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

Life's transition into new habitats has profoundly shaped its trajectory through a reciprocal relationship with Earth's changing environments. To understand how ancient metabolisms co-evolved with these environments, it is essential to identify the ecological and evolutionary processes that structured their underlying genetic networks. Here, we focus on nitrogen (N_2) fixation, one of life's most critical and ancient metabolisms, to investigate the drivers of complexity in its associated genetic network. We used a large-scale comparative genomics framework to construct a comprehensive catalog of today's 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 restructuring of its genetic network, driven by feedbacks between microbial evolution and changing 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 shaping N_2 fixation in both historical and modern contexts remain unclear [14].

The broad distribution of diazotrophs and their habitats reflects their evolution under diverse ecological and physiological demands. The origin of nitrogenase remains unsolved [15], but its core structural and functional features have remained largely conserved over its long evolutionary history [4, 16]. In contrast, numerous genes involved in its function and regulation have diversified [3, 17]. The composition of this genetic "machinery" varies widely across the tree of life, from a small gene set in some genomes, to more than 80 nitrogenase-associated genes in others [7, 18]. This variability suggests that the environmental context and physiological requirements govern the nitrogenase-related genes distribution among taxa [3, 19].

An integrated synthesis that brings together the diversity of N_2 fixation and its associated genes within a broader ecological and evolutionary framework is only beginning to take shape. This integration is essential to connect how the molecular evolution of nitrogenase has been shaped by the interplay of genes within its broader interaction network.

Here, we employed large-scale comparative genomics to investigate how organismal lifestyles influence evolutionary trajectories, particularly in in relation to the varying N_2 fixation-associated gene content. Our findings demonstrate that the evolutionary diversification of the N_2 fixation machinery has facilitated the expansion of diazotrophs into increasingly broader ecological niches over time.

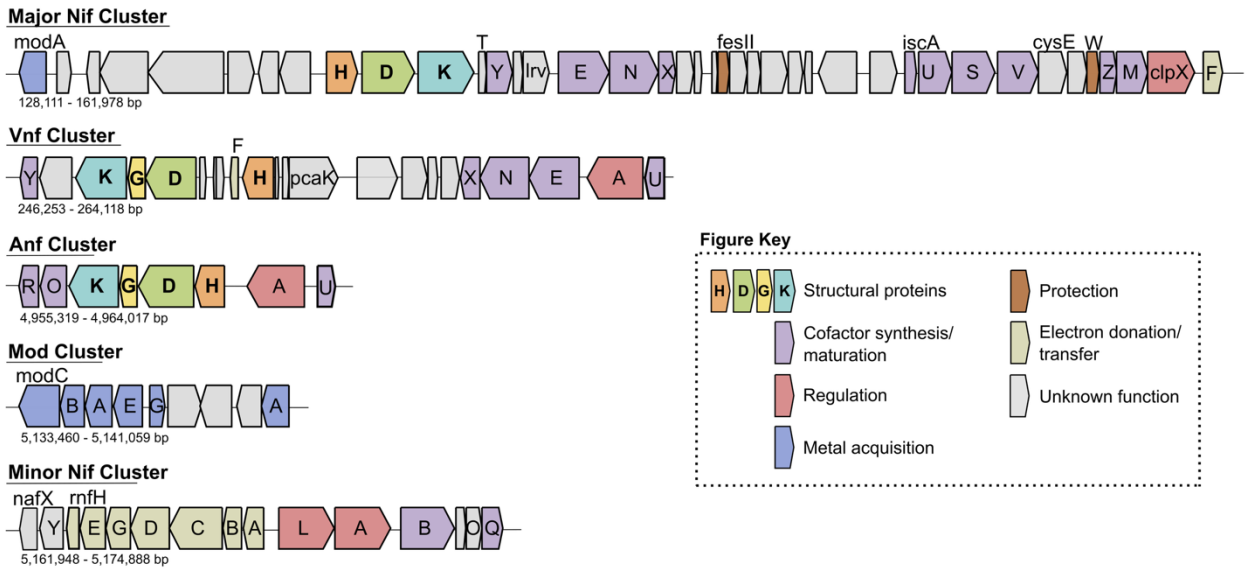


Figure 1. Structure of the major and minor molybdenum nitrogenase (Nif) clusters, vanadium nitrogenase (Vnf) and iron nitrogenase (Anf) clusters, and the molybdenum acquisition (Mod) cluster in model diazotroph *Azotobacter vinelandii* DJ. Genome positions and gene sizes for each cluster are based on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

Curation of nitrogen fixation gene catalog

Protein fasta files were functionally annotated using the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [23]. Proteomes 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 a literature search (**Supplemental Table 1**).

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 [25], NCBI BioProject [26], or GOLD JGI [27] as well as matching the NCBI taxon IDs to the Madin et al (2020) phenotypic trait database [28]. Trait data was also predicted for all genomes. OGTs were predicted using the OGT_prediction tool [29]. 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 [30]. Predictions for pH were done with the Ramoneda et al. (2023) prediction model which is based on genes that influence pH tolerance [31]. Habitat preference and coverage were determined with ProkAtlas [32] by mapping the 16S rRNA gene sequences to the ProkAtlas metagenome database. Barrnap version 0.9 [33] 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 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 investigated by ProkAtlas. The presence of KEGG modules was predicted using the KEGG KO annotations with MetaPathPredict (version 0) [34]. Metabolic ~~breadth~~ coverage was calculated as the total number of KEGG modules present in a genome divided by the total number of modules 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

We compared the mean number of N_2 fixation genes across specific traits (oxygen, temperature, and habitat preferences) with non-parametric methods using the *stat_compare_means* function in the ggpubr R package (version 0.6.0) [35]. For multi-group traits habit and oxygen preference, we first applied a Kruskal-Wallis test, followed by pairwise Wilcoxon rank-sum tests with Benjamini-Hochberg correction. For temperature preference (categories: mesophilic and thermophilic) we used the Wilcoxon rank-sum test directly. Compact letter display was used to report pairwise groups whose means are discernibly different, based on a p-value of 0.05. The *distRoot* function in the adephylo R package [36] was used to calculate the root-to-tip patristic distances by measuring the total length of branches (in terms of the number of substitutions per site) from the root of the species tree to each tip. Non-phylogenetic linear regressions were performed using the *lm* function in the stats R package [37].

The gene presence-absence matrix and trait data were mapped onto the species phylogeny using the ggTree R package (version 3.6.2) [38]. This matrix was also used in phylogenetic principal-component analysis (PCA) performed with the *phyl.pca* function (method = “BM”, mode =”corr”) in phytools R package [39], to visualize similarities and dissimilarities between genomes. The first two components were retained for visualization.

Evolutionary trends analysis

Phylogenetic regression models were fit using the *phylolm* function in the phylolm R package (version 2.62) [40], using Pagel’s lambda model of phylogenetic correlation. The

138 lambda parameter (λ) ranges from 0 (no phylogenetic signal) to 1 (high phylogenetic
139 signal) in the response trait when the model has no covariates, or in the residuals when
140 covariates are included. Gene gain/loss was determined with Count [41] using a non-
141 binary count matrix [42] for all accessory (non-structural) Nif genes. Wagner parsimony
142 with a gain/loss penalty of 2 was applied to infer the most parsimonious scenario of genes
143 present at the internal nodes, as done previously [43]. Ancestral trait reconstructions were
144 performed on an ultrametric and dichotomous tree, which was made using the *multi2di*
145 and *chronos* (model = "relaxed") functions of the APE package [44] in R. The *fitdiscrete*
146 function in geiger [45] was used for discrete reconstructions of oxygen preference, which
147 were mapped onto the phylogeny with the *simmap* function in phytools [39]. The *fastAnc*
148 function in phytools was used for continuous variable reconstructions of metabolic and
149 habitat diversity, which was mapped onto the tree using the *contMap* function in phytools.

150 ***Limitations of comparative genomics***

151 As with any comparative genomic approach, there are inherent limitations when working
152 to identify the genes and traits of genomes. One limitation of this study is the selection of
153 genes that have been identified as playing a role in N_2 fixation, which means there may
154 be numerous other genes involved in the process that remain undiscovered due to the
155 lack of comprehensive functional characterization of many diazotrophs. Additionally, the
156 use of KEGG KO numbers presents another caveat, as not all genes could be assigned
157 KO numbers, and some KO numbers cannot adequately distinguish between genes with
158 high sequence similarities, e.g. *nifA*/*vnfA*/*anfA*. Future improvements in functional
159 annotation of genes and database curation is required to obtain a complete view of the
160 capabilities of diazotrophs [46].

161 ***Data availability***

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

165

166 **Results**

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

168 We used a large prokaryote reference phylogeny, "WoL" [20], containing over 7,000 high-
169 quality genomes [21] to first assess the representative diversity and distribution of
170 nitrogenase today. A total of 673 genomes (~6.4%) were found to contain nitrogenase
171 (**Supplemental Table 2**) based on the presence of the catalytic HDK subunits of
172 nitrogenase (**Figure 2A**). Our estimate of nitrogenase diversity falls within range of recent
173 estimates (~5.9% to 16%) [19, 47–49]. Here, approximately 46.9% of these genomes
174 belonged to Proteobacteria (in order of abundance:
175 Alpha>Delta>Gamma>Beta>Epsilon>Acidithiobacillia), while the remaining top three
176 phyla included Firmicutes (17.3%), Cyanobacteria (12.6%), and Euryarchaeota (11.4%).
177 Amongst the nitrogenase-containing genomes, molybdenum (Mo)-only nitrogenase made
178 up the majority (90.3%), followed by genomes with both Mo-nitrogenase and iron (Fe)
179 nitrogenase (5.9%), genomes with both Mo-nitrogenase and vanadium (V) nitrogenase
180 (2.2%), and lastly with 1.5% of genomes containing all three nitrogenases (Figure 1)
181 (**Figure 2B**). Genomes with V-nitrogenase were found in 6 phyla: Euryarchaeota,
182 Firmicutes, Alphaproteobacteria, Deltaproteobacteria, Cyanobacteria, and
183 Gammaproteobacteria. Fe-nitrogenases were found in 11 phyla: Firmicutes,
184 Epsilonproteobacteria, Alphaproteobacteria, Gammaproteobacteria, Verrucomicrobia,
185 Euryarchaeota, Deltaproteobacteria, Gammaproteobacteria, Bacteroidetes,
186 Spirochaetes, and Chlorobi. Euryarchaeota represented 50% of the genomes with all
187 three nitrogenases; the remaining phyla included Alphaproteobacteria, Firmicutes, and
188 Gammaproteobacteria.

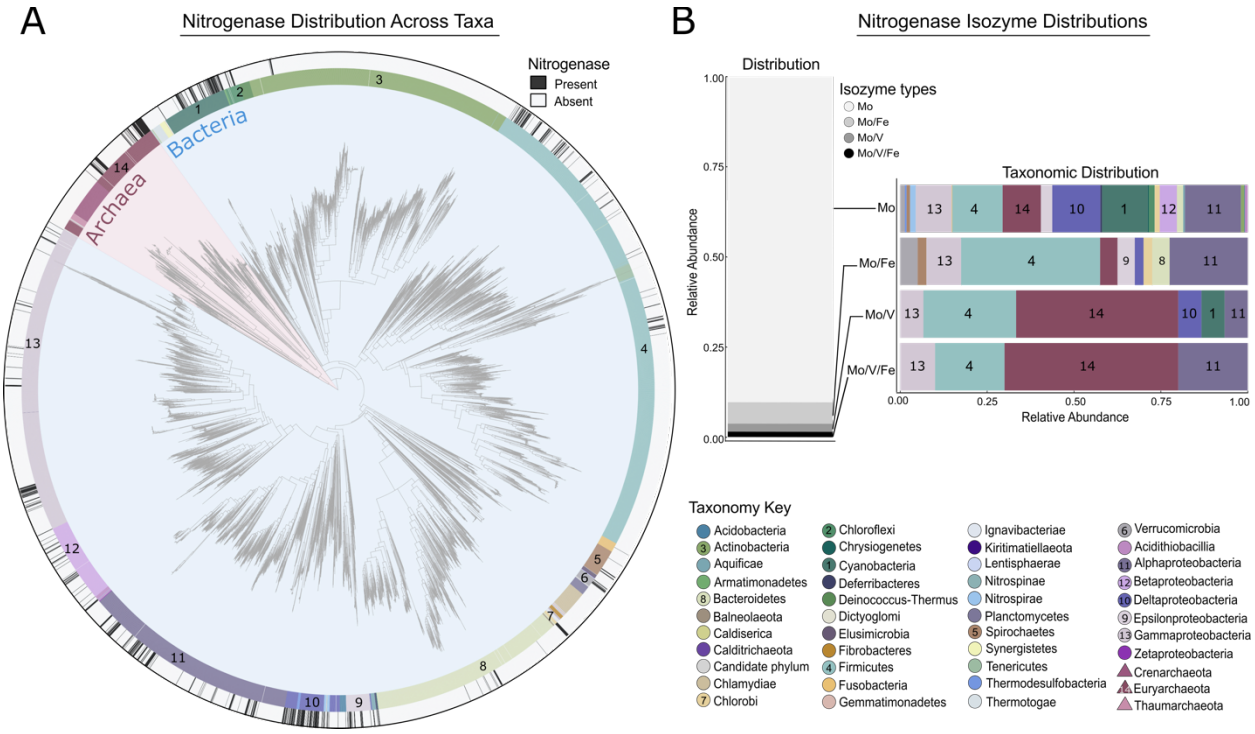


Figure 2. Diversity and phylogenetic distribution of nitrogenase. **A)** The overall distribution of nitrogenase across the microbial diversity from the Zhu et al (2020) Web of Life (WoL) phylogeny. Black bars on the outer ring represent the presence of nitrogenase. Numbers reflect the top 14 taxonomic groups in the dataset with the highest proportion of genomes with nitrogenase. **B)** Proportions and taxonomic distributions of the different nitrogenase isozymes across WoL.

An Ecological and Genomic Survey of N_2 Fixation Genes

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). Throughout the entire species phylogeny, we observed a general trend of increasing number of genes scaled with increasing taxa branch lengths (**Supplemental Figure 2**).

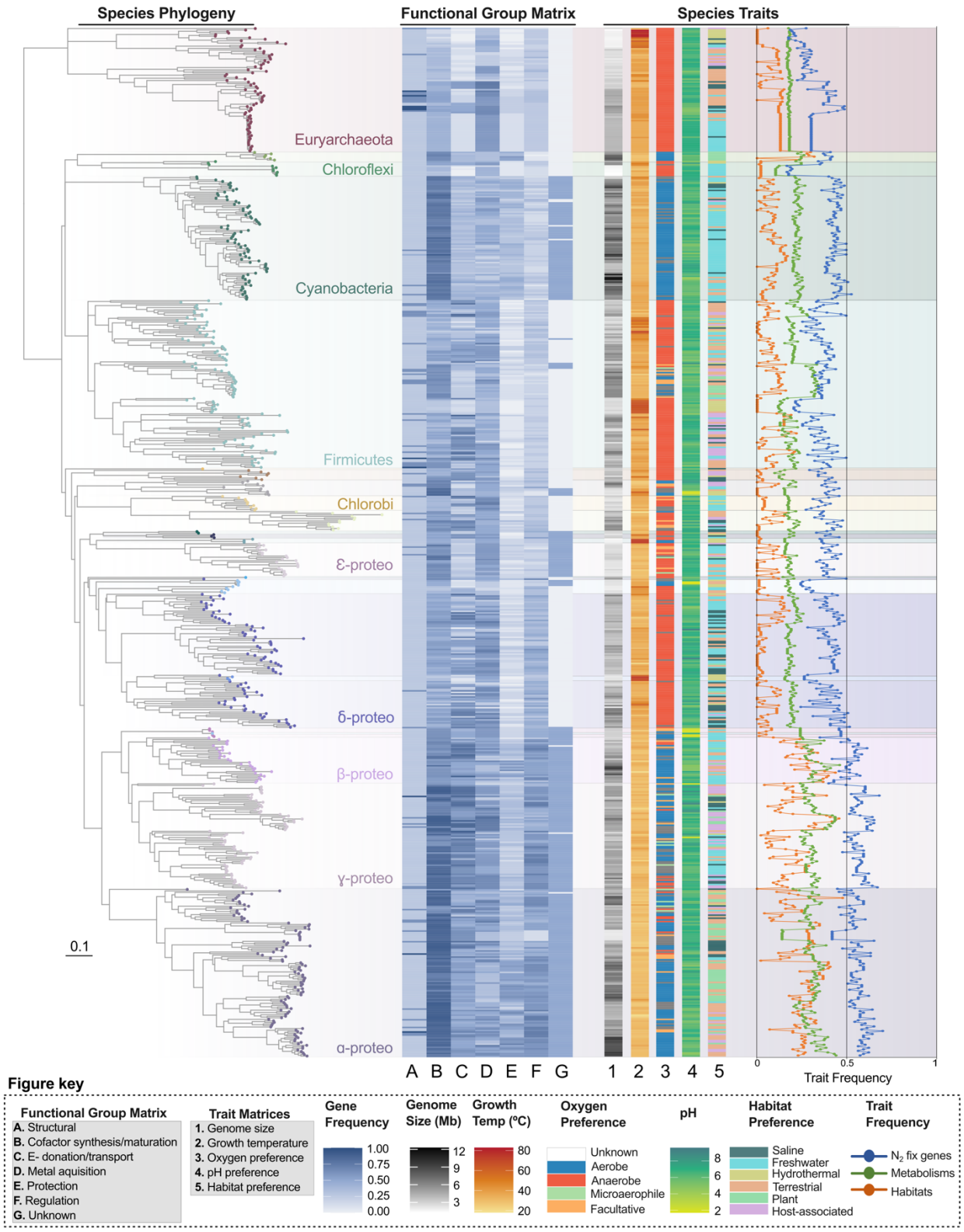


Figure 3. Distribution of N_2 fixation machinery mapped onto the phylogeny of nitrogenase-containing genomes, highlighting associations with species-associated ecological and physiological traits. This pattern

also 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 frequency of genes for each category calculated. The full gene matrix visualization can be found in Supplemental Figure 1 and the gene matrix and associated categories in Supplemental Table 3.

For each genome, we gathered trait data for optimal growth temperature and genome size, as well as preferences for oxygen, pH, and habitats (**Figure 3**). In addition, we calculated the frequency of metabolic pathways encoded in each genome based on the total number of pathways considered (i.e., metabolic coverage) and the proportion of habitats a genome was assigned based on the total number of habitats surveyed (i.e., habitat coverage). To improve trait assignments, we combined data from multiple databases and predicted traits based on genomic features (see methods).

Amongst the nitrogenase-containing genomes, 47.5% of the genomes were anaerobes, 31.6% were aerobes, 5.3% were facultative, and 1.5% were microaerophiles. The remaining 14% could not be determined due to missing data or uncertainty in predictions. Temperature preferences further define nitrogenase distribution, with most genomes (91.2%) belonging to mesophiles that favor temperatures from 18ºC to 45ºC. The remaining ~8.8% belonged to thermophiles that preferred temperatures above 45ºC. Similarly, neutrophiles (pH 5.5-8.0) dominated the pH preferences (87.2%). Acidophiles and alkaliphiles, though present, were not as common. Habitat preferences varied amongst genomes. Here, freshwater and terrestrial environments were particularly favored by 32.5% and 26.2% of genomes, respectively. The remaining genomes were associated with saline environments (14.3%), non-plant host associations (11.4%), associations with plants (8.8%), and hydrothermal systems (6.8%). In addition to these traits, we assessed differences in metabolic and habitat coverage to highlight the ecological versatility of diazotrophs. Metabolic coverage ranged 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.

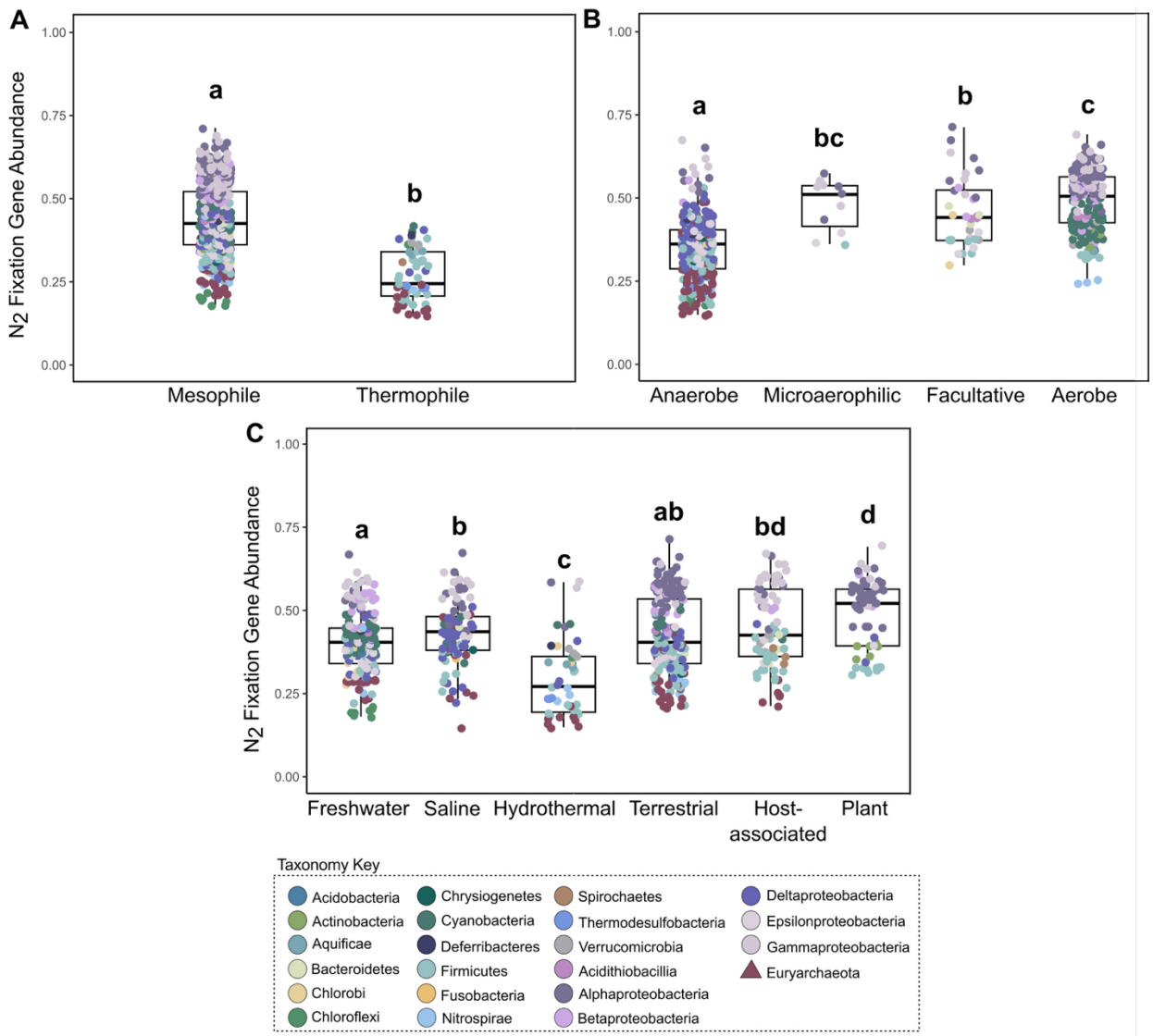


Figure 4. N_2 fixation gene richness varies across traits. The number of genes was compared across **A)** temperature**, B)** oxygen, and **C)** habitat preferences**.** Pairwise differences between means were calculated using non-parametric Wilcoxon rank-sum tests. For each panel, the means of those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

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

Next, we looked at relationships between the number of N_2 fixation genes in a genome and their environmental and physiological traits to understand factors that have altered

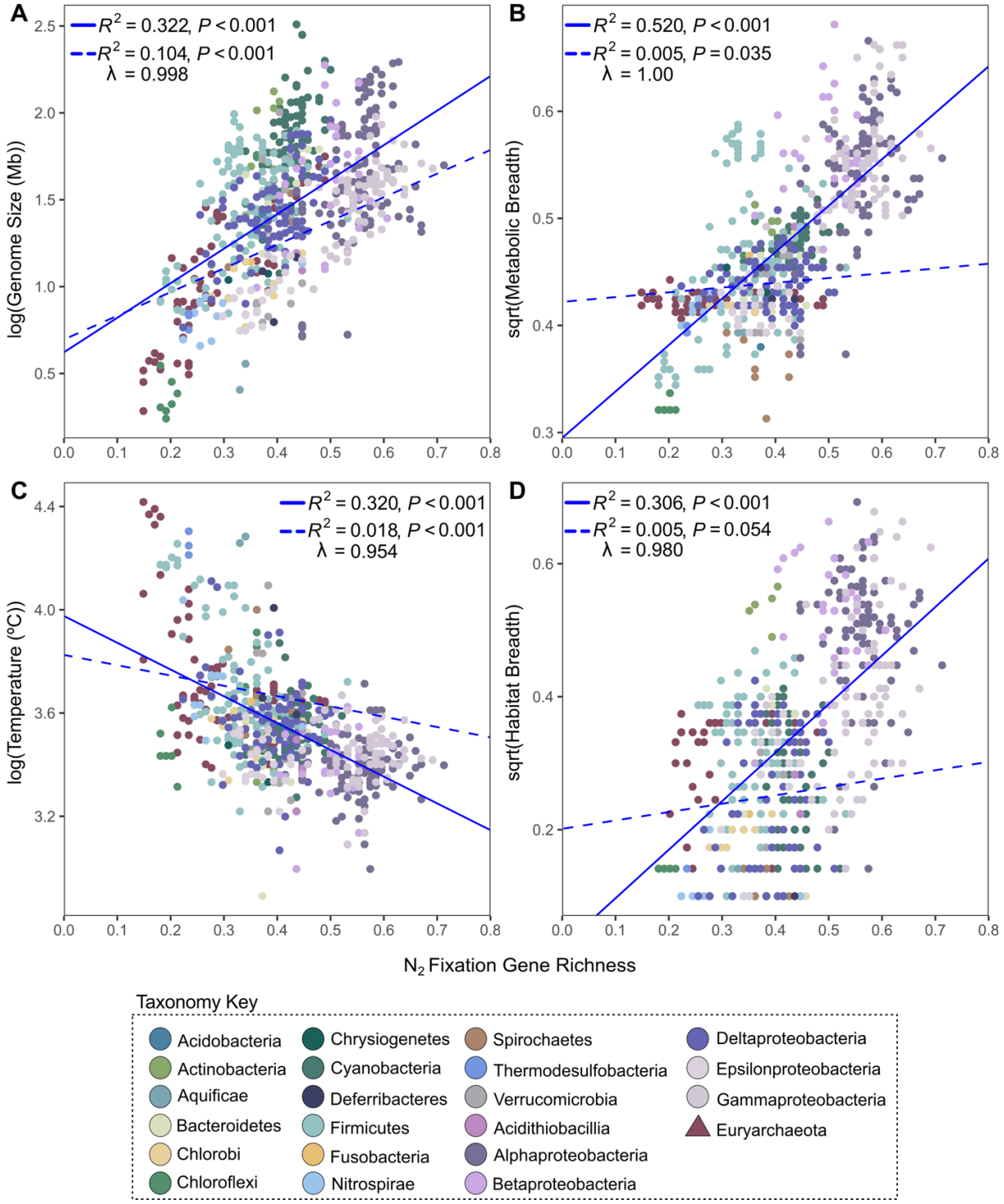
252 the complexity of machinery needed for N_2 fixation. We observed that genomes with
253 similar preferences for temperature, oxygen, and habitat type had comparable numbers
254 of N_2 fixation genes (**Figure 3**). Specifically, thermophiles and anaerobes have discernibly
255 fewer genes compared to their mesophilic (p.adj=8.60×10^{-23}) and aerobic (p.adj=$4.80\times10^{-}$
256 ^{45}) counterparts (**Fig. 4A,B**). Accordingly, organisms who prefer hydrothermal
257 environments have discernibly fewer genes on average compared to all other preferred
258 environments (**Fig. 4C**). There were no discernible differences in gene richness between
259 terrestrial and freshwater, saline, or host-associated environments. Plant-associated taxa
260 had the highest average of N_2 fixation genes, which was discernibly higher than all
261 environments, except host-associated.

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

263 To test the extent to which traits are evolutionarily conserved or have evolved
264 independently, we measured the phylogenetic signal for gene richness, habitat coverage,
265 and metabolic coverage, as well as genome size, temperature, pH, and oxygen
266 preference across the species phylogeny using Pagel's lambda (λ) with the *phylolm* R
267 package [40]. All traits had a moderate to high (λ≥0.69) phylogenetic signal
268 (**Supplemental Table 4**), meaning that closely related genomes demonstrated greater
269 similarity in trait values compared to genomes randomly selected from the phylogeny.

270 Next, we assessed whether certain traits shared an evolutionary history with gene
271 richness via phylogenetic linear regression models (**Figure 5, Supplemental Table 5**).
272 Gene richness was chosen as the independent variable to determine if the complexity of
273 N_2 fixation machinery in a genome could be used to predict other trait values. Response
274 (trait) variables were transformed using either natural log or square root to stabilize
275 variance prior to analysis. Genome size (**Figure 5A**) and metabolic coverage (**Figure 5B**)
276 had a positive, significant linear relationship with N_2 fixation gene richness. As for
277 temperature, a significant, negative linear relationship with gene richness was measured
278 (**Figure 5C**). Outliers (n=15) were identified following log transformation of the
279 temperature variable, but we confirmed that removing these outliers prior to the
280 regression did not influence the strength of statistical evidence, nor the direction of the
281 relationship. Habitat richness was marginally significantly dependent (p=0.054) on gene

number, and had a positive slope, indicating a general relationship between the two traits (**Figure 5D**).



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

290

291 We assessed whether phylogenetic signal at the individual phylum level remained in order
292 to test for phylum-specific evolutionary histories and differing selection pressures. This
293 analysis helped to determine whether the overall phylogenetic signal measured across
294 the entire species phylogeny was driven by specific phyla more so than others. Due to
295 some phyla having too few genomes (<20), phyla-specific analysis was only carried out
296 on genomes assigned to Alpha-, Beta-, Delta-, Gammaproteobacteria, Cyanobacteria,
297 Firmicutes, and Euryarchaeota. Genome size, habitat coverage, and metabolic coverage
298 maintained strong phylogenetic signal across these major nitrogenase-containing phyla
299 ($\lambda \geq 0.8$) (**Supplemental Table 6**). Gene richness had a high λ (>0.8) within most phyla,
300 except Cyanobacteria, where λ was 0.34. No phylogenetic signal for temperature was
301 observed in Gammaproteobacteria ($\lambda=0.001$) as they are all are mesophiles, whereas the
302 other phyla had high signal ($\lambda>0.8$). Oxygen did not have a strong signal in
303 Gammaproteobacteria nor Firmicutes. As for pH, the strength of phylogenetic
304 conservation was high for Cyanobacteria, Euryarcheaota, and Gamma- and
305 Betaproteobacteria, but was low for Deltaproteobacteria.

306 ***Differences in* N_2 *fixation machinery distribution***

307 Phylogenetic principal component analysis (PCA) was used to cluster genomes together
308 based on similarity of N_2 fixation gene presence-absence patterns, while accounting for
309 phylogenetic relationships. Indeed, we found that phylogenetic relationships accounted
310 for most of the variation, indicated by the low explained variance in the phylogenetic PCA
311 (**Supplemental Figure 3**).

312

313 The top ten most influential genes from each component of the phylogeny-corrected PCA
314 (**Supplemental Figure 3**) included the different regulation elements (*nifI*$_2$, *nifL*, *regRS*,
315 *ntrB*), two different electron-donating/transfer mechanisms (*rnfABCDEG, fixABCX*),
316 nodulation and heterocyst differentiation genes (*nodABCD*, *devC*), oxidative protection
317 mechanisms (*sodN*, *cooA*), cofactor synthesis and iron acquisition genes (*nifM*, *sufD*,
318 *feoC2*), and the genes that make up both the Fe- and V-nitrogenase (*vnfHGDKE*,
319 *anfHGDK*). The post-translational *nifI*$_2$ regulator is primarily found in anaerobes, such as
320 Archaea and Firmicutes, whereas *regS*, *nifL*, and *ntrB* are primarily found in aerobic
321 Proteobacteria (**Supplemental Figure 1**). The preference for nitrogenase regulation
322 mechanisms therefore drives part of the separation of aerobes and anaerobes in the PCA
323 (**Supplemental Figure 3**). Pearson correlation of gene co-occurrence (**Supplemental**
324 **Figure 4**) also showed that *nifI*$_2$ and *nifI*$_1$ had strong negative correlations with genes
325 *nifTQWXZ*, *fnr*, *fdx*, *oxyR,* and *ntrBC*, which are genes primarily found in aerobes.

326 A preference for either *rnfACDEG* or *fixACX* electron donation systems was evident in
327 the PCA where the genes pointed opposite from one another (**Supplemental Figure 3**).
328 This anti-correlation was also seen in **Supplemental Figure 2**. Nodulation genes
329 *nodABC* also pointed opposite to *rnfACDEG* (**Supplemental Figure 3**) and instead had
330 a positive correlation with *fixACX* (**Supplemental Figure 4**). This indicates that
331 diazotrophs forming symbioses with plants through nodulation may prefer Fix over Rnf
332 for electron donation. None of the N_2 fixation genes were unique to thermophiles and
333 neither Fe- nor V- nitrogenases were found in genomes with optimal growth temperatures
334 greater than 41ºC.

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

336 Although previous studies have noted the potential dynamic nature of N_2 fixation gene
337 acquisitions and losses [19, 50], none have specifically analyzed these patterns in detail,
338 nor in relation to their traits. By observing gene gain/loss patterns across the phylogeny
339 and linking them to shifts in lifestyle changes, we can begin to observe the selective
340 pressures on the retention or loss of particular genes. We estimated net gain or loss at
341 each node of the nitrogenase-containing species tree using Wagner parsimony (see
342 methods) (**Figure 6**). Evolutionary trajectories of N_2 fixation machinery complexity indeed

343 indicated a dynamic history of gene gain and loss over time. Specifically, early branching
344 nodes imply net gene gain, suggesting an initial expansion of genes early in diazotroph
345 evolution, whereas gene loss primarily occurs closer to the terminal tree branches.

346 The nodes with the highest number of gene gains (+11) occurs at the most recent
347 common ancestor (MRCA) of Bacteria and at the MRCA of the Euryarchaeota clade
348 Methanosarcina (not shown). Some of the clades that experienced large gene losses
349 include Acidithiobacillia (-18), the Firmicute-Clostridia clade Thermoanaerobacterales (-
350 18) (not shown), Thermodesulfobacteria (-17) (not shown), Nitrospirae (-16), and the
351 MRCA of Aquificae and Epsilonproteobacteria (-11) (**Supplemental Tables 7-8**).
352 Acidothiobacillia's loss of genes *cowN*, *fixJL*, and the Suf and Rnf systems occurs in
353 tandem with a loss of habitat coverage (**Figure 6**). The loss of post-transcriptional
354 regulators $nifI_1I_2$ unsurprisingly coincides with a gain in aerobic lifestyles at the MRCA of
355 Cyanobacteria, Actinobacteria and the MRCA of Gamma- and Alphaproteobacteria
356 (**Figure 6**). The appearance of transcriptional regulation by *nifA* predates the loss of $nifI_1I_2$
357 in the MRCA of Gamma- and Alphaproteobacteria (**Figure 6**). The gain of genes involved
358 in protection against oxidants (*katG*, *oxyR*, and *cowN*) and oxygen response regulators
359 *fixJL* occurs early along the MRCA nodes of Verrucomicrobia, Chlorobi, Bacteroidetes,
360 and Proteobacteria (**Figure 6**). We noted two distinct gains of *rnfACDEG*, *nifTWZ*, *katG*
361 and *fixCX* in distantly related diazotrophs, indicating they were either inherited from a
362 common ancestor that does not have nitrogenase (and therefore would not be
363 represented in this tree), and/or they were gained through horizontal gene transfer.

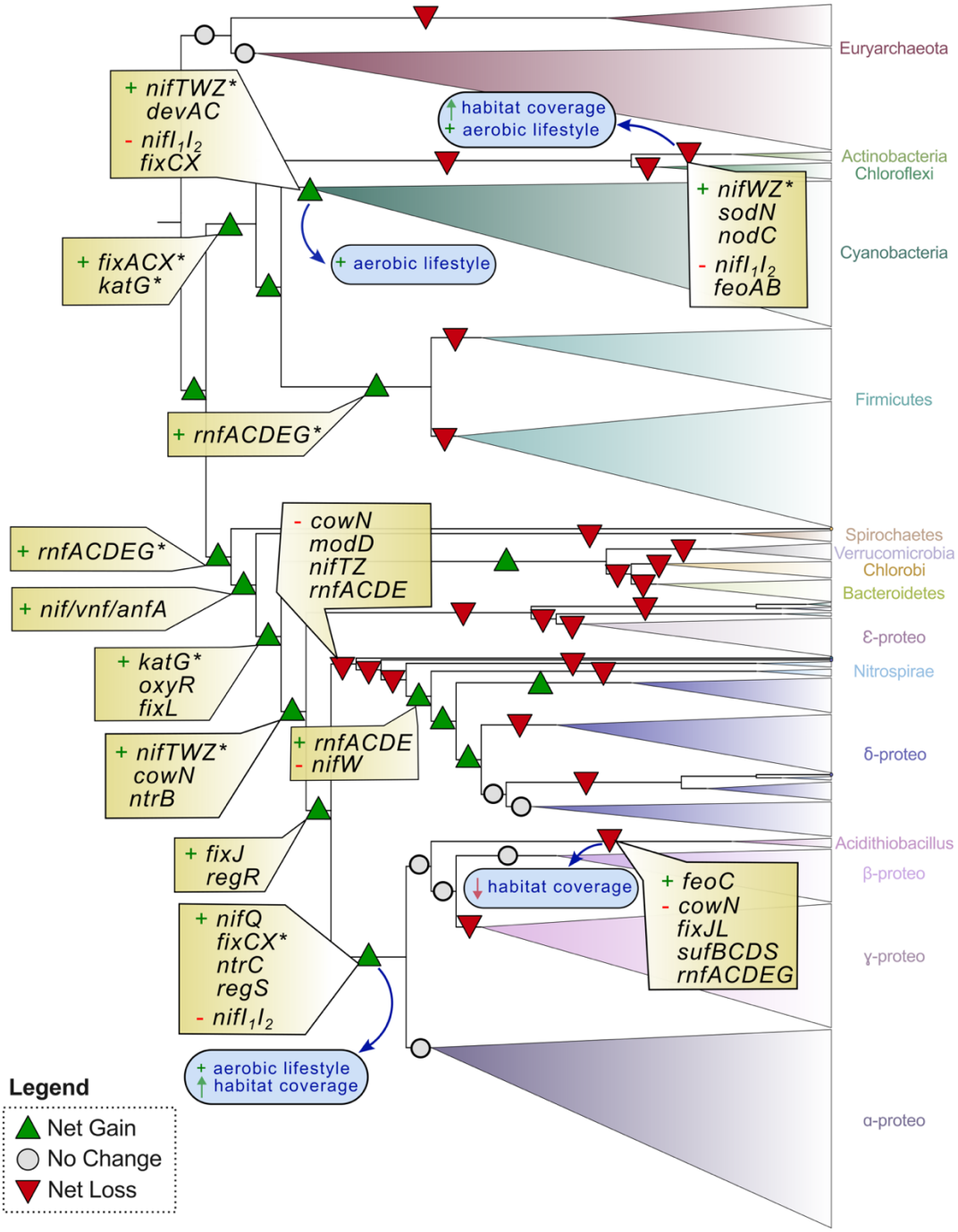


Figure 6. N_2 fixation gene gain (green triangles) and gene loss (red triangles) throughout the evolution of diazotrophs. No change (grey circle) represents no gain or loss at that node, or an equal number or gains and losses. Asterisks denote candidate genes for HGT, or genes potentially derived from non-diazotrophic ancestors not represented in this subsetted phylogeny. All gene events can be found in Supplemental table 8. Key transitional events of gene richness and habitat coverage can be found in Supplemental Figure 5.

370 **Discussion**

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

372 Nitrogenase is widespread across prokaryotes, appearing within phylogenetically and
373 ecologically diverse Bacteria and Archaea (**Figures 2 and 3**). Mo-nitrogenase is the most
374 prevalent form, reported everywhere from the rhizosphere [51] to deep-sea sediments
375 kilometers below the ocean surface [52–54]. Although V- and Fe-nitrogenases occur less
376 frequently, they also span across diverse taxa (**Figure 2**; [55]). However, focusing solely
377 on nitrogenase's distribution overlooks the evolutionary and genetic complexity of N_2
378 fixation.

379 To address this, we developed a comprehensive catalog of nitrogenase-containing
380 genomes and associated N_2 fixation machinery, mapped to environmental preferences
381 and phylogenetic history (**Figure 3**). Expanding on prior work [19], our work incorporates
382 ~2.5× more N_2 fixation genes and genomes, along with data on oxygen metabolism,
383 temperature, pH, genome size, habitat range. Whereas previous comparative studies
384 often targeted single clades or ecosystems [56, 57], our framework unifies nitrogenase
385 and its supporting genes across the entire microbial tree.

386 Oxygen is a well-known inhibitor of nitrogenase [58], supporting its origin under anoxic
387 conditions and later diversification following planetary oxygenation [5, 17]. Consistent with
388 this, aerobic diazotrophs harbor significantly more N_2 fixation genes than anaerobes
389 (**Figure 4B**), likely due to the need for oxygen protection and regulation under aerobic
390 growth (**Figure 3**).

391 Temperature also shapes the complexity of N_2 fixation gene repertoires. Mesophiles
392 harbor more N_2 fixation genes than thermophiles (**Figure 4A**) and hydrothermal
393 organisms exhibit the lowest average counts (**Figure 4B**). These findings align with
394 previous research showing temperature-driven influences on molecular evolution [59–
395 61]. Cold-adapted diazotrophs often switch from Mo- to V-based nitrogenases [62–64]. In
396 further support of this view, we detected a temperature-dependent trend in gene richness
397 across evolutionary histories of nitrogenase-encoding organisms (**Figure 5C**).

398 Notably, several anaerobic thermophiles lack *nifE*, *nifN*, or both (**Figure 3**), genes
399 involved in FeMo cofactor assembly and part of the typical minimal nif genetic set
400 (*nifHDKENB)* required for active N_2 fixation [7]. These lineages include Euryarcheaota,
401 such as members of the ANME groups [65], Chloroflexi, and Firmicutes. Some have
402 verified N_2 fixation activity [66–69], while others, like *Roseiflexus castenholzii*, do not [70].
403 Interestingly, recent work suggests that *nifEN* itself may have originated from a duplicated
404 ancestral nitrogenase [15], raising the possibility that some extant diazotrophs lacking
405 *nifEN* represent primitive traits of early nitrogenase evolution or that they developed
406 genetic strategies to maximize energy conservation [71].

407 Plant-associated genomes showed the highest N_2 fixation gene counts (**Figure 4C**),
408 consistent with their pivotal role in host nitrogen supply [72, 73]. In symbiosis, plants
409 regulate oxygen to initiate the FixL and FixJ pathway, ultimately driving the expression of
410 nitrogenase [13], while *nod* genes mediate nodulation [74]. Terrestrial habitats, being
411 more variable than oceans [75], impose fluctuating environmental conditions that
412 necessitate dynamic nitrogenase regulation [76–78]. Molybdenum scarcity in soils also
413 favors alternative nitrogenases [79], likely contributing to gene expansion in terrestrial
414 environments.

415 ***Reconstructing the Eco-Evolutionary History of Nitrogenase-Associated Genes***

416 Central to our framework is that N_2 fixation gene complexity is not an isolated trait, rather,
417 it fits into a larger evolutionary framework that includes genome expansion, metabolic
418 versatility, and habitat diversity (**Figure 5**). As such, gene complexity may serve as an
419 ecological marker—distinguishing generalists, which inhabit diverse environments and
420 possess large, versatile genomes [80–82], from specialist, which typically occupy
421 narrower niches and have smaller genomes [83, 84]. Distinguishing the ecological
422 strategies of diazotrophs may be especially important today, as specialist diazotrophs
423 may contribute disproportionately to total soil N_2 fixation [85].

424 We find that as diazotrophs diversified and specialized over time, so did their N_2 fixation
425 machinery (**Figure 3**). Early evolution of diazotrophs was dominated by gene gain; later
426 stages were marked by loss (**Figure 6**), particularly in terminal branches, pointing to

427 lineage specific streamlining or ecological specialization. We observed this pattern most
428 notably in Chloroflexi, Nitrospirae, Aquificae, Acidithiobacillia, Thermodesulfobacterota,
429 and Epsilonproteobacteria, groups often characterized with either anaerobic,
430 thermophilic, acidophilic, or host-adapted lifestyles (**Supplemental Table 2**). These traits
431 suggest niche-specialization from once generalist ancestors.

432 In Epsilonproteobacteria and Acidithiobacillia, gene loss correlates with reduced habitat
433 breath, consistent with previous findings [86–88]. Whether via adaptation to deep-sea
434 hydrothermal vents [89], shifting toward heterotrophy [87], or acidophilic transitions from
435 a neutrophile ancestor [90], these patterns indicate broader evolutionary trends. In
436 obligate symbioses, gene loss often reflects host compensation [91]. A well-known case
437 is UCYN-A, a marine symbiotic cyanobacterium with a reduced genome [92, 93], which
438 was recently identified as an early stage N_2-fixing organelle ("Nitroplast") in
439 *Braarudosphaera bigelowii* [94, 95]. This and other examples underscore the central
440 importance of gene loss in shaping diazotroph evolution.

441 Examining the distribution of diverse diazotroph genetic machinery offers insight into
442 which genes were critical for enabling transitions into new environments over evolutionary
443 time. The rise of atmospheric oxygen, for example, fundamentally reshaped electron-
444 donor availability [96] and, along with changes in in metal availability [14, 97], likely
445 exerted strong selective pressures on N_2 fixation machinery. Certain genes, such as the
446 G-subunit of alternative nitrogenases [17], appear to have emerged during this transition
447 to an oxic atmosphere, underscoring how environmental transitions, particularly the Great
448 Oxidation Event, shaped molecular evolution in deep time.

449 Reconstructing historical patterns from the modern genetic records is thus pivotal for
450 understanding the processes that govern evolutionary change [98]. Global trends in
451 genome evolution not only reveal how organisms responded to past environmental
452 fluctuations but connects them to their present-day ecological niches. In particular, N_2
453 fixation complexity may help infer a diazotroph's ecological role, highlighting how
454 environmental factors and cellular demands intersect to enable colonization of new
455 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 [99]. 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 and a UW-Madison Research Forward Award to BK. 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.

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 cycle. *Science* 2010; **330**: 192–196.
3. Mus F, Colman DR, Peters JW, Boyd ES. Geobiological feedbacks, oxygen, and the evolution of nitrogenase. *Free Radic Biol Med* 2019; **140**: 250–259.
4. Garcia AK, McShea H, Kolaczkowski B, Kaçar B. Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization.

483 *Geobiology* 2020; **18**: 394–411.

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

487 6. Lyons TW, Tino CJ, Fournier GP, Anderson RE, Leavitt WD, Konhauser KO, et al.
488 Co-evolution of early Earth environments and microbial life. *Nat Rev Microbiol* 2024;
489 **22**: 572–586.

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

493 8. Navarro-González R, McKay CP, Mvondo DN. A possible nitrogen crisis for
494 Archaean life due to reduced nitrogen fixation by lightning. *Nature* 2001; **412**: 61–
495 64.

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

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

500 11. Zehr JP, Capone DG. Changing perspectives in marine nitrogen fixation. *Science*
501 2020; **368**: eaay9514.

502 12. Marcarelli AM, Fulweiler RW, Scott JT. Nitrogen fixation: a poorly understood
503 process along the freshwater-marine continuum. *Limnol Oceanogr Lett* 2022; **7**: 1–
504 10.

505 13. Mus F, Crook MB, Garcia K, Garcia Costas A, Geddes BA, Kouri ED, et al.

506 Symbiotic nitrogen fixation and the challenges to its extension to nonlegumes. *Appl*
507 *Environ Microbiol* 2016; **82**: 3698–3710.
508 14. Rucker HR, Kaçar B. Enigmatic evolution of microbial nitrogen fixation: insights
509 from Earth's past. *Trends Microbiol* 2024; **32**: 554–564.
510 15. Garcia AK, Kolaczkowski B, Kaçar B. Reconstruction of nitrogenase predecessors
511 suggests origin from maturase-like proteins. *Genome Biol Evol* 2022; **14**.
512 16. Harris DF, Rucker HR, Garcia AK, Yang Z-Y, Chang SD, Feinsilber H, et al. Ancient
513 nitrogenases are ATP dependent. *MBio* 2024; **15**: e0127124.
514 17. Cuevas-Zuviría B, Garcia AK, Rivier AJ, Rucker HR, Carruthers BM, Kaçar B.
515 Emergence of an orphan nitrogenase protein following atmospheric oxygenation.
516 *Mol Biol Evol* 2024; **41**.
517 18. Rubio LM, Ludden PW. Biosynthesis of the iron-molybdenum cofactor of
518 nitrogenase. *Annu Rev Microbiol* 2008; **62**: 93–111.
519 19. Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW. Evolution of molybdenum
520 nitrogenase during the transition from anaerobic to aerobic metabolism. *J Bacteriol*
521 2015; **197**: 1690–1699.
522 20. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, et al. Phylogenomics
523 of 10,575 genomes reveals evolutionary proximity between domains Bacteria and
524 Archaea. *Nat Commun* 2019; **10**: 5477.
525 21. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy
526 TBK, et al. Minimum information about a single amplified genome (MISAG) and a
527 metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol*
528 2017; **35**: 725–731.

529 22. Nadkarni PM, Parikh CR. An eUtils toolset and its use for creating a pipeline to link
530 genomics and proteomics analyses to domain-specific biomedical literature. *J Clin*
531 *Bioinforma* 2012; **2**: 9.
532 23. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference
533 resource for gene and protein annotation. *Nucleic Acids Res* 2016; **44**: D457-62.
534 24. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al.
535 KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive
536 score threshold. *Bioinformatics* 2020; **36**: 2251–2252.
537 25. Reimer LC, Sardà Carbasse J, Koblitz J, Ebeling C, Podstawka A, Overmann J.
538 BacDive in 2022: the knowledge base for standardized bacterial and archaeal data.
539 *Nucleic Acids Res* 2022; **50**: D741–D746.
540 26. Barrett T, Clark K, Gevorgyan R, Gorelenkov V, Gribov E, Karsch-Mizrachi I, et al.
541 BioProject and BioSample databases at NCBI: facilitating capture and organization
542 of metadata. *Nucleic Acids Res* 2012; **40**: D57-63.
543 27. Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Sundaramurthi JC, Lee J, et
544 al. Genomes OnLine Database (GOLD) v.8: overview and updates. *Nucleic Acids*
545 *Res* 2021; **49**: D723–D733.
546 28. Madin JS, Nielsen DA, Brbic M, Corkrey R, Danko D, Edwards K, et al. A synthesis
547 of bacterial and archaeal phenotypic trait data. *Sci Data* 2020; **7**: 170.
548 29. Sauer DB, Wang D-N. Predicting the optimal growth temperatures of prokaryotes
549 using only genome derived features. *Bioinformatics* 2019; **35**: 3224–3231.
550 30. Jabłońska J, Tawfik DS. The number and type of oxygen-utilizing enzymes
551 indicates aerobic vs. anaerobic phenotype. *Free Radic Biol Med* 2019; **140**: 84–92.

552 31. Ramoneda J, Stallard-Olivera E, Hoffert M, Winfrey CC, Stadler M, Niño-García JP,
553 et al. Building a genome-based understanding of bacterial pH preferences. *bioRxiv* .
554 2023. , 2023.01.24.524446
555 32. Mise K, Iwasaki W. Environmental atlas of prokaryotes enables powerful and
556 intuitive habitat-based analysis of community structures. *iScience* 2020; **23**:
557 101624.
558 33. Seemann T. barrnap: Bacterial ribosomal RNA predictor.
559 https://github.com/tseemann/barrnap. Accessed 2 Dec 2024.
560 34. Geller-McGrath D, Konwar KM, Edgcomb VP, Pachiadaki M, Roddy JW, Wheeler
561 TJ, et al. Predicting metabolic modules in incomplete bacterial genomes with
562 MetaPathPredict. *Elife* 2024; **13**.
563 35. Kassambara A. ggpubr:'ggplot2'based publication ready plots. *R package version*
564 2018; 2.
565 36. Jombart T, Balloux F, Dray S. Adephylo: New tools for investigating the
566 phylogenetic signal in biological traits. *Bioinformatics* 2010; **26**: 1907–1909.
567 37. R Core Team. R: A Language and Environment for Statistical Computing. Vienna,
568 Austria: R Foundation for Statistical Computing; 2014. R Foundation for Statistical
569 Computing. 2015.
570 38. Yu G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc*
571 *Bioinformatics* 2020; **69**: e96.
572 39. Revell LJ. phytools 2.0: an updated R ecosystem for phylogenetic comparative
573 methods (and other things). *PeerJ* 2024; **12**: e16505.
574 40. Ho L si T, Ané C. A linear-time algorithm for Gaussian and non-Gaussian trait

575 evolution models. *Syst Biol* 2014; **63**: 397–408.
576 41. Csűrös M. Ancestral reconstruction by asymmetric Wagner parsimony over
577 continuous characters and squared parsimony over distributions. *Comparative*
578 *Genomics*. 2008. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 72–86.
579 42. Kapust N, Nelson-Sathi S, Schönfeld B, Hazkani-Covo E, Bryant D, Lockhart PJ, et
580 al. Failure to recover major events of gene flux in real biological data due to method
581 misapplication. *Genome Biol Evol* 2018; **10**: 1198–1209.
582 43. Oyserman BO, Moya F, Lawson CE, Garcia AL, Vogt M, Heffernen M, et al.
583 Ancestral genome reconstruction identifies the evolutionary basis for trait
584 acquisition in polyphosphate accumulating bacteria. *ISME J* 2016; **10**: 2931–2945.
585 44. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and
586 evolutionary analyses in R. *Bioinformatics* 2019; **35**: 526–528.
587 45. Pennell MW, Eastman JM, Slater GJ, Brown JW, Uyeda JC, FitzJohn RG, et al.
588 Geiger v2.0: An expanded suite of methods for fitting macroevolutionary models to
589 phylogenetic trees. *Bioinformatics* 2014; **30**: 2216–2218.
590 46. de Crécy-Lagard V, Amorin de Hegedus R, Arighi C, Babor J, Bateman A, Blaby I,
591 et al. A roadmap for the functional annotation of protein families: a community
592 perspective. *Database (Oxford)* 2022; **2022**: baac062.
593 47. Koirala A, Brözel V. Phylogeny of nitrogenase structural and assembly components
594 reveals new insights into the origin and distribution of nitrogen fixation across
595 bacteria and Archaea. *Microorganisms* 2021; **9**: 1662.
596 48. Pi H-W, Lin J-J, Chen C-A, Wang P-H, Chiang Y-R, Huang C-C, et al. Origin and
597 evolution of nitrogen fixation in prokaryotes. *Mol Biol Evol* 2022; **39**.

598 49. Bellanger M, Figueroa JL 3rd, Tiemann L, Friesen ML, Iii RAW. NFixDB (Nitrogen
599 Fixation DataBase)-a comprehensive integrated database for robust ’omics analysis
600 of diazotrophs. *NAR Genom Bioinform* 2024; **6**: lqae063.
601 50. Nichio BT de L, Chaves RBR, Pedrosa FO, Raittz RT. Exploring diazotrophic
602 diversity: Unveiling nif core distribution and evolutionary patterns in nitrogen-fixing
603 organisms. *Research Square* . 2024.
604 51. Smercina DN, Evans SE, Friesen ML, Tiemann LK. To fix or not to fix: Controls on
605 free-living nitrogen fixation in the rhizosphere. *Appl Environ Microbiol* 2019; **85**.
606 52. Reese BK, Zinke LA, Sobol MS, LaRowe DE, Orcutt BN, Zhang X, et al. Nitrogen
607 cycling of active bacteria within oligotrophic sediment of the mid-Atlantic ridge flank.
608 *Geomicrobiol J* 2018; **35**: 468–483.
609 53. Kapili BJ, Barnett SE, Buckley DH, Dekas AE. Evidence for phylogenetically and
610 catabolically diverse active diazotrophs in deep-sea sediment. *ISME J* 2020; **14**:
611 971–983.
612 54. Semler AC, Paris ER, Salvador M, Dekas AE. Abundance, identity, and potential
613 diazotrophic activity of nifH-containing organisms at marine cold seeps. *Environ
614 Microbiol* 2025; **27**: e70058.
615 55. McRose DL, Zhang X, Kraepiel AML, Morel FMM. Diversity and activity of
616 alternative nitrogenases in sequenced genomes and coastal environments. *Front
617 Microbiol* 2017; **8**: 267.
618 56. Xie J-B, Du Z, Bai L, Tian C, Zhang Y, Xie J-Y, et al. Comparative genomic analysis
619 of N2-fixing and non-N2-fixing Paenibacillus spp.: organization, evolution and
620 expression of the nitrogen fixation genes. *PLoS Genet* 2014; **10**: e1004231.

621 57. Higdon SM, Pozzo T, Kong N, Huang BC, Yang ML, Jeannotte R, et al. Genomic
622 characterization of a diazotrophic microbiota associated with maize aerial root
623 mucilage. *PLoS One* 2020; **15**: e0239677.
624 58. Gallon JR. Tansley review no. 44. Reconciling the incompatible: N_2 fixation and O_2.
625 *New Phytol* 1992; **122**: 571–609.
626 59. Friedman R, Drake JW, Hughes AL. Genome-wide patterns of nucleotide
627 substitution reveal stringent functional constraints on the protein sequences of
628 thermophiles. *Genetics* 2004; **167**: 1507–1512.
629 60. Drake JW. Avoiding dangerous missense: thermophiles display especially low
630 mutation rates. *PLoS Genet* 2009; **5**: e1000520.
631 61. Groussin M, Gouy M. Adaptation to environmental temperature is a major
632 determinant of molecular evolutionary rates in archaea. *Mol Biol Evol* 2011; **28**:
633 2661–2674.
634 62. Miller RW, Eady RR. Molybdenum and vanadium nitrogenases of Azotobacter
635 chroococcum. Low temperature favours N2 reduction by vanadium nitrogenase.
636 *Biochem J* 1988; **256**: 429–432.
637 63. Walmsley J, Kennedy C. Temperature-dependent regulation by molybdenum and
638 vanadium of expression of the structural genes encoding three nitrogenases in
639 Azotobacter vinelandii. *Appl Environ Microbiol* 1991; **57**: 622–624.
640 64. Darnajoux R, Bradley R, Bellenger J-P. In vivo temperature dependency of
641 molybdenum and vanadium nitrogenase activity in the heterocystous Cyanobacteria
642 Anabaena variabilis. *Environ Sci Technol* 2022; **56**: 2760–2769.
643 65. Laso-Pérez R, Wu F, Crémière A, Speth DR, Magyar JS, Zhao K, et al.

644 Evolutionary diversification of methanotrophic ANME-1 archaea and their expansive
645 virome. *Nat Microbiol* 2023; **8**: 231–245.
646 66. Mehta MP, Baross JA. Nitrogen fixation at 92 degrees C by a hydrothermal vent
647 archaeon. *Science* 2006; **314**: 1783–1786.
648 67. Dekas AE, Poretsky RS, Orphan VJ. Deep-sea archaea fix and share nitrogen in
649 methane-consuming microbial consortia. *Science* 2009; **326**: 422–426.
650 68. Ivanovsky RN, Lebedeva NV, Keppen OI, Tourova TP. Nitrogen metabolism of an
651 anoxygenic filamentous phototrophic bacterium oscillocholris trichoides strain DG-6.
652 *Microbiology* 2021; **90**: 428–434.
653 69. Chen Y, Nishihara A, Haruta S. Nitrogen-fixing ability and nitrogen fixation-related
654 genes of thermophilic fermentative bacteria in the genus Caldicellulosiruptor.
655 *Microbes Environ* 2021; **36**.
656 70. Thiel V, Wood JM, Olsen MT, Tank M, Klatt CG, Ward DM, et al. The dark side of
657 the Mushroom Spring microbial mat: Life in the shadow of chlorophototrophs. :I.
658 microbial diversity based on 16S RRNA gene amplicons and metagenomic
659 sequencing. 2016. CRC Press, Boca Raton, FL.
660 71. Sabath N, Ferrada E, Barve A, Wagner A. Growth temperature and genome size in
661 bacteria are negatively correlated, suggesting genomic streamlining during thermal
662 adaptation. *Genome Biol Evol* 2013; **5**: 966–977.
663 72. Franche C, Lindström K, Elmerich C. Nitrogen-fixing bacteria associated with
664 leguminous and non-leguminous plants. *Plant Soil* 2009; **321**: 35–59.
665 73. Bennett AB, Pankievicz VCS, Ané J-M. A model for nitrogen fixation in cereal crops.
666 *Trends Plant Sci* 2020; **25**: 226–235.

667 74. Ferguson BJ, Indrasumunar A, Hayashi S, Lin M-H, Lin Y-H, Reid DE, et al.
668 Molecular analysis of legume nodule development and autoregulation. *J Integr*
669 *Plant Biol* 2010; **52**: 61–76.
670 75. Steele JH, Brink KH, Scott BE. Comparison of marine and terrestrial ecosystems:
671 suggestions of an evolutionary perspective influenced by environmental variation.
672 *ICES J Mar Sci* 2019; **76**: 50–59.
673 76. Dynarski KA, Houlton BZ. Nutrient limitation of terrestrial free-living nitrogen fixation.
674 *New Phytol* 2018; **217**: 1050–1061.
675 77. Zheng M, Zhou Z, Zhao P, Luo Y, Ye Q, Zhang K, et al. Effects of human
676 disturbance activities and environmental change factors on terrestrial nitrogen
677 fixation. *Glob Chang Biol* 2020; **26**: 6203–6217.
678 78. Yan G, Xing Y, Han S, Zhang J, Wang Q, Mu C. Long-time precipitation reduction
679 and nitrogen deposition increase alter soil nitrogen dynamic by influencing soil
680 bacterial communities and functional groups. *Pedosphere* 2020; **30**: 363–377.
681 79. Bellenger JP, Darnajoux R, Zhang X, Kraepiel AML. Biological nitrogen fixation by
682 alternative nitrogenases in terrestrial ecosystems: a review. *Biogeochemistry* 2020;
683 **149**: 53–73.
684 80. Sriswasdi S, Yang C-C, Iwasaki W. Generalist species drive microbial dispersion
685 and evolution. *Nat Commun* 2017; **8**: 1162.
686 81. Chen Y-J, Leung PM, Wood JL, Bay SK, Hugenholtz P, Kessler AJ, et al. Metabolic
687 flexibility allows bacterial habitat generalists to become dominant in a frequently
688 disturbed ecosystem. *ISME J* 2021; **15**: 2986–3004.
689 82. Carbonero F, Oakley BB, Purdy KJ. Metabolic flexibility as a major predictor of

690 spatial distribution in microbial communities. *PLoS One* 2014; **9**: e85105.
691 83. Van Tienderen PH. Evolution of generalists and specialists in spatially
692 heterogeneous environments. *Evolution* 1991; **45**: 1317–1331.
693 84. Poisot T, Bever JD, Nemri A, Thrall PH, Hochberg ME. A conceptual framework for
694 the evolution of ecological specialisation: Evolution of ecological specialisation. *Ecol*
695 *Lett* 2011; **14**: 841–851.
696 85. Hu W, Wang X, Xu Y, Wang X, Guo Z, Pan X, et al. Biological nitrogen fixation and
697 the role of soil diazotroph niche breadth in representative terrestrial ecosystems.
698 *Soil Biol Biochem* 2024; **189**: 109261.
699 86. Kloesges T, Popa O, Martin W, Dagan T. Networks of gene sharing among 329
700 proteobacterial genomes reveal differences in lateral gene transfer frequency at
701 different phylogenetic depths. *Mol Biol Evol* 2011; **28**: 1057–1074.
702 87. Waite DW, Vanwonterghem I, Rinke C, Parks DH, Zhang Y, Takai K, et al.
703 Comparative Genomic Analysis of the Class Epsilonproteobacteria and Proposed
704 Reclassification to Epsilonbacteraeota (phyl. nov.). *Front Microbiol* 2017; **8**: 682.
705 88. Moya-Beltrán A, Beard S, Rojas-Villalobos C, Issotta F, Gallardo Y, Ulloa R, et al.
706 Genomic evolution of the class Acidithiobacillia: deep-branching Proteobacteria
707 living in extreme acidic conditions. *ISME J* 2021; **15**: 3221–3238.
708 89. Campbell BJ, Smith JL, Hanson TE, Klotz MG, Stein LY, Lee CK, et al. Adaptations
709 to submarine hydrothermal environments exemplified by the genome of Nautilia
710 profundicola. *PLoS Genet* 2009; **5**: e1000362.
711 90. González-Rosales C, Vergara E, Dopson M, Valdés JH, Holmes DS. Integrative
712 genomics sheds light on evolutionary forces shaping the Acidithiobacillia class

acidophilic lifestyle. *Front Microbiol* 2021; **12**: 822229.

91. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2011; **10**: 13–26.
92. Tripp HJ, Bench SR, Turk KA, Foster RA, Desany BA, Niazi F, et al. Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature* 2010; **464**: 90–94.
93. Zehr JP, Shilova IN, Farnelid HM, Muñoz-MarínCarmen MDC, Turk-Kubo KA. Unusual marine unicellular symbiosis with the nitrogen-fixing cyanobacterium UCYN-A. *Nat Microbiol* 2016; **2**: 16214.
94. Cornejo-Castillo FM, Inomura K, Zehr JP, Follows MJ. Metabolic trade-offs constrain the cell size ratio in a nitrogen-fixing symbiosis. *Cell* 2024; **187**: 1762-1768.e9.
95. Coale TH, Loconte V, Turk-Kubo KA, Vanslembrouck B, Mak WKE, Cheung S, et al. Nitrogen-fixing organelle in a marine alga. *Science* 2024; **384**: 217–222.
96. Anbar AD, Knoll AH. Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science* 2002; **297**: 1137–1142.
97. Rico KI, Garcia AK, Saito MA, Kaçar B, Anbar AD. Mysteries of metallome evolution: Integrating insights from the Earth and life sciences. In: Anbar A, Weis D (eds). *Treatise on Geochemistry*. 2025. Elsevier, pp 337–364.
98. Kaçar B. Reconstructing early microbial life. *Annu Rev Microbiol* 2024; **78**: 463–492.
99. Falkowski PG. Life’s engines: How microbes made earth habitable. *Life’s Engines: How Microbes Made Earth Habitable* 2015.