Sequencing our way to more accurate community abundance

Georgina Brennan[1]

[1]Institute of Marine Sciences

August 23, 2022

Abstract

Over the last two decades, there has been a huge increase in our understanding of microbial diversity, structure and composition enabled by high throughput sequencing (HTS) technologies. Yet, it is unclear how the number of sequences translates to the number of cells or species within the community. Additional observational data may be required to ensure relative abundance patterns from sequence reads are biologically meaningful or presence absence data may be used instead of abundance. The goal is to obtain robust community abundance data, simultaneously, from environmental samples. In this issue of Molecular Ecology Resources, Karlusich et al., (2022) describe a new method for quantifying phytoplankton cell abundance. Using Tara Oceans datasets, the authors propose the photosynthetic gene psbO for reporting accurate relative abundance of the entire phytoplankton community from metagenomic data. The authors demonstrate improved correlations with traditional optical methods including microscopy and flow cytometry, improving upon current molecular identification typically using rRNA markers genes. Furthermore, to facilitate application of their approach, the authors curated a psbO gene database for accessible taxonomic queries. This is an important step towards improving species abundance estimates from molecular data and eventually reporting of absolute species abundance, enhancing our understanding of community dynamics.

Sequencing our way to more accurate community abundance

Georgina L. Brennan

Institute of Marine Sciences (ICM-CSIC), Passeig Marítim de la Barceloneta 37-49, 08003, Barcelona, Spain

Over the last two decades, there has been a huge increase in our understanding of microbial diversity, structure and composition enabled by high throughput sequencing (HTS) technologies. Yet, it is unclear how the number of sequences translates to the number of cells or species within the community. Additional observational data may be required to ensure relative abundance patterns from sequence reads are biologically meaningful or presence absence data may be used instead of abundance. The goal is to obtain robust community abundance data, simultaneously, from environmental samples. In this issue of Molecular Ecology Resources, Karlusich et al., (2022) describe a new method for quantifying phytoplankton cell abundance. Using *Tara* Oceans datasets, the authors propose the photosynthetic gene *psbO* for reporting accurate relative abundance of the entire phytoplankton community from metagenomic data. The authors demonstrate improved correlations with traditional optical methods including microscopy and flow cytometry, improving upon current molecular identification typically using rRNA markers genes. Furthermore, to facilitate application of their approach, the authors curated a *psbO* gene database for accessible taxonomic queries. This is an important step towards improving species abundance estimates from molecular data and eventually reporting of absolute species abundance, enhancing our understanding of community dynamics.

High-throughput sequencing (HTS) technologies for identification of taxa from environmental samples have significantly improved our understanding of biodiversity and community assembly processes. However, quantification of species abundance from sequence reads is not a straight forward task. This is because biases

from DNA extraction, PCR amplification and sequencing will affect the number of sequence reads obtained for each taxonomic unit and therefore the representation within the environmental sample (Bik et al., 2012). In addition, multi-copy genes are often targeted to increase detection sensitivity of target DNA from environmental samples for example, prokaryote (16S) and eukaryote (18S) rRNA marker genes. However, large variations in copy number within and between taxa reduce our ability to quantify taxon abundance. Karlusich et al. (2022) explains that whilst many HTS studies report the relative abundance of the gene sequences, this may not be an accurate measure of the relative abundance of the organisms containing those sequences. Yet, accurate relative abundance measurements are crucial to our understanding of community composition simply because when one taxonomic unit increases in relative abundance, another necessarily decreases (figure 1).

Inaccurate assessments of abundance will have serious consequences to our understanding and management of ecosystems. For example, Karlusich et al. (2022) highlights the ecological importance of marine phytoplankton including, their position at the foundation of ocean ecosystems and roles in primary productivity and biogeochemical cycles (Field, Behrenfeld, Randerson, & Falkowski, 1998). Under future global change species sorting will potentially alter the composition of functional groups within marine microbial communities (Di Pane, Wiltshire, McLean, Boersma, & Meunier, 2022), which in turn feeds back into the biogeochemical cycles. It is therefore important to know how these communities will be composed in the future, and the consequences to ecosystem services they provide. Targeted amplicon sequencing (a.k.a**metabarcoding**) is now routinely used for the characterization of complex assemblages of prokaryotic and eukaryotic organisms (Creer et al., 2016) and we are now in a position where we can reliably identify most of the abundant taxa in complex assemblages (albeit with some exceptions) and provide "semi-quantitative" data of taxa abundance from complex mixtures (e.g. ocean microbiome (Giner et al., 2016), soil microbiome (Delgado-Baquerizo et al., 2018), air microbiome (Drautz-Moses et al., 2022)). However, it is well documented that metabarcoding suffers from biases associated with PCR amplification of target genes (Bik et al., 2012). HTS-based metagenomics (the sequencing of genomic fragments from many members of the community) is a non-targeted, PCR-free method and as costs decline, is an emerging solution to taxonomic identification without biases introduced by PCR. Whilst traditional methods, such as microscopy and flowcytometry are better at providing quantitative data and are well validated, they often lack the ability to scale up to whole communities, especially in systems or methods that rely on human expertise instead of automation (Makiola et al., 2020). The goal is to obtain reliable abundance data for each taxonomic unit, from the number of sequences reads obtained from the environmental sample.

Karlusich et al. (2022) propose a straightforward solution to robustly measure relative abundance from environmental samples and describe each step of their selection and validation process. Using datasets from the*Tara* Oceans (global expedition sampling global plankton in the upper layers of the world ocean (Sunagawa et al., 2020)), Karlusich et al. (2022) target nuclear-encoded single-copy, core, photosynthetic genes obtained from metagenomes to circumvent the limitations of targeted gene sequencing (metabarcoding) and multicopy markers. The authors focused on the *psbO* gene, which is essential for photosynthetic activity and does not have non-photosynthetic homologs, thus is can be used to measure abundance of the total photosynthetic group and has the added benefit covering the whole phytoplankton community. Similarly, both cyanobacteria and eukaryotic phytoplankton can be measured by combining two rRNA marker genes (e.g. prokaryotic 16S and eukaryotic 18S) however, relative abundances derived from different amplicon libraries cannot be directly compared (Tkacz, Hortala, & Poole, 2018). Importantly, cross domain comparisons can be made using the *psbO* gene.

Karlusich et al. (2022), found that the *psbO* gene is a robust marker for estimating relative abundance of phytoplankton and were able to examine the biogeography of the entire phytoplankton community simultaneously. To validate their approach, the authors used *Tara*Oceans data including, imaging datasets (microscopy and flow cytometry) and molecular datasets from metabarcoding, metagenomics and metatranscriptomics. Using imaging datasets (flow cytometry, microscopy) they demonstrated the accuracy of their approach and even confirmed the presence colony formation and symbiosis in some of the smallest phytoplankton cells that were found in the largest size-fractioned water samples. Armed with the evidence to

demonstrate that the *psbO*gene accurately provides relative abundance data, the authors compared their results with the commonly used rRNA marker genes 16S and 18S (rRNA gene miTags from metagenome data and rRNA gene metabarcoding). Here they show that the *psbO* gene outperformed rRNA gene datasets in reporting accurate relative abundance of phytoplankton. Furthermore, the authors demonstrate that *psbO* gene improves measures of microbial community diversity, structure, and composition as compared to rRNA genes and identified biases in metabarcoding datasets. However, they report that diversity indices such as Shannon diversity (that accounts for both species richness and evenness), were sufficiently robust to account for biases introduced by the rRNA marker methods. Furthermore, they confirm that neither rRNA gene markers nor *psbO* could accurately report biovolume.

This is an exciting tool since we still do not have a clear understanding of the abundance of phytoplankton groups from the ocean. Similarly, the same steps can be followed from Karlusich et al. (2022), in order to identify suitable genes for other study systems. There are many research avenues where the use of good quality abundance data would be enormously impactful. For example, to make more accurate assessment of floral resource use from pollen grains found in honey (Jones et al., 2021) or the bodies of pollinators (Lowe, Jones, Brennan, Creer, & de Vere, 2022), exploring how the abundance of allergenic airborne pollen correlates with human health (Rowney et al., 2021) and to gain insights into the relationship between gut microbiome and human health (Proctor et al., 2019). However, it is important that new markers are accompanied by well populated genetic databases in order to avoid biases during taxonomic assignment. A measure of absolute abundance is the ultimate goal and future investigations using this approach can achieve absolute abundance using careful sampling design and DNA internal standards ('spike in') (Tkacz et al., 2018).

Hosted file

`image1.emf` available at https://authorea.com/users/458916/articles/582993-sequencing-our-way-to-more-accurate-community-abundance

FIGURE 1 Abundance of two marine microbes illustrated by three hypothetical scenarios. (a) equal relative abundance does not match cell abundance due to copy number variation of the target molecular marker (e.g., rRNA marker genes), (b) accurate relative abundance matches cell abundance when using single copy molecular marker genes (e.g.,*psb0* gene) and (c) accurate absolute abundance when using single copy molecular marker and a DNA spike in.

References

Bik, H. M., Porazinska, D. L., Creer, S., Caporaso, J. G., Knight, R., & Thomas, W. K. (2012). Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* ,*27* (4), 233–243. doi: 10.1016/j.tree.2011.11.010

Creer, S., Deiner, K., Frey, S., Porazinksa, D., Taberlet, P., Kelley Thomas, W. K., ... Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution* , *7* (9), 1008–1018. doi: 10.1111/2041-210X.12574

Delgado-Baquerizo, M., Oliverio, A. M., Brewer, T. E., Benavent-González, A., Eldridge, D. J., Bardgett, R. D., ... Fierer, N. (2018). A global atlas of the dominant bacteria found in soil. *Science* , *359* (6373), 320–325. doi: 10.1126/science.aap9516

Di Pane, J., Wiltshire, K. H., McLean, M., Boersma, M., & Meunier, C. L. (2022). Environmentally induced functional shifts in phytoplankton and their potential consequences for ecosystem functioning. *Global Change Biology* , *28* (8), 2804–2819. doi: 10.1111/gcb.16098

Drautz-Moses, D. I., Luhung, I., Gusareva, E. S., Kee, C., Gaultier, N. E., Premkrishnan, B. N. V., ... Schuster, S. C. (2022). Vertical stratification of the air microbiome in the lower troposphere.*Proceedings of the National Academy of Sciences* , *119* (7), e2117293119. doi: 10.1073/pnas.2117293119

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* , *281* (5374), 237–240. doi:

10.1126/science.281.5374.237

Giner, C. R., Forn, I., Romac, S., Logares, R., de Vargas, C., & Massana, R. (2016). Environmental Sequencing Provides Reasonable Estimates of the Relative Abundance of Specific Picoeukaryotes. *Applied and Environmental Microbiology* , *82* (15), 4757–4766. doi: 10.1128/AEM.00560-16

Jones, L., Brennan, G. L., Lowe, A., Creer, S., Ford, C. R., & de Vere, N. (2021). Shifts in honeybee foraging reveal historical changes in floral resources. *Communications Biology* , *4* (1), 37. doi: 10.1038/s42003-020-01562-4

Karlusich, J. J. P., Pelletier, E., Zinger, L., Lombard, F., Colin, S., Gasol, J., ... Bowler, C. (2022). A robust approach to estimate relative phytoplankton cell abundances from metagenomes. *Molecular Ecology Resources* , *XX* (XX), XXX–XXX.

Lowe, A., Jones, L., Brennan, G., Creer, S., & de Vere, N. (2022). Seasonal progression and differences in major floral resource use by bees and hoverflies in a diverse horticultural and agricultural landscape revealed by DNA metabarcoding. *Journal of Applied Ecology* , *59* (6), 1484–1495. doi: 10.1111/1365-2664.14144

Makiola, A., Compson, Z. G., Baird, D. J., Barnes, M. A., Boerlijst, S. P., Bouchez, A., ... Bohan, D. A. (2020). Key Questions for Next-Generation Biomonitoring. *Frontiers in Environmental Science* , *7* . Retrieved from https://www.frontiersin.org/articles/10.3389/fenvs.2019.00197

Rowney, F. M., Brennan, G. L., Skjøth, C. A., Griffith, G. W., McInnes, R. N., Clewlow, Y., ... Creer, S. (2021). Environmental DNA reveals links between abundance and composition of airborne grass pollen and respiratory health. *Current Biology* , *31* (9), 1995-2003.e4. doi: 10.1016/j.cub.2021.02.019

Sunagawa, S., Acinas, S. G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., ... de Vargas, C. (2020). Tara Oceans: Towards global ocean ecosystems biology. *Nature Reviews Microbiology* ,*18* (8), 428–445. doi: 10.1038/s41579-020-0364-5

Tkacz, A., Hortala, M., & Poole, P. S. (2018). Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* ,*6* (1), 110. doi: 10.1186/s40168-018-0491-7