. https://doi.org/10.1038/nmeth.3176

Campo, J. del, Sieracki, M. E., Molestina, R., Keeling, P., Massana, R., & Ruiz-Trillo, I. (2014). The others: Our biased perspective of eukaryotic genomes. *Trends in Ecology & Evolution*, *29*(5), 252–259. https://doi.org/10.1016/j.tree.2014.03.006

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., Bell, C. J., Bharti, A., Dyhrman, S. T., Guida, S. M., & others. (2017). Probing the evolution, ecology and physiology of marine protists using transcriptomics. *Nature Reviews Microbiology*, *15*(1), 6–20. https://doi.org/10.1038/nrmicro.2016.160

Caron, D. A., & Hu, S. K. (2019). Are we overestimating protistan diversity in nature? *Trends in Microbiology*, *27*(3), 197–205. https://doi.org/10.1016/j.tim.2018.10.009

Carradec, Q., Pelletier, E., Da Silva, C., Alberti, A., Seeleuthner, Y., Blanc-Mathieu, R., Lima-Mendez, G., Rocha, F., Tirichine, L., Labadie, K., Kirilovsky, A., Bertrand, A., Engelen, S., Madoui, M.-A., Meheust, R., Poulain, J., Romac, S., Richter, D. J., Yoshikawa, G., … Wincker, P. (2018). A global ocean atlas of eukaryotic genes. *Nature Communications*, *9*(1), 373. https://doi.org/10.1038/s41467-017-02342-1

Das, S., Lyla, P., & Khan, S. A. (2006). Marine microbial diversity and ecology: Importance and future perspectives. *Current Science*, 1325–1335.

Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L. F., Nenarokov, S., Massana, R., Guillou, L., Simpson, A., Berney, C., Vargas, C. de, & others. (2018). EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biology*, *16*(9), e2005849. https://doi.org/10.1371/journal.pbio.2005849

Delmont, T. O., Quince, C., Shaiber, A., Esen, O. C., Lee, S. T., Rappe, M. S., McLellan, S. L., Lucker, S., & Eren, A. M. (2018). Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, *3*(7), 804–813. https://doi.org/10.1038/s41564-018-0176-9

Fell, J., Statzell-Tallman, A., Lutz, M., & Kurtzman, C. (1992). Partial rRNA sequences in marine yeasts: a model for identification of marine eukaryotes. *Molecular Marine Biology and Biotechnology*, *1*(3), 175–186.

Geisen, S., Tveit, A. T., Clark, I. M., Richter, A., Svenning, M. M., Bonkowski, M., & Urich, T. (2015). Metatranscriptomic census of active protists in soils. *The ISME Journal*, *9*(10), 2178–2190. https://doi.org/10.1038/ismej.2015.30

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., Boutte, C., Burgaud, G., De Vargas, C., Decelle, J., & others. (2012). The protist ribosomal reference database (PR2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated

taxonomy. *Nucleic Acids Research*, *41*(D1), D597–D604. https://doi.org/10.1093/nar/gks1160

Hehemann, J.-H., Arevalo, P., Datta, M. S., Yu, X., Corzett, C. H., Henschel, A., Preheim, S. P., Timberlake, S., Alm, E. J., & Polz, M. F. (2016). Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nature Communications*, *7*(1), 1–10. https://doi.org/10.1038/ncomms12860

Hillebrand, H., Brey, T., Gutt, J., Hagen, W., Metfies, K., Meyer, B., & Lewandowska, A. (2018). *Climate change: Warming impacts on marine biodiversity* (pp. 353–373). Springer. https://doi.org/10.1007/978-3-319-60156-4_18

Johnson, L. K., Alexander, H., & Brown, C. T. (2019). Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes. *GigaScience*, *8*(4), giy158. https://doi.org/10.1093/gigascience/giy158

Keegan, K. P., Glass, E. M., & Meyer, F. (2016). MG-RAST, a metagenomics service for analysis of microbial community structure and function. In *Microbial Environmental Genomics (MEG)* (pp. 207–233). Springer. https://doi.org/10.1093/bioinformatics/btv351

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., & others. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS Biol*, *12*(6), e1001889. https://doi.org/10.1371/journal.pbio.1001889

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research*, *12*(4), 656–664. https://doi.org/10.1101/gr.229202

Klemetsen, T., Raknes, I. A., Fu, J., Agafonov, A., Balasundaram, S. V., Tartari, G., Robertsen, E., & Willassen, N. P. (2018). The MAR databases: development and implementation of databases specific for marine metagenomics. *Nucleic Acids Research*, *46*(D1), D692–D699. https://doi.org/10.1093/nar/gkx1036

Kranzler, C. F., Krause, J. W., Brzezinski, M. A., Edwards, B. R., Biggs, W. P., Maniscalco, M., McCrow, J. P., Van Mooy, B. A., Bidle, K. D., Allen, A. E., & others. (2019). Silicon limitation facilitates virus infection and mortality of marine diatoms. *Nature Microbiology*, *4*(11), 1790–1797. https://doi.org/10.1038/s41564-019-0502-x

Krinos, A., Hu, S., Cohen, N., & Alexander, H. (2020). EUKulele: Taxonomic identification of pesky eukaryotes. In *GitHub repository*. GitHub. https://github.com/AlexanderLabWHOI/EUKulele

Leray, M., & Knowlton, N. (2016). Censusing marine eukaryotic diversity in the twenty-first century. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1702), 20150331. https://doi.org/10.1098/rstb.2015.0331

Liu, Z., & Hu, S. (2020). EukZoo database. In *GitHub repository*. GitHub. https://github.com/zxl124/EukZoo-database

Liu, Z., & Hu, S. (2020). EukZoo database. In *GitHub repository*. GitHub. https://github.com/zxl124/EukZoo-database

Marcelino, V. R., Clausen, P. T., Buchmann, J. P., Wille, M., Iredell, J. R., Meyer, W., Lund, O., Sorrell, T. C., & Holmes, E. C. (2020). CCMetagen: comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data. *Genome Biology*, *21*(1), 1–15. https://doi.org/10.1186/s13059-020-02014-2

Menzel, P., Ng, K. L., & Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with kaiju. *Nature Communications*, *7*(1), 1–9. https://doi.org/10.1038/ncomms11257

Mock, T., Daines, S. J., Geider, R., Collins, S., Metodiev, M., Millar, A. J., Moulton, V., & Lenton, T. M. (2016). Bridging the gap between omics and earth system science to better understand how environmental change impacts marine microbes. *Global Change Biology*, *22*(1), 61–75. https://doi.org/10.1111/gcb.12983

Obiol, A., Giner, C. R., Sánchez, P., Duarte, C. M., Acinas, S. G., & Massana, R. (2020). A metagenomic assessment of microbial eukaryotic diversity in the global ocean. *Molecular Ecology Resources*, *20*(3). https://doi.org/10.1111/1755-0998.13147

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Piganeau, G., Eyre-Walker, A., Grimsley, N., & Moreau, H. (2011). How and why DNA barcodes underestimate the diversity of microbial eukaryotes. *PloS One*, *6*(2), e16342. https://doi.org/10.1371/journal.pone.0016342

Price, D. C., & Bhattacharya, D. (2017). Robust dinoflagellata phylogeny inferred from public transcriptome databases. *Journal of Phycology*, *53*(3), 725–729. https://doi.org/10.1111/jpy.12529

Rasheed, Z., & Rangwala, H. (2012). Metagenomic taxonomic classification using extreme learning machines. *Journal of Bioinformatics and Computational Biology*, *10*(05), 1250015. https://doi.org/10.1142/S0219720012500151

Reaka-Kudla, M. L., Wilson, D. E., & Wilson, E. O. (1996). *Biodiversity II: Understanding and protecting our biological resources*. Joseph Henry Press. https://doi.org/10.17226/4901

Richter, D. J., Berney, C., Strassert, J. F., Burki, F., & De Vargas, C. (2020). EukProt: a database of genome-scale predicted proteins across the diversity of eukaryotic life. *BioRxiv*. https://doi.org/10.1101/2020.06.30.180687

Salazar, G., & Sunagawa, S. (2017). Marine microbial diversity. *Current Biology*, *27*(11), R489–R494. https://doi.org/10.1016/j.cub.2017.01.017

Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., & others. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, *75*(23), 7537–7541. https://doi.org/10.1128/AEM.01541-09

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, *31*(19), 3210–3212.

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203. https://doi.org/10.1038/sdata.2017.203

UGA, M. L. (2020). *MarineRefII*. http://roseobase.org/data/

Zhang, H., & Ning, K. (2015). The Tara Oceans project: new opportunities and greater challenges ahead. *Genomics, Proteomics & Bioinformatics*, *13*(5), 275. https://doi.org/10.1016/j.gpb.2015.08.003