Environmental genomics points to non-diazotrophic *Trichodesmium* species abundant and widespread in the open ocean

Tom O. Delmont

Génomique Métabolique, Genoscope, Institut François Jacob, CEA, CNRS, Univ Evry, Université Paris-Saclay, 91057 Evry, France

Correspondence: tomodelmont@gmail.com

Abstract: Filamentous and colony-forming cells within the cyanobacterial genus *Trichodesmium* might account for nearly half of nitrogen fixation in the sunlit ocean, a critical mechanism that sustains plankton's primary productivity at large-scale. Here, we report the genome-resolved metagenomic characterization of two newly identified marine species we tentatively named 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*'. Near-complete environmental genomes for those closely related candidate species revealed unexpected functional features including a lack of the entire nitrogen fixation gene apparatus and hydrogen recycling genes concomitant with the enrichment of nitrogen assimilation genes and apparent acquisition of the *nirb* gene from a non-cyanobacterial lineage. These comparative genomic insights were cross-validated by complementary metagenomic investigations. Our results contrast with the current paradigm that *Trichodesmium* species are necessarily capable of nitrogen fixation. The black queen hypothesis could explain gene loss linked to nitrogen fixation among *Trichodesmium* species, possibly triggered by gene acquisitions from the colony epibionts. Critically, the candidate species are not only widespread in the 3-2000 μm planktonic size fraction of the surface of the oceans and seas, but might also substantially expand the ecological niche of *Trichodesmium*, stressing the need to disconnect taxonomic signal for this genus from a microbial community's ability to fix nitrogen. Especially, differentiating diazotrophic from non-diazotrophic populations when counting *Trichodesmium* filaments and colonies might help refine our understanding of the marine nitrogen balance. While culture representatives are needed to move beyond metagenomic insights, we are reminded that established links between taxonomic lineages and functional traits might not always hold true.

Key words:

Trichodesmium, environmental genomics, metagenomics, nitrogen fixation gene loss, ecology and evolution, climate, black queen hypothesis

Introduction

Plankton in the sunlit ocean includes a wide range of microbial lineages with different functional capabilities that influence global biogeochemical cycles and climate[1–6]. The primary productivity of plankton is constrained by the amount of bioavailable nitrogen[7,8], a critical element for cellular growth and division. Few bacterial and archaeal populations can code for the catalytic (*nifHDK*) and biosynthetic (*nifENB*) proteins required for biological nitrogen fixation, transferring a valuable source of nitrogen from the atmosphere to the plankton[9–11]. These populations are called diazotrophs and represent key marine players that sustain plankton primary productivity in large oceanic regions[9].

Cyanobacterial species within the genus *Trichodesmium* first described in 1830[12] are among the most prominent marine nitrogen fixers[13], possibly accounting for half the biological nitrogen fixation in the sunlit ocean[14–16]. While other makers provide different trends (e.g., in[17]), phylogeny of the 16S rRNA genes points to three distinct *Trichodesmium* clades covering *T. thiebautii* and *T. hildebrandtii* (clade I), *T. tenue* and *T. contortum* (clade II), and *T. erythraeum* and *T. havanum* (clade III)[18]. For now, only cultures of *T. erythraeum* and *T. thiebautti* have been characterized with genomics. Insights from culture representatives and oceanic expeditions have provided a wealth of information regarding their biogeography, functional life styles and nitrogen fixation regulation mechanisms[14,19]. Most notably, *Trichodesmium* cells are capable of forming large blooms of filaments and colonies in the sunlit ocean[13,20–22], regulate nitrogen fixation rates on a daily basis using a dedicated circadian rhythm[23,24] as well as an associated microbiome[25–27], and possess genes linked to the recycling of hydrogen, a by-product of nitrogen fixation[28–30]. Decades of scientific insights have shaped a dogma depicting the numerous *Trichodesmium* cells as carbon and nitrogen fixers with a highly beneficial role for plankton productivity.

Recent large-scale single cell and genome-resolved metagenomic surveys have dramatically expanded the genomic characterization of free-living marine microbes[31–34] and lead to new insights into primary processes in the surface of the open ocean, including nitrogen fixation[34]. However, a major focus on small planktonic size fractions typically excluded organisms such as *Trichodesmium* that form filaments and colonies and occur in larger size fractions[21]. This gap is partially filled by a recent metagenomic survey that reconstructed and manually curated more than one thousand bacterial genomes abundant in large size fractions of the *Tara* Oceans expeditions[35], which included five near-complete environmental *Trichodesmium* genomes. To our knowledge, they correspond to the first *Trichodesmium* genomes recovered without the need for cultivation, providing a new venue to study the ecology and evolution of populations within this genus and occurring in the open ocean. Here, we show that in addition to three genomes that resolve to *T. erythraeum* and *T. thiebautti*, this collection reveals two widespread *Trichodesmium* species most closely related to *T. tenue* and *T. contortum* (clade II) that display functional repertoires denoting a lack of the key ability fixing nitrogen.

Results

A first set of *Trichodesmium* environmental genomes from the sunlit ocean.

A genome-resolved metagenomic survey was recently performed to target planktonic populations abundant in polar, temperate, and tropical sunlit oceans using nearly one thousand metagenomes (total of 280 billion reads) derived from the *Tara* Oceans expeditions and encompassing eight plankton size fractions ranging from 0.8 µm to 2 mm[36] (Table S1). Notably, bacterial metagenome-assembled genomes (MAGs) characterized from this data set covered five distinct *Trichodesmium* populations[35] (Figure 1, Table S2).

The *Trichodesmium* MAGs were affiliated to *T. erythraeum* (one population), *T. thiebautii* (two closely related populations), and two candidate species we tentatively named 'Ca. *Trichodesmium miru*' (one population) and 'Ca. *Trichodesmium nobis*' (one population) that form a distinct evolutionary clade (ANI of 94%) distantly related from both *T. thiebautii* and *T. erythraeum* (ANI <91%) (Figure S1, Table S2). The five MAGs have a length ranging between 5.4 Mbp and 6.8 Mbp, with an estimated completion >90% (average of 96%). They were mostly detected in the Indian Ocean, with the exception of 'Ca. *Trichodesmium miru*' that occurred mostly in the Atlantic Ocean (Table S3). Furthermore, their niche partitioning between the Pacific Ocean, Red Sea and Mediterranean Sea revealed different distribution patterns between the four species. For instance, 'Ca. *Trichodesmium nobis*' was markedly more detected in the Mediterranean Sea. As expected, signal across size fractions indicates that all five populations mostly occur under the form of filaments and colonies in the sunlit ocean. We note that 'Ca. *Trichodesmium miru*' was substantially more detected in the largest size fraction (180-2000 µm), suggesting it forms larger aggregates in the Atlantic Ocean compared to the other *Trichodesmium* populations mostly detected in the Indian Ocean. In terms of overall signal across metagenomes regardless of oceanic regions or size fractions, the two candidate species were abundant but substantially less so than compared to *T. thiebautii* and *T. erythraeum* populations (Table S3). On the other hand, 'Ca. *Trichodesmium miru*' was detected in 29 *Tara* Ocean stations, being markedly more widespread compare to MAGs from the other species, which were detected in a number of stations ranging from 19 to 23 (Tables S2 and S3).

Computing the metagenome-centric consensus of 16S rRNA genes

16S rRNA gene sequences are missing in most MAGs (including within *Trichodesmium*) due to their high evolutionary stability and occurrence in multi-copy (e.g.,[34]). While in uncharted territory, it is in theory possible to retrieve the missing gene of a bacterial population using a closely related reference to recruit metagenomic reads, provided this population dominates the signal within the range of the affiliated genus in the sequence space. Since each of the four *Trichodesmium* species dominated the metagenomic signal at times within the scope of this genus (e.g., in parts of the Atlantic Ocean for 'Ca. *Trichodesmium miru*'), we took this opportunity to recover their 16S rRNA genes. We used the 16S rRNA gene of *T. thiebautii* as bait and performed a stringent read recruitment (to minimize non-specific mapping) for metagenomic triplicates targeting each species (Table S4). For each metagenome, gene positions that differed from the reference among recruited reads were changed accordingly to the most prevalent nucleotide type. Using this approach, we could create 12 metagenome-centric consensus 16S rRNA gene sequences for *Trichodesmium*. As expected, metagenomes dominated by *T. thiebautii* were highly coherent with the reference sequence, requiring only 5 to 8 changes that denote slight differences between the culture and dominant population in one hypervariable region (Figure 1). In contrast, metagenomes dominated by other species required between 25 and 35 changes covering multiple regions of the gene (Table S4), denoting this time greater distances between *Trichodesmium* species.

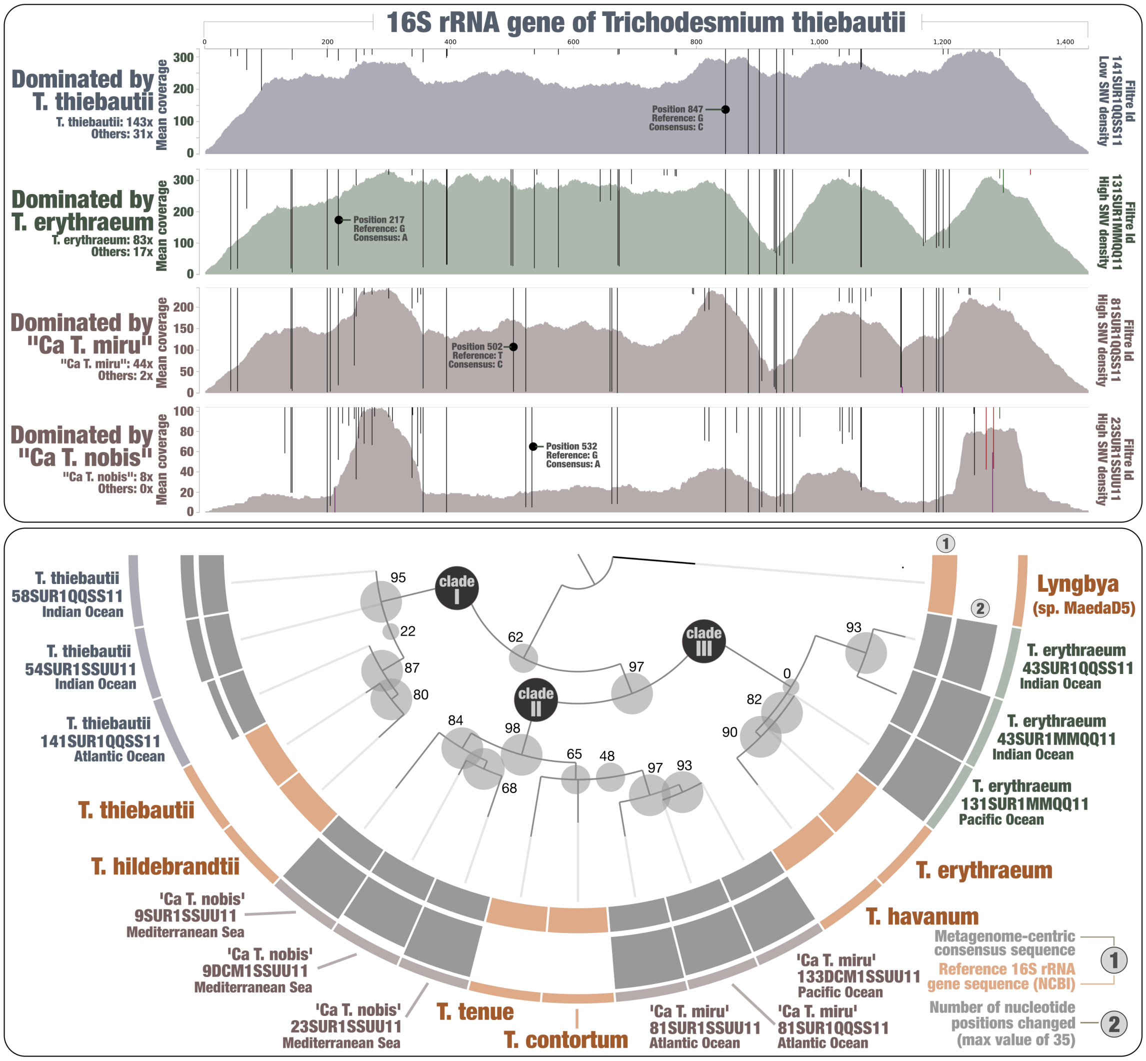


Figure 1: Hybrid phylogeny of the 16S rRNA gene of *Trichodesmium* species using references and metagenomes. Top panel displays read recruitments for the *T. thiebautii* 16S rRNA gene across four metagenomes, each dominated by a single *Trichodesmium* species (mapping stringency of >95% identity over >95% of the read length). Single nucleotide variants identified by anvi'o (when using

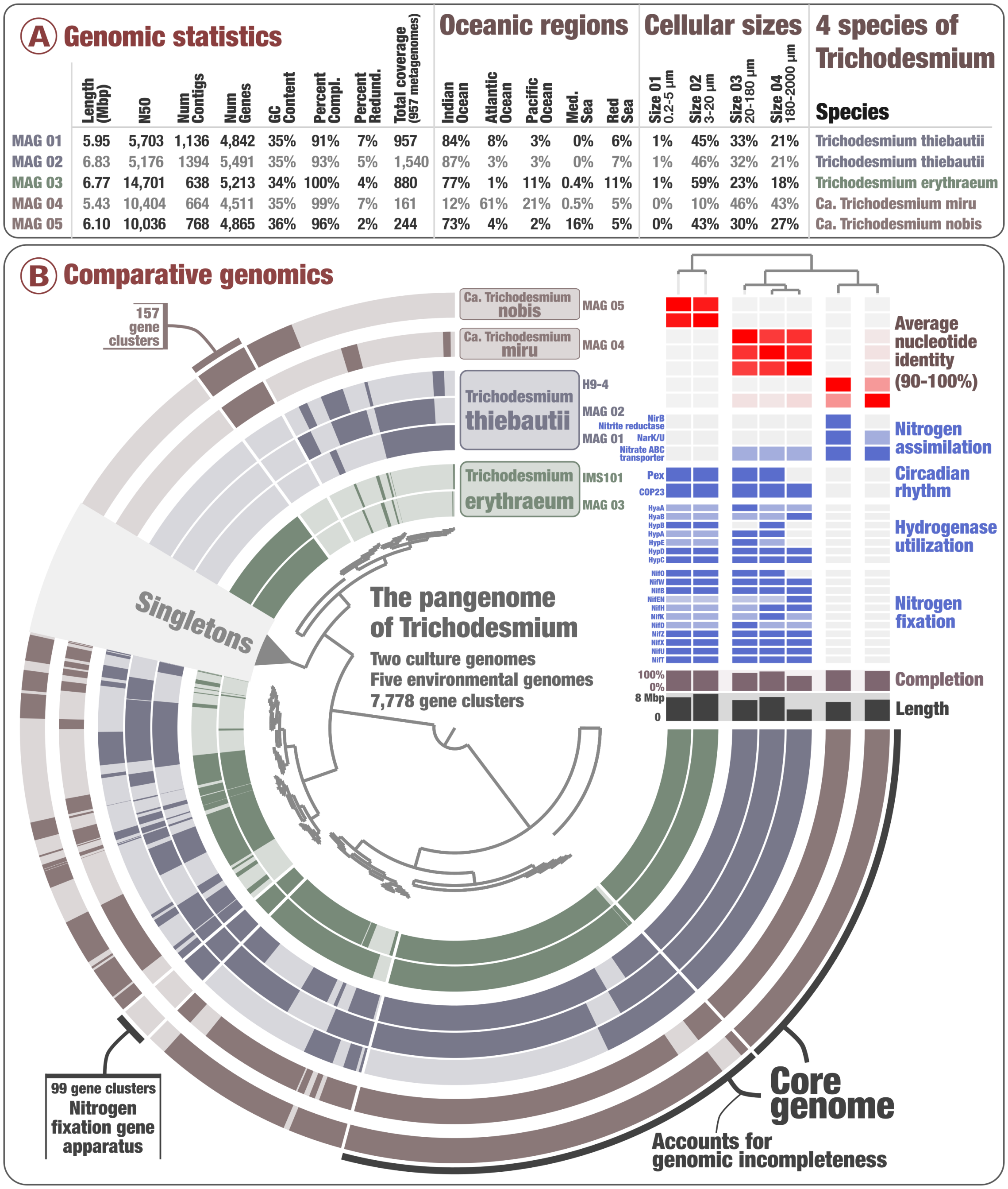
default parameters) were visualized with an amplitude reaching 100% when all reads contained the same nucleotide type. Bottom panel displays a hybrid phylogenetic analysis using both reference sequences and 12 metagenome-centric consensus sequences. The number of changed nucleotide positions compared to the reference (*T. thiebautii* 16S rRNA gene) is presented for each consensus sequence. The phylogenetic tree was decorated with associated data and visualized using anvi'o.

We then performed a phylogenetic analysis using the 12 consensus sequences along with reference 16S rRNA genes retrieved from NCBI, recapitulating the three *Trichodesmium* clades while nesting the candidate species into clade II. NCBI blast confirmed that the candidate species contain 16S rRNA gene signal most closely related to *T. tenue* and *T. contortum* among the metagenomes considered (percent identity >99% for 'Ca. *Trichodesmium miru*' and >98.5% for 'Ca. *Trichodesmium nobis*'). Yet, we detected nucleotide differences in multiple regions of the 16S rRNA gene when using *T. tenue* and *T. contortum* as baits for mapping (Figure S2), suggesting that 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*' correspond to previously uncharacterized species within clade II of this genus.

The case for *Trichodesmium* species lacking nitrogen fixation gene apparatus

We performed a comparative genomic survey of the genus *Trichodesmium* by considering the five newly identified MAGs plus two reference genomes from cultivation and accessed from NCBI: *T. erythraeum* IMS101 (closed genome) and *T. thiebautii* H9-4 (fragmented and only 72% complete). Our pangenomic analysis of these seven genomes with a total of 33,249 genes resulted in 7,778 gene clusters (Figure 1, Table S5). We collapsed singletons (2,542 gene clusters only detected in a single genome) and grouped some of the remaining gene clusters into bins based on their occurrence across genomes: (1) a core-genome (2,183 gene clusters), (2) gene clusters characteristic of *T. erythraeum* and *T. thiebautii* ("*erythraeum/thiebautii*" bin; n=99), (3) and gene clusters characteristic of the two candidate species ("*miru/nobis*" bin; n=157). The overall pangenomic trends for these four *Trichodesmium* species revealed a relatively large pan-genome but also denoted species-specific gene clusters that might be linked to different life styles for *Trichodesmium* clades I, II and III.

In order to provide a global view of functional capabilities across four *Trichodesmium* species, we accessed functions in their gene content using Pfam[37] within the anvi'o pangenomic workflow[38] (Table S5), COG20 functions, categories and pathways[39], KOfam[40], KEGG modules and classes[41] within the anvi'o genomic workflow[42] (Table S6), and RAST annotation[43] (Table S7). The two most prominent functional capabilities of *Trichodesmium* are photosynthesis and nitrogen fixation. As expected, a large set of photosynthetic genes occurred in the seven genomes. On the other hand, multiple lines of evidence - emerging from the inspection of all functional annotations and cross-validated by complementary metagenomic investigations - point to 'Ca. *Trichodesmium nobis*' and 'Ca. *Trichodesmium miru*' lacking the ability to fix nitrogen from the atmosphere, providing a first case for the occurrence, in plain sight, of non-diazotrophic marine *Trichodesmium* species.



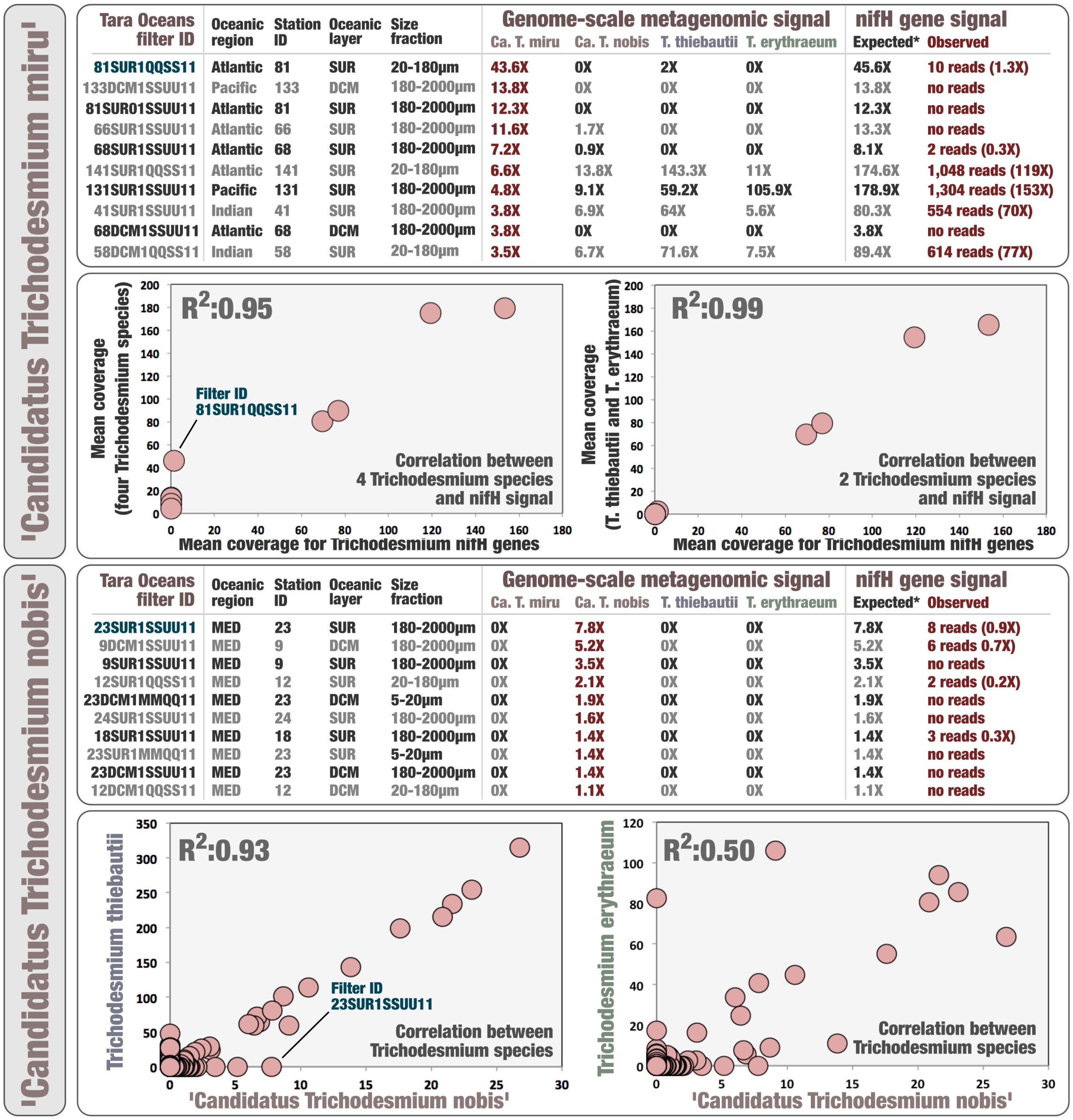
A Genomic statistics

	Genomic statistics								Oceanic regions					Cellular sizes				4 species of Trichodesmium
	Length (Mbp)	N50	Num Contigs	Num Genes	GC Content	Percent Compl.	Percent Redund.	Total coverage (957 metagenomes)	Indian Ocean	Atlantic Ocean	Pacific Ocean	Med. Sea	Red Sea	Size 01 0.2-5 µm	Size 02 3-20 µm	Size 03 20-180 µm	Size 04 180-2000 µm	Species
MAG 01	5.95	5,703	1,136	4,842	35%	91%	7%	957	84%	8%	3%	0%	6%	1%	45%	33%	21%	Trichodesmium thiebautii
MAG 02	6.83	5,176	1394	5,491	35%	93%	5%	1,540	87%	3%	3%	0%	7%	1%	46%	32%	21%	Trichodesmium thiebautii
MAG 03	6.77	14,701	638	5,213	34%	100%	4%	880	77%	1%	11%	0.4%	11%	1%	59%	23%	18%	Trichodesmium erythraeum
MAG 04	5.43	10,404	664	4,511	35%	99%	7%	161	12%	61%	21%	0.5%	5%	0%	10%	46%	43%	Ca. Trichodesmium miru
MAG 05	6.10	10,036	768	4,865	36%	96%	2%	244	73%	4%	2%	16%	5%	0%	43%	30%	27%	Ca. Trichodesmium nobis

Figure 2: The pangenome of *Trichodesmium*. Panel A displays genomic statistics for the five environmental *Trichodesmium* MAGs, along with their environmental signal among 937 *Tara* Oceans metagenomes. Panel B displays the *Trichodesmium* pangenome that covers 33,249 genes and 7,776 gene clusters from seven genomes of *Trichodesmium* corresponding to four distinct species. The top-right corner of panel B holds average nucleotide identity between genomes and a selection of RAST functional features decorate this pangenome, which was visualized with the anvi'o interactive interface[42]. Finally, genomes were organized based on the average nucleotide identity metric.

First and foremost, MAGs corresponding to the two candidate species lacked the entire nitrogen fixation gene apparatus, with the corresponding gene clusters occurring in the "*erythraeum/thiebautii*" pangenomic bin (Figure 2). This lack of signal was recapitulated across all functional annotations (Table S8) after removing few false positives easily identified by NCBI blast (Table S6). For instance, we found that COG20 functions incorrectly annotated as "Nitrogenase ATPase subunit NifH / coenzyme F430 biosynthesis subunit CfbC" a gene corresponding in reality to "ferredoxin : protochlorophyllide reductase". Just two *nifU* related genes were detected (with NCBI blast confirmation) in those MAGs, however their occurrence in non-diazotrophic lineages indicates they are not reliable markers for nitrogen fixation. More relevant nitrogen fixation gene markers (*nifHDK and nifENB*) were successfully detected in a wide range of newly identified marine diazotrophic MAGs spanning multiple phyla[34,35], suggesting our functional workflow could detect those markers in newly identified *Trichodesmium* species. It did not.

Perhaps more problematic, reconstructing genomes from metagenomes often suffers from quality and completion issues, which could explain the lack of *nif* genes in some *Trichodesmium* MAGs. As a first effort to address this, we investigated the occurrence of a comprehensive marine *nifH* gene database (see methods) across the *Tara* Oceans metagenomes, which includes *nifH* genes for T. *erythraeum* and *T. thiebautii* with >95% sequence identity. We used a mapping stringency sufficiently low (>80% sequence identity over 80% of the read length) to capture this genus in the nucleotidic sequence space. The *nifH* gene sequence has long been documented to display very low genetic diversity among *Trichodesmium* species[18,44,45]. Yet, we found that *Tara* Oceans metagenomes with high signal for 'Ca *Trichodesmium miru*' did not contain the expected signal for *Trichodesmium nifH* genes (Figure 3 and Table S3). For instance, at Station 133 in the Pacific Ocean (deep chlorophyll maximum layer, 180-2000 µm size fraction) where only 'Ca *Trichodesmium miru*' was detected with a mean coverage of 13.8X (i.e., its genome was sequenced nearly 14 times in this metagenome), not a single metagenomic read matched to the >750 nucleotides long *nifH* genes from *T. erythraeum* and *T. thiebautii*, or any of the other *nifH* genes in the database. Overall, the correlation between the coverages of *Trichodesmium* genomes and their *nifH* genes better correlated in those samples when excluding the two candidate species (R^2:0.99). 'Ca *Trichodesmium nobis*' displayed a higher niche overlap with *T. erythraeum* and *T. thiebautii*, nevertheless we could detect a clear discrepancy between its genomic occurrence and signal for *Trichodesmium nifH* genes in metagenomes of the Mediterranean Sea (Figure 3, Table S3). To expand this search for missing signal beyond the *nifH* gene, we then mapped metagenomic reads against the entire closed genome of *T. erythraeum* using an even lower mapping stringency (70% identity over 70% of the read length), revealing this time a lack of signal for the entire *nif* operon in samples dominated by the candidate species (Figure 4). In contrast, in samples that were dominated by *T. thiebautii*, the entire nif operon was perfectly recovered, except for the divergent intergenic regions. Thus, metagenomic signal for *nifH* and related genes were coherent with the comparative genomic insights, both pointing to a lack of the nitrogen fixation gene apparatus in the two closely related candidate species.



Tara Oceans filter ID	Oceanic region	Station ID	Oceanic layer	Size fraction	Genome-scale metagenomic signal				nifH gene signal	
					Ca. T. miru	Ca. T. nobis	T. thiebautii	T. erythraeum	Expected*	Observed
81SUR1QQSS11	Atlantic	81	SUR	20-180µm	43.6X	0X	2X	0X	45.6X	10 reads (1.3X)
133DCM1SSUU11	Pacific	133	DCM	180-2000µm	13.8X	0X	0X	0X	13.8X	no reads
81SUR01SSUU11	Atlantic	81	SUR	180-2000µm	12.3X	0X	0X	0X	12.3X	no reads
66SUR1SSUU11	Atlantic	66	SUR	180-2000µm	11.6X	1.7X	0X	0X	13.3X	no reads
68SUR1SSUU11	Atlantic	68	SUR	180-2000µm	7.2X	0.9X	0X	0X	8.1X	2 reads (0.3X)
141SUR1QQSS11	Atlantic	141	SUR	20-180µm	6.6X	13.8X	143.3X	11X	174.6X	1,048 reads (119X)
131SUR1SSUU11	Pacific	131	SUR	180-2000µm	4.8X	9.1X	59.2X	105.9X	178.9X	1,304 reads (153X)
41SUR1SSUU11	Indian	41	SUR	180-2000µm	3.8X	6.9X	64X	5.6X	80.3X	554 reads (70X)
68DCM1SSUU11	Atlantic	68	DCM	180-2000µm	3.8X	0X	0X	0X	3.8X	no reads
58DCM1QQSS11	Indian	58	SUR	20-180µm	3.5X	6.7X	71.6X	7.5X	89.4X	614 reads (77X)

R^2:0.95
Mean coverage (four Trichodesmium species)
200
180
160
140
120
100
80
60
40
20
0
Filter ID 81SUR1QQSS11
Correlation between 4 Trichodesmium species and nifH signal
0
20
40
60
80
100
120
140
160
180
Mean coverage for Trichodesmium nifH genes
R^2:0.99
Mean coverage (T. thiebautii and T. erythraeum)
200
180
160
140
120
100
80
60
40
20
0
Correlation between 2 Trichodesmium species and nifH signal
0
20
40
60
80
100
120
140
160
180
Mean coverage for Trichodesmium nifH genes
'Candidatus Trichodesmium nobis'

Tara Oceans filter ID	Oceanic region	Station ID	Oceanic layer	Size fraction	Genome-scale metagenomic signal				nifH gene signal	
					Ca. T. miru	Ca. T. nobis	T. thiebautii	T. erythraeum	Expected*	Observed
23SUR1SSUU11	MED	23	SUR	180-2000µm	0X	7.8X	0X	0X	7.8X	8 reads (0.9X)
9DCM1SSUU11	MED	9	DCM	180-2000µm	0X	5.2X	0X	0X	5.2X	6 reads 0.7X)
9SUR1SSUU11	MED	9	SUR	180-2000µm	0X	3.5X	0X	0X	3.5X	no reads
12SUR1QQSS11	MED	12	SUR	20-180µm	0X	2.1X	0X	0X	2.1X	2 reads (0.2X)
23DCM1MMQQ11	MED	23	DCM	5-20µm	0X	1.9X	0X	0X	1.9X	no reads
24SUR1SSUU11	MED	24	SUR	180-2000µm	0X	1.6X	0X	0X	1.6X	no reads
18SUR1SSUU11	MED	18	SUR	180-2000µm	0X	1.4X	0X	0X	1.4X	3 reads 0.3X)
23SUR1MMQQ11	MED	23	SUR	5-20µm	0X	1.4X	0X	0X	1.4X	no reads
23DCM1SSUU11	MED	23	DCM	180-2000µm	0X	1.4X	0X	0X	1.4X	no reads
12DCM1QQSS11	MED	12	DCM	20-180µm	0X	1.1X	0X	0X	1.1X	no reads

Figure 3: Metagenomic signal for *nifH* gene in the context of 'Ca *Trichodesmium miru*' and 'Ca *Trichodesmium nobis*'. Top panel summarizes metagenomic read recruitment results for five *Trichodesmium* MAGs (genome-scale with mapping stringency of >90% identity over >80% of the read length) and three known *Trichodesmium nifH* genes (gene-centric with mapping stringency of >80% identity over >80% of the read length) across ten *Tara* Oceans metagenomes with highest mean coverage for 'Ca *Trichodesmium miru*'. It also shows the correlation between mapped metagenomic reads for Trichodesmium *nifH* genes and either the cumulative mean coverage of all four Trichodesmium species (five MAGs) or the cumulative mean coverage of just *T. thiebautii* and *T. erythraeum* (three MAGs) across the same set of ten metagenomes. Bottom panel displays the correlation between 'Ca *Trichodesmium nobis*' and two *Trichodesmium* species across all *Tara* Oceans metagenomes. It also summarizes metagenomic read recruitment results for five *Trichodesmium* MAGs (genome-scale) and three known *Trichodesmium nifH* genes (gene-centric) across ten *Tara* Oceans metagenomes with highest mean coverage for 'Ca *Trichodesmium miru*' and no detection of the other MAGs (MED: Mediterranean Sea).

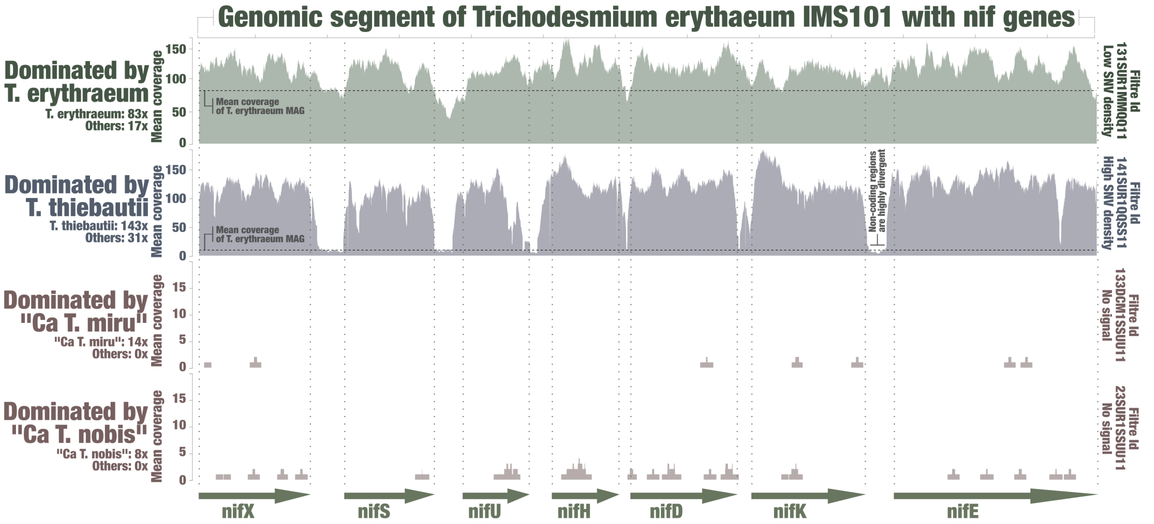


Figure 4: Metagenomic signal for the nif operon of *Trichodesmium erythraeum*. The figure displays read recruitment for the *T. erythraeum* IMS101 nif operon across four metagenomes, each dominated by a single *Trichodesmium* species (low mapping stringency of >70% identity over >70% of the read length; anvi'o visualization).

Hydrogen is a by-product of nitrogen fixation that cyanobacterial diazotrophs reutilize to gain energy[28–30]. We found that functions related to the recycling of hydrogen (*hyaABD* and *hypABCDE*,) were only missing in MAGs corresponding to the two candidate species (Figure 2B, top right panel; Table S8), which was also supported by a lack of metagenomic read recruitments for the *hyaABD* and *hypABCDE* genes of *T. erythraeum* in relevant metagenomes with low mapping stringency, echoing results for the *nif* operon (Figure S3). Non-diazotrophs in the surface ocean may have no reasons to contain genes for the recycling of hydrogen, however they do need to assimilate biologically available nitrogen molecules. We found the nitrite/nitrate transporter gene *nark* in single copy in *T. erythraeum* and *T. thiebautii* (COG20 functions), yet it occurred in two copies in 'Ca *Trichodesmium miru*' and in three copies in 'Ca *Trichodesmium nobis*'. This function provides a means to catalyze the uptake of nitrogen-enriched small molecules[46]. RAST annotation also detected an enrichment of this functional annotation in the candidate species. Furthermore, COG20 functions and RAST annotation identified the gene *nirB* allowing dissimilatory nitrate reduction (anaerobic respiration, nitrogen metabolism) in 'Ca *Trichodesmium miru*'. NCBI best blast hits linked this gene to the distantly related phyla Lentisphaera and Verrucomicrobia (including *Rubritalea marina* isolated from a sponge[47]), rather than more closely related lineages within Cyanobacteria. In addition, the gene appears next to a cyanobacterial transposase. Thus, we observed a lack of hydrogen recycling genes concomitant with the enrichment of genes related to nitrogen assimilation and metabolism in the candidate species, coherent with a non-diazotrophic life style.

Some cyanobacterial diazotrophs possess circadian rhythms that regulate nitrogen fixation and photosynthesis activities temporally and spatially[48–51]. While all four *Trichodesmium* species contained genes coding for the core circadian clock proteins (*kaiABC*) linked to photosynthesis regulation[51], the pangenomic analysis revealed two gene clusters corresponding to the circadian oscillating protein *COP23* and only missing in MAGs corresponding to the candidate species (Table S8). Furthermore, RAST annotation only identified one gene for *COP23* and one gene for Pex (another function linked to circadian rhythm), which were only detected in *T. erythraeum* and *T. thiebautii* (Figure 2B, top right panel). The cyanobacterial nitrogen fixation rhythm can be disconnected from light variations in laboratory experiments[23,52], however genes related to this mechanism have yet to be fully understood. Our results echo previous work on the *COP23* and *Pex* genes[53–56], suggesting that at least some of these genes play a central role in the circadian rhythm of nitrogen fixation in *T. erythraeum* and *T. thiebautii*.

Finally, we cross-validated these insights by performing six additional *Tara* Oceans genome-resolved metagenomic surveys guided by the known distribution of four *Trichodesmium* species and designed to only target either 'Ca *Trichodesmium miru*' or 'Ca *Trichodesmium nobis*' by means of single assemblies or small co-assemblies (Table S9). First, we used HMM models for the *nifHDK* and *nifENB* genes designed to cover the entire bacterial spectrum and found no trace of these gene markers in the six raw metagenomic assemblies, which contain contigs as short as 1,000nt (as a perspective, the *nifB* gene of *T. erythraeum* is 1,470 nt long). Then, we characterized and manually curated the *Trichodesmium* MAG of the targeted candidate species in each metagenomic assembly. Their completion ranged between 90.1% and 98.6%, with N50 as high as 26,010nt (Table S9). Their genetic content agreed with the previous lines of evidence as they were enriched in genes for nitrate/nitrite transporter *nark*, contained this time the gene *nirB* in both species (systematically linking this gene to Lentisphaera and Verrucomicrobia), and most importantly lacked genes related to the nitrogen fixation apparatus (except for the *nifU* related genes) and hydrogen recycling (*hyaABD* and *hypABCDE*) (Table S10). Thus, our lines of evidence for non-diazotrophic *Trichodesmium* species were reproducible using various metagenomic combinations and are supported by a lack of signal for *nifHDK* and *nifENB* genes in the raw metagenomic assemblies.

Discussion:

The association between *Trichodesmium* and nitrogen fixation has been deeply rooted in our minds due to considerable culture and fieldwork legacies, and as a result *Trichodesmium* is routinely referred to as a diazotrophic genus (e.g.,[14]). Here, we provide multiple lines of evidence indicating that the newly identified species 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*', characterized by means of genome-resolved metagenomics and found abundant in multiple regions of the open ocean, do not have the ability to fix nitrogen from the atmosphere. These species form a sister genomic clade most closely related to *T. tenue* and *T. contortum* based on metagenome-centric 16S rRNA gene consensus sequences (Figure 1). Near-complete environmental genomes for those candidate species denoted a lack of the entire nitrogen fixation gene apparatus and hydrogen recycling genes concomitant with the enrichment of genes related to nitrogen assimilation. These comparative genomic insights were supported by the absence of metagenomic signal for *Trichodesmium nifH* genes, the reference genome *T. erythraeum* IMS101, and targeted metagenomic assemblies and binning efforts, contrasting with the current paradigm that *Trichodesmium* species are necessarily capable of nitrogen fixation.

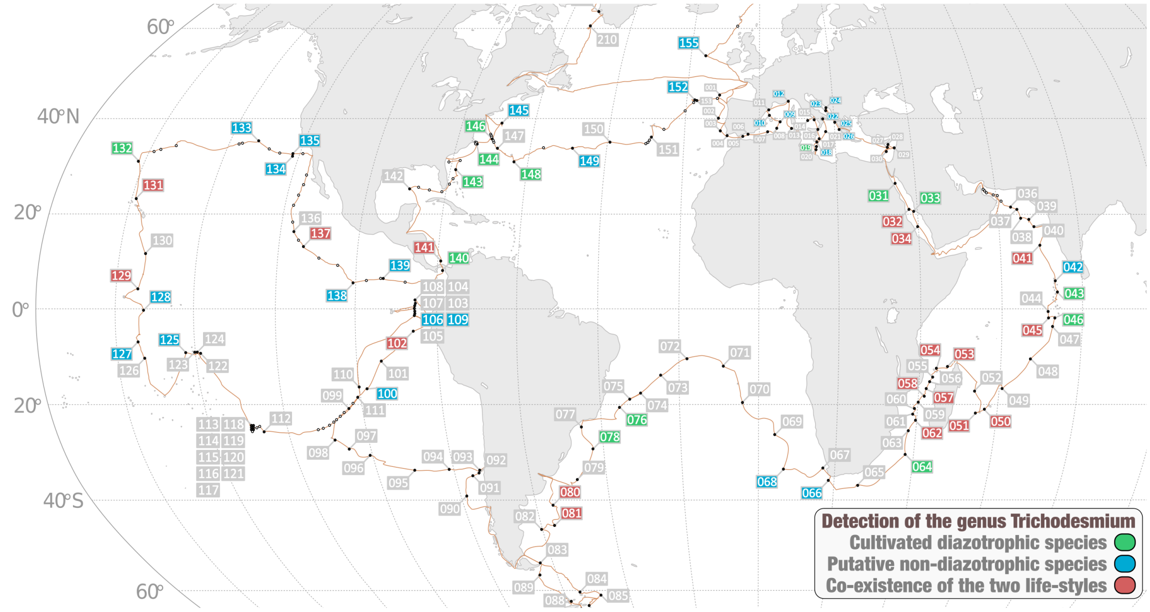


Figure 5: Detection of *Trichodesmium* species across the *Tara* Oceans stations. The world map describes stations in which we detected (1) only *T. erythraeum* and/or *T. thiebautii* (diazotrophic species), (2) only 'Ca. *Trichodesmium miru*' and/or 'Ca. *Trichodesmium nobis*' (putative non-diazotrophic species), (3) or species from both groups. This surveys covers all size fractions >0.2 µm.

When considering 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*' as non-diazotrophic species, their detection mostly restricted to large size fractions echoes trends for known *Trichodesmium* species and suggests filaments and colonies are the norm for this genus regardless of nitrogen fixation, albeit this has yet to be demonstrated. But most importantly, a critical consequence of this gene loss is that *Trichodesmium* might have a broader marine niche distribution compared to its nitrogen fixation capability. In fact, *T. erythraeum* and *T. thiebautii* remained undetected in 27 stations with clear signal for the putative non-diazotrophic *Trichodesmium* species (based on genome-wide metagenomic read recruitment, see Figures 5 and S4, Table S3). These stations cover the Indian, Atlantic and Pacific Oceans as well as the Mediterranean Sea, which indicates that 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*' are not only widespread but might also substantially expand the ecological niche of this genus. In our view, this stresses the need to disconnect taxonomic signal for *Trichodesmium* from the ability of a microbial community to fix nitrogen. Far from contesting the paramount importance of *T. erythraeum* and *T. thiebautii* for nitrogen fixation in the open ocean (in the context of dozens of other abundant cyanobacterial and heterotopic bacterial populations[35]), we merely suggest that differentiating diazotrophic from non-diazotrophic *Trichodesmium* populations might be needed to refine our understanding of the nitrogen balance in the oceans and seas. Counting *Trichodesmium* cells to survey the biomass of marine diazotrophs, as routinely performed for decades (e.g.,[57]), might lead to erroneous nitrogen fixation rate estimations, provided of course that non-diazotrophic *Trichodesmium* species indeed do exist and are not just a metagenomic mirage. In addition, the joint study of diazotrophic and non-diazotrophic *Trichodesmium* populations under culture conditions could bolster our understanding of the underlying genetic mechanisms for nitrogen fixation in *Trichodesmium* and beyond. As a humble step in this direction, it could be hypothesized that some circadian rhythm genes missing in 'Ca. *Trichodesmium miru*' and 'Ca. *Trichodesmium nobis*' but present in *T. erythraeum* and *T. thiebautii* are specifically linked to the regulation of nitrogen fixation, rather than photosynthesis, among *Trichodesmium* diazotrophs.

Non-diazotrophic *Trichodesmium* populations might have diverged from a diazotrophic lineage in order to fill distinct ecological niches at the surface of the open ocean (gene loss hypothesis), echoing within a genus results observed at the level of Cyanobacteria that revealed its complex evolutionary history[14,58,59] and suggested repeated losses of the *nif* genes within this phylum[60]. Interestingly, 'Ca. *Trichodesmium nobis*' strongly correlated with *T. thiebautii* (R^2=0.93), the most abundant of the four species within the scope of *Tara* Oceans metagenomes. The considerable overlap between these two species in the Indian Ocean especially, with one prevalent diazotroph and a less abundant one (ratio of about 1/15) lacking the ability to fix nitrogen fits well with the principles behind the black queen hypothesis[61]. This hypothesis states that gene loss can provide a selective advantage as long as the function is dispensable. 'Ca. *Trichodesmium nobis*' might have lost its ability to fix nitrogen in part because it could benefit from the nitrogen fixation of *T. thiebautii,* the "leaky helper" mentioned by Morris et al.[61]. Furthermore, the *nirB* gene linked to dissimilatory nitrate reduction in low-oxygen environments might provide a clue to understanding mechanisms behind this ecologically important evolutionary process, which based on our lines of evidence occurred in some *Trichodesmium* species but not in others. First, this gene was only detected in the two candidate species and is most closely related to Lentisphaera and Verrucomicrobia related genes. Second, *nirB* genes have been observed in the genomic content of non-cyanobacterial epibionts living at the surface of *Trichodesmium* colonies[62,63], suggesting a possible complementary functional role contributing to their interactions. The temporality of *nirB* gene acquisition and loss of nitrogen fixation and hydrogen recycling genes in the candidate species are currently unknown. Nevertheless, one could wonder if lateral gene transfers between *Trichodesmium* colonies and their epibionts have provided a non-diazotrophic evolutionary path for the common ancestor of 'Ca *Trichodesmium miru*' and 'Ca *Trichodesmium nobis*', opening exciting prospects regarding the ecology and evolution of this model planktonic lineage in the context of gene flows.

Conclusion:

The cyanobacterial genus *Trichodesmium* includes some of the most prominent marine nitrogen fixing species, which have been extensively studied in culture and the field for decades. Here, we explored the genome-resolved metagenomic content of two previously uncharacterized *Trichodesmium* species relatively abundant in the surface of oceans and seas. Critically, multiple lines of evidence point to the existence of non-diazotrophic *Trichodesmium* species with distinct ecological niches. Establishing that newly identified microbial populations lack critical functional traits from the sole perspective of environmental genomics could be perceived as a precarious endeavor. This is especially true when observations are not in line with those resulting from decades of cultivation. Yet, this approach comes with its own strengths, including reference-free de novo metagenomic assemblies on one hand and metagenomic read recruitments using reference genomes and gene markers as bait on the other. In the case of ‘Ca. *Trichodesmium miru*’ and ‘Ca. *Trichodesmium nobis*’, genome-resolved metagenomics and read recruitments applied to the considerable metagenomic legacy of *Tara* Oceans provided strong evidence that two closely related *Trichodesmium* species have lost the ability to fix nitrogen from the atmosphere. Culture representatives for these candidate species are needed to confront and move beyond those metagenomic insights. Until then, we are reminded that some long-established links between taxonomic lineages and functional traits supporting our understanding of the ocean microbiome might not always hold true.

Material and methods:

***Tara* Oceans metagenomes.** We analyzed a total of 937 *Tara Oceans* metagenomes available at the EBI under project PRJEB402 (https://www.ebi.ac.uk/ena/browser/view/PRJEB402). Table S1 reports general information (including the number of reads and environmental metadata) for each metagenome.

Biogeography of MAGs. We performed a mapping of all metagenomes to calculate the mean coverage and detection of MAGs. Briefly, we used BWA v0.7.15 (minimum identity of >90% over >80% of the read length) and a FASTA file containing the 1,888 non-redundant MAGs from Delmont et al.[35] to recruit short reads from all 937 metagenomes. We considered MAGs were detected in a given filter when >25% of their length was covered by reads to minimize non-specific read recruitments[34]. The number of recruited reads below this cut-off was set to 0 before determining vertical coverage and percent of recruited reads.

Metagenome-centric consensus 16S rRNA gene sequences. We performed a mapping of metagenomes against a reference 16S rRNA gene in order to retrieve corresponding signal for each of the four *Trichodesmium* species. Briefly, we used BWA v0.7.15 (minimum identity of >95% over >95% of the read length) and a FASTA file containing the 16S rRNA gene sequence T. *thiebautii* (AF013027) to recruit short reads from 12 metagenomes. We then used the anvi'o v.7 metagenomic workflow[42] to create a CONTIG database (the reference 16S rRNA gene sequence) and PROFILE databases for each metagenome, with the flag "--report-variability-full" in order to report every single nucleotide variations (full mode), or without it (default mode). After merging the PROFILE databases, we used the anvi'o interactive interface with inspection mode to visualize the coverage of this gene across metagenomes in the context of single nucleotide variants (default mode). We then used the program "anvi-gen-gene-consensus-sequences" with the flag "--contigs-mode" to generate metagenome-centric consensus 16S rRNA gene sequences (full mode).

Phylogenetic inferences using 16S rRNA gene sequences. We performed a phylogenetic analysis of the metagenome-centric consensus 16S rRNA gene sequences (see previous section) and reference 16S rRNA genes retrieved from NCBI. Briefly, we used the online platform GenomeNet (https://www.genome.jp/) to generate a phylogenetic tree at the nucleotide level and using as parameters the function "build" of ETE3 v3.1.1[64], MAFFT v6.861b with the linsi options[65] for alignment. Columns with more than ten percent of gaps were removed from the alignment using trimAl v1.4.rev6[66]. Finally, ML tree was inferred using PhyML v20160115 ran with model GTR and parameters "-o tlr --alpha e --bootstrap 100 --pinv e --nclasses 4 -f m"[67]. Branch supports were computed out of 100 bootstrapped trees. We used anvi'o to visualize the phylogenomic tree in the context of additional information.

Phylogenomic analysis of cyanobacterial genomes. We used PhyloSift[68] v1.0.1 with default parameters to infer associations between genomes in a phylogenomic context. Briefly, PhyloSift (1) identifies a set of 37 marker gene families in each genome, (2) concatenates the alignment of each marker gene family across genomes, and (3) computes a phylogenomic tree from the concatenated alignment using FastTree[69] v2.1. We used anvi'o to visualize the phylogenomic tree in the context of additional information.

Pangenomic analysis of MAGs. We used the anvi'o pangenomic workflow[38] to compute and visualize the pangenome of *Trichodesmium*. Briefly, the workflow consisted of three main steps: (1) we generated an anvi'o genome database to store DNA and amino acid sequences, as well as functional annotations of each gene in the seven *Trichodesmium* genomes under consideration, (2) we computed the *Trichodesmium* pangenome from a genome database by identifying 'gene clusters', and (3) we displayed the pangenome to visualize the distribution of gene clusters across genomes. The gene clusters represent sequences of one or more predicted open reading frames grouped together based on their homology at the translated DNA sequence level. To compute the *Trichodesmium* pangenome, we used the program'anvi-pan-genome'with the flag'--use-ncbi-blast'and default parameters. This program (1) calculates similarities of each amino acid sequence in every genome against every other amino acid sequence using blastp[70], (2) removes weak hits using the 'minbit heuristic', which was originally described in ITEP[71], (3) uses the MCL algorithm[72] to identify gene clusters in the remaining blastp search results, (4) computes the occurrence of gene clusters across genomes and the total number of genes they contain, (5) performs hierarchical clustering analyses for gene clusters (based on their distribution across genomes) and for genomes (based on gene clusters they share) using Euclidean distance and Ward clustering by default, and finally (6) generates an anvi'o pan database that stores all results for downstream analyses and was used to visualize the *Trichodesmium* pangenome in the interactive interface.

Functional inferences of MAGs. We inferred functions among the *Trichodesmium* genes using (1) Pfam[37] from within the anvi'o pangenomic workflow, (2) COG20 functions, categories and pathways[39], KOfam[40], KEGG modules and classes[41] within the anvi'o genomic workflow[42] and (3) the RAST online platform[43]. Regarding the KEGG modules, we calculated their level of completeness in each genomic database using the anvi'o program "anvi-estimate-metabolism" with default parameters. The URL https://merenlab.org/m/anvi-estimate-metabolism describes this program in more detail. Lastly, we used online NCBI blasts to identify false positives regarding specific nitrogen fixation gene markers. False positives correspond to functional annotations for which NCBI blast identified a different function with lower e-values and bit scores.

Metagenomic signal for the extended nifH gene database. We performed a mapping of metagenomes to calculate the mapped reads and mean coverage of sequences in an extended *nifH* gene database (see[35] for more details). Briefly, we used BWA v0.7.15 (minimum identity of 80%) and a FASTA file containing the sequences to recruit short reads from 937 *Tara* Oceans metagenomes.

Metagenomic signal for the reference *Trichodesmium erythraeum* IMS101. We performed a mapping of metagenomes against the reference genome *Trichodesmium erythraeum* IMS101 in order to retrieve corresponding signal for each of the four *Trichodesmium* species. Briefly, we used BWA v0.7.15 (minimum identity of >70% over >70% of the read length) and a FASTA file containing the entire genome *Trichodesmium erythraeum* IMS101 to recruit short reads from 12 metagenomes. We then used the anvi'o metagenomic workflow to create a CONTIG database (the genome) and PROFILE databases for each metagenome. After merging the PROFILE databases, we used the anvi'o interactive interface with inspection mode to visualize the coverage of regions of interest (e.g., the nif operon) across metagenomes.

Reproducing the recovery of MAGs for the candidate species. As a supplement of the initial large-scale genome-resolved metagenomic survey[35,36], we also performed additional genome-resolved metagenomic surveys by taking advantage of our knowledge of their distribution across the *Tara* Oceans metagenomes. Briefly, we used metagenomic reads as inputs for six metagenomic single assemblies or co-assemblies using MEGAHIT[73] v1.1.1, and simplified the scaffold header names in the resulting assembly outputs using anvi'o[42]. We then completed the anvi'o manual binning workflow to extract *Trichodesmium* MAGs from these assemblies. For each targeted species, we used a distinct set of 8 relevant metagenomes for mapping in order to effectively use differential coverage for binning. Finally, we studied the functional repertoire of recovered *Trichodesmium* MAGs using anvi'o, as described in the "Functional inferences of MAGs" section.

Search for *nifHDK* and *nifENB* genes in raw assemblies. We used HMM models designed to cover the entire bacterial spectrum of *nifHDK* and *nifENB* genes to search for these gene markers in raw metagenomic assemblies performed to extract the genomic content of the candidate species. The e-value cut-off was set to e-100.

Supplemental information. We provided a supplemental information document describing the anvi'o workflows used in the study. Each section of the document includes the list of anvi'o programs and a breif explanation of the workflow.

Data availability. First, the genomic resource (bacterial and archaeal MAGs from the surface of the oceans) is publicly available at http://www.genoscope.cns.fr/tara/. The link provides access to the 11 raw metagenomic co-assemblies from Delmont et al.[36] as well as the FASTA file for 1,888 MAGs, including the five corresponding to *Trichodesmium*. In addition, the doi https://doi.org/10.6084/m9.figshare.14207321 provides access to (0) FASTA files for the initial five *Trichodesmium* MAGs, (1) anvi'o files corresponding to the metagenomic read recruitments used to generate 16S rRNA metagenomic consensus sequences, (2) anvi'o CONTIGS databases for reference *Trichodesmium* MAGs and isolate genomes (including functional annotations), (3) anvi'o files corresponding to the *Trichodesmium* pangenome, (4) anvi'o files corresponding to the metagenomic read recruitments for reference genome *Trichodesmium* erythraeum IMS101 (different mapping stringencies included), (5) targeted genome-resolved metagenomic surveys for the candidate *Trichodesmium* species (contigs >2,5kbp, anvi'o files and summaries), (6) anvi'o CONTIGS databases for the *Trichodesmium* MAGs extracted from the targeted genome-resolved metagenomic surveys (including functional annotations), (7) HMM models for six nitrogen fixation gene markers, (8) the supplemental tables and information, (9) and finally the raw assemblies (contigs >1kbp) for targeted genome-resolved metagenomic surveys.

Acknowledgments

I would like to thank Michael D. Lee, A. Murat Eren, Eric Pelletier and Jed A. Fuhrman for technical support, editorial suggestions and constructive discussions regarding *Trichodesmium* and nitrogen fixation in the sunlit ocean. I also would like to thank Paul Frémont regarding world map production for the relative distribution of *Trichodesmium* species. In addition, this survey was especially made possible by two scientific endeavors: the sampling and sequencing efforts by the *Tara* Oceans consortium, and the bioinformatics and visualization capabilities afforded by anvi'o. As a result, I would like to thank everyone that contributed to *Tara* Oceans and anvi'o over the years[74,75]. In addition, *Tara* Oceans (which includes the *Tara* Oceans and *Tara* Oceans Polar Circle expeditions) would not exist without the leadership of the *Tara* Oceans Foundation and the continuous support of 23 institutes (https://oceans.taraexpeditions.org/). Finally, part of the computation was performed using the platine, titane and curie HPC machine provided through GENCI grants (t2011076389, t2012076389, t2013036389, t2014036389, t2015036389 and t2016036389).

Supplemental figure

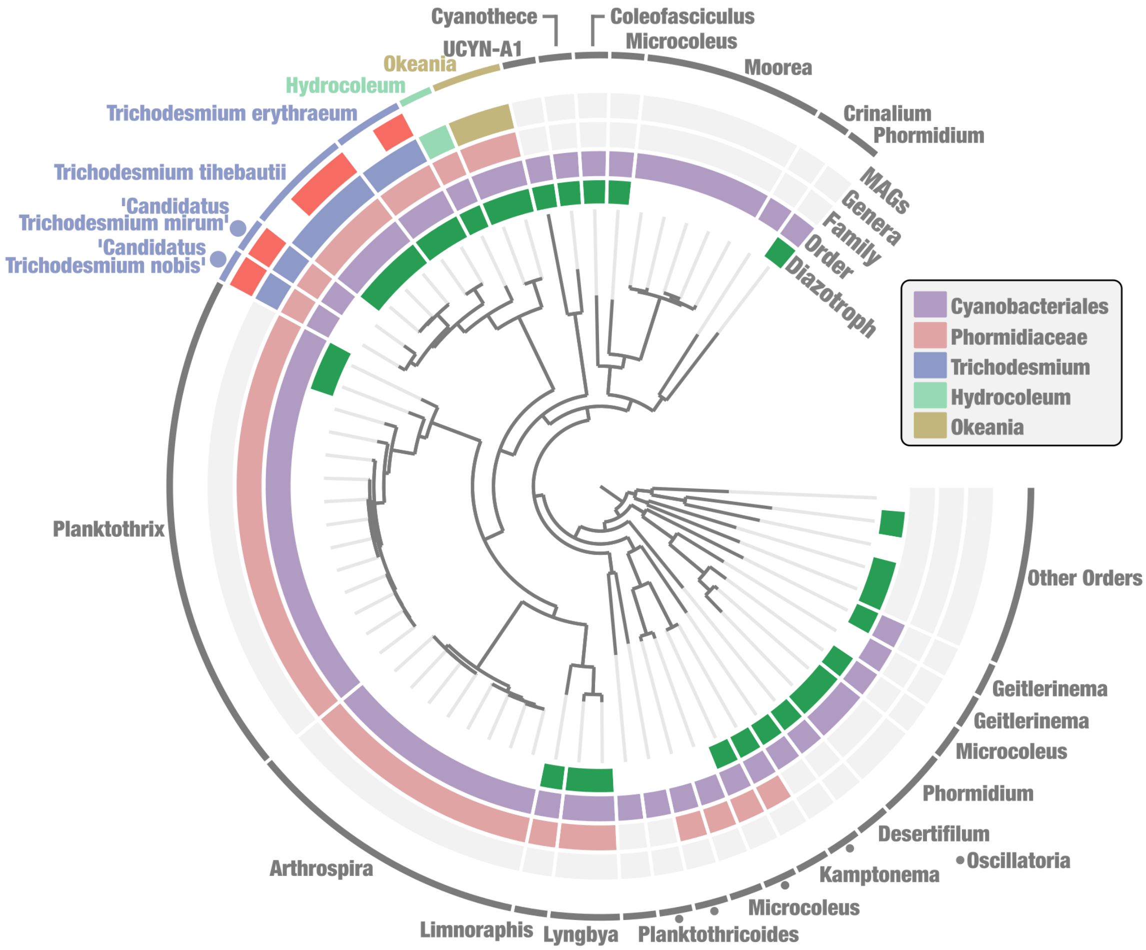


Figure S01: Phylogenomic analysis of cyanobacterial lineages closely related to *Trichodesmium*. This phylogenomic analysis includes xx reference genomes affiliated to Cyanobacteriales, including the five *Trichodesmium* population genomes characterized in this study ('MAGs' layer). The tree is rooted using five additional genomes affiliated to closely related orders within Cyanobacteria. Genomes containing nitrogen fixation genes are highlighted ('diazotroph' layer).

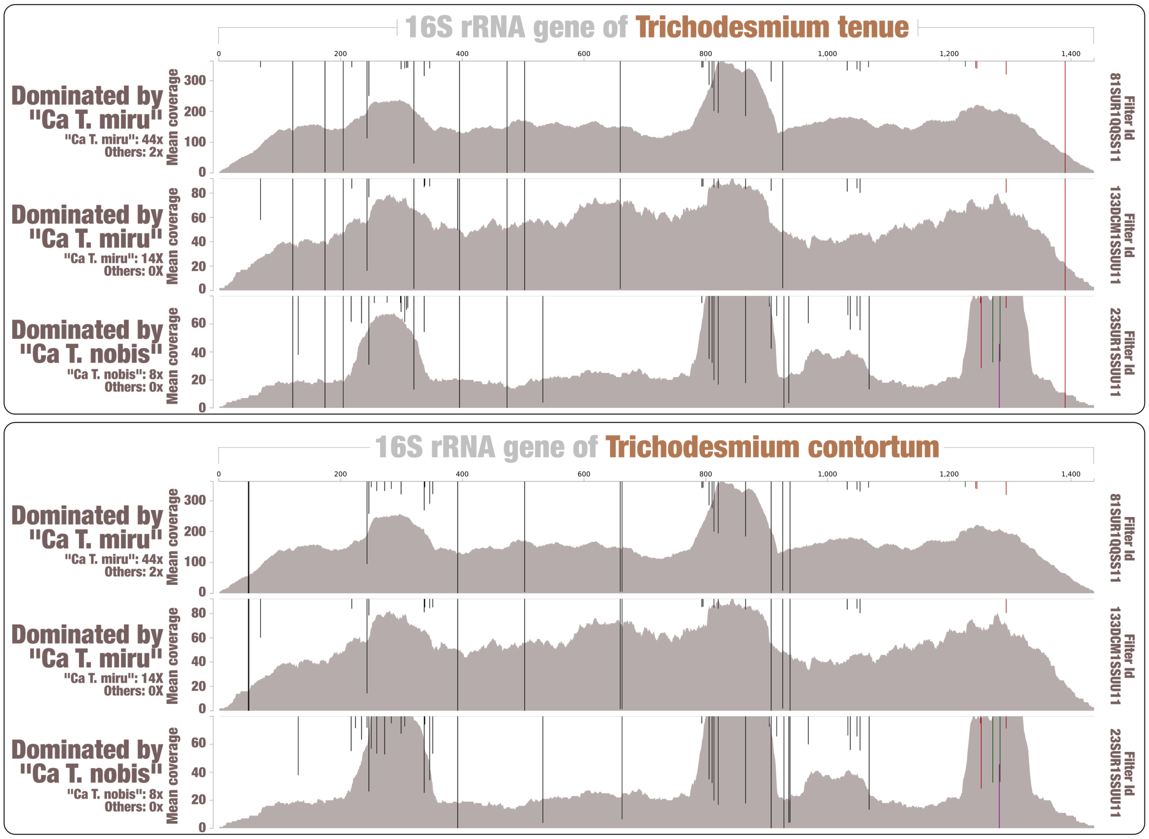


Figure S02: 16S rRNA gene recruitment using Trichodesmium tenue and Trtichodesmium contortum as references. The figure displays read recruitment for the *T. tenue* and *T. contortum* 16S rRNA gene across three metagenomes dominated by candidate species (mapping stringency of >95% identity over >95% of the read length). Single nucleotide variants identified by anvi'o (when using default parameters) were visualized with an amplitude reaching 100% when all reads contained the same nucleotide type.

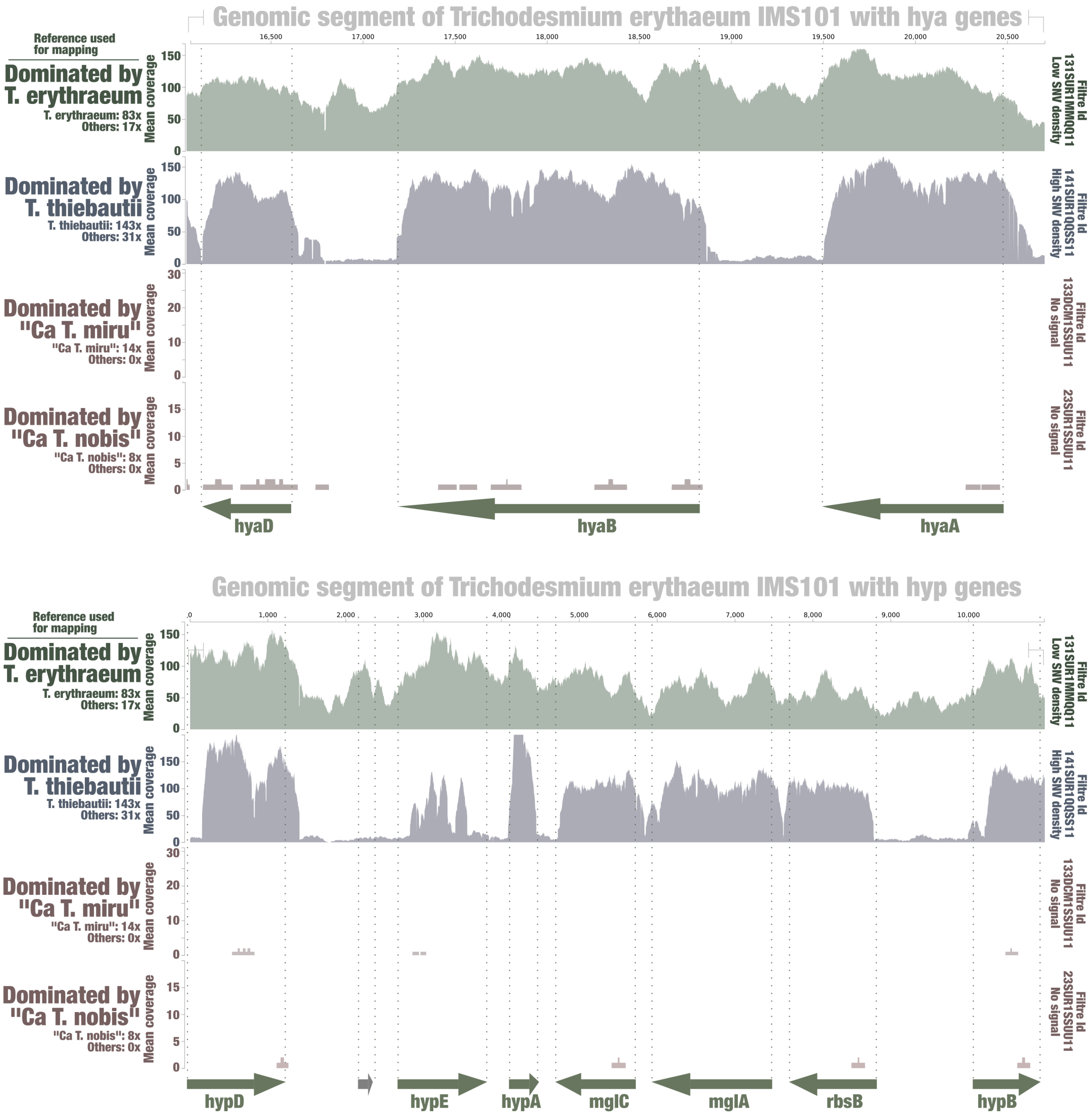


Figure S03: Metagenomic signal for the hya and hyp operons for hydrogen recycling of *Trichodesmium erythraeum*. The figure displays read recruitment for the *T. erythraeum* IMS101 hya and hyp operons across four metagenomes, each dominated by a single *Trichodesmium* species (low mapping stringency of >70% identity over >70% of the read length; anvi'o visualization).

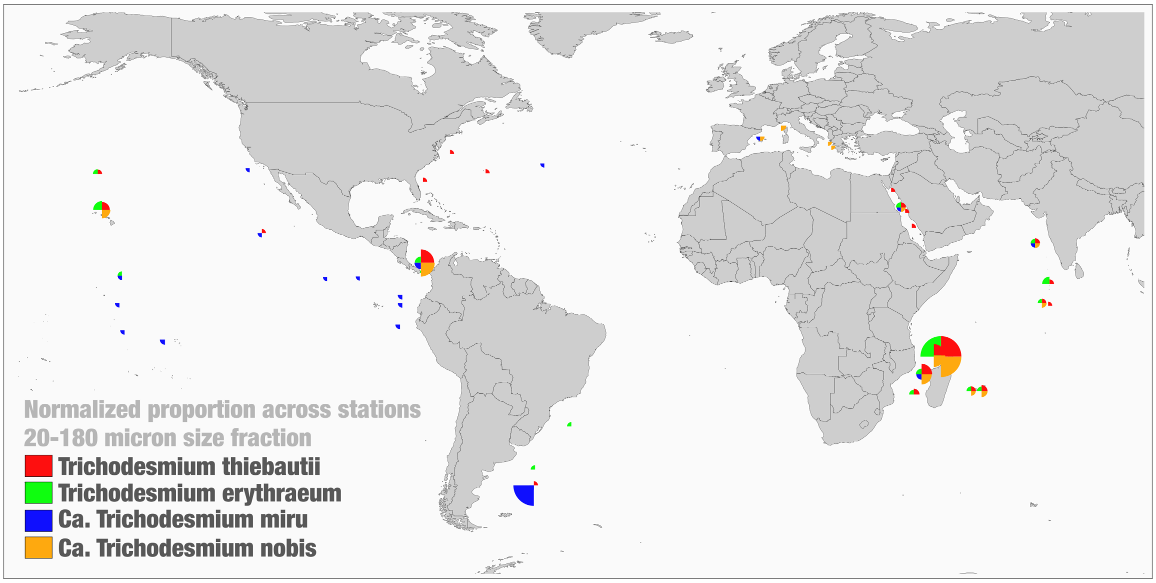

Normalized proportion across stations
20-180 micron size fraction
🟥 Trichodesmium thiebautii
🟩 Trichodesmium erythraeum
🟦 Ca. Trichodesmium miru
🟧 Ca. Trichodesmium nobis

Figure S04: Normalized relative distribution of four *Trichodesmium* species. The world map displays the relative distribution of four *Trichodesmium* species across *Tara* Oceans metagenomes corresponding to the 20-180 microns size fraction (normalized so that each species has a full quarter of circle when it reaches its maximal value). The figure aims at providing some perspective regarding the occurrence of key marine *Trichodesmium* species in the surface of the open oceans and seas.

Supplemental Tables

Table S01: Statistics for 937 *Tara* Oceans metagenomes organized by depth, size fraction and oceanic region. The table also contains environmental conditions across *Tara* Oceans stations.

Table S02: Statistics for the 1,888 bacterial and archaeal MAGs. The table contains genomic statistics (e.g., completion and length), taxonomic information, general mapping trends such as the cosmopolitan score, as well as additional information regarding the five *Trichodesmium* MAGs (e.g., ANI comparisons).

Table S03: Metagenomic read recruitment statistics for the five *Trichodesmium* MAGs (90% identity) and reference *nifH* genes (80% identity). The table contains mean genomic coverage values across the 937 *Tara* Oceans metagenomes.

Table S04: Metagenomic set used to generate the 12 metagenome-centric consensus sequences. The table describes the mean coverage of each *Trichodesmium* MAG and also contains reference and consensus 16S rRNA gene sequences.

Table S05: The pangenome of *Trichodesmium*. The table contains the anvi'o pangenomic summary outcome of seven *Trichodesmium* genomes (MAGs and isolates), connecting genes to gene clusters, genomes, bins and Pfam functional annotations.

Table S06: Functional annotation of seven *Trichodesmium* genomes (MAGs and isolates) using COG20 functions, categories and pathways, KOfam, KEGG modules and classes processed within anvi'o.

Table S07: Functional annotation of seven *Trichodesmium* genomes (MAGs and isolates) using the online RAST annotation platform.

Table S08: Summary of the lines of evidence across functional annotations for seven *Trichodesmium* genomes (MAGs and isolates) with respect to nitrogen fixation.

Table S09: Additional genome-resolved metagenomic surveys to recover MAGs from the two candidate *Trichodesmium* species. The table describes the sets of metagenomes used for assembly, along with the sets of metagenomes used for mapping.

Table S10: Functional annotation of redundant *Trichodesmium* MAGs for 'Ca *Trichodesmium miru*' (n=5) and 'Ca *Trichodesmium nobis*' (n=1) using using COG20 functions, categories and pathways, KOfam, KEGG modules and classes processed within anvi'o.

References

1. Sanders, R. *et al.* The Biological Carbon Pump in the North Atlantic. *Prog. Oceanogr.* (2014). doi:10.1016/j.pocean.2014.05.005
2. Boyd, P. W. Toward quantifying the response of the oceans' biological pump to climate change. *Front. Mar. Sci.* (2015). doi:10.3389/fmars.2015.00077
3. Charlson, R. J., Lovelock, J. E., Andreae, M. O. & Warren, S. G. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature* **326**, 655–661 (1987).
4. Falkowski, P. G., Barber, R. T. & Smetacek, V. Biogeochemical controls and feedbacks on ocean primary production. *Science (80-.).* **281**, 200–206 (1998).
5. Arrigo, K. R. Marine microorganisms and global nutrient cycles. *Nature* **437**, 349–355 (2005).
6. De Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science (80-.).* (2015). doi:10.1126/science.1261605
7. Moore, C. M. *et al.* Processes and patterns of oceanic nutrient limitation. *Nat. Geosci* **6**, 701–710 (2013).
8. Tyrrell, T. The relative influences of nitrogen and phosohorus on oceanic primary production. *Nature* **400**, 525–531 (1999).
9. Zehr, J. P. & Capone, D. G. Changing perspectives in marine nitrogen fixation. *Science* (2020). doi:10.1126/science.aay9514
10. Zehr, J. P., Jenkins, B. D., Short, S. M. & Steward, G. F. Nitrogenase gene diversity and microbial community structure: a cross-system comparison. *Env. Microbiol* **5**, 539–554 (2003).

11. Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C. & Dixon, R. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* **13**, 162 (2012).
12. Ehrenberg, C. G. Neue Beobachlungen über blutartige Erscheinungen in Aegypten, Arabien und Sibirien, nebst einer Uebersicht und Kritik der früher bekannnten. *Ann. Phys.* (1830). doi:10.1002/andp.18300940402
13. Capone, D. G. Trichodesmium, a Globally Significant Marine Cyanobacterium. *Science (80-.).* **276**, 1221–1229 (1997).
14. Bergman, B., Sandh, G., Lin, S., Larsson, J. & Carpenter, E. J. Trichodesmium - a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiology Reviews* (2013). doi:10.1111/j.1574-6976.2012.00352.x
15. Karl, D. *et al.* The role of nitrogen fixation in biogeochemical cycling in the subtropical North Pacific Ocean. *Nature* **388**, 533–538 (1997).
16. Dore, J. E., Brum, J. R., Tupas, L. M. & Karl, D. M. Seasonal and interannual variability in sources of nitrogen supporting export in the oligotrophic subtropical North Pacific Ocean. *Limnol. Oceanogr.* (2002). doi:10.4319/lo.2002.47.6.1595
17. Orcutt, K. M. *et al.* Characterization of Trichodesmium spp. by genetic techniques. *Appl. Environ. Microbiol.* (2002). doi:10.1128/AEM.68.5.2236-2245.2002
18. Janson, S., Bergman, B., Carpenter, E. J., Giovannoni, S. J. & Vergin, K. Genetic analysis of natural populations of the marine diazotrophic cyanobacterium Trichodesmium. *FEMS Microbiol. Ecol.* (1999). doi:10.1016/S0168-6496(99)00041-0
19. McKinna, L. I. W. Three decades of ocean-color remote-sensing Trichodesmium spp. in the World's oceans: A review. *Progress in Oceanography* (2015). doi:10.1016/j.pocean.2014.12.013
20. Dyhrman, S. T. *et al.* Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*. *Nature* **439**, 68–71 (2006).
21. Pierella Karlusich, J. J. *et al.* Global distribution patterns of marine nitrogen-fixers by imaging and molecular methods. *bioRxiv* (2020). doi:10.1101/2020.10.17.343731
22. Klawonn, I. *et al.* Distinct nitrogen cycling and steep chemical gradients in Trichodesmium colonies. *ISME J.* (2020). doi:10.1038/s41396-019-0514-9
23. Chen, Y. B., Dominic, B., Mellon, M. T. & Zehr, J. P. Circadian rhythm of nitrogenase gene expression in the diazotrophic filamentous nonheterocystous cyanobacterium Trichodesmium sp. strain IMS 101. *J. Bacteriol.* (1998). doi:10.1128/jb.180.14.3598-3605.1998
24. Rodriguez, I. B. & Ho, T. Y. Diel nitrogen fixation pattern of Trichodesmium: The interactive control of light and Ni. *Sci. Rep.* (2014). doi:10.1038/srep04445
25. Frischkorn, K. R., Haley, S. T. & Dyhrman, S. T. Coordinated gene expression between Trichodesmium and its microbiome over day-night cycles in the North Pacific Subtropical Gyre. *ISME J.* (2018). doi:10.1038/s41396-017-0041-5

26. Lee, M. D. *et al.* The Trichodesmium consortium: Conserved heterotrophic co-occurrence and genomic signatures of potential interactions. *ISME J.* (2017). doi:10.1038/ismej.2017.49
27. Frischkorn, K. R., Rouco, M., Van Mooy, B. A. S. & Dyhrman, S. T. The Trichodesmium microbiome can modulate host N 2 fixation . *Limnol. Oceanogr. Lett.* (2018). doi:10.1002/lol2.10092
28. Wilson, S. T., Foster, R. A., Zehr, J. P. & Karl, D. M. Hydrogen production by Trichodesmium erythraeum Cyanothece sp. and Crocosphaera watsonii. *Aquat. Microb. Ecol.* (2010). doi:10.3354/ame01407
29. Wilson, S. T., Kolber, Z. S., Tozzi, S., Zehr, J. P. & Karl, D. M. Nitrogen fixation, hydrogen cycling, and electron transport kinetics in trichodesmium erythraeum (cyanobacteria) strain ims101. *J. Phycol.* (2012). doi:10.1111/j.1529-8817.2012.01166.x
30. Eichner, M., Basu, S., Gledhill, M., De Beer, D. & Shaked, Y. Hydrogen dynamics in trichodesmiumcolonies and their potential role in mineral iron acquisition. *Front. Microbiol.* (2019). doi:10.3389/fmicb.2019.01565
31. Parks, D. H. *et al.* Author Correction: Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology* 1 (2017). doi:10.1038/s41564-017-0083-5
32. Tully, B. J., Graham, E. D. & Heidelberg, J. F. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci. Data* **5**, 1–8 (2018).
33. Pachiadaki, M. G. *et al.* Charting the Complexity of the Marine Microbiome through Single-Cell Genomics. *Cell* (2019). doi:10.1016/j.cell.2019.11.017
34. Delmont, T. O. *et al.* Nitrogen-fixing populations of Planctomycetes and Proteobacteria are abundant in surface ocean metagenomes. *Nat. Microbiol.* 1 (2018). doi:10.1038/s41564-018-0176-9
35. Delmont, T. O. & et al. Heterotrophic bacterial diazotrophs are more abundant than their cyanobacterial counterparts in metagenomes covering most of the sunlit ocean. *Preprint* (2021).
36. Delmont, T. O. *et al.* Functional repertoire convergence of distantly related eukaryotic plankton lineages revealed by genome-resolved metagenomics. *bioRxiv* 2020.10.15.341214 (2020). doi:10.1101/2020.10.15.341214
37. Bateman, A. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **28**, 263–266 (2000).
38. Delmont, T. O. & Eren, A. M. Linking pangenomes and metagenomes: The Prochlorococcus metapangenome. *PeerJ* **2018**, (2018).
39. Galperin, M. Y. *et al.* COG database update: focus on microbial diversity, model organisms, and widespread pathogens. *Nucleic Acids Res.* (2021). doi:10.1093/nar/gkaa1018
40. Aramaki, T. *et al.* KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* (2020). doi:10.1093/bioinformatics/btz859
41. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: New perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* (2017). doi:10.1093/nar/gkw1092

42. Eren, A. M. *et al.* Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**, e1319 (2015).
43. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).
44. Ben-Porath, J., Carpenter, E. J. & Zehr, J. P. GENOTYPIC RELATIONSHIPS IN TRICHODESMIUM (CYANOPHYCEAE) BASED ON nifH SEQUENCE COMPARISONS. *J. Phycol.* (1993). doi:10.1111/j.0022-3646.1993.00806.x
45. Zehr, J. P., Limberger, R. J., Ohki, K. & Fujita, Y. Antiserum to nitrogenase generated from an amplified DNA fragment from natural populations of Trichodesmium spp. *Appl. Environ. Microbiol.* (1990). doi:10.1128/aem.56.11.3527-3531.1990
46. Kolesnikow, T., Schroder, I. & Gunsalus, R. P. Regulation of narK gene expression in Escherichia coli in response to anaerobiosis, nitrate, iron, and molybdenum. *J. Bacteriol.* (1992). doi:10.1128/jb.174.22.7104-7111.1992
47. Scheuermayer, M., Gulder, T. A. M., Bringmann, G. & Hentschel, U. Rubritalea marina gen. nov., sp. nov., a marine representative of the phylum 'Verrucomicrobia', isolated from a sponge (porifera). *Int. J. Syst. Evol. Microbiol.* (2006). doi:10.1099/ijs.0.64360-0
48. Cohen, S. E. & Golden, S. S. Circadian Rhythms in Cyanobacteria. *Microbiol. Mol. Biol. Rev.* (2015). doi:10.1128/mmbr.00036-15
49. Bell-Pedersen, D. *et al.* Circadian rhythms from multiple oscillators: Lessons from diverse organisms. *Nature Reviews Genetics* (2005). doi:10.1038/nrg1633
50. Reimers, A. M., Knoop, H., Bockmayr, A. & Steuer, R. Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. *Proc. Natl. Acad. Sci. U. S. A.* (2017). doi:10.1073/pnas.1617508114
51. Kondo, T. & Ishiura, M. The circadian clock of cyanobacteria. *BioEssays* (2000). doi:10.1002/(SICI)1521-1878(200001)22:1<10::AID-BIES4>3.0.CO;2-A
52. Gaudana, S. B. *et al.* Rhythmic and sustained oscillations in metabolism and gene expression of Cyanothece sp. ATCC 51142 under constant light. *Front. Microbiol.* (2013). doi:10.3389/fmicb.2013.00374
53. Muñoz-Marín, M. D. C. *et al.* The transcriptional cycle is suited to daytime N 2 fixation in the unicellular cyanobacterium "candidatus atelocyanobacterium thalassa" (UCYN-A). *MBio* (2019). doi:10.1128/mBio.02495-18
54. Takai, N., Ikeuchi, S., Manabe, K. & Kutsuna, S. Expression of the circadian clock-related gene pex in Cyanobacteria increases in darkness and is required to delay the clock. *J. Biol. Rhythms* (2006). doi:10.1177/0748730406289400
55. Kutsuna, S., Kondo, T., Aoki, S. & Ishiura, M. A period-extender gene, pex, that extends the period of the circadian clock in the cyanobacterium Synechococcus sp. strain PCC 7942. *J. Bacteriol.* (1998). doi:10.1128/jb.180.8.2167-2174.1998
56. Chen, H. M., Chien, C. Y. & Huang, T. C. Regulation and molecular structure of a circadian oscillating protein located in the cell membrane of the prokaryote Synechococcus RF- 1. *Planta* (1996). doi:10.1007/BF00195182
57. Luo, Y. W. *et al.* Database of diazotrophs in global ocean: Abundance, biomass and nitrogen fixation rates. *Earth Syst. Sci. Data* (2012). doi:10.5194/essd-4-

47-2012

58. Tomitani, A., Knoll, A. H., Cavanaugh, C. M. & Ohno, T. The evolutionary diversification of cyanobacteria: Molecular-phylogenetic and paleontological perspectives. *Proc. Natl. Acad. Sci. U. S. A.* (2006). doi:10.1073/pnas.0600999103
59. Bothe, H., Schmitz, O., Yates, M. G. & Newton, W. E. Nitrogen Fixation and Hydrogen Metabolism in Cyanobacteria. *Microbiol. Mol. Biol. Rev.* (2010). doi:10.1128/mmbr.00033-10
60. Latysheva, N., Junker, V. L., Palmer, W. J., Codd, G. A. & Barker, D. The evolution of nitrogen fixation in cyanobacteria. *Bioinformatics* (2012). doi:10.1093/bioinformatics/bts008
61. Morris, J. J., Lenski, R. E. & Zinser, E. R. The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *MBio* (2012). doi:10.1128/mBio.00036-12
62. Frischkorn, K. R., Rouco, M., Van Mooy, B. A. S. & Dyhrman, S. T. Epibionts dominate metabolic functional potential of Trichodesmium colonies from the oligotrophic ocean. *ISME J.* (2017). doi:10.1038/ismej.2017.74
63. Lee, M. D. *et al.* Transcriptional activities of the microbial consortium living with the marine nitrogenfixing cyanobacterium Trichodesmium reveal potential roles in community-level nitrogen cycling. *Appl. Environ. Microbiol.* (2018). doi:10.1128/AEM.02026-17
64. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* (2016). doi:10.1093/molbev/msw046
65. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* (2013). doi:10.1093/molbev/mst010
66. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp348
67. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
68. Darling, A. E. *et al.* PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ* **2**, e243 (2014).
69. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 - Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).
70. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
71. Benedict, M. N., Henriksen, J. R., Metcalf, W. W., Whitaker, R. J. & Price, N. D. ITEP: an integrated toolkit for exploration of microbial pan-genomes. *BMC Genomics* **15**, 8 (2014).
72. Van Dongen, S. & Abreu-Goodger, C. Using MCL to extract clusters from networks. *Methods Mol. Biol.* **804**, 281–295 (2012).
73. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via

succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2014).

74. Eren, A. M. *et al.* Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology* (2021). doi:10.1038/s41564-020-00834-3

75. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* 1–18 (2020). doi:10.1038/s41579-020-0364-5