Ecological constraints and evolutionary trade-offs shape nitrogen fixation across habitats

Morgan S. Sobol[1,2*], Cécile Ané[3,4,] Katherine D. McMahon[5] and Betül Kaçar[1*]

1. Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA
2. Department of Biology, Texas State University, San Marcos, TX, USA
3. Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA
4. Department of Botany, University of Wisconsin-Madison, Madison, WI, USA
5. Department of Civil and Environmental Engineering, University of Wisconsin–Madison, Madison, WI, USA

Correspondence: msobol@txstate.edu, bkacar@wisc.edu

Abstract

From its earlier beginnings, life's expansion into new habitats has been profoundly shaped by its reciprocal interactions with Earth's changing environments. Understanding how ancient metabolisms co-evolved with their environments require uncovering the ecological and evolutionary processes that structured the functionally linked genes and networks underlying these metabolisms. Here, we focus on nitrogen (N_2) fixation, one of life's most critical metabolisms to all known life, and investigate the drivers of complexity in its associated gene machinery today. We used a large-scale comparative genomics framework to construct a comprehensive catalog of extant N_2 fixation-associated genes and assessed their distribution across diverse microbial genomes and environmental backgrounds. Genomes enriched in N_2 fixation genes generally have larger genome sizes, broader metabolic capabilities, wider habitat ranges, and are predominantly associated with mesophilic and aerobic lifestyles. Evolutionary reconstructions reveal a pattern of early gene gains in ancestral diazotrophs followed by lineage specific gene losses in later diverging taxa, suggesting evolutionary trade-offs shaped by changing environments. These findings demonstrate that the evolution of N_2 fixation has been

intertwined with the composition and organization of the genes supporting the overarching N_2 metabolism, driven by feedbacks between genome evolution and shifting environmental and ecological conditions.

Introduction

Biological nitrogen fixation ("N_2 fixation") is an ancient metabolism tightly coupled to Earth's geosphere [1–6]. Nitrogenases, catalytically responsible for reducing N_2 to NH_3, are the only known enzymes capable of this reaction found in microorganisms, collectively known as diazotrophs [7]. By supplying essential nitrogen for life, nitrogenase has been transformative for biological productivity and Earth's environment [1, 2, 8–10].

While the global significance of diazotrophs is well established through studies of marine and freshwater cyanobacteria [11] [12], and plant-associated symbioses [13], the factors driving the evolution of N_2 fixation in both historical and modern contexts remain unclear [14]. It's been revealed that despite nitrogenase's ~3 billion year evolutionary history, its core structural and functional features have remained largely conserved over time [4, 15–17]. In contrast, numerous genes involved in its function and regulation have diversified [3, 18].

The function of nitrogenase is orchestrated by a myriad of components in the cell. Nitrogenase is a metalloenzyme found in three different forms, differentiated by their metal cofactors, molybdenum (Mo), vanadium (V), or iron (Fe) [19–21]. Of these, is Mo-nitrogenase (*nif*), the most well-studied and catalytically efficient [22], encoded by the *nifHDK* structural genes. Nitrogenase activity depends on additional proteins for cofactor assembly, including maturases (*nifBEN*) [23] as well as numerous nif genes that support cellular functions such as electron transfer (*nifFJ*), transcription ($nifALI_1I_2$), and biosynthesis (*nifXQUVYS*) [24, 25].

Beyond the nif genes, other gene clusters contribute to N_2 fixation, including those involved in metal transport (*mod, feo, cys*), iron-sulfur cluster assembly (*suf, isc, fdx*) [25, 26], and nitrogen-carbon metabolism (*gln, nir, nap, hox*) [27, 28]. Many additional genes remain functionally uncharacterized [25, 29]. Nitrogenase interacts directly with several

of these proteins [30–32] and can indirectly with others, for example, through pathways for nitrogen and amino acid metabolism [24, 27], oxygen protection [33, 34], and electron transport [35]. **Figure 1** illustrates the genetic framework underlying nitrogenase in the model diazotroph *Azotobacter vinelandii*.

The composition of this complex genetic "machinery" varies widely across the tree of life, from a minimal gene set in some organisms, to more than 80 nitrogenase-associated genes in others [7, 25]. This variability underscores the need to examine the environmental and physiological contexts that shape nitrogenase gene distribution among taxa [3, 36].

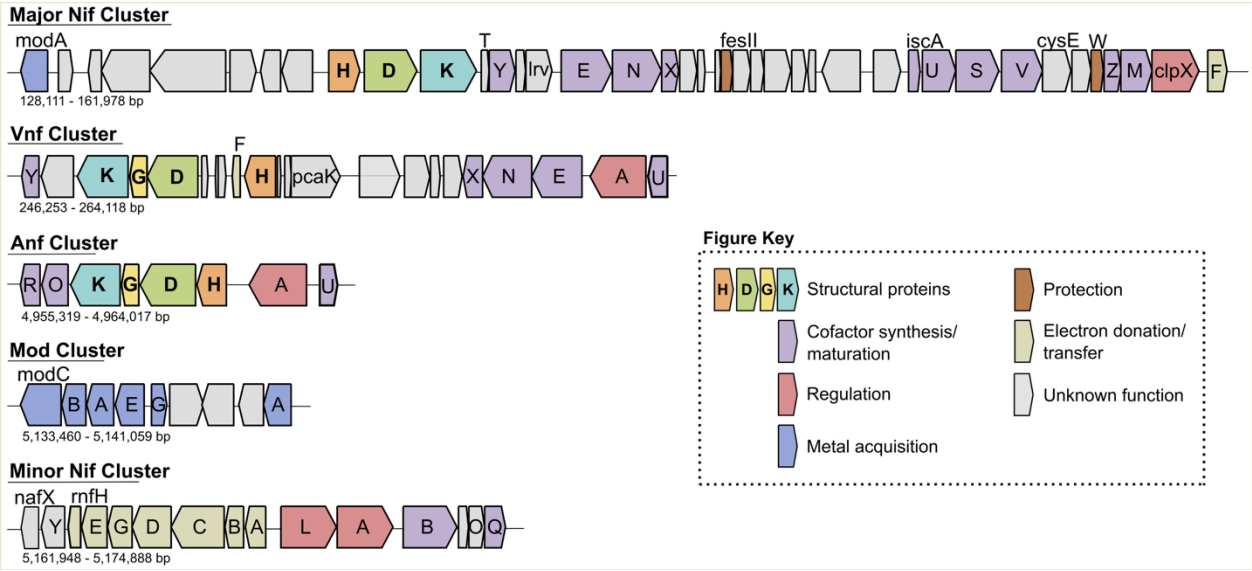


Figure 1. Structure of the major and minor molybdenum nitrogenase (Nif) clusters, vanadium nitrogenase (Vnf) and iron nitrogenase (Anf) clusters, and the molybdenum acquisition (Mod) cluster in model diazotroph *Azotobacter vinelandii* DJ. Genome positions and gene sizes for each cluster are based on KEGG genome reference T00885 for *A. vinelandii* DJ.

An integrated synthesis that brings together the diversity of N_2 fixation and its associated genes within a broader ecological and evolutionary framework would be essential for understanding how this key metabolism has been shaped across time and environments.

While previous studies have recognized the potentially dynamic nature of N_2 fixation gene acquisitions and losses [36, 37], few have examined these patterns in depth. Here, we address this gap by employing large-scale comparative genomics to explore how present-day organismal lifestyles shape evolutionary trajectories, especially in the context of varying N_2 fixation-associated gene content. Ultimately, our study reveals how genome-scale evolution underlies the expansion of diazotrophs into new and diverse ecological niches by reshaping N_2 fixation machinery over time.

Methods

Genome selection

The reference phylogeny Web of Life (WoL) [38] was selected as the basis of our study because of its extensive genome coverage and phylogenetic reconstruction. For our analysis, we kept WoL genomes with ≥90% completeness and <5% contamination, following quality thresholds established by [39].

Protein fasta files associated with the genome accessions (named "*_protein.faa.gz") were downloaded from NCBI with eutils [40] (accessed July 17th, 2023). Genomes missing available NCBI protein fasta files were excluded from analysis. This reduced the WoL genome dataset from 10,575 genomes to 7,458. Nucleotide sequences of noncoding RNA genes were downloaded if they were available at GenBank or RefSeq and the 16S rRNA genes were extracted.

Nitrogen fixation gene catalog curation

Protein fasta files were functionally annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [41]. Specifically, the translated amino acid sequences were annotated with KEGG orthology (KO) identifiers by profiling each genome against KEGG hidden Markov models (KEGG release 106) for all KOs [42] (cutoff E-value = 1e-5) using *hmmscan* in HMMER version 3.2.2 (http://hmmer.org/). Custom Python scripts were used to parse through each genome's *hmmscan* output and generate the final KO presence-absence matrix. The presence/absence matrix was subset to genes known to

be involved in N_2 fixation (and which had KO identifiers), which were collected through an extensive literature search that included foundational studies with thorough analyses of genes involved in N_2 fixation (**Supplemental Table 1**). Gene involvement was supported by functional confirmation, for example by evidence from protein interactions with nitrogenase, gene knockout phenotypes, or differential gene expression under N_2 fixation.

Species tree construction and visualization

The WoL species tree was subsetted to genomes annotated with nifHDK using the *drop.tip* function from the APE package [43] in R. We additionally confirmed the presence of nitrogenase by phylogenetic placement of concatenated nifDK sequences, as described in reference [4], from each genome. The resulting species tree, the gene presence-absence matrix, and trait data were plotted using the ggTree R package (version 3.6.2) [44].

Trait data curation

Optimal growth temperature (OGT), oxygen, pH preference, and habitat preference trait data were collected for each genome by matching the Genbank accession number to BacDive [45], NCBI BioProject [46], or GOLD JGI [47] as well as matching the NCBI taxon IDs to the Madin et al (2020) phenotypic trait database [48]. Trait data was also predicted for all genomes. OGTs were predicted using the OGT_prediction tool [49]. The regression model used included both Bacteria and Archaea, but excluded using information of the 16S rRNA gene and genome size (*superkingdom-all_species-all_features_ex_genome_size_and_rRNA.txt*). The *Oxyphen* package was used for oxygen metabolism prediction based on the number of oxygen utilizing genes in each genome [50]. Predictions for pH were done with the Ramoneda et al. (2023) prediction model which is based on genes that influence pH tolerance [51]. Habitat preference and coverage were determined with ProkAtlas [52] by mapping the 16S rRNA gene sequences to the ProkAtlas metagenome database. Barrnap version 0.9 [53] was used to extract 16S rRNA sequences for genomes with missing annotations. Using a custom python script, the longest 16S rRNA sequence was extracted for each genome as its representative 16S rRNA gene sequence to correct for errors in sequence length due to

136 sequencing and assembly bias. Twenty-one genomes without a 16S rRNA gene
137 sequence were excluded. A nucleotide identity of 97% and a query length match of at
138 least 150 base pairs were used as thresholds. Habitat preference, categorized into
139 freshwater, saline, terrestrial, plant-associated, non-plant host-associated, and
140 hydrothermal, was defined as the environment with the highest 16S abundance for each
141 genome. Habitat coverage was determined by summing up the total number of habitats
142 where a genome was considered present and dividing it by the total number of habitats
143 (n=100) investigated by ProkAtlas. The presence of KEGG modules was predicted using
144 the KEGG KO annotations with MetaPathPredict (version 0) [54]. Metabolic coverage was
145 calculated as the total number of KEGG modules present in a genome divided by the total
146 number of modules (n=194) searched for with MetaPathPredict. Data collected from
147 databases and predicted data were cross-referenced and manually annotated to provide
148 the final trait data for each genome.

149 ***Data analysis and statistics***

150 We compared the mean number of N_2 fixation genes across specific traits (oxygen,
151 temperature, and habitat preferences) with non-parametric methods using the
152 *stat_compare_means* function in the ggpubr R package (version 0.6.0) [55]. For the
153 categorical traits habit, oxygen, and temperature preference, we first applied a Kruskal-
154 Wallis test, followed by pairwise Wilcoxon rank-sum tests with Benjamini-Hochberg
155 correction. For temperature preference (categories: mesophilic and thermophilic) we
156 used the Wilcoxon rank-sum test directly. Compact letter display was used to report
157 pairwise groups whose means are discernibly different, based on a p-value of 0.05.

158 The gene presence-absence matrix was used in phylogenetic principal-component
159 analysis (PCA) performed with the *phyl.pca* function (method = “BM”, mode =”corr”) in
160 phytools R package [56]. This was done to visualize similarities and dissimilarities
161 between genomes in regard to N_2 fixation gene content while accounting for phylogenetic
162 relationships. This approach reduces the risk of over-interpreting similarities simple due
163 to common ancestry among taxa. The first two components were retained for
164 visualization. Pearson correlation between the presence-absence of N_2 fixation genes
165 was calculated with the *cor* function in the stats R package [58] and plotted using corrplot

166 [6] with hierarchical clustering and default p-value of 0.05, to identify patterns of gene co-
167 occurrence.

168 ***Evolutionary trends analysis***

169 In order to test whether certain traits shared a coevolutionary history with N_2 fixation gene
170 abundance, phylogenetic linear regression models were fit using the *phylolm* function in
171 the phylolm R package (version 2.62) [57], using Pagel's lambda model of phylogenetic
172 correlation. The lambda parameter (λ) ranges from 0 (no phylogenetic signal) to 1 (high
173 phylogenetic signal) in the response trait when the model has no covariates, or in the
174 residuals when covariates are included. Non-phylogenetic linear regressions were
175 performed using the *lm* function in the stats R package [58]. The *distRoot* function in the
176 adephylo R package [59] was used to calculate the root-to-tip patristic distances by
177 measuring the total length of branches (in terms of the number of substitutions per site)
178 from the root of the species tree to each tip.

179 Gene gain/loss was determined with Count [60] using a non-binary count matrix [61] for
180 all accessory (non-structural) Nif genes. Wagner parsimony with a gain/loss penalty of 2
181 was applied to infer the most parsimonious scenario of genes present at the internal
182 nodes, as done previously [62]. Ancestral trait reconstructions were performed on an
183 ultrametric and dichotomous tree, which was made using the *multi2di* and *chronos* (model
184 = "relaxed") functions of the APE package [62] in R. The *fitdiscrete* function in geiger [63]
185 was used for discrete reconstructions of oxygen preference, which were mapped onto the
186 phylogeny with the *simmap* function in phytools [56]. The *fastAnc* function in phytools was
187 used for continuous variable reconstructions of metabolic and habitat diversity, which was
188 mapped onto the tree using the *contMap* function in phytools [56].

189 **Results**

190 ***Distribution of nitrogenases across the tree of life***

191 We used a large prokaryote reference phylogeny, "WoL" [39], containing over 7,000 high-
192 quality genomes [36, 63–65] to first assess the representative diversity and distribution of
193 nitrogenase today. A total of 673 genomes (~6.4%) were found to contain nitrogenase

194 (**Supplemental Table 2**) based on the presence of the catalytic HDK subunits of
195 nitrogenase (**Figure 2A**). Our estimate of nitrogenase diversity falls within range of recent
196 estimates (~5.9% to 16%) [36, 64–66]. Here, approximately 46.9% of these genomes
197 belonged to Proteobacteria (in order of abundance:
198 Alpha>Delta>Gamma>Beta>Epsilon>Acidithiobacillia), while the remaining top three
199 phyla included Firmicutes (17.3%), Cyanobacteria (12.6%), and Euryarchaeota (11.4%).
200 Amongst the nitrogenase-containing genomes, molybdenum (Mo)-only nitrogenase made
201 up the majority (90.3%), followed by genomes with both Mo-nitrogenase and iron (Fe)
202 nitrogenase (5.9%), genomes with both Mo-nitrogenase and vanadium (V) nitrogenase
203 (2.2%), and lastly with 1.5% of genomes containing all three nitrogenases (**Figure 1**)
204 (**Figure 2B**). Genomes with V-nitrogenase were found in 6 phyla: Euryarchaeota,
205 Firmicutes, Alphaproteobacteria, Deltaproteobacteria, Cyanobacteria, and
206 Gammaproteobacteria. Fe-nitrogenases were found in 11 phyla: Firmicutes,
207 Epsilonproteobacteria, Alphaproteobacteria, Gammaproteobacteria, Verrucomicrobia,
208 Euryarchaeota, Deltaproteobacteria, Gammaproteobacteria, Bacteroidetes,
209 Spirochaetes, and Chlorobi. Euryarchaeota represented 50% of the genomes with all
210 three nitrogenases; the remaining phyla included Alphaproteobacteria, Firmicutes, and
211 Gammaproteobacteria.

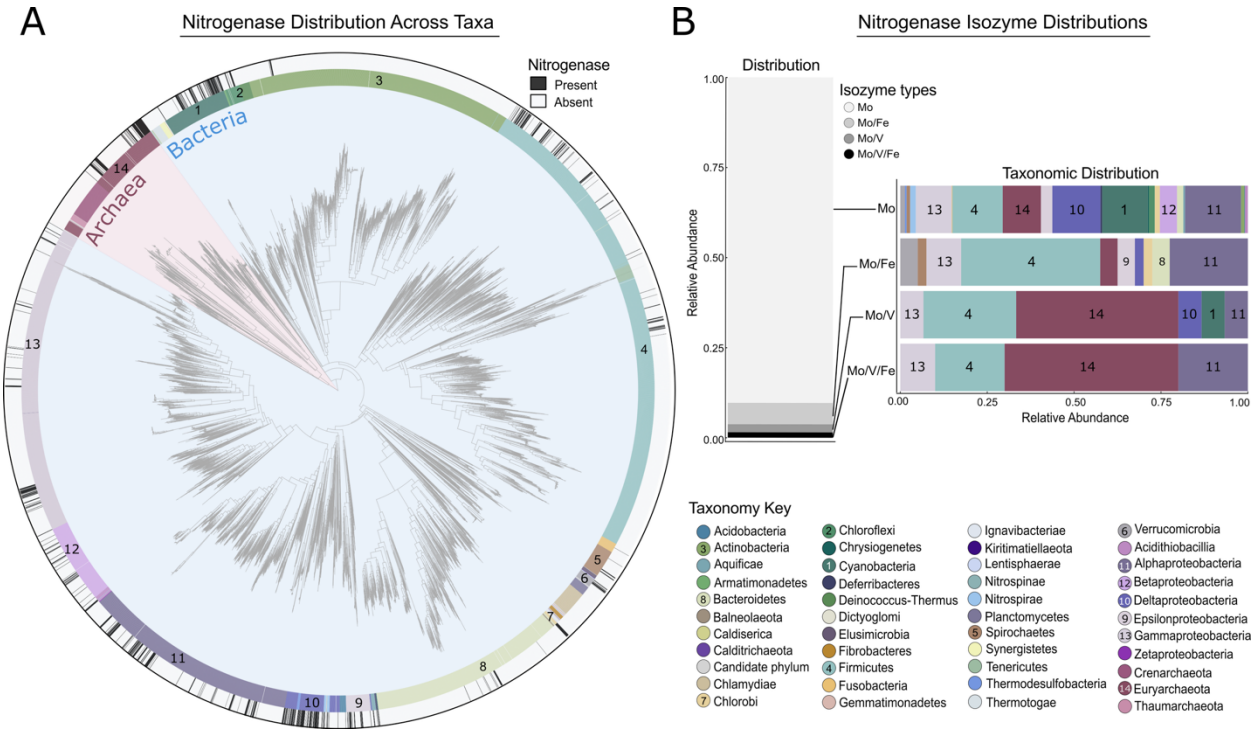
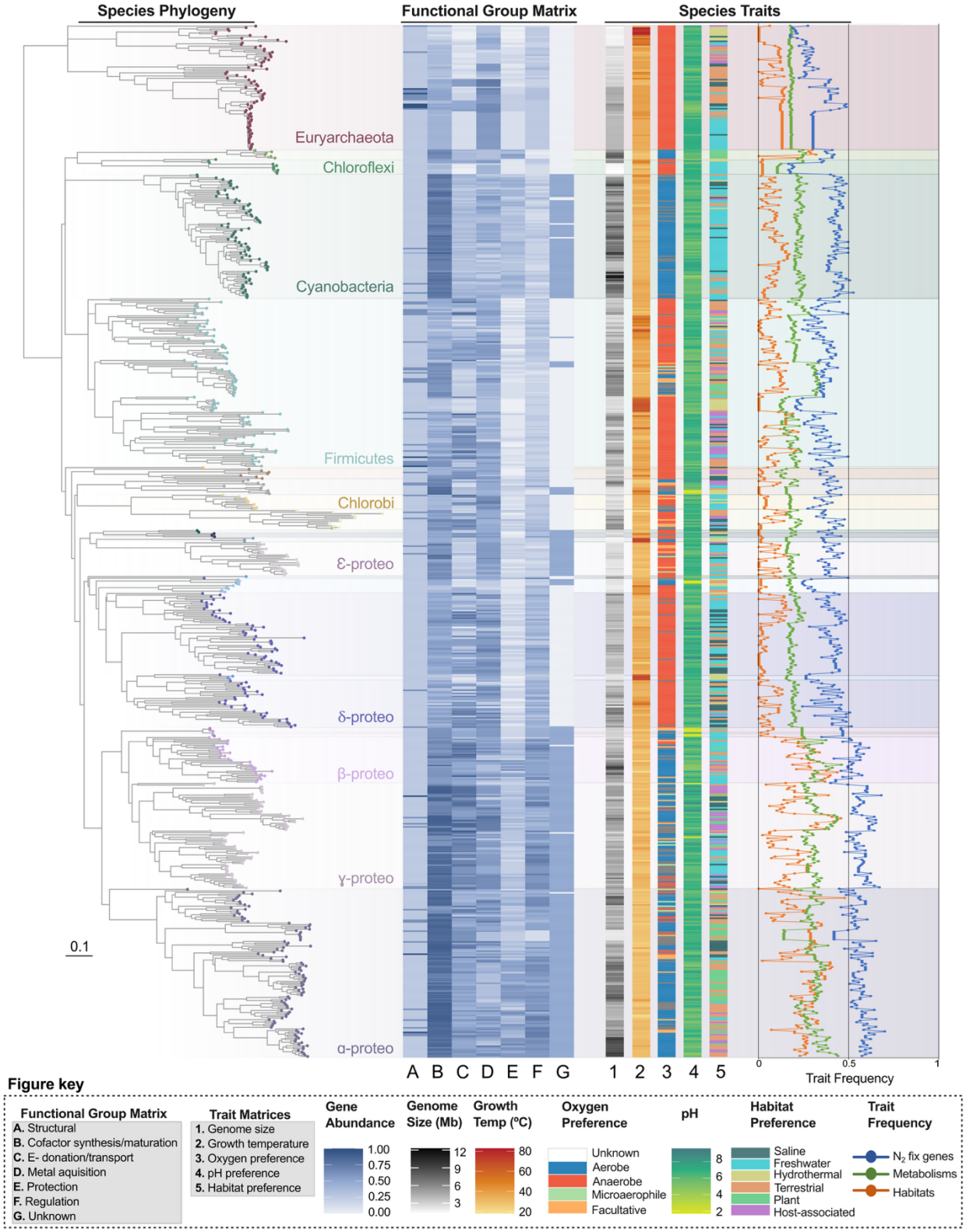


Figure 2. Diversity and phylogenetic distribution of nitrogenase. **A)** The overall distribution of nitrogenase across the microbial diversity from the Zhu et al (2019) Web of Life (WoL) phylogeny [38]. Black bars on the outer ring represent the presence of nitrogenase. Numbers reflect the top 14 taxonomic groups in the dataset with the highest proportion of genomes with nitrogenase. The phylogeny was plotted with iTOL [68]. **B)** Proportions and taxonomic distributions of the different nitrogenase isozymes across WoL.

An N_2 fixation genomic and gene-centric ecological catalog

We screened all 673 genomes for 93 known genes ("machinery") involved in N_2 fixation (**Supplemental Table 3, Supplemental Figure 1**). In Figure 3, genes were organized into categories and the relative abundance of genes within each category was calculated. At the individual genome level, without considering *nifHDK*, the total number of N_2 fixation genes within a genome ranged from 14 genes to 67 genes. The highest number of genes were found in the Alphaproteobacteria *Rhodopseudomonas palustris CGA009,* followed closely by *Azotobacter vinelandii* DJ with 65 genes. Archaea *Methanococcus aeolicus* Nankai-3, *Methanothermococcus okinawensis* IH1 and *Methanocaldococcus infernus* ME all tied for the lowest number of genes (14). Across the species phylogeny, the number of N_2 fixation genes per genome scaled with increasing species branch lengths, indicating

that later diverged taxa tend to have larger N_2 fixation gene repertoires (**Supplemental Figure 2**).



233 **Figure 3**. Distribution of N_2 fixation machinery mapped onto the phylogeny of nitrogenase-containing
234 genomes, highlighting associations with species-associated ecological and physiological traits. This pattern
235 also correlates with those that are mesophilic, prefer aerobic and/or facultative lifestyles, possess larger
236 genome sizes, broader metabolic capabilities, and occupy a wider range of habitats. For this figure, N_2
237 fixation genes were organized into functional categories and the relative abundance of genes for each
238 category calculated. The full gene matrix visualization can be found in Supplemental Figure 1 and the gene
239 matrix and associated categories in Supplemental Table 3. The phylogeny and matrices were plotted using
240 the ggTree R package (version 3.6.2) [44].

241

242 To explore the physiological and ecological diversity of nitrogenase-containing genomes,
243 we examined the habitats they occupy and prefer, their physiological tolerances to
244 oxygen, pH, and temperature, their degree of metabolic versatility, and the variation in
245 their genome sizes (**Figure 3**).

246 Amongst the nitrogenase-containing genomes, 47.5% of the genomes were anaerobes,
247 31.6% were aerobes, 5.3% were facultative, and 1.5% were microaerophiles. The
248 remaining 14% could not be determined due to missing data or uncertainty in predictions.
249 Temperature preferences further define nitrogenase distribution, with most genomes
250 (91.2%) belonging to mesophiles that favor temperatures from 18ºC to 45ºC. The
251 remaining ~8.8% belonged to thermophiles that preferred temperatures above 45ºC.
252 Similarly, neutrophiles (pH 5.5-8.0) dominated the pH preferences (87.2%). Acidophiles
253 and alkaliphiles, though present, were not as common. Genomes differed in their inferred
254 habitat preferences, with some habitats being more commonly represented than others.
255 Here, freshwater and terrestrial environments were particularly favored by 32.5% and
256 26.2% of genomes, respectively. The remaining genomes were associated with saline
257 environments (14.3%), non-plant host associations (11.4%), plant associations (8.8%),
258 and hydrothermal systems (6.8%). These patterns likely reflect not only genuine
259 ecological preferences, but also the influence of uneven environmental sampling, which
260 can skew our perception of where nitrogenase-containing organisms most thrive in the
261 nature. In addition to these traits, we assessed differences in metabolic and habitat
262 coverage to highlight the ecological versatility of diazotrophs. Metabolic coverage ranged
263 from 9.8 to 46.4%, with an average of 23.2%. Habitat coverage ranged from 0 to 48%,

with an average of 12%, when not accounting for the ~3% of genomes whose habitat coverage could not be determined. We also assessed variations in genome size. Here, genome size ranged from 1.3 to 12.3 Mbps, with the average being 4.6 Mbps.

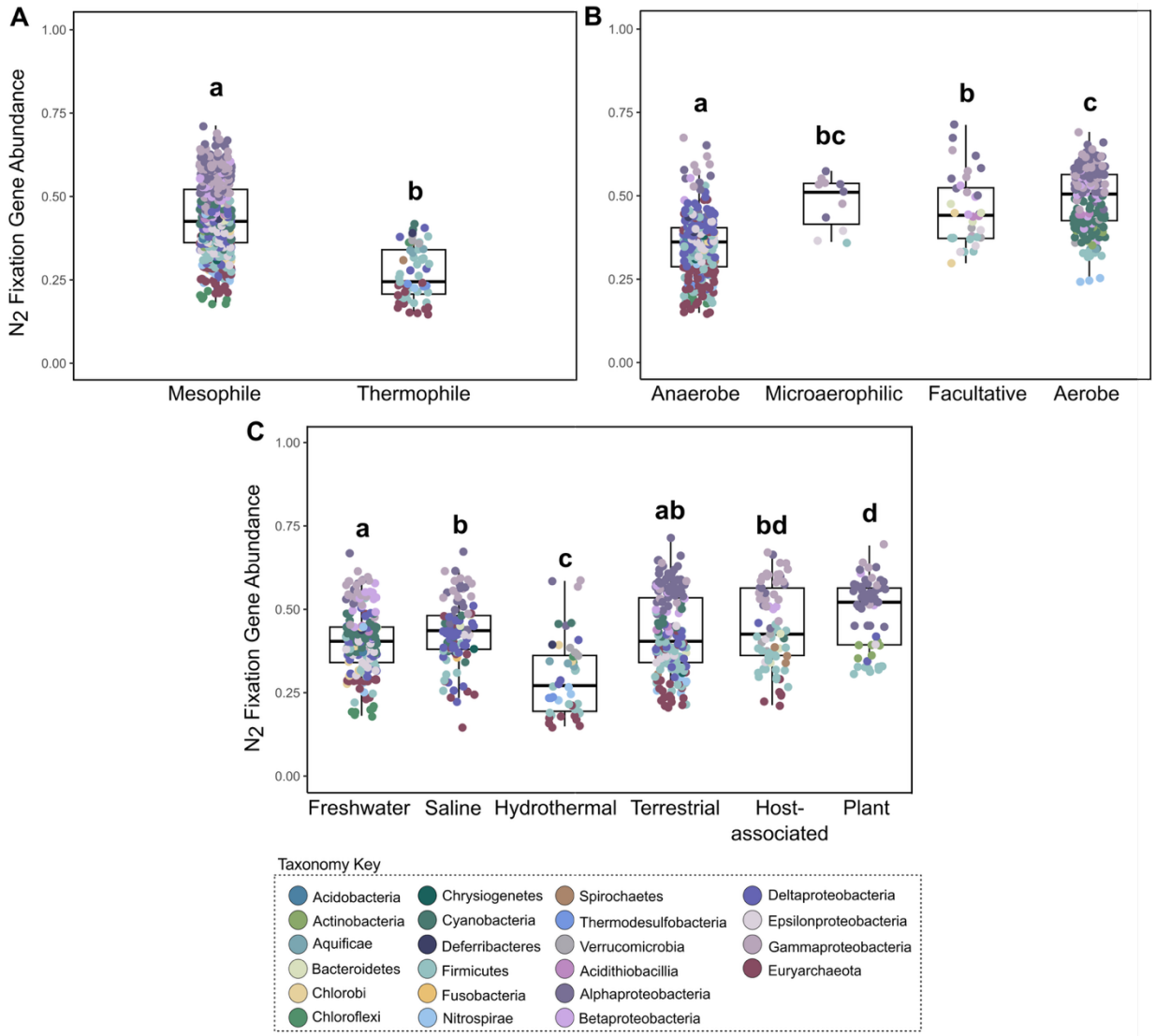


Figure 4. N_2 fixation gene abundance varies across traits. The number of genes was compared across **A)** temperature**, B)** oxygen, and **C)** habitat preferences**.** Pairwise differences between means were calculated using non-parametric Wilcoxon rank-sum tests. For each panel, the means of those that do not differ with statistical evidence share the same letter, whereas those that do not share a letter are discernibly different ($p<0.05$). Boxplots were plotted with the ggpubr R package [55].

Physiological, metabolic, and habitat preferences vary with N_2 fixation genetic complexity

Next, we examined how the number of N_2 fixation genes within a genome relates to its environmental and physiological traits, aiming to uncover the factors that have influenced the complexity of the genetic machinery required for N_2 fixation. We observed that genomes with similar preferences for temperature, oxygen, and habitat type had comparable numbers of N_2 fixation genes (**Figure 3**). Specifically, thermophiles and anaerobes have discernibly fewer genes compared to their mesophilic (p.adj=8.60×10^{-23}) and aerobic (p.adj=4.80×10^{-45}) counterparts (**Fig. 4A,B**). Accordingly, organisms who prefer hydrothermal environments have discernibly fewer N_2 fixation genes on average compared to all other preferred environments (**Fig. 4C**). There were no discernible differences in N_2 fixation gene abundances between terrestrial and freshwater, saline, or host-associated environments. Plant-associated taxa had the highest average of N_2 fixation genes, which was discernibly higher than all environments, except host-associated.

N_2 fixation genetic complexity as a predictor of co-evolutionary trait associations

To test the extent to which traits are evolutionarily conserved or have evolved independently, we measured the phylogenetic signal for gene abundance, habitat coverage, and metabolic coverage, as well as genome size, temperature, pH, and oxygen preference across the species phylogeny using Pagel's lambda (λ) with the *phylolm* R package [67]. All traits had a moderate to high (λ≥0.69) phylogenetic signal (**Supplemental Table 4**), meaning that closely related genomes demonstrated greater similarity in trait values compared to genomes randomly selected from the phylogeny.

We further assessed whether phylogenetic signal persisted at the individual phylum level to test for phylum-specific evolutionary histories and differing selection pressures. This analysis clarified whether the overall phylogenetic signal measured across the species phylogeny was disproportionately driven by particular phyla. Because some phyla had too few genomes (<20), phylum-specific analyses were restricted to Alpha-, Beta-, Delta-, Gammaproteobacteria, as well as Cyanobacteria, Firmicutes, and Euryarchaeota.

303 Genome size, habitat coverage, and metabolic coverage all exhibited strong phylogenetic
304 signal across these major nitrogenase-containing phyla, ($\lambda \geq 0.8$, **Supplemental Table 6**),
305 indicating that these traits are broadly conserved across these lineages, rather than
306 confined to a single clade. N_2 fixation gene abundance likewise showed a high λ (>0.8)
307 within most phyla, with the exception of Cyanobacteria, ($\lambda = 0.34$), likely reflecting limited
308 variability in their N_2 fixation gene sets.

309 By contrast, no phylogenetic signal for temperature was detected in
310 Gammaproteobacteria (λ=0.001), consistent with their mesophilic lifestyle, while other
311 phyla retained high signal (λ>0.8). Oxygen tolerance showed little phylogenetic signal in
312 Gammaproteobacteria nor Firmicutes, whereas pH exhibited strong phylogenetic
313 conservation in Cyanobacteria, Euryarcheaota, and Gamma- and Betaproteobacteria, but
314 low conservation in Deltaproteobacteria.

315 Next, we assessed whether certain traits shared an evolutionary history with N_2 fixation
316 gene abundance via phylogenetic linear regression models (**Figure 5, Supplemental**
317 **Table 5**). Gene abundance was chosen as the independent variable to determine if the
318 complexity of N_2 fixation genes in a genome could be used to predict other trait values.
319 Response (trait) variables were transformed using either natural log or square root to
320 stabilize variance prior to analysis. Genome size (**Figure 5A**) and metabolic coverage
321 (**Figure 5B**) had a positive, significant linear relationship with N_2 fixation gene abundance,
322 indicating that N_2 fixation gene complexity is impacted by broader genomic and metabolic
323 expansion or subtraction forces. For temperature, we initially observed a significant,
324 negative linear relationship with N_2 fixation gene abundance (**Figure 5C**). Because prior
325 work has shown that adaptation to high temperatures is often linked to genome reduction
326 [70], we repeated the regression using N_2 fixation gene abundance normalized to genome
327 size (**Supplemental Table 2**). When normalized, the negative relationship with
328 temperature disappeared, suggesting that the original pattern likely reflected genome size
329 constraints rather than a temperature specific effect on N_2 fixation. A similar pattern
330 emerged for habitat coverage: we found a marginally significant positive relationship
331 (p=0.054) with gene number (**Figure 5D**), which likewise disappeared after normalizing
332 N_2 fixation gene abundance to genome size.

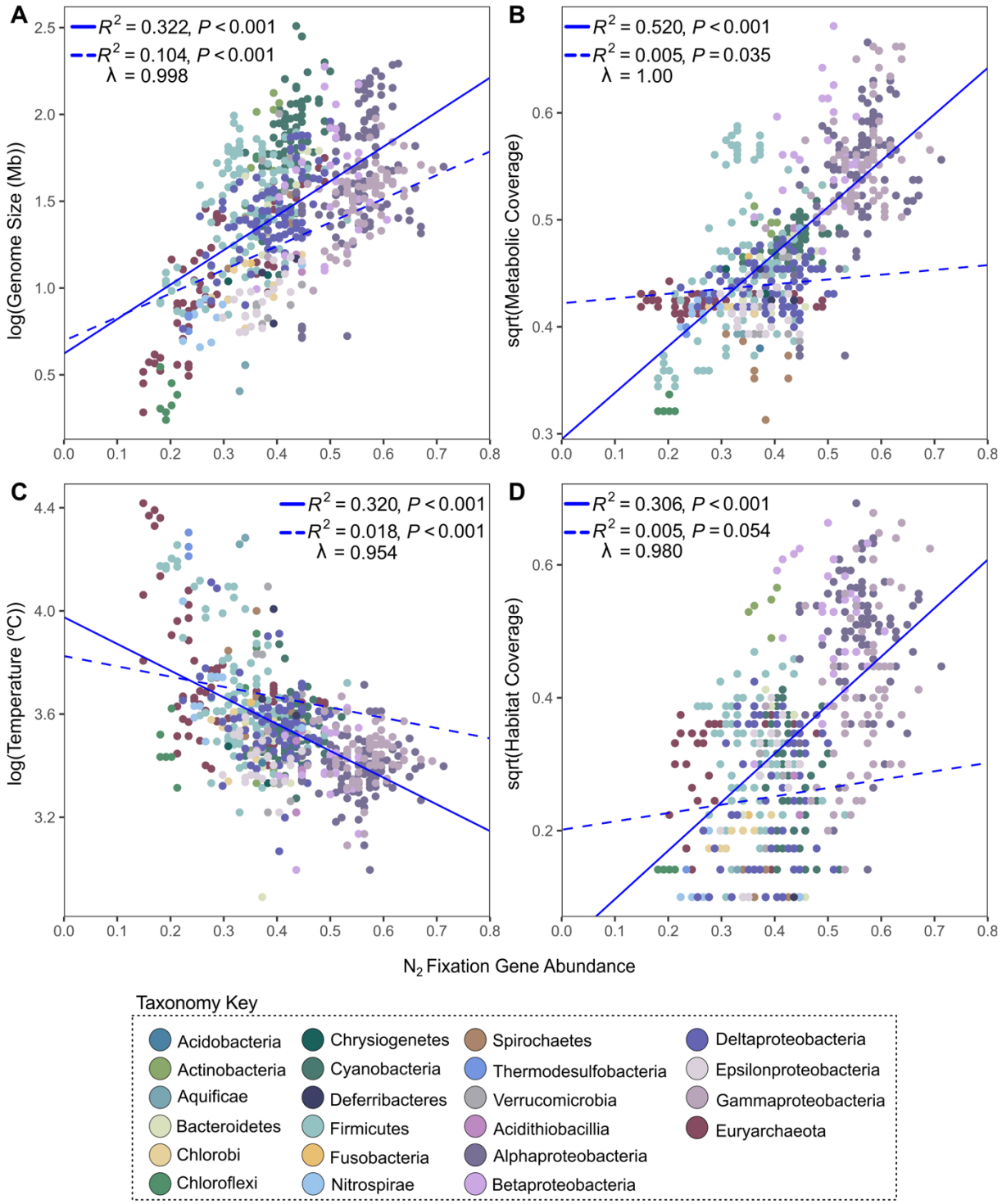


Figure 5. Phylogenetic linear regressions assessed whether the number of N_2 fixation genes in a genome could predict other traits, such as **A)** genome size, **B)** metabolic coverage, **C)** optimal growth temperature, and **D)** habitat coverage. Response variables were either transformed using natural log or square root to stabilize variance prior analysis. Solid blue lines represent non-phylogenetic linear model coefficients,

338 dashed blue lines represent phylogenetic linear coefficients. Scatter plots were plotted using ggplot2 in R
339 [69].

340

341 ***Differences in* N_2 *fixation machinery distribution***

342 We used phylogenetic principal component analysis (PCA) to cluster genomes based on
343 the similarity of their N_2 fixation gene presence-absence patterns, while accounting for
344 shared evolutionary history. Phylogenetic relationships accounted for most of the
345 variation, as indicated by the low variance captured by the phylogenetic PCA
346 (**Supplemental Figure 3**). This finding suggests that clades are clustered primarily
347 because they share conserved gene sets inherited from common ancestors.

348 We then examined the key genes driving separation among genomes in the phylogeny-
349 corrected PCA and interpreted their distribution across taxa and oxygen preferences. The
350 ten genes with the strongest influence on each component of the phylogeny-corrected
351 PCA included the different regulation elements (*nifI*$_2$, *nifL*, *regRS*, *ntrB*), two different
352 electron-donating/transfer mechanisms (*rnfABCDEG, fixABCX*), nodulation and
353 heterocyst differentiation genes (*nodABCD*, *devC*), oxidative protection mechanisms
354 (*sodN*, *cooA*), cofactor synthesis and iron acquisition genes (*nifM*, *sufD*, *feoC2*), and the
355 genes that make up both the Fe- and V-nitrogenase (*vnfHGDKE*, *anfHGDK*). The post-
356 translational *nifI*$_2$ regulator is primarily found in anaerobes, such as Archaea and
357 Firmicutes, confirming that other regulatory mechanisms *regS*, *nifL*, and *ntrB* are primarily
358 found in aerobic Proteobacteria (**Supplemental Figure 1**) [68, 69]. Pearson correlation
359 of gene co-occurrence (**Supplemental Figure 4**) confirming as well that *nifI*$_2$ and *nifI*$_1$ had
360 strong negative correlations with genes *nifTQWXZ*, *fnr*, *fdx*, *oxyR,* and *ntrBC*, which are
361 genes primarily found in aerobes.

362 A preference for either *rnfACDEG* or *fixACX* electron donation systems was evident in
363 the PCA where the genes pointed opposite from one another (**Supplemental Figure 3,4**).
364 Nodulation genes *nodABC* also pointed opposite to *rnfACDEG* and instead had a positive
365 correlation with *fixACX* (**Supplemental Figure 4**). This is confirmed by the fact that
366 diazotrophs forming symbioses with plants through nodulation prefer Fix over Rnf for

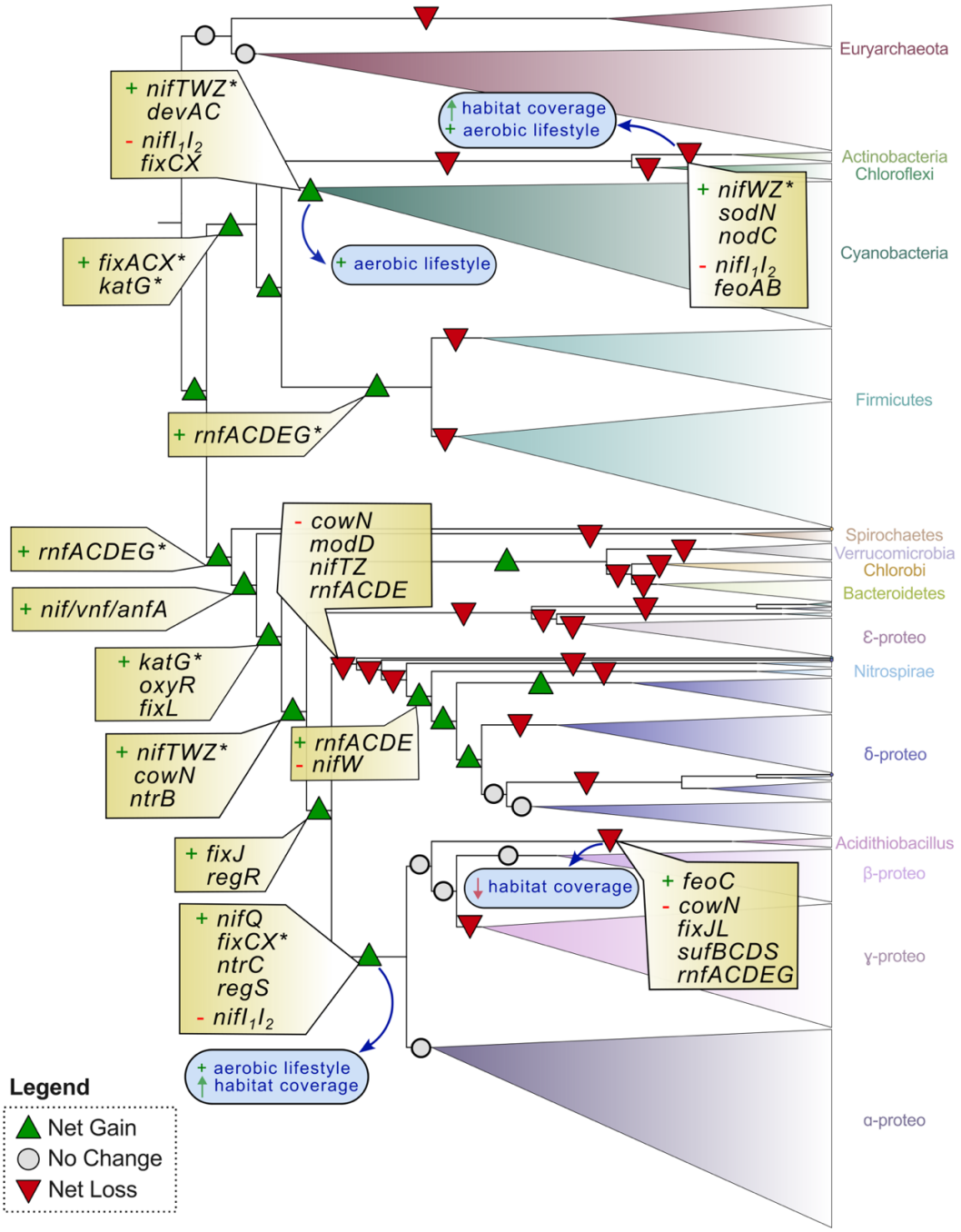
367 electron donation [71, 72]. None of the N_2 fixation genes were unique to thermophiles and
368 neither Fe- nor V- nitrogenases were found in genomes with optimal growth temperatures
369 greater than 41ºC.

370 ***Dynamic history of gene gain and loss***

371 Although previous studies have noted the potential dynamic nature of N_2 fixation gene
372 acquisitions and losses [70], none have specifically analyzed these patterns in detail, nor
373 in relation to their traits. By observing gene gain/loss patterns across the phylogeny and
374 linking them to shifts in lifestyle changes, we can begin to observe the selective pressures
375 on the retention or loss of particular genes. We estimated net gain or loss at each node
376 of the nitrogenase-containing species tree using Wagner parsimony (see methods)
377 (**Figure 6**). Evolutionary trajectories of N_2 fixation machinery complexity indeed indicated
378 a dynamic history of gene gain and loss over time. Specifically, early branching nodes
379 imply net gene gain, suggesting an initial expansion of genes early in diazotroph
380 evolution, whereas gene loss primarily occurs closer to the terminal tree branches.

381 The nodes with the highest number of gene gains (+11) occurs at the most recent
382 common ancestor (MRCA) of Bacteria and at the MRCA of the Euryarchaeota clade
383 Methanosarcina (not shown). Some of the clades that experienced large gene losses
384 include Acidithiobacillia (-18), the Firmicute-Clostridia clade Thermoanaerobacterales (-
385 18) (not shown), Thermodesulfobacteria (-17) (not shown), Nitrospirae (-16), and the
386 MRCA of Aquificae and Epsilonproteobacteria (-11) (**Supplemental Tables 7-8**).
387 Acidothiobacillia's loss of genes *cowN*, *fixJL*, and the Suf and Rnf systems occurs in
388 tandem with a loss of habitat coverage (**Figure 6**). The loss of post-transcriptional
389 regulators $nifI_1I_2$ unsurprisingly coincides with a gain in aerobic lifestyles at the MRCA of
390 Cyanobacteria, Actinobacteria and the MRCA of Gamma- and Alphaproteobacteria
391 (**Figure 6**) as $nifI_1I_2$ is the main regulatory mechanism in anaerobes. The appearance of
392 transcriptional regulation by *nifA*, mostly used by aerobes, predates the loss of $nifI_1I_2$ in
393 the MRCA of Gamma- and Alphaproteobacteria (**Figure 6**). The gain of genes involved
394 in protection against oxidants (*katG*, *oxyR*, and *cowN*) and oxygen response regulators
395 *fixJL* occurs early along the MRCA nodes of Verrucomicrobia, Chlorobi, Bacteroidetes,
396 and Proteobacteria (**Figure 6**). We noted two distinct gains of *rnfACDEG*, *nifTWZ*, *katG*

397 and *fixCX* in distantly related diazotrophs, indicating they were either inherited from a
398 common ancestor that does not have nitrogenase (and therefore would not be
399 represented in this tree), and/or they were gained through horizontal gene transfer.



400

401 **Figure 6.** N_2 fixation gene gain (green triangles) and gene loss (red triangles) throughout the evolution of
402 diazotrophs. No change (grey circle) represents no gain or loss at that node, or an equal number or gains

403 and losses. Asterisks highlight potential candidates for HGT or genes derived from non-diazotroph
404 ancestors not in this subsetted phylogeny. All gene events can be found in Supplemental table 8. Key
405 transitional events of gene abundance and habitat coverage can be found in Supplemental Figure 5.

406

407 **Discussion**

408 ***Nitrogenase Distribution, Diversity, and Evolution Across Prokaryotic Lineages***

409 Nitrogenase is widespread across prokaryotes, appearing within phylogenetically and
410 ecologically diverse Bacteria and Archaea (**Figures 2 and 3**). Mo-nitrogenase is the
411 dominate form encoded by diazotrophs, reported everywhere from the rhizosphere [73]
412 to deep-sea sediments kilometers below the ocean surface [74–76]. Although V- and Fe-
413 nitrogenases occur less frequently, they also span across diverse taxa (**Figure 2**; [77]).
414 However, focusing solely on nitrogenase's distribution overlooks the evolutionary and
415 genetic complexity of N_2 fixation.

416 To address this, we developed a comprehensive catalog of nitrogenase-containing
417 genomes and associated N_2 fixation machinery, mapped to environmental preferences
418 and phylogenetic history (**Figure 3**). Expanding on prior work [75–77], our work
419 incorporates ~2.5× more N_2 fixation genes and genomes, along with data on oxygen
420 metabolism, temperature, pH, genome size, habitat range. Whereas previous
421 comparative studies often targeted single clades or ecosystems [78–80], our framework
422 unifies nitrogenase and its supporting genes across the entire microbial tree.

423 Oxygen inhibits nitrogenase activity [81], and the enzyme's sensitivity to oxygen has long
424 shaped how diazotrophs function and evolve [78, 79]. Our expanded analysis confirms
425 that the majority of nitrogenase-containing genomes belong to anaerobic organisms,
426 while aerobic diazotrophs harbor significantly more N_2 fixation genes than anaerobes
427 (**Figure 4B**), likely reflecting the additional genetic machinery needed for oxygen
428 protection and regulation under aerobic growth [36]. We found that genes *katG*, *oxyR*,
429 and *cowN*, which provide protection against oxidants [82, 83], along with the oxygen
430 response regulators *fixJL* [81] were acquired along the lineages leading up to the
431 divergence of Proteobacteria (**Figure 6**), a period that coincides with the estimated timing

of the Great Oxidation Event (GOE; see Figure 3 in [84]). This pattern indicates these genes played a role enabling ancient diazotrophs to expand into increasingly oxygenated environments. Our recent work further supports the divergence of the alternative nitrogenases during this transition to an oxic atmosphere [66], further underscoring how environmental transitions, particularly the GOE, reshaped molecular evolution of N_2 fixation in deep time.

We also confirmed the relationship between N_2 fixation gene abundance and temperature previously reported by Boyd et al [67]: thermophiles, on average, possess the simplest N_2 fixation gene sets (**Figure 4AB, Figure 5C**). Our analyses indicate that this pattern is in part influenced by genome reduction associated with long-term adaptation to high temperatures [70]. Yet, if the earliest diazotrophs were indeed thermophilic [3, 85], then temperature's role shaping the evolution of N_2 fixation cannot be dismissed. For example, some cold-adapted diazotrophs switch from Mo- to V-based nitrogenases [86–88], indicating how shifts in temperature regimes may have shaped the functional diversification of nitrogenase over evolutionary time.

Building on these patterns, we next examined unusual derivations from the typical nitrogenase gene set. Notably, several anaerobic thermophiles lack *nifE*, *nifN*, or both (**Figure 3**), genes involved in FeMo cofactor assembly and part of the cononical minimal *nif* genetic set (*nifHDKENB)* required for N_2 fixation [86]. These lineages include Euryarcheaota, such as members of the ANME groups [89], Chloroflexi, and Firmicutes. Some have verified N_2 fixation activity [90–93], while others, like *Roseiflexus castenholzii*, do not [94].

Interestingly, recent work suggests that *nifDK* may have originated from a duplicated maturase-like *nifEN* ancestor [67], raising the possibility that some extant diazotrophs lacking *nifEN* evolved alternative cofactor maturation strategies, perhaps as a way to maximize energy conservation under thermophilic conditions [70]. Support for this idea comes from *Roseiflexus* sp. RS-1, which lacks *nifEN* and appears to use *nifDK* to fulfill both the cofactor maturase and catalytic roles [95]. Further work will be required to determine whether this streamlined mechanism occurs in other diazotrophs missing *nifEN*.

462 Plant-associated genomes showed the highest N_2 fixation gene counts (**Figure 4C**),
463 consistent with their pivotal role in host nitrogen supply [96, 97]. Our reconstructions
464 captured the divergence of Actinobacteria *Frankia* from its ancestor, marked by a
465 transition to aerobic metabolism, accompanying genome and habitat expansion, the loss
466 of *nifI*$_1$*I*$_2$ regulators, and the gain of both the nodulation gene *nodC* and oxidative stress
467 response gene *sodN* to support its dual free-living and symbiosis states (**Figure 6**).

468 We observed that *Frankia* lacks key regulators such as like *nifAL* [95]. However, we noted
469 the presence of the post-translational regulator *draG*, without its partner *draT*, indicating
470 that draG-draT regulation is likely non-functional and perhaps a novel, as-yet-unknown
471 mechanism for nitrogenase inactivation is present [98] – a possibility that warrants further
472 study.

473 In many plant symbioses, oxygen levels are tightly regulated by the host to activate the
474 FixL - FixJ pathway, ultimately driving nitrogenase expression [96], while *nod* genes
475 mediate nodulation [99]. Terrestrial habitats, being more variable than oceans [100],
476 impose fluctuating environmental conditions that demand dynamic nitrogenase regulation
477 [101–103]. Soil molybdenum scarcity likely favors the deployment of alternative
478 nitrogenases [104], contributing to gene repertoire expansion in terrestrial environments.
479 Our results illustrate this pattern. For example, several lichen-symbionts within the
480 Nostocales order of Cyanobacteria possess V-nitrogenase [105], likely as a way to
481 combat limited Mo in high latitude boreal forests [106]. Together, these findings highlight
482 how diazotroph N_2 fixation gene repertoires have been shaped by environmental
483 constraints and the selective pressures of symbiotic lifestyles.

484

485 ***Eco-evolutionary history of nitrogenase-associated genes***

486 Central to our framework is that N_2 fixation gene complexity is not an isolated trait, rather,
487 it is embedded within a larger evolutionary framework of genome expansion, metabolic
488 versatility, and habitat diversity (**Figure 5**). Gene complexity can even serve as an
489 ecological marker—distinguishing generalists, with broad habitats and large, versatile

genomes [107–109], from specialists with narrow niches and smaller genomes [110, 111]. This distinction is increasingly relevant today, as specialist diazotrophs may contribute disproportionately to total soil N_2 fixation [112].

Our reconstructions show that as diazotrophs diversified, so did their N_2 fixation machinery (**Figure 3**). Gene gain dominated early evolution, with clades such as Alpha-, Gamma-, and Betaproteobacteria maintaining extensive N_2 fixation genes (**Figure 6**). In contrast, later stages were marked by lineage-specific losses, particularly in terminal branches, pointing ecological specialization and genome streamlining. Losses were most pronounced in Chloroflexi, Nitrospirae, Aquificae, Acidithiobacillia, Thermodesulfobacterota, and Epsilonproteobacteria, groups often associated with anaerobic, thermophilic, acidophilic, or host-adapted lifestyles (**Supplemental Table 2**).

In some cases, N_2 fixation gene loss tracked closely with reduced habitat coverage, as in Epsilonproteobacteria and Acidithiobacillia (**Figure 6**), consistent with prior work on genome streamlining in these clades [113–115]. Adaptation to deep-sea hydrothermal vents [116], shifts toward heterotrophy [114], and transitions to acidophily [117] all show how niche transitions often coincide with simplification of N_2 fixation gene sets.

Yet, gene loss was not universal in streamlined genomes. *Zymomonas mobilis* (Alphaproteobacteria) maintains the highest proportion of N_2 fixation genes (~2-3% of all coding genes) despite having a compact ~2Mb genome (**Supplemental Table 2**). Similarly, in symbiont systems, genome reduction often reflects host compensation [118]. UCYN-A2, a marine symbiont recently identified as an early stage “nitroplast” in *Braarudosphaera bigelowii* [119, 120], has lost CO_2 fixation and photosystem II [121] [121, 122], yet retains an intact *nif* cluster (*nifTZVBSHDKENXW)* [123].

Horizontal gene transfer (HGT) has further shaped N_2 fixation evolution [5, 80], likely explaining some gene acquisitions we observed (e.g., *rnfACDEG*, *nifTWZ*, *katG* and *fixCX* in Figure 6). Previous studies have documented entire *nif* clusters moving across major clades – including a >20kb transfer from Gammaproteobacterium in to the diatom symbiont “*Candidatus* Tectiglobus diatomicola” [124]. In our dataset, *Coleofasciculus*

518 *chthonoplastes* PCC 7420 (formerly *Microcoleus chthonoplastes*) appears to have
519 acquired *nifHDKEN* from a Deltaproteobacterium [125].

520 Together, these patterns show that the complexity of N_2 fixation genes reflects a balance
521 between the forces of gene loss and gain, specialization, horizontal transfer - all
522 embedded in the broader trajectory of diazotroph genome evolution.

523 ***Limitations of comparative genomics***

524 As with any comparative genomic approach, there are inherent limitations in identifying
525 the full complement of genes and traits within genomes. One limitation is that our analysis
526 focuses on genes already recognized to be playing a role in N_2 fixation, meaning
527 additional, yet-undiscovered genes may contribute to the process, particularly in non-
528 model diazotrophs, where many functional mechanisms remain uncharacterized. Another
529 caveat is our reliance on KEGG KO number. Not all genes could be assigned KO
530 numbers, and some KO identifiers fail to distinguish between closely genes, such as
531 *nifA*/*vnfA*/*anfA*. Continued improvements in gene functional annotation and database
532 curation will be essential for building a more complete undersatnding of of diazotroph
533 capabilities [126].

534 ***Conclusion***

535 Examining the distribution of diverse N_2 fixation machinery offers insight into which genes
536 were critical for enabling transitions into new environments over evolutionary time. The
537 rise of atmospheric oxygen, for example, fundamentally reshaped electron-donor
538 availability [127] and, along with changes in in metal availability [14, 128], likely exerted
539 strong selective pressures on N_2 fixation machinery as we have shown here.

540 Reconstructing historical patterns from the modern genetic records is thus pivotal for
541 understanding the processes that govern evolutionary change [129]. Global trends in
542 genome evolution not only reveal how organisms responded to past environmental
543 fluctuations but connects them to their present-day ecological niches. In particular, N_2
544 fixation complexity may help infer a diazotroph's ecological role, highlighting how

environmental factors and cellular demands intersect to enable colonization of new habitats.

In his book *Life's Engines*, Paul Falkowski likens nitrogenase and its core machinery to a computer's hardware, while accessory genes function as the software, essential for maintenance, optimization and updates [130]. This analogy is compelling, capturing the deep interconnection between core biochemical processes and their regulatory networks. These elements must evolve in concert to navigate the challenges posed by changing environments, a dynamic reflected in the expansions and contractions of diazotrophs which we have documented across both ancient and modern contexts.

Acknowledgements

We gratefully acknowledge support from the NASA Interdisciplinary Consortium for Astrobiology Research (ICAR): Metal Utilization and Selection Across Eons, MUSE under grant no 80NSSC17K0296, and the NASA Exobiology Program under grant no NNH23ZDA001N. M.S. was supported by the NASA Postdoctoral Program, administered by Oak Ridge Associated Universities under contract with NASA. We thank Patricia Tran, Shaomei He, colleagues in the Kaçar Lab, members of NASA MUSE ICAR, and colleagues at Texas State University for helpful feedback on manuscript; and the Center for High Throughput Computing (CHTC) and the Department of Bacteriology at the University of Wisconsin–Madison for providing computing resources. We would like to finally thank the two anonymous reviewers for their constructive feedback.

Data availability

Supplementary Table 2 contains genome accession, taxonomy, and trait data information. Custom scripts, tree files, raw data, and PDF versions of the figures are available at https://github.com/kacarlab/Sobol_2025_NifGeneEcoEvo.

References

1. Falkowski PG. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* 1997; **387**: 272–275.
2. Canfield DE, Glazer AN, Falkowski PG. The evolution and future of Earth's nitrogen

574 cycle. *Science* 2010; **330**: 192–196.

575 3. Mus F, Colman DR, Peters JW, Boyd ES. Geobiological feedbacks, oxygen, and
576 the evolution of nitrogenase. *Free Radic Biol Med* 2019; **140**: 250–259.

577 4. Garcia AK, McShea H, Kolaczkowski B, Kaçar B. Reconstructing the evolutionary
578 history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization.
579 *Geobiology* 2020; **18**: 394–411.

580 5. Parsons C, Stüeken EE, Rosen CJ, Mateos K, Anderson RE. Radiation of nitrogen-
581 metabolizing enzymes across the tree of life tracks environmental transitions in
582 Earth history. *Geobiology* 2021; **19**: 18–34.

583 6. Lyons TW, Tino CJ, Fournier GP, Anderson RE, Leavitt WD, Konhauser KO, et al.
584 Co-evolution of early Earth environments and microbial life. *Nat Rev Microbiol* 2024;
585 **22**: 572–586.

586 7. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R. Distribution of nitrogen
587 fixation and nitrogenase-like sequences amongst microbial genomes. *BMC*
588 *Genomics* 2012; **13**: 162.

589 8. Navarro-González R, McKay CP, Mvondo DN. A possible nitrogen crisis for
590 Archaean life due to reduced nitrogen fixation by lightning. *Nature* 2001; **412**: 61–
591 64.

592 9. Raymond J, Siefert JL, Staples CR, Blankenship RE. The natural history of nitrogen
593 fixation. *Mol Biol Evol* 2004; **21**: 541–554.

594 10. Allen JF, Thake B, Martin WF. Nitrogenase inhibition limited oxygenation of earth's
595 Proterozoic atmosphere. *Trends Plant Sci* 2019; **24**: 1022–1031.

596 11. Zehr JP, Capone DG. Changing perspectives in marine nitrogen fixation. *Science*

597 2020; **368**: eaay9514.
598 12. Marcarelli AM, Fulweiler RW, Scott JT. Nitrogen fixation: a poorly understood
599 process along the freshwater-marine continuum. *Limnol Oceanogr Lett* 2022; **7**: 1–
600 10.
601 13. Mus F, Crook MB, Garcia K, Garcia Costas A, Geddes BA, Kouri ED, et al.
602 Symbiotic nitrogen fixation and the challenges to its extension to nonlegumes. *Appl*
603 *Environ Microbiol* 2016; **82**: 3698–3710.
604 14. Rucker HR, Kaçar B. Enigmatic evolution of microbial nitrogen fixation: insights
605 from Earth's past. *Trends Microbiol* 2024; **32**: 554–564.
606 15. Garcia AK, Kolaczkowski B, Kaçar B. Reconstruction of nitrogenase predecessors
607 suggests origin from maturase-like proteins. *Genome Biol Evol* 2022; **14**.
608 16. Harris DF, Rucker HR, Garcia AK, Yang Z-Y, Chang SD, Feinsilber H, et al. Ancient
609 nitrogenases are ATP dependent. *MBio* 2024; **15**: e0127124.
610 17. Cuevas-Zuviría B, Detemple F, Amritkar K, Garcia AK, Seefeldt L, Einsle O, et al.
611 Nitrogenase structural evolution across Earth's history. *eLife* 2025; **14**.
612 18. Cuevas-Zuviría B, Garcia AK, Rivier AJ, Rucker HR, Carruthers BM, Kaçar B.
613 Emergence of an orphan nitrogenase protein following atmospheric oxygenation.
614 *Mol Biol Evol* 2024; **41**.
615 19. Bishop PE, Joerger RD. Genetics and molecular biology of alternative nitrogen
616 fixation systems. *Annu Rev Plant Physiol Plant Mol Biol* 1990; **41**: 109–125.
617 20. Eady RR. Structureminus signFunction relationships of alternative nitrogenases.
618 *Chem Rev* 1996; **96**: 3013–3030.
619 21. Howard JB, Rees DC. How many metals does it take to fix N_2? A mechanistic

620 overview of biological nitrogen fixation. *Proc Natl Acad Sci U S A* 2006; **103**:
621 17088–17093.
622 22. Seefeldt LC, Yang Z-Y, Lukoyanov DA, Harris DF, Dean DR, Raugei S, et al.
623 Reduction of substrates by nitrogenases. *Chem Rev* 2020; **120**: 5082–5106.
624 23. Kaiser JT, Hu Y, Wiig JA, Rees DC, Ribbe MW. Structure of precursor-bound
625 NifEN: a nitrogenase FeMo cofactor maturase/insertase. *Science* 2011; **331**: 91–94.
626 24. Dixon R, Kahn D. Genetic regulation of biological nitrogen fixation. *Nat Rev*
627 *Microbiol* 2004; **2**: 621–631.
628 25. Rubio LM, Ludden PW. Biosynthesis of the iron-molybdenum cofactor of
629 nitrogenase. *Annu Rev Microbiol* 2008; **62**: 93–111.
630 26. Mus F, Alleman AB, Pence N, Seefeldt LC, Peters JW. Exploring the alternatives of
631 biological nitrogen fixation. *Metallomics* 2018; **10**: 523–538.
632 27. Kleinschmidt JA, Kleiner D. Relationships between nitrogenase, glutamine
633 synthetase, glutamine, and energy charge in Azotobacter vinelandii. *Arch Microbiol*
634 1981; **128**: 412–415.
635 28. Luxem KE, Kraepiel AML, Zhang L, Waldbauer JR, Zhang X. Carbon substrate re-
636 orders relative growth of a bacterium using Mo-, V-, or Fe-nitrogenase for nitrogen
637 fixation. *Environ Microbiol* 2020; **22**: 1397–1408.
638 29. Nonaka A, Yamamoto H, Kamiya N, Kotani H, Yamakawa H, Tsujimoto R, et al.
639 Accessory Proteins of the Nitrogenase Assembly, NifW, NifX/NafY, and NifZ, Are
640 Essential for Diazotrophic Growth in the Nonheterocystous Cyanobacterium
641 Leptolyngbya boryana. *Front Microbiol* 2019; **10**: 495.
642 30. Lery LMS, Bitar M, Costa MGS, Rössle SCS, Bisch PM. Unraveling the molecular

643 mechanisms of nitrogenase conformational protection against oxygen in
644 diazotrophic bacteria. *BMC Genomics* 2010; **11 Suppl 5**: S7.
645 31. Schlesier J, Rohde M, Gerhardt S, Einsle O. A conformational switch triggers
646 nitrogenase protection from oxygen damage by Shethna protein II (FeSII). *J Am*
647 *Chem Soc* 2016; **138**: 239–247.
648 32. Van Stappen C, Jiménez-Vicente E, Pérez-González A, Yang Z-Y, Seefeldt LC,
649 DeBeer S, et al. A conformational role for NifW in the maturation of molybdenum
650 nitrogenase P-cluster. *Chem Sci* 2022; **13**: 3489–3500.
651 33. Sabra W, Zeng A-P, Lünsdorf H, Deckwer W-D. Effect of oxygen on formation and
652 structure of Azotobacter vinelandii alginate and its role in protecting nitrogenase.
653 *Appl Environ Microbiol* 2000; **66**: 4037–4044.
654 34. Milligan AJ, Berman-Frank I, Gerchman Y, Dismukes GC, Falkowski PG. Light-
655 dependent oxygen consumption in nitrogen-fixing cyanobacteria plays a key role in
656 nitrogenase protection[1]. *J Phycol* 2007; **43**: 845–852.
657 35. Poudel S, Colman DR, Fixen KR, Ledbetter RN, Zheng Y, Pence N, et al. Electron
658 transfer to nitrogenase in different genomic and metabolic backgrounds. *J Bacteriol*
659 2018; **200**.
660 36. Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW. Evolution of molybdenum
661 nitrogenase during the transition from anaerobic to aerobic metabolism. *J Bacteriol*
662 2015; **197**: 1690–1699.
663 37. Nichio BT de L, Chaves RBR, Pedrosa FO, Raittz RT. Exploring diazotrophic
664 diversity: Unveiling nif core distribution and evolutionary patterns in nitrogen-fixing
665 organisms. *Research Square* . 2024.

666 38. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, et al. Phylogenomics
667 of 10,575 genomes reveals evolutionary proximity between domains Bacteria and
668 Archaea. *Nat Commun* 2019; **10**: 5477.
669 39. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy
670 TBK, et al. Minimum information about a single amplified genome (MISAG) and a
671 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*
672 2017; **35**: 725–731.
673 40. Nadkarni PM, Parikh CR. An eUtils toolset and its use for creating a pipeline to link
674 genomics and proteomics analyses to domain-specific biomedical literature. *J Clin*
675 *Bioinforma* 2012; **2**: 9.
676 41. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference
677 resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**: D457-62.
678 42. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al.
679 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive
680 score threshold. *Bioinformatics* 2020; **36**: 2251–2252.
681 43. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and
682 evolutionary analyses in R. *Bioinformatics* 2019; **35**: 526–528.
683 44. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc*
684 *Bioinformatics* 2020; **69**: e96.
685 45. Reimer LC, Sardà Carbasse J, Koblitz J, Ebeling C, Podstawka A, Overmann J.
686 BacDive in 2022: the knowledge base for standardized bacterial and archaeal data.
687 *Nucleic Acids Res* 2022; **50**: D741–D746.
688 46. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al.

689 BioProject and BioSample databases at NCBI: facilitating capture and organization
690 of metadata. *Nucleic Acids Res* 2012; **40**: D57-63.
691 47. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, et
692 al. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids*
693 *Res* 2021; **49**: D723–D733.
694 48. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A synthesis
695 of bacterial and archaeal phenotypic trait data. *Sci Data* 2020; **7**: 170.
696 49. Sauer DB, Wang D-N. Predicting the optimal growth temperatures of prokaryotes
697 using only genome derived features. *Bioinformatics* 2019; **35**: 3224–3231.
698 50. Jabłońska J, Tawfik DS. The number and type of oxygen-utilizing enzymes
699 indicates aerobic vs. anaerobic phenotype. *Free Radic Biol Med* 2019; **140**: 84–92.
700 51. Ramoneda J, Stallard-Olivera E, Hoffert M, Winfrey CC, Stadler M, Niño-García JP,
701 et al. Building a genome-based understanding of bacterial pH preferences. *bioRxiv* .
702 2023. , 2023.01.24.524446
703 52. Mise K, Iwasaki W. Environmental atlas of prokaryotes enables powerful and
704 intuitive habitat-based analysis of community structures. *iScience* 2020; **23**:
705 101624.
706 53. Seemann T. barrnap: Bacterial ribosomal RNA predictor.
707 https://github.com/tseemann/barrnap. Accessed 2 Dec 2024.
708 54. Geller-McGrath D, Konwar KM, Edgcomb VP, Pachiadaki M, Roddy JW, Wheeler
709 TJ, et al. Predicting metabolic modules in incomplete bacterial genomes with
710 MetaPathPredict. *Elife* 2024; **13**.
711 55. Kassambara A. ggpubr:'ggplot2'based publication ready plots. *R package version*

712 2018; 2.
713 56. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative
714 methods (and other things). *PeerJ* 2024; **12**: e16505.
715 57. R Core Team. R: A Language and Environment for Statistical Computing. Vienna,
716 Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical
717 Computing. 2015.
718 58. Wei T, Simko V. R package “corrplot”: Visualization of a Correlation Matrix. 2024.
719 59. Ho L si T, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait
720 evolution models. *Syst Biol* 2014; **63**: 397–408.
721 60. Jombart T, Balloux F, Dray S. Adephylo: New tools for investigating the
722 phylogenetic signal in biological traits. *Bioinformatics* 2010; **26**: 1907–1909.
723 61. Csűrös M. Ancestral reconstruction by asymmetric Wagner parsimony over
724 continuous characters and squared parsimony over distributions. *Comparative*
725 *Genomics*. 2008. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 72–86.
726 62. Kapust N, Nelson-Sathi S, Schönfeld B, Hazkani-Covo E, Bryant D, Lockhart PJ, et
727 al. Failure to recover major events of gene flux in real biological data due to method
728 misapplication. *Genome Biol Evol* 2018; **10**: 1198–1209.
729 63. Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, et al.
730 Ancestral genome reconstruction identifies the evolutionary basis for trait
731 acquisition in polyphosphate accumulating bacteria. *ISME J* 2016; **10**: 2931–2945.
732 64. Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, et al.
733 Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to
734 phylogenetic trees. *Bioinformatics* 2014; **30**: 2216–2218.

735 65. Koirala A, Brözel V. Phylogeny of nitrogenase structural and assembly components
736 reveals new insights into the origin and distribution of nitrogen fixation across
737 bacteria and Archaea. *Microorganisms* 2021; **9**: 1662.
738 66. Pi H-W, Lin J-J, Chen C-A, Wang P-H, Chiang Y-R, Huang C-C, et al. Origin and
739 evolution of nitrogen fixation in prokaryotes. *Mol Biol Evol* 2022; **39**.
740 67. Bellanger M, Figueroa JL 3rd, Tiemann L, Friesen ML, Iii RAW. NFixDB (Nitrogen
741 Fixation DataBase)-a comprehensive integrated database for robust ’omics analysis
742 of diazotrophs. *NAR Genom Bioinform* 2024; **6**: lqae063.
743 68. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the
744 phylogenetic tree display and annotation tool. *Nucleic Acids Res* 2024; **52**: W78–
745 W82.
746 69. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in
747 bacteria are negatively correlated, suggesting genomic streamlining during thermal
748 adaptation. *Genome Biol Evol* 2013; **5**: 966–977.
749 70. Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis (2nd ed.).
750 *Measurement (Mahwah NJ)* 2019; **17**: 160–167.
751 71. Earl CD, Ronson CW, Ausubel FM. Genetic and structural analysis of the
752 Rhizobium meliloti fixA, fixB, fixC, and fixX genes. *J Bacteriol* 1987; **169**: 1127–
753 1136.
754 72. Gubler M, Hennecke H. Regulation of the fixA gene and fixBC operon in
755 Bradyrhizobium japonicum. *J Bacteriol* 1988; **170**: 1205–1214.
756 73. Smercina DN, Evans SE, Friesen ML, Tiemann LK. To fix or not to fix: Controls on
757 free-living nitrogen fixation in the rhizosphere. *Appl Environ Microbiol* 2019; **85**.

758 74. Reese BK, Zinke LA, Sobol MS, LaRowe DE, Orcutt BN, Zhang X, et al. Nitrogen
759 cycling of active bacteria within oligotrophic sediment of the mid-Atlantic ridge flank.
760 *Geomicrobiol J* 2018; **35**: 468–483.
761 75. Kapili BJ, Barnett SE, Buckley DH, Dekas AE. Evidence for phylogenetically and
762 catabolically diverse active diazotrophs in deep-sea sediment. *ISME J* 2020; **14**:
763 971–983.
764 76. Semler AC, Paris ER, Salvador M, Dekas AE. Abundance, identity, and potential
765 diazotrophic activity of nifH-containing organisms at marine cold seeps. *Environ*
766 *Microbiol* 2025; **27**: e70058.
767 77. McRose DL, Zhang X, Kraepiel AML, Morel FMM. Diversity and activity of
768 alternative nitrogenases in sequenced genomes and coastal environments. *Front*
769 *Microbiol* 2017; **8**: 267.
770 78. Xie J-B, Du Z, Bai L, Tian C, Zhang Y, Xie J-Y, et al. Comparative genomic analysis
771 of N2-fixing and non-N2-fixing Paenibacillus spp.: organization, evolution and
772 expression of the nitrogen fixation genes. *PLoS Genet* 2014; **10**: e1004231.
773 79. Higdon SM, Pozzo T, Kong N, Huang BC, Yang ML, Jeannotte R, et al. Genomic
774 characterization of a diazotrophic microbiota associated with maize aerial root
775 mucilage. *PLoS One* 2020; **15**: e0239677.
776 80. Wang S, Meade A, Lam H-M, Luo H. Evolutionary timeline and genomic plasticity
777 underlying the lifestyle diversity in Rhizobiales. *mSystems* 2020; **5**.
778 81. Gallon JR. Tansley review no. 44. Reconciling the incompatible: N_2 fixation and O_2.
779 *New Phytol* 1992; **122**: 571–609.
780 82. Zhao Y, Nickels LM, Wang H, Ling J, Zhong Z, Zhu J. OxyR-regulated catalase

activity is critical for oxidative stress resistance, nodulation and nitrogen fixation in Azorhizobium caulinodans. *FEMS Microbiol Lett* 2016; **363**: fnw130.

83. Medina MS, Bretzing KO, Aviles RA, Chong KM, Espinoza A, Garcia CNG, et al. CowN sustains nitrogenase turnover in the presence of the inhibitor carbon monoxide. *J Biol Chem* 2021; **296**: 100501.

84. Davín AA, Woodcroft BJ, Soo RM, Morel B, Murali R, Schrempf D, et al. A geological timescale for bacterial evolution and oxygen adaptation. *Science* 2025.

85. Shock EL, Boyd ES. Principles of Geobiochemistry. *Elements (Que)* 2015; **11**: 395–401.

86. Miller RW, Eady RR. Molybdenum and vanadium nitrogenases of Azotobacter chroococcum. Low temperature favours N2 reduction by vanadium nitrogenase. *Biochem J* 1988; **256**: 429–432.

87. Walmsley J, Kennedy C. Temperature-dependent regulation by molybdenum and vanadium of expression of the structural genes encoding three nitrogenases in Azotobacter vinelandii. *Appl Environ Microbiol* 1991; **57**: 622–624.

88. Darnajoux R, Bradley R, Bellenger J-P. In vivo temperature dependency of molybdenum and vanadium nitrogenase activity in the heterocystous Cyanobacteria Anabaena variabilis. *Environ Sci Technol* 2022; **56**: 2760–2769.

89. Laso-Pérez R, Wu F, Crémière A, Speth DR, Magyar JS, Zhao K, et al. Evolutionary diversification of methanotrophic ANME-1 archaea and their expansive virome. *Nat Microbiol* 2023; **8**: 231–245.

90. Mehta MP, Baross JA. Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. *Science* 2006; **314**: 1783–1786.

91. Dekas AE, Poretsky RS, Orphan VJ. Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. *Science* 2009; **326**: 422–426.

92. Ivanovsky RN, Lebedeva NV, Keppen OI, Tourova TP. Nitrogen metabolism of an anoxygenic filamentous phototrophic bacterium oscillocholris trichoides strain DG-6. *Microbiology* 2021; **90**: 428–434.

93. Chen Y, Nishihara A, Haruta S. Nitrogen-fixing ability and nitrogen fixation-related genes of thermophilic fermentative bacteria in the genus Caldicellulosiruptor. *Microbes Environ* 2021; **36**.

94. Thiel V, Wood JM, Olsen MT, Tank M, Klatt CG, Ward DM, et al. The dark side of the Mushroom Spring microbial mat: Life in the shadow of chlorophototrophs. :I. microbial diversity based on 16S RRNA gene amplicons and metagenomic sequencing. 2016. CRC Press, Boca Raton, FL.

95. Payá-Tormo L, Echavarri-Erasun C, Makarovsky-Saavedra N, Pérez-González A, Yang Z-Y, Guo Y, et al. Iron-molybdenum cofactor synthesis by a thermophilic nitrogenase devoid of the scaffold NifEN. *Proc Natl Acad Sci U S A* 2024; **121**: e2406198121.

96. Franche C, Lindström K, Elmerich C. Nitrogen-fixing bacteria associated with leguminous and non-leguminous plants. *Plant Soil* 2009; **321**: 35–59.

97. Bennett AB, Pankievicz VCS, Ané J-M. A model for nitrogen fixation in cereal crops. *Trends Plant Sci* 2020; **25**: 226–235.

98. Bueno Batista M, Dixon R. Manipulating nitrogen regulation in diazotrophic bacteria for agronomic benefit. *Biochem Soc Trans* 2019; **47**: 603–614.

99. Ferguson BJ, Indrasumunar A, Hayashi S, Lin M-H, Lin Y-H, Reid DE, et al.

827 Molecular analysis of legume nodule development and autoregulation. *J Integr*
828 *Plant Biol* 2010; **52**: 61–76.
829 100. Steele JH, Brink KH, Scott BE. Comparison of marine and terrestrial ecosystems:
830 suggestions of an evolutionary perspective influenced by environmental variation.
831 *ICES J Mar Sci* 2019; **76**: 50–59.
832 101. Dynarski KA, Houlton BZ. Nutrient limitation of terrestrial free-living nitrogen
833 fixation. *New Phytol* 2018; **217**: 1050–1061.
834 102. Zheng M, Zhou Z, Zhao P, Luo Y, Ye Q, Zhang K, et al. Effects of human
835 disturbance activities and environmental change factors on terrestrial nitrogen
836 fixation. *Glob Chang Biol* 2020; **26**: 6203–6217.
837 103. Yan G, Xing Y, Han S, Zhang J, Wang Q, Mu C. Long-time precipitation
838 reduction and nitrogen deposition increase alter soil nitrogen dynamic by influencing
839 soil bacterial communities and functional groups. *Pedosphere* 2020; **30**: 363–377.
840 104. Bellenger JP, Darnajoux R, Zhang X, Kraepiel AML. Biological nitrogen fixation
841 by alternative nitrogenases in terrestrial ecosystems: a review. *Biogeochemistry*
842 2020; **149**: 53–73.
843 105. Chen M-Y, Teng W-K, Zhao L, Han B-P, Song L-R, Shu W-S. Phylogenomics
844 uncovers evolutionary trajectory of nitrogen fixation in Cyanobacteria. *Mol Biol Evol*
845 2022; **39**.
846 106. Darnajoux R, Constantin J, Miadlikowska J, Lutzoni F, Bellenger J-P. Is
847 vanadium a biometal for boreal cyanolichens? *New Phytol* 2014; **202**: 765–771.
848 107. Sriswasdi S, Yang C-C, Iwasaki W. Generalist species drive microbial dispersion
849 and evolution. *Nat Commun* 2017; **8**: 1162.

850 108. Chen Y-J, Leung PM, Wood JL, Bay SK, Hugenholtz P, Kessler AJ, et al.
851 Metabolic flexibility allows bacterial habitat generalists to become dominant in a
852 frequently disturbed ecosystem. *ISME J* 2021; **15**: 2986–3004.
853 109. Carbonero F, Oakley BB, Purdy KJ. Metabolic flexibility as a major predictor of
854 spatial distribution in microbial communities. *PLoS One* 2014; **9**: e85105.
855 110. Van Tienderen PH. Evolution of generalists and specialists in spatially
856 heterogeneous environments. *Evolution* 1991; **45**: 1317–1331.
857 111. Poisot T, Bever JD, Nemri A, Thrall PH, Hochberg ME. A conceptual framework
858 for the evolution of ecological specialisation: Evolution of ecological specialisation.
859 *Ecol Lett* 2011; **14**: 841–851.
860 112. Hu W, Wang X, Xu Y, Wang X, Guo Z, Pan X, et al. Biological nitrogen fixation
861 and the role of soil diazotroph niche breadth in representative terrestrial
862 ecosystems. *Soil Biol Biochem* 2024; **189**: 109261.
863 113. Kloesges T, Popa O, Martin W, Dagan T. Networks of gene sharing among 329
864 proteobacterial genomes reveal differences in lateral gene transfer frequency at
865 different phylogenetic depths. *Mol Biol Evol* 2011; **28**: 1057–1074.
866 114. Waite DW, Vanwonterghem I, Rinke C, Parks DH, Zhang Y, Takai K, et al.
867 Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed
868 Reclassification to Epsilonbacteraeota (phyl. nov.). *Front Microbiol* 2017; **8**: 682.
869 115. Moya-Beltrán A, Beard S, Rojas-Villalobos C, Issotta F, Gallardo Y, Ulloa R, et
870 al. Genomic evolution of the class Acidithiobacillia: deep-branching Proteobacteria
871 living in extreme acidic conditions. *ISME J* 2021; **15**: 3221–3238.
872 116. Campbell BJ, Smith JL, Hanson TE, Klotz MG, Stein LY, Lee CK, et al.

873 Adaptations to submarine hydrothermal environments exemplified by the genome of
874 Nautilia profundicola. *PLoS Genet* 2009; **5**: e1000362.
875 117. González-Rosales C, Vergara E, Dopson M, Valdés JH, Holmes DS. Integrative
876 genomics sheds light on evolutionary forces shaping the Acidithiobacillia class
877 acidophilic lifestyle. *Front Microbiol* 2021; **12**: 822229.
878 118. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria.
879 *Nat Rev Microbiol* 2011; **10**: 13–26.
880 119. Cornejo-Castillo FM, Inomura K, Zehr JP, Follows MJ. Metabolic trade-offs
881 constrain the cell size ratio in a nitrogen-fixing symbiosis. *Cell* 2024; **187**: 1762-
882 1768.e9.
883 120. Coale TH, Loconte V, Turk-Kubo KA, Vanslembrouck B, Mak WKE, Cheung S, et
884 al. Nitrogen-fixing organelle in a marine alga. *Science* 2024; **384**: 217–222.
885 121. Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, et al. Metabolic
886 streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 2010; **464**:
887 90–94.
888 122. Zehr JP, Shilova IN, Farnelid HM, Muñoz-MarínCarmen MDC, Turk-Kubo KA.
889 Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium
890 UCYN-A. *Nat Microbiol* 2016; **2**: 16214.
891 123. Kantor EJH, Robicheau BM, Tolman J, Archibald JM, LaRoche J. Metagenomics
892 reveals the genetic diversity between sublineages of UCYN-A and their algal host
893 plastids. *ISME Commun* 2024; **4**: ycae150.
894 124. Tschitschko B, Esti M, Philippi M, Kidane AT, Littmann S, Kitzinger K, et al.
895 Rhizobia-diatom symbiosis fixes missing nitrogen in the ocean. *Nature* 2024; **630**:

896 899–904.
897 125. Bolhuis H, Severin I, Confurius-Guns V, Wollenzien UIA, Stal LJ. Horizontal
898 transfer of the nitrogen fixation gene cluster in the cyanobacterium Microcoleus
899 chthonoplastes. *ISME J* 2010; **4**: 121–130.
900 126. de Crécy-Lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, Blaby
901 I, et al. A roadmap for the functional annotation of protein families: a community
902 perspective. *Database (Oxford)* 2022; **2022**: baac062.
903 127. Anbar AD, Knoll AH. Proterozoic ocean chemistry and evolution: a bioinorganic
904 bridge? *Science* 2002; **297**: 1137–1142.
905 128. Rico KI, Garcia AK, Saito MA, Kaçar B, Anbar AD. Mysteries of metallome
906 evolution: Integrating insights from the Earth and life sciences. In: Anbar A, Weis D
907 (eds). *Treatise on Geochemistry*. 2025. Elsevier, pp 337–364.
908 129. Kaçar B. Reconstructing early microbial life. *Annu Rev Microbiol* 2024; **78**: 463–
909 492.
910 130. Falkowski PG. Life's engines: How microbes made earth habitable. *Life's*
911 *Engines: How Microbes Made Earth Habitable* 2015.