Ten metagenome-assembled genomes show region-robust associations with dissolved nutrients across the global surface ocean: a compositional, false-discovery-controlled, and leave-one-region-out analysis of Tara Oceans

Analysis of Tara Oceans metagenome-assembled genomes (Figshare `10.6084/m9.figshare.4902938.v3`), PANGAEA nutrient measurements (`10.1594/PANGAEA.875575`) and the Tara sample registry (`10.1594/PANGAEA.875582`).

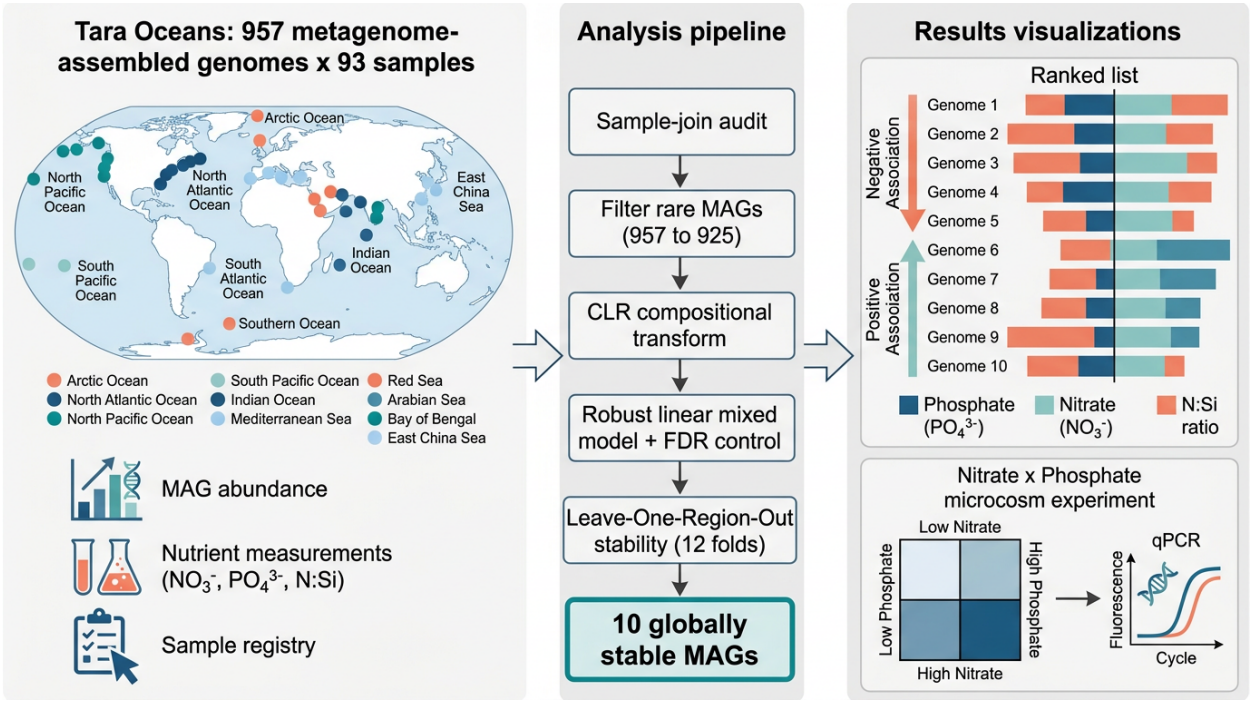

Figure 1: Graphical abstract. Three publicly archived Tara Oceans resources—a matrix of mean coverage for 957 non-redundant metagenome-assembled genomes (MAGs) across 93 surface-ocean metagenomes, dissolved-nutrient measurements, and the sample registry—are joined, audited, and analysed through a reproducible pipeline. Rare MAGs are removed, abundances are centred-log-ratio (CLR) transformed, and region-aware robust linear mixed models with Benjamini–Hochberg false-discovery-rate (FDR) control identify nutrient associations that are then required to hold under leave-one-region-out (LORO) cross-validation. Ten globally stable MAGs emerge, each tied to phosphate, nitrate+nitrite, or the N:Si ratio, and a 2×2 nitrate $\times$ phosphate microcosm with a per-MAG qPCR / metatranscriptomic readout is designed to test the inferred dependencies.

Abstract

Dissolved inorganic nutrients structure the biogeography of marine prokaryotes, but linking specific, genome-resolved populations to nutrient regimes across ocean basins is complicated by compositional sequencing data, uneven regional sampling, and the ease with which spatial co-variation can be mistaken for ecological dependence. Here we reanalyse publicly archived Tara Oceans surface-ocean data—mean coverage for 957 non-redundant metagenome-assembled genomes (MAGs) across 93 metagenomes, harmonised with PANGAEA nutrient measurements and the Tara sample registry—to identify MAGs whose distributions are robustly associated with nitrate+nitrite, phosphate, silicate, or nutrient stoichiometry. After a full sample-join audit (63 stations; 93 station×depth-layer combinations; 100% join resolution), we removed low-abundance MAGs (957→925), applied a centred-log-ratio transform, and fitted 5,550 region-aware linear mixed models (one per MAG × predictor) with temperature, salinity and depth as covariates and ocean region as a random intercept. Benjamini–Hochberg FDR control ($q < 0.05$) yielded 314 significant MAG–nutrient associations, dominated by phosphate (125) and nitrate+nitrite (124). Requiring each association to remain directionally and statistically significant across all twelve leave-one-region-out folds collapsed these to 168 doubly stable associations, from which we ranked the ten most robust. The ten MAGs span Proteobacteria, Verrucomicrobia, Bacteroidetes, *Candidatus* Marinimicrobia and Euryarchaeota; seven are negatively associated with a nutrient (consistent with oligotroph niches and high-affinity scavenging), and three track nitrate positively. Because the predictors are strongly collinear (nitrate, phosphate and log N:Si co-vary at $\rho \geq 0.8$), a single-nutrient coefficient cannot by itself isolate the causal nutrient from field data; and because none of the ten fall within the supplement's gene-level functional tables, metabolic interpretation is explicitly lineage-based and correlational. We enumerate physical-transport, niche-covariation and biotic-interaction explanations that stability alone cannot exclude. Finally, we specify a 2×2 nitrate × phosphate microcosm experiment (four treatments, $n = 4$, six timepoints over 96 h) with MAG-specific qPCR / metatranscriptomic readouts of *pstS*, *phoX*, *glnA*, *nasA* and *amtB* that converts each observational association into a falsifiable, directional hypothesis. The result is a ranked, auditable, and experimentally testable set of candidate nutrient-responsive ocean genomes.

Keywords: marine microbiome; metagenome-assembled genomes; Tara Oceans; nutrient limitation; compositional data analysis; false discovery rate; cross-validation; microcosm experiment; qPCR

1 Introduction

The surface ocean is a vast, dilute chemostat in which the availability of nitrogen, phosphorus, iron and silicon sets the ceiling on primary production and shapes the composition of microbial communities that mediate roughly half of global carbon fixation. Decades of biogeochemical work have established that low-latitude subtropical gyres are broadly nitrogen- or nitrogen–iron-co-limited, whereas high-macronutrient upwelling and polar regions are frequently iron-limited, with phosphorus limitation intensifying in stratified basins such as the subtropical North Atlantic [1–3]. The elemental stoichiometry of marine plankton, far from the fixed Redfield ratio, varies systematically with latitude and nutrient supply, reaching C:N:P ratios far above 106:16:1 in warm oligotrophic gyres [4, 5]. These macro-scale patterns imply that individual microbial populations should partition along nutrient gradients according to their uptake affinities, storage capacities and reliance on regenerated versus new nutrients—yet connecting a specific, genome-resolved population to a specific nutrient axis at global scale remains difficult.

The Tara Oceans expedition transformed this problem from one of inference-by-proxy into one of genome-resolved biogeography. The global ocean microbiome reference catalogue revealed that epipelagic community structure is governed primarily by temperature, with a core microbiome sharing a broad functional repertoire across basins [6, 7]. Genome-resolved reconstruction then

recovered nearly a thousand non-redundant prokaryotic MAGs from surface metagenomes and, unexpectedly, showed that non-cyanobacterial heterotrophic bacterial diazotrophs—Proteobacteria and Planctomycetes carrying *nifH*—are abundant and biogeographically structured in the sunlit ocean [8, 9]. Companion resources expanded the catalogue of ocean MAGs to thousands of draft genomes [10] and demonstrated that community turnover and gene expression respond to environmental gradients across the global metatranscriptome [11]. A recurring theme of this body of work is that function and taxonomy are partially decoupled: environmental conditions structure functional-group abundance more tightly than the taxonomic identity of the organisms performing those functions [12], and biotic interactions are themselves major, non-random determinants of community structure [13, 14].

Against this backdrop, three methodological hazards make naive "abundance-versus-nutrient" correlations untrustworthy. First, metagenomic abundance data are *compositional*: sequencing yields relative, not absolute, abundances, so an apparent increase in one taxon mechanically depresses the apparent abundance of others, generating spurious negative correlations and non-reproducible associations unless log-ratio transforms are used [15–18]. Second, Tara sampling is spatially clustered into ocean provinces; a single peculiar region can drive a "global" correlation, and unmodelled water-mass structure confounds any nutrient signal [19, 20]. Third, and most fundamentally, an association—however statistically clean—is not a mechanism: currents co-transport cells and nutrients, unmeasured co-limiting factors (iron, light, dissolved organic matter) covary with measured nutrients, and biotic partners such as diatoms or phytoplankton hosts can mediate a bacterium's distribution without any direct nutrient dependency [20, 21].

Here we confront all three hazards directly. We reanalyse the archived Tara Oceans MAG coverage matrix together with PANGAEA nutrient and registry tables, and we (i) audit every sample join before integrating datasets, (ii) filter rare MAGs and apply a centred-log-ratio transform to respect compositionality, (iii) fit region-aware robust models with false-discovery-rate control, and (iv) demand that every reported association survive leave-one-region-out cross-validation, a stability criterion in the spirit of resampling-based variable selection [22]. We connect the resulting robust hits to the functional annotations available in the resource—being explicit that none of our ten top MAGs carries gene-level annotation in the supplement, so metabolic interpretation is lineage-based and hypothesis-generating rather than confirmatory. We then translate the observational findings into a concrete, falsifiable 2×2 nitrate $\times$ phosphate microcosm experiment with MAG-tailored molecular readouts. Our aim is not to claim causation from field data, but to deliver a ranked, auditable, and directly testable shortlist of nutrient-responsive ocean genomes.

2 Materials and Methods

2.1 Data sources

Three publicly archived resources were used without modification of their primary contents. (1) MAG distributions were taken from the Tara Oceans non-redundant MAG resource on Figshare (`10.6084/m9.figshare.4902938.v3`); the mean-coverage matrix (supplementary `Table_S3`, sheet *Mean coverage non-redundant data*) provides coverage for 957 non-redundant MAGs across 93 surface-ocean metagenomes, with associated genome-quality and taxonomy metadata [8, 23, 24]. (2) Dissolved nutrient measurements (nitrate, nitrite, nitrate+nitrite, phosphate, silicate) were obtained from PANGAEA (`10.1594/PANGAEA.875575`), comprising 34,516 records across 202 stations. (3) Contextual environmental sensor data and sample provenance were obtained from the Tara registry

collection (10.1594/PANGAEA.875582; TARA_ENV_DEPTH_SENSORS.tab), comprising 34,730 records across 203 stations [25]. All downloads were checksummed and logged for provenance.

2.2 Sample-join audit

Because the three resources use different identifier schemes, joins were audited before any statistical integration. Each MAG metagenome label of the form <OS-REGION>_<STATION>_<DEPTH>m (e.g. ANE_004_05M) was parsed into an ocean region, a Tara station number (mapped to TARA_%03d), and a depth in metres. Depth was discretised to a layer using the rule "depth $\leq 10\,\mathrm{m} \rightarrow$ surface (SRF); otherwise deep chlorophyll maximum (DCM)," matching the environmental-feature vocabulary used by PANGAEA. Because the MAG metagenomes derive from the prokaryote-enriched size fraction while PANGAEA environmental values are repeated across size-fraction rows, the contextual tables were de-duplicated to unique (station, layer) combinations before joining. The audit confirmed that the 93 metagenomes span 63 stations and 93 unique (station, layer) combinations (61 SRF, 32 DCM) across twelve ocean regions, and that *all* 63 stations and *all* 93 (station, layer) combinations resolved to both the nutrient and registry tables (100% join resolution; no missing stations). Residual limitations—the metres-to-layer approximation and the absence of a shared sample accession—were recorded in the audit report rather than silently ignored.

2.3 Preprocessing and compositional transform

Environmental covariates were harmonised to the 93 aligned samples; 92 carried at least one contextual measurement. Missing nutrient and hydrographic values were median-imputed per variable (missing before imputation: nitrite 8, phosphate 10, nitrate+nitrite 13, silicate 8, temperature 1, salinity 1). Three nutrient-stoichiometry predictors (N:P, N:Si, P:Si) were computed using nitrate+nitrite as the dissolved inorganic nitrogen proxy; raw zeros were replaced by one-half the smallest non-zero value of that nutrient (a detection-limit proxy rather than an arbitrary mathematical epsilon), and ratios were natural-log transformed when their absolute skew exceeded 1.0, which reduced skew from $6.96 \rightarrow -0.20$ (N:P), $4.53 \rightarrow -1.24$ (N:Si) and $5.05 \rightarrow -1.63$ (P:Si).

MAGs were filtered to remove rare populations that cannot support a stable association test: a MAG was retained only if it was present (non-zero coverage) in $\geq 10\%$ of samples ($\geq$10 metagenomes) *and* had a mean relative abundance $\geq 10^{-4}$. This removed 32 of 957 MAGs (all failing the abundance threshold; none failed prevalence), leaving 925 MAGs. Retained coverage values were converted to proportions and centred-log-ratio (CLR) transformed [15, 16]: zeros were replaced by a pseudocount equal to one-half the global minimum non-zero proportion ($\approx 4.9 \times 10^{-10}$) and the composition re-closed prior to the CLR, whose per-sample geometric-mean centring was verified (maximum absolute row mean $< 3 \times 10^{-14}$). The CLR was chosen in preference to library-size rarefaction or naive relative abundances because it treats the data as intrinsically compositional and is less prone to the spurious-correlation and false-positive artefacts that afflict count-based differential-abundance tests [17, 26].

2.4 Nutrient-association modelling with FDR control

For each retained MAG and each of six predictors (nitrate+nitrite, phosphate, silicate, $\log$ N:P, $\log$ N:Si, $\log$ P:Si) we fitted a linear mixed-effects model of the MAG's CLR abundance on the z-scored predictor, adjusting for temperature, salinity and sampling depth as fixed covariates and including ocean region (twelve levels) as a random intercept to absorb broad biogeographic

and water-mass structure [27, 28]. This yielded $925 \times 6 = 5{,}550$ models, all of which were fitted (0 failures) with an overall convergence rate of 97.7%. Per predictor, p-values for the predictor coefficient were adjusted for multiple testing using the Benjamini–Hochberg false-discovery-rate procedure, and associations with $q < 0.05$ were called significant [29, 30]. As a sensitivity check against high-leverage points, the leading associations per predictor were refitted with a Huber robust (M-estimator) regression using region as fixed-effect dummies [31], and the fraction of estimates stable within 25% of the mixed-model coefficient was recorded. Analyses used a fixed random seed (42) for full reproducibility.

2.5 Leave-one-region-out (LORO) stability

Statistical significance in the full model does not guarantee that an association generalises across ocean basins. We therefore subjected every significant association to leave-one-region-out cross-validation: each of the twelve regions was held out in turn and the mixed model refitted on the remaining eleven, generating twelve out-of-province coefficient and q-value estimates per association (66,600 planned refits; 66,588 completed; 97.1% fold convergence). Two orthogonal stability criteria were imposed. *Directional stability* required the sign of the coefficient to match the full-model sign in all twelve folds. *Significance stability* required $q < 0.05$ in all twelve folds. Associations meeting both criteria ("doubly stable") were ranked by the absolute value of their median across-fold coefficient, and the top ten were carried forward. This resampling-based screen follows the logic of stability selection, in which reproducibility across perturbations of the data guards against data-set-specific artefacts [22].

2.6 Functional annotation and interpretation

Taxonomy and genome-quality statistics (anvi'o and CheckM completion and redundancy, genome size, GC content, N50, contig count) for the ten MAGs were drawn from the resource's non-redundant metadata sheet [23, 24]. We then queried the supplement's gene-level functional tables (KEGG-module / RAST functions and *nifH* diazotroph tables), which in this resource are populated only for the heterotrophic-bacterial-diazotroph reference set and a small number of specific MAGs. None of our ten top MAGs appears in these gene-level tables; consequently, nutrient-metabolism capacity was inferred from lineage physiology (e.g. high-affinity phosphate transport *pstSCAB* and alkaline phosphatases *phoX*/*phoA* in oligotrophs; ammonium transport *amtB* and glutamine synthetase *glnA* in regenerated-N specialists; assimilatory nitrate/nitrite reduction *nasA*/*nirA*/*narB* in nitrate exploiters) and is reported as hypothesis-generating rather than confirmatory [21, 32–36].

2.7 Microcosm experimental design

To convert the observational associations into causal tests, we programmatically designed a 2×2 factorial nitrate $\times$ phosphate enrichment microcosm. Each of the ten MAGs was assigned to one of four functional response groups directly from its (predictor, association-sign) pair, and each group was given a directional expected response across the four treatments and a primary/secondary target-gene panel. Treatment, sampling, molecular-readout, replication and analysis specifications follow established nutrient-bioassay and quantitative-PCR reporting standards [1, 6, 8, 37] and are detailed in Results (Section 3.6).

3 Results

3.1 Sample joins resolve completely and rare MAGs are removed

The sample-join audit established a clean basis for integration: the 93 Tara surface metagenomes map to 63 stations and 93 unique (station, layer) combinations, distributed across twelve ocean regions (ranging from 3 metagenomes in the Southern Ocean to 16 in the South Pacific East), and every station and every (station, layer) combination resolved to both the PANGAEA nutrient and registry tables. Prevalence and abundance filtering (Fig. 2) retained 925 of 957 MAGs; the 32 removed genomes all failed the mean-relative-abundance threshold of 10^{-4} while none failed the 10% prevalence threshold, indicating that the MAG catalogue is dominated by prevalent-but-sometimes-rare populations rather than by patchily distributed ones. Environmental predictors were moderately to strongly collinear (Fig. 3): nitrate+nitrite correlated with phosphate (Spearman $\rho = 0.80$) and with $\log$ N:Si ($\rho = 0.91$), and phosphate with $\log$ P:Si ($\rho = 0.75$), underscoring why single-nutrient coefficients must be interpreted with explicit caveats about co-limitation and shared water-mass structure.

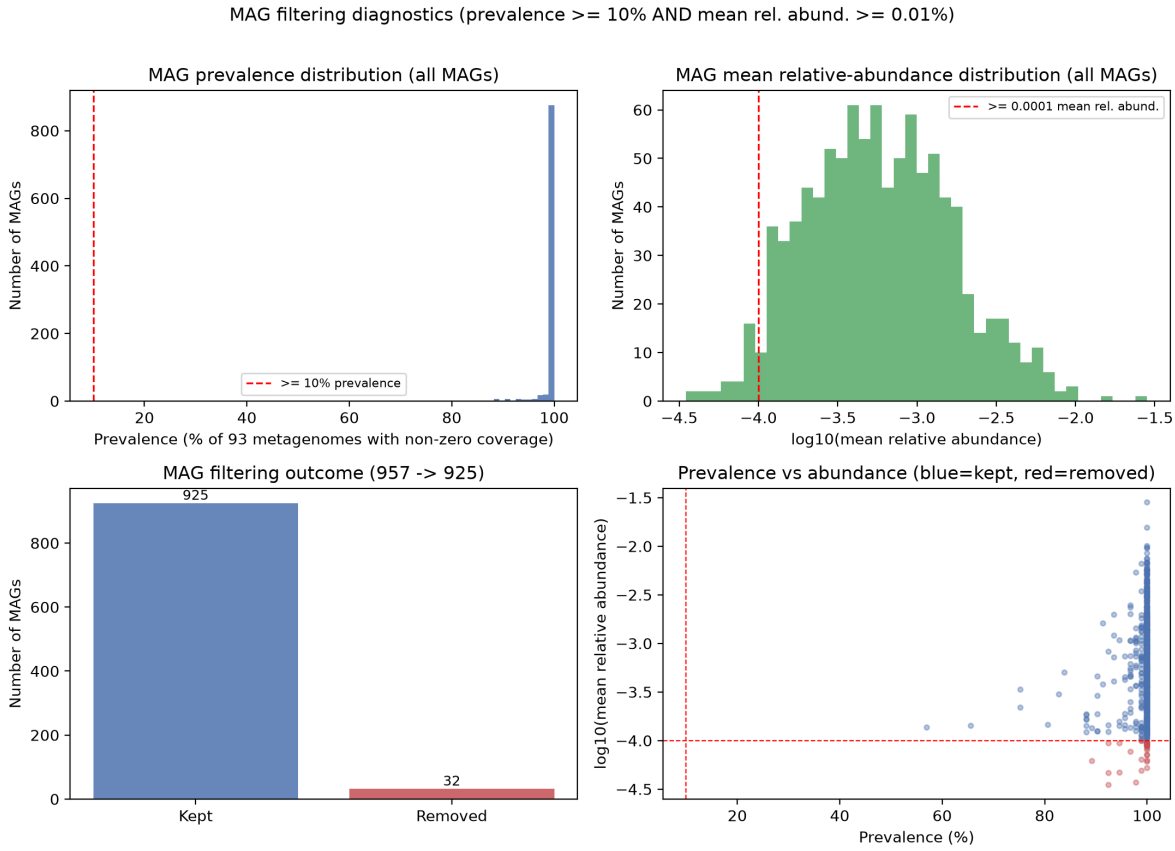

Figure 2: Rare-MAG filtering diagnostics. Prevalence (percentage of 93 metagenomes with non-zero coverage) and mean relative abundance distributions for all 957 MAGs, with the retention thresholds ($\geq 10\%$ prevalence *and* $\geq 10^{-4}$ mean relative abundance) marked in red. Bottom left: filtering outcome (925 kept, 32 removed). Bottom right: prevalence-versus-abundance scatter, with removed MAGs (red) clustered at low mean abundance despite frequently high prevalence.

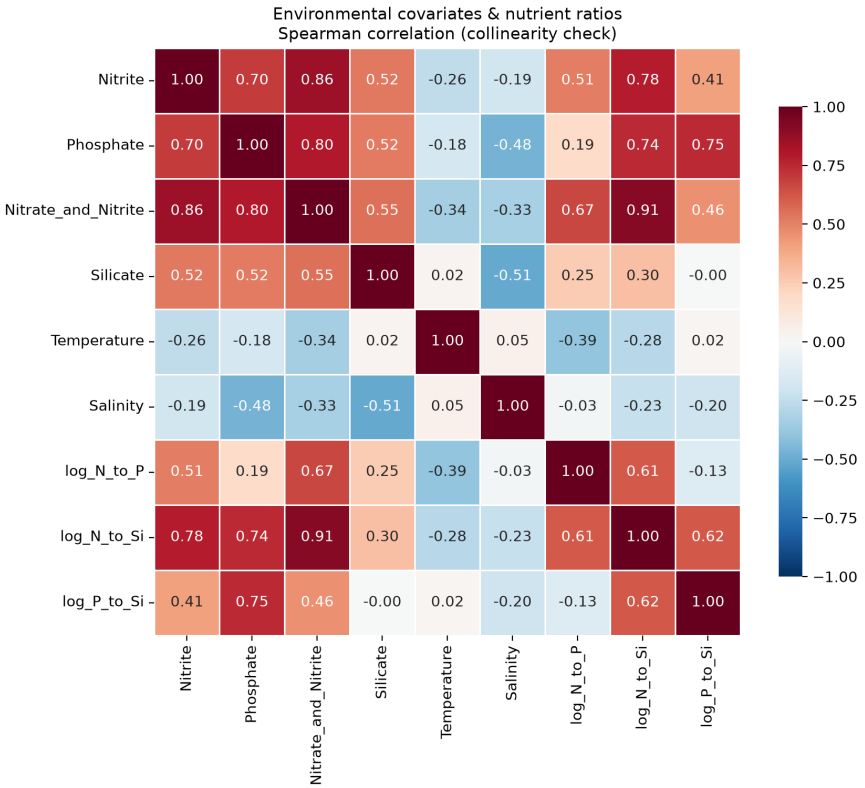

Figure 3: Collinearity among environmental predictors and nutrient ratios. Spearman correlation matrix across the 93 aligned samples. Strong positive coupling among nitrate+nitrite, phosphate and log N:Si ($\rho \geq 0.8$) means that a MAG associated with one nutrient will often appear associated with correlated nutrients; the modelling and stability framework, together with the microcosm design, is intended to mitigate but cannot fully resolve this ambiguity from field data alone.

3.2 The genome-wide association landscape is dominated by phosphate and nitrate

Across the 5,550 region-aware mixed models, 314 MAG–nutrient associations passed FDR control at $q < 0.05$ (Fig. 4, 5). The signal was concentrated on the two macronutrients most directly tied to microbial growth: phosphate (125 significant MAGs) and nitrate+nitrite (124), followed by log N:Si (39), silicate (19) and log P:Si (7). No MAG was significantly associated with log N:P after FDR control, consistent with the near-symmetric, weakly skewed distribution of that ratio and with N and P covarying so tightly that their ratio carries little independent structure. For the two dominant nutrients the significant associations were split between positive and negative signs (nitrate: 57 positive / 67 negative; phosphate: 45 positive / 80 negative), whereas the stoichiometric-ratio associations were predominantly negative. Model convergence was uniformly high across predictors (≥ 0.97; Fig. 5, right), and the Huber robust refits reproduced the mixed-model coefficients within 25% for the majority of the strongest hits (e.g. 15/20 for phosphate, 13/20 for nitrate), indicating that the leading associations are not artefacts of a few high-leverage samples.

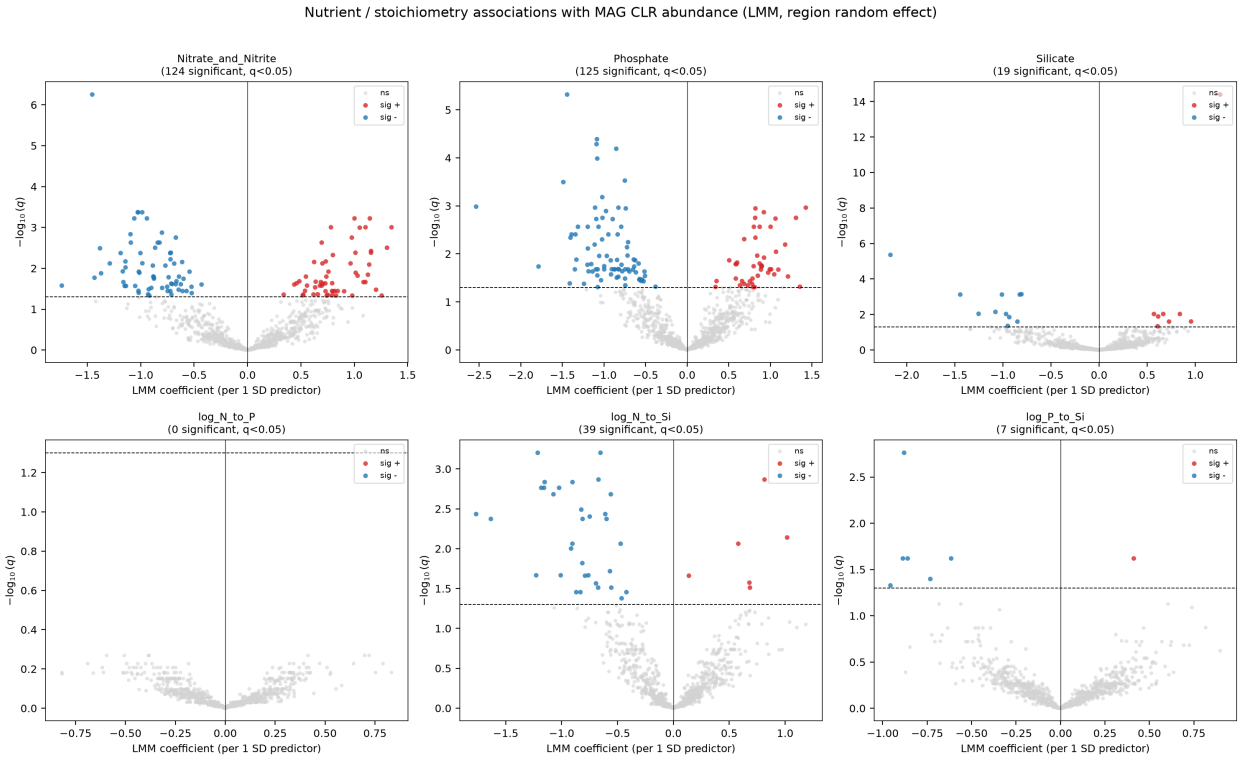

Figure 4: Volcano plots of MAG–nutrient associations. Each panel shows, for one predictor, the standardized mixed-model coefficient (per 1 SD of predictor) against $-\log_{10}(q)$ for all 925 MAGs. Red and blue mark FDR-significant positive and negative associations, respectively ($q < 0.05$); grey marks non-significant MAGs. Phosphate and nitrate+nitrite carry the largest number of significant hits; the ratio log N:P yields none.

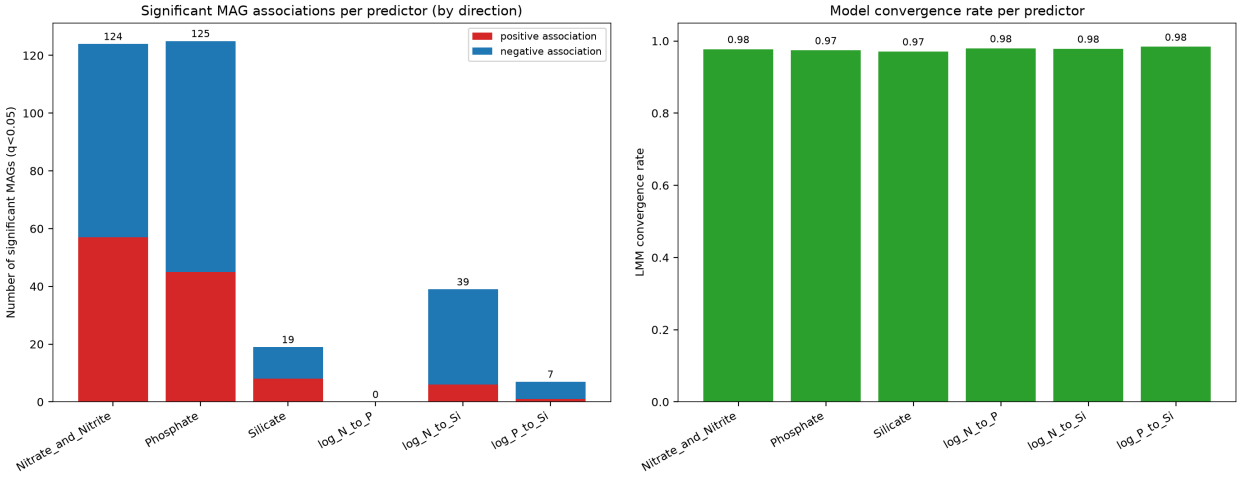

Figure 5: Significant associations per predictor and model convergence. Left: number of FDR-significant MAG associations ($q < 0.05$) per predictor, coloured by direction (positive/negative). Right: linear-mixed-model convergence rate per predictor, uniformly ≥ 0.97.

3.3 Ten MAGs are stable under leave-one-region-out cross-validation

Requiring associations to generalise across ocean basins sharply reduced the candidate set. Of the 314 FDR-significant associations, 297 were directionally stable (correct sign in all twelve LORO folds), 179 were significance-stable ($q < 0.05$ in all twelve folds), and 168 satisfied both criteria. The ten most robust doubly stable associations, ranked by absolute median across-fold coefficient, are reported in Table 1 and visualised in Figs. 6–8. Every one of the ten is significant and correctly signed in all twelve folds (12/12), with tight across-fold coefficient distributions (Fig. 7), and none depends on any single region for either its significance or its direction (Fig. 8).

The strongest and most stable association is a Gammaproteobacterial MAG (`TARA_PSE_MAG_00096`) negatively associated with phosphate (standardized coefficient -2.54; $q = 1.0 \times 10^{-3}$), followed by an alphaproteobacterial *Rhizobiales* MAG (`TARA_MED_MAG_00059`) negatively associated with nitrate+nitrite (-1.45; $q = 5.6 \times 10^{-7}$, the lowest full-model q-value in the top ten). The ranked set spans five phyla—Proteobacteria (six MAGs), Verrucomicrobia, Bacteroidetes, *Candidatus* Marinimicrobia (SAR406), and Euryarchaeota (a Marine-Group-II/III archaeon)—and three predictor types: phosphate (three MAGs, all negative), nitrate+nitrite (five MAGs, three negative and two positive), and log N:Si (two MAGs, both negative). Seven of the ten associations are negative, i.e. the MAG is relatively enriched where the nutrient (or ratio) is low—the expected signature of oligotroph specialists and high-affinity nutrient scavengers. Genome-quality statistics (Table 1) confirm that the ranked MAGs are largely credible: CheckM completion ranges from 59–97% with low redundancy, and genome sizes (0.89–3.25 Mbp) are consistent with the streamlined-to-moderate genomes typical of open-ocean lineages. The sole exception is the nitrate-positive Marine-Group-II/III euryarchaeon (`TARA_IOS_MAG_00039`), whose 32% completion is the lowest in the set; its coverage-based association remains directionally and significance-stable in all twelve folds, but the partial genome weakens (without invalidating) that signal, and we flag it as the one hit most in need of an improved reconstruction before functional follow-up.

Table 1: The ten globally stable Tara Oceans MAGs. Ranked by absolute median leave-one-region-out (LORO) coefficient. All ten are FDR-significant in the full model and directionally *and* significance-stable in all twelve LORO folds (12/12). Coefficients are standardized (per 1 SD of predictor) mixed-model estimates; sign indicates the direction of association. Compl./Red. are CheckM completion and redundancy (%).

Rank	MAG	Predictor	Coef.	q-value	Sig. folds	Taxonomy (Phylum/Class)	Compl
1	`TARA_PSE_MAG_00096`	Phosphate	-2.54	1.0×10^{-3}	12/12	Gammaproteobacteria	
2	`TARA_MED_MAG_00059`	Nitrate+Nitrite	-1.45	5.6×10^{-7}	12/12	Alphaproteo. (Rhizobiales)	
3	`TARA_RED_MAG_00096`	Phosphate	-1.49	3.2×10^{-4}	12/12	Verrucomicrobia (Opitutae)	
4	`TARA_IOS_MAG_00039`	Nitrate+Nitrite	$+1.35$	9.9×10^{-4}	12/12	Euryarchaeota	
5	`TARA_ANE_MAG_00012`	Nitrate+Nitrite	-1.38	3.2×10^{-3}	12/12	Gammaproteo. (Legionellales)	
6	`TARA_SOC_MAG_00012`	Nitrate+Nitrite	$+1.31$	3.1×10^{-3}	12/12	Bacteroidetes (Flavobacteriia)	
7	`TARA_ION_MAG_00049`	log N:Si	-1.21	6.3×10^{-4}	12/12	Gammaproteobacteria	
8	`TARA_MED_MAG_00110`	Nitrate+Nitrite	$+1.15$	6.0×10^{-4}	12/12	Gammaproteobacteria	
9	`TARA_IOS_MAG_00016`	log N:Si	-1.16	1.7×10^{-3}	12/12	*Ca.* Marinimicrobia	
10	`TARA_MED_MAG_00070`	Phosphate	-1.20	2.7×10^{-3}	12/12	Proteobacteria	

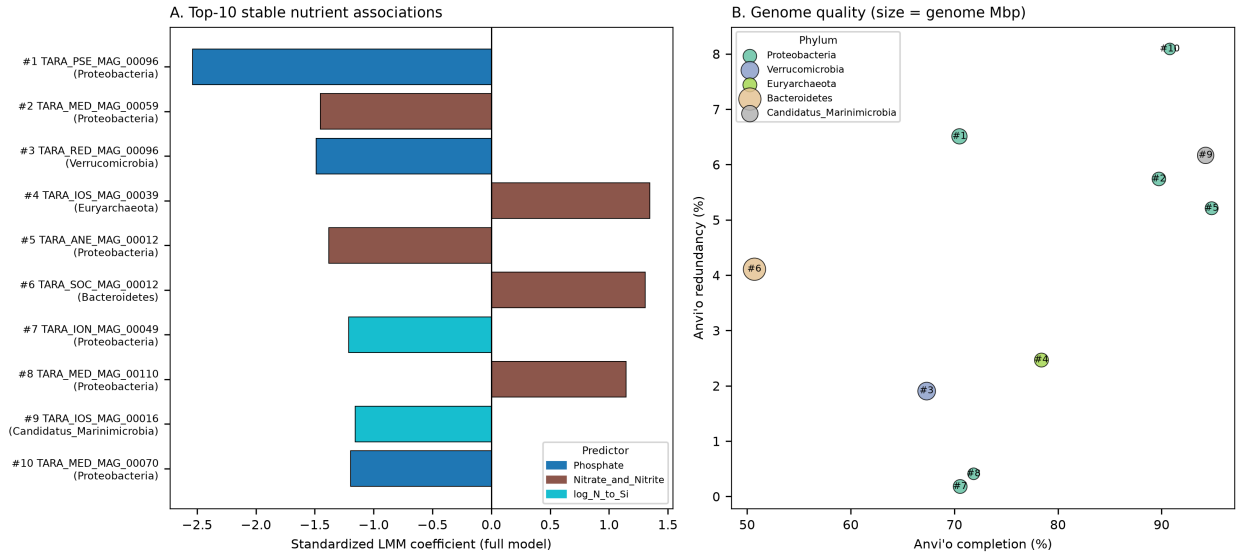

Figure 6: Top-ten stable associations and genome quality. (A) Standardized full-model coefficient for each of the ten ranked MAGs, coloured by predictor (phosphate, nitrate+nitrite, log N:Si). Negative bars indicate enrichment where the nutrient/ratio is low. (B) Genome-quality space (anvi'o completion versus redundancy; symbol size $\propto$ genome size), coloured by phylum, showing that the ranked MAGs are of credible quality across five phyla.

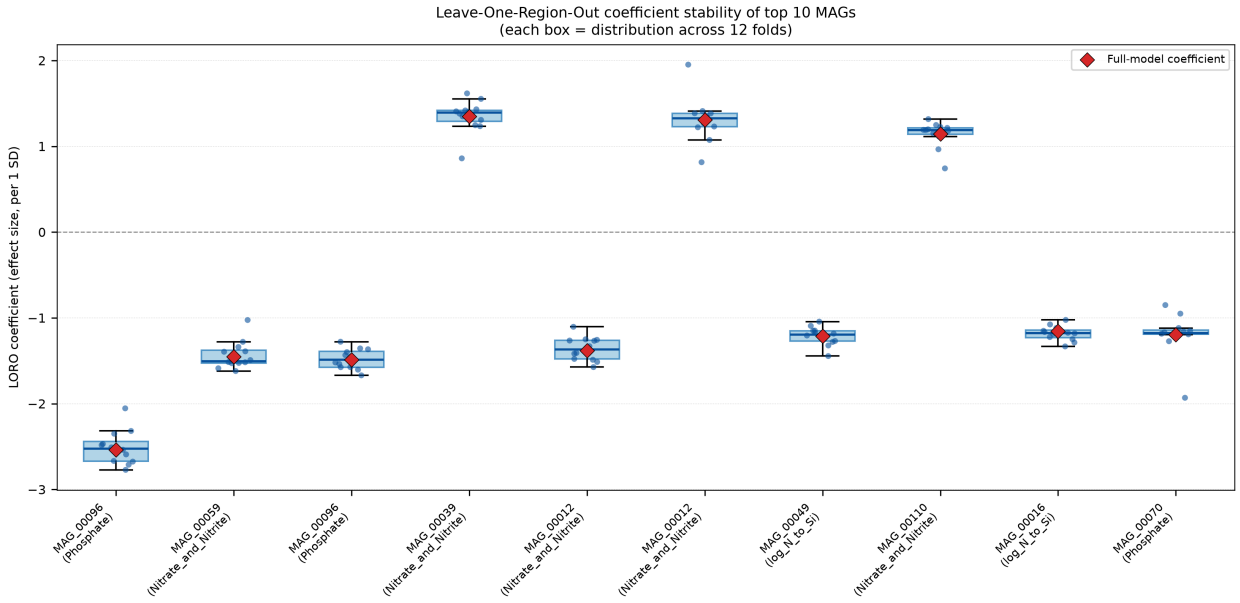

Figure 7: Coefficient stability across leave-one-region-out folds. For each of the ten MAGs, the box shows the distribution of the standardized coefficient across the twelve LORO folds; the red diamond marks the full-model coefficient. Distributions are tight and never cross zero, i.e. neither the magnitude nor the direction of any association depends on the inclusion of a particular ocean region.

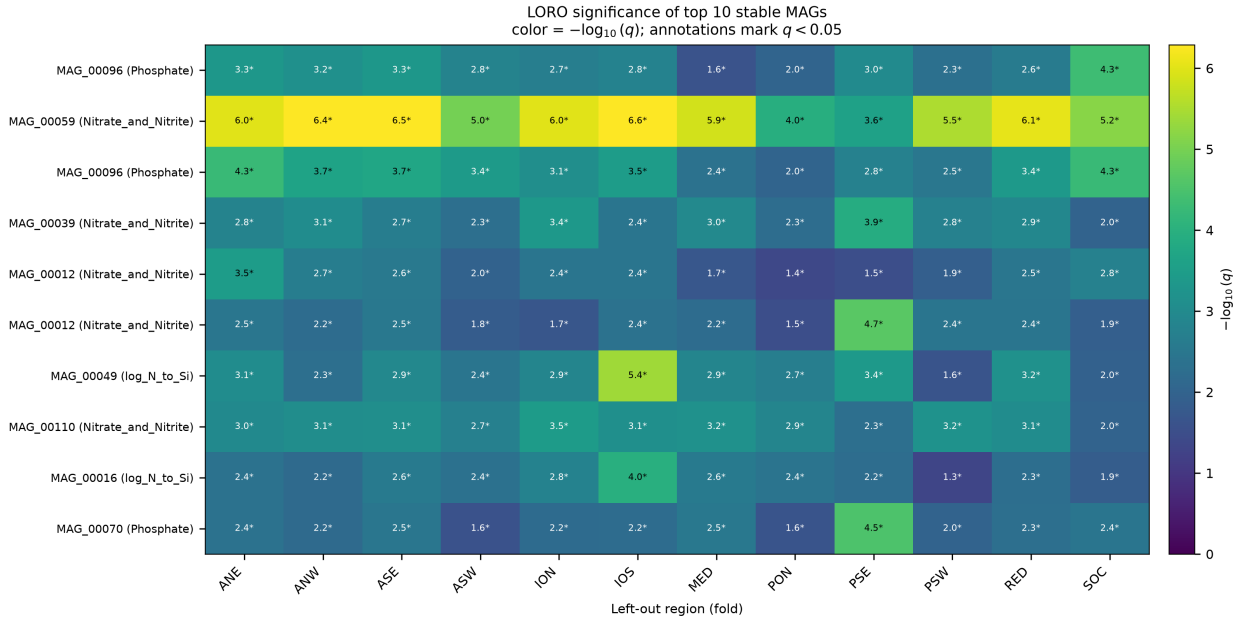

Figure 8: Per-fold significance of the ten stable MAGs. Colour encodes $-\log_{10}(q)$ for each MAG when each ocean region (columns) is held out; asterisks mark $q < 0.05$. Every cell is significant, demonstrating that no association is carried by a single region. `TARA_MED_MAG_00059` (nitrate+nitrite) is the most strongly and uniformly significant.

3.4 Lineage-based functional interpretation—explicitly correlational

Because none of the ten MAGs carries gene-level annotation in the resource's functional tables (0/10 in the KEGG/RAST and *nifH* tables), we interpret each association through the physiology of its lineage, treating the interpretation as a hypothesis to be tested rather than a demonstrated mechanism. The three phosphate-negative MAGs (`TARA_PSE_MAG_00096`, `TARA_RED_MAG_00096`, `TARA_MED_MAG_00070`) are enriched in P-depleted waters, a pattern consistent with the possession of high-affinity phosphate transport (*pstSCAB*), organic-phosphorus scavenging via alkaline phosphatases (*phoX*/*phoA*) or C–P lyase, hallmarks of oligotrophic P-stress adaptation documented in streamlined marine heterotrophs and *Prochlorococcus* [32–34]. The three nitrate-positive MAGs (a Marine-Group-II/III euryarchaeon, a Flavobacterium, and a Gammaproteobacterium) track oxidised-N-rich waters, compatible with assimilatory nitrate/nitrite reduction (*nasA*/*nirA*/*narB*)—although for the photoheterotrophic archaeon and the polysaccharide-degrading Flavobacterium, a niche/co-occurrence explanation (tracking upwelling or phytoplankton-derived organic matter) is at least as plausible as direct uptake [21, 35]. The two nitrate-negative MAGs (a *Rhizobiales* alphaproteobacterium and a Legionellales gammaproteobacterium) favour low-nitrate, stratified surface waters, as expected of populations relying on regenerated nitrogen (ammonium via *amtB*/*glnA*) or, potentially, on N_2 fixation. The two log N:Si-negative MAGs—a Gammaproteobacterium and a *Candidatus* Marinimicrobia (SAR406) genome—are associated with a stoichiometric ratio rather than a single nutrient, most likely reflecting community context (e.g. diatom-driven silicate drawdown shifting N:Si, or a nutricline niche) rather than a direct single-nutrient transporter dependency [36, 38].

3.5 Alternative, non-causal explanations that stability cannot exclude

Leave-one-region-out stability establishes that our associations are spatially reproducible and not artefacts of a single ocean province, but it does *not* establish causation or a direct metabolic dependency. We consider three classes of alternative explanation explicitly. First, *physical transport and advection*: ocean currents co-transport cells and dissolved nutrients within water masses, so passive tracers can appear "associated" without any metabolic link. Our region random intercept and the temperature/salinity/depth covariates absorb broad water-mass identity, and LORO removes signals driven by a single advective province, but sub-regional mesoscale covariation is not fully removed. Second, *niche overlap and unmeasured covariation*: iron, light, mixed-layer depth and dissolved organic matter co-vary with nutrients and are absent from the model; because these are strong open-ocean niche axes, and because linear models cannot capture unimodal niche optima, a measured-nutrient coefficient may be a proxy for an unmeasured driver. This risk is highest for the two ratio-based ($\log$ N:Si) associations. Third, *biotic interactions*: heterotrophic bacteria frequently depend on phytoplankton-derived organic matter, and Bacteroidetes and many Proteobacteria bloom with algal hosts, so a nutrient association may be mediated by a biological partner rather than by direct nutrient metabolism [13, 20, 21, 35]. We did not model prokaryote–eukaryote co-occurrence, so this explanation remains open, particularly for the silicate/N:Si associations, which can act as proxies for diatom abundance [38]. These caveats motivate the manipulative experiment below, which is the only way to separate direct nutrient dependence from correlated confounders.

3.6 A nitrate × phosphate microcosm to test the associations

To move from stable correlation toward causal inference, we designed a 2×2 factorial nitrate × phosphate enrichment microcosm (Fig. 9). Oligotrophic surface (or DCM) water in which the target MAGs are confirmed abundant—candidate sources include Mediterranean, South Pacific gyre, Indian Ocean and Red Sea stations that supplied the ranked MAGs—is collected by Niskin rosette, gently pre-screened through 200 μm mesh, and dispensed into acid-washed 2.4 L polycarbonate bottles. Four treatments are established: an unamended control, +nitrate ($NaNO_3$, +10 μM), +phosphate (KH_2PO_4, +1 μM), and +nitrate+phosphate, each in $n = 4$ independent bottles, incubated at in-situ temperature under a 12h:12h light cycle (∼30% surface PAR) for 96 h with gentle daily mixing. Sampling at six timepoints (0, 12, 24, 48, 72, 96 h) captures chemical (DIN, soluble reactive phosphorus, silicate), biological (flow- cytometric cell counts, chlorophyll) and molecular (DNA and RNA on 0.2 μm Sterivex filters) responses, with RNA prioritised at early timepoints when regulatory signals are strongest.

Each MAG is assigned to one of four functional response groups directly from its (predictor, sign) pair, and each group is given a directional prediction and a tailored marker-gene panel (Table 2). The three phosphate-negative MAGs (group P-negative) are predicted to decline, with repression of *pstS*/*phoX*/*phoA*, under phosphate addition; the three nitrate-positive MAGs (N-positive) are predicted to increase, with induction of *glnA*/*nasA*, under nitrate addition; the two nitrate-negative MAGs (N-negative) are predicted to decline, with repression of the ammonium transporter *amtB*, under nitrate addition; and the two $\log$ N:Si-negative MAGs (N:Si-negative), which behave like low-N-favouring taxa on the manipulated nitrogen axis, are predicted to decline under nitrate addition, with silicate/diatom coupling monitored but not manipulated. Readouts follow MIQE reporting standards [37]: MAG-specific qPCR/RT-qPCR primers are designed against each target-gene allele in the MAG assembly, verified for uniqueness against the other nine genomes and a metagenomic background, and normalised to total 16S rRNA gene/transcript copies and

a single-copy marker (*recA*/*rpoB*); where per-MAG primers are infeasible, shotgun metagenomic read recruitment (coverage/RPKM) and metatranscriptomic TPM provide the relative-abundance and expression readouts. Responses are analysed per MAG and per gene with a two-way ANOVA / mixed model (response $\sim$ N $\times$ P with time), testing the nitrogen main effect, phosphorus main effect and their interaction, with Benjamini–Hochberg FDR correction across the ten MAG $\times$ gene tests and effect sizes (η^2/ω^2 with 95% CI) reported alongside p-values [29, 39]. With four cells and $n = 4$ (residual df = 12) the design has >80% power to detect large main/interaction effects (Cohen $f \geq 0.7$) typical of enrichment-driven shifts.

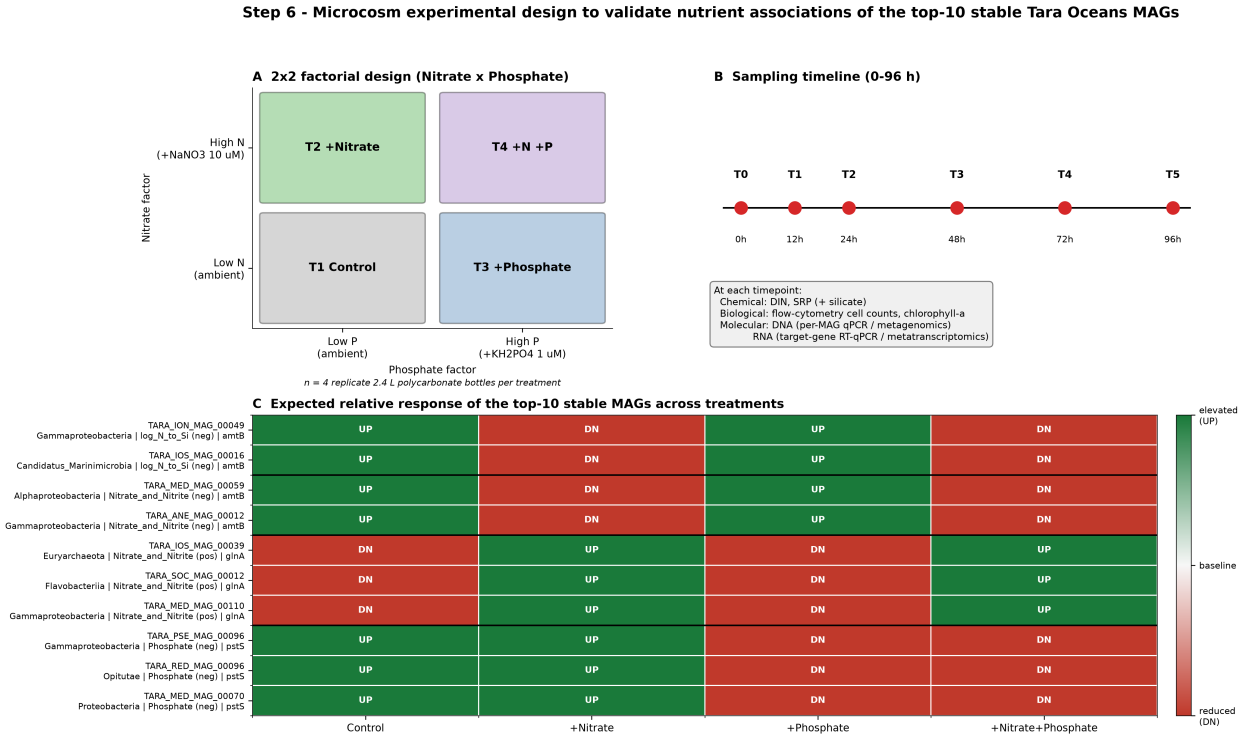

Figure 9: Microcosm experimental design. (A) 2×2 factorial layout crossing nitrate ($NaNO_3$, $+10\,\mu M$) and phosphate (KH_2PO_4, $+1\,\mu M$), $n = 4$ bottles per treatment. (B) Sampling timeline over 96 h (T0–T5) with chemical, biological and molecular measurements at each point. (C) Expected relative response of each of the ten stable MAGs across the four treatments (green: enriched above baseline; red: reduced below baseline), derived directly from each MAG's field association and functional group. The design converts each observational association into a falsifiable directional hypothesis.

Table 2: Functional response groups and molecular readouts for the microcosm. Groups are assigned from each MAG's field (predictor, sign) pair. Expected response under the four treatments (Ctrl/+N/+P/+NP) codes the predicted direction relative to baseline in each treatment (+ enriched above baseline, − reduced below baseline). Target genes are lineage-inferred and must be confirmed by direct re-annotation of each MAG genome before primer design.

Group	MAGs (n)	Axis	Ctrl/+N/+P/+NP	Primary / secondary genes	Prediction
P-negative	3	Phosphate	+ / + / − / −	*pstS* / *pstCAB*, *phoX*, *phoA*, *phnJ*	Decline and P-uptake-gene repression under +P
N-positive	3	Nitrate	− / + / − / +	*glnA* / *nasA*, *nirA*, *narB*, *nrtA*	Increase and N-assimilation induction under +N
N-negative	2	Nitrate	+ / − / + / −	*amtB* / *glnA*, *glnK*, *ureC*	Decline and *amtB* re-pression under +N
N:Si-negative	2	Nitrate (via N:Si)	+ / − / + / −	*amtB* / *nasA*, *glnA*	Decline under +N; silicate context monitored only

4 Discussion

Starting from three independently archived Tara Oceans resources, we identified ten metagenome-assembled genomes whose surface-ocean distributions are associated with dissolved nutrients in a way that survives compositional correction, false-discovery control, robust re-estimation, and—most importantly—leave-one-region-out cross-validation across twelve ocean basins. The central methodological claim of this study is that *stability*, not mere significance, is the appropriate bar for a global association: of 314 FDR-significant MAG–nutrient links, only about half (168) were both directionally and statistically stable across all twelve regional hold-outs, and the ten we prioritise are significant and correctly signed in every fold. This is a deliberately conservative screen. It discards associations that, while real in the full dataset, are carried by one or two peculiar provinces, and it therefore yields a shortlist that is far more likely to reflect a genuinely pan-oceanic ecological relationship than a single-basin artefact.

The biology of the shortlist is coherent with decades of marine biogeochemistry. Seven of the ten MAGs are *negatively* associated with a nutrient or nutrient ratio, meaning they are relatively more abundant where that nutrient is scarce—the expected signature of oligotroph specialists that win in dilute waters through high-affinity uptake and scavenging rather than through fast growth on pulses [32, 33]. The three phosphate-negative populations echo the phosphorus-acquisition adaptations that structure *Prochlorococcus* and SAR11 populations along basin-scale P gradients [33, 34], while the nitrate-negative *Rhizobiales* and Legionellales MAGs are consistent with populations that thrive on regenerated nitrogen in stratified, low-nitrate surface waters. The positive nitrate associations of a Marine-Group-II/III euryarchaeon, a Flavobacterium and a Gammaproteobacterium are more ambiguous: for the photoheterotrophic archaeon and the polysaccharide-degrading Flavobacterium in particular, tracking oxidised-N-rich upwelling or phytoplankton-derived organic matter is at least as plausible as direct nitrate uptake [14, 21, 35]. This ambiguity is not a flaw in the analysis but a fundamental property of observational ocean data, and it is precisely why we frame the functional interpretation as lineage-based and hypothesis-generating.

Two limitations deserve emphasis. First, none of the ten MAGs carries gene-level functional annotation in the resource's supplement; the KEGG/RAST and *nifH* tables were curated for the heterotrophic-diazotroph subset, not for these general populations. Our marker-gene interpretations (*pstS*, *phoX*, *nasA*, *amtB*, *glnA*) are therefore inferred from lineage physiology and must be confirmed by direct re-annotation of each genome (e.g. with Prokka, DRAM, or KOfam) before they can be treated as established metabolic capacities. Second, and more deeply, the entire field analysis is correlational. The strong collinearity among nitrate, phosphate and log N:Si (Fig. 3) means that a single-nutrient coefficient cannot, on its own, identify the causal driver, and the alternative explanations we enumerate—advective co-transport, unmeasured co-limiting factors such as iron and light, and biotic mediation by eukaryotic partners—remain live for every one of the ten hits. Function–taxonomy decoupling in the ocean microbiome [12] further cautions against reading a taxon's identity as a guarantee of its metabolism.

These limitations define the value of the microcosm experiment. By crossing nitrate and phosphate in a factorial design and reading out both abundance and target-gene expression per MAG, the experiment converts each stable association into a directional, falsifiable prediction: phosphate-negative MAGs should decline and repress *pstS*/*phoX* under added phosphate; nitrate-positive MAGs should increase and induce *glnA*/*nasA* under added nitrate; nitrate- and N:Si-negative MAGs should decline under added nitrate. Crucially, the factorial structure also tests for N×P interactions and, because the manipulation is direct, breaks the field collinearity that no amount of statistical adjustment can fully resolve. The design's own limitations are acknowledged: bottle and containment effects over 96 h, the fact that silicate/diatom coupling is monitored but not manipulated (so the N:Si associations are only partially addressed), possible micronutrient co-limitation, and the risk that enrichment selects for fast responders so that slow-growing oligotrophs—most obviously the archaeal MAG—may respond weakly within four days. A full test of the N:Si associations would require a three-factor design that also manipulates silicate.

Several extensions follow naturally. Re-annotating the ten genomes would convert lineage-based marker predictions into genome-verified capacities and sharpen primer design. Fitting a jointly regularised model (e.g. Elastic Net) across the correlated nutrients could help attribute each MAG to its primary driver, and adding iron, light and dissolved-organic-matter proxies as covariates would shrink the unmeasured-confounder space. Building prokaryote–eukaryote co-occurrence networks [13, 20] would directly test biotic-partner mediation, especially for the silicate and N:Si associations. More broadly, the workflow demonstrated here—audit joins, respect compositionality, control the false-discovery rate, and demand cross-regional stability before believing a global association—is transferable to any large environmental-omics survey in which spatial clustering and compositional data threaten to manufacture reproducible-looking but non-generalisable correlations.

In conclusion, we deliver a ranked, auditable and directly testable set of ten ocean genomes whose distributions are robustly tied to phosphate, nitrate or nutrient stoichiometry across the global surface ocean, together with an explicit accounting of what the associations can and cannot mean and a concrete experiment to resolve the difference. Treating association as a starting hypothesis rather than a conclusion, this study turns a global biogeographic pattern into a program of falsifiable, genome-resolved ecophysiology.

Data and code availability

All primary data are publicly archived: Tara Oceans non-redundant MAG distributions (Figshare `10.6084/m9.figshare.4902938.v3`), PANGAEA nutrient measurements (`10.1594/PANGAEA.875575`)

and the Tara sample registry (`10.1594/PANGAEA.875582`). The analysis pipeline (data acquisition and sample-join audit, preprocessing and CLR transform, association modelling with FDR control, LORO stability analysis, functional interpretation, and microcosm design) was executed with a fixed random seed (42); intermediate tables (sample-join audit, transformed abundances, association results, LORO stability summaries, the annotated top-ten table, and the machine-readable microcosm design) accompany this manuscript.

Author contributions

The computational analysis, figures, and experimental design described here were produced by the automated research workflow that accompanies this manuscript.

Acknowledgements

This work reuses data generated by the Tara Oceans consortium and archived in PANGAEA and Figshare; we acknowledge the consortium and the sampling, sequencing and data-management teams whose open data made this reanalysis possible.

References

[1] C. M. Moore, M. M. Mills, K. R. Arrigo, I. Berman-Frank, L. Bopp, P. W. Boyd, E. D. Galbraith, R. J. Geider, C. Guieu, S. L. Jaccard, T. D. Jickells, J. La Roche, T. M. Lenton, N. M. Mahowald, E. Marañón, I. Marinov, J. K. Moore, T. Nakatsuka, A. Oschlies, M. A. Saito, T. F. Thingstad, A. Tsuda, and O. Ulloa. Processes and patterns of oceanic nutrient limitation. *Nature Geoscience*, 6(9):701–710, 2013. doi: 10.1038/ngeo1765.

[2] Toby Tyrrell. The relative influences of nitrogen and phosphorus on oceanic primary production. *Nature*, 400(6744):525–531, 1999. doi: 10.1038/22941.

[3] Thomas J. Browning and C. Mark Moore. Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation. *Nature Communications*, 14:5014, 2023. doi: 10.1038/s41467-023-40774-0.

[4] Alfred C. Redfield. The biological control of chemical factors in the environment. *American Scientist*, 46(3):205–221, 1958.

[5] Adam C. Martiny, Chau T. A. Pham, Francois W. Primeau, Jasper A. Vrugt, J. Keith Moore, Simon A. Levin, and Michael W. Lomas. Strong latitudinal patterns in the elemental ratios of marine plankton and organic matter. *Nature Geoscience*, 6(4):279–283, 2013. doi: 10.1038/ngeo1757.

[6] Shinichi Sunagawa, Luis Pedro Coelho, Samuel Chaffron, Jens Roat Kultima, Karine Labadie, Guillem Salazar, Bardya Djahanschiri, Georg Zeller, Daniel R. Mende, Adriana Alberti, Francisco M. Cornejo-Castillo, Paul I. Costea, Corinne Cruaud, Francesco d'Ovidio, Stefan Engelen, Isabel Ferrera, Josep M. Gasol, Lionel Guidi, Falk Hildebrand, Florian Kokoszka, Cyrille Lepoivre, Gipsi Lima-Mendez, Julie Poulain, Bonnie T. Poulos, Marta Royo-Llonch, Hugo Sarmento, Sara Vieira-Silva, Céline Dimier, Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis,

Tara Oceans coordinators, Chris Bowler, Colomban de Vargas, Gabriel Gorsky, Nigel Grimsley, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Jean Weissenbach, Patrick Wincker, Eric Karsenti, Jeroen Raes, Silvia G. Acinas, and Peer Bork. Structure and function of the global ocean microbiome. *Science*, 348(6237):1261359, 2015. doi: 10.1126/science.1261359.

[7] Shinichi Sunagawa, Silvia G. Acinas, Peer Bork, Chris Bowler, Tara Oceans Coordinators, Damien Eveillard, Gabriel Gorsky, Lionel Guidi, Daniele Iudicone, Eric Karsenti, Fabien Lombard, Hiroyuki Ogata, Stephane Pesant, Matthew B. Sullivan, Patrick Wincker, and Colomban de Vargas. Tara Oceans: towards global ocean ecosystems biology. *Nature Reviews Microbiology*, 18(8):428–445, 2020. doi: 10.1038/s41579-020-0364-5.

[8] Tom O. Delmont, Christopher Quince, Alon Shaiber, Özcan C. Esen, Sonny T. M. Lee, Michael S. Rappé, Sandra L. McLellan, Sebastian Lücker, and A. Murat Eren. Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nature Microbiology*, 3(7):804–813, 2018. doi: 10.1038/s41564-018-0176-9.

[9] Tom O. Delmont, Juan José Pierella Karlusich, Iva Veseli, Jessika Fuessel, A. Murat Eren, Rachel A. Foster, Chris Bowler, Patrick Wincker, and Eric Pelletier. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *The ISME Journal*, 16(4):927–936, 2022. doi: 10.1038/s41396-021-01135-1.

[10] Benjamin J. Tully, Elaina D. Graham, and John F. Heidelberg. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, 5:170203, 2018. doi: 10.1038/sdata.2017.203.

[11] Guillem Salazar, Lucas Paoli, Adriana Alberti, Jaime Huerta-Cepas, Hans-Joachim Ruscheweyh, Miguelangel Cuenca, Christopher M. Field, Luis Pedro Coelho, Corinne Cruaud, Stefan Engelen, Ann C. Gregory, Karine Labadie, Claudie Marec, Eric Pelletier, Marta Royo-Llonch, Simon Roux, Pablo Sánchez, Hideya Uehara, Ahmed A. Zayed, Georg Zeller, Margaux Carmichael, Céline Dimier, Joannie Ferland, Stefanie Kandels, Marc Picheral, Sergey Pisarev, Julie Poulain, Tara Oceans Coordinators, Silvia G. Acinas, Marcel Babin, Peer Bork, Emmanuel Boss, Chris Bowler, Guy Cochrane, Colomban de Vargas, Michael Follows, Gabriel Gorsky, Nigel Grimsley, Lionel Guidi, Pascal Hingamp, Daniele Iudicone, Olivier Jaillon, Stefanie Kandels, Lee Karp-Boss, Eric Karsenti, Fabrice Not, Hiroyuki Ogata, Stéphane Pesant, Nicole Poulton, Jeroen Raes, Christian Sardet, Sabrina Speich, Lars Stemmann, Matthew B. Sullivan, Shinichi Sunagawa, and Patrick Wincker. Gene expression changes and community turnover differentially shape the global ocean metatranscriptome. *Cell*, 179(5):1068–1083, 2019. doi: 10.1016/j.cell.2019.10.014.

[12] Stilianos Louca, Laura Wegener Parfrey, and Michael Doebeli. Decoupling function and taxonomy in the global ocean microbiome. *Science*, 353(6305):1272–1277, 2016. doi: 10.1126/science.aaf4507.

[13] Gipsi Lima-Mendez, Karoline Faust, Nicolas Henry, Johan Decelle, Sébastien Colin, Fabrizio Carcillo, Samuel Chaffron, J. César Ignacio-Espinosa, Simon Roux, Flora Vincent, Lucie Bittner, Youssef Darzi, Jun Wang, Stéphane Audic, Leo Berline, Gianluca Bontempi, Ana M. Cabello, Laurent Coppola, Francisco M. Cornejo-Castillo, Francesco d'Ovidio, Luc De Meester, Isabel Ferrera, Marie-José Garet-Delmas, Lionel Guidi, Elena Lara, Stéphane Pesant, Marta Royo-Llonch, Guillem Salazar, Pablo Sánchez, Marta Sebastian, Caroline Souffreau, Céline Dimier,

Marc Picheral, Sarah Searson, Stefanie Kandels-Lewis, Tara Oceans coordinators, Gabriel Gorsky, Fabrice Not, Hiroyuki Ogata, Sabrina Speich, Lars Stemmann, Jean Weissenbach, Patrick Wincker, Silvia G. Acinas, Shinichi Sunagawa, Peer Bork, Matthew B. Sullivan, Eric Karsenti, Chris Bowler, Colomban de Vargas, and Jeroen Raes. Determinants of community structure in the global plankton interactome. *Science*, 348(6237):1262073, 2015. doi: 10.1126/science.1262073.

[14] Jed A. Fuhrman, Jacob A. Cram, and David M. Needham. Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, 13(3):133–146, 2015. doi: 10.1038/nrmicro3417.

[15] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome datasets are compositional: and this is not optional. *Frontiers in Microbiology*, 8:2224, 2017. doi: 10.3389/fmicb.2017.02224.

[16] John Aitchison. The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160, 1982. doi: 10.1111/j.2517-6161.1982.tb01195.x.

[17] Thomas P. Quinn, Ionas Erb, Mark F. Richardson, and Tamsyn M. Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018. doi: 10.1093/bioinformatics/bty175.

[18] David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS Computational Biology*, 11(3):e1004075, 2015. doi: 10.1371/journal.pcbi.1004075.

[19] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5:27, 2017. doi: 10.1186/s40168-017-0237-y.

[20] Karoline Faust and Jeroen Raes. Microbial interactions: from networks to models. *Nature Reviews Microbiology*, 10(8):538–550, 2012. doi: 10.1038/nrmicro2832.

[21] Alison Buchan, Gary R. LeCleir, Christopher A. Gulvik, and José M. González. Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology*, 12(10):686–698, 2014. doi: 10.1038/nrmicro3326.

[22] Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010. doi: 10.1111/j.1467-9868.2010.00740.x.

[23] A. Murat Eren, Özcan C. Esen, Christopher Quince, Joseph H. Vineis, Hilary G. Morrison, Mitchell L. Sogin, and Tom O. Delmont. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, 3:e1319, 2015. doi: 10.7717/peerj.1319.

[24] Donovan H. Parks, Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, 25(7):1043–1055, 2015. doi: 10.1101/gr.186072.114.

[25] Stéphane Pesant, Fabrice Not, Marc Picheral, Stefanie Kandels-Lewis, Noan Le Bescot, Gabriel Gorsky, Daniele Iudicone, Eric Karsenti, Sabrina Speich, Romain Troublé, Céline Dimier, Sarah Searson, and Tara Oceans Consortium Coordinators. Open science resources for the discovery and analysis of Tara Oceans data. *Scientific Data*, 2:150023, 2015. doi: 10.1038/sdata.2015.23.

[26] Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11:3514, 2020. doi: 10.1038/s41467-020-17041-7.

[27] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015. doi: 10.18637/jss.v067.i01.

[28] Skipper Seabold and Josef Perktold. statsmodels: econometric and statistical modeling with Python. *Proceedings of the 9th Python in Science Conference*, pages 92–96, 2010. doi: 10.25080/Majora-92bf1922-011.

[29] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[30] John D. Storey. The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics*, 31(6):2013–2035, 2003. doi: 10.1214/aos/1074290335.

[31] Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

[32] Stephen J. Giovannoni, H. James Tripp, Scott Givan, Mircea Podar, Kevin L. Vergin, Damon Baptista, Lisa Bibbs, Jonathan Eads, Toby H. Richardson, Michiel Noordewier, Michael S. Rappé, Jay M. Short, James C. Carrington, and Eric J. Mathur. Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, 309(5738):1242–1245, 2005. doi: 10.1126/science.1114057.

[33] Sarah M. Sowell, Larry J. Wilhelm, Angela D. Norbeck, Mary S. Lipton, Carrie D. Nicora, Douglas F. Barofsky, Craig A. Carlson, Richard D. Smith, and Stephen J. Giovannoni. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *The ISME Journal*, 3(11):1260–1270, 2009. doi: 10.1038/ismej.2009.70.

[34] Maureen L. Coleman and Sallie W. Chisholm. Ecosystem-specific selection pressures revealed through comparative population genomics. *Proceedings of the National Academy of Sciences*, 107(43):18634–18639, 2010. doi: 10.1073/pnas.1009480107.

[35] Hanno Teeling, Bernhard M. Fuchs, Dörte Becher, Christine Klockow, Antje Gardebrecht, Christin M. Bennke, Mariette Kassabgy, Sixing Huang, Alexander J. Mann, Jost Waldmann, Michael Weber, Anna Klindworth, Andreas Otto, Jana Lange, Jörg Bernhardt, Christine Reinsch, Michael Hecker, Jörg Peplies, Frank D. Bockelmann, Ulrich Callies, Gunnar Gerdts, Antje Wichels, Karen H. Wiltshire, Frank Oliver Glöckner, Thomas Schweder, and Rudolf Amann. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science*, 336(6081):608–611, 2012. doi: 10.1126/science.1218344.

[36] Christian Rinke, Patrick Schwientek, Alexander Sczyrba, Natalia N. Ivanova, Iain J. Anderson, Jan-Fang Cheng, Aaron Darling, Stephanie Malfatti, Brandon K. Swan, Esther A. Gies, Jeremy A. Dodsworth, Brian P. Hedlund, George Tsiamis, Stefan M. Sievert, Wen-Tso Liu, Jonathan A. Eisen, Steven J. Hallam, Nikos C. Kyrpides, Ramunas Stepanauskas, Edward M.

Rubin, Philip Hugenholtz, and Tanja Woyke. Insights into the phylogeny and coding potential of microbial dark matter. *Nature*, 499(7459):431–437, 2013. doi: 10.1038/nature12352.

[37] Stephen A. Bustin, Vladimir Benes, Jeremy A. Garson, Jan Hellemans, Jim Huggett, Mikael Kubista, Reinhold Mueller, Tania Nolan, Michael W. Pfaffl, Gregory L. Shipley, Jo Vandesompele, and Carl T. Wittwer. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clinical Chemistry*, 55(4):611–622, 2009. doi: 10.1373/clinchem.2008.112797.

[38] Colleen A. Durkin, Julie A. Koester, Sara J. Bender, and E. Virginia Armbrust. The evolution of silicon transporters in diatoms. *Journal of Phycology*, 52(5):716–731, 2016. doi: 10.1111/jpy.12441.

[39] Himel Mallick, Ali Rahnavard, Lauren J. McIver, Siyuan Ma, Yancong Zhang, Long H. Nguyen, Timothy L. Tickle, George Weingart, Boyu Ren, Emma H. Schwager, Suvo Chatterjee, Kelsey N. Thompson, Jeremy E. Wilkinson, Ayshwarya Subramanian, Yiren Lu, Levi Waldron, Joseph N. Paulson, Eric A. Franzosa, Hector Corrada Bravo, and Curtis Huttenhower. Multivariable association discovery in population-scale meta-omics studies. *PLoS Computational Biology*, 17 (11):e1009442, 2021. doi: 10.1371/journal.pcbi.1009442.