Towards quantitative microbiome community profiling using internal standards

Yajuan Lin [a,b]#, Scott Gifford [c], Hugh Ducklow [d], Oscar Schofield [e],

and Nicolas Cassar [a,b]#

[a] Division of Earth and Ocean Sciences, Nicholas School of the Environment, Duke University, Durham, NC 27708, USA

[b] Université de Brest - UMR 6539 CNRS/UBO/IRD/Ifremer, Laboratoire des sciences de l'environnement marin – IUEM, Rue Dumont D'Urville, Plouzané, 29280, France

[c] Department of Marine Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA

[d] Lamont Doherty Earth Observatory of Columbia University, Palisades NY 10964 USA

[e] Rutgers University's Center for Ocean Observing Leadership (RU COOL), Department of Marine and Coastal Sciences, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 80901, USA

Running Title: Quantitative microbiome profiling

#Address correspondence to Yajuan Lin (yajuan.lin@duke.edu) and Nicolas Cassar (Nicolas.Cassar@duke.edu)

Abstract

An inherent issue of high-throughput rRNA gene tag sequencing microbiome surveys is that they provide compositional data in relative abundances. This often leads to spurious correlations making the interpretation of relationships to biogeochemical rates challenging. To overcome this issue, we quantitatively estimated the abundance of microorganisms by spiking in known amounts of internal DNA standards. Using a 3-year sample set of diverse microbial communities from the Western Antarctica Peninsula, we demonstrated that the internal standard method yielded community profiles and taxa co-occurrence patterns substantially different from those derived using relative abundances. We found that the method provided results consistent with the traditional CHEMTAX analysis of pigments and total bacterial counts by flow cytometry. Using the internal standard method, we also showed that chloroplast 16S rRNA gene data in microbial surveys can be used to estimate abundances of certain eukaryotic phototrophs such as cryptophytes and diatoms. In *Phaeocystis*, scatter in the 16S/18S rRNA gene ratio may be explained by physiological adaptation to environmental conditions. We conclude that the internal standard method, when applied to rRNA gene microbial community profiling, is quantitative and that its application will substantially improve our understanding of microbial ecosystems.

Importance High-throughput sequencing based marine microbiome profiling is rapidly expanding and changing how we study the oceans. Although powerful, the technique is not fully quantitative - it only provides taxon counts in relative abundances. In order to address this issue, we presented a method to quantitatively estimate microbial abundances per unit volume of seawater filtered by spiking in known amounts of internal DNA standards to each sample. We validated this method by comparing the calculated abundances to other independent estimates

45 including chemical markers (pigments) and total bacterial cell counts by flow cytometry. The
46 internal standard approach allows us to quantitatively estimate and compare marine microbial
47 community profiles, with important implications for linking environmental microbiomes to
48 quantitative processes such as metabolic and biogeochemical rates.
49

Introduction

Since the first application of Roche 454 pyrosequencing to marine 16S rRNA gene amplicon samples (1), high-throughput sequencing of environmental PCR-amplified marker genes has transformed the study of marine microbiomes. It has been at the core of multiple recent programs varying in scale and breadth, including the International Census of Marine Microbes (2), TARA expeditions (3, 4), Malaspina 2010 Expedition (5), Ocean Sampling Day (6) and the Long Term Ecological Research sites (Palmer, HOT, Tahiti and other sites). These studies and other programs have revealed unprecedented microbial diversity and biogeographic patterns and advanced our understanding of marine microbial ecology (7) and biogeochemistry (4, 8).

An important limitation of the rRNA gene tag based DNA sequencing approach is that it only provides compositional data, i.e., taxonomical profiles in relative proportions. While useful, compositional data is incomplete. As an example, should species A be equally abundant in two samples, its relative abundance in the first sample will be double that in the second sample if the total cell concentration is twice as high in the second sample. More broadly, compositional data can lead to various statistical issues mainly due to two geometric features (9). First, the distance between two points has no absolute scale, e.g., counts of 1 and 2 have the same information as 100 and 200 (10), and thus the counts from different samples could have different uncertainties, making it difficult to identify statistically significant differences by standard tests (11). Second, compositional data is constrained by the 'sum of 1' and its projection in space is restricted to a simplex for which common statistical analyses based on Euclidean space may not be applicable (12–15). For example, it has long been realized that correlation analyses on compositional data can yield spurious correlations (16). This problem is particularly severe when communities have dominant taxa (14), as commonly observed in some environmental samples (1, 17). These issues

hinder cross-study comparisons of the rapidly expanding communal rRNA gene data sets. Various transformation (e.g., centered log-ratio transformation) and specialized data analysis routines have been developed to overcome these issues [e.g., programs such as DESeq2 consider the weighting of each taxon (18, 19)]. However, such routines make it difficult to interpret the underlying biological and ecological mechanisms and absolute quantification provides a very valuable piece of information.

To palliate the artifacts associated with compositional data or relative microbiome profiling (RMP), two approaches have recently been developed for quantitative microbiome profiling (QMP). The first approach is to normalize the 16S rRNA gene OTU counts to total bacterial counts estimated by flow cytometry (FCM) (20) (21). The second approach, internal standard normalization (ISN), consists of spiking-in known concentrations of internal standards (DNA or cells) into samples before DNA extraction (22). This approach was adapted from internal RNA standards in metatranscriptomics (23). ISN has recently been applied to study prokaryotic community composition in soils (24) and in the human gastrointestinal tract (25). In this study, as a proof of concept, we estimated the QMP of oceanic prokaryotes and eukaryotic plankton sampled from the Western Antarctica Peninsula (WAP) (Figure 1A) using 16S and 18S rRNA gene amplicon sequencing combined with internal DNA standards. The large environmental gradients (e.g., coast vs. open ocean and open water ice covered regions) at the WAP lead to diverse and highly variable microbial communities (8, 26), thereby providing an ideal stage to test the ISN. Below, we present the internal standard normalization method (ISN) as applied to marine samples. In order to validate the method, we 1) assessed the precision of ISN by spiking in varying amounts of standards; 2) compared phytoplankton abundances based on ISN to those based on CHEMTAX estimates, a program to calculate phytoplankton

abundances based on pigment analyses (27); and 3) compared total bacterial counts estimated by the 16S rRNA gene ISN to direct measurements by cell counting flow cytometry. As an example of the numerous applications of this new approach, we demonstrated how the QMP and the relation of phytoplankton chloroplast 16S to genomic 18S rRNA genes abundances may provide insight into plankton ecology and photophysiology.

Results and discussion

Brief description of the method

A thorough description of the method is presented in the material and methods section. Briefly, known amounts of genomic DNA from organisms not expected in the natural seawater samples, i.e., *Schizosaccharomyces pombe* for 18S rRNA genes and *Thermus thermophiles* for 16S rRNA genes, were added to each sample before DNA extraction. The abundance of OTU_i (in 16S or 18S rRNA gene copies per ml seawater) in sample j was calculated as:

$$A_{i,j} = \frac{R_{i,j} \times C_s}{R_{s,j} \times V_j}$$

where $R_{i,j}$ is the number of reads of OTU_i in sample j, $R_{s,j}$ is the number of 16S or 18S rRNA gene standard reads sequenced in sample j, C_s is the total number of 16S or 18S rRNA gene copies spiked into each sample, and V_j is the filtered sea water volume in ml. For double-stranded DNA, assuming the average weight of a base pair is 650 Daltons (650 g per mole), C_s can be calculated as:

$$C_s = \frac{gDNA\ amount\ (ng) \times 6.022 \times 10^{23}\ (copies\ mol^{-1}) \times rrn\ s}{length\ of\ gDNA\ (bp) \times 1 \times 10^{9}\ (ng\ g^{-1}) \times 650\ (g\ mol^{-1} bp^{-1})}$$

where rrn_s is the 16S or 18S rRNA gene copy number per cell. In our study, the spiked 16S rRNA gene standard was 14.85 ng of *T. thermophilus* gDNA, with rrn = 2 and genome size = 2.13 Mb

(28), thus $C_s = 1.29 \times 10^7$ rRNA gene copies per sample. For the 18S rRNA gene standard *S. pombe*, the rrn could vary from 100 – 120 copies per cell (29). For our 18S rRNA gene calculation we used a median number rrn = 110, which may introduce up to a 10% bias. However, this bias should be the same across all the samples and thus should not influence the comparison between samples. With 16.1 ng of spiked 18S standard per sample and the genome size of 13.8 Mb (29), $C_s = 1.19 \times 10^8$ copies per sample. With a known number of rRNA gene copy number per cell rrn_i for OTU_i (e.g., 1 copy per cell for SAR11), cell abundance in sample j (in cells ml^{-1}) can be calculated as $A_{i,j} / rrn_i$. We note that this is only possible when the rrn_i is known and assuming single genome per cell.

Validation of the method

To validate the method, 56 samples were collected at the WAP on three Palmer LTER annual cruises (years 2012, 2013 and 2015) (Figure 1A). Internal standard recoveries averaged 0.8% (0.2% - 2.9%) of total prokaryotic 16S rRNA gene reads, and 2.4% (0.7% - 5.7%) of total eukaryotic 18S rRNA gene reads, well within the range appropriate for detection (i.e., $\geq 0.1\%$) without overwhelming the environmental reads. Based on ISN, the abundance of rRNA genes between stations varied by 16- and 27-fold for eukaryotes and prokaryotes, respectively (Figure 2). Using rrn from the rrnDB database (30), we converted OTU2 (SAR11) and OTU5 (*Polaribacter*) rRNA gene counts to cell abundances (Supplementary Figure S1). The average cell abundance of the SAR11 OTU in our samples was 2.0×10^5 cells ml^{-1}, in line with SAR11 estimates reported by other studies in the Southern Ocean (31)(32)(33). Below, we assessed the precision of the ISN by spiking in two different amounts of internal standards. We also corroborated our results with abundance estimates using two independent methods, CHEMTAX pigment analyses for the 18S rRNA gene and FCM for the 16S rRNA gene abundances.

Precision of ISN

In a test-sequencing run to optimize the standard amount, we added the eukaryotic internal standards at two different concentrations (1:5) into representative samples (see details in Material and methods). The response was proportional to the spiked-in level (Figure 1B) with a maximum deviation estimated at 25% (averaged 18%) across the varying communities sampled at the coastal and open ocean sites. For comparison, the traditional qPCR methods can yield errors as large as the signal (34) with typical coefficient of variation (CV) values ranging from 15% to 50% (35)(36). This comparison should be interpreted with caution because the precision of qPCR has been verified over a wider range of concentrations (i.e., 7-9 orders of magnitude) (37, 38) than most internal standard studies (39). To test the reproducibility of the sequencing technique, we also barcoded and sequenced a coastal sample in duplicates (Coastal_2A and 2B), and the resulting community profiles are highly similar (Figure 1B). The CV for estimated taxa abundance was 2.8% on average and 12.3% at maximum (supplementary Table S1) with higher uncertainties for rarer taxa.

Method comparison:

a. Phytoplankton 18S rRNA gene ISN vs. CHEMTAX abundance

We compared phytoplankton QMP estimated by ISN with the traditional CHEMTAX analysis of High Performance Liquid Chromatography (HPLC) pigment profiles (26, 40) for three phytoplankton groups commonly observed at the WAP, i.e., cryptophytes, diatoms, and *Phaeocystis*. The cryptophyte abundances calculated by 18S rRNA gene and CHEMTAX were highly correlated (Pearson's $R^2 = 0.98$, $P < 0.0001$) (Figure 3A). Significant correlations were also observed for diatoms ($R^2 = 0.42$, $P < 0.0001$) (Figure 3C) and *Phaeocystis* ($R^2 = 0.57$, $P < 0.0001$) (Figure 3E), although the relationships were weaker. Because alloxanthin is only present

in cryptophytes, their CHEMTAX estimates are likely more robust than the ones for diatoms and *Phaeocystis.* In addition, alloxanthin was the most abundant pigment in our sample set, with an average concentration of 0.61 μg/L. In comparison, the other accessory pigments were substantially less abundant (19' butanoyloxyfucoxanthin (0.01 μg/L), chlorophyll c2 (0.18 μg/L), chlorophyll c3 (0.02 μg/L), chlorophyll b (0.01 μg/L), fucoxanthin (0.13 μg/L), hexanoyloxyfucoxanthin (0.13 μg/L)). Low concentrations of accessory pigments could introduce errors in CHEMTAX estimates of diatoms and *Phaeocystis*. Using RMP, a significant but weaker correlation was observed for cryptophytes ($R^2 = 0.51$, $P < 0.001$) (Figure 3B). No significant correlation between RMP and CHEMTAX estimates was observed for diatoms (Figure 3D) and *Phaeocystis* (Figure 3F).

b. Bacterial 16S rRNA gene ISN vs. FCM bacterial abundance

The total prokaryotic 16S rRNA gene abundances were significantly correlated with the bacterial FCM counts albeit with a small correlation coefficient (Pearson's $R^2 = 0.19$, $P < 0.001$; or $R^2 = 0.20$, $P < 0.001$ after log-transformation) (Figure 4A). In general, rRNA gene copy numbers were much higher than the FCM cell counts. A variety of factors may explain this. First, for the four points circled in grey in Figure 4B), the FCM estimates of $\geq 2.0 \times 10^6$ cells ml^{-1} were anomalously high compared to the corresponding leucine incorporation rates or Chl *a* concentrations. Second, while bacteria associated with particles were efficiently captured by DNA sequencing, they may have been missed by FCM counts if the vortex step did not break down the particle-bacteria-associations. In polar and coastal regions, a significant proportion of bacteria could be attached to particles (41). Corroborating this hypothesis, we found that samples where ISN predicted a higher abundance of bacteria than FCM tended to have a higher percentage of particle-associated OTUs (Figure 4A). Finally, the difference in rrn for different

OTUs could also explain the discrepancy between the ISN and FCM bacterial abundance. For example, the rrns in SAR11 and *Marinomonas* sp. MWYL1 are 1 and 8, respectively (30). Populations with larger rrn should have higher 16S rRNA gene to FCM counts ratios. In addition, the fact that multiple genomes may exist within a single cell (42) could also contribute to the discrepancy. To estimate cell abundances, top 20 classified OTU QMP in 16S rRNA gene copies per ml were divided by their rrn estimated by rrnDB (30) and the resulting OTU cell abundances were summed up for each sample. Taxa identified as particle-associated bacteria through size-fractionated filtration in (41) were then excluded. After discarding the four potential outliers and correcting for rrn and particle – association effects, cell abundances estimated by rRNA gene and FCM counts displayed a substantially higher correlation coefficient ($R^2 = 0.61$, $P < 0.001$; or $R^2 = 0.44$, $P < 0.001$ after log-transformation) and were close to the 1:1 line (Figure 4B).

We note that the rrn correction is not only important for ISN but also for the normalization of FCM (e.g. (20)). The absolute cell abundance of OTU x in a particular sample should be calculated as $\frac{C_x/rrn_x}{\sum_i^n C_i/rrn_i} \times FCM$, where C_x is the rRNA gene counts for OTU x . Should rrn_x be constant for a particular taxa, changes in the numerator introduces a systematic bias when comparing relative changes in absolute abundances between samples. However, because $\sum_i^n C_x/rrn_x \neq \sum_i^n C_x$, the denominator may lead to uneven biases across samples. A simple example using two OTUs commonly found in the WAP is presented in supplementary Table S2. Without taking into account the rrn, the estimates of absolute OTU abundances based on FCM normalization could be off by 5 fold, and the estimated abundance variation between two samples could be off by 3.6 fold in this particular example. Caution should therefore be taken in applying the FCM normalization method without resolving the community rrn profile.

One approach to estimating the rrn profile is to use the phylogenetic information to predict the rrn of OTUs based on existing rrn databases such as rrnDB (30)(43). A recent human microbiome study corrected the 16S rRNA gene matrix using rrnDB (21). However, substantial uncertainties associated with the rrn correction remain as 1) a significant portion of the OTUs are unclassified and 2) the limited number of known rrn from sequenced genomes likely does not reflect the natural variability in rrn.

When applying the FCM normalization method to marine samples, the difference in sampling volume for DNA and FCM should be considered. Cells for DNA analyses are generally filtered from liters of seawater, while FCM samples are generally estimated from less than 1 ml of seawater. In patchy environments, these two volumes may reflect different communities.

Application: Case study at the WAP

In our WAP case study, the estimated total eukaryotic rRNA gene abundance was significantly correlated with environmental variables including the distance to shore (Pearson's $R = -0.6$, $P < 0.001$; Spearman's $\rho = -0.6$, $P < 0.001$), Chl *a* concentration ($R = 0.8$, $P < 0.001$; $\rho = 0.7$, $P < 0.001$), and primary production rate ($R = 0.7$, $P < 0.001$; $\rho = 0.5$, $P < 0.001$). Conversely, the estimated total prokaryotic rRNA gene abundance was not significantly correlated with distance to shore ($R = -0.3$, $P > 0.05$; $\rho = -0.2$, $P > 0.1$), but was significantly correlated with Chl *a* ($R = 0.6$, $P < 0.001$; not significant by Spearman, $\rho = 0.3$, $P > 0.05$) and significantly correlated with bacterial production measured by ^{3}H-Leucine incorporation ($R = 0.7$, $P < 0.001$; $\rho = 0.6$, $P < 0.001$). Looking at specific taxa, the abundance of *Polaribacter* OTU5 increased significantly with increasing Chl *a* ($R = 0.8$, $P < 0.001$; $\rho = 0.5$, $P < 0.001$) (Supplementary Figure S1), which is consistent with the observations that *Polaribacter* thrives during phytoplankton blooms (44, 45). The SAR11 OTU2 cell abundances did not show a clear

232 trend across Chl *a* gradients (R = -0.02, P = 0.9; ρ = -0.01, P = 0.9). This could be a result of
233 patterns at finer taxonomic scales, e.g., amplicon sequence variants resolved down to the single-
234 nucleotide level (46). The relative abundance of SAR11 OTU decreased with increasing Chl *a* (R
235 = -0.5, P < 0.001; ρ = -0.5, P < 0.001), but this could be a spurious correlation stemming from an
236 increase in the total bacterial abundance.

237 Community co-occurrence matrices based on Spearman's correlation coefficients (Figure
238 5) showed that QMP and RMP matrices were significantly different (P < 0.001) by Jennrich test
239 (47) and Steiger test (48). QMP resulted in more positive correlations (270 vs. 218 for RMP)
240 mostly appearing within the prokaryotic communities, and fewer negative correlations overall
241 (124 vs. 172 for RMP). Interestingly, similar differences in co-occurrence patterns based on RMP
242 and QMP have also been observed in human gut microbiome studies using the FCM
243 normalization method (21).

244 *Quantitatively estimating eukaryotic phytoplankton abundances using chloroplast 16S*
245 *rRNA gene abundances*

246 The QMP of five eukaryotic phytoplankton groups calculated from internal standard
247 normalized 18S rRNA gene abundances and the corresponding chloroplast 16S rRNA gene
248 counts were compared (Figure 6). Strong linear correlations using the type-II least-square fit
249 were observed between the chloroplast 16S rRNA gene counts and genomic 18S rRNA gene
250 counts for Cryptophytes (R^2 = 0.87, P < 0.0001), and diatoms, including *Fragilariopsis* (R^2 =
251 0.55, P < 0.0001), *Corethron* (R^2 = 0.72, P < 0.0001) and *Proboscia* (R^2 = 0.40, P < 0.0001). A
252 weak correlation was observed for *Phaeocystis* using the type-II least-square fit (R^2 = 0.06, P <
253 0.0001) but not with a Pearson coefficient (R^2 = 0.06, P = 0.09). These results show that
254 eukaryotic autotroph abundances can be reliably estimated from their corresponding chloroplast

16S rRNA gene abundances for the three phytoplankton groups examined, i.e., Cryptophytes, Diatoms and *Phaeocystis*.

Chloroplast-16S rRNA genes can represent a large fraction of total community 16S rRNA gene library reads, especially in productive oceanic regions where phototrophic eukaryotes tend to dominate. For example, 52% of the total 16S rRNA gene reads were annotated as chloroplasts at our study site (averaged over all sampled stations). While these chloroplast reads are generally discarded, they may provide valuable information about the phototrophic eukaryote abundance without incurring the additional cost of 18S rRNA gene amplicon sequencing. Several recent studies inferred eukaryotic phytoplankton relative abundances from the chloroplast 16S rRNA gene reads (41, 49). The method described herein may allow us to estimate the host phytoplankton abundances from the ISN chloroplast sequences (Figure 6).

18S to 16S rRNA gene ratios as measure of phytoplankton ecophysiology

ISN can also be used to quantify variability in the ratio of chloroplast 16S rRNA gene / genomic 18S rRNA gene, and thus gain insight into phytoplankton ecophysiology. Compared to diatoms and cryptophytes, laboratory data suggest that *Phaeocystis* is well adapted to variability in light availability (50). This photoacclimation capacity could result from a greater plasticity in pigments per chloroplast (51), or chloroplasts per cell under different light regimes. The latter strategy could explain the variability in chloroplast 16S vs. genomic 18S rRNA gene reads in *Phaeocystis* observed in our study. As shown in Figure 6E, the ratios of *Phaeocystis* chloroplast 16S/ genomic 18S rRNA gene generally decreased from north to south. Phytoplankton physiology is influenced by sea ice dynamics at the WAP (52)(53). Considering that the ice generally retreated from north to south, the southern communities closer to the ice edge might have been more recently exposed to higher light levels. The northern communities on the other

hand had been in open water for a longer period of time, being exposed to stronger wind-induced vertical mixing, and were therefore more likely to be light-limited. This may explain the higher chloroplast 16S/ genomic18S rRNA gene ratios in the south. These geographic variations were consistent with changes in the relative abundances of two *Phaeocystis* subclades (Figure 6F) which may be adapted to different light conditions. The correlation to mixed-layer depth was not as strong as to the geographic gradients (Figure S2). Overall, the chloroplast 16S/ genomic 18S rRNA gene ratio could prove to be a valuable indicator of *in situ* algal photophysiology adaptations when combined with laboratory experiments for further validation.

Limitations of ISN

There are several limitations to ISN. The first issue is associated with the extraction efficiency. Since the extraction efficiency is never 100%, the calculated rRNA gene abundance represents a lower bound on the true abundance. This could partially be addressed by spiking in cells instead of genomic DNA, although cell standards could also introduce biases due to 1) differences in extraction efficiency between the standard cells and the natural cells, and 2) variability in number of genomes per cell (42). A second issue is the high uncertainty in rrn correction (54), which is only relevant when converting rRNA gene copy numbers to cell numbers or when combining groups with mixed rrn. For example, large eukaryotes such as some dinoflagellates could have high rrn (> 1000 copies per cell) (55) and thus their 18S rRNA gene abundances could be orders of magnitude higher than their cell numbers. However, should a specific OTU have a constant rrn, the relative changes in absolute abundances across samples will still be captured because the copy numbers are proportional to the cell density. As the rrn is more comparable at finer taxonomic levels (56), it is best to apply the rrn normalization down to single genotypes. Defining OTUs at coarse taxonomic levels may combine groups with differing

301 rrns. In this case, the rRNA gene copy numbers are no longer proportional to the true cell
302 numbers thus complicating the interpretation of the rRNA gene counts. Finally, a third issue is
303 that some eukaryotic species have high plasticity in rrn (57). Variability in their 18S rRNA gene
304 counts may not reflect variability in their cell numbers. On the other hand, positive correlation of
305 rrn versus cell biovolume have been reported across different eukaryotic plankton taxon
306 including diatoms and dinoflagellates (54,52). If this relation is valid, groups with different rrn
307 could be combined, and the rRNA gene copy numbers could be used as an index for group
308 specific biomass. This is important because biomass is often of more relevance to
309 biogeochemical budgets (e.g. carbon, nitrogen) than cell numbers.
310 PCR bias could skew the relative abundances of mixed community members estimated
311 from the PCR products (58, 59). One main concern specific to our approach is the biased PCR
312 amplification caused by the varying template GC contents. Due to the triple hydrogen bonds
313 between G and G, templates with higher GC contents have higher melting temperatures and are
314 less efficiently amplified (59, 60). *T. thermophilus*, the 16S rRNA internal standard used in our
315 study, has a high GC content (69% for whole genome (61) and 65% for the amplified V4 region).
316 High GC content can cause underestimation of the internal standard abundance and
317 overestimation of the natural community member abundance. A second concern is the
318 amplification bias introduced by the degenerate primers. DNA sequences with G/C at the
319 degenerate position can be over-amplified compared to sequences with A/T. The deviation in
320 PCR product due to a single base difference at the priming site could be over 100% after 35 PCR
321 cycles (58). Various methods have been developed to reduce PCR biases: combining PCR
322 replicates (combined triplicates in this study), minimizing PCR cycle numbers and the
323 degeneracy of primers, and reconditioning PCR (62). On the other hand, despite the significant

PCR biases, inter-sample variability could still be precisely captured by the PCR method (58). A time-series study reported that PCR primer selection affects the estimated population abundances but not the community dynamic patterns (63). Although the abundance estimates by PCR based ISN may deviate from the absolute cell numbers due to PCR bias and rrn issues, the estimated inter-sample variability is less affected. Hence, this may not be as much of an issue for correlation analyses, e.g., time series community dynamics, community co-occurrence, and correlations to environmental variables.

Conclusions

Addition of internal standards to the amplicon rRNA gene sequencing approach allowed us to quantitatively compare microbial communities across different samples, as well as phytoplankton chloroplast 16S and genomic 18S rRNA gene abundances. Conceptually, the ISN could provide information equivalent to qPCR measurements targeting rRNA genes but with the advantage of examining a diverse community in a single assay. In our case study at the WAP, significant correlations observed in phytoplankton abundances based on 18S rRNA gene vs. CHEMTAX abundances and in total bacteria abundances based on 16S rRNA gene vs. FCM counts confirm that the ISN is quantitative. Our study also shows that chloroplast 16S rRNA gene sequences could be used to estimate phytoplankton abundances, and that the chloroplast 16S to genomic 18S rRNA gene ratio may be an insightful indicator of phytoplankton *in situ* photophysiology. The ISN comes at a minimal cost of implementation, and could be applied in conjunction with metagenomics (64). Overall, the ISN allows for an improved statistical, and ultimately ecological, interpretation of the rich and rapidly expanding marine microbiome datasets. More broadly, this approach could be valuable to researchers interested in relating

347 microbial ecology to quantitative processes such as microbial interactions, metabolic rates,
348 energy and material fluxes, and eventually quantitative ecosystem modeling.
349

350 **Materials and methods**

351 *DNA extraction with internal standard DNA addition*

352 Samples for DNA extraction were collected by seawater filtration (details see
353 Supplementary Information). Each filter with recorded filtration volume (4 L for most samples)
354 was split into two with one half for DNA extraction and the other half stored for later RNA work.
355 We note that this step could introduce errors due to uneven cell distribution on filter. Just prior to
356 DNA extraction, gDNA from two organisms representing eukaryotic and prokaryotic taxa not
357 expected to be present in marine surface water samples were added to the tube containing the
358 sample filter and lysis buffer (see below for optimization of internal standard addition). For the
359 18S rRNA gene internal standard, 50 µl of *Schizosaccharomyces pombe* gDNA (ATCC
360 #24843D-5, Manassas, VA, USA) at 0.322 ng/µl was spiked into each sample. For the 16S rRNA
361 gene internal standard, 50 µl of *Thermus thermophiles* gDNA (ATCC 27634D-5) at 0.297 ng/µl
362 was added to each sample. The internal standard working solutions were made in single use
363 aliquots to avoid DNA being lost during freeze-thaw cycles. gDNA standard stock solutions and
364 dilution concentrations were measured using a Qubit 3.0 fluorometer (Thermo Fisher Scientific,
365 Waltham, MA, USA). After spiking in internal standards, DNA extraction was performed as
366 described in (8).
367 *Optimizing the amount of internal standard added to a sample*

In order to get enough standard signal without overwhelming the environmental signal, we added the internal genomic DNA (gDNA) standards targeting a final concentration of around 1% of the total 16S and 18S rRNA gene reads. The amount of the prokaryotic genomic internal standard to spike in was based on the anticipated total extracted DNA mass as estimated with trial samples (22). For example, if we expected 10 μg of total genomic DNA in the sample, we added 100 ng of prokaryotic gDNA internal standard. Because the fraction of eukaryotic gDNA in total community DNA and the eukaryotic rRNA gene copy numbers per unit weight of gDNA are highly variable in different marine environments, a test sequencing run was conducted to optimize the internal standard amount to be spiked in. Test libraries were constructed with representative samples spiked with different amounts of internal eukaryotic genomic standard (16.1 ng or 3.22 ng) *Schizosaccharomyces pombe* gDNA (Figure 1B). The test amplicon libraries were subsequently sequenced using Illumina MiSeq platform (nano format) as a customized run at Duke Institute for Genomic Sciences and Policy (IGSP) with 300 bp single coverage forward reads and 10bp reverse reads to read the reverse barcodes. The averaged read count per sample was 50,661 after demultiplexing (see supplementary Table S4).

Amplicon library construction

16S rRNA genes were amplified by PCR using V4 primer set 515F (5' – GTGYCAGCMGCCGCGGTAA – 3') (65) and 805R (5' – GACTACNVGGGTATCTAAT – 3') modified from (66) and (67). 18S rRNA genes were amplified by PCR using V4 primer set EukF (5' – CCAGCASCYGCGGTAATTCC – 3') (70) and EukR (5' – ACTTTCGTTCTTGAT – 3') modified from (70) as described in (8) to increase coverage for Haptophytes.

Dual indexed fusion primers had 6-bp barcodes at each end constructed using error proof Hamming codes (71). In order to improve the "low sequence diversity" issue of the rRNA

amplicon library, 0 – 5 bp heterogeneity spacers were added to each primer (72). PCR were performed in triplicates for each sample. 18S rRNA gene PCR and library pooling were performed as described in (8). 16S rRNA gene library construction was similar to that of 18S rRNA gene except that 2U of Platinum Taq DNA Polymerase High Fidelity (Invitrogen) were added to each reaction, and PCR annealing temperature was 60 °C.

Amplicon libraries were sequenced at Duke IGSP using Illumina MiSeq 250PE platform for 16S rRNA amplicons and MiSeq 300PE platform for 18S rRNA amplicons. For each library, reads per sample after multiplexing were reported in Supplementary Table S4.

Bioinformatic analysis

For each library, paired-end reads were assembled using VSEARCH v2.3.4 (73) with quality score of the merged bases calculated following (74). Assembled reads were further processed using USEARCH (75) and QIIME (76) following (8). In brief, 16S or 18S rRNA gene reads were quality-controlled including quality filtering and chimera checking, and then were trimmed for barcodes and primer sequences. Singletons were discarded. OTUs (97% similarity) were then clustered using USEARCH and the representative sequences were assigned taxonomy based on the SILVA SSU database 128 using QIIME.

For 16S rRNA gene library, sequences identified in SILVA as mitochondria were removed. Sequences identified as chloroplast were filtered out as a separate data set. In order to further identify the phytoplankton host taxonomy from the chloroplast sequences, representative chloroplast sequences were blasted against the NCBI nucleotide collection database using BLAST+ 2.6.0 (77). The top three hits for each sequence were reported in Table S5.

Accession numbers

Sequences were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive under the BioProject accession numbers PRJNA508517 and PRJNA508514.

Acknowledgements

This work was supported by "Laboratoire d'Excellence" LabexMER (ANR-10-LABX-19) in France, and the US NSF OPP-1043339 to Cassar. This research was undertaken during the January 2012, 2013 and 2015 cruises by the Palmer Antarctica Long Term Ecological Research (PAL-LTER) project, supported by NSF awards OPP-0823101 and 1440435 to Ducklow, and the NASA ROSES award NNX14AL86G to Schofield. We thank Bruce Barnett, Rachel Eveleth and Naomi Shelton for sampling, and all the scientists and the crew on R/V L.M.Gould for shipboard assistance. We are also grateful to Damien Eveillard and Samuel Chaffron for insightful discussion and comments.

Author contributions

Y. L. and N. C. conceived and designed the study. H. D., O. S., and Y. L. collected field samples and underway data. Y. L. processed the DNA samples and analyzed the data. Y. L., N. C., and S. G. wrote the manuscript with contributions from all other authors.

Author Information

The authors declare that they have no competing financial interests to the work described.

Correspondence and requests for materials should be addressed to Y. L. (yajuan.lin@duke.edu) and N.C. (Nicolas.Cassar@duke.edu).

References

1. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored “rare biosphere.” Proc Natl Acad Sci 103:12115–12120.

2. Zinger L, Amaral-Zettler LA, Fuhrman JA, Horner-Devine MC, Huse SM, Welch DBM, Martiny JBH, Sogin M, Boetius A, Ramette A. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS One 6:e24570.

3. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science (80-) 348:1261605.

4. Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, Darzi Y, Audic S, Berline L, Brum JR. 2016. Plankton networks driving carbon export in the oligotrophic ocean. Nature.

5. Pernice MC, Giner CR, Logares R, Perera-Bel J, Acinas SG, Duarte CM, Gasol JM, Massana R. 2016. Large variability of bathypelagic microbial eukaryotic communities across the world’s oceans. ISME J 10:945–958.

6. Kopf A, Bicak M, Kottmann R, Schnetzer J, Kostadinov I, Lehmann K, Fernandez-Guerra A, Jeanthon C, Rahav E, Ullrich M. 2015. The ocean sampling day consortium. Gigascience 4:27.

7. Fuhrman JA, Cram JA, Needham DM. 2015. Marine microbial community dynamics and their ecological interpretation. Nat Rev Microbiol 13:133–146.

8. Lin Y, Cassar N, Marchetti A, Moreno C, Ducklow H, Li Z. 2017. Specific eukaryotic plankton are good predictors of net community production in the Western Antarctic Peninsula. Sci Rep 1.

9. Quinn T, Richardson MF, Lovell D, Crowley T. 2017. propr: An R-package for Identifying Proportionally Abundant Features Using Compositional Data Analysis. bioRxiv 104935.

10. van den Boogaart KG, Tolosana-Delgado R. 2008. “compositions”: A unified R package to analyze compositional data. Comput Geosci 34:320–338.

11. McMurdie PJ, Holmes S. 2014. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol 10:e1003531.

12. Aitchison J. 1982. The statistical analysis of compositional data. J R Stat Soc Ser B 139–177.

13. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, Xia LC, Xu ZZ, Ursell L, Alm EJ. 2016. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. ISME J 10:1669–1681.

14. Li H. 2015. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Its Appl 2:73–94.

15. Lin W, Shi P, Feng R, Li H. 2014. Variable selection in regression with compositional covariates. Biometrika 101:785–797.

16. Pearson K. 1896. Mathematical contributions to the theory of evolution. III. Regression, heredity, and panmixia. Philos Trans R Soc London Ser A, Contain Pap a Math or Phys character 187:253–318.

17. Logares R, Audic S, Bass D, Bittner L, Boutte C, Christen R, Claverie J-M, Decelle J, Dolan JR, Dunthorn M, Edvardsen B, Gobet A, Kooistra WHCF, Mahé F, Not F, Ogata H, Pawlowski J, Pernice MC, Romac S, Shalchian-Tabrizi K, Simon N, Stoeck T, Santini S, Siano R, Wincker P, Zingone A, Richards TA, de Vargas C, Massana R. 2014. Patterns of Rare and Abundant Marine Microbial Eukaryotes. Curr Biol 24:813–821.

18. Anders S, Huber W. 2010. Differential expression analysis for sequence count data. Genome Biol 11.

19. Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.

20. Props R, Kerckhof F-M, Rubbens P, De Vrieze J, Sanabria EH, Waegeman W, Monsieurs P, Hammes F, Boon N. 2017. Absolute quantification of microbial taxon abundances. ISME J 11:584.

21. Vandeputte D, Kathagen G, D'hoe K, Vieira-Silva S, Valles-Colomer M, Sabino J, Wang J, Tito RY, De Commer L, Darzi Y, Vermeire S, Falony G, Raes J. 2017. Quantitative microbiome profiling links gut community variation to microbial load. Nature.

22. Satinsky BM, Gifford SM, Crump BC, Moran MA. 2013. Use of internal standards for quantitative metatranscriptome and metagenome analysis. Methods Enzymol 531:237–250.

23. Gifford SM, Sharma S, Rinta-Kanto JM, Moran MA. 2011. Quantitative analysis of a deeply sequenced marine microbial metatranscriptome. ISME J 5:461–472.

24. Smets W, Leff JW, Bradford MA, McCulley RL, Lebeer S, Fierer N. 2016. A method for

simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. Soil Biol Biochem 96:145–151.

25. Stämmler F, Gläsner J, Hiergeist A, Holler E, Weber D, Oefner PJ, Gessner A, Spang R. 2016. Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. Microbiome 4:28.

26. Huang K, Ducklow H, Vernet M, Cassar N, Bender ML. 2012. Export production and its regulating factors in the West Antarctica Peninsula region of the Southern Ocean. Global Biogeochem Cycles 26.

27. Mackey MD, Mackey DJ, Higgins HW, Wright SW. 1996. CHEMTAX—a program for estimating class abundances from chemical markers: application to HPLC measurements of phytoplankton. Mar Ecol Prog Ser 265–283.

28. Henne A, Brüggemann H, Raasch C, Wiezer A, Hartsch T, Liesegang H, Johann A, Lienard T, Gohl O, Martinez-Arias R, Jacobi C, Starkuviene V, Schlenczeck S, Dencker S, Huber R, Klenk H-P, Kramer W, Merkl R, Gottschalk G, Fritz H-J. 2004. The genome sequence of the extreme thermophile Thermus thermophilus. Nat Biotechnol 22:547.

29. Wood V, Gwilliam R, Rajandream M-A, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S. 2002. The genome sequence of Schizosaccharomyces pombe. Nature 415:871–880.

30. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. 2001. rrndb: the ribosomal RNA operon copy number database. Nucleic Acids Res 29:181–184.

31. Straza TRA, Ducklow HW, Murray AE, Kirchman DL. 2010. Abundance and single-cell

activity of bacterial groups in Antarctic coastal waters. Limnol Oceanogr 55:2526–2536.

32. Wietz M, Gram L, Jørgensen B, Schramm A. 2010. Latitudinal patterns in the abundance of major marine bacterioplankton groups. Aquat Microb Ecol 61:179–189.

33. Thiele S, Fuchs BM, Ramaiah N, Amann R. 2012. Microbial community response during the iron fertilization experiment LOHAFEX. Appl Environ Microbiol 78:8803–8812.

34. Smith CJ, Osborn AM. 2009. Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. FEMS Microbiol Ecol 67:6–20.

35. Karlen Y, McNair A, Perseguers S, Mazza C, Mermod N. 2007. Statistical significance of quantitative PCR. BMC Bioinformatics 8:131.

36. Morrison TB, Weis JJ, Wittwer CT. 1998. Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. Biotechniques 24.

37. Zinser ER, Coe A, Johnson ZI, Martiny AC, Fuller NJ, Scanlan DJ, Chisholm SW. 2006. Prochlorococcus ecotype abundances in the North Atlantic Ocean as revealed by an improved quantitative PCR method. Appl Environ Microbiol 72:723–732.

38. Labrenz M, Brettar I, Christen R, Flavier S, Bötel J, Höfle MG. 2004. Development and application of a real-time PCR approach for quantification of uncultured bacteria in the central Baltic Sea. Appl Environ Microbiol 70:4971–4979.

39. Gifford SM, Becker JW, Sosa OA, Repeta DJ, DeLong EF. 2016. Quantitative transcriptomics reveals the growth-and nutrient-dependent response of a streamlined marine methylotroph to methanol and naturally occurring dissolved organic matter. MBio 7:e01279-16.

40. Schofield O, Saba G, Coleman K, Carvalho F, Couto N, Ducklow H, Finkel Z, Irwin A, Kahl A, Miles T. 2017. Decadal variability in coastal phytoplankton community composition in a changing West Antarctic Peninsula. Deep Sea Res Part I Oceanogr Res Pap 124:42–54.

41. Delmont TO, Hammar KM, Ducklow HW, Yager PL, Post AF. 2014. Phaeocystis antarctica blooms strongly influence bacterial community structures in the Amundsen Sea polynya. Front Microbiol 5:646.

42. Sargent EC, Hitchcock A, Johansson SA, Langlois R, Moore CM, LaRoche J, Poulton AJ, Bibby TS. 2016. Evidence for polyploidy in the globally important diazotroph Trichodesmium. FEMS Microbiol Lett 363.

43. Stoddard SF, Smith BJ, Hein R, Roller BRK, Schmidt TM. 2014. rrn DB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. Nucleic Acids Res 43:D593–D598.

44. Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J. 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science (80-) 336:608–611.

45. Williams TJ, Wilkins D, Long E, Evans F, DeMaere MZ, Raftery MJ, Cavicchioli R. 2013. The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. Environ Microbiol 15:1302–1317.

46. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016.

564 DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods
565 13:581.

566 47. Jennrich RI. 1970. An asymptotic χ2 test for the equality of two correlation matrices. J Am
567 Stat Assoc 65:904–912.

568 48. Steiger JH. 1980. Tests for comparing elements of a correlation matrix. Psychol Bull
569 87:245.

570 49. Needham DM, Sachdeva R, Fuhrman JA. 2017. Ecological dynamics and co-occurrence
571 among marine phytoplankton, bacteria and myoviruses shows microdiversity matters.
572 ISME J.

573 50. Arrigo KR, Mills MM, Kropuenske LR, van Dijken GL, Alderkamp A-C, Robinson DH.
574 2010. Photophysiology in two major Southern Ocean phytoplankton taxa: photosynthesis
575 and growth of Phaeocystis antarctica and Fragilariopsis cylindrus under different
576 irradiance levels. Integr Comp Biol 50:950–966.

577 51. Moisan TA, Ellisman MH, Buitenhuys CW, Sosinsky GE. 2006. Differences in chloroplast
578 ultrastructure of Phaeocystis antarctica in low and high light. Mar Biol 149:1281–1290.

579 52. Montes-Hugo M, Doney SC, Ducklow HW, Fraser W, Martinson D, Stammerjohn SE,
580 Schofield O. 2009. Recent changes in phytoplankton communities associated with rapid
581 regional climate change along the western Antarctic Peninsula. Science (80-) 323:1470–
582 1473.

583 53. Obryk MK, Doran PT, Friedlaender AS, Gooseff MN, Li W, Morgan-Kiss RM, Priscu JC,
584 Schofield O, Stammerjohn SE, Steinberg DK. 2016. Responses of Antarctic marine and

freshwater ecosystems to changing ice conditions. Bioscience 66:864–879.

54. Louca S, Doebeli M, Parfrey LW. 2018. Correcting for 16S rRNA gene copy numbers in microbiome surveys remains an unsolved problem. Microbiome 6:41.

55. Godhe A, Asplund ME, Härnström K, Saravanan V, Tyagi A, Karunasagar I. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Appl Environ Microbiol 74:7174–7182.

56. Angly FE, Dennis PG, Skarshewski A, Vanwonterghem I, Hugenholtz P, Tyson GW. 2014. CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. Microbiome 2:11.

57. Gibbons JG, Branco AT, Godinho SA, Yu S, Lemos B. 2015. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. Proc Natl Acad Sci 112:2485–2490.

58. Polz MF, Cavanaugh CM. 1998. Bias in template-to-product ratios in multitemplate PCR. Appl Environ Microbiol 64:3724–3730.

59. Suzuki MT, Giovannoni SJ. 1996. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. Appl Environ Microbiol 62:625–630.

60. Reysenbach A-L, Giver LJ, Wickham GS, Pace NR. 1992. Differential amplification of rRNA genes by polymerase chain reaction. Appl Environ Microbiol 58:3417–3418.

61. Oshima T, Imahori K. 1974. Description of Thermus thermophilus (Yoshida and Oshima) comb. nov., a nonsporulating thermophilic bacterium from a Japanese thermal spa. Int J Syst Evol Microbiol 24:102–112.

606 62. Thompson JR, Marcelino LA, Polz MF. 2002. Heteroduplexes in mixed-template
607 amplifications: formation, consequence and elimination by ‘reconditioning PCR.’ Nucleic
608 Acids Res 30:2083–2088.

609 63. Wear EK, Wilbanks EG, Nelson CE, Carlson CA. 2018. Primer selection impacts specific
610 population abundances but not community dynamics in a monthly time-series 16S rRNA
611 gene amplicon analysis of coastal marine bacterioplankton. Environ Microbiol.

612 64. Satinsky BM, Crump BC, Smith CB, Sharma S, Zielinski BL, Doherty M, Meng J, Sun S,
613 Medeiros PM, Paul JH. 2014. Microspatial gene expression patterns in the Amazon River
614 Plume. Proc Natl Acad Sci 111:11085–11090.

615 65. Parada AE, Needham DM, Fuhrman JA. 2016. Every base matters: assessing small
616 subunit rRNA primers for marine microbiomes with mock communities, time series and
617 global field samples. Environ Microbiol 18:1403–1414.

618 66. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. 2011.
619 PrimerProspector: de novo design and taxonomic analysis of barcoded polymerase chain
620 reaction primers. Bioinformatics 27:1159–1161.

621 67. Apprill A, McNally S, Parsons R, Weber L. 2015. Minor revision to V4 region SSU rRNA
622 806R gene primer greatly increases detection of SAR11 bacterioplankton.

623 68. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ,
624 Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of
625 sequences per sample. Proc Natl Acad Sci 108:4516–4522.

626 69. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO.

2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

70. Stoeck T, Bass D, Nebel M, Christen R, Jones MDM, BREINER H, Richards TA. 2010. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. Mol Ecol 19:21–31.

71. Bystrykh L V. 2012. Generalized DNA barcode design based on Hamming codes. PLoS One 7:e36852.

72. Fadrosh DW, Ma B, Gajer P, Sengamalay N, Ott S, Brotman RM, Ravel J. 2014. An improved dual-indexing approach for multiplexed 16S rRNA gene sequencing on the Illumina MiSeq platform. Microbiome 2:1–7.

73. Rognes T, Flouri T, Nichols B, Quince C, Mahé F. 2016. VSEARCH: a versatile open source tool for metagenomics. PeerJ 4:e2584.

74. Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics 31:3476–3482.

75. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

76. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI. 2010. QIIME allows analysis of high-throughput community sequencing data. Nat Methods 7:335–336.

77. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

648