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

Morgan S. Sobol[1,2*], Aya S. Klos[1], 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 earliest 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 requires 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, 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 feedback between genome evolution and shifting environmental and ecological conditions.

Introduction

Biological nitrogen fixation ("N_2 fixation") is an ancient metabolism characterized by the reduction of atmospheric N_2 to bioavailable ammonia [1, 2]. Nitrogenases are the only known enzymes capable of catalyzing this energy-intensive reaction, giving them a crucial historical role expanding the biosphere through nitrogen supply [3, 4].

Phylogenetic reconstructions support a shared ancestry among five major clades of nitrogenases that have diversified over time [5]. These include three isozymes defined by the metal in their active site cofactors, molybdenum (Mo), vanadium (V), or iron (Fe), and the three Mo-nitrogenase clades that differ in host ecology and taxonomy [2, 6, 7]. Structural variations distinguish the Mo-nitrogenase clades and align with historic transitions in environmental oxygen levels [6]. Despite this diversification, nitrogenase's core structure [6] and catalytic mechanism [8, 9] have remained largely conserved over ~3 billion years of evolution [10, 11].

Outside the main nitrogenase complex (*nifHDK* for Mo-nitrogenases), N_2 fixation depends on a myriad of other proteins that have coevolved with it throughout the history of this metabolism [12, 13]. These include proteins involved in cofactor assembly (*nifBEN*) [14], electron transfer (*nifFJ*), transcriptional regulation (*nifA*LI_1I_2), and biosynthesis (*nifXQUVYS*) [15, 16]. Beyond the *nif* cluster, numerous other gene families contribute to N_2 fixation, including those involved in oxygen protection (*nod, ahp, coo, cow, sod*), electron donation and transfer (*fix, rng, fdx*), transcriptional regulation (*draGT, fixJLK, regRS, ntc, ntr*), metal transport (*mod, feo, cys*), iron-sulfur cluster assembly (*suf, isc, fdx*) [17], as well as nitrogen metabolism (*gln, nir, nap, hox*) [18] (**Figure 1**). Additional genes with potential roles in N_2 fixation remain functionally uncharacterized [19].

The composition of the complex genetic machinery responsible for N_2-fixation varies widely among diazotrophs, with the simplest gene sets found in hyperthermophilic methanogens and the most elaborate networks in aerobes [12]. Previous studies have

highlighted the diverse evolutionary trajectories of nitrogenase-associated genes [13] and revealed a dynamic history of gene acquisitions and losses linked to the emergence of aerobic metabolisms [12]. Nevertheless, a central challenge remains to reconstruct the evolutionary history of N_2 fixation in deep time and relate it to the ecological diversification of diazotrophs. Here, we address this challenge using large-scale comparative genomics and show how the evolution of N_2 fixation genes enabled diazotrophs to expand into diverse and increasingly oxygenated environments.

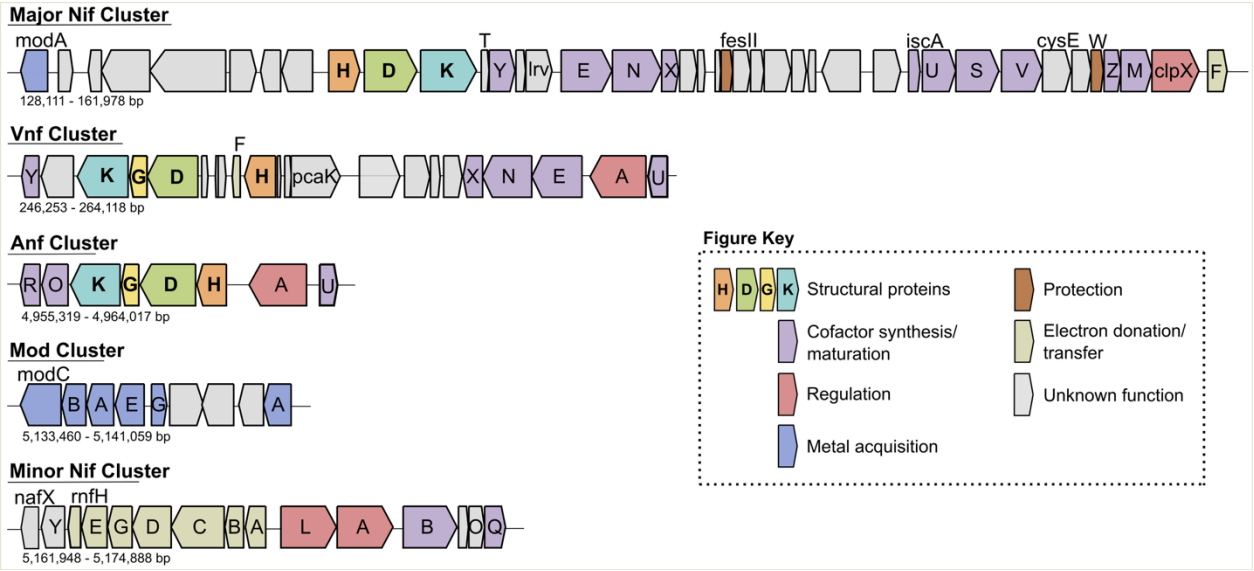


Figure 1. Structure of the major and minor molybdenum nitrogenase (Nif) clusters, vanadium nitrogenase (Vnf) and iron nitrogenase (Anf) clusters, and the molybdenum acquisition (Mod) cluster in model diazotroph *Azotobacter vinelandii* DJ. Genome positions and gene sizes for each cluster are based on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

Methods

Genome selection

The reference phylogeny Web of Life (WoL) [20] 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 [21]. Protein fasta files associated with the genome accessions (named "*_protein.faa.gz") were downloaded from NCBI with eutils

[22] (accessed July 17th, 2023). Genomes missing available NCBI protein fasta files were excluded from analysis. After filtering for high-quality genomes and removing those without annotated fasta files, the WoL genome dataset was reduced 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 [23]. 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 [24] (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 [25] in R. We additionally confirmed the presence of nitrogenase by phylogenetic placement of concatenated *nifDK* sequences, as described in reference [5], 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) [26].

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 [27], NCBI BioProject [28], or GOLD JGI [29] as well as matching the NCBI taxon IDs to the Madin et al (2020) phenotypic trait database [30]. Trait data was also predicted for all genomes. OGTs were predicted using the OGT_prediction tool [31]. 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 [32]. Predictions for pH were done with the Ramoneda et al. (2023) prediction model which is based on genes that influence pH tolerance [33]. Habitat preference and coverage were determined with ProkAtlas [34] by mapping the 16S rRNA gene sequences to the ProkAtlas metagenome database. Barrnap version 0.9 [35] 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 sequencing and assembly bias. Twenty-one genomes without a 16S rRNA gene sequence were excluded. A nucleotide identity of 97% and a query length match of at least 150 base pairs were used as thresholds. Habitat preference, categorized into freshwater, saline, terrestrial, plant-associated, non-plant host-associated, and hydrothermal, was defined as the environment with the highest 16S abundance for each genome. Habitat coverage was determined by summing up the total number of habitats where a genome was considered present and dividing it by the total number of habitats (n=100) investigated by ProkAtlas. The presence of KEGG modules was predicted using the KEGG KO annotations with MetaPathPredict (version 0) [36]. Metabolic coverage was calculated as the total number of KEGG modules present in a genome divided by the total number of modules (n=194) searched for with MetaPathPredict. Data collected from databases and predicted data were cross-referenced and manually annotated to provide the final trait data for each genome.

Data analysis and statistics

138 We compared the mean number of N_2 fixation genes across specific traits (oxygen,
139 temperature, and habitat preferences) with non-parametric methods using the
140 *stat_compare_means* function in the ggpubr R package (version 0.6.0) [37]. For the
141 categorical traits habitat, oxygen, and temperature preference, we first applied a Kruskal-
142 Wallis test, followed by pairwise Wilcoxon rank-sum tests with Benjamini-Hochberg
143 correction. For temperature preference (categories: mesophilic and thermophilic) we
144 used the Wilcoxon rank-sum test directly. Compact letter display was used to report
145 pairwise groups whose means are discernibly different, based on a *P*-value of 0.05.

146 The gene presence-absence matrix was used in phylogenetic principal-component
147 analysis (PCA) performed with the *phyl.pca* function (method = "BM", mode ="corr") in
148 phytools R package [38]. This was done to visualize similarities and dissimilarities
149 between genomes regarding N_2 fixation gene content while accounting for phylogenetic
150 relationships. This approach reduces the risk of over-interpreting similarities simple due
151 to common ancestry among taxa. The first two components were retained for
152 visualization. Pearson correlation between the presence-absence of N_2 fixation genes
153 was calculated with the *cor* function in the stats R package [39] and plotted using corrplot
154 [40] with hierarchical clustering and default *P*-value of 0.05, to identify patterns of gene
155 co-occurrence.

156 ***Evolutionary trends analysis***

157 In order to test whether certain traits shared a coevolutionary history with N_2 fixation gene
158 abundance, phylogenetic linear regression models were fit using the *phylolm* function in
159 the phylolm R package (version 2.62) [41], using Pagel's lambda model of phylogenetic
160 correlation. The lambda parameter (λ) ranges from 0 (no phylogenetic signal) to 1 (high
161 phylogenetic signal) in the response trait when the model has no covariates, or in the
162 residuals when covariates are included. When assessing phylogenetic signal within
163 individual clades, those represented by fewer than 20 genomes were not included in the
164 analysis, as statistical power to detect signal decreases substantially in trees with fewer
165 than 20 branches [42, 43]. Non-phylogenetic linear regressions were performed using the
166 *lm* function in the stats R package [39]. The *distRoot* function in the adephylo R package
167 [44] was used to calculate the root-to-tip patristic distances by measuring the total length

168 of branches (in terms of the number of substitutions per site) from the root of the species
169 tree to each tip.

170 Gene gain/loss was determined with Count [45] using a non-binary count matrix [46] for
171 all accessory (non-structural) *nif* genes. Wagner parsimony with a gain-to-loss penalty
172 ratio of 2 was applied to infer the most parsimonious scenario of gene presence at internal
173 nodes, as described previously [47]. This ratio is widely used because recurring
174 observations indicate that gene loss occurs at least twice as frequently than gene gain
175 [48–50]. Ancestral trait reconstructions were performed on an ultrametric and
176 dichotomous tree, which was made using the *multi2di* and *chronos* (model = "relaxed")
177 functions of the APE package [25] in R. The *fitdiscrete* function in geiger [51] was used
178 for discrete reconstructions of oxygen preference, which were mapped onto the
179 phylogeny with the *simmap* function in phytools [38]. The *fastAnc* function in phytools was
180 used for continuous variable reconstructions of metabolic and habitat diversity, which was
181 mapped onto the tree using the *contMap* function in phytools [38].

182 ***Sampling genomes from non-diazotrophs***

183 To assess whether the observed co-evolutionary trends between traits and N_2 fixation
184 gene abundance were specific to diazotrophs, we analyzed a random sample of non-
185 diazotrophic genomes using the same phylogenetic linear regression methodology
186 described above. Among the 6,785 genomes missing one or more of the catalytic *nifHDK*
187 subunits, 700 were randomly selected with the base R *sample* function (version 4.4.2)
188 [39], for further analysis (**Supplemental Table 9**). Trait data (including OGT, oxygen
189 status, pH preference, and habitat preference) and gene presence/absence for N_2
190 fixation-related genes were collected as described for diazotrophs (**Supplemental**
191 **Tables 9, 10**). Fifty-one genomes without 16S rRNA gene sequences were excluded from
192 analyses involving habitat coverage.
193

194 **<u>Results</u>**

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

196 We used a large prokaryote reference phylogeny, “WoL” [20], containing over 7,000 high-
197 quality genomes [21] to first assess the representative diversity and distribution of
198 nitrogenase today. A total of 673 genomes (~9.0%) were found to contain nitrogenase
199 (**Supplemental Table 2**) based on the presence of the catalytic HDK subunits of
200 nitrogenase (**Figure 2A**). Our estimate of nitrogenase diversity falls within the range of
201 recent estimates (~5.9% to 16%) [7, 12, 52, 53]. Here, approximately 46.9% of these
202 genomes belonged to Proteobacteria (in order of abundance:
203 Alpha>Delta>Gamma>Beta>Epsilon>Acidithiobacillia), while the remaining top three
204 phyla included Firmicutes (17.3%), Cyanobacteria (12.6%), and Euryarchaeota (11.4%).
205 Amongst the nitrogenase-containing genomes, molybdenum (Mo)-only nitrogenase made
206 up the majority (90.3%), followed by genomes with both Mo-nitrogenase and iron (Fe)
207 nitrogenase (5.9%), genomes with both Mo-nitrogenase and vanadium (V) nitrogenase
208 (2.2%), and lastly with 1.5% of genomes containing all three nitrogenases (**Figure 1**)
209 (**Figure 2B**). Genomes with V-nitrogenase were found in 6 phyla: Euryarchaeota,
210 Firmicutes, Alphaproteobacteria, Deltaproteobacteria, Cyanobacteria, and
211 Gammaproteobacteria. Fe-nitrogenases were found in 11 phyla: Firmicutes,
212 Epsilonproteobacteria, Alphaproteobacteria, Gammaproteobacteria, Verrucomicrobia,
213 Euryarchaeota, Deltaproteobacteria, Gammaproteobacteria, Bacteroidetes,
214 Spirochaetes, and Chlorobi. Euryarchaeota represented 50% of the genomes with all
215 three nitrogenases; the remaining phyla included Alphaproteobacteria, Firmicutes, and
216 Gammaproteobacteria.

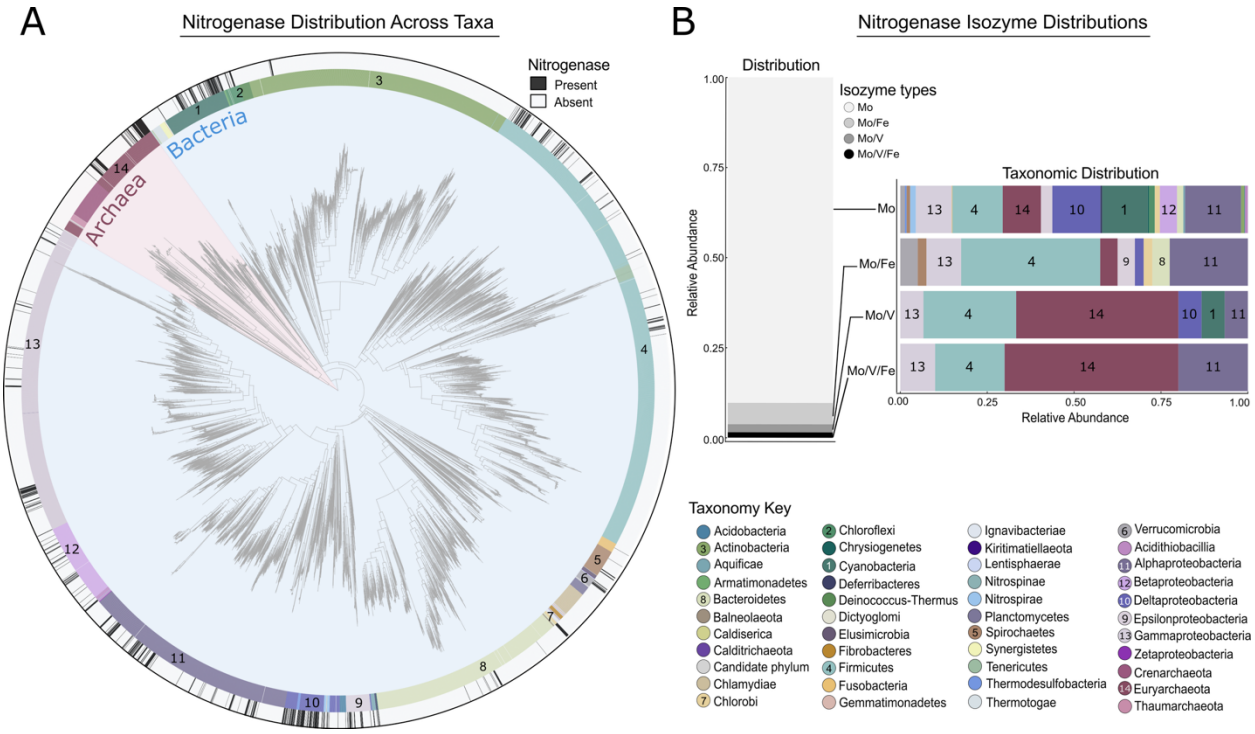


Figure 2. Diversity and phylogenetic distribution of nitrogenase. **A)** The overall distribution of nitrogenase across the microbial diversity from the Zhu et al (2019) Web of Life (WoL) phylogeny [20]. A total of 7,458 high-quality genomes were assessed. Of these, ~9.0% of all genomes (n=673) encoded genes for the full nitrogenase complex (*nifHDK*). 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 [54]. **B)** Proportions and taxonomic distributions of the different nitrogenase isozymes across WoL. Mo-nitrogenase constitutes the majority of nitrogenases identified, followed by Fe-nitrogenase, and then V-nitrogenase in abundance. Only ~1% of genomes contained all three nitrogenases, but ~50% of genomes with all three isozymes belonged to Euryarchaeota.

An N_2 fixation genomic and gene-centric ecological catalog

We screened all 673 genomes with the full *nifHDK* complement for 93 known genes ("machinery") involved in N_2 fixation (**Table 1, Supplemental Table 3, Supplemental Figure 1**). In Figure 3, genes are grouped by functional category in relation to N_2 fixation, and the relative abundance of genes within each category is shown. Excluding the structural genes *nifHDK*, the total number of N_2 fixation genes detected per genome ranged from 14 to 67 genes. The largest N_2 fixation gene repertoires were observed in

235 the Alphaproteobacteria *Rhodopseudomonas palustris CGA009* and the
236 Gammaproteobacteria *Azotobacter vinelandii* DJ, which contained 67 and 65 genes,
237 respectively. In contrast, the archaea taxa *Methanococcus aeolicus* Nankai-3,
238 *Methanothermococcus okinawensis* IH1 and *Methanocaldococcus infernus* ME each
239 possessed the smallest repertoires with 14 genes. Across the species phylogeny, the
240 number of N_2 fixation genes per genome scaled positively with branch lengths, indicating
241 lineages with greater cumulative evolutionary change tend to possess larger N_2 fixation
242 gene complements (**Supplemental Figure 2**).
243
244 **Table 1.** Functional groupings of N_2 fixation machinery genes.

Functional group	**N_2 fixation machinery genes**
Structural	*nifHDK*, *vnfHDGK*, *anfHDGK*
Cofactor synthesis and maturation	*nifBEMNPQVXZ*, *vnfE*, *iscAR*, *iscU/nifU*, *iscS/nifS*, *sufABCDES*
Electron donation and transfer	*fdx*, *fdxA*, *fer*, *fixABCX*, *fldA*, *fnr*, *nifJ*, *rnfABCDEG*
Metal acquisition	*feoABC*, *modABCDEF*, *mosAB*
Protection	*nodABCD*, *ahpCD*, *cooA*, *cowN*, *devAC*, *hghD*, *katG*, *nifW*, *oxyR*, *sod1,2*, *sodN*
Regulation	*draGT*, *fixJLK*, *nif/anf/vnfA*, *nifL*, *nifI$_1$I$_2$*, *nifR3*, *ntcA*, *ntrBC*, *regRS*
Unknown	*nifO*, *nifT*

245

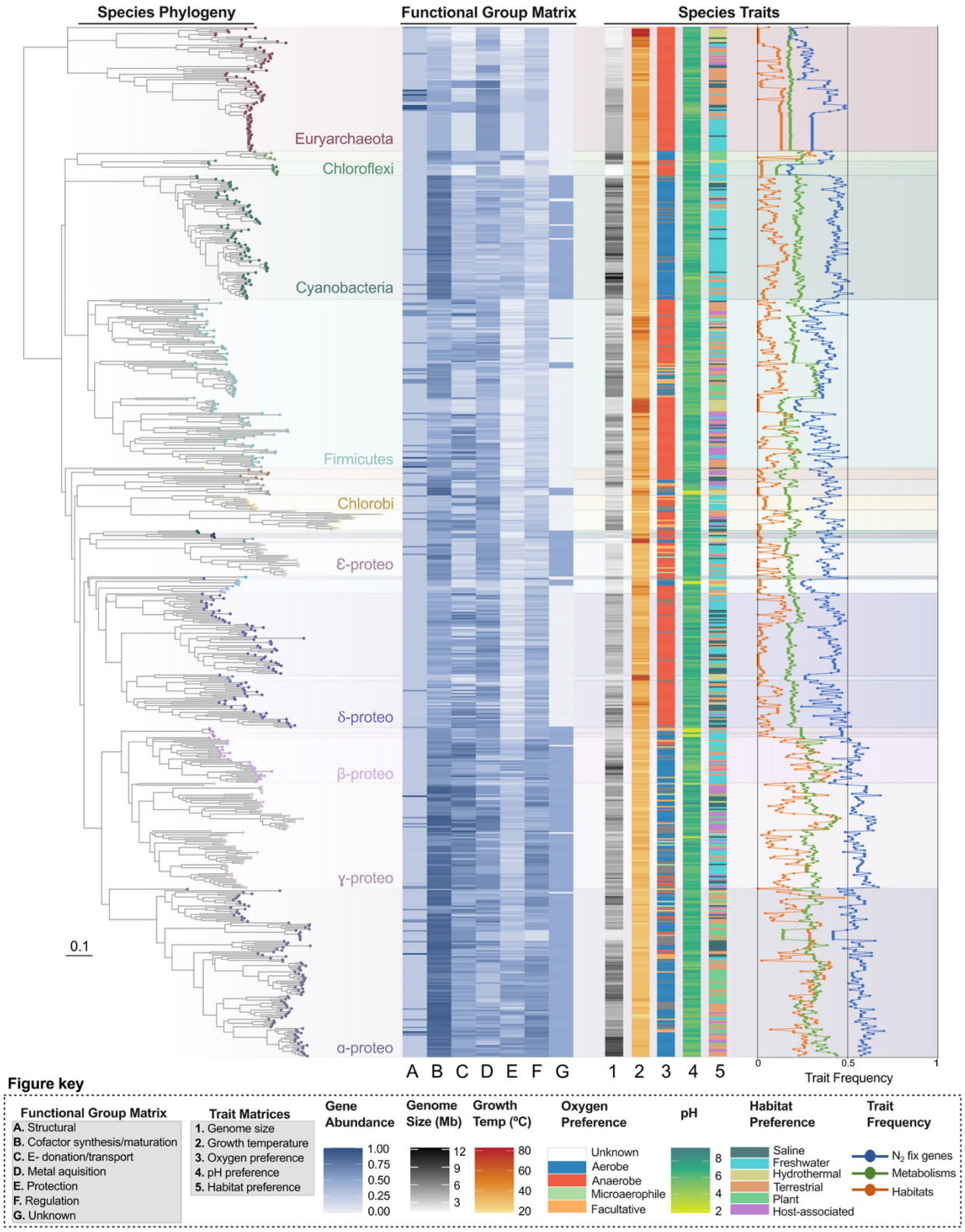


Figure 3. Distribution of N_2 fixation machinery mapped onto the phylogeny of nitrogenase-containing genomes, highlighting associations with species-associated ecological (habitat preference, genome size,

coverage of metabolic pathways) and physiological traits (oxygen, temperature, and pH preferences). N_2 fixation gene abundance correlates with those that are mesophilic, prefer aerobic and/or facultative lifestyles, possess larger genome sizes, broader metabolic capabilities, and occupy a wider range of habitats. For this figure, N_2 fixation genes were organized into functional categories and the relative abundance of genes for each category calculated. The full gene matrix visualization can be found in Supplemental Figure 1 and the gene matrix and associated trait categories in Supplemental Table 3. Trait frequencies were calculated as the relative abundances of N_2 fixation genes (blue line), metabolic pathways (green line), and habitats (orange line) associated with genome hosts. The phylogeny and matrices were plotted using the ggTree R package (version 3.6.2) [26].

To explore the physiological and ecological diversity of diazotrophs identified in this study, we mapped their habitat preferences, oxygen tolerance, optimal growth pH and temperature, metabolic versatility, and genome size onto a phylogenetic tree depicting relationships among diazotroph lineages (**Figure 3**).

Amongst nitrogenase-containing genomes, 47.5% belonged anaerobes, 31.6% to aerobes, 5.3% to facultative species, and 1.5% to microaerophiles. The remaining 14% could not be classified due to missing data or uncertainty in predictions. Temperature preferences further define nitrogenase distribution with most genomes (91.2%) classified as mesophiles favoring growth between 18ºC and 45ºC. The remaining ~8.8% nitrogenase-containing genomes belonged to thermophiles that grow best at temperatures above 45ºC. Similarly, neutrophiles (pH 5.5-8.0) dominated the pH preference dataset (87.2%), whereas acidophiles and alkaliphiles were comparatively rare.

Habitat distributions varied across taxa. Freshwater and terrestrial environments were the most common inferred habitats of preference, encompassing 32.5% and 26.2% of genomes, respectively. Other environments included saline systems (14.3%), non-plant host associations (11.4%), plant-associated niches (8.8%), and hydrothermal systems (6.8%). These patterns likely both reflect genuine ecological preferences and uneven environmental sampling, which may skew our understanding of where nitrogenase-containing organisms are most abundant in nature.

Diazotrophs also differed in their metabolic and ecological breadth. Metabolic coverage ranged from 9.8 to 46.4% with an average of 23.2%, and habitat coverage varied from 0 to 48% with an average of 12%, excluding the ~3% of genomes for which habitat data were unavailable. Genome size spanned from 1.3 to 12.3 Mbp with a mean of 4.6 Mbps.

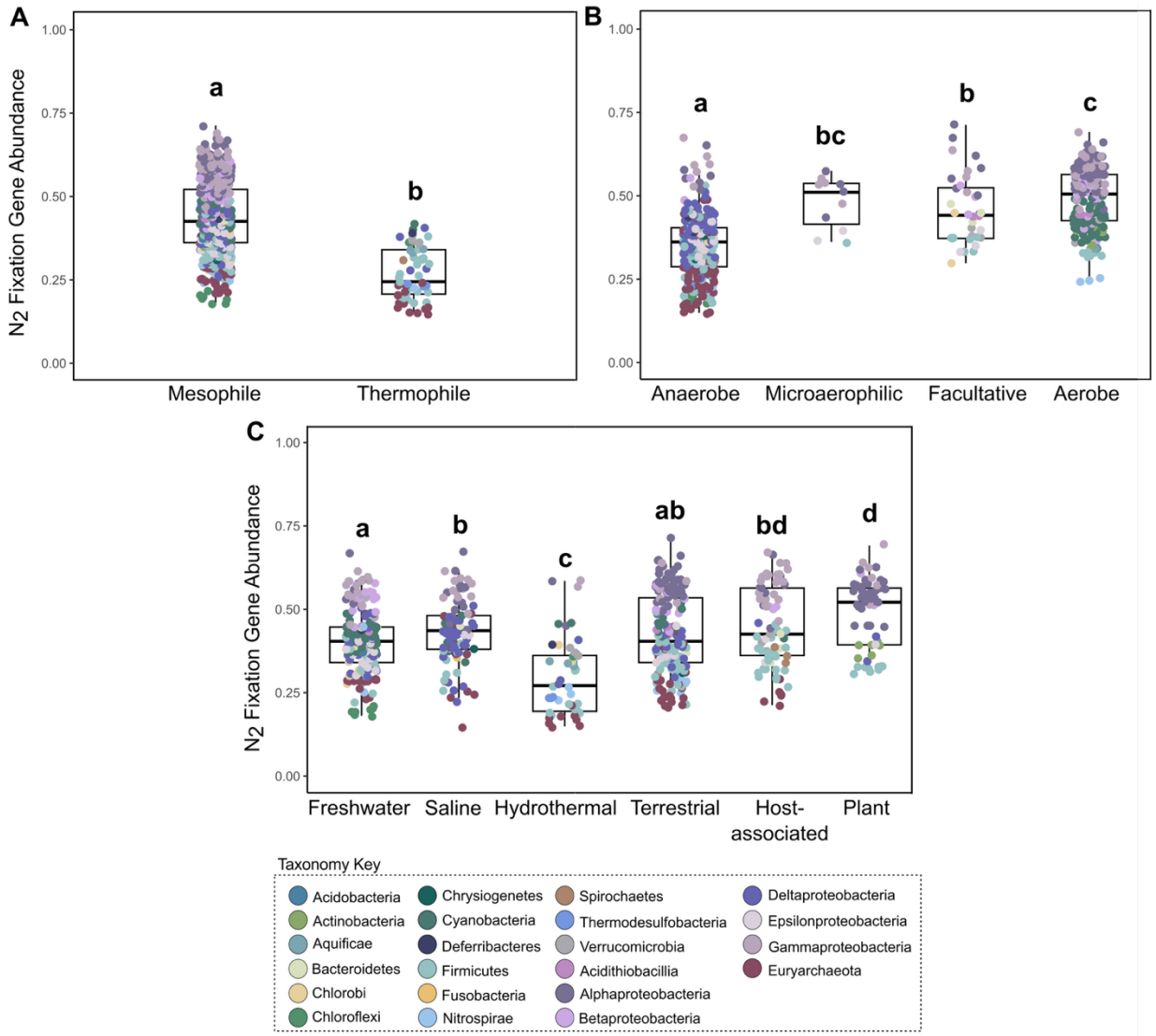


Figure 4. N_2 fixation gene abundance varies across traits. The number of genes was compared across **A)** temperature**, B)** oxygen, and **C)** habitat preferences**.** Thermophiles and anaerobes have discernably fewer N_2 fixation genes than mesophiles and all other oxygen phenotypes, respectively. Genomes associated with hydrothermal environments have on average, the fewest number of N_2 fixation genes, whereas plant-associated genomes have the highest average. 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 significantly different (P<0.05). Boxplots were plotted with the ggpubr R package [37].

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

Next, we examined how the number of N_2 fixation genes per genome relates to environmental and physiological traits, to identify the factors influencing the genetic complexity of the N_2 fixation machinery. Here, genetic complexity is defined as the proportion of N_2 fixation genes detected within a given genome.

We observed that genomes sharing similar temperature, oxygen, and habitat preferences had comparable numbers of N_2 fixation genes (**Figure 3**). Specifically, thermophiles and anaerobes contained significantly fewer genes than their mesophilic (p.adj=8.60×10^{-23}) and aerobic (p.adj=4.80×10^{-45}) counterparts (**Figure 4AB**). Consistently, taxa preferring hydrothermal environments on average have fewer N_2 fixation genes than those from other habitats (**Figure 4C**). No significant differences were detected among terrestrial and freshwater, saline, or non-plant host-associated environments. Plant-associated taxa exhibited the highest average of N_2 fixation gene counts, exceeding those from other environments except non-plant host-associated taxa.

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

To test the degree to which traits are phylogenetically conserved versus independently evolved, we measured phylogenetic signal for gene abundance, habitat coverage, metabolic coverage, genome size, temperature, pH, and oxygen preference across the species phylogeny using Pagel's lambda (λ), as implemented in the *phylolm* R package [41]. All traits had a moderate to high phylogenetic signal (λ≥0.69; **Supplemental Table 4**), indicating closely related genomes tend to share more similar trait values than those drawn randomly from the phylogeny.

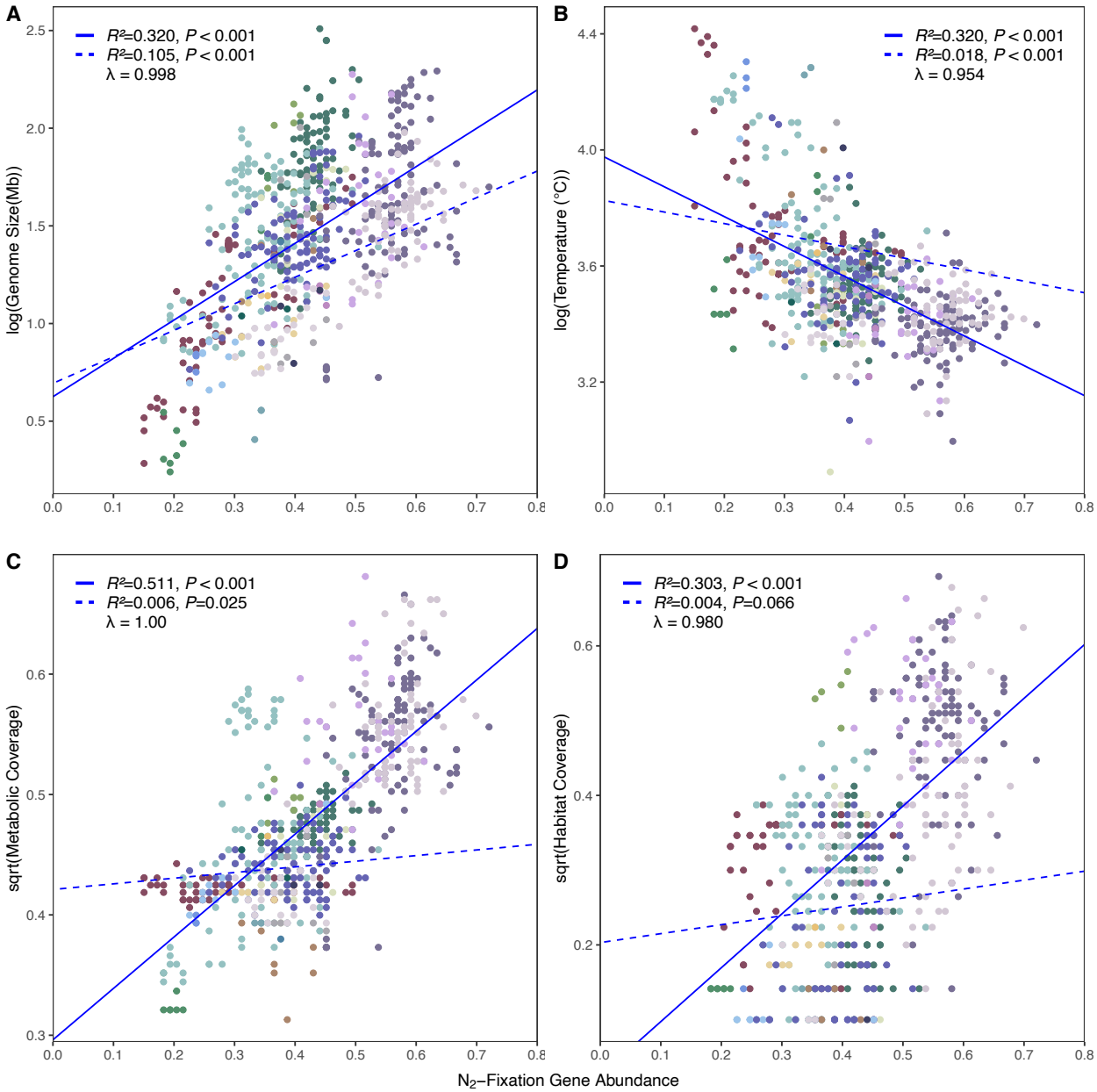
We further assessed phylogenetic signal within individual phyla to test for lineage-specific evolutionary patterns and differences in selection pressures. This analysis clarified

318 whether the overall phylogenetic signal measured across the species phylogeny was
319 disproportionately driven by particular phyla. Because some phyla had too few genomes
320 (<20), phylum-specific analyses were restricted to Alpha-, Beta-, Delta-,
321 Gammaproteobacteria, as well as Cyanobacteria, Firmicutes, and Euryarchaeota.
322 Genome size, habitat coverage, and metabolic coverage all exhibited strong phylogenetic
323 signal across these major nitrogenase-containing phyla, ($\lambda \geq 0.8$, **Supplemental Table 6**),
324 indicating that these traits are broadly conserved across these lineages, rather than
325 confined to a single clade. N_2 fixation gene abundance likewise showed a high λ (>0.8)
326 within most phyla with the exception of Cyanobacteria ($\lambda = 0.34$), potentially due to limited
327 variability of N_2 fixation gene sets within this clade.

328 By contrast, no phylogenetic signal for temperature was detected in
329 Gammaproteobacteria ($\lambda=0.001$), consistent with their largely mesophilic lifestyle,
330 whereas other phyla retained strong signals ($\lambda>0.8$). Oxygen tolerance showed little
331 phylogenetic structure in either Gammaproteobacteria or Firmicutes. In contrast, pH
332 exhibited strong phylogenetic conservation in Cyanobacteria, Euryarcheaota, and
333 Gamma- and Betaproteobacteria, but low conservation in Deltaproteobacteria.

334 Next, we assessed whether certain traits shared an evolutionary history with N_2 fixation
335 gene abundance via phylogenetic linear regression models (**Figure 5, Supplemental**
336 **Figure 3, Supplemental Tables 5, 11**). Gene abundance (calculated as the proportion
337 of the ninety-three N_2 fixation machinery genes screened in this study, per genome) was
338 chosen as the independent variable to determine if the complexity of N_2 fixation genes in
339 a genome could be used to predict other trait values. Response (trait) variables were
340 transformed using either natural log or square root to stabilize variance prior to analysis.
341 Genome size (**Figure 5A, Supplemental Figure 3A**) had a positive, significant linear
342 relationship with N_2 fixation gene abundance for diazotrophs (R^2=0.104, P=2.2×10^{-16}).
343 The same trend was observed screening a random sample of non-diazotrophs from our
344 dataset (R^2=0.23, P=8.3×10^{-42}), suggesting N_2 fixation gene complexity may interact with
345 broader genomic expansion/subtraction forces, irrespective of an organism's ability to fix
346 nitrogen.

347 For temperature, we initially observed a significant, negative linear relationship with N_2
348 fixation gene abundance (**Figure 5B, Supplemental Figure 3B**). Because prior work has
349 shown that adaptation to high temperatures is often linked to genome reduction [55], we
350 repeated the regression using N_2 fixation gene abundance normalized to genome size
351 (**Supplemental Table 2**). When normalized, the negative relationship with temperature
352 disappeared, suggesting that the original pattern likely reflected genome size constraints
353 rather than a temperature specific effect on N_2 fixation. A similar pattern emerged for
354 metabolic and habitat coverage. Both traits showed positive relationships with N_2 fixation
355 gene number, significant for metabolic coverage (R^2=0.006, P=0.025) and marginally
356 significant for habitat coverage (R^2=0.004, P=0.066) (**Figures 5CD, Supplemental**
357 **Figures 3CD**). After accounting for genome size, these relationships disappeared,
358 suggesting that gene abundance scales broadly with genomic content rather than specific
359 ecological traits.



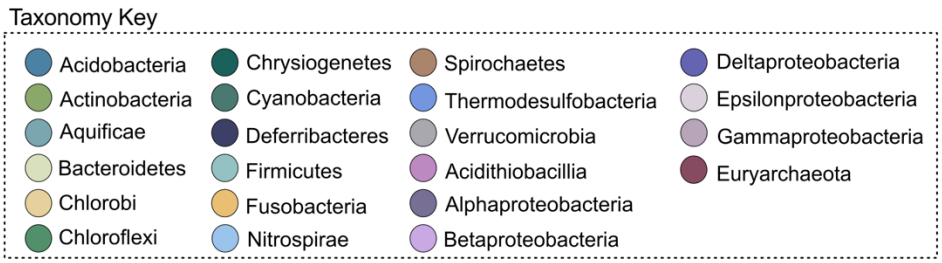


Figure 5. Phylogenetic linear regressions assessed whether the number of N_2 fixation genes in a genome could predict other traits, such as **A)** genome size, **B)** metabolic coverage, **C)** optimal growth temperature,

364 and **D)** habitat coverage. Positive relationships were observed between N_2 fixation gene abundance with
365 genome size, metabolic, and habitat coverage. Response variables were either transformed using natural
366 log or square root to stabilize variance prior analysis. Solid blue lines represent non-phylogenetic linear
367 models, dashed blue lines represent phylogenetic linear models. Correlation coefficients (R^2), associated
368 *P*-values for both regression models, and lambda values are shown for each plot. Scatter plots were plotted
369 using ggplot2 in R [56].

370

371 ***Differences in* N_2 *fixation machinery distribution***

372 We used phylogenetic principal component analysis (PCA) to cluster genomes based on
373 the similarity of their N_2 fixation gene presence-absence patterns, while accounting for
374 shared evolutionary history. Phylogenetic relationships accounted for most of the
375 variation, as indicated by the low variance captured by the phylogenetic PCA
376 (**Supplemental Figure 10**). This finding suggests that clades are clustered primarily
377 because they share conserved gene sets inherited from common ancestors.

378 We then examined the key genes driving separation among genomes in the phylogeny-
379 corrected PCA and interpreted their distribution across taxa and oxygen preferences. The
380 ten genes with the strongest influence on each component of the phylogeny-corrected
381 PCA included the different regulation elements (*nifI*$_2$, *nifL*, *regRS*, *ntrB*), two different
382 electron-donating/transfer mechanisms (*rnfABCDEG, fixABCX*), nodulation and
383 heterocyst differentiation genes (*nodABCD*, *devC*), oxidative protection mechanisms
384 (*sodN*, *cooA*), cofactor synthesis and iron acquisition genes (*nifM*, *sufD*, *feoC2*), and the
385 genes that make up both the Fe- and V-nitrogenases (*vnfHGDKE*, *anfHGDK*). The post-
386 translational *nifI*$_2$ regulator is primarily found in anaerobes, such as Archaea and
387 Firmicutes [12, 13], while other regulatory mechanisms such as *regS*, *nifL*, and *ntrB* are
388 primarily found in aerobic/facultative Proteobacteria (**Supplemental Figure 1**). Pearson
389 correlations of gene co-occurrence (**Supplemental Figure 11**) confirmed that *nifI*$_2$ and
390 *nifI*$_1$ had strong negative correlations with genes *nifTQWXZ*, *fnr*, *fdx*, *oxyR,* and *ntrBC*,
391 many of which were correlated with the ability to use oxygen in diazotrophs [12].

392 A preference for either *rnfACDEG* or *fixACX* electron donation systems was evident in
393 the PCA where the genes pointed opposite from one another (**Supplemental Figure**

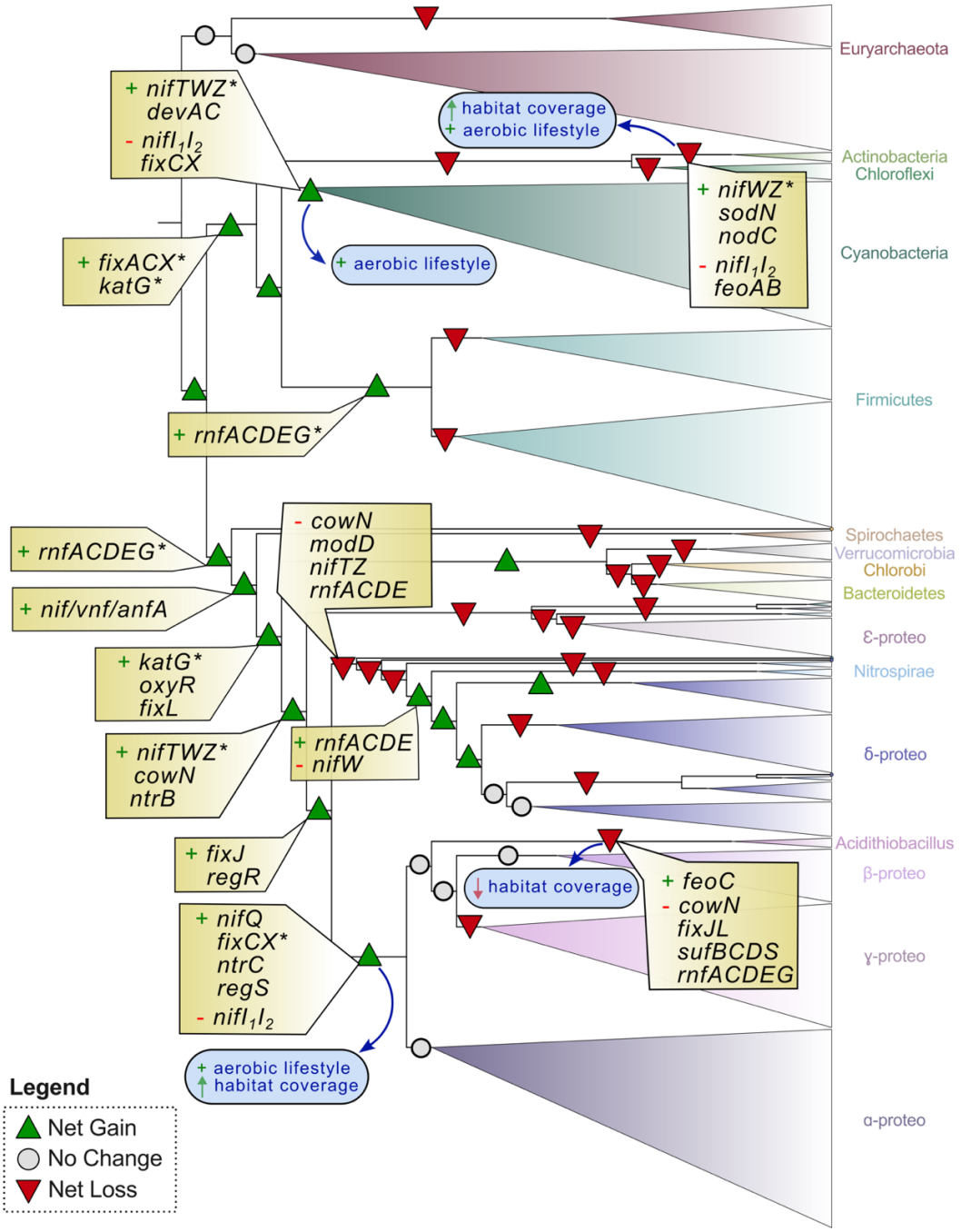
394 **10,11**). Nodulation genes *nodABC* also pointed opposite to *rnfACDEG* and instead had a
395 positive correlation with *fixACX* (**Supplemental Figure 11**). This is confirmed by the fact
396 that diazotrophs forming symbioses with plants through nodulation prefer *fix* genes over
397 *rnf* genes for electron donation [57, 58]. None of the N_2 fixation genes were unique to
398 thermophiles and neither Fe- nor V- nitrogenases were found in genomes with optimal
399 growth temperatures greater than 41ºC.

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

401 We estimated net gain or loss of N_2 fixation machinery genes at each node of the
402 nitrogenase-containing species tree using Wagner parsimony (see methods). Our results
403 show a dynamic history of gene gain and loss over time (**Figure 6**). Specifically, net gene
404 gain was predicted at many of the early branching nodes, whereas gene loss primarily
405 occurred closer to terminal tree branches. This suggests an initial expansion of genes
406 early in diazotroph evolution, followed by successive gene loss events, potentially
407 corresponding to an expansion of diazotroph lineages into new environments where
408 certain components of the N_2 fixation machinery became obsolete.

409 The nodes with the highest number of gene gains (+11) occurs at the most recent
410 common ancestor (MRCA) of Bacteria and at the MRCA of the Euryarchaeota clade
411 Methanosarcina (not shown). Some of the clades that experienced large gene losses
412 include Acidithiobacillia (-18), the Firmicute-Clostridia clade Thermoanaerobacterales (-
413 18) (not shown), Thermodesulfobacteria (-17) (not shown), Nitrospirae (-16), and the
414 MRCA of Aquificae and Epsilonproteobacteria (-11) (**Supplemental Tables 7-8**).
415 Acidothiobacillia's loss of genes *cowN*, *fixJL*, and the Suf and Rnf systems occurs in
416 tandem with a loss of habitat coverage (**Figure 6**). The loss of post-transcriptional
417 regulators $nifI_1I_2$ unsurprisingly coincides with a gain in aerobic lifestyles at the MRCA of
418 Cyanobacteria, Actinobacteria and the MRCA of Gamma- and Alphaproteobacteria
419 (**Figure 6**) as $nifI_1I_2$ is the main regulatory mechanism in anaerobes. The appearance of
420 transcriptional regulation by *nifA*, mostly used by aerobes, predates the loss of $nifI_1I_2$ in
421 the MRCA of Gamma- and Alphaproteobacteria (**Figure 6**). The gain of genes involved
422 in protection against oxidants (*katG*, *oxyR*, and *cowN*) and oxygen response regulators
423 *fixJL* occurs early along the MRCA nodes of Verrucomicrobia, Chlorobi, Bacteroidetes,

424 and Proteobacteria (**Figure 6**). We noted two distinct gains of *rnfACDEG*, *nifTWZ*, *katG*
425 and *fixCX* in distantly related diazotrophs, indicating they were either inherited from a
426 common ancestor that does not have nitrogenase (and therefore would not be
427 represented in this tree), and/or they were gained through horizontal gene transfer.



428

429 **Figure 6.** N_2 fixation gene gain (green triangles) and gene loss (red triangles) throughout the evolution of
430 diazotrophs. Events are mapped to the WoL [20] species tree which was pruned to include only tips
431 associated with diazotrophs identified in this study (n=673). Gene gain dominated the earliest nodes,
432 whereas gene loss was most common near terminal branches. No change (grey circle) represents no gain
433 or loss at that node, or an equal number or gains and losses. Asterisks highlight potential candidates for
434 HGT or genes derived from non-diazotroph ancestors not in this subsetted phylogeny. All gene events can
435 be found in Supplemental table 8. Key transitional events of gene abundance and habitat coverage can be
436 found in Supplemental Figure 12.

437

438 ## Discussion

439 ### *Nitrogenase Distribution, Diversity, and Evolution Across Prokaryotic Lineages*

440 Nitrogenase is found across a taxonomically diverse range of microbes with ecologies
441 spanning root-associated symbiosis in the rhizosphere [59] to deep-sea sediments
442 kilometers below the ocean surface [60–62]. The factors underlying this broad yet uneven
443 distribution remain unclear but likely reflect a complex evolutionary history shaped by
444 dynamic interplay between nitrogenase, its host cell, and the surrounding environment.

445 To investigate these patterns, we developed a comprehensive catalog of 673
446 nitrogenase-encoding genomes and their associated N_2 fixation machinery, tracing
447 historical genomic changes that shaped the evolution of this important metabolism
448 (**Figure 3**). In addition to the *nif* cluster and its accessory genes, we included 63 genes
449 outside of this cluster with reported roles in N_2 fixation, yielding a total of 93 genes
450 representing the N_2 fixation genetic machinery (**Table 1, Supplemental Table 1**). These
451 data were integrated with host-associated trait information (e.g., oxygen metabolism,
452 temperature, pH, genome size, habitat coverage), linking the evolutionary history of
453 nitrogenase and its supporting genes to the environmental diversity of N_2 fixation
454 observed today.

455 Oxygen inhibits nitrogenase activity [63], and modern diazotrophs employ a range of
456 strategies to protect the enzyme from this interference [19, 64, 65]. Consistent with
457 previous findings [12], we observed that on average, aerobic diazotrophs harbor more N_2
458 fixation-associated genes than anaerobes (**Figure 4B**), likely reflecting the additional

459 genetic components required for oxygen protection and regulation under aerobic
460 conditions. For example, we found that genes *katG*, *oxyR*, and *cowN*, which provide
461 protection against oxidants [66, 67], along with the oxygen response regulators *fixJL* [15]
462 were acquired along the lineages leading up to the divergence of Proteobacteria (**Figure**
463 **6**), a period that coincides with the historic accumulation of oxygen in the atmosphere
464 ~2.45 billion years ago known as the Great Oxidation Event (GOE) [68]. This pattern
465 suggests these genes played a role enabling ancient diazotrophs to expand into
466 increasingly oxygenated environments. Recent studies further support the divergence of
467 the alternative nitrogenases during this transition to an oxic atmosphere [69],
468 underscoring how environmental transitions, particularly the historical rise of atmospheric
469 oxygen levels, reshaped the molecular evolution of N_2 fixation in deep time.

470 We recapitulate the relationship between N_2 fixation gene abundance and temperature
471 previously reported by Boyd et al. [12], that on average, thermophiles possess the
472 simplest N_2 fixation-related gene sets (**Figure 4AB, Figure 5B**). Our analyses indicate
473 that this pattern is in part due to genome reduction associated with long-term adaptation
474 to high temperatures [55]. Yet, if the earliest diazotrophs were indeed thermophilic [12],
475 then temperature's role shaping the evolution of N_2 fixation cannot be dismissed. For
476 example, some cold-adapted diazotrophs switch from Mo- to V-based nitrogenases [70–
477 72], indicating how shifts in temperature regimes may have shaped the functional
478 diversification of nitrogenase over evolutionary time.

479 Outside our set of nitrogenase-containing genomes, we identified genomes with partial
480 detection of the core N_2 fixation genes. Notably, several anaerobic thermophiles lack *nifE*,
481 *nifN*, or both (**Figure 3**). These genes are involved in FeMo cofactor assembly and are
482 part of a proposed minimal *nif* gene set (*nifHDKENB)* required for N_2 fixation [73].
483 Lineages exhibiting this partial gene set include Euryarcheaota (such as members of the
484 ANME groups [74]), Chloroflexi, and Firmicutes. Some have verified N_2 fixation activity ,
485 while others, like *Roseiflexus castenholzii*, do not [75–78]. Ancestral sequence
486 reconstructions suggest that *nifDK* may have originated from a duplicated maturase-like
487 *nifEN* ancestor [79], raising the possibility that some extant diazotrophs lacking *nifEN*
488 evolved alternative cofactor maturation strategies, perhaps as a way to maximize energy

489 conservation under thermophilic conditions [55]. Support for this idea comes from
490 *Roseiflexus* sp. RS-1, which lacks *nifEN* and appears to use *nifDK* to fulfill both the
491 cofactor maturase and catalytic roles [80]. Further work will be required to determine
492 whether this streamlined mechanism occurs in other diazotrophs missing *nifEN*.

493 Plant-associated genomes showed the highest N_2 fixation gene counts (**Figure 4C**),
494 consistent with the pivotal role of rhizobia in plant host nitrogen supply [81, 82]. Our
495 reconstructions capture the divergence of Actinobacteria *Frankia* from its ancestor,
496 marked by a transition to aerobic metabolism, accompanying genome and habitat
497 expansion, the loss of $nifI_1I_2$ regulators, and the gain of both nodulation (*nodC*) and
498 oxidative stress response (*sodN*) genes to support its dual free-living and symbiosis
499 states (**Figure 6**).

500 We observed that *Frankia* lacks key regulators such as *nifAL* [17]. However, we noted the
501 presence of the post-translational regulator *draG*, without its partner *draT*, indicating that
502 draG-draT regulation is likely non-functional and perhaps a novel, as-yet-unknown
503 mechanism for nitrogenase inactivation is present [83] – a possibility that warrants further
504 study.

505 In many plant symbioses, oxygen levels are tightly regulated by the host to activate the
506 FixL - FixJ pathway, ultimately driving nitrogenase expression [84], while *nod* genes
507 mediate nodulation [85]. Terrestrial habitats, being more variable than oceans [86],
508 impose fluctuating environmental conditions that demand dynamic nitrogenase regulation
509 [87–89]. Soil molybdenum scarcity likely favors the deployment of alternative nitrogenases
510 [90], contributing to the expansion of gene repertoire in terrestrial environments. Our
511 results illustrate this pattern. For example, several lichen-symbionts within the Nostocales
512 order of Cyanobacteria possess V-nitrogenase [91], likely as a way to cope with limited
513 Mo in high latitude boreal forests [92]. Together, these findings highlight how diazotroph
514 N_2 fixation gene repertoires have been shaped by environmental constraints and the
515 selective pressures of symbiotic lifestyles.

516

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

518 Central to our framework is that N_2 fixation gene complexity is not an isolated trait, rather,
519 it is embedded within a larger evolutionary framework of genome expansion, metabolic
520 versatility, and habitat diversity (**Figures 4, 5**). Gene complexity can even serve as an
521 ecological marker distinguishing generalists, with broad habitats and large, versatile
522 genomes, from specialists with narrow niches and smaller genomes [93–95]. This
523 distinction is particularly relevant today as specialist diazotrophs may contribute
524 disproportionately to total soil N_2 fixation and thereby differentially influence nitrogen input
525 into these environments [96].

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

535 In some cases, N_2 fixation gene loss tracked closely with reduced habitat coverage, as in
536 Epsilonproteobacteria and Acidithiobacillia (**Figure 6**), consistent with prior work on
537 genome streamlining in these clades [97–99]. Adaptation to deep-sea hydrothermal vents
538 [100], shifts toward heterotrophy [97], and transitions to acidophily [101] all show how
539 niche transitions often coincide with simplification of N_2 fixation gene sets. There are
540 exceptions to this trend between genome size and the size of the N_2 fixation machinery.
541 For example, *Zymomonas mobilis* (Alphaproteobacteria) maintains the highest proportion
542 of N_2 fixation genes (~2-3% of all coding genes) despite having a compact ~2Mb genome
543 (**Supplemental Table 2**). Similarly, in symbiont systems, genome reduction often reflects
544 host compensation [102]. UCYN-A2, a marine symbiont recently identified as an early

545 stage “nitroplast” in *Braarudosphaera bigelowii* [103, 104], has lost CO_2 fixation and
546 photosystem II [105, 106], yet retains an intact *nif* cluster (*nifTZVBSHDKENXW)* [107].

547 Horizontal gene transfer (HGT) has further shaped N_2 fixation evolution [11, 108], likely
548 explaining several gene acquisitions we observed (e.g., *rnfACDEG*, *nifTWZ*, *katG* and
549 *fixCX* in **Figure 6**). Previous studies have documented entire *nif* clusters moving across
550 major clades – including a >20kb transfer from Gammaproteobacterium in to the diatom
551 symbiont “*Candidatus* Tectiglobus diatomicola” [86, 109]. In our dataset, *Coleofasciculus*
552 *chthonoplastes* PCC 7420 (formerly *Microcoleus chthonoplastes*) appears to have
553 acquired *nifHDKEN* from a Deltaproteobacterium [88, 110]. Together, these patterns
554 show that the complexity of N_2 fixation genes reflects a balance between the forces of
555 gene loss and gain, specialization, horizontal transfer - all embedded in the broader
556 trajectory of diazotroph genome evolution.

557 ***Limitations of comparative genomics***

558 As with any comparative genomic approach, there are inherent limitations in identifying
559 the full complement of genes and traits within genomes. One potential limitation is that
560 our analysis focuses on genes already recognized to be playing a role in N_2 fixation,
561 meaning additional, yet-undiscovered genes may contribute to the process, particularly
562 in non-model diazotrophs, where many functional mechanisms remain uncharacterized.
563 Another caveat is our reliance on KEGG KO number. Not all genes could be assigned
564 KO numbers, and some KO identifiers fail to distinguish between genes with high
565 sequence similarity, such as *nifA*/*vnfA*/*anfA*. Continued improvements in gene functional
566 annotation and database curation will be essential for building a more complete
567 understanding of diazotroph capabilities [111].

568 ***Conclusion***

569 Examining the distribution of diverse N_2 fixation machinery offers insight into which genes
570 were critical for enabling transitions into new environments throughout evolutionary
571 history. The rise of atmospheric oxygen, for example, fundamentally reshaped electron-
572 donor availability [112] and, together with changes in metal availability [113, 114], likely
573 exerted strong selective pressures on N_2 fixation systems, as demonstrated here.

Reconstructing historical patterns from modern genetic records is thus pivotal for understanding the processes that govern evolution [115]. Global trends in genome evolution not only reveal how organisms responded to past environmental fluctuations but connects them to their present-day ecological niches.

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 [116]. 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 within diazotroph genomes 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 four anonymous reviewers for their constructive feedback.

Data availability

All custom scripts, tree files, raw data, and PDF versions of the figures are available at https://github.com/kacarlab/Sobol_2025_NifGeneEcoEvo. All supplemental tables are available in the uploaded datasets. Specifically, supplementary Tables 2 and 9 include genome accession numbers, taxonomy, and trait data information.

References

1. Canfield DE, Glazer AN, Falkowski PG. The Evolution and Future of Earth's Nitrogen Cycle. *Science* 2010. https://doi.org/10.1126/science.1186120
2. Raymond J et al. The Natural History of Nitrogen Fixation. *Mol Biol Evol* 2004;**21**:541–554. https://doi.org/10.1093/molbev/msh047
3. Navarro-González R, McKay CP, Mvondo DN. A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. *Nature* 2001;**412**:61–64. https://doi.org/10.1038/35083537
4. Koehler MC et al. Transient surface ocean oxygenation recorded in the ~2.66-Ga Jeerinah Formation, Australia. *Proc Natl Acad Sci* 2018;**115**:7711–7716. https://doi.org/10.1073/pnas.1720820115
5. Garcia AK et al. Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization. *Geobiology* 2020;**18**:394–411. https://doi.org/10.1111/gbi.12381
6. Cuevas Zuviría B et al. Structural evolution of nitrogenase over 3 billion years. *eLife* 2025;**14**:RP105613. https://doi.org/10.7554/eLife.105613
7. Koirala A, Brözel VS. Phylogeny of Nitrogenase Structural and Assembly Components Reveals New Insights into the Origin and Distribution of Nitrogen Fixation across Bacteria and Archaea. *Microorganisms* 2021;**9**:1662. https://doi.org/10.3390/microorganisms9081662
8. Garcia AK et al. Nitrogenase resurrection and the evolution of a singular enzymatic mechanism. *eLife* 2023;**12**:e85003. https://doi.org/10.7554/eLife.85003
9. Harris DF et al. Ancient nitrogenases are ATP dependent. *mBio* 2024;**15**:e01271-24. https://doi.org/10.1128/mbio.01271-24
10. Stüeken EE et al. Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature* 2015;**520**:666–669. https://doi.org/10.1038/nature14180

628 11. Parsons C et al. Radiation of nitrogen-metabolizing enzymes across the tree of life tracks
629 environmental transitions in Earth history. *Geobiology* 2021;**19**:18–34.
630 https://doi.org/10.1111/gbi.12419
631 12. Boyd ES et al. Evolution of Molybdenum Nitrogenase during the Transition from
632 Anaerobic to Aerobic Metabolism. *J Bacteriol* 2015;**197**:1690–1699.
633 https://doi.org/10.1128/jb.02611-14
634 13. Nichio BT de L et al. Exploring diazotrophic diversity: unveiling Nif core distribution and
635 evolutionary patterns in nitrogen-fixing organisms. *BMC Genomics* 2025;**26**:81.
636 https://doi.org/10.1186/s12864-024-10994-9
637 14. Kaiser JT et al. Structure of Precursor-Bound NifEN: A Nitrogenase FeMo Cofactor
638 Maturase/Insertase. *Science* 2011;**331**:91–94. https://doi.org/10.1126/science.1196954
639 15. Dixon R, Kahn D. Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol*
640 2004;**2**:621–631. https://doi.org/10.1038/nrmicro954
641 16. Rubio LM, Ludden PW. Biosynthesis of the Iron-Molybdenum Cofactor of Nitrogenase.
642 *Annu Rev Microbiol* 2008;**62**:93–111.
643 https://doi.org/10.1146/annurev.micro.62.081307.162737
644 17. Mus F et al. Exploring the alternatives of biological nitrogen fixation. *Metallomics*
645 2018;**10**:523–538. https://doi.org/10.1039/c8mt00038g
646 18. Kleinschmidt JA, Kleiner D. Relationships between nitrogenase, glutamine synthetase,
647 glutamine, and energy charge in Azotobacter vinelandii. *Arch Microbiol* 1981;**128**:412–
648 415. https://doi.org/10.1007/BF00405923
649 19. Nonaka A et al. Accessory Proteins of the Nitrogenase Assembly, NifW, NifX/NafY, and
650 NifZ, Are Essential for Diazotrophic Growth in the Nonheterocystous Cyanobacterium
651 Leptolyngbya boryana. *Front Microbiol* 2019;**10**. https://doi.org/10.3389/fmicb.2019.00495

652 20. Zhu Q et al. Phylogenomics of 10,575 genomes reveals evolutionary proximity between
653 domains Bacteria and Archaea. *Nat Commun* 2019;**10**:5477.
654 https://doi.org/10.1038/s41467-019-13443-4
655 21. Bowers RM et al. Minimum information about a single amplified genome (MISAG) and a
656 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*
657 2017;**35**:725–731. https://doi.org/10.1038/nbt.3893
658 22. Nadkarni PM, Parikh CR. An eUtils toolset and its use for creating a pipeline to link
659 genomics and proteomics analyses to domain-specific biomedical literature. *J Clin*
660 *Bioinforma* 2012;**2**:9. https://doi.org/10.1186/2043-9113-2-9
661 23. Kanehisa M et al. KEGG as a reference resource for gene and protein annotation.
662 *Nucleic Acids Res* 2016;**44**:D457–D462. https://doi.org/10.1093/nar/gkv1070
663 24. Aramaki T et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and
664 adaptive score threshold. *Bioinformatics* 2020;**36**:2251–2252.
665 https://doi.org/10.1093/bioinformatics/btz859
666 25. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary
667 analyses in R. *Bioinformatics* 2019;**35**:526–528.
668 https://doi.org/10.1093/bioinformatics/bty633
669 26. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinforma*
670 2020;**69**:e96. https://doi.org/10.1002/cpbi.96
671 27. Reimer LC et al. BacDive in 2022: the knowledge base for standardized bacterial and
672 archaeal data. *Nucleic Acids Res* 2022;**50**:D741–D746.
673 https://doi.org/10.1093/nar/gkab961
674 28. Barrett T et al. BioProject and BioSample databases at NCBI: facilitating capture and
675 organization of metadata. *Nucleic Acids Res* 2012;**40**:D57–D63.
676 https://doi.org/10.1093/nar/gkr1163

29. Mukherjee S et al. Genomes OnLine Database (GOLD) v.10: new features and updates. *Nucleic Acids Res* 2025;**53**:D989–D997. https://doi.org/10.1093/nar/gkae1000

30. Madin JS et al. A synthesis of bacterial and archaeal phenotypic trait data. *Sci Data* 2020;**7**:170. https://doi.org/10.1038/s41597-020-0497-4

31. Sauer DB, Wang D-N. Predicting the optimal growth temperatures of prokaryotes using only genome derived features. *Bioinformatics* 2019;**35**:3224–3231. https://doi.org/10.1093/bioinformatics/btz059

32. Jabłońska J, Tawfik DS. The number and type of oxygen-utilizing enzymes indicates aerobic vs. anaerobic phenotype. *Free Radic Biol Med* 2019;**140**:84–92. https://doi.org/10.1016/j.freeradbiomed.2019.03.031

33. Ramoneda J et al. Building a genome-based understanding of bacterial pH preferences. *Sci Adv* 2023;**9**:eadf8998. https://doi.org/10.1126/sciadv.adf8998

34. Mise K, Iwasaki W. Environmental Atlas of Prokaryotes Enables Powerful and Intuitive Habitat-Based Analysis of Community Structures. *iScience* 2020;**23**:101624. https://doi.org/10.1016/j.isci.2020.101624

35. Seemann T. tseemann/barrnap. 2025. 2025.

36. Geller-McGrath D et al. Predicting metabolic modules in incomplete bacterial genomes with MetaPathPredict. *eLife* 2024;**13**:e85749. https://doi.org/10.7554/eLife.85749

37. Alboukadel Kassambara. ggpubr: ‘ggplot2’ Based Publication Ready Plots. 2025. 2025.

38. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative methods (and other things). *PeerJ* 2024;**12**:e16505. https://doi.org/10.7717/peerj.16505

39. R CORE TEAM (2014) R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. - References - Scientific Research Publishing. https://www.scirp.org/reference/referencespapers?referenceid=2145566. .

40. Taiyun Wei, Viliam Simko. R package ‘corrplot’: Visualization of a Correlation Matrix. 2024. 2024.

703 41. Tung Ho L si, Ané C. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait
704 Evolution Models. *Syst Biol* 2014;**63**:397–408. https://doi.org/10.1093/sysbio/syu005
705 42. Rezende EL, Diniz-Filho JAF. Phylogenetic Analyses: Comparing Species to Infer
706 Adaptations and Physiological Mechanisms. *Compr Physiol* 2012;**2**:639–674.
707 https://doi.org/10.1002/j.2040-4603.2012.tb00401.x
708 43. Freckleton RP, Rees M. Comparative analysis of experimental data. *Methods Ecol Evol*
709 2019;**10**:1308–1321. https://doi.org/10.1111/2041-210X.13164
710 44. Jombart T, Balloux F, Dray S. adephylo: new tools for investigating the phylogenetic
711 signal in biological traits. *Bioinformatics* 2010;**26**:1907–1909.
712 https://doi.org/10.1093/bioinformatics/btq292
713 45. Csűrös M. Ancestral Reconstruction by Asymmetric Wagner Parsimony over Continuous
714 Characters and Squared Parsimony over Distributions. In: Nelson CE, Vialette S (eds.)
715 *Comp. Genomics*. 2008. Berlin, Heidelberg: Springer, 2008, pp 72–86.
716 46. Kapust N et al. Failure to Recover Major Events of Gene Flux in Real Biological Data Due
717 to Method Misapplication. *Genome Biol Evol* 2018;**10**:1198–1209.
718 https://doi.org/10.1093/gbe/evy080
719 47. Oyserman BO et al. Ancestral genome reconstruction identifies the evolutionary basis for
720 trait acquisition in polyphosphate accumulating bacteria. *ISME J* 2016;**10**:2931–2945.
721 https://doi.org/10.1038/ismej.2016.67
722 48. Pál C, Papp B, Lercher MJ. Adaptive evolution of bacterial metabolic networks by
723 horizontal gene transfer. *Nat Genet* 2005;**37**:1372–1375. https://doi.org/10.1038/ng1686
724 49. Puigbò P et al. Genomes in turmoil: quantification of genome dynamics in prokaryote
725 supergenomes. *BMC Biol* 2014;**12**:66. https://doi.org/10.1186/s12915-014-0066-4
726 50. Albalat R, Cañestro C. Evolution by gene loss. *Nat Rev Genet* 2016;**17**:379–391.
727 https://doi.org/10.1038/nrg.2016.39

728 51. Pennell MW et al. geiger v2.0: an expanded suite of methods for fitting macroevolutionary
729 models to phylogenetic trees. *Bioinforma Oxf Engl* 2014;**30**:2216–2218.
730 https://doi.org/10.1093/bioinformatics/btu181
731 52. Pi H-W et al. Origin and Evolution of Nitrogen Fixation in Prokaryotes. *Mol Biol Evol*
732 2022;**39**:msac181. https://doi.org/10.1093/molbev/msac181
733 53. Bellanger M et al. NFixDB (Nitrogen Fixation DataBase)-a comprehensive integrated
734 database for robust ’omics analysis of diazotrophs. *NAR Genomics Bioinforma*
735 2024;**6**:lqae063. https://doi.org/10.1093/nargab/lqae063
736 54. Letunic I, Bork P. Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic
737 tree display and annotation tool. *Nucleic Acids Res* 2024;**52**:W78–W82.
738 https://doi.org/10.1093/nar/gkae268
739 55. Sabath N et al. Growth Temperature and Genome Size in Bacteria Are Negatively
740 Correlated, Suggesting Genomic Streamlining During Thermal Adaptation. *Genome Biol*
741 *Evol* 2013;**5**:966–977. https://doi.org/10.1093/gbe/evt050
742 56. Villanueva RAM, Chen ZJ. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Meas*
743 *Interdiscip Res Perspect* 2019;**17**:160–167.
744 https://doi.org/10.1080/15366367.2019.1565254
745 57. Earl CD, Ronson CW, Ausubel FM. Genetic and structural analysis of the Rhizobium
746 meliloti fixA, fixB, fixC, and fixX genes. *J Bacteriol* 1987;**169**:1127–1136.
747 https://doi.org/10.1128/jb.169.3.1127-1136.1987
748 58. Gubler M, Hennecke H. Regulation of the fixA gene and fixBC operon in Bradyrhizobium
749 japonicum. *J Bacteriol* 1988;**170**:1205–1214. https://doi.org/10.1128/jb.170.3.1205-
750 1214.1988
751 59. Smercina DN et al. To Fix or Not To Fix: Controls on Free-Living Nitrogen Fixation in the
752 Rhizosphere. *Appl Environ Microbiol* 2019;**85**:e02546-18.
753 https://doi.org/10.1128/AEM.02546-18

754 60. Reese BK et al. Nitrogen Cycling of Active Bacteria within Oligotrophic Sediment of the
755 Mid-Atlantic Ridge Flank. *Geomicrobiol J* 2018;**35**:468–483.
756 https://doi.org/10.1080/01490451.2017.1392649
757 61. Kapili BJ et al. Evidence for phylogenetically and catabolically diverse active diazotrophs
758 in deep-sea sediment. *ISME J* 2020;**14**:971–983. https://doi.org/10.1038/s41396-019-
759 0584-8
760 62. Semler AC et al. Abundance, Identity, and Potential Diazotrophic Activity of nifH-
761 Containing Organisms at Marine Cold Seeps. *Environ Microbiol* 2025;**27**:e70058.
762 https://doi.org/10.1111/1462-2920.70058
763 63. Gallon JR. Reconciling the incompatible: N2 fixation And O2. *New Phytol* 1992;**122**:571–
764 609. https://doi.org/10.1111/j.1469-8137.1992.tb00087.x
765 64. Milligan AJ et al. Light-dependent oxygen consumption in nitrogen-fixing cyanobacteria
766 plays a key role in nitrogenase protection. *J Phycol* 2007;**43**:845–852.
767 https://doi.org/10.1111/j.1529-8817.2007.00395.x
768 65. Maier RJ, Moshiri F. Role of the Azotobacter vinelandiiNitrogenase-Protective Shethna
769 Protein in Preventing Oxygen-Mediated Cell Death. *J Bacteriol* 2000;**182**:3854–3857.
770 https://doi.org/10.1128/jb.182.13.3854-3857.2000
771 66. Medina MS et al. CowN sustains nitrogenase turnover in the presence of the inhibitor
772 carbon monoxide. *J Biol Chem* 2021;**296**:100501.
773 https://doi.org/10.1016/j.jbc.2021.100501
774 67. Zhao Y et al. OxyR-regulated catalase activity is critical for oxidative stress resistance,
775 nodulation and nitrogen fixation in Azorhizobium caulinodans. *FEMS Microbiol Lett*
776 2016;**363**:fnw130. https://doi.org/10.1093/femsle/fnw130
777 68. Poulton SW et al. A 200-million-year delay in permanent atmospheric oxygenation.
778 *Nature* 2021;**592**:232–236. https://doi.org/10.1038/s41586-021-03393-7

69. Cuevas-Zuviría B et al. Emergence of an Orphan Nitrogenase Protein Following Atmospheric Oxygenation. *Mol Biol Evol* 2024;**41**:msae067. https://doi.org/10.1093/molbev/msae067

70. 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. https://doi.org/10.1042/bj2560429

71. 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. https://doi.org/10.1128/aem.57.2.622-624.1991

72. 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. https://doi.org/10.1021/acs.est.1c05279

73. Dos Santos PC et al. Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics* 2012;**13**:162. https://doi.org/10.1186/1471-2164-13-162

74. Laso-Pérez R et al. Evolutionary diversification of methanotrophic ANME-1 archaea and their expansive virome. *Nat Microbiol* 2023;**8**:231–245. https://doi.org/10.1038/s41564-022-01297-4

75. Mehta MP, Baross JA. Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. *Science* 2006;**314**:1783–1786. https://doi.org/10.1126/science.1134772

76. Dekas AE, Poretsky RS, Orphan VJ. Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. *Science* 2009;**326**:422–426. https://doi.org/10.1126/science.1178223

804 77. Ivanovsky RN et al. Nitrogen Metabolism of an Anoxygenic Filamentous Phototrophic
805 Bacterium Oscillocholris trichoides Strain DG-6. *Microbiology* 2021;**90**:428–434.
806 https://doi.org/10.1134/S0026261721040068
807 78. Chen Y, Nishihara A, Haruta S. Nitrogen-fixing Ability and Nitrogen Fixation-related
808 Genes of Thermophilic Fermentative Bacteria in the Genus Caldicellulosiruptor. *Microbes*
809 *Environ* 2021;**36**:ME21018. https://doi.org/10.1264/jsme2.ME21018
810 79. Garcia AK, Kolaczkowski B, Kaçar B. Reconstruction of Nitrogenase Predecessors
811 Suggests Origin from Maturase-Like Proteins. *Genome Biol Evol* 2022;**14**:evac031.
812 https://doi.org/10.1093/gbe/evac031
813 80. Payá-Tormo L et al. Iron-molybdenum cofactor synthesis by a thermophilic nitrogenase
814 devoid of the scaffold NifEN. *Proc Natl Acad Sci* 2024;**121**:e2406198121.
815 https://doi.org/10.1073/pnas.2406198121
816 81. Franche C, Lindström K, Elmerich C. Nitrogen-fixing bacteria associated with leguminous
817 and non-leguminous plants. *Plant Soil* 2009;**321**:35–59. https://doi.org/10.1007/s11104-
818 008-9833-8
819 82. Bennett AB, Pankievicz VCS, Ané J-M. A Model for Nitrogen Fixation in Cereal Crops.
820 *Trends Plant Sci* 2020;**25**:226–235. https://doi.org/10.1016/j.tplants.2019.12.004
821 83. Bueno Batista M, Dixon R. Manipulating nitrogen regulation in diazotrophic bacteria for
822 agronomic benefit. *Biochem Soc Trans* 2019;**47**:603–614.
823 https://doi.org/10.1042/BST20180342
824 84. Mus F et al. Symbiotic Nitrogen Fixation and the Challenges to Its Extension to
825 Nonlegumes. *Appl Environ Microbiol* 2016;**82**:3698–3710.
826 https://doi.org/10.1128/AEM.01055-16
827 85. Ferguson BJ et al. Molecular analysis of legume nodule development and autoregulation.
828 *J Integr Plant Biol* 2010;**52**:61–76. https://doi.org/10.1111/j.1744-7909.2010.00899.x

86. Steele JH, Brink KH, Scott BE. Comparison of marine and terrestrial ecosystems: suggestions of an evolutionary perspective influenced by environmental variation. *ICES J Mar Sci* 2019;**76**:50–59. https://doi.org/10.1093/icesjms/fsy149

87. Zheng M et al. Effects of human disturbance activities and environmental change factors on terrestrial nitrogen fixation. *Glob Change Biol* 2020;**26**:6203–6217. https://doi.org/10.1111/gcb.15328

88. Dynarski KA, Houlton BZ. Nutrient limitation of terrestrial free-living nitrogen fixation. *New Phytol* 2018;**217**:1050–1061. https://doi.org/10.1111/nph.14905

89. Yan G et al. Long-time precipitation reduction and nitrogen deposition increase alter soil nitrogen dynamic by influencing soil bacterial communities and functional groups. *Pedosphere* 2020;**30**:363–377. https://doi.org/10.1016/S1002-0160(19)60834-9

90. Bellenger JP et al. Biological nitrogen fixation by alternative nitrogenases in terrestrial ecosystems: a review. *Biogeochemistry* 2020;**149**:53–73. https://doi.org/10.1007/s10533-020-00666-7

91. Chen M-Y et al. Phylogenomics Uncovers Evolutionary Trajectory of Nitrogen Fixation in Cyanobacteria. *Mol Biol Evol* 2022;**39**:msac171. https://doi.org/10.1093/molbev/msac171

92. Darnajoux R et al. Is vanadium a biometal for boreal cyanolichens? *New Phytol* 2014;**202**:765–771. https://doi.org/10.1111/nph.12777

93. Sriswasdi S, Yang C, Iwasaki W. Generalist species drive microbial dispersion and evolution. *Nat Commun* 2017;**8**:1162. https://doi.org/10.1038/s41467-017-01265-1

94. Chen Y-J et al. Metabolic flexibility allows bacterial habitat generalists to become dominant in a frequently disturbed ecosystem. *ISME J* 2021;**15**:2986–3004. https://doi.org/10.1038/s41396-021-00988-w

95. Carbonero F, Oakley BB, Purdy KJ. Metabolic Flexibility as a Major Predictor of Spatial Distribution in Microbial Communities. *PLOS ONE* 2014;**9**:e85105. https://doi.org/10.1371/journal.pone.0085105

855 96. Hu W et al. Biological nitrogen fixation and the role of soil diazotroph niche breadth in
856 representative terrestrial ecosystems. *Soil Biol Biochem* 2024;**189**:109261.
857 https://doi.org/10.1016/j.soilbio.2023.109261
858 97. Waite DW et al. Comparative Genomic Analysis of the Class Epsilonproteobacteria and
859 Proposed Reclassification to Epsilonbacteraeota (phyl. nov.). *Front Microbiol* 2017;**8**.
860 https://doi.org/10.3389/fmicb.2017.00682
861 98. Kloesges T et al. Networks of Gene Sharing among 329 Proteobacterial Genomes
862 Reveal Differences in Lateral Gene Transfer Frequency at Different Phylogenetic Depths.
863 *Mol Biol Evol* 2011;**28**:1057–1074. https://doi.org/10.1093/molbev/msq297
864 99. Moya-Beltrán A et al. Genomic evolution of the class Acidithiobacillia: deep-branching
865 Proteobacteria living in extreme acidic conditions. *ISME J* 2021;**15**:3221–3238.
866 https://doi.org/10.1038/s41396-021-00995-x
867 100. Campbell BJ et al. Adaptations to Submarine Hydrothermal Environments Exemplified by
868 the Genome of Nautilia profundicola. *PLOS Genet* 2009;**5**:e1000362.
869 https://doi.org/10.1371/journal.pgen.1000362
870 101. González-Rosales C et al. Integrative Genomics Sheds Light on Evolutionary Forces
871 Shaping the Acidithiobacillia Class Acidophilic Lifestyle. *Front Microbiol* 2022;**12**.
872 https://doi.org/10.3389/fmicb.2021.822229
873 102. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev*
874 *Microbiol* 2011;**10**:13–26. https://doi.org/10.1038/nrmicro2670
875 103. Cornejo-Castillo FM et al. Cyanobacterial symbionts diverged in the late Cretaceous
876 towards lineage-specific nitrogen fixation factories in single-celled phytoplankton. *Nat*
877 *Commun* 2016;**7**:11071. https://doi.org/10.1038/ncomms11071
878 104. Coale TH et al. Nitrogen-fixing organelle in a marine alga. *Science* 2024;**384**:217–222.
879 https://doi.org/10.1126/science.adk1075

880 105. Tripp HJ et al. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium.
881 *Nature* 2010;**464**:90–94. https://doi.org/10.1038/nature08786
882 106. Zehr JP et al. Unusual marine unicellular symbiosis with the nitrogen-fixing
883 cyanobacterium UCYN-A. *Nat Microbiol* 2016;**2**:16214.
884 https://doi.org/10.1038/nmicrobiol.2016.214
885 107. Kantor EJH et al. Metagenomics reveals the genetic diversity between sublineages of
886 UCYN-A and their algal host plastids. *ISME Commun* 2024;**4**:ycae150.
887 https://doi.org/10.1093/ismeco/ycae150
888 108. Wang S et al. Evolutionary Timeline and Genomic Plasticity Underlying the Lifestyle
889 Diversity in Rhizobiales. *mSystems* 2020;**5**:10.1128/msystems.00438-20.
890 https://doi.org/10.1128/msystems.00438-20
891 109. Zehr JP, Capone DG. Changing perspectives in marine nitrogen fixation. *Science* 2020.
892 https://doi.org/10.1126/science.aay9514
893 110. Xie J-B et al. Comparative genomic analysis of N2-fixing and non-N2-fixing Paenibacillus
894 spp.: organization, evolution and expression of the nitrogen fixation genes. *PLoS Genet*
895 2014;**10**:e1004231. https://doi.org/10.1371/journal.pgen.1004231
896 111. de Crécy-Lagard V et al. A roadmap for the functional annotation of protein families: a
897 community perspective. *Database J Biol Databases Curation* 2022;**2022**:baac062.
898 https://doi.org/10.1093/database/baac062
899 112. Anbar AD, Knoll AH. Proterozoic Ocean Chemistry and Evolution: A Bioinorganic Bridge?
900 *Science* 2002;**297**:1137–1142. https://doi.org/10.1126/science.1069651
901 113. Rucker H, Kacar B. Enigmatic evolution of microbial nitrogen fixation: insights from
902 Earth's past. *Trends Microbiol* 2023. https://doi.org/10.1016/j.tim.2023.03.011
903 114. Rico KI et al. Mysteries of metallome evolution: Integrating insights from the Earth and life
904 sciences. *Treatise Geochem* 2024.

115. Kaçar B. Reconstructing Early Microbial Life. *Annu Rev Microbiol* 2024. https://doi.org/10.1146/annurev-micro-041522-103400

116. Falkowski PG. Life's Engines: How Microbes Made Earth Habitable. Princeton University Press, 2015.