Molecular Biology and Evolution, 2025, **42**, 1–15
https://doi.org/10.1093/molbev/msaf181
Advance access publication 6 August 2025
Discoveries

MBE OXFORD

Selection Maintains Photosynthesis in a Symbiotic Cyanobacterium Despite Redundancy With its Fern Host

Liam Friar [1],* **Kyle Keepers** [2] **Arkadiy I. Garber** [3] **John P. McCutcheon** [3,4] **Boswell Wing** [1] **Nolan C. Kane** [2],*

[1]Department of Geological Sciences, University of Colorado, Boulder, CO, USA
[2]Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO, USA
[3]Biodesign Institute and School of Life Sciences, Arizona State University, Tempe, AZ, USA
[4]Howard Hughes Medical Institute, Chevy Chase, MD, USA

***Corresponding authors:** E-mails: liamfriar@gmail.com; nckane@gmail.com.
Associate editor: Aurelien Tellier

Abstract

Vertically inherited symbionts experience different physical, chemical, and population genetic environments than free-living organisms. As a result, they can experience long-term reductions in effective population size (N_e) and weaker purifying selection on genes that are less important in the host-associated environment. Over time, these forces result in gene loss. A comparative genomic approach using independently evolved symbiotic bacteria and free-living relatives can reveal which genes are important in the symbiotic state. We apply this approach to understand why some diazotrophic cyanobacteria evolving as vertically inherited symbionts of photosynthetic eukaryotic hosts have lost their ancestral capacity for photosynthesis while others have retained that capacity. We look specifically at *Trichormus azollae*, a diazotrophic cyanobacterium that remains photosynthetic after 50 to 90 Ma as a vertically inherited symbiont of *Azolla* ferns. We show that gene loss is ongoing, with different genes lost across the eight *T. azollae* strains examined. We apply molecular evolutionary models to genomes of *T. azollae* and free-living relatives, finding genome-wide signatures of drift in *T. azollae* consistent with long-term reductions in N_e. Ribosomal proteins and proteins from the energy-capturing photosynthetic light reactions are under stronger purifying selection than genes from other pathways, including nitrogen fixation and photosynthetic carbon fixation. Strong purifying selection is expected for the ribosome given its extraordinary levels of conservation, even in ancient vertically inherited symbionts. That genes in the light reactions are under strong purifying selection and never lost in any strain suggests that energy capture, likely required for energy-intensive nitrogen fixation, remains important to this symbiont.

Keywords: molecular evolution, *Trichormus azollae*, symbiotic genome reduction, comparative genomics, genetic drift, *Azolla*.

Introduction

Symbioses span the tree of life (Rosenberg and Zilber-Rosenberg 2011; McCutcheon et al. 2024). They fill key ecological niches, including as the main sources of bioavailable nitrogen in many ecosystems (Davies-Barnard and Friedlingstein 2020; Zehr and Capone 2020; Tschitschko et al. 2024). Symbioses are also drivers of evolutionary complexity, including as the route by which eukaryotes have gained the capacities for photosynthesis and carbon fixation (McFadden 2001; Gabr et al. 2020). In some symbioses, the smaller partner (symbiont) is recruited by the larger partner (host) from the environment at each host generation, while in others, the symbiont is strictly passed from parent to offspring ("vertical inheritance"). When symbionts are vertically inherited there is often no free-living form of either partner. Vertically inherited symbionts lose ancestral cellular functions and, over time, can become obligately associated with their hosts (Moran et al. 2008). At the extreme, this loss of independent function and increase in host integration has resulted in symbiont-derived organelles including mitochondria, chloroplasts, and chromatophores (Gray and Archibald 2012).

Loss of cellular functions in symbionts is caused by the erosion and loss of functional genes (McCutcheon et al. 2024). In the canonical trajectory, initially, mobile elements proliferate in the symbiont genome and genes break, causing an accumulation of nonfunctional genes (pseudogenes) (McCutcheon and Moran 2012). Mobile elements and pseudogenes are later lost, such that the genome shrinks first in number of functional (intact) genes and then in total length (McCutcheon and Moran 2012). Gene loss can be adaptive, in which case the loss should be inevitable given the deletional bias in bacterial genomes and how easy it is for a mutation to "break" or "turn off" a gene (Giovannoni et al. 2014). Alternatively, neutral or even deleterious gene loss can occur due to probabilistic fluctuations in allele frequencies (drift) given the frequent population bottlenecks and lack of recombination that lead to persistent reductions in effective population sizes (N_e) in vertically inherited bacterial symbionts (Mira and Moran 2002; Bobay and Ochman 2017).

While pseudogenization and gene deletion are pronounced in bacteria experiencing strong drift due to small N_e, this, of course, is counteracted by selection for certain genes to remain functional (Bobay and Ochman 2017). Gene loss is determined, then, by the balance of drift and the selective advantage of one allele (here, a functional gene) over another (here, a lost or broken gene). Indeed, while there is strong evidence that many vertically inherited symbionts have very small N_e (McCutcheon and Moran 2012), efficacious purifying selection is evident in the continued viability of symbionts over tens to hundreds of millions of years and patterns in which

Received: March 4, 2025. **Revised:** May 30, 2025. **Accepted:** June 24, 2025

genes are maintained across independently evolving symbiont lineages (McCutcheon et al. 2024). Thus, genes retained by long-term symbionts reveal, sometimes in astonishing detail, which functions the symbionts perform in the host environment.

Cyanobacteria are an ancient and diverse group of bacteria that have formed symbioses with hosts from across the tree of life (Rai et al. 2000; Adams et al. 2013). Cyanobacteria and cyanobacteria-derived organelles perform the great majority of global carbon fixation and all oxygenic photosynthesis (Gray and Archibald 2012; Fischer et al. 2016; Crockford et al. 2023). Some cyanobacteria are also diazotrophic, fixing nitrogen on a globally significant scale (Warshan et al. 2016; Zehr and Capone 2020). Many of the most highly host-integrated vertically inherited symbionts and symbiont-derived organelles are descended from cyanobacteria that fix carbon or nitrogen for their hosts (Nakayama and Inagaki 2017; Gabr et al. 2020; Coale et al. 2024). Independent lineages of diazotrophic cyanobacteria have formed highly integrated, vertically inherited symbioses with multiple independent lineages of photosynthetic eukaryotic hosts (Becking 1987; Nakayama and Inagaki 2017; Nieves-Morión et al. 2023; Coale et al. 2024). In some of these, the symbiont ("diazo-cyanobiont") has evolved to lose parts or all of photosynthesis, which comprises the energy-capturing "light reactions" and carbon-fixing Calvin–Benson–Bassham (Calvin or CBB) Cycle, living heterotrophically off of photosynthate from the host, while in others, the symbiont remains photosynthetic (Tripp et al. 2010; Nakayama and Inagaki 2017; Foster et al. 2022).

One diazo-cyanobiont that retains full photosynthesis is *Trichormus azollae*, a heterocystous cyanobacterium of the order Nostocales that lives as a vertically inherited symbiont of *Azolla*, a genus of aquatic ferns (Ray et al. 1979; Kaplan and Peters 1988, Ran et al. 2010). *T. azollae* is alternatively *Anabaena azollae* or *Nostoc azollae* (Pereira and Vasconcelos 2014); we use *Trichormus* in keeping with the trend in recent publications. *T. azollae* supplies *Azolla* with fixed nitrogen (Brouwer et al. 2017), and there is evidence that *Azolla* provides fixed carbon to *T. azollae* (Ray et al. 1979; Kaplan and Peters 1988), which means that *T. azollae* would have a host-derived source of reduced carbon were it to lose photosynthetic capacity. *Azolla* is unique among the hosts of the vertically inherited diazo-cyanobionts in being a land plant (Embryophyta) as opposed to an alga. In fact, *Azolla* is the only embryophyte known to have a strictly vertically inherited symbiont (not including plastids and mitochondria) (Rai et al. 2000; Adams et al. 2013). *T. azollae* is also unusual because it is one of the few known extracellular vertically inherited symbionts (Salem et al. 2015; Salem et al. 2017), living as populations within specialized compartments in the leaves of *Azolla* spp. (Peters and Mayne 1974; Ran et al. 2010; Li et al. 2018). Codiversification has been ongoing for ~50 to 90 Ma (Metzgar and Pryer 2007; Testo and Sundue 2016; Li et al. 2018), and research on the first complete *T. azollae* genome found a large number of pseudogenes and low coding density, suggesting that *T. azollae* is at a well-described midpoint on the typical evolutionary trajectory of a vertically inherited symbiont (Ran et al. 2010; McCutcheon et al. 2019), where gene disruption has begun in earnest, but large-scale gene loss events have not culminated in a very small genome. Those genomic characteristics, along with observations of bottlenecks during symbiont propagation by the host (Becking 1987), suggest a small N_e, although models of sequence evolution have not previously been applied to support or refute this idea.

Here, we use comparative genomic approaches to leverage 57 cyanobacterial genomes, including recently developed *dN/dS*-based models of sequence evolution to test the hypothesis of a small N_e for *T. azollae*. We expand on previous analyses of gene loss and pseudogenization in *T. azollae*. We then determine which cellular functions are evolving under more drift-affected (less selection) or selection-dominated regimes. Because a gene can be affected by drift and weak selection even if its total functional elimination might be lethal (Moran et al. 2008), sensitive tests of the selective regime acting on intact genes can provide new information beyond inventories of pseudogenes and lost genes. Given the unique qualities of *T. azollae*, this analysis improves our understanding of what drives loss or retention of photosynthesis in diazo-cyanobionts, and of functions in vertically inherited symbionts generally.

Materials and Methods

Genome Annotations, Pseudogene Prediction, and Orthogroup Clustering

All sample accessions and abbreviations used are listed in supplementary table S1, Supplementary Material online. A total of 57 genomes are compared in this study. Our focal group is the *T. azollae* clade, for which we used eight genomes, including one complete reference-level genome from NCBI (*Nostoc Azollae* 0708, accession GCF_000196515.1; Ran et al. 2010) and seven contig-level metagenome-assembled genomes (MAGs) from ENA (project accession PRJEB45214; Dijkhuizen et al. 2021). A second set of *T. azollae* MAGs from Li et al. (2018) were not used because each of those strains was already included in the set from Dijkhuizen et al. 2021 (more detail in supplementary Material "Methods Sensitivity," Supplementary Material online). Data from Song et al. (2025) were not used in order to avoid oversampling certain species of *Azolla* and because MAGs from individual samples were not available. We compared the symbiont ingroup to the genomes of 47 free-living Nostocales and two outgroup *Gloeobacter* genomes, which were obtained by searching for all complete "Nostocales" and "Gloeobacter" genomes on RefSeq (O'leary et al. 2016; Tatusova et al. 2016), released since 2010 January 1 using the command-line NCBI datasets tool (*ncbi-datasets-cli v. 15.6.0*; Sayers et al. 2022). Quality of MAGs was assessed using *checkM v1.0.18* (Parks et al. 2015) implemented in *KBase v1.4.0* (Arkin et al. 2018; Chivian et al. 2023).

Open reading frames (ORFs) were determined for all genomes by *prokka v. 1.14.6* (Seeman 2014), including for consistency the RefSeq genomes, which already had ORFs as annotated by the prokaryotic genome annotation pipeline (PGAP; Haft et al. 2018; Li et al. 2021). We determined that *prokka* ORF predictions were generally in strong agreement with RefSeq-predicted ORFs (supplementary Material "Methods Sensitivity," Supplementary Material online) so we believe our results would be robust using the PGAP pipeline as an alternate annotation method.

Pseudogenes are genes that mutations have rendered nonfunctional. These can be identified based on homology to intact reference sequences. Pseudogenes were predicted by *Pseudofinder v. 1.1.0* (Syberg-Olsen et al. 2022). *T. azollae* genomes were run against a database of the intact genes from the 47 free-living Nostocales. The free-living Nostocales were run against the same database, but with each genome's own sequences removed. The *T. azollae* genomes were not included in the reference databases due to the high proportion of pseudogenes. All genes reported as pseudogenes in this manuscript were manually inspected to confirm the *Pseudofinder* prediction. We found that the number of genes predicted intact by *Pseudofinder* was on average within 2% of the number of intact genes as annotated on RefSeq (supplementary Material "Methods Sensitivity," Supplementary Material online).

In order to determine gene loss and pseudogenization events among the taxa, as well as to perform sequence-based analyses of gene evolution, it was necessary to characterize all orthologous genes (orthologs) among the 57 taxa in this study. Groups of orthologs (orthogroups) were predicted using *OrthoFinder v. 2.5.5* (Emms and Kelly 2019). *OrthoFinder* uses a phylogenomic tree to improve orthogroup prediction. *OrthoFinder* generated an initial tree as part of its workflow, with concatenated gene alignment performed in *MAFFT v. 7.520* (Katoh and Standley 2013) and tree inference using maximum likelihood in *FastTree v. 2.1.11* (Price et al. 2010). The tree was then rooted on the two *Gloeobacter* genomes using the "*reroot*" function in *Gotree v. 0.4.3* (Lemoine and Gascuel 2021) and *OrthoFinder* was re-run with the rooted tree using the "*-ft*" flag. Except for the two *Gloeobacter* genomes, for which all ORFs were submitted, only ORFs that were predicted intact by *Pseudofinder* were used because of the possibility that the inclusion of pseudogenes would break the assumptions of *OrthoFinder*.

Seven of the free-living Nostocales genomes were deeply diverged; these and the two *Gloeobacter* genomes were removed from all subsequent analyses to maintain phylogenetic proximity between the *T. azollae* and the free-living Nostocales (supplementary fig. S1, Supplementary Material online). This left eight *T. azollae* and 40 free-living Nostocales genomes.

Predicted pseudogenes were assigned to orthogroups by *DIAMOND v. 2.1.7 blastx* (Buchfink et al. 2021) search with pseudogenes as queries and a database of all intact genes from the 40 free-living and eight *T. azollae* genomes. Pseudogenes were assigned to the orthogroup that contained the locus with the lowest *e-value* hit to the pseudogene and functional annotations were propagated from the orthogroup to the pseudogenes (see supplementary Material "Methods Sensitivity," Supplementary Material online for more detail).

Functional annotation was performed by submitting all intact ORFs to eggnog-mapper online *v. 2.1.12* (Cantalapiedra et al. 2021) using default parameters. Gene names, gene descriptions, and KEGG numbers (Kanehisa and Goto 2000; Kanehisa 2019; Kanehisa et al. 2023) were taken directly from eggnog-mapper. Modules and pathways were determined by mapping KEGG numbers using the *KEGGREST v. 1.42.0* (Tenebaum and Bioconductor Package Maintainer 2023) package in *R* (R Core Team 2025). Annotations were propagated from loci to orthogroups inclusively: all annotations assigned to at least one locus in an orthogroup were propagated to the entire orthogroup. If functional annotations of different loci within an orthogroup disagreed with each other, that was noted (supplementary table S5, Supplementary Material online), but this was rarely the case and never the case for the loci discussed in the main manuscript.

Because the seven *T. azollae* MAGs are incomplete, genes could be falsely identified as absent or, in the case of an ORF that runs off the end of a contig, falsely identified as a pseudogene. This is not a systemic concern given the near completeness of the MAGs (*Fig.* 2; supplementary table S2, Supplementary Material online), and steps were taken to mitigate these shortcomings (supplementary Material Methods," Supplementary Material online). In brief, MAGs were reassembled using different thresholds for read filtering and loci were considered intact if they were intact in any of the reassemblies from that sample.

Strength of Selection

All of our strength of selection analyses are based on the ratio of the rate of selectively relevant nonsynonymous mutations (*dN*) to the rate of selectively neutral synonymous mutations (*dS*). The relative strength of selection acting on the *T. azollae* loci compared with orthologs in the free-living Nostocales was determined with *HyPhy RELAX v. 2.5.51(MP)* (Wertheim et al. 2015). *RELAX* utilizes a branch-site model of sequence evolution to test the hypothesis that selection is more "intense" or more "relaxed" in an a priori ingroup of branches on a gene or species tree compared with an a priori outgroup (Wertheim et al. 2015). "Intense" selection means that there is a high probability that allele frequency will increase over generations for the fittest allele, while "relaxed" selection implies that allele frequencies are changing more randomly (Wertheim et al. 2015). Thus, "relaxed" selection could be the result of small N_e, small |*s*|, or both (Charlesworth 2009). We want to emphasize that *RELAX* would call both stronger purifying selection ($dN/dS << 1$) and stronger directional selection ($dN/dS > 1$) "intensified selection", despite the opposite nature of these two selective regimes. Moreover, and adding to the complex terminology, both purifying and directional selection could be present at different sites within the same gene (Wertheim et al. 2015). As such, *RELAX* does not test purifying versus directional selection—it only tests whether the balance of drift and selection is more tilted toward drift in the ingroup ("relaxed") or more tilted toward selection in the ingroup ("intense"), relative to in the outgroup. Because we find it confusing to discuss "relaxed" selection (a function of N_e and |*s*|) and simultaneously discuss the constituent |*s*| as "weak" or "strong" or any such adjective, in this manuscript, we replace "relaxed" with "drift-affected" and "intense" with "selection-dominated." We also hope that these choices of terms emphasize that drift and selection are not mutually exclusive.

We ran *RELAX* on each orthogroup. The ingroup included all *T. azollae* terminal branches of the species tree as well as all internal branches that were ancestral to only the *T. azollae* clade. The outgroup consisted of all other branches. The root branch of the *T. azollae* clade was excluded from both groups because we do not know where along that branch the vertically inherited symbiosis began. To be included in this analysis, an orthogroup had to contain exactly one locus from each of at least three *T. azollae* genomes and at least three free-living Nostocales genomes. This threshold was a functional requirement for *RELAX* to create a distribution of *dN/dS* values. If a genome contained multiple loci assigned to a single orthogroup, none of those loci were included in the *RELAX* analyses because paralogs would break assumptions of the evolutionary model.

For each orthogroup included in this analysis, a multiple sequence alignment (MSA) was generated using *MACSE v. 2.07* (Ranwez et al. 2018). *MACSE* indicates possible frameshifts; this was very uncommon, and we determined that it was most straightforward to eliminate these sequences.

For each orthogroup analyzed with *RELAX*, the species tree was pruned to include only relevant genomes using the "*prune*" function in *Gotree*. The ingroup and outgroup were labeled using the "*rename*" function in *Gotree*. For each orthogroup, *RELAX* gave a *P*-value, which we corrected for multiple hypothesis testing with Benjamini and Hochberg (1995). Orthogroups were assigned results of "drift-affected," "selection-dominated," or "not significant" with corrected $P_{\text{orthogroup}} < 0.1$.

To determine if the number of drift-affected orthogroups was unevenly distributed among KEGG pathways, a new P_{pathway} was calculated for each pathway from a two-sided binomial test where the number of positives is the number of drift-affected orthogroups in that pathway, the number of tests is the total number of *RELAX* results for that pathway, and the hypothesized probability of success is the genome-wide portion of drift-affected orthogroups. These were corrected using Benjamini and Hochberg. Orthogroups were counted for each pathway to which they were assigned, and many were assigned to multiple pathways. The same analysis was performed for selection-dominated orthogroups. In addition to KEGG pathways, these analyses were performed for multicopy genes and orthogroups that contained pseudogenes. Four-letter gene names and KEGG KOs were the most granular functional annotations so if one of those annotations was assigned to multiple loci in the same genome, that indicated a multicopy gene. An orthogroup was considered multicopy for *T. azollae* if it was multicopy for at least one *T. azollae* genome.

In addition to looking at selection acting on individual orthogroups within the *T. azollae*, we were curious to see a single genome-wide comparison, to estimate the overall selective regime acting on the level of the organism. To that end, a concatenated alignment of all single-copy orthogroups was generated by concatenating the *MACSE* alignments for all orthogroups that had exactly one locus from each of the 48 genomes after elimination of sequences that contained frame-shift accommodations in the MSA. The species tree was used with ingroup/outgroup labeling as described above, but without pruning. Genome-wide *dN*/*dS* values were calculated using *hyphy-analyses FitMG94* (https://github.com/veg/hyphy-analyses/tree/master/FitMG94; Kosakovsky Pond et al. 2010) on this same concatenated alignment of single-copy orthogroups with the species tree.

Previous literature has established that *dN*/*dS* can be an unreliable signal of natural selection regimes at low *dS* (Mugal et al. 2014) and that comparing *dN*/*dS* across phylogenomic branches of different lengths may introduce biases where shorter branches are more likely to have elevated values (Wolf et al. 2009). We took three approaches to address possible biases introduced to our analyses of natural selection by the difference in distributions of branch lengths between *T. azollae* and the free-living Nostocales. First, we used different ingroups to repeat the process of running individual orthogroups through *RELAX*. The selection of different ingroups was based on clades that subjectively looked like the *T. azollae* clade in terms of total divergence and numbers of leaves (Fig. 1). For each ingroup, the outgroup was defined as the entire tree except the ingroup and the *T. azollae* clade, which was excluded to not skew the selective regimes in the outgroups. Second, we subsampled the genome set (removing four *T. azollae* and six free-living genomes) so that each pairwise distance between genomes had $dS > 0.01$ (supplementary fig. S5, Supplementary Material online) and then ran *RELAX* on the concatenated alignment of single-copy orthogroups with *T. azollae* as the ingroup as previously described for the full genome set, as well as calculating *dN*, *dS*, and *dN*/*dS* for the concatenated alignment of single-copy orthogroups along each branch in the subsampled tree. Third, we compared *dN*/*dS* values between *T. azollae* and the free-living Nostocales for only genomes with low *dS*.

Other Analyses

A binomial test was also performed for enrichment of KEGG pathways for pseudogenes. Of necessity, only orthogroups that contained at least one pseudogene and had a KEGG pathway annotation were considered. In addition, orthogroups had to have loci present from at least six genomes (combined *T. azollae* and free-living) to parallel the thresholds for *RELAX*.

Supplementary table S5, Supplementary Material online includes the "Dates of minimum gene loss events" for each gene that is in the free-living Nostocales core, but not in the *T. azollae* core. These were determined using parsimony for the fewest total number of gene loss events, considering pseudogenes equivalent to absent genes.

Results

Initial Pipeline

A total of 57 genomes were obtained from public repositories (supplementary table S1, Supplementary Material online), of which, nine taxa were removed from the final analysis for being too phylogenetically distant for comparison to our ingroups. All following results and discussion refer only to the remaining 48 taxa (eight *T. azollae* and 40 free-living Nostocales). Though highly fragmented (264 to 368 contigs), the *T. azollae* MAGs are all highly complete (≥97.89% according to *checkM* cyanobacteria marker genes) and minimally contaminated (≤0.22%) (supplementary table S2, Supplementary Material online).

After ORF prediction with *prokka* and pseudogene prediction with *Pseudofinder*, a phylogenomic tree was generated using *FastTree* implemented in *OrthoFinder*, and rooted using *Gloeobacter* as the outgroup (Fig. 1). All bootstrap support values are one at each node. Some genera are not monophyletic, but that is to be expected given the difficulty of assigning genera within Nostocales (Huo et al. 2021; Mishra et al. 2021). *T. azollae* are monophyletic and their phylogeny matches the *T. azollae* phylogeny generated by Li et al. (2018), which used many of the same reference strains as our analysis but independently extracted DNA from different individuals. Notably, both our *T. azollae* phylogeny and that of Li et al. (2018) diverge from the host phylogeny of Li et al. (2018) in the placement of *A. caroliniana* and its symbiont. However, when a short internal branch is collapsed, the host and symbiont topologies agree. Although all bootstrap support values were one, given the importance of tree topology to selection analyses, we wanted to confirm the tree topology using an alternative method. To that end, we inferred trees using two alternative methods (*RAxML-ng*; Kozlov et al. 2019) using the same alignments as *FastTree* in *OrthoFinder* and *GTDB-tk* (Chaumeil et al. 2022), which uses a smaller set of marker genes, which increased our confidence in the tree that we used (see detailed discussion in supplementary Material "Methods Sensitivity," supplementary fig. S1, and supplementary fig. S2, Supplementary Material online).

A KEGG identifier or four-letter gene name was assigned to 28.3% of orthogroups and 44.9% of loci, while KEGG pathways were assigned to 14.2% of orthogroups and 25.5% of loci. A small portion of loci were not placed into orthogroups

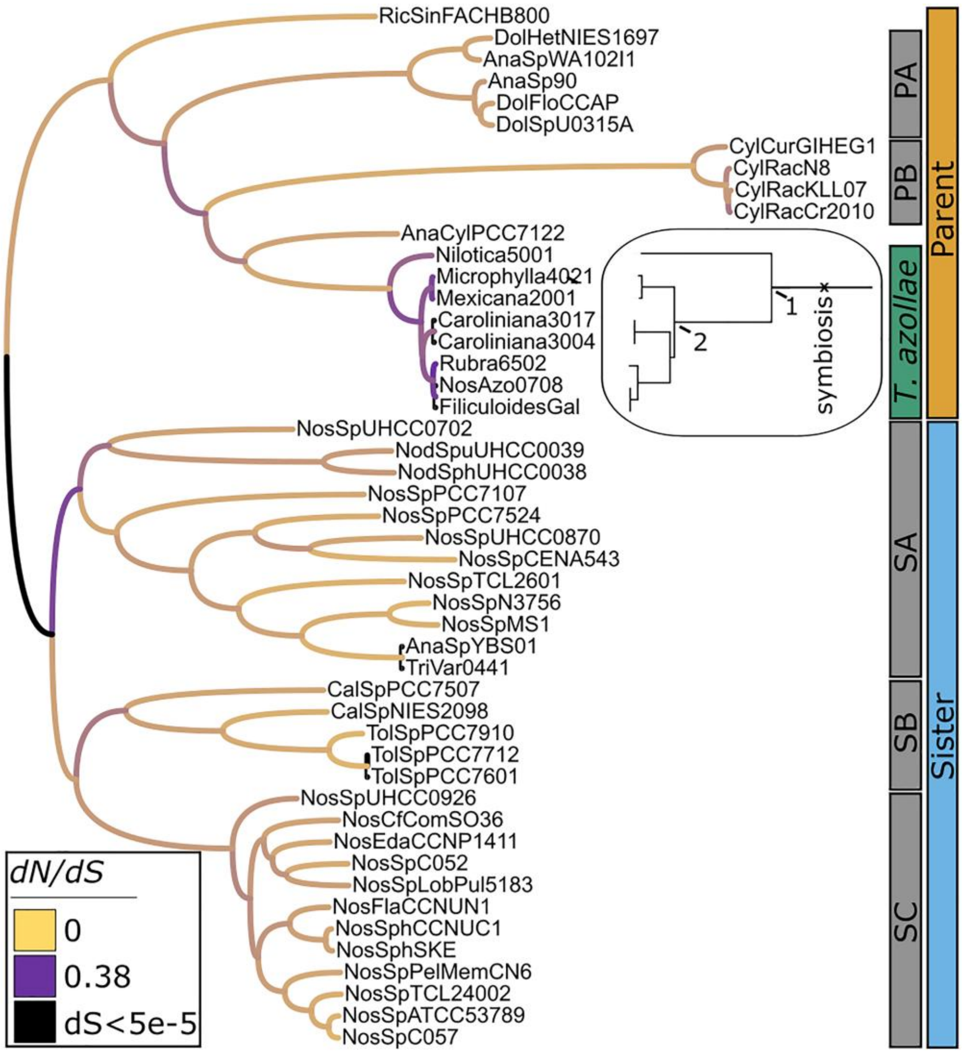

Fig. 1. Genome-wide *dN/dS* on phylogenomic tree: Maximum likelihood species tree generated by *FastTree* implemented in *OrthoFinder*, rooted on the *gloeobacter* outgroups and pruned to include only the 48 taxa included in downstream analyses. Bootstrap values are one at each node. Focal *T. azollae* clade is labeled with green bar. Branches are colored by *dN/dS* calculated on a concatenated alignment of 1,015 single-copy *OrthoFinder*-generated orthogroups. Black branches have extremely low *dS* (<5e-5) so *dN/dS* calculations would be unreliable. All other branches have 5.1e-3 ≤ *dS* ≤ 1.1. "Parent" and "Sister" clades do not necessarily represent a meaningful biological partition but are useful for comparing the broader evolutionary context of the Nostocales to which *T. azollae* is most closely related to the rest of the Nostocales. The other subclades do not necessarily represent a meaningful biological partition but are useful for testing methods sensitivity as described in the main text and in Fig. 3. Inset: stretched view of *T. azollae* clade. Divergence times as estimated from Metzgar and Pryer 2007 and Testo and Sundue 2016: (symbiotic origin) 50 to 90 Ma; (1) ~50 Ma; (2) ~16 Ma.

by *OrthoFinder*. This represented 0.031% to 5.1% of intact sequences from the *T. azollae* genomes and 0.060% to 8.3% of intact sequences from the free-living genomes. The large majority of these sequences are poorly annotated and none of them affects any analyses presented in this manuscript.

We identify 1,042 "single-copy" orthogroups that each contain exactly one intact locus from each of the 48 genomes. After elimination of 27 such orthogroups in which *MACSE* detected frameshifts, alignments for the remaining 1,015 single-copy orthogroups were concatenated, generating one sequence per genome of 0.98 to 1.00 Mbp to be used in downstream analyses of single-copy orthogroups. This represents 30% to 33% of the orthogroups in each *T. azollae* genome and 27% to 31% of the loci in each *T. azollae* genome.

Genome Statistics

The Parent clade genomes all have lower GC% than the Sister clade genomes and the Parent clade are also shorter with three exceptions (Mann–Whitney–Wilcoxon test $P < 10^{-5}$ for each) (Fig. 2; supplementary table S3, Supplementary Material online). The only complete *T. azollae* genome, *NosAzo0708*, is within the free-living ranges in terms of GC%, genome length, and number of intact genes (Fig. 2; supplementary table S3, Supplementary Material online), although only the *Cylindrospermopsis*, which have the shortest genomes, have fewer intact genes.

All *T. azollae* are major outliers in terms of pseudogenes (Fig. 2), especially when accounting for genome length (supplementary fig. S3, Supplementary Material online). Counting pseudogenes, which are often fragmented, is difficult, but we are confident that this signal is real given its magnitude. Identifying mobile elements is outside of the scope of this study, but a significant number of mobile elements have previously been documented in *NosAzo0708* (Ran et al. 2010).

The 40 free-living Nostocales genomes show extremely strong linear correlation between genome length and number of intact genes (i.e. genes per Mbp, $R^2 = 0.99$) and between

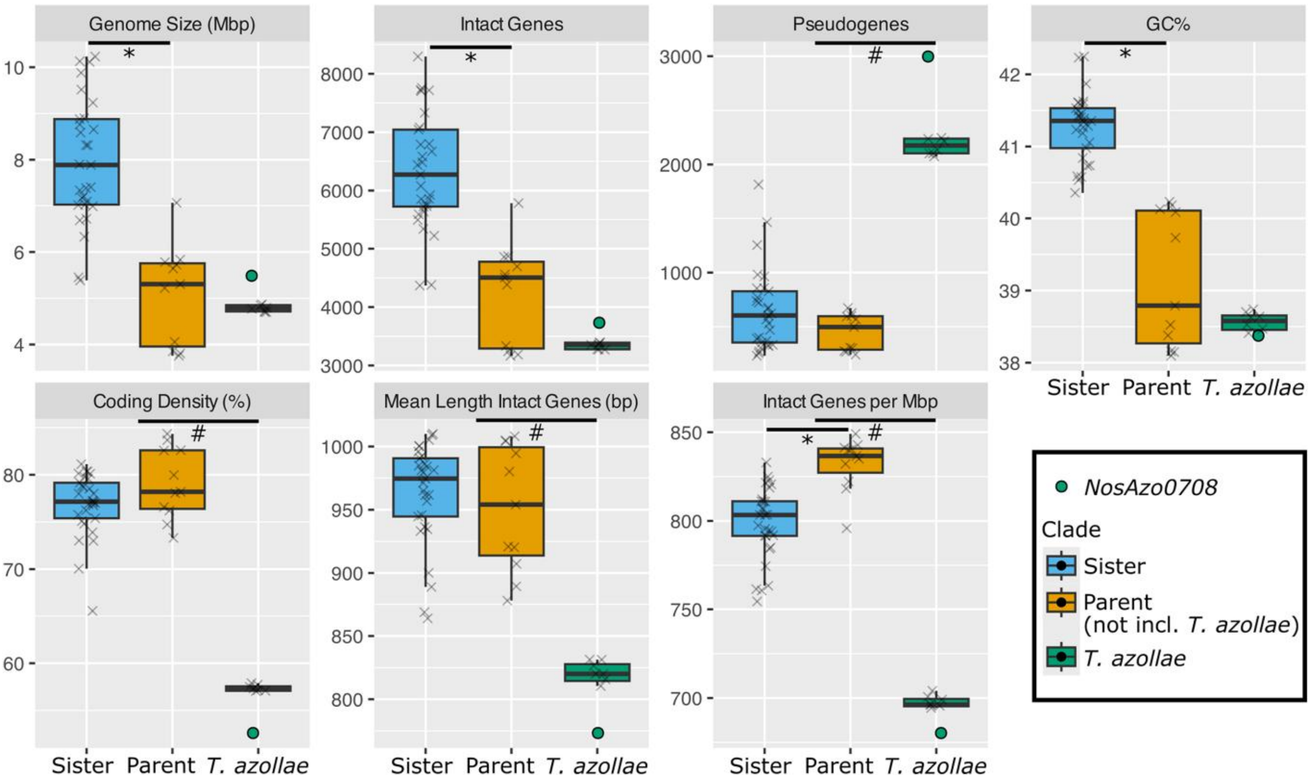

Fig. 2. Genome statistics: genome statistics for 40 complete Nostocales genomes, one complete *T. azollae* genome, and seven *T. azollae* MAGs. Green circle represents *NosAzo0708*, the only complete *T. azollae* genome. Clades as labeled in Fig. 1. Parent clade does not include the *T. azollae* subclade. *: $P < 10^{-5}$ (all others $P > 0.05$), Mann–Whitney–Wilcoxon Test for difference between Sister and Parent clades (not including *T. azollae*). #: value for *NosAzo0708* is greater or less than the range for all free-living Nostocales by >50% of the full free-living range.

genome length and the cumulative length of all intact genes (i.e. coding density, $R^2 = 0.97$) (supplementary fig. S3, Supplementary Material online). *T. azollae* have the lowest coding density, due to having both the fewest intact genes per Mbp and the shortest mean length of intact genes (Fig. 2). The four *Cylindrospermopsis* have the highest coding density (supplementary table S3, Supplementary Material online).

Core Nostocales Genes Missing From *T. azollae*

We define a core genome as genes that are intact in every genome within a given set of genomes, so, for instance, genes in the free-living core are intact in every free-living genome, but not necessarily in every *T.* azollae genome. The pan genome consists of all genes that are intact in at least one genome in a set of genomes, so the pangenome subsumes the core genome. For a summary of the core and pan genomes of *T. azollae* and of the free-living Nostocales as determined by orthogroups, see supplementary fig. S7, Supplementary Material online. We focus here on core and pan genomes as defined by functional annotations instead of orthogroups because this is more relevant to understanding which cellular functions might be compromised in *T. azollae*. So, for instance, if multiple orthogroups are annotated as *gene A*, no individual orthogroup has to be intact in every genome for *gene A* to be intact in every genome and thus part of the core genome. Thirty-nine genes (as identified by gene name and KEGG KO) from the free-living core were found to be pseudogenized or absent in at least one *T. azollae* genome. Some are highlighted here. All are summarized in supplementary table S5, Supplementary Material online, along with an estimated time since the functional gene was lost. We did not identify any functional annotations that were unique to *T. azollae*.

All of the *T. azollae* genomes contain a pseudogenized copy of *nifJ*, and no intact copies. *nifJ* is required for diazotrophic growth of *Anabaena sp. PCC 7120* only when iron is limited (Bauer et al. 1993). *moaB* and *moaC* are pseudogenes in all *T. azollae* genomes and *moaE* is additionally pseudogenic in *Nilotica5001*. The *moaABCDE* operon is involved in biosynthesis of the molybdenum cofactor MoCo, which is needed for nitrate reduction and all other functions that require a molybdenum catalyst except for nitrogen fixation, which requires a different molybdenum-bearing cofactor, FeMo (Rubio et al. 1998; Mendel 2013; Ringel et al. 2013). While the precise role of the heme oxygenase *hutZ/hugZ* (possibly intact in *Nilotica5001*, pseudogenic in all others) is not well established, it is involved in heme utilization, including heme degradation and possibly iron acquisition (Guo et al. 2008; Dojun et al. 2020). The vitamin B12 transporter *btuB* is absent from all *T. azollae* genomes. *T. azollae* genomes are compromised for sugar uptake including ABC sugar transporter proteins and a melibiose permease (*melB*, pseudogenic in all) that is responsible for glucose uptake and is ubiquitous among cyanobacteria (Moreno-Cabezuelo et al. 2019). Phosphorus acquisition is compromised, including alkaline phosphatases (*phoA*/*phoD*), a phosphonate transporter (*phnD*), and glycerophosphoryl diester phosphodiesterase (*glpQ*/*ugpQ*). Finally, *cheR*, a chemotaxis protein methyltransferase, is absent in all *T. azollae* genomes.

We also looked at some genes of interest including those involved in nitrogen fixation and photosystems I and II of the photosynthetic light reactions. All eight *T. azollae* have the same two orthologous copies of nitrogenase subunit *nifH* except for *Nilotica5001*, in which one of those copies is predicted to be a pseudogene due to an internal stop codon.

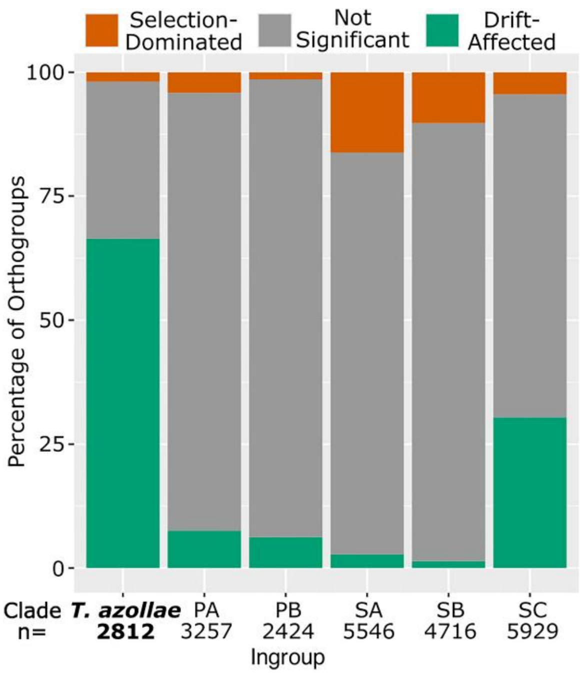

Fig. 3. Percentage of orthogroups evolving under selection-dominated or drift-affected selective regimes: clades as shown in Fig. 1. Percentages of orthogroups with different *RELAX* results (corrected $P \leq 0.1$) using different ingroups and using all noningroup branches except for *T. azollae* as the outgroup.

The photosystem II gene *psbZ* is more than 50% longer in *T. azollae* than in the free-living Nostocales, but the sequences are highly conserved within the *T. azollae*, suggesting that this gene has diverged from the free-living form, but are still functional.

Genome-wide Strength of Selection

A concatenated alignment of all single-copy orthogroups was used to calculate genome-wide *dN/dS* for each branch in the species tree (Fig. 1) and for a genome-wide *RELAX* model using *T. azollae* as the ingroup and free-living Nostocales as the outgroup. Eight branches have exceedingly low *dS* values (<5e-5) that we felt might be susceptible to noise in a *dN/dS* calculation. Removal of those eight branches leaves 85 branches with $5.1\text{e-}3 \leq dS \leq 1.1$. Genome-wide *dN/dS* is elevated within the *T. azollae* clade: the 10 branches within the *T. azollae* clade (excluding the subclade root) have 10 of the 11 highest *dN/dS* values in the tree. The root of the *T. azollae* clade has a below-average *dN/dS* (69th highest of 85 branches). *RELAX* determines that the single-copy orthogroups are significantly more drift-affected within the *T. azollae* clade than in the free-living Nostocales ($k = 0.50$, $P <$ 10e-4).

To improve the granularity of our understanding of selection in *T. azollae*, *RELAX* was run on individual orthogroups, again using *T. azollae* as the ingroup. Of the 2,825 orthogroups that met the thresholds described in *Methods*, 13 did not converge so were removed from downstream analyses. The remaining 2,812 orthogroups represent 81% to 85% of orthogroups and 73% to 82% of loci for the *T. azollae* genomes. Using a corrected *P*-value cutoff of $P < 0.1$ for each orthogroup, *RELAX* determined that 66.5% of orthogroups were significantly more drift-affected in *T. azollae* compared with in the free-living Nostocales, while 1.9% were more selection-dominated in *T. azollae*, and 31.7% did not differ significantly between ingroup and outgroup (Fig. 3).

We took three approaches to control for possible biasing effects of variable branch lengths in our selection analyses. Wolf et al. 2009 finds that *dN/dS* decreases in both mean and variance as *dS* increases, which is clearly replicated in our data, as visualized on graphs of genome-wide *dN/dS* versus *dS* for both the full genome set and a subsampled genome set for which the pairwise distance between any two genomes is $dS > 0.01$ (supplementary figs. S4a and S6a, Supplementary Material online). However, the *T. azollae* genomes clearly display elevated *dN/dS* even when compared only to free-living genomes of similar *dS* (supplementary figs. S4b and S6b, Supplementary Material online). *RELAX* run on the subsampled genome set determines that the single-copy orthogroups are significantly more drift-affected within the *T. azollae* clade than in the free-living Nostocales ($k = 0.33$, $P <$ 10e-4). The gross results of the orthogroup-level *RELAX* results were essentially unchanged, with the percentage of drift-affected orthogroups in *T. azollae* changing from 66.5% to 66.3%, the percentage selection-dominated changing from 1.9% to 1.2%, and the percentage not significant changing from 31.7% to 32.5%. Finally, we repeated the individual orthogroup *RELAX* analysis using different ingroups (Fig. 3). We did not find any obvious bias toward one result over another. Three ingroups (including *T. azollae*) have more drift-affected orthogroups than selection-dominated orthogroups and three ingroups have the opposite. Clade SC has 30.4% drift-affected orthogroups, a much higher rate than the other four ingroups, which have 1.4% to 7.5% drift-affected orthogroups. Of the 12 strains in clade SC, four are nonsymbiotic, four are facultative symbionts of embryophytes, and four are facultative lichen symbionts (supplementary table S1, Supplementary Material online). In contrast, of the 28 strains outside of the *T. azollae* and SC clades, one is a facultative symbiont of an embryophyte, 23 are nonsymbiotic, and four are undetermined (supplementary table S1, Supplementary Material online). We still refer to facultative symbionts as free-living because they are free-living sometimes, unlike strictly vertically inherited symbionts, which are never free-living.

Decreasing the corrected *P*-value cutoff for significance decreases the percentage of significant results for each ingroup, but does not alter the main result that *T. azollae* has a much greater percentage of drift-affected orthogroups than do the other ingroups (42.0% for *T. azollae* at $P < 0.01\%$ and 26.6% at $P < 0.001$; supplementary table S4, Supplementary Material online).

Evolutionary Trends in Individual Cellular Functions

Orthogroups affected by drift in *T. azollae* are nonrandomly distributed among KEGG pathways (Fig. 4 and Table 1). KEGG modules are subsets of KEGG pathways so could allow more granular analysis. However, most KEGG modules are so small that even extreme results (i.e. 100% or 0% drift-affected) would be insignificant; zero modules show significant divergence from the genome-wide percentage of drift-affected orthogroups. Though not significant, each of the four modules within the pathway "Photosynthesis" (light reactions) has fewer drift-affected orthogroups than the genome-wide average. Those modules are "Photosystem I," "Photosystem II," "Cytochrome b6f complex," and "F-type ATPase, prokaryotes, and chloroplasts." Similarly, though not individually significant, each of the five KEGG pathways

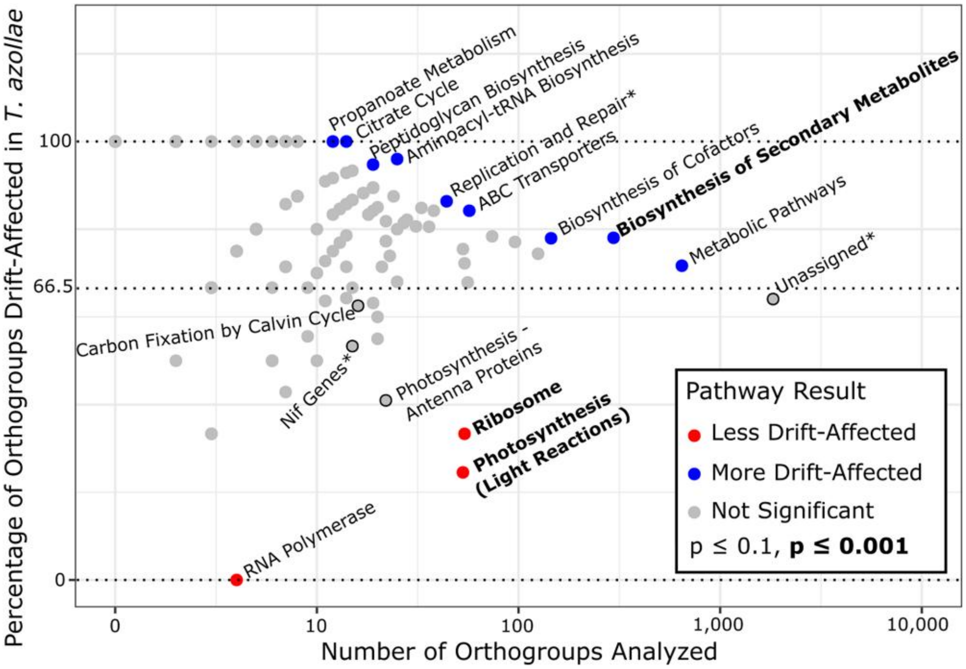

Fig. 4. KEGG pathway enrichment for drift-affected orthogroups: orthogroups that are more drift-affected in *T. azollae* genomes than in the free-living Nostocales are nonrandomly distributed among cellular functions as defined by KEGG pathways. ***** are not KEGG pathways: "Replication and Repair" comprises five KEGG pathways that are grouped under this label on the KEGG database, "Nif genes" comprises all genes involved in nitrogen fixation with gene name abbreviations that begin with "*nif*" and "Unassigned" comprises all orthogroups not assigned to a pathway. A total of 982 orthogroups were analyzed. The central dashed line represents the genome-wide percentage of drift-affected orthogroups (66.5%). Pathways that differ significantly from that genome-wide percentage are labeled and are represented as blue and red circles. Pathways of interest with insignificant results are labeled and are represented as gray circles with black outlines.

grouped in "Replication and Repair" have more drift-affected orthogroups than the genome-wide average.

No KEGG pathway or module showed significant divergence from the genome-wide proportion of selection-dominated orthogroups, and manual examination of functional annotations did not reveal any obvious patterns. Five of the 53 selection-dominated orthogroups were well-annotated and intact in all 47 genomes. Those five are: *folK* (involved in folate biosynthesis; Romine et al. 2017); *lgt/umpA* (lipid modification; Mao et al. 2016); *atpG* (involved in biosynthesis of ATP synthase; Chaux et al. 2023); *apcC* (involved in biosynthesis of phycobilisomes; Steiner et al. 2003), and *nifE* (part of the *NifNE* protein complex which is involved in biosynthesis of FeMo cofactor of nitrogenase; Fani et al. 2000).

There were 304 orthogroups that contained at least one *T. azollae* pseudogene that were also analyzed in *RELAX* (with pseudogenes removed from *RELAX* analysis, as described in *Materials and Methods*). The percentage of these that were drift-affected (66.8%) did not diverge significantly from the genome-wide percentage (66.5%). The percentage of these that were selection-dominated (4.9%, 15 orthogroups), however, was significantly greater than the genome-wide percentage (1.9%, uncorrected $P = 7.8e\text{-}4$). The 533 multicopy gene orthogroups that were analyzed in *RELAX* were significantly more likely to be drift-affected than the genome-wide rate (72%, uncorrected $P = 4.4e\text{-}3$).

We find that there is only one KEGG pathway that has significantly fewer pseudogene-containing orthogroups than the genome-wide proportion (using 19.7% of the 237 orthogroups that meet the thresholds described in *Materials and Methods*, not 35.0% of all orthogroups) and none that exceed it. That pathway is "Ribosome," in which zero of 56 orthogroups contain a pseudogene from any *T. azollae* genomes ($P = 9.7e\text{-}4$) and all *T. azollae* genomes share the exact same set of orthogroups. All other pathways have corrected $P > 0.35$. Although not individually significant, if the two pathways involved in the light reactions, "Photosynthesis—antenna proteins" and "Photosynthesis" are combined, they have fewer pseudogenes than expected ($P = 5.3e\text{-}2$, five of 86 orthogroups). We acknowledge the small sample size and that pseudogene assignment to orthogroups and thus pathways was done with a less refined methodology than was used for intact loci.

Discussion

We were motivated to understand the balance of drift and selection in shaping the ongoing genomic evolution of a vertically inherited symbiont. We were particularly interested in the selective regime under which the capacity for photosynthesis has been maintained in *T. azollae*, a diazotrophic cyanobacterial vertically inherited symbiont of the fern genus *Azolla*. We (i) confirm that loss of ancestral genes has occurred in *T. azollae* and that different strains have lost different genes, indicating that gene loss is ongoing; (ii) discover genome-wide effects of drift; (iii) discover that drift-affected genes are significantly nonrandomly distributed across cellular functions, which must indicate strong purifying selection on those functions that are less affected by drift, including the light reactions of photosynthesis. Thus, we answer both motivating questions: the *T. azollae* are evolving under the influence of both drift and selection, and photosynthesis is retained due to strong purifying selection on the light reactions. We explore each of these points in more detail in the sections below. At the end, we leverage our knowledge of other vertically inherited diazocyanobionts of photosynthetic eukaryotic hosts to propose

Table 1 KEGG pathway enrichment for drift-affected orthogroups: all pathways with a significant (corrected $P \leq 0.1$) result from the analysis of enrichment of KEGG pathways for drift-affected orthogroups (OGs), as visualized in Fig. 4

	P	% Drift-Affected	# OGs Analyzed	# OGs NOT Analyzed
Lower percentage of drift-affected OGs than genome-wide percentage				
Photosynthesis (light reactions)	8.8e-08	25	53	40
Ribosome	3.7e-05	33	54	10
RNA polymerase	1.0e-01	0	4	5
Higher percentage of drift-affected OGs than genome-wide percentage				
Biosynthesis of secondary metabolites	4.8e-04	78	296	359
Aminoacyl-tRNA biosynthesis	2.3e-02	96	25	8
ABC transporters	5.4e-02	84	57	203
Citrate cycle	5.5e-02	100	14	9
Metabolic pathways	5.5e-02	72	646	965
Replication and repair[a]	6.1e-02	86	44	54
Peptidoglycan biosynthesis	6.2e-02	95	19	15
Biosynthesis of cofactors	6.8e-02	78	145	130
Propanoate metabolism	1.0e-01	100	12	10
Pathways of interest with nonsignificant results				
Photosynthesis—antenna proteins	1.6e-01	41	22	43
Unassigned[a]	1.7e-01	64	1830	14006
Nif genes[a]	4.3e-01	53	15	11
Carbon fixation by Calvin cycle	8.1e-01	62	16	7

[a]Are not KEGG pathways as described in Fig. 4 caption. A small set of pathways of interest with insignificant results are also shown. The lack of significant divergence in the "Unassigned" orthogroups from the genome-wide percentage of drift-affected orthogroups eases concerns that proteins with assigned functions in KEGG are biased toward particular evolutionary constraints. The category "% Drift-Affected" is based only on analyzed orthogroups. Orthogroups not analyzed were excluded as described in Methods.

that the heterocystous morphology of *T. azollae* is the reason that selection favors retention of photosynthesis.

Absent and Broken Genes Indicate Compromised Cellular Functions

While photosynthesis has not been lost in *T. azollae*, other functions have. Here, we briefly discuss some of the genes that are core to all free-living Nostocales, but not to *T. azollae*. Knockout of *cheR* caused a loss of motility and host competency in the typically motile differentiated morphology (hormogonia) in a Nostocales that is a facultative symbiont of a different plant (Duggan et al. 2013). *T. azollae* lacks *cheR*, but its hormogonia are still motile (Adams et al. 2013). The loss of some phosphate and carbohydrate transporters has been noted in a previous genomic analysis of *T. azollae* (Ran et al. 2010); we expect, however, that there must remain some functional phosphate and carbohydrate transporters given the high ATP demands of nitrogen fixation (Inomura et al. 2020), evidence that *T. azollae* receives photosynthate from the host (Kaplan and Peters 1988), and genomic evidence of other carbohydrate transporters (Ran et al. 2010). Several missing or pseudogenized genes relate to uptake and utilization of metals. Biosynthesis of MoCo, which appears to be compromised, is required by many cellular functions, including nitrate reduction (Mendel 2013; Ringel et al. 2013). Indeed, *T. azollae* likely does not reduce nitrate (Holst and Yopp 1979; Ran et al. 2010). Interestingly, *nifE*, which is involved in biosynthesis of the FeMo cofactor of nitrogenase, the one molybdenum-bearing cofactor that does not require MoCo (Fani et al. 2000; Mendel 2013), is one of the few selection-dominated orthogroups in *T. azollae*. The pseudogenization of *nifJ* may indicate that the host ensures an adequate supply of iron to the symbiont. Biosynthesis of vitamin B12 (which contains cobalt) was lost in one, but not the other, of two vertically inherited diazo-cyanobionts of diatoms (Nakayama and Inagaki 2017), but potentially compromised B12 uptake suggests that this is not the case in *T. azollae*.

Further evidence for the loss of ancestral genes is revealed by low coding density in the *T. azollae* genomes, which is due in part to the large number of previously coding, but now noncoding pseudogenes. Low coding density and pseudogene accumulation are hallmarks of an intermediate stage in the canonical evolution of vertically inherited symbionts (McCutcheon and Moran 2012). Although the *T. azollae* have lost genes, their genomes are not (yet) superlatively reduced in terms of absolute number of intact genes relative to the free-living Nostocales: the *Cylindrospermopsis* (clade PB, also known as *Raphidopsis*), which are bloom-forming, freshwater, and planktonic (Aguilera et al. 2018), have the fewest intact genes. Strikingly, *T. azollae* have the lowest coding density while *Cylindrospermopsis* have the highest. While high coding density is observed in late-stage vertically inherited symbionts (McCutcheon and Moran 2012), in free-living bacteria like *Cylindrospermopsis*, it is more likely the result of selection for a streamlined genome, which could also explain the small number of intact genes in *Cylindrospermopsis* (Giovannoni et al. 2014).

T. azollae is not an outlier for GC% or genome length. While low GC% and small genomes are common characteristics of vertically inherited symbiont genomes (McCutcheon and Moran 2012; McCutcheon et al. 2019), including those that are extracellular (Salem et al. 2015; Salem et al. 2017), and of genomes from organisms with low N_e more generally (Rocha and Feil 2010; Bobay and Ochman 2017), those are not universal correlations (Van Leuven and McCutcheon 2012).

Similar analyses were performed by Ran et al. (2010), and some of our findings confirmed their results, as cited above. However, Ran et al. (2010) used only one *T. azollae* genome (*NosAzo0708*) compared with our eight and their core genome was not Nostocales specific. Our broader taxonomic sampling of *T. azollae* reveals their diversity of gene content (supplementary table S5, Supplementary Material online). Placed in taxonomic context, these patterns of diversity give clues to the order of gene loss and reveal that gene loss is ongoing, with some core Nostocales genes being differentially lost on only one side of recently diverged *T. azollae* lineages (supplementary table S5, Supplementary Material online). The fact that photosynthesis has been retained across all eight independently evolving *T. azollae* lineages analyzed is thus all the more remarkable and strengthens the case that photosynthesis is maintained by strong purifying selection.

Models of Molecular Evolution Reveal Genome-wide Drift Which Suggests Small N_e

This is the first application of *dN*/*dS*-based molecular evolutionary models to *T. azollae* genomes. We find elevated *dN*/*dS* across a concatenated alignment of all single-copy orthogroups, which is characteristic of vertically inherited symbionts, suggests effects of drift, and generally indicates small N_e (Bobay and Ochman 2017). Although elevated *dN*/*dS* can also indicate strong directional selection, this is unlikely to be the case genome-wide, as genome-wide histograms of *dN*/*dS* have a modal value close to zero (Buschiazzo et al. 2012). Drift is further confirmed by the *RELAX* model, which determines that the single-copy orthogroups are drift-affected. To determine if this signal is truly genome-wide, as opposed to being driven by a small subset of orthogroups, we ran *RELAX* on individual orthogroups with the result that 66.5% of orthogroups were drift-affected (corrected $P < 0.1$). The fact that 66.5% of orthogroups are drift-affected strongly suggests that *T. azollae* have a reduced N_e: if the effects of drift were driven primarily by weak selection (low |s|) on individual genes, we would expect a more confined signal. To that point, even the KEGG pathway that is least drift-affected, "Photosynthesis," still contains 25% drift-affected orthogroups.

The *T. azollae* have the lowest average gene length, which we believe supports that drift is driving gene loss: under pure drift, a mutation would be equally likely to fix at all loci, so longer genes, which provide a larger target for mutations, would be more likely to accrue mutations and break faster. Although *T. azollae* is not experiencing pure drift, many genes are experiencing significant drift such that this mechanism may be relevant.

Graphs of *dN*/*dS* by *dS* (supplementary fig. S4a and S6a, Supplementary Material online) clearly demonstrate the potential for biases when comparing *dN*/*dS* among branches of greatly different *dS*. However, the central results that *T. azollae* have an elevated genome-wide *dN*/*dS*, that the concatenated alignment of single-copy orthogroups is drift-affected, and that ~66.5% of orthogroups are drift-affected ($P \leq 0.1$) are replicated when comparing only to similarly short *dS* branches of the free-living Nostocales (supplementary fig. S4b and S6b, Supplementary Material online) and when subsampling the genome set so that all pairwise distances between genomes have $dS > 0.01$. Use of alternative ingroups demonstrates that the orthogroup-level *RELAX* results are not biased toward drift-affected in every case where the ingroup is significantly less diverged than the outgroup. The sensitivity of this analysis is further supported by the fact that clade SC, which contains facultative symbionts, has a percentage of drift-affected orthogroups that is intermediate between *T. azollae* and the strictly free-living clades, considering that facultative symbionts can exhibit similar genomic signatures to vertically inherited symbionts, though less extremely (Moran et al. 2008). While we cannot rule out some effects of the biases introduced by uneven branch lengths between in and outgroup, we believe that these controls demonstrate the central results of these analyses should be robust to any biases.

Drift-affected Orthogroups are Significantly Nonrandomly Distributed Among Cellular Functions, Indicating Differences in Strength of Selection

Due to the lack of recombination in genomes of other vertically inherited symbionts (McCutcheon et al. 2019), we assume that N_e is the same for all loci in the *T. azollae* genome (Charlesworth 2009). Thus, any differences among cellular functions in the balance of drift and selection must be due to different strengths of selection (Charlesworth 2009). While the *RELAX* model does not specify purifying selection, an intact gene must be under purifying selection at least to the degree that nonfunctional alleles have been selected against.

Several pathways have a significantly elevated percentage of drift-affected orthogroups, indicating weak purifying selection on these pathways (Fig. 4). The identities of these pathways suggest that genes with more confined effects experience weaker selection on average: secondary metabolites are peripheral to central metabolites and "Metabolic pathways" at large might have fewer network effects than the ribosome, for instance. Contrary to that pattern, all of the genes within the "Propanoate metabolism" pathway are also involved in other pathways of central carbon metabolism including pyruvate metabolism and the citrate cycle. Weak selection on those genes as well as the "Citrate cycle" is interesting given the unusual and poorly understood carbon economy of this symbiosis. We note that cyanobacteria generally lack the complete canonical citrate cycle (Zhang and Bryant 2011).

Weak selection on "Biosynthesis of cofactors" and "ABC transporters" parallels some of the missing genes discussed earlier. Poor DNA repair is a hallmark of vertically inherited symbionts (Moran et al. 2008) and degradation of both replication and repair machinery may facilitate the accumulation of mutations that cause pseudogenes and rapid molecular evolution (Moran et al. 2008; McCutcheon et al. 2024). Peptidoglycan biosynthesis is lost in several intracellular vertically inherited symbionts of insects (McCutcheon and Moran 2012); genes involved in peptidoglycan synthesis may also be involved in heterocyst development (Videau et al. 2016). The significantly elevated rate of drift-affected orthogroups among multicopy genes is in line with what we know about evolution in multicopy genes (Copley 2020); we do not think this phenomenon affects our pathway-level results because such multicopy genes are not obviously over- or under-represented in the KEGG pathways.

There are very few selection-dominated orthogroups and no significant signal in their distribution. This makes sense given that *RELAX* uses free-living Nostocales to set the null expectation, so the strength of selection would have to be so much stronger in *T. azollae* to overcome small N_e.

Ribosomal Proteins and the Light Reactions of Photosynthesis Experience the Strongest Selection

The KEGG pathways "RNA polymerase" (subunits of that enzyme) and "Ribosome" (ribosomal proteins) have significantly lower percentages of drift-affected orthogroups, indicating strong purifying selection. This makes sense given the central importance of transcription and translation to all cellular functions and the fact that these functions are not easily outsourced to a host. It also matches expectations from other vertically inherited symbionts: genes encoding ribosomal proteins, subunits of RNA polymerase, and aminoacyl-tRNA synthetases are among the few protein-coding genes that are generally intact even in the smallest vertically inherited symbiont genomes (Tripp et al. 2010, Graf et al. 2021; Rangel-Chávez et al. 2021; McCutcheon et al. 2024). This result thus increases our confidence in our approach. Paradoxically, given that tRNAs work in tandem with ribosomes, orthogroups in “Aminoacyl-tRNA biosynthesis,” which consists mainly of tRNA synthetases, are marginally significantly more likely to be drift-affected than the genome-wide average. Loss of sequence conservation has previously been observed in aminoacyl-tRNA synthetases of parasites and vertically inherited symbionts, specifically in the editing domains (Melnikov et al. 2018). We did not perform domain-level analyses, so that should be explored in future work.

“Photosynthesis” (light reactions), appears to be under the strongest purifying selection. The results for “Carbon Fixation by Calvin Cycle” and for the *nif* genes (nitrogen fixation), however, are not significant, indicating that they are not under particularly strong purifying selection. If organismal fitness is proportional to metabolic flux and selection is strongest on flux-limiting steps (Dykhuizen and Dean 1990), these results suggest that this system is more often limited not by carbon and nitrogen fixation themselves, but by the light reactions, which generate cellular energy and reducing power for carbon and nitrogen fixation (Magnuson 2019). Remembering that these results are in comparison to free-living Nostocales, this makes sense if *T. azollae* is receiving some amount of fixed carbon from the host (Kaplan and Peters 1988) and the symbiosis is not growth limited by nitrogen fixation (Brouwer et al. 2017). Our hypothesis is obviously not that nitrogen fixation is not important to the symbiosis, but that the light reactions are the rate-limiting steps. Likewise, carbon fixation may be important but not as important as the light reactions. Alternatively, carbon fixation may be maintained to facilitate the light reactions and not because carbon fixation is itself important (Thornton 2014; Braakman et al. 2017). “Glycolysis (Embden–Meyerhof pathway)” and “Pentose phosphate, oxidative phase” are KEGG modules that are involved in converting photosynthate to cellular energy and reductants, respectively (Magnuson 2019); as with all other modules, neither differed significantly from the genome-wide percentage of drift-affected orthogroups.

The importance of ribosomes and the light reactions are supported by other analyses. Although we find no general correlation between an orthogroup containing pseudogenes and being drift-affected, ribosomal proteins and proteins from the combination of “Photosynthesis” and “Photosynthesis—antenna proteins” are significantly less likely to be pseudogenes. Furthermore, none of the missing Nostocales core genes belong to either of these functions.

We have thus answered our motivating questions: the *T. azollae* genomes are evolving under the influence of both drift and selection, and strong purifying selection on the light reactions drives the retention of photosynthesis.

Evolutionary Convergence of Cyanobacterial Symbionts Depends on Compartmentalization of Photosynthesis and Nitrogen Fixation

The *Azolla*-*T. azollae* symbiosis is one of at least four independently evolved vertically inherited symbioses between a diazotrophic cyanobacteria and a photosynthetic eukaryotic host. This allows a comparative approach to understanding why photosynthesis experiences strong purifying selection in *T. azollae* when photosynthesis has been lost in some of these comparable symbionts. Unlike *T. azollae*, the other three symbionts, *UCYN-A*, spheroid bodies, and *Richelia euintracellularis*, live intracellularly within non-Viridiplantae algal hosts (haptophyte algae including *Braarudosphaera bigelowii*, diatoms of the family Rhopalodiaceae, and *Hemiaulus hauckii*, respectively) (Tripp et al. 2010; Hilton et al. 2013; Nakayama and Inagaki 2017; Caputo et al. 2019; Nieves-Morión et al. 2023). All of these symbionts, including *T. azollae*, appear to receive photosynthate from their hosts (Ray et al. 1979; Kaplan and Peters 1988; Nakayama et al. 2014; Foster et al. 2022). While precise dating of the onsets of these symbioses is difficult, all are thought to be in the range of ∼50 to 100Ma (Metzgar and Pryer 2007; Cornejo-Castillo et al. 2016; Testo and Sundue 2016; Caputo et al. 2019), except for spheroid bodies, which might be as young as ∼25 Ma (Caputo et al. 2019) or even ∼12 Ma (Nakayama et al. 2011). *R. euintracellularis* and *T. azollae* retain full photosynthesis while spheroid bodies have lost photosystem I, photosystem II, and carbon fixation, and *UCYN-A* retains only photosystem I (Ray et al. 1979; Kaplan and Peters 1988; Jahson et al. 1995; Zehr et al. 2008; Nakayama and Inagaki 2017; Nieves-Morión et al. 2020; Flores et al. 2022).

Strikingly, *R. euintracellularis* and *T. azollae*, both of which retain photosynthesis, are heterocystous Nostocales, while *UCYN-A* and spheroid bodies, each of which have lost at least photosystem II and carbon fixation, are unicellular representatives of other cyanobacterial clades. The fact that retention of photosynthesis by the symbiont correlates with whether the symbiont is a heterocystous Nostocales or unicellular, and does not correlate with whether the host is a multicellular land plant or a unicellular alga, nor with the location of the symbiont relative to the host (intracellular vs. extracellular), strongly indicates that retention of photosynthesis is at least partly determined by whether the symbiont is a heterocystous Nostocales. Importantly, other heterocystous Nostocales that form facultative symbioses with embryophytes fully downregulate photosynthesis to live heterotrophically off photosynthate from the host (Rai et al. 2000). This suggests that loss of photosynthesis would not be fatal for *T. azollae* and *R. euintracellularis* and that there must be a selective advantage to retaining photosynthesis that does not exist for unicellular symbionts.

We believe that the presence of heterocysts in *T. azollae* and *R. euintracellularis* has facilitated the retention of photosynthesis. Nitrogenase, the enzyme that fixes nitrogen, is extremely sensitive to O_2, so diazotrophic cyanobacteria must separate nitrogen fixation from photosynthesis, particularly O_2-evolving photosystem II and carbon fixation to which it is closely tied (Kumar et al. 2010). Heterocystous cyanobacteria separate the two processes spatially, with nitrogen fixation occurring in specialized cells called heterocysts that lack carbon fixation and active photosystem II (Zeng and Zhang 2022). Unicellular diazotrophic cyanobacteria instead achieve separation temporally by fixing nitrogen only at night (Kumar et al. 2010). Thus, there may be a selective advantage to losing photosystem II and carbon fixation that is unique to a unicellular diazo-cyanobiont: relying solely on photosynthate from the host allows them to fix nitrogen during the day when the host is photosynthesizing. Indeed, *UCYN-A* and spheroid bodies fix nitrogen during the day (Gradoville et al. 2021, Abresch et al. 2024) and *UCYN-A* maintains photosystem I which does not generate O_2 (Zehr et al. 2008). Diurnal nitrogen fixation is advantageous because it allows direct transfer of ATP and reductant from photosynthesis, unlike nocturnal nitrogen fixation, which requires photosynthate to be stored and later broken down (Scherer et al. 1988). Additionally, the lack of differentiated cell types in the unicellular diazocyanobionts may facilitate the transition to a simpler, nitrogen-fixing pseudo-organelle compared with a filamentous Nostocales, which, for instance, requires vegetative cells to reproduce but those vegetative cells cannot fix nitrogen (Zeng and Zhang 2022).

This comparative analysis suggests that the inability of unicellular cyanobacteria to spatially separate nitrogen fixation and photosynthesis within themselves creates a selective advantage, possibly in terms of the cellular energy economy, to separating those processes by losing photosynthesis and outsourcing it to the host. Our molecular evolutionary analysis revealed strong purifying selection acting on the light reactions of photosynthesis in *T. azollae*, suggesting that *T. azollae* is most often limited for cellular energy and reducing power, which feed both carbon and nitrogen fixation. More work is needed to develop detailed models of these tradeoffs to understand the optimum, most efficient metabolic strategies of these symbionts, as well as to quantify the symbionts' photosynthetic metabolism under various conditions.

Supplementary Material

Supplementary material is available at *Molecular Biology and Evolution* online.

Acknowledgments

We thank all who were involved in producing the genomes we used and in making them publicly available. Thank you, Heidi Abresch, for illuminating conversations about diazocyanobionts of algae. Thank you, Jeffrey C. Cameron, for inspiring us to work on the *Azolla*—*T. azollae* symbiosis, for your enthusiasm for all things cyanobacteria, and for your friendship.

Author Contributions

L.F., K.K., B.W., and N.C.K. conceived the project. L.F., K.K., and A.I.G. developed the bioinformatic pipeline and analyses with input from J.P.M. and N.C.K. L.F. analyzed the data. L.F. wrote the manuscript with input, editing, and revisions from all.

Funding

This work was funded by a "Future Investigators in NASA Earth and Space Science and Technology" fellowship (#80NSSC23K1389) from the National Aeronautics and Space Administration, awarded to L.F. and B.W., and by a grant from the Interdisciplinary Consortia for Astrobiology Research from the National Aeronautics and Space Administration (#80NSSC23K1357) awarded to J.P.M. and B.W. Publication of this article was funded by the University of Colorado Boulder Libraries Open Access Fund.

Data Availability

Accessions for all genomes used, including MAGs, are in supplementary table S1, Supplementary Material online. Specific commands used for different programs are available in supplementary methods, Supplementary Material online and all code is publicly available on GitHub (https://github.com/KaneLab/Tazollae-selection).

References

Abresch H, Bell T, Miller SR. Diurnal transcriptional variation is reduced in a nitrogen-fixing diatom endosymbiont. *ISME J*. 2024:18(1):wrae064. https://doi.org/10.1093/ismejo/wrae064.

Adams DG, Bergman B, Nierzwicki-Bauer SA, Duggan PS, Rai AN, Schüßler A. Cyanobacterial-plant symbioses. In: Rosenberg E, DeLong EF, Lory S, Stackebrandt E, Thompson F, editors. *The prokaryotes*. Heidelberg: Springer Berlin Heidelberg; 2013.

Aguilera A, Gómez EB, Kaštovský J, Echenique RO, Salerno GL. The polyphasic analysis of two native *Raphidiopsis* isolates supports the unification of the genera *Raphidiopsis* and *Cylindrospermopsis* (Nostocales, Cyanobacteria). *Phycologia*. 2018:57(2):130–146. https://doi.org/10.2216/17-2.1.

Arkin AP, Cottingham RW, Henry CS, Harris NL, Stevens RL, Maslov S, Dehal P, Ware D, Perez F, Canon S, *et al*. KBase: the United States department of energy systems biology knowledgebase. *Nat Biotechnol*. 2018:36(7):566–569. https://doi.org/10.1038/nbt.4163.

Bauer CC, Scappino L, Haselkorn R. Growth of the cyanobacterium Anabaena on molecular nitrogen: nifJ is required when iron is limited. *Proc Natl Acad Sci U S A*. 1993:90(19):8812–8816. https://doi.org/10.1073/pnas.90.19.8812.

Becking JH. Endophyte transmission and activity in the Anabaena-Azolla association. In: Van Diest A, editor. *Plant and soil interfaces and interactions*. Dordrecht: Springer Netherlands; 1987. p. 183–212.

Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Stat Methodol*. 1995:57(1):289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Bobay L-M, Ochman H. The evolution of bacterial genome architecture. *Front Genet*. 2017:8:72. https://doi.org/10.3389/fgene.2017.00072.

Braakman R, Follows MJ, Chisholm SW. Metabolic evolution and the self-organization of ecosystems. *Proc Natl Acad Sci U S A*. 2017:114(15):E3091–E3100. https://doi.org/10.1073/pnas.1619573114.

Brouwer P, Bräutigam A, Buijs VA, Tazelaar AOE, van der Werf A, Schlüter U, Reichart G-J, Bolger A, Usadel B, Weber APM, *et al*. Metabolic adaptation, a specialized leaf organ structure and vascular responses to diurnal N2 fixation by nostoc Azollae sustain the astonishing productivity of Azolla ferns without nitrogen fertilizer. *Front Plant Sci*. 2017:8:442. https://doi.org/10.3389/fpls.2017.00442.

Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods*. 2021:18(4):366–368. https://doi.org/10.1038/s41592-021-01101-x.

Buschiazzo E, Ritland C, Bohlmann J, Ritland K. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. *BMC Evol Biol*. 2012:12(1):8. https://doi.org/10.1186/1471-2148-12-8.

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale. *Mol Biol Evol*. 2021:38(12):5825–5829. https://doi.org/10.1093/molbev/msab293.

Caputo A, Nylander JAA, Foster RA. The genetic diversity and evolution of diatom-diazotroph associations highlights traits favoring symbiont integration. *FEMS Microbiol Lett*. 2019:366(2):fny297. https://doi.org/10.1093/femsle/fny297.

Charlesworth B. Effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 2009:10(3):195–205. https://doi.org/10.1038/nrg2526.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk v2: memory friendly classification with the genome taxonomy database. *Bioinformatics*. 2022:38(23):5315–5316. https://doi.org/10.1093/bioinformatics/btac672.

Chaux F, Jarrige D, Rodrigues-Azevedo M, Bujaldon S, Caspari OD, Ozawa S-I, Drapier D, Vallon O, Choquet Y, de Vitry C, *et al.* Chloroplast ATP synthase biogenesis requires peripheral stalk subunits AtpF and ATPG and stabilization of *atpE* mRNA by OPR protein MDE1. *Plant J.* 2023:116(6):1582–1599. https://doi.org/10.1111/tpj.16448.

Chivian D, Jungbluth SP, Dehal PS, Wood-Charlson EM, Canon RS, Allen BH, Clark MM, Gu T, Land ML, Price GA, *et al.* Metagenome-assembled genome extraction and analysis from microbiomes using KBase. *Nat Protoc.* 2023:18(1):208–238. https://doi.org/10.1038/s41596-022-00747-x.

Coale TH, Loconte V, Turk-Kubo KA, Vanslembrouck B, Mak WKE, Cheung S, Ekman A, Chen J-H, Hagino K, Takano Y, *et al.* Nitrogen-fixing organelle in a marine alga. *Science.* 2024: 384(6692):217–222. https://doi.org/10.1126/science.adk1075.

Copley SD. Evolution of new enzymes by gene duplication and divergence. *FEBS J.* 2020:287(7):1262–1283. https://doi.org/10.1111/febs.15299.

Cornejo-Castillo FM, Cabello AM, Salazar G, Sánchez-Baracaldo P, Lima-Mendez G, Hingamp P, Alberti A, Sunagawa S, Bork P, de Vargas C, *et al.* Cyanobacterial symbionts diverged in the late cretaceous towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat Commun.* 2016:7(1):11071. https://doi.org/10.1038/ncomms11071.

Crockford PW, Bar On YM, Ward LM, Milo R, Halevy I. The geologic history of primary productivity. *Curr Biol.* 2023:33(21): 4741–4750.e5. https://doi.org/10.1016/j.cub.2023.09.040.

Davies-Barnard T, Friedlingstein P. The global distribution of biological nitrogen fixation in terrestrial natural ecosystems. *Global Biogeochem Cycles.* 2020:34(3):e2019GB006387. https://doi.org/10.1029/2019GB006387.

Dijkhuizen LW, Tabatabaei BES, Brouwer P, Rijken N, Buijs VA, Güngör E, Schluepmann H. Far-red light-induced Azolla Filiculoides symbiosis sexual reproduction: responsive transcripts of symbiont nostoc Azollae encode transporters whilst those of the fern relate to the angiosperm floral transition. *Front Plant Sci.* 2021:12:693039. https://doi.org/10.3389/fpls.2021.693039.

Dojun N, Muranishi K, Ishimori K, Uchida T. A single mutation converts Alr5027 from Cyanobacteria nostoc Sp. PCC 7120 to a heme-binding protein with heme-degrading ability. *J Inorg Biochem.* 2020:203:110916. https://doi.org/10.1016/j.jinorgbio.2019.110916.

Duggan PS, Thiel T, Adams DG. Symbiosis between the cyanobacterium nostoc and the liverwort Blasia requires a CheR-type MCP methyltransferase. *Symbiosis.* 2013:59(2):111–120. https://doi.org/10.1007/s13199-012-0216-9.

Dykhuizen DE, Dean AM. Enzyme activity and fitness: evolution in solution. *Trends Ecol Evol.* 1990:5(8):257–262. https://doi.org/10.1016/0169-5347(90)90067-N.

Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* 2019:20(1):238. https://doi.org/10.1186/s13059-019-1832-y.

Fani R, Gallo R, Liò P. Molecular evolution of nitrogen fixation: the evolutionary history of the nifD, nifK, nifE, and nifN genes. *J Mol Evol.* 2000:51(1):1–11. https://doi.org/10.1007/s002390010061.

Fischer WW, Hemp J, Johnson JE. Evolution of oxygenic photosynthesis. *Annu Rev Earth Planet Sci.* 2016:44(1):647–683. https://doi.org/10.1146/annurev-earth-060313-054810.

Flores E, Romanovicz DK, Nieves-Morión M, Foster RA, Villareal TA. Adaptation to an intracellular lifestyle by a nitrogen-fixing, heterocyst-forming cyanobacterial endosymbiont of a diatom. *Front Microbiol.* 2022:13:799362. https://doi.org/10.3389/fmicb.2022.799362.

Foster RA, Tienken D, Littmann S, Whitehouse MJ, Kuypers MMM, White AE. The rate and fate of N2 and C fixation by marine diatom-diazotroph symbioses. *ISME J.* Feb. 2022:16(2):477–487. https://doi.org/10.1038/s41396-021-01086-7.

Gabr A, Grossman AR, Bhattacharya D. *Paulinella*, a model for understanding plastid primary endosymbiosis. *J Phycol.* 2020:56(4): 837–843. https://doi.org/10.1111/jpy.13003.

Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J.* 2014:8(8): 1553–1565. https://doi.org/10.1038/ismej.2014.60.

Gradoville MR, Cabello AM, Wilson ST, Turk-Kubo KA, Karl DM, Zehr JP. Light and depth dependency of nitrogen fixation by the non-photosynthetic, symbiotic cyanobacterium UCYN-A. *Environ Microbiol.* 2021:23(8):4518–4531. https://doi.org/10.1111/1462-2920.15645.

Graf JS, Schorn S, Kitzinger K, Ahmerkamp S, Woehle C, Huettel B, Schubert CJ, Kuypers MMM, Milucka J. Anaerobic endosymbiont generates energy for Ciliate host by denitrification. *Nature.* 2021:591(7850):445–450. https://doi.org/10.1038/s41586-021-03297-6.

Gray MW, Archibald JM. Origins of mitochondria and plastids. In: Bock R, Knoop V, editors. *Genomics of chloroplasts and mitochondria.* Dodrecht: Springer Netherlands; 2012. p. 1–30.

Guo Y, Guo G, Mao X, Zhang W, Xiao J, Tong W, Liu T, Xiao B, Liu X, Feng Y, *et al.* Functional identification of HugZ, a heme oxygenase from Helicobacter Pylori. *BMC Microbiol.* 2008:8(1):226. https://doi.org/10.1186/1471-2180-8-226.

Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li W, Chitsaz F, Derbyshire MK, Gonzales NR, *et al.* RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.* 2018:46(D1):D851–D860. https://doi.org/10.1093/nar/gkx1068.

Hilton JA, Foster RA, James Tripp H, Carter BJ, Zehr JP, Villareal TA. Genomic deletions disrupt nitrogen metabolism pathways of a cyanobacterial diatom symbiont. *Nat Commun.* 2013:4(1):1767. https://doi.org/10.1038/ncomms2748.

Holst RW, Yopp JH. Effect of Various nitrogen sources on growth and the nitrate-nitrite reductase system of the Azolla Mexicana-Anabaena Azollae symbiosis. *Aquat Bot.* 1979:7:359–367. https://doi.org/10.1016/0304-3770(79)90035-4.

Huo D, Li H, Cai F, Guo X, Qiao Z, Wang W, Yu G, Li R. Genome evolution of filamentous cyanobacterium nostoc Species: from facultative symbiosis to free living. *Microorganisms.* 2021:9(10):2015. https://doi.org/10.3390/microorganisms9102015.

Inomura K, Deutsch C, Masuda T, Prášil O, Follows MJ. Quantitative models of nitrogen-fixing organisms. *Comput Struct Biotechnol J.* 2020:18:3905–3924. https://doi.org/10.1016/j.csbj.2020.11.022.

Jahson S, Rai AN, Bergman B. Intracellular cyanobiont richelia Intracellularis: ultrastructure and immuno-localisation of phycoerythrin, nitrogenase, rubisco and glutamine synthetase. *Mar Biol.* 1995:124(1):1–8. https://doi.org/10.1007/BF00349140.

Kanehisa M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* 2019:28(11):1947–1951. https://doi.org/10.1002/pro.3715.

Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* 2023:51(D1):D587–D592. https://doi.org/10.1093/nar/gkac963.

Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 2000:28(1):27–30. https://doi.org/10.1093/nar/28.1.27.

Kaplan D, Peters G. Interaction of carobon metabolism in the Azolla-Anabaena symbiosis. *Symbiosis.* 1988:6:53–67.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol.* 2013:30(4):772–780. https://doi.org/10.1093/molbev/mst010.

Kosakovsky Pond S, Delport W, Muse SV, Scheffler K. Correcting the bias of empirical frequency parameter estimators in Codon models. *PLoS One.* 2010:5(7):e11230. https://doi.org/10.1371/journal.pone.0011230.

Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for Maximum likelihood phylogenetic inference. *Bioinformatics.* 2019:35(21):4453–4455. https://doi.org/10.1093/bioinformatics/btz305.

Kumar K, Mella-Herrera RA, Golden JW. Cyanobacterial heterocysts. *Cold Spring Harb Perspect Biol.* 2010:2(4):a000315. https://doi.org/10.1101/cshperspect.a000315.

Lemoine F, Gascuel O. Gotree/goalign: toolkit and go API to facilitate the development of phylogenetic workflows. *NAR Genom Bioinform*. 2021:3(3):lqab075. https://doi.org/10.1093/nargab/lqab075.

Li F-W, Brouwer P, Carretero-Paulet L, Cheng S, de Vries J, Delaux P-M, Eily A, Koppers N, Kuo L-Y, Li Z, *et al.* Fern genomes elucidate land plant evolution and cyanobacterial symbioses. *Nat Plants*. 2018:4(7):460–472. https://doi.org/10.1038/s41477-018-0188-8.

Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvernin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, *et al.* RefSeq: expanding the prokaryotic genome annotation pipeline reach with protein family model curation. *Nucleic Acids Res*. 2021:49(D1): D1020–D1028. https://doi.org/10.1093/nar/gkaa1105.

Magnuson A. Heterocyst thylakoid bioenergetics. *Life*. 2019:9(1):13. https://doi.org/10.3390/life9010013.

Mao G, Zhao Y, Kang X, Li Z, Zhang Y, Wang X, Sun F, Sankaran K, Zhang XC. Crystal structure of *E. Coli* lipoprotein diacylglyceryl transferase. *Nat Commun*. 2016:7(1):10198. https://doi.org/10.1038/ncomms10198.

McCutcheon JP, Boyd BM, Dale C. The life of an insect endosymbiont from the cradle to the grave. *Curr Biol*. 2019:29(11):R485–R495. https://doi.org/10.1016/j.cub.2019.03.032.

McCutcheon JP, Garber AI, Spencer N, Warren JM. How do bacterial endosymbionts work with so few genes? *PLoS Biol*. 2024:22(4): e3002577. https://doi.org/10.1371/journal.pbio.3002577.

McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic Bacteria. *Nat Rev Microbiol*. 2012:10(1):13–26. https://doi.org/10.1038/nrmicro2670.

McFadden GI. Chloroplast origin and integration. *Plant Physiol*. 2001:125(1):50–53. https://doi.org/10.1104/pp.125.1.50.

Melnikov SV, van den Elzen A, Stevens DL, Thoreen CC, Söll D. Loss of protein synthesis quality control in host-restricted organisms. *Proc Natl Acad Sci U S A*. 2018:115(49):E11505–E11512. https://doi.org/10.1073/pnas.1815992115.

Mendel RR. The molybdenum cofactor. *J Biol Chem*. 2013:288(19): 13165–13172. https://doi.org/10.1074/jbc.R113.455311.

Metzgar JS, Pryer KM. Phylogeny and divergence time estimates for the fern genus *Azolla* (Salviniaceae). *Int J Plant Sci*. 2007:168(7): 1045–1053. https://doi.org/10.1086/519007.

Mira A, Moran NA. Estimating population size and transmission bottlenecks in maternally transmitted endosymbiotic Bacteria. *Microb Ecol*. Sept. 2002:44(2):137–143. https://doi.org/10.1007/s00248-002-0012-9.

Mishra D, Saraf A, Kumar N, Pal S, Singh P. Issues in cyanobacterial taxonomy: comprehensive case study of unbranched, false branched and true branched heterocytous Cyanobacteria. *FEMS Microbiol Lett*. 2021:368(4):fnab005. https://doi.org/10.1093/femsle/fnab005.

Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*. 2008:42(1):165–190. https://doi.org/10.1146/annurev.genet.41.110306.130119.

Moreno-Cabezuelo JÁ, López-Lozano A, Díez J, García-Fernández JM. Differential expression of the glucose transporter gene *glcH* in response to glucose and light in marine picocyanobacteria. *PeerJ*. 2019:6:e6248. https://doi.org/10.7717/peerj.6248.

Mugal CF, Wolf JBW, Kaj I. Why time matters: codon evolution and the temporal dynamics of dN/dS. *Mol Biol Evol*. Jan. 2014:31(1): 212–231. https://doi.org/10.1093/molbev/mst192.

Nakayama T, Ikegami Y, Nakayama T, Ishida K, Inagaki Y, Inouye I. Spheroid bodies in rhopalodiacean diatoms were derived from a single endosymbiotic cyanobacterium. *J Plant Res*. 2011:124(1): 93–97. https://doi.org/10.1007/s10265-010-0355-0.

Nakayama T, Inagaki Y. Genomic divergence within non-photosynthetic cyanobacterial endosymbionts in rhopalodiacean diatoms. *Sci Rep*. 2017:7(1):13075. https://doi.org/10.1038/s41598-017-13578-8.

Nakayama T, Kamikawa R, Tanifuji G, Kashiyama Y, Ohkouchi N, Archibald JM, Inagaki Y. Complete genome of a nonphotosynthetic cyanobacterium in a diatom reveals recent adaptations to an intracellular lifestyle. *Proc Natl Acad Sci U S A*. 2014:111(31): 11407–11412. https://doi.org/10.1073/pnas.1405222111.

Nieves-Morión M, Camargo S, Bardi S, Ruiz MT, Flores E, Foster RA. Heterologous expression of genes from a cyanobacterial endosymbiont highlights substrate exchanges with its diatom host. *PNAS Nexus*. 2023:2(6):pgad194. https://doi.org/10.1093/pnasnexus/pgad194.

Nieves-Morión M, Flores E, Foster RA. Predicting substrate exchange in marine diatom-heterocystous Cyanobacteria symbioses. *Environ Microbiol*. 2020:22(6):2027–2052. https://doi.org/10.1111/1462-2920.15013.

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, *et al.* Reference sequence (RefSeq) database at NCBI: current Status, taxonomic expansion, and functional annotation. *Nucleic Acids Res*. 2016: 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. Checkm: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res*. 2015:25(7):1043–1055. https://doi.org/10.1101/gr.186072.114.

Pereira AL, Vasconcelos V. Classification and phylogeny of the cyanobiont Anabaena Azollae strasburger: an answered question? *Int J Syst Evol Microbiol*. 2014:64(Pt_6):1830–1840. https://doi.org/10.1099/ijs.0.059238-0.

Peters GA, Mayne BC. The *Azolla, Anabaena Azollae* relationship: I. Initial characterization of the association. *Plant Physiol*. 1974:53(6):813–819. https://doi.org/10.1104/pp.53.6.813.

Price MN, Dehal PS, Arkin AP. FastTree 2—approximately Maximum-likelihood trees for large alignments. *PLoS One*. 2010:5(3):e9490. https://doi.org/10.1371/journal.pone.0009490.

Rai AN, Söderbäck E, Bergman B. Tansley review No. 116: cyanobacterium–plant symbioses. *New Phytol*. 2000:147(3):449–481. https://doi.org/10.1046/j.1469-8137.2000.00720.x.

Ran L, Larsson J, Vigil-Stenman T, Nylander JAA, Ininbergs K, Zheng W-W, Lapidus A, Lowry S, Haselkorn R, Bergman B. Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PLoS One*. 2010:5(7):e11486. https://doi.org/10.1371/journal.pone.0011486.

Rangel-Chávez CP, Galán-Vásquez E, Pescador-Tapia A, Delaye L, Martínez-Antonio A. RNA polymerases in strict endosymbiont Bacteria with extreme genome reduction show distinct erosions that might result in limited and differential promoter recognition. *PLoS One*. 2021:16(7):e0239350. https://doi.org/10.1371/journal.pone.0239350.

Ranwez V, Douzery EJP, Cambon C, Chantret N, Delsuc F. MACSE v2: toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. *Mol Biol Evol*. 2018:35(10):2582–2584. https://doi.org/10.1093/molbev/msy159.

Ray TB, Mayne BC, Toia RE, Peters GA. *Azolla-Anabaena* relationship: VIII. Photosynthetic characterization of the association and individual partners. *Plant Physiol*. 1979:64(5):791–795. https://doi.org/10.1104/pp.64.5.791.

R Core Team. *R: a language and environment for statistical computing*. R Foundation for Statistical Computing; 2023 [accessed 2025]. https://www.R-project.org/.

Ringel P, Krausze J, van den Heuvel J, Curth U, Pierik AJ, Herzog S, Mendel RR, Kruse T. Biochemical characterization of molybdenum cofactor-free nitrate reductase from neurospora Crassa. *J Biol Chem*. May 2013:288(20):14657–14671. https://doi.org/10.1074/jbc.M113.457960.

Rocha EPC, Feil EJ. Mutational patterns cannot explain genome composition: are there any neutral sites in the genomes of Bacteria? *PLoS Genet*. 2010:6(9):e1001104. https://doi.org/10.1371/journal.pgen.1001104.

Romine MF, Rodionov DA, Maezato Y, Anderson LN, NandhikondaP, Rodionova IA, Carre A, Li X, Xu C, Clauss TRW, *et al.* Elucidation of roles for vitamin B_{12} in regulation of folate, ubiquinone, and methionine metabolism. *Proc Natl Acad Sci U S A*. 2017:114(7):E1205–E1214. https://doi.org/10.1073/pnas.1612360114.

Rosenberg E, Zilber-Rosenberg I. Symbiosis and development: the hologenome concept. *Birth Defects Res C Embryo Today*. 2011:93(1): 56–66. https://doi.org/10.1002/bdrc.20196.

Rubio LM, Flores E, Herrero A. The *narA* locus of *Synechococcus* Sp. Strain PCC 7942 consists of a cluster of molybdopterin biosynthesis genes. *J Bacteriol*. 1998:180(5):1200–1206. https://doi.org/10.1128/JB.180.5.1200-1206.1998.

Salem H, Bauer E, Kirsch R, Berasategui A, Cripps M, Weiss B, Koga R, Fukumori K, Vogel H, Fukatsu T, *et al*. Drastic genome reduction in an Herbivore's pectinolytic symbiont. *Cell*. 2017:171(7):1520–1531.e13. https://doi.org/10.1016/j.cell.2017.10.029.

Salem H, Florez L, Gerardo N, Kaltenpoth M. An out-of-body experience: the extracellular dimension for the transmission of mutualistic Bacteria in insects. *Proc R Soc Lond B Biol Sci*. 2015:282(1804):20142957. https://doi.org/10.1098/rspb.2014.2957.

Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S, *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res*. 2022:50(D1):D20–D26. https://doi.org/10.1093/nar/gkab1112.

Scherer S, Almon H Böger P. Interaction of photosynthesis, respiration and nitrogen fixation in Cyanobacteria. *Photosynth Res*. 1988:15(2):95–114. https://doi.org/10.1007/BF00035255.

Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics*. 2014:30(14):2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

Song MJ, Freund F, Tribble CM, Toffelmier E, Miller C, Bradley SH, Li F-W, Rothfels CJ. The nitrogen-fixing fern *Azolla* has a complex microbiome characterized by varying degrees of cophylogenetic signal. *Am J Bot*. 2025:112(3). https://doi.org/10.1002/ajb2.70010.

Steiner JM, Pompe JA, Löffelhardt W. Characterization of apcC, the nuclear gene for the phycobilisome core linker polypeptide L c 7.8 from the glaucocystophyte alga Cyanophora Paradoxa. Import of the precursor into isolated cyanelles and integration of the mature protein into intact phycobilisomes. *Curr Genet*. Nov. 2003:44(3):132–137. https://doi.org/10.1007/s00294-003-0433-0.

Syberg-Olsen MJ, Garber AI, Keeling PJ, McCutcheon JP, Husnik F. Pseudofinder: detection of pseudogenes in prokaryotic genomes. *Mol Biol Evol*. July 2022:39(7):msac153. https://doi.org/10.1093/molbev/msac153.

Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res*. 2016:44(14):6614–6624. https://doi.org/10.1093/nar/gkw569.

Tenebaum D; Bioconductor Package Maintainer. 2023. KEGGREST: Client-Side REST Access to the Kyoto Encyclopedia of Genes and Genomes (KEGG). R package version 1.42.0. [*Computer software*]. https://bioconductor.org/packages/KEGGREST. *DOI.org* (*Datacite*). https://doi.org/10.18129/B9.BIOC.KEGGREST.

Testo W, Sundue M. A 4000-Species dataset provides new insight into the evolution of ferns. *Mol Phylogenet Evol*. 2016:105:200–211. https://doi.org/10.1016/j.ympev.2016.09.003.

Thornton DCO. Dissolved organic matter (DOM) release by phytoplankton in the contemporary and future ocean. *Eur J Phycol*. 2014:49(1):20–46. https://doi.org/10.1080/09670262.2013.875596.

Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, Affourtit JP, Zehr JP. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*. 2010:464(7285):90–94. https://doi.org/10.1038/nature08786.

Tschitschko B, Esti M, Philippi M, Kidane AT, Littmann S, Kitzinger K, Speth DR, Li S, Kraberg A, Tienken D, *et al*. Rhizobia–diatom symbiosis fixes missing nitrogen in the ocean. *Nature*. 2024:630(8018):899–904. https://doi.org/10.1038/s41586-024-07495-w.

Van Leuven JT, McCutcheon JP. An AT mutational bias in the tiny GC-rich endosymbiont genome of hodgkinia. *Genome Biol Evol*. 2012:4(1):24–27. https://doi.org/10.1093/gbe/evr125.

Videau P, Rivers OS, Ushijima B, Oshiro RT, Kim MJ, Philmus B, Cozy LM. Mutation of the *murC* and *murB* genes impairs heterocyst differentiation in Anabaena Sp. Strain PCC 7120. *J Bacteriol*. 2016:198(8):1196–1206. https://doi.org/10.1128/JB.01027-15.

Warshan D, Bay G, Nahar N, Wardle DA, Nilsson M-C, Rasmussen U. Seasonal variation in *nifH* abundance and expression of cyanobacterial communities associated with boreal feather mosses. *ISME J*. 2016:10(9):2198–2208. https://doi.org/10.1038/ismej.2016.17.

Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: detecting relaxed selection in a phylogenetic framework. *Mol Biol Evol*. 2015:32(3):820–832. https://doi.org/10.1093/molbev/msu400.

Wolf JBW, Künstner A, Nam K, Jakobsson M, Ellegren H. Nonlinear dynamics of nonsynonymous (dN) and synonymous (dS) substitution rates affects inference of selection. *Genome Biol Evol*. 2009:1:308–319. https://doi.org/10.1093/gbe/evp030.

Zehr JP, Bench SR, Carter BJ, Hewson I, Niazi F, Shi T, Tripp HJ, Affourtit JP. Globally distributed uncultivated oceanic N_2 -fixing Cyanobacteria lack oxygenic photosystem II. *Science*. 2008:322(5904):1110–1112. https://doi.org/10.1126/science.1165340.

Zehr JP, Capone DG. Changing perspectives in marine nitrogen fixation. *Science*. 2020:368(6492):eaay9514. https://doi.org/10.1126/science.aay9514.

Zeng X, Zhang C-C. The making of a heterocyst in Cyanobacteria. *Annu Rev Microbiol*. 2022:76(1):597–618. https://doi.org/10.1146/annurev-micro-041320-093442.

Zhang S, Bryant DA. The tricarboxylic acid cycle in Cyanobacteria. *Science*. 2011:334(6062):1551–1553. https://doi.org/10.1126/science.1210858.