Contribution Evenness: A functional redundancy metric for microbially-mediated biogeochemical rates and processes

Taylor M. Royalty (1,2) and Andrew D. Steen (2,1)*

University of Tennessee Departments of (1) Earth and Planetary Sciences and (2) Microbiology

Running title:

Measuring the functional redundancy of microbiomes

*Corresponding: asteen1@utk.edu

Abstract

The concept of functional redundancy has received considerable attention in both the macroecology and microbial ecology literature. As a result, multiple metrics of functional redundancy have been proposed. These vary in how they weight trait levels, species abundance, functional richness, and species richness. Trait-based models are also increasingly popular ways to represent the function of microbial ecosystems. Here we present a new functional redundancy measurement tailored to be useful for community-aggregated parameters, which are summed ecosystem processes contributed by species, such as rates of chemical transformations. We focus in particular on community-aggregated parameters mediated by microbial communities. We call this approach contribution evenness (CE). CE measures how evenly species contribute to a community-aggregated parameter based on species contribution levels and abundances. As CE is an evenness measurement, quantitative representation of functional redundancy spans between 0 and 1, where 0 corresponds to a single species contributing and 1 corresponds to all species contributing equally to a community-aggregated parameter. CE is sensitive to the resilience of an ecosystem function to species extinction—an ecological consequence of functional redundancy. To illustrate how CE can be used, we analyzed the functional redundancy of eight nitrogen-transforming pathways using 2,631 metagenome-assembled genomes from 47 marine sites. We found that, on average, NH_4^+ assimilation pathway was the most functionally redundant (0.44 ± 0.08) while dissimilatory nitrate reduction had the lowest functional redundancy (0.005 ± 0.005). As demonstrated here, CE is a metric that is sensitive to trait resilience and allows for comparative analysis of functional redundancy for different traits.

Introduction

Functional redundancy was originally proposed as a community property that reflected the resilience of an ecosystem function (Naeem, 1998). Species performing a similar function could compensate for one another in the event of local species extinction, and thus, those functions are more resilient to taxa extinction. The purpose of measuring functional redundancy was to focus conservation efforts on vulnerable ecological groups that provide essential and unique ecological functions—i.e., keystone species (Walker, 1992). Since its proposal, interest in measuring functional redundancy has extended beyond conservation biology. In both the macroecology and microbial ecology literature, functional redundancy has measured in studies evaluating community succession, intra-community resource competition, biogeochemical fluxes, and metabolism biodiversity (Louca et al., 2018). Since measurements of functional redundancy have been used for diverse purposes in the ecological literature, it is helpful to identify metrics for functional redundancy that are tailored to purpose.

Several metrics for functional redundancy exist in the literature. Functional redundancy has been taken as the ratio of community richness and the theoretical maximum number of functional entities (Mouillot et al., 2014), median metabolic overlap of genomic traits (Hester, Jetten, Welte, & Lücker, 2019), expected variability in trait space among community members (Ricotta et al., 2016), the degree of covariance in functional diversity and biodiversity (e.g., Galand et al., 2018; Micheli & Halpern, 2005; Miki, Yokokawa, & Matsui, 2013), and the simple fraction of a community with a given trait (Trivedi et al., 2020; Tully, Wheat, Glazer, & Huber, 2018; Wohl, Arora, & Gladstone, 2004). Notably, these different metrics treat community structure differently. For instance, some metrics are sensitive to species abundances (Galand et al., 2018; Ricotta et al., 2016), richness (Micheli & Halpern, 2005; Mouillot et al., 2014), or

neither (Hester et al., 2019). The diversity of metrics reflects the diverse sources of data used to calculate functional redundancy and the diverse research purposes for how this concept is applied in research.

With exception of measuring a fraction of a community with a trait, all the functional redundancy metrics listed above represent ecological function in terms of niche, where a niche is thought of as a point in a hyperdimensional trait space. Communities with higher species density in trait space are interpreted as being more functionally redundant. For the metrics mentioned above, high species richness relative to unique combinations of traits, high overlap of metabolic pathways, low variance of community members' position in trait space, and resilience of community function to disturbance of community structure all suggest functionally redundant communities.

An ecological function we are interested in understanding is the relative contribution by species towards community-aggregated parameters. Community-aggregated parameters are ecosystem functions (e.g. primary productivity, carbon turnover, nitrogen fixation, etc.) observed on the community level, where the contribution by individual species is weighted by species abundance and summed together by species-level contributions (Violle et al., 2007). Thus, we argue one way to understand a species' ecological function is its contribution towards a community-aggregated parameter, or the community-level ecosystem function. Describing ecological function in terms of community-aggregated parameters would be particularly valuable to metagenomic studies of microbial communities (Fierer, Barberán, & Laughlin, 2014; Krause et al., 2014).

However, it is challenging to assign functions to microbes. Most microbial cells on Earth belong to genera which lack cultured representatives, so the phenotypes for these organisms are

often unknown (Lloyd, Steen, Ladau, Yin, & Crosby, 2018; Staley & Konopka, 1985). Function can sometimes be inferred experimentally (e.g., Orsi et al., 2018), but in most cases, information about microbial traits comes exclusively from sequencing studies, where genomic traits are proxies for microbial metabolisms. For microbial communities, we are often most concerned with traits that relate to environmental processes, such as elemental transformations or secondary metabolite production (Louca et al., 2018). Thus, it may be preferential to evaluate a small subset of specific traits versus the full hyperdimensional trait space (thousands of genes, and a difficult-to-describe number of metabolic processes, per genome) of a community.

As functional redundancy is a measure of species density with respect to an ecological function, enumerating the number of "contributors" towards a community-aggregated parameter may suffice as a measure of functional redundancy. Although not explicitly stated as such, some methods we mentioned before use this approach (Trivedi et al., 2020; Tully, Wheat, et al., 2018; Wohl et al., 2004). We note that simple enumeration does not weight the variable contributions by different species, which may be desirable. Here, we propose a functional redundancy measurement that accounts for variable contributions by different species towards community-aggregated parameters. With this approach, the "true diversity" of contributors to a community-aggregated parameter is calculated with traditional diversity metrics, where entropy is converted into an effective richness of contributors (Hill, 1973; Jost, 2006). The observed diversity is normalized to a theoretical maximum diversity and made "unrelated" to community richness (Chao & Ricotta, 2019). Importantly, the normalization steps allow for meaningful comparisons between communities with different species richness. Our measure of functional redundancy, which we term contribution evenness (CE) for conciseness, is a measure of how evenly species contribute to a community-aggregated parameter.

To demonstrate the efficacy of CE, we performed two *in silico* experiments with artificial communities in parallel. Both experiments evaluated the extent that CE recapitulated the resilience of a community-aggregated parameter to extinction of species. The first experiment sought to minimize variance between CE and trait resilience to species extinction by changing the diversity order (Hill, 1973; Jost, 2006) used to calculate CE. The next experiment compared CE with other functional redundancy metrics as a demonstration that CE is fundamentally different from previously proposed metrics. Last, we analyzed functional redundancy for eight nitrogen-transforming pathways found in metagenomic data (Pajares & Ramos, 2019) in *TARA* Oceans draft metagenome-assembled genomes (MAGs) (Tully, Graham, & Heidelberg, 2018). We found CE appealing because it related to trait resilience to species extinction and was easily generalizable—the functional redundancy among different genomic traits were comparable to one another while not making assumptions about multi-trait guilds. We hope that the formulation of functional redundancy fills a void in the genomic ecology literature by providing a generalized tool for generating higher quality and comparable measurements of functional redundancy.

Methods and Theory

Definition of Functional Redundancy

Here, we define a microbial community, C, with S species, where the ith species has a relative abundance, p_i. (For simplicity, we refer to ecological units as species, although it would be possible to perform this analysis at other taxonomic levels – strains, say, or genera – or even to perform the analysis at the level of guilds in order to evaluate resource competition among guilds, e.g., sulfate-reducers competing with methanogens for hydrogen electrons). We therefore represent C as:

123 $$C = \{p | p_i > 0 \; \forall \; p_i; \sum_{i=1}^{S} p_i = 1\} \quad (1$$

124 where p is the set of relative abundances of species, such that any given relative abundance for
125 the ith species is greater than zero and all species relative abundances sum to one. Equation 1
126 represents a traditional abundance distribution (McGill et al., 2007).

127 Our goal was to calculate how evenly different species in a community, C (equation 1),
128 contribute to a community-aggregated parameter, or net ecosystem function, T, at a fixed time
129 and space. To do this, we first determined the relative contribution that the ith species makes
130 towards T, defined here as τ_i, such that:

131 $$\tau_i = \frac{p_i t_i}{\sum_{i=1}^{n} p_i t_i} \quad 2)$$

132 where t_i is the species-specific contribution level and p_i is the relative abundance the ith species in
133 C. In equation 2, τ_i represents the relative contribution that the ith species, or population,
134 contributes to a community-aggregated parameter. Units of t are flexible such that t can be
135 expressed in terms of rates (e.g., g C m^{-2} $year^{-1}$ for net primary productivity), quantities (e.g., net
136 gene copies in a metagenome), or dimensionless numbers (e.g., a relative proportion of rates or
137 quantities); however, T must be quantitative.

138 As an example, imagine a hypothetical community with three species, a Cyanobacterium
139 (p=0.5), an Alphaproteobacterium (p=0.25), and a Gammaproteobacterium (p=0.25). We wish to
140 calculate the functional redundancy of this community with respect to production of some
141 secondary metabolite. We use the presence of a gene, *xxxA*, as a proxy for an organisms'
142 potential to produce this secondary metabolite. The average number of *xxxA* gene copies in the
143 genomes of a Cyanobacterium, Alphaproteobacterium, and Gammaproteobacterium, are 5, 5,

144 and 0, respectively. Thus, the relative contribution from each species to secondary metabolite

145 production, when using *xxxA* as a proxy, is τ=0.67, τ=0.33, and τ=0, respectively

146 Next, we reformulate the abundance distribution for community, *C*, as an abundance

147 distribution with respect to τ (relative contribution distribution), such that:

148 $$C_{\tau,all} = \left\{\tau \middle| \tau_i \geq 0 \; \forall \; \tau_i; \sum_{i=1}^{n} \tau_j = 1\right\} \quad (3.1$$

149 and

150 $$C_{\tau} = \left\{\tau \in C_{\tau,all} \middle| \tau_i > 0 \; \forall \; \tau_i; \sum_{i=1}^{\sigma} \tau_i = 1\right\} \quad (3.2$$

151 where again, τ_i is the relative contribution by the ith species to *T*, such that all τ sum to equal 1.

152 The community represented by equation 3.1 includes species not contributing to the community-

153 aggregated parameter. The community represented by equation 3.2 is a subset of equation 3.1,

154 where C_τ contains only species from $C_{\tau,all}$ contributing towards a community-aggregated

155 parameter, which we refer to as "contributors." As such, we adopted the variable, σ, which

156 represents the richness of contributors. The effective richness of contributors towards *T* was

157 calculated using equation 3.2 and the HCDT entropy index (Jost, 2006) such that:

158 $$D_q = \begin{cases} \left(\sum_{i=1}^{\sigma} {\tau_i}^q\right)^{\frac{1}{(1-q)}} & q \neq 1 \\ exp\left(\sum_{i=1}^{\sigma} -\tau_i \log \tau_i\right) & q = 1 \end{cases} \quad (4$$

159 where τ_i is the relative contribution by i^{th} species in C_τ towards *T*, and D_q is the effective richness

160 of contributors, and *q* is the diversity order, which weights the importance of species contributing

161 more or less towards *T* when calculating D_q. This is analogous to diversity order weighting the

162 importance of rare versus abundant species when measuring diversity of species abundances

163 (Hill, 1973). With diversity order *q*=0, equation 4 simplifies to the number of contributors in

equation 3.2 (i.e., σ). Diversity order q=0 removes any influence in the variability of τ when calculating D_q. Increasing the diversity order subsequently decreases the perceived contribution of contributors with smaller τ. Diversity orders q=1 and q=2 correspond to Shannon's index. and Simpson's index.

We then normalize D_q by community richness (equal to assuming every species contributes to T equally), S, and adjust evenness to range from 0 to 1 (Chao & Ricotta, 2019), such that:

$$R = \frac{D_q - 1}{S - 1} \quad (5$$

Here, functional redundancy, R, is taken as how evenly species contribute to a community-aggregated parameter, T. For the hypothetical community described above, the effective richness (D_q, q=0.5—see *Results*) of species possessing *xxxA* was ~1.55, and the functional redundancy, with respect to *xxxA*, was R=0.47 (q=0.5; see *Results*).

Simulations of Functional Redundancy as a Function of Resilience to Species Extinction

In order to test the fidelity with which CE measures resilience of a community-aggregated parameter, we simulated artificial communities with varying values of species richness (S), richness of contributors (σ), and contribution levels by contributors (t). First, the richness of contributors was randomly chosen from a uniformly distributed range of 1 to 100. Next, the richness of species not contributing to the community-aggregated parameter was randomly chosen from a uniformly distributed range of 0 to 100, so that some artificial communities were dominated by contributors while others by non-contributors. The relative abundance, p, of individual species was modeled log-normally (Galand, Casamayor, Kirchman, & Lovejoy, 2009; Magurran, 1988) such that:

186 $$R(r) = R_0 e^{-a^2 r^2} \quad (6$$

187 where R was the rth rank, R_0 was the maximum abundance, and a was the inverse width of the
188 lognormal distribution. Values of a were randomly selected from 0 (even) to 0.4 (more uneven),
189 at 0.01 intervals. After establishing species relative abundances, we randomly chose σ
190 contributing taxa from our lognormally distributed community. The contribution level, t, was
191 then randomly assigned to these species from a uniformly distributed range of 0.01 to 1, at
192 intervals of 0.01. We generated and calculated R for 1,000 artificial communities.

193 The resilience of a community-aggregated parameter to extinction was tested by a
194 simulation in which resilience to extinction was quantified as the average fraction of species
195 richness (S) that must go extinct in order to decrease community-aggregated parameter by 99%.
196 Species were randomly removed from the community until summed τ was reduced to 1% of its
197 original level. The extinction simulation was performed 100 times for each of the individual
198 1,000 artificial communities. The resilience to extinction generated from the 100 simulations
199 were then averaged for each of the individual 1,000 artificial communities. We made no *a priori*
200 assumptions about the optimal diversity order (i.e., q) to use in equations 4 and 5. We performed
201 the simulation described above varying diversity orders from 0 to 2 at 0.1 intervals. The best
202 diversity order was determined by fitting generalized additive model (GAM) regressions
203 between R and resilience to extinction, and then choosing the diversity order which resulted in a
204 GAM fit with the lowest mean absolute error among residuals. All GAM fits were performed
205 using the gam function from the R package mgcv (Wood, 2017). Last, we performed the same
206 simulations and calculated R using other methods described in the literature. A summary of these
207 measurements is provided in Table 1. An R package which calculates functional redundancy

208 (equations 1-5) and performs these simulations is available at

209 https://github.com/taylorroyalty/functional_redundancy_script.

210 *A Case Study of Functional Redundancy in the Global Ocean Nitrogen Cycle*

211 Draft metagenome-assembled genomes (MAGs) from (Tully, Graham, et al., 2018) were

212 downloaded from NCBI (Federhen, 2012). Data on metagenome sequencing effort (number of

213 short reads), MAG relative abundance estimates (pre-corrected for genome size), ocean depths,

214 and ocean regions, defined in (Tully, Graham, et al., 2018), were downloaded from Figshare

215 (https://doi.org/10.6084/m9.figshare.5188273). MAG relative abundances from each site were

216 multiplied by the respective site's sequencing effort to convert relative abundances into

217 abundance count data. Abundance counts for all sites were normalized by the lowest abundance

218 (greater than 0) among all sites. MAG abundances were then rarefied using the rrarefy function

219 from the R package vegan (Philip, 2003). We chose a sampling effort of 14,265, as this sampling

220 effort corresponded to the site with lowest sampling effort (i.e., sequencing effort). We tabulated

221 nitrogen-cycle marker genes in individual MAGs by blasting MAG proteins against nitrogen-

222 cycle marker gene databases using the BLASTP algorithm (*E*-value=10^{-10}) (Clark, Karsch-

223 Mizrachi, Lipman, Ostell, & Sayers, 2019). Nitrogen cycle marker genes were categorized into

224 eight categories: N_2 fixation (*nifH*), nitrification (*amoA*, *hao*), NO_3 reduction (*napA*, *narG*),

225 dissimilatory NO_3 reduction (*nrfA*), assimilatory NO_3/NO_2 reduction (*narB*, *nasA*, *nasB*, and

226 *nirB*), NH_4^+ assimilation (*gdhA*, *glnA*), N remineralization (*gltB*), and denitrification (*nirK*, *nirS*,

227 *norB*, and *nosZ*) (Pajares & Ramos, 2019). Individual BLAST databases were constructed for

228 each nitrogen cycle category. Gene names associated with individual categories were searched

229 for in the Swiss-Prot database (Bairoch & Apweiler, 2000). All proteins matching gene names

230 were subsequently downloaded and used in BLAST database construction. Functional

231 redundancy was then calculated for individual stations, at individual depths, for each nitrogen
232 cycle category using equations 1-5. We used a diversity order, q=0.5, based on the GAM
233 sensitivity analysis. Table 2 contains protein names, gene names, corresponding EC, and entry
234 name for all genes used in the above analysis.

235 **Results and Discussion**

236 *Representing microbial communities as species abundance, trait abundance, and relative*
237 *contribution distributions*

238 We represent microbial communities using two distributions: the distribution of species
239 abundances within a community, and the distribution of contribution levels among species. Here,
240 we evaluate genomic traits as these are regularly adopted as proxies for microbially-mediated
241 biogeochemistry (Anantharaman et al., 2016; Baker, Lazar, Teske, & Dick, 2015; Falkowski,
242 Fenchel, & Delong, 2008; Krause et al., 2014; Malik et al., 2019; Salazar et al., 2019; Teeling et
243 al., 2012). Because trait levels are measured as non-negative, real numbers (Violle et al., 2007),
244 the relative contribution of individual species to a community-aggregated parameter can be
245 represented as a relative abundance distribution (Figure 1). We determined the relative
246 contribution of individual species to the community-aggregated parameter from the product of
247 species relative abundances and their respective trait abundance distributions (equation 2). We
248 refer to the distribution of the relative contribution of each species to total as the relative
249 contribution distribution (equation 3). Thus, an abundant species that expresses a small amount
250 of trait per cell, and a rarer species that expresses a large amount of trait per cell, can contribute
251 equally with respect to a community-aggregated parameter (Figure 1).

252 *Defining, Quantifying Functional Redundancy, and Choosing a Diversity Order q*

We measured functional redundancy as how evenly species in a community contribute to a community-aggregated parameter and termed this metric as contribution evenness (CE). This definition is consistent with descriptions of functional redundancy as multiple species sharing similar ecological roles or functions (Naeem, 1998; Rosenfeld & Mall, 2002). The relative contribution by species towards a community aggregated-parameter (equation 2) provides a framework to calculate the effective number, or richness, of species that contribute to that community-aggregated parameter as an ecological function. This effective richness is calculated with traditional diversity indices (equations 3-4) (Hill, 1973; Jost, 2006). To determine the effective proportion of a community contributing to the community-aggregated parameter, the richness of contributors is normalized by species richness (Hill, 1973). Last, to make meaningful comparisons of evenness between communities with different levels of richness, we made our evenness metric unrelated to community richness (equation 5) (Chao & Ricotta, 2019).

Numerous diversity indices have been proposed in the literature. Three common measures of diversity include species richness, Shannon's diversity index, and Simpson's diversity index, are manifestations of equation 4 when the q term—commonly referred to as the diversity order—is q=0, q=1, and q=2, respectively. The diversity order changes the relative weight placed on rare versus common species when determining the effective richness. Here, we adopted a diversity order q=0.5, based on empirical results from our numerical extinction simulations (see *Functional Redundancy and Trait Resilience to Species Extinction* for details). We found that GAM fits between functional redundancy and resilience to extinction followed positive, monotonic functions for all diversity orders spanning q=0 to q=2. However, at q≈0.5, CE was most constrained (mean absolute error) with respect to resilience to extinction, our

benchmark for functional redundancy efficacy (Figure 2). As such, we thought it appropriate to select the diversity order maximizing efficacy with respect to our benchmark.

For $0.1<q\leq2$, diversity order only weakly affected mean absolute error. Diversity orders q=0 (richness), q=1 (Shannon's Index), and q=2 (Simpson's Index) correspond to the reciprocal harmonic, geometric, and arithmetic means of the taxa proportions (p_i or τ_i), respectively. These diversity orders are perceived as both intuitive and preferential diversity orders relative to empirically derived diversity orders (Hill, 1973). Given diversity orders $0.1<q\leq2$ do not alter our conclusions, we allow investigators to select a custom diversity order in the R package.

Functional Redundancy and Trait Resilience to Species Extinction

A tangible ecological consequence of functional redundancy is resilience of an ecosystem function to the extinction of species performing that function. When functions are more broadly distributed across species, a function is less likely to disappear due to the extinction of any subset of species. This interpretation is often highlighted in macro-ecological studies (McPherson, Yeager, & Baum, 2018; Naeem, 1998; Ricotta et al., 2016; Rosenfeld & Mall, 2002; Tilman, 2001). The conceptual relationship between functional redundancy and resilience of an ecosystem function provides an opportunity to demonstrate the fidelity of CE. To do this, CE was calculated for randomly generated communities. Individual species were randomly assigned a contribution levels to a community aggregated parameter. We simulated random species extinction events in the artificial communities until >99% of the original ecosystem function was removed from the community. This choice of threshold allowed for the evaluation of communities where all species were contributors. In such communities, a 100% extinction rate was necessary to remove the ecosystem function.

It is important to emphasize that these simulations were not designed to mimic how extinction occurs in nature. Instead, these simulations empirically demonstrate the degree to which an ecosystem function is concentrated in species. The more widely an ecosystem function is distributed across species, the more species must go extinct before the ecosystem function is lost. We observed a positive, monotonic relationship between the resilience to random species extinction events and CE (Figure 3A). When CE was ~0, resilience to species extinction was ~0.5, meaning ~50% of the community, on average, went extinct before an ecosystem function was removed from the community. This is equivalent to only one species performing the ecosystem function. When CE was ~1, resilience to species extinct was ~1. Again, this means 100% of the community, on average, had to go extinct to completely remove a function from an ecosystem. In general, the relationship between CE and trait resilience to extinction resembled a saturation curve—although we opted to use GAMs when calculating mean absolute error (previous section) as we made no assumption about the true relationship between CE and function resilience, other than it should be a positive, monotonic function.

CE Measures Resilience Differently than Other Functional Redundancy Metrics

Several methods for evaluating community-level functional redundancy exist in the literature, as detailed in Table 1. While Table 1 certainly does not include all published metrics, the highlighted metrics reflect the diversity in how traits, community properties (i.e., richness and evenness), as well as the desired metric utility (by design) vary among different functional redundancy metrics. Two of the metrics evaluate niche overlap, where niches are defined in multivariate trait space (Hester et al., 2019; Ricotta et al., 2016). In the case of Ricotta et al. (2016), functional redundancy weights species abundance, leading to a measure of the expected niche dissimilarity from the mean community niche. In contrast, Hester et al. (2019) evaluates

320 the median metabolic overlap which considers neither species abundance nor trait levels. One
321 metric evaluates niche occupancy, where niches are defined in multivariate trait space (Mouillot
322 et al., 2014). In this approach, functional redundancy is taken to be as the ratio of species
323 richness to plausible number niches (unique combinations of trait states), called functional
324 entities. The last metric is informal but has been utilized by various authors. In this approach,
325 functional redundancy is simply interpreted as the proportion of species in a community with a
326 trait (e.g., Trivedi, 2020; Tully, Wheat, et al., 2018; Wohl, Arora, & Gladstone, 2004).

327 To highlight the difference between our metric and the others described in Table 1, we
328 repeated the extinction simulations and compared all estimates of functional redundancy with
329 trait resilience to species extinction (Figure 3B-E). In these simulations, only one trait was
330 considered when calculating functional redundancy with all metrics. For the method presented in
331 Mouillot et al. (2014), we assumed there were two functional entities (trait absence and
332 presence). Generally, all functional redundancy metrics trended differently to resilience to
333 species extinction. This observation was not surprising, as most of these methods evaluate niche
334 overlap in some fashion. Among the metrics in Table 1, the fraction of the community
335 possessing a trait behaved most similar to CE (Figure 3B). We note that the fraction of the
336 community possessing a trait is a special case of CE, with diversity order q=0. Lastly, one
337 inherent limitation of CE is that it only evaluates single community-aggregated parameters at a
338 time. To address functional redundancy with respect to multiple traits perform dimensional
339 reduction (e.g., PCA) on multiple traits and use the resultant factors as trait levels in equation 2.
340 Factors in this case would be interpreted as different niches.

341 *Case Study: Redundancy of Nitrogen Cycle Marker Genes in Marine Microbiomes*

342 To demonstrate a use of CE, we measured functional redundancy in the marine nitrogen
343 cycle using genomic data from the *TARA Oceans* project (Sunagawa et al., 2015). Recently,
344 (Tully, Graham, et al., 2018) published 2,361 draft metagenome-assembled genomes (MAGs)
345 which represent genomic content from seven oceanographic provinces, three different depths,
346 and four filter size fractions. Relative abundance data used here were reported in (Tully, Graham,
347 et al., 2018) and were determined by the proportion of sequence reads mapping to assembled
348 MAGs. For simplicity, we limited our analysis to only MAGs identified in the bacterial size
349 fraction (<3 μm), which corresponds to free-living microbes. The nitrogen cycle was simplified
350 to eight pathways (Pajares & Ramos, 2019) and included N_2 fixation, nitrification, NO_3^-
351 reduction, dissimilatory NO_3^- reduction, assimilatory NO_3^-/NO_2^- reduction, NH_4^+ assimilation, N
352 remineralization, and denitrification. A general summary of the marker genes (downloaded from
353 Swiss-Prot; see Methods) used for quantifying nitrogen cycle pathways in MAGs is summarized
354 in Table 2.

355 Overall, the functional redundancy for individual genomic traits was consistent with our
356 conceptual understanding of primary nitrogen-transforming pathways in the ocean (Figure 4). A
357 clear example was the low functional redundancy in the nitrification pathway. We expected this
358 observation as nitrification metabolisms in oceans are primarily restricted to *Crenarchaeota*
359 (Wuchter et al., 2006). Similarly, the N_2 fixation pathway, which is predominantly facilitated by
360 a small group of *Cyanobacteria*, exhibits low functional redundancy. Conversely, the uptake of
361 inorganic N via NH_4^+ and NO_3^-/NO_2^- assimilation are known to be broadly expressed traits
362 (Pajares & Ramos, 2019), and accordingly exhibit high CE. Last, the low functional redundancy
363 in dissimilatory NO_3^- reduction and NO_3^- reduction were not surprising considering these

364 pathways occur in anaerobic and low-oxygen environments, respectively (Pajares & Ramos,
365 2019).

366 Some variance existed across oceanographic provinces and depth for intra-pathway
367 functional redundancy. However, inter-pathway variance in functional redundancy was generally
368 larger than intra-pathway variance (Figure 4). This suggests that the marine nitrogen cycle
369 marker genes considered here are consistently distributed among communities for MAGs
370 retrieved in Earth's ocean at middle and lower latitudes. This was consistent with the lack of
371 relationship between the metagenome companion metadata (e.g., O_2, inorganic N, PO_4^{3-}, latitude,
372 longitude, etc.) and functional redundancy. The low intra-pathway variance in functional
373 redundancy was not particularly surprising, as saltwater biomes exhibit the lowest alpha-
374 diversity and small beta-diversity among free-living microbiomes (Sunagawa et al., 2015;
375 Thompson et al., 2017). Minor shifts in community composition among lower taxonomic ranks
376 (e.g., genus or species) will not drastically alter redundancy in genomic traits as changes in
377 genomic composition is buffered by traits conserved at higher taxonomic ranks (e.g., phylum
378 level) (Royalty & Steen, 2019). This observation was in agreement with an earlier analysis
379 evaluating cluster of orthologous groups (COGs) composition across *TARA Oceans*
380 metagenomes (Sunagawa et al., 2015).

381 A more recent *TARA Oceans* analysis expanded sampling coverage to include more
382 metagenomes at polar latitudes. Community composition was found to be significantly different
383 at polar latitudes relative to lower and middle latitudes (Salazar et al., 2019). In this analysis,
384 eggNOG functional richness was compared to OTU richness across all metagenomes, leading to
385 the conclusion that metagenomes had low functional redundancy—the opposite conclusion from
386 Sunagawa et al. (2015). The discrepancy between these two analyses probably reflects

387 differences in how genes were annotated (COGs versus eggNOG). eggNOG orthologous groups
388 are defined by sequence homology and phylogeny, so gene clusters from different clades can
389 have the same function but different eggNOG annotation (Huerta-Cepas et al., 2019). Thus,
390 changes in biodiversity should give rise to a corresponding change in functional diversity if using
391 eggNOG annotations—again, annotations should correlate strongly with taxonomy. More
392 broadly speaking, this emphasizes the importance of properly contextualizing genomic trait
393 functional redundancy to account for annotation scheme.

394 In our analysis of *TARA Oceans* metagenomes, we used genomic traits as proxies for
395 nitrogen metabolisms. As a point of emphasis, measurable properties (e.g., nitrogen uptake)
396 provide direct measures of ecosystem function that are more easily integrated into predictive
397 models. In an ideal scenario, the latter would be discernible in all ecological studies, for all
398 ecosystem functions. However in the case of microbes, a disproportionate number of lineages
399 that lack cultured representatives highlights the continued importance of nucleotide sequencing
400 in microbial ecology (Lloyd et al., 2018). Consequently, the microbiological community relies
401 on 'omics technology (genes, transcripts, or proteins) to characterize community function.
402 Conceivably, other 'omics techniques such as metatranscriptomics and metaproteomics can be
403 evaluated with CE as well. In these approaches, the observed metatranscript and metapeptide
404 sequence abundances implicitly incorporate relative abundances of species, and do not require
405 scaling lineage abundances as is done in equation 2. Furthermore, technological advances with
406 single-cell nanoSIMS may provide a method for quantifying lineage-specific metabolic rates for
407 *in situ* inoculation studies (Musat, Foster, Vagner, Adam, & Kuypers, 2012). Coupling these
408 metabolic rates with abundance data, we believe CE could measure functional redundancy of
409 phenotypes.

Data Availability

In order to facilitate the use of CE functional redundancy measurements, we created an R package to calculate CE functional redundancy and run extinction simulations. The package is available at https://github.com/taylorroyalty/functional_redundancy_script.

Acknowledgments

We gratefully acknowledge the comments of three anonymous reviewers, whose constructive comments substantially improved the manuscript. Funding for this project was provided by a C-DEBI graduate fellowship to TMR and by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomic Science Program under Award Number DE-SC0020369 for TMR and ADS. This is C-DEBI contribution number [to be assigned upon manuscript acceptance].

Author Contribution

TMR designed the research, performed research, analyzed data, developed the R package, and wrote the paper. ADS designed the research, developed the R package, and wrote the paper.

References

Anantharaman, K., Brown, C. T., Hug, L. A., Sharon, I., Castelle, C. J., Probst, A. J., … Banfield, J. F. (2016). Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nature Communications*, *7*, 13219. doi: 10.1038/ncomms13219

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. doi: 10.1093/nar/28.1.45

Baker, B. J., Lazar, C. S., Teske, A. P., & Dick, G. J. (2015). Genomic resolution of linkages in carbon, nitrogen, and sulfur cycling among widespread estuary sediment bacteria. *Microbiome*, *3*(1), 14. doi: 10.1186/s40168-015-0077-6

Chao, A., & Ricotta, C. (2019). *Quantifying evenness and linking it to diversity , beta diversity , and similarity*. *100*(12), 1–15. doi: 10.1002/ecy.2852

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2019). GenBank. *Nucleic Acids Research*, *47*(D1), D94–D99. doi: 10.1093/nar/gky989

Falkowski, P. G., Fenchel, T., & Delong, E. F. (2008). The Microbial Engines That Drive Earth's Biogeochemical Cycles. *Science*, *320*(5879), 1034–1039. doi: 10.1126/science.1153213

Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(D1), 136–143. doi: 10.1093/nar/gkr1178

Fierer, N., Barberán, A., & Laughlin, D. C. (2014). Seeing the forest for the genes: Using metagenomics to infer the aggregated traits of microbial communities. *Frontiers in*

Microbiology, *5*(NOV), 1–6. doi: 10.3389/fmicb.2014.00614

Galand, P. E., Casamayor, E. O., Kirchman, D. L., & Lovejoy, C. (2009). Ecology of the rare microbial biosphere of the Arctic Ocean. *Proceedings of the National Academy of Sciences*, *106*(52), 22427–22432. doi: 10.1073/pnas.0908284106

Galand, P. E., Pereira, O., Hochart, C., Auguet, J. C., & Debroas, D. (2018). A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME Journal*, *12*(10), 2470–2478. doi: 10.1038/s41396-018-0158-1

Hester, E. R., Jetten, M. S. M., Welte, C. U., & Lücker, S. (2019). Metabolic overlap in environmentally diverse microbial communities. *Frontiers in Genetics*, *10*(OCT), 1–10. doi: 10.3389/fgene.2019.00989

Hill, M. O. (1973). Diversity and Evenness : A Unifying Notation and Its Consequences. *Ecology*, *54*(2), 427–432.

Huerta-Cepas, J., Szklarczyk, D., Heller, D., Hernández-Plaza, A., Forslund, S. K., Cook, H., … Bork, P. (2019). EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Research*, *47*(D1), D309–D314. doi: 10.1093/nar/gky1085

Jost, L. (2006). *Entropy and diversity*. *2*.

Krause, S., Le Roux, X., Niklaus, P. A., van Bodegom, P. M., Lennon T., J. T., Bertilsson, S., … Bodelier, P. L. E. (2014). Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Frontiers in Microbiology*, *5*(MAY), 1–10. doi: 10.3389/fmicb.2014.00251

Lloyd, K. G., Steen, A. D., Ladau, J., Yin, J., & Crosby, L. (2018). Phylogenetically Novel Uncultured Microbial Cells Dominate Earth Microbiomes. *MSystems*, *3*(5), 1–12. doi: 10.1128/mSystems.00055-18

Louca, S., Martiny, A. C., Polz, M. F., Doolittle, W. F., Nelson, M. B., Hahn, A. S., … Hallam, S. J. (2018). Function and functional redundancy in microbial communities. *Nature Ecology and Evolution*, *2*, 936–943.

Magurran, A. E. (1988). *Ecological Diversity and Its Measurement* (1st ed.). doi: 10.1007/978-94-015-7358-0

Malik, A. A., Martiny, J. B. H., Brodie, E. L., Martiny, A. C., Treseder, K. K., & Allison, S. D. (2019). Defining trait-based microbial strategies with consequences for soil carbon cycling under climate change. *ISME Journal*, 445866. doi: 10.1101/445866

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., … White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, *10*(10), 995–1015. doi: 10.1111/j.1461-0248.2007.01094.x

McPherson, J. M., Yeager, L. A., & Baum, J. K. (2018). A simulation tool to scrutinise the behaviour of functional diversity metrics. *Methods in Ecology and Evolution*, *9*(1), 200–206. doi: 10.1111/2041-210X.12855

Micheli, F., & Halpern, B. S. (2005). Low functional redundancy in coastal marine assemblages. *Ecology Letters*, *8*(4), 391–400. doi: 10.1111/j.1461-0248.2005.00731.x

Miki, T., Yokokawa, T., & Matsui, K. (2013). Biodiversity and multifunctionality in a microbial

community: A novel theoretical approach to quantify functional redundancy. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1776). doi: 10.1098/rspb.2013.2498

Mouillot, D., Villéger, S., Parravicini, V., Kulbicki, M., Arias-González, J. E., Bender, M., … Bellwood, D. R. (2014). Functional over-redundancy and high functional vulnerability in global fish faunas on tropical reefs. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(38), 13757–13762. doi: 10.1073/pnas.1317625111

Musat, N., Foster, R., Vagner, T., Adam, B., & Kuypers, M. M. M. (2012). Detecting metabolic activities in single cells, with emphasis on nanoSIMS. *FEMS Microbiology Reviews*, *36*(2), 486–511. doi: 10.1111/j.1574-6976.2011.00303.x

Naeem, S. (1998). *Species Redundancy and Ecosystem Reliability*. *12*(1), 39–45.

Orsi, W. D., Wilken, S., del Campo, J., Heger, T., James, E., Richards, T. A., … Santoro, A. E. (2018). Identifying protist consumers of photosynthetic picoeukaryotes in the surface ocean using stable isotope probing. *Environmental Microbiology*, *20*(2), 815–827. doi: 10.1111/1462-2920.14018

Pajares, S., & Ramos, R. (2019). Processes and Microorganisms Involved in the Marine Nitrogen Cycle: Knowledge and Gaps. *Frontiers in Marine Science*, *6*(November). doi: 10.3389/fmars.2019.00739

Philip, D. (2003). Computer program review VEGAN , a package of R functions for community ecology. *Journal of Vegetation Science*, *14*, 927–930. doi: 10.1111/j.1654-1103.2003.tb02228.x

Ricotta, C., de Bello, F., Moretti, M., Caccianiga, M., Cerabolini, B. E. L., & Pavoine, S. (2016).

Measuring the functional redundancy of biological communities: a quantitative guide. *Methods in Ecology and Evolution*, *7*(11), 1386–1395. doi: 10.1111/2041-210X.12604

Rosenfeld, J. S., & Mall, M. (2002). Functional redundancy in ecology and conservation. *OIKOS*, *1*.

Royalty, T. M., & Steen, A. D. (2019). Quantitatively Partitioning Microbial Genomic Traits among Taxonomic Ranks across the Microbial Tree of Life. *MSphere*, *4*(4), e00446-19. doi: 10.1128/mSphere.00446-19

Salazar, G., Paoli, L., Alberti, A., Huerta-Cepas, J., Ruscheweyh, H. J., Cuenca, M., … Wincker, P. (2019). Gene Expression Changes and Community Turnover Differentially Shape the Global Ocean Metatranscriptome. *Cell*, *179*(5), 1068-1083.e21. doi: 10.1016/j.cell.2019.10.014

Staley, J., & Konopka, A. (1985). Measurement of In Situ Activities of Nonphotosynthetic Microorganisms in Aquatic and Terrestrial Habitats. *Annual Review of Microbiology*, *39*(1), 321–346. doi: 10.1146/annurev.micro.39.1.321

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1–10. doi: 10.1126/science.1261359

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., … Amann, R. (2012). Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science*, *336*(May), 608–611. doi: 10.1126/science.1218344

Thompson, L. R., Sanders, J. G., McDonald, D., Amir, A., Ladau, J., Locey, K. J., … Zhao, H. (2017). A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*, *551*(7681), 457–463. doi: 10.1038/nature24621

Tilman, D. (2001). Functional Diversity. *Encyclopedia of Biodiversity: Second Edition*, *3*, 587–596. doi: 10.1016/B978-0-12-384719-5.00061-7

Trivedi, C. B., Stamps, B. W., Lau, G. E., Grasby, S. E., Templeton, A. S., & Spear, J. R. (2020). Microbial Metabolic Redundancy Is a Key Mechanism in a Sulfur-Rich Glacial Ecosystem. *MSystems*, *5*(4), 1–16. doi: https://doi.org/10.1128/mSystems .00504-20.

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*, 1–8. doi: 10.1038/sdata.2017.203

Tully, B. J., Wheat, C. G., Glazer, B. T., & Huber, J. A. (2018). A dynamic microbial community with high functional redundancy inhabits the cold, oxic subseafloor aquifer. *ISME Journal*, *12*(1), 1–16. doi: 10.1038/ismej.2017.187

Violle, C., Navas, M., Vile, D., Kazakou, E., Fortunel, C., Hummel, I., … Fortunel, C. (2007). Let the concept of trait be functional! *Oikos*, *116*(5), 882–892. doi: https://doi.org/10.1111/j.0030-1299.2007.15559.x

Walker, B. H. (1992). Biodiversity and Ecological Redundancy. *Conservation Biology*, *6*(1), 18–23. doi: 10.1046/j.1523-1739.1992.610018.x

Wohl, D. L., Arora, S., & Gladstone, J. R. (2004). Functional Redundancy Supports Biodiversity and Ecosystem Function in a Closed and Constant Environment. *Ecology*, *85*(6), 1534–

1540.

Wood, S. (2017). *mgcv: Mixed GAM Computation Vehicle with GCV/AIC/REML Smoothness Estimation*.

Wuchter, C., Abbas, B., Coolen, M. J. L., Herfort, L., Van Bleijswijk, J., Timmers, P., … Damsté, J. S. S. (2006). Archaeal nitrification in the ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(33), 12317–12322. doi: 10.1073/pnas.0600756103

Data Accessibility

We provided a function for calculating functional redundancy with CE and an example script at https://github.com/taylorroyalty/functional_redundancy_script.

Author Contribution

TMR designed the research, performed research, analyzed data, and wrote the paper. ADS designed the research, performed research, and wrote the paper.

Tables

Table 1. A summary of metrics of community-level functional redundancy.

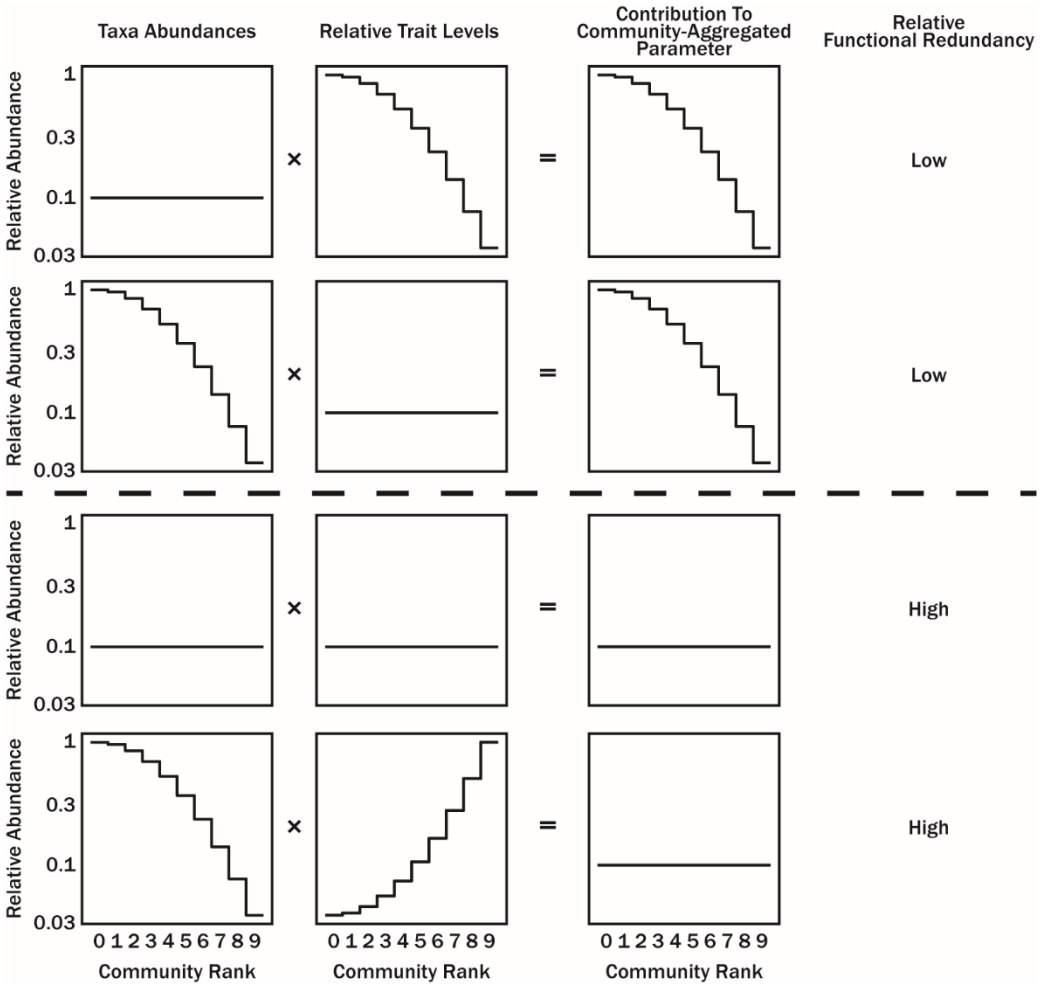
Authors	Interpretation	Equation	Notes
Here	Effective proportion of species with a trait	$R = \frac{D_q - 1}{S - 1}$	• Accounts for species abundances • Continuous trait states • Individual traits • R spans 0 to 1
(Ricotta et al., 2016)	Expected trait dissimilarity among species in a community	$R = 1 - \frac{\sum_i p_i \sum_j p_j \delta_{ji}}{D_2}$; $\delta_{ji} = trait\ distance$; $D_2 =$ Taxonomic diversity order ($q = 2$)	• Accounts for species abundances • Continuous trait states • Individual or multivariate traits • Normalized by a maximum theoretical dissimilarity • R spans 0 to 1
(Mouillot et al., 2014)	Ratio of species to number of functional entities (states)	$R = \frac{S}{FE}$; FE = # of possible combinations of trait states for all traits under consideration	• Binary trait states • Individual or multivariate traits • R spans ~0 ($\lim_{FE \to \infty} \frac{S}{FE} = 0$) to S
Various (see text for examples)	Fraction of species with a trait	$R = \frac{v}{S}$	• Binary trait states • Individual traits • R spans 0 to 1
(Hester et al., 2019)	Median overlap of traits between species	$R = med(O)$; $O = \{o \mid o_{i,j} = \lvert g_i \cap g_j \rvert; i \neq j\}$; $g = taxa$; O = set containing # of overlaps among species	• Binary trait states • $\lvert g_i \cap g_j \rvert$represents the number of overlapping traits g between species i and species j • Individual and multivariate traits • R spans 0 to ∞

567 **Table 2. A summary of marker gene sequences used in the BLASTP analysis of *TARA***
568 ***Oceans* MAGs. All marker genes were downloaded from Swiss-Prot.**

Pathway	Gene Names	EC Number	# Marker Genes
N Remineralization	*gltB*	1.4.1.13	14
N Fixation	*nifH*	1.18.6.1	127
Nitrification	*amoA*	1.14.99.39	2
	hoa	1.7.2.6	1
NO_3^- Reduction	*napA*	1.9.6.1	136
	narG	1.7.5.1	13
Dissimilatory NO_3^- Reduction	*nrfA*	1.7.2.2	77
Assimilatory NO_3^-/NO_2^- Reduction	*narB*	1.7.5.1	3
	nasA	1.7.7.2	1
	nasB	1.7.1.4	3
	nirB	1.7.1.15	5
NH_4^+ Assimilation	*gdhA*	1.4.1.3	14
	glnA	6.3.1.2	173
Denitrification	*nirK*	1.7.2.1	7
	nirS	1.7.2.1; 1.7.99.1	4
	norB	1.7.2.5	2
	nosZ	1.7.2.4	11

569

570 **Figures**



571

572 **Figure 1.**

573 Community taxon abundance and relative contribution level are scalable and provide a measure
574 for the contribution that different species have towards a community-aggregated parameter. The
575 evenness of species-level contribution to the community-aggregated parameter is taken as the
576 functional redundancy.

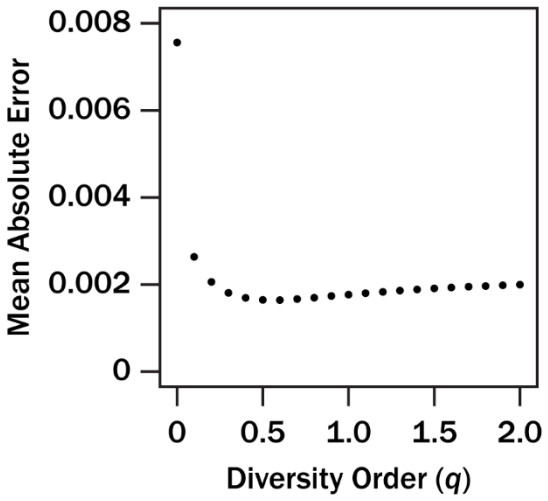


Figure 2.

The sensitivity of mean absolute error between functional redundancy and trait resilience to species extinction in response changing diversity order q in equation 4.

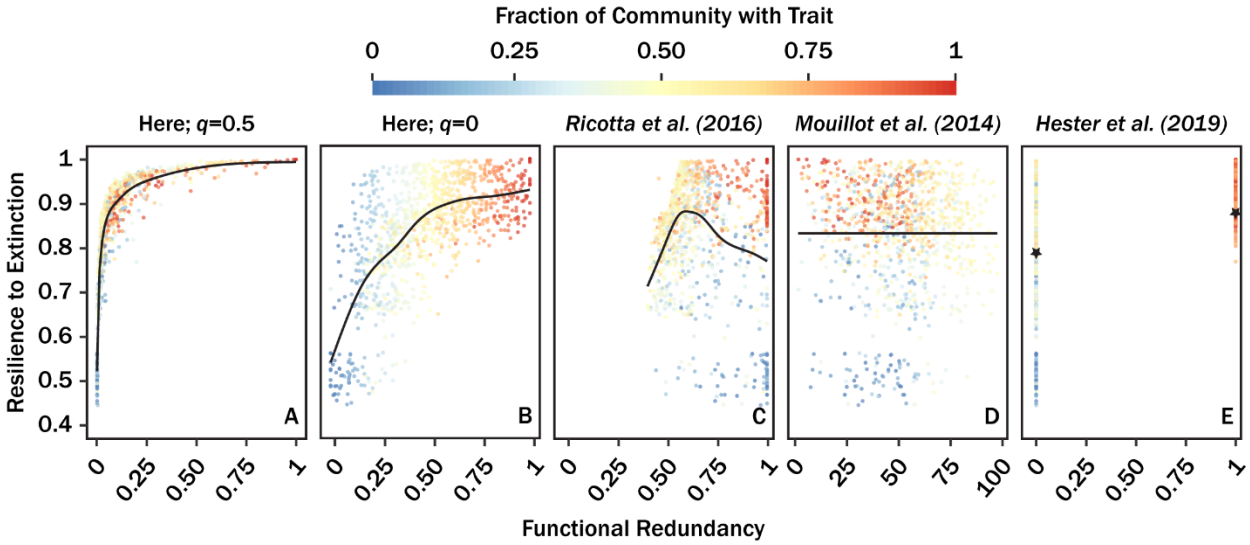


Figure 3.

A comparison of the relationship between selected different functional redundancy metrics and trait resilience (see methods). Functional redundancy calculated with equation 5 (q=0.5) (**A**), equation 5 (q=0) (**B**), methodology from Ricotta et al. (2016) (**C**), methodology from Mouillot et al. (2014) (**D**), and methodology from Hester et al. (2019). Black lines in panels **A-D** correspond to best fit GAMs. The stars in panel **E** correspond to the average trait resilience at R=0 and R=1. The color scale corresponds to the fraction of a community with a trait versus.

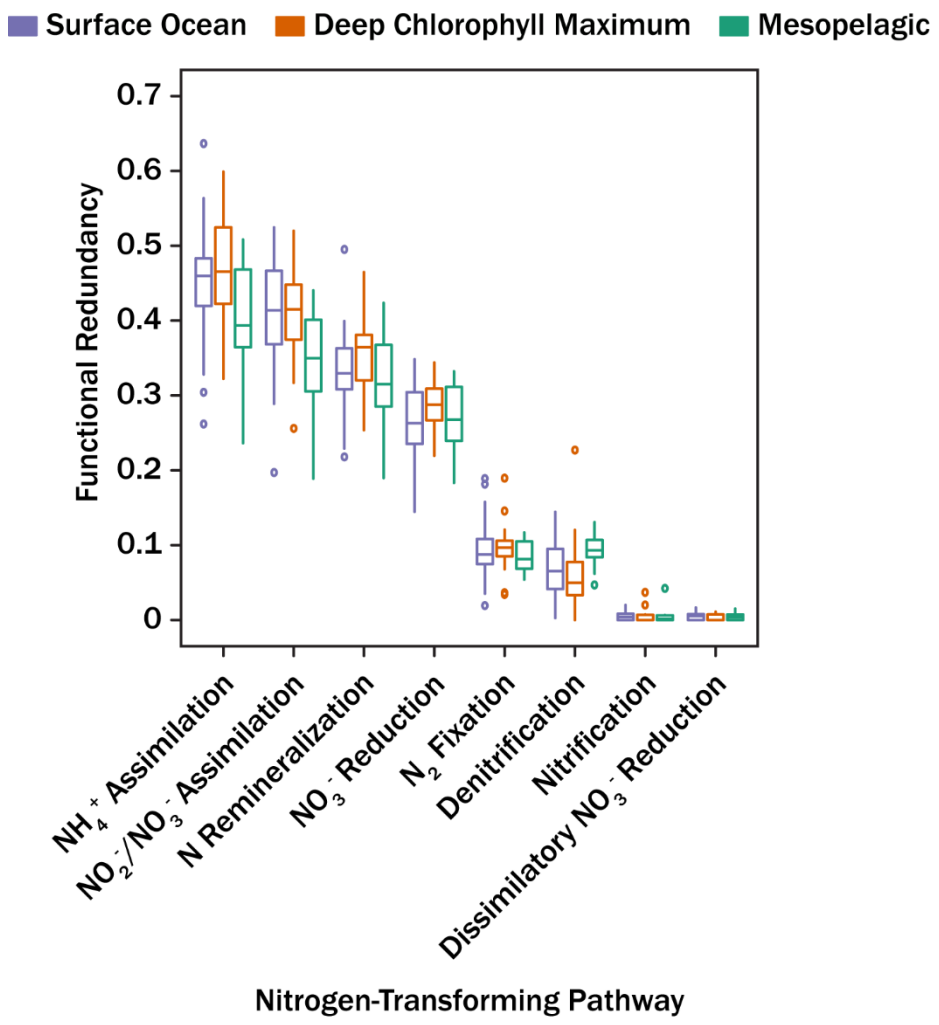


Figure 4.

Boxplots showing functional redundancy calculated for eight nitrogen-transforming pathways in *TARA Oceans* draft MAGs. Different colors correspond to sample depth. Whiskers correspond to 150% of the inter-quantile region.