ARTICLE
DISCOVERIES

Reconstruction of nitrogenase predecessors suggests origin from maturase-like proteins

Amanda K. Garcia[1], Bryan Kolaczkowski[2], Betül Kaçar[1]

[1]Department of Bacteriology, University of Wisconsin – Madison, USA;
[2]Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida, USA;

Corresponding author: Betül Kaçar
Email: betul@wisc.edu

ABSTRACT

The evolution of biological nitrogen fixation, uniquely catalyzed by nitrogenase enzymes, has been one of the most consequential biogeochemical innovations over life's history. Though understanding the early evolution of nitrogen fixation has been a longstanding goal from molecular, biogeochemical, and planetary perspectives, its origins remain enigmatic. In this study, we reconstructed the evolutionary histories of nitrogenases, as well as homologous maturase proteins that participate in the assembly of the nitrogenase active-site cofactor but are not able to fix nitrogen. We combined phylogenetic and ancestral sequence inference with an analysis of predicted functionally divergent sites between nitrogenases and maturases to infer the nitrogen-fixing capabilities of their shared ancestors. Our results provide phylogenetic constraints to the emergence of nitrogen fixation and are consistent with a model wherein nitrogenases emerged from maturase-like predecessors. Though the precise functional role of such a predecessor protein remains speculative, our results highlight evolutionary contingency as a significant factor shaping the evolution of a biogeochemically essential enzyme.

SIGNIFICANCE STATEMENT

The origin of nitrogenase-catalyzed nitrogen fixation was a transformative event in life's history, garnering long-term study from molecular, biogeochemical, and planetary perspectives. Reconstruction of ancestral nitrogenases suggests that the protein sequence space capable of yielding a nitrogen-fixing enzyme in the past was likely more constrained than previously thought. Specifically, here we show that nitrogenases likely evolved from ancestors that resemble maturases, homologs that today participate in nitrogenase cofactor assembly, contrary to the commonly accepted view that maturases evolved from a nitrogenase ancestor. We further submit that the molecular architecture that may have been required for nitrogenase origins was unlikely to have been shaped by the same environmental drivers often implicated in the evolution of nitrogen fixation. If this decoupling is found to be a recurring pattern in metabolic origins, then the presented results would undercut the common, systems-focused rationale of using ancient environmental conditions to explain the timing of critical and singular biogeochemical innovations in life's past.

KEYWORDS

nitrogenase, nitrogen fixation, maturase, ancestral sequence reconstruction, early life, historical contingency

INTRODUCTION

The modern biosphere is shaped by a variety of essential and ancient enzymes that have co-evolved with the Earth environment for billions of years. Though general mechanisms for the gain of novel enzymatic functions have been explored (Ohno 1970; Gerlt and Babbitt 2001; Copley 2015; Noda-Garcia et al. 2018; Copley 2021), the co-evolutionary steps toward the origins of many specific, key enzymes during Earth's early history remain unresolved. An unavoidable task in addressing this challenge is constraining the ancestral functions of early-evolved enzyme families and their precursors (Benner et al. 2007, Kacar and Garcia 2019).

Biological nitrogen fixation is a notable example of a critical metabolic process with ancient and enigmatic origins. All life requires fixed, or bioavailable, nitrogen. For much of Earth history, this biologically vital element has primarily been acquired by organisms via the activities of nitrogenase metalloenzymes, early evolved and conserved catalysts that uniquely reduce dinitrogen (N_2) to ammonia (NH_3) (Hoffman et al. 2014; Einsle and Rees 2020). The evolution of nitrogenases has constrained the long-term productivity of the biosphere and has itself been shaped by the co-evolving biogeochemistry of Earth's environment (Falkowski 1997; Glass et al. 2009; Stüeken et al. 2016; Luo et al. 2018; Allen et al. 2019; Mus et al. 2019; Garcia et al. 2020). Nitrogenases represent the only known biomolecular solution for the reduction of N_2, a remarkable innovation given that the N≡N bond is one of the most inert in nature. The answer to how biology converged on this solution billions of years ago remains elusive (Boyd and Peters 2013; Mus et al. 2019; Garcia et al. 2020).

Insights into the origins of biological nitrogen fixation can be gained by reconstructing the protein ancestors of nitrogenases and their close homologs. Nitrogenases do not operate alone, but instead exist within a larger protein network required for their assembly and function. The closest homologs to nitrogenases are themselves a key player in this larger assembly network, serving as maturases (also referred to as assembly scaffold proteins) for the final steps in nitrogenase cofactor biosynthesis (Fay et al. 2016; Buren et al. 2020) (fig. 1a). These maturases are

considered necessary in most nitrogenase assembly pathways to modify a complex metal cluster precursor that, when matured, serves as the nitrogenase active site for N_2 reduction. (The only forms of nitrogenases that are confirmed to assemble without maturases are those that only incorporate iron into their active-site cofactors (Perez-Gonzalez et al. 2021)). Though not themselves known to reduce N_2, these maturases reduce a variety of other substrates *in vitro* under highly reducing conditions, including C_2H_2, CO, and CN^-, that also serve as alternative, non-physiological substrates of nitrogenases (Hu et al. 2010; Fay et al. 2016; Seefeldt et al. 2020). These findings establish maturases as catalytically similar to nitrogenases with the exception of their inability to reduce N_2. The divergent protein features between nitrogenases and maturases must therefore account for this difference in N_2-reduction capability.

To investigate the origins of biological nitrogen fixation, we reconstructed the evolutionary history of nitrogenases and maturases. By comparing patterns of sequence conservation between nitrogenases and maturases, we identified divergent residues that might account for their functional differentiation—namely, their ability or inability to reduce N_2, respectively. These sequence features were then leveraged to infer the N_2-reduction capability of reconstructed ancestral proteins and phylogenetically map the origins of biological nitrogen fixation within the nitrogenase family evolutionary history. The relative timing of the evolutionary relationship between nitrogenases and maturases is debated, with some studies suggesting that maturases are evolutionarily derived from nitrogenases (Boyd et al. 2011a). Our findings support an origin of the canonical nitrogenase clade from predecessor proteins that were unlikely to have been capable of N_2 reduction and more closely resemble extant maturases—proteins that are today only ancillary in biological nitrogen fixation. Nitrogenases may therefore represent a case in molecular evolution where a pre-existing but already complex protein architecture, adapted to an alternative role, shaped the origins of one of the most consequential biomolecular innovations in Earth history.

RESULTS AND DISCUSSION

Uncharacterized homologs root canonical nitrogenases within maturase protein clades

We reconstructed the phylogenetic history of nitrogenase and maturase homologs to explore ancestral states for this protein family. There are three forms of nitrogenase—Nif, Vnf, and Anf—

103 that each differ in the composition of their active-site iron-sulfur cluster ("M-cluster"; incorporating
104 molybdenum, vanadium, or additional iron, respectively). The heterotetrameric ($\alpha_2\beta_2$) catalytic
105 protein component of nitrogenase (Nif/Vnf/AnfDK) has its counterpart in a homologous maturase
106 protein complex (Nif/VnfEN; the Anf nitrogenase system does not have dedicated maturase
107 proteins (Perez-Gonzalez et al. 2021)) (fig. 1a). We compiled a comprehensive dataset including
108 nitrogenase Nif/Vnf/AnfDK and maturase NifEN protein sequences (fig. 1b), as well as outgroup
109 dark-operative protochlorophyllide oxidoreductase homologs that share the $\alpha_2\beta_2$ subunit
110 arrangement (BchNB) (Fujita and Bauer 2000; Moser and Brocker 2011). Four maximum
111 likelihood phylogenies were built to test the robustness of tree topology and downstream ancestral
112 sequence inference to sequence sampling and alternate alignment methods (table 1): 1) 2,425
113 nitrogenase, maturase, and outgroup homologs aligned by MAFFT (Katoh and Standley 2013),
114 Tree-1; 2) removal of "uncharacterized" nitrogenase and maturase homologs, Tree-2 (see
115 definition and discussion of "uncharacterized" homologs below); 3) removal of β-subunit
116 nitrogenase, maturase, and outgroup homologs, Tree-3; and 4) alignment with MUSCLE (Edgar
117 2004) instead of MAFFT, Tree-4.
118
119 General shared features across the reconstructed phylogenies include clustering of α-subunit
120 sequences sister to β-subunit sequences (except in Tree-3, which lacks β-subunit sequences),
121 reproducing the $\alpha_2\beta_2$ structural distinction for nitrogenases and maturases (fig. 1c, supplementary
122 fig. S1). These α- and β-subunit clades themselves each segregate into nitrogenase and
123 maturase protein sequences. This topology is consistent with an initial gene duplication event that
124 resulted in separate α- and β-subunits, followed by a secondary duplication event that resulted in
125 functionally distinct nitrogenase and maturase proteins (Fani et al. 2000; Raymond et al. 2004;
126 Boyd et al. 2011a; Boyd and Peters 2013). Within the nitrogenase clade, vanadium- and iron-
127 nitrogenase sequences nest within molybdenum-nitrogenase clades, as has been previously
128 observed (Raymond et al. 2004; Boyd et al. 2011a; Garcia et al. 2020). By contrast, the
129 phylogenetic clustering of maturase sequences associated with different metal-dependent forms
130 of nitrogenases do not reproduce this nesting pattern. Vanadium-maturase sequences are each
131 split into two groups: one forms a small clade with a relatively long branch that diverges prior to
132 all other nitrogenase and maturase sequences, whereas another diverges relatively recently
133 within nitrogenase clades associated with aerobic or facultative bacteria. This topology suggests
134 that maturases for the vanadium nitrogenase system originated independently at least twice, with
135 one origin associated with an early divergence from ancestors of unknown function and another
136 from a recent nitrogenase ancestor.

For the phylogenies that include them (i.e., Tree-1, Tree-3, Tree-4), several maturase-like homologs diverge prior to nitrogenases and root the latter within the maturase clade (fig. 1c, supplementary fig. S1). These homologs belong to “uncharacterized” bacterial and archaeal taxa that lack extensive experimental characterization regarding the metal dependence and N_2-reducing capability of their nitrogenase-like proteins (McGlynn et al. 2012; Garcia et al. 2020). We obtained preliminary functional predictions for these uncharacterized, maturase-like sequences by KEGG’s BlastKOALA (Kanehisa et al. 2016), including a control subset of sequences as well that branch within canonical maturase clades. BlastKOALA returned a mix of maturase and nitrogenase annotations, even for certain control sequences (supplementary table S1). Given this discrepancy and the absence of experimental data, we assign these uncharacterized maturase-like sequences as putative maturases (rather than nitrogenases) based on three lines of evidence. First, the genes that encode these homologs in uncharacterized taxa are located closely downstream of nitrogenase-like genes, as is frequently the case with *bona fide* maturases (fig. 1b). Second, certain uncharacterized taxa (including those that are missing a NifN-like maturase subunit gene) have been shown to fix nitrogen, evidencing a functioning nitrogenase and, by extension, maturase (Mehta and Baross 2006; Dekas et al. 2009; Chen et al. 2021). Third, these homologs conserve a Cys48 residue (numbering here and hereafter from aligned *Azotobacter vinelandii* nitrogenase NifD) present in most maturase proteins (with the exception of certain VnfE homologs) and considered important for binding the cluster precursor prior to maturation (fig. 2a) (Kaiser et al. 2011). At the same time, these sequences all lack the strictly conserved nitrogenase His442 residue that ligates the active-site M-cluster and is critical for N_2 reduction (Kim and Rees 1992; Lee et al. 1998; Li 2002; Jimenez-Vicente et al. 2018). Together, these observations suggest that the uncharacterized maturase-like homologs are unlikely to be functioning as nitrogenases and are more likely operating as canonical maturases.

Our alternate tree reconstructions demonstrate that the rooting of nitrogenases within maturase clades is primarily determined by the inclusion of uncharacterized homologs. Though the phylogenetic position of certain uncharacterized clades is ambiguous (e.g., one clade diverges immediately prior to nitrogenases in Tree-1, but prior to both nitrogenases and other maturases in Tree-3 and Tree-4), we do not observe rooting of maturases within nitrogenase sequences in any of these trees (fig. 1c, supplementary fig. S1). These topological features are also unaffected by trimming the Tree-1 alignment (supplementary fig. S2). This consistency suggests that the observed rooting pattern is robust to the exclusion of the β-subunit protein subtree (Tree-3) and

171 variation in tested alignment methodology (Tree-4). By contrast, the exclusion of uncharacterized
172 homologs from the sequence dataset results in reciprocal monophyly between α-subunit
173 nitrogenase and maturase clades and nesting of β-subunit maturase sequences within
174 nitrogenase sequences (Tree-2; supplementary fig. S1). The α-subunit topology is therefore more
175 stable and less affected by the presence of uncharacterized homologs than the β-subunit
176 topology, which is consistent with the comparatively low branch support values among β-subunit
177 sequences in Tree-2.
178
179 The rooting of nitrogenases within the broader maturase protein clade might parsimoniously
180 suggest that the functionality of the common ancestor of both protein groups more likely
181 resembled that of extant maturases. These phylogenetic observations present a hypothesis that
182 can be tested by evaluating sequence features of reconstructed ancestral proteins inferred to
183 contribute to the functional divergence between nitrogenases and maturases.
184
185
186 **Phylogenetically divergent protein sites between extant nitrogenases and maturases map**
187 **to functionally important structural regions**
188
189 We performed a comprehensive, comparative sequence-structure analysis to identify protein sites
190 that likely contribute to the functional divergence (i.e., (in)ability to reduce N_2) between
191 nitrogenases and maturases. Our goal was to subsequently leverage this analysis for
192 identification of similar sites in reconstructed protein ancestors and phenotypic inference. We
193 limited our analysis to α-subunit nitrogenase (NifD) and maturase (NifE) sequences in part due to
194 the greater topological uncertainty within β-subunit subtrees across our phylogenetic
195 reconstructions (fig. 1c, supplementary fig. S1). In addition, α-subunit sequences host the active-
196 site M-cluster or cluster precursor. We therefore expected that functional differences between
197 nitrogenases and maturases are more likely to be modulated by sequence- and structural-level
198 differences between α-subunit proteins.
199
200 To predict functionally divergent protein sites, we first calculated the amino acid frequency
201 distributions of each alignment column for nitrogenase and maturase sequences (Materials and
202 Methods). The Jensen-Shannon (J-S) distance between the two protein groups for every
203 alignment column was calculated, where larger distances indicate greater divergence for that site
204 between nitrogenases and maturases. We estimated the expected distribution of site-wise J-S

205 distances by randomly partitioning our protein sequences into 2 groups 10,000 times and
206 calculating site-wise J-S distances from each random partition. The p-value for each site's J-S
207 distance from the nitrogenase-maturase partition was calculated from the distribution of J-S
208 distances for that site across random sequence partitions. We defined functionally divergent sites
209 as those exceeding the 75th-percentile distance across all alignment columns, as well as having
210 an FDR-corrected *p*-value <0.0001. This analysis was repeated for all four alignments used to
211 build phylogenetic trees (table 1), yielding 117 (116 for the Tree-4 alignment) predicted sites
212 (supplementary table S2).
213
214 Putative functionally divergent sites identified from the sequence alignments cluster with known
215 functionally important structural regions of the nitrogenase subunit (fig. 2b). These include the M-
216 cluster active site for N_2 reduction, the interface between nitrogenase NifD and NifK subunits, and
217 the M-cluster insertion funnel (Hu et al. 2008). This observed correlation suggests that the set of
218 putative functionally divergent sites is likely enriched for positions contributing to the functional
219 divergence between nitrogenases and maturases. Six sites are of the most divergent across all
220 sequence alignments: 69, 189, 362, 383, 440, and 444 (supplementary table S2). Two additional
221 sites, 442 and 445, are highly divergent in alignments for Tree-1, Tree-2, and Tree-3, but not for
222 Tree-4, likely because Tree-4 was constructed by a different alignment method that would have
223 impacted downstream distance calculations (table 1). All are in or proximal to the nitrogenase
224 active site or M-cluster insertion funnel. Some have specific inferred or experimentally determined
225 functional roles in nitrogenases, including as an M-cluster ligand (site 442) (Kim and Rees 1992),
226 a "lock" to hold the M-cluster within the active site (site 444) (Hu et al. 2008; Solomon et al. 2020),
227 and a "lid" at the cluster insertion funnel opening (site 362) (Hu et al. 2008). Finally, our set of
228 divergent sites includes the maturase cluster precursor ligand, site 48 (Kaiser et al. 2011). The
229 clustering of highly divergent sites with these protein regions highlights the differential interaction
230 of nitrogenases or maturases with the M-cluster. Specifically, maturase function does not require
231 the conservation of residues that permit the insertion and stabilization of the cluster within the
232 active site (Hu et al. 2008), nor the fine-tuning of residues in the active site for catalysis.
233
234 In addition to divergent sites, we identified several conserved residues among nitrogenase and
235 maturase sequences (fig. 2a). His83, Cys88, Cys154, Gly160, Gly185, Gly194, and Cys275
236 residues are conserved in both nitrogenases and maturases. These sites are likely essential to
237 both groups and arose prior to their evolutionary divergence. 23 sites are conserved only in
238 nitrogenases, compared to just three sites that are conserved only in maturases: Cys62, Gly246,

239 and Gly455. However, these three residues are still present in most nitrogenase sequences. The
240 greater number of uniquely conserved residues in nitrogenases relative to maturases may reflect
241 the stronger selective constraint associated with N_2-reduction functionality.
242
243
244 **Putative functionally divergent sites of oldest ancestors resemble extant maturases more**
245 **than nitrogenases**
246
247 With a set of predicted functionally divergent protein sites between extant α-subunit nitrogenase
248 and maturase proteins, we used a probabilistic approach to compare features of ancestral
249 proteins inferred from all reconstructed phylogenies (table 1, supplementary fig. S1). For divergent
250 sites, J-S distances were calculated between ancestral amino acid posterior probability
251 distributions and extant amino acid frequency distributions for either extant nitrogenase or
252 maturase homologs. These distance scores were then normalized to yield a value between -1
253 and +1, here called the "D-score" (supplementary table S3). Positive D-scores indicate greater
254 similarity to α-subunit nitrogenase NifD homologs and negative scores indicate greater similarity
255 to NifE homologs (see Materials and Methods for additional details). D-scores were averaged
256 across divergent sites for each ancestral node.
257
258 Ancestral protein sequences inferred from oldest phylogenetic nodes on average resemble extant
259 maturase homologs more than nitrogenase homologs at predicted functionally divergent sites.
260 Mean D-scores across divergent sites for nodes ancestral to all nitrogenase and maturase
261 homologs (fig. 3, "nitrogenase/maturase last common ancestor 1") range between -0.08 and 0.00.
262 These values are low-magnitude relative to the full range of mean D-scores across all ancestral
263 nodes (~0.2 to +0.2), which is expected due to the mixing of both nitrogenase- and maturase-like
264 sequence features at oldest nodes. Nevertheless, the primarily negative mean D-score values at
265 these nodes indicate greater sequence-level similarity to maturase Nif/VnfE sequences than to
266 nitrogenase Nif/Vnf/AnfD sequences. The ambiguity of the Tree-3 node (D-score = 0.00) may
267 result from the removal of β-subunit sequences for this phylogeny that would otherwise form an
268 outgroup to the α-subunit clade and constrain ancestral sequence composition at this node. At
269 more recent ancestral nodes that exclude early diverged VnfE clades (which have long, less
270 highly supported branches and, thus, ambiguous evolutionary context; "nitrogenase/maturase last
271 common ancestor 2"), mean D-scores range between -0.10 and +0.02. The only tree that yields
272 positive D-scores for these more recent nodes is that reconstructed from the MUSCLE alignment

273 (Tree-4), indicating that this analysis is more sensitive to the tested alignment method than
274 sequence sampling. The sensitivity of ancestral sequence inference to alignment method has
275 been observed previously with simulated data, also finding that MUSCLE produces less accurate
276 inferences than MAFFT (Vialle et al. 2018). Finally, nodes associated with the "nitrogenase last
277 common ancestor" yield positive D-scores for all tree, and therefore resemble nitrogenases more
278 than maturases at predicted functionally divergent sites.
279
280 In addition to surveying mean ancestral similarity to extant nitrogenases or maturases across all
281 putative functionally divergent sites, we also investigated site-wise D-score values, some of which
282 have been experimentally determined to contribute to either nitrogenase or maturase function.
283 These site-wise D-scores were assessed specifically along the phylogenetic transect between the
284 last common ancestor of nitrogenases and the last common ancestor of all nitrogenase and
285 maturases (supplementary table S3). Certain divergent sites that become "nitrogenase-like" early
286 (i.e., prior to the nitrogenase ancestor) include site 195 (important for N_2 substrate binding (Kim
287 et al. 1995)), site 444 (locks the M-cluster in the nitrogenase active site (Hu et al. 2008; Solomon
288 et al. 2020)), and site 359 (helps form the cluster insertion funnel (Hu et al. 2006)). By contrast,
289 site 48 (involved with L-cluster binding in maturases (Kaiser et al. 2011)), as well as sites 361 and
290 362 (involved with M-cluster insertion at the nitrogenase active site (Hu et al. 2008)), remain
291 primarily "maturase-like" until the nitrogenase last common ancestor.
292
293 The site 442 histidine M-cluster ligand conserved across all extant canonical nitrogenase proteins
294 is considered critical for nitrogenase activity (Kim and Rees 1992; Lee et al. 1998; Li 2002;
295 Jimenez-Vicente et al. 2018) and, notably, remains maturase-like prior to the nitrogenase last
296 common ancestor (mean D-score ≈ -0.40 across alternate alignments, compared to a minimum
297 D-score of -0.76 for all divergent sites; supplementary table S3). Site 442 is only not predicted to
298 be functionally divergent for the tree reconstructed by a MUSCLE alignment (Tree-4). This
299 difference is likely due to an inferred homology, unique to the MUSCLE alignment, between the
300 nitrogenase His442 residue and a frequently observed histidine residue in maturases. However,
301 the other alignments, which instead infer homology with a neighboring arginine residue in many
302 maturases (fig. 2a), is supported by studies which indicate that this arginine residue is structurally
303 aligned to nitrogenase His442 (Kaiser et al. 2011). Thus, it is likely that the MUSCLE alignment
304 is erroneous at this site, consistent with the observed reduced accuracy of MUSCLE compared
305 to MAFFT for ancestral sequence inference (Vialle et al. 2018). Due to its functional significance

306 and conservation among extant nitrogenases, the appearance of a histidine residue at site 442
307 may have been critical for the origins of N_2 reduction.
308
309
310 **Nitrogenases likely originated from a non-N_2-reducing maturase-like protein with possible**
311 **biosynthetic or alternate catalytic roles**
312
313 Our exploration of nitrogenase and maturase ancestry, coupled with the investigation of global
314 sequence and structural features that account for their functional divergence, is generally
315 consistent with the hypothesis that the last common ancestor of nitrogenase and maturase
316 proteins was unlikely to have functioned as a nitrogenase. This inference is also supported by the
317 absence of residue-level similarity at divergent sites that have empirically been shown to be
318 critical for nitrogenase function. These results are robust to phylogenetic uncertainty stemming
319 fromsequence sampling of early diverged uncharacterized lineages and incorporates statistical
320 uncertainty associated with ancestral sequence inference. Though we find that an alternative
321 alignment method, MUSCLE, does modulate these sequence-based functional inferences at one
322 ancestral node (fig. 3), there is reason to suspect reduced alignment accuracy given probable
323 misalignment of at least one key nitrogenase residue and decreased performance in simulated
324 data (Vialle et al. 2018) (see above).
325
326 Conservatively, our results suggest that phylogenetic inference of ancient nitrogen fixation can
327 reliably extend only to within the canonical nitrogenase clade, and that inferences of earlier
328 nitrogen fixation activity require further evidence. More definitive assessment of the N_2-reduction
329 capability of the common ancestor of nitrogenases and maturases awaits experimental
330 investigation, which can be directly achieved through the laboratory resurrection of ancestral
331 proteins inferred in this study (Thornton 2004; Benner 2017; Garcia and Kacar 2019).
332 Nevertheless, a model that posits a maturase-like ancestry for nitrogenases deviates from existing
333 hypotheses regarding their early evolution (fig. 4). Previous models are based on parsimonious
334 interpretations of nitrogenase and maturase phylogenetic topology that is not observed in the
335 trees reconstructed here. For instance, previous studies root the maturase clade within
336 nitrogenase sequences, suggesting that the former evolved via gene duplication of nitrogenase
337 ancestors (Boyd et al. 2011a; Boyd and Peters 2013). An updated phylogenetic analysis
338 incorporating the Boyd et al. (2011a) dataset, lacking uncharacterized sequences, produces a
339 topology similar to Tree-2 where nitrogenase and maturase α-subunit clades are reciprocally

340 monophyletic (supplementary fig. S2). However, with the inclusion of uncharacterized nitrogenase
341 and maturase homologs, maturases instead root nitrogenases, and support a model wherein
342 nitrogen fixation is instead a derived feature of a maturase-like ancestor. Our reconstruction of
343 ancestral states within the nitrogenase and maturase phylogeny provides additional constraints
344 on ancestral phenotypes within a maximum likelihood framework, extending beyond inferences
345 drawn from phylogenetic topology alone. These results constrain the likely origin of nitrogen
346 fixation to a relatively well-resolved lineage within the nitrogenase/maturase topology, rather than
347 to a deeper history that bridges nitrogenases with more distantly related homologs (e.g.,
348 coenzyme F430 biosynthesis proteins (Mus et al. 2019)).
349
350 Our results suggest that ancestral maturase-like proteins may have provided the molecular
351 architecture for the origin of nitrogen fixation. However, the precise functional role of such an
352 ancestor is not clear. Candidate phenotypic attributes, shared between extant nitrogenases and
353 maturases, may shed light on how biological N_2 reduction evolved. For example, both are the only
354 proteins known to bind the nitrogenase active-site M-cluster (Fay et al. 2016). In addition, both
355 extant nitrogenases and maturases can reduce several non-physiological, alternative substrates
356 including C_2H_2, CO, and CN^-, albeit in highly reducing experimental conditions (Fay et al. 2016;
357 Seefeldt et al. 2020). It has previously been argued that the ability of nitrogenases to reduce
358 alternative substrates may simply be the byproduct of overcoming the significant activation barrier
359 required for N_2 reduction (Boyd and Peters 2013). However, combined with the evidence
360 presented here for a maturase-like ancestry, a plausible scenario is an ancestral protein capable
361 of reducing the shared substrates of extant nitrogenases and maturases at an M-cluster-like
362 active site (perhaps in a role as detoxyases, as has previously been proposed ((Silver and
363 Postgate 1973); fig. 4). This scenario would provide a stepwise path for the evolution of
364 nitrogenases from ancestral proteins capable of catalyzing less ATP-intensive reactions (Hu et
365 al. 2010), and requires only residue-level tuning of an already complex peptide environment to
366 achieve the earliest whiffs of N_2 reduction. Our reconstruction of ancestral residues at predicted
367 functionally divergent sites provides possible mutational trajectories toward the evolution of
368 nitrogenases.
369
370 The origin of nitrogenases from proteins involved with cofactor biosynthesis such as the
371 maturases studied here might be expected given the prevalence of biosynthetic proteins
372 associated with the broader family of nitrogenase-like homologs. These include chlorophyll
373 biosynthesis proteins used as phylogenetic outgroups in our study (Fujita and Bauer 2000; Moser

374 and Brocker 2011), as well as coenzyme F430 biosynthesis proteins that are conserved among
375 methanogens (Staples et al. 2007; Zheng et al. 2016). In addition, more distantly related and
376 poorly studied nitrogenase-like homologs may have roles in assembly of hydrogenase
377 metalloclusters and metal transport (reviewed in Ghebreamlak and Mansoorabadi (2020)). It is
378 not clear whether a hypothetical, maturase-like nitrogenase ancestor may have had a central
379 function in cofactor biosynthesis as its descendants and several related homologs do today. Other
380 distantly related nitrogenase-like homologs have putative catalytic roles, including a recently
381 reported homolog suggested to participate in a methionine salvage pathway that forms ethylene
382 (North et al. 2020). It is possible that a maturase-like predecessor would have been promiscuous,
383 a suggested general feature of early-evolved proteins (Copley 2015; De Tarafder et al. 2021),
384 and capable of both providing a scaffold for cluster maturation as well as catalysis using the same
385 matured cluster. Gene duplication and divergence might then have subsequently specialized
386 maturases to only function as a scaffold and evolve residues that permit the release of the
387 matured cluster. In parallel, nitrogenases would have specialized to function in a solely catalytic
388 role.
389
390 The possible capability of an ancestral maturase-like protein to bind an M-cluster-like cofactor
391 would have important implications for the coevolutionary trajectory of nitrogenases and the
392 biogeochemical environment (Anbar and Knoll 2002; Glass et al. 2009; Boyd et al. 2011b). A
393 defining role of extant maturases in the molybdenum-dependent nitrogenase system is to provide
394 a scaffold for the incorporation of molybdenum into the active-site M-cluster. Thus, the origin of
395 maturase proteins has previously been suggested to coincide with the origin of molybdenum
396 dependence in nitrogenases (Boyd et al. 2011a). However, if maturase-like proteins predate
397 nitrogenases, the molybdenum-containing M-cluster itself may predate nitrogenases as well (fig.
398 4). Inferences for the age of nitrogenases extend to more than 3 billion years ago (Stueken et al.
399 2015; Parsons et al. 2020) when molybdenum in Earth's oceans was likely exceedingly scarce
400 (Anbar 2008). A molybdenum-incorporating maturase-like protein existing prior to 3 billion years
401 ago would suggest that the bulk geochemistry of the early Earth environment may not necessarily
402 have provided strict constraints on enzyme evolution (Garcia et al. 2020), and would be in
403 agreement with geochemical evidence of early molybdenum-dependent nitrogenase activity
404 (Stueken et al. 2015). Rather, localized environments may have provided sufficient molybdenum
405 for the function of nitrogenases and their predecessors (Stueken et al. 2015), or molybdenum
406 may have simply been selected despite its scarcity due to its invaluable chemical features.

407 Another possibility is that earliest maturases-like ancestors did not incorporate molybdenum, but
408 rather the iron-only cluster precursor that is matured to the M-cluster today (Mus et al. 2019).
409
410
411 **The role of contingency and subsumed complexity in nitrogenase evolution**
412
413 The alternative model of nitrogenase origins that we propose exemplifies a case for molecular
414 evolution in which a novel and consequential metabolic function was built off a pre-existing,
415 complex molecular architecture. An open question is whether the pre-existing complexity and
416 functional role of a maturase-like protein was needed to evolve an enzyme capable of reducing
417 N_2. The role of evolutionary contingency in shaping biological diversity has long been examined
418 (Gould 1989; Vermeij 2006; Blount et al. 2018), particularly to envision future evolutionary
419 scenarios or alternate trajectories on other worlds characterized by distinct environmental
420 parameters (Kaçar and Gaucher 2012). Regarding the evolution of biological nitrogen fixation in
421 particular, it has been argued that necessity—i.e., the need for bioavailable sources of nitrogen—
422 or environmental geochemistry likely controlled the timing of early nitrogenase evolution and
423 diversification on Earth (Navarro-Gonzalez et al. 2001; Anbar and Knoll 2002; Mus et al. 2018).
424 However, another possibility is that the origin of biological nitrogen fixation required the subsumed
425 complexity of a protein predecessor (Adam et al. 2018), which was initially positively selected for
426 an entirely different metabolic role. The origin of nitrogenases thus may not have occurred (or
427 may have been significantly delayed) without a suitable protein on which to build, despite the
428 scarcity of bioavailable nitrogen, an abundance of possible metal cofactors, or both. Testing these
429 possibilities would require experimentally replaying the evolutionary path that led to the origin of
430 biological nitrogen fixation.
431
432 There is no evidence that Terran biology evolved nitrogen fixation more than once. Whether this
433 is the product of very exceptional circumstances of origination or of a survivorship bias so
434 pronounced that there is only one remaining functional example supported across all of biology,
435 nitrogenases are therefore on par with other singular molecular-level innovations such as the
436 ribosome (Fox 2010) and oxygenic photosynthesis (Blankenship and Hartman 1998). The existing
437 variation among different metal-dependent forms of nitrogenase enzymes do not constitute truly
438 independent evolutionary experiments, but variations on a theme that was determined and
439 uniquely constrained by the common ancestral form. Even within this narrow range of constraints,

440 the degree to which contingent amino-acid substitutions shaped the diversification of nitrogenase
441 metal preference and specificity remains unknown.
442
443 Identifying the singular circumstances that left Earth with a single common ancestor for all
444 nitrogenase function may be critical for understanding the pervasiveness of nitrogen fixation as a
445 universal biological capability. This is particularly important for assessing whether metal
446 availability significantly guides protein evolution or whether other internal biophysical constraints
447 lead to background-dependent, epistatic interactions (Williams and da Silva 1996; Anbar and
448 Knoll 2002; Moore et al. 2017; Smethurst and Shcherbik 2021). Efforts to generate artificial
449 nitrogenases and nitrogenase metalloclusters (Tanifuji et al. 2015; Sickerman et al. 2017) may
450 expand the suite of molecular structures capable of reducing N_2, but biotic experiments integrating
451 gene regulatory and protein-protein interaction constraints are needed to test different
452 macroevolutionary hypotheses of nitrogenase emergence. A survey of such functional constraints
453 on nitrogenase and maturase predecessors could reveal the sequence of biomolecular functions
454 conducive for the evolution of nitrogen fixation, which could then be integrated into a more
455 comprehensive accounting of internal selective forces, geochemical features and planetary
456 environments that can host similar evolutionary pathways (Kacar et al. 2020).
457
458 Perhaps most intriguingly, our results suggest that nitrogen fixation may have emerged from
459 natural selection acting on a maturase-like protein whose ancestral function was largely
460 decoupled from extracellular conditions frequently implicated as drivers for the origins of nitrogen
461 fixation. This scenario would thus cast the origin of nitrogen fixation as an act of extreme
462 contingency bordering on happenstance, betraying its utility as one of the most evolutionarily
463 significant and biologically limiting metabolic pathways on Earth. If borne out by further study or
464 found to be a recurring pattern for other critical molecular innovations, enzyme origins largely
465 decoupled from putative environmental drivers may severely compromise the soundness of
466 systems-focused hypotheses that tie organismal or ecological need to bulk geochemical substrate
467 or cofactor availabilities. The paleobiology of molecular innovations would require disciplinary
468 approaches and conceptual foundations quite distinct from the study of their more recent
469 evolution.
470
471
472 **MATERIALS AND METHODS**
473

474 **Phylogenetic reconstruction and ancestral sequence inference**

475

476 Protein homologs were identified from the National Center for Biotechnology Information non-
477 redundant protein database by BLASTp (Camacho et al. 2009) with an expect value cutoff of <1e-
478 5 (accessed January 2020). Query sequences from *Azotobacter vinelandii* (NifD: WP_01270336,
479 NifK: WP_012698833; NifE: WP_012698838, NifN: WP_012698839) were used for nitrogenase
480 and maturase homolog identification, and sequences from *Synechococcus elongatus* (BchN:
481 WP_126148028, BchB: WP_126147769) for outgroup dark-operative protochlorophyllide
482 oxidoreductase homolog identification. Sequences from this relatively permissive BLASTp search
483 were aligned using MAFFT v7.450 (Katoh and Standley 2013) to build a preliminary phylogeny
484 with FastTree v2.1.11 (Price et al. 2010). Putative nitrogenase homologs were identified based
485 on previously published phylogenies (Boyd et al. 2011a; Garcia et al. 2020) as well as sequence
486 features known to be critical for N_2 reduction (e.g., Cys275, His442). Putative maturase homologs
487 were only retained if the encoding genes were co-localized with nitrogenase genes in the same
488 genome. Finally, sequences in overrepresented clades were pruned to obtain a roughly equal
489 number of nitrogenase and maturase versus outgroup sequences, so as not to bias subsequent
490 ancestral sequence inference.

491

492 A final untrimmed MAFFT alignment was used as input for phylogenetic reconstruction by RAxML
493 v8.2.11 (Stamatakis 2014) using 100 rapid bootstrap searches and the best-fit LG+G+F
494 evolutionary model determined by ModelFinder (Kalyaanamoorthy et al. 2017) (an additional
495 phylogeny was also built from an alignment trimmed by TrimAl (Capella-Gutierrez et al. 2009)
496 (supplementary fig. S2)). The tree was further optimized by applying nearest-neighbor-
497 interchanges before calculation of SH-like branch support values (Guindon et al. 2010), resulting
498 in the final phylogeny, Tree-1. Additional trees incorporated in ancestral sequenc inference were
499 generated by altering sequence sampling or alignment with MUSCLE v3.8.425 (Edgar 2004)
500 instead of MAFFT (see Results and Discussion, table 1). Finally, the Bayesian phylogenetic
501 analysis by Boyd et al. (Boyd et al. 2011a) was replicated using their reported methods
502 (supplementary fig. S2).

503

504 Ancestral sequences were inferred by maximum likelihood marginal reconstruction in PAML v4.9j
505 (Yang 2007) using the same evolutionary model parameters described above for RAxML.
506 Sequence gaps were reconstructed in PAML using the binary encoding approach described in
507 Aadland et al. (2019). Briefly, the protein sequence alignment was recoded as a 'presence-

508 absence' alignment matrix, and the posterior-probability of the presence (amino-acid residue) or
509 absence (gap) state at each position in each ancestral sequence was calculated using maximum-
510 likelihood reconstruction, assuming a binary character model with state frequencies inferred by
511 maximum likelihood. All phylogenetic data, including sequence alignments, trees, and ancestral
512 sequence inference outputs can be found at https://github.com/kacarlab/maturase2021.
513
514
515 **Prediction of functionally divergent protein sites between nitrogenases and maturases**
516
517 For each position in the sequence alignment, we calculated the Jensen-Shannon (J-S) distance
518 between the amino-acid frequency distribution estimated from extant nitrogenase Nif/Vnf/AnfD
519 sequences in the alignment and that estimated from extant maturase Nif/VnfE sequences, with
520 nitrogenase and maturase sequences being defined based on monophyly in the tree topology
521 and gene location within the *nif*, *vnf*, or *anf* loci (see Results and Discussion, fig. 1b). Briefly, the
522 J-S distance is calculated as the average Kullback-Leibler divergence, or "relative entropy," which
523 estimates the loss of information when one frequency distribution is used to represent another.
524 Intuitively, the site-wise J-S distance between nitrogenase and maturase amino-acid frequency
525 distributions describes how dissimilar the distribution of extant amino-acids is between
526 nitrogenase and maturase sequences for each alignment site.
527
528 We estimated the expected distribution of site-wise J-S distances, given our sequence alignment,
529 by randomly partitioning the alignment into two sequence groups of sizes equivalent to the sizes
530 of our actual nitrogenase and maturase groups, respectively, and calculating site-wise J-S
531 distances between these randomly partitioned groups. We performed 10,000 random partitions
532 and site-wise J-S distance calculations. For each alignment column *i*, we calculated the probability
533 of observing J-S(nitrogenase, maturase)$_i$, given the distribution of J-S-distances at column *i* in the
534 randomly partitioned dataset (i.e., *p*-value). We enriched for highly divergent positions using an
535 FDR-corrected *p*-value cutoff of 0.0001. We additionally excluded any sites with J-S distances in
536 the lower 75th percentile of the J-S distance distribution across all alignment positions.
537
538
539 **Probabilistic assessment of nitrogenase-like ancestral sequence features**
540

Nitrogenase-like ancestral sequence features were assessed by incorporating the statistical uncertainty of ancestral sequence inference into comparisons between those of extant nitrogenases and maturases. For putative functionally divergent protein sites identified as described above, J-S distances were calculated between the ancestral amino acid posterior probability distributions and the extant amino acid frequency distributions across either extant nitrogenase (Nif/Vnf/AnfD) or maturase (Nif/VnfE) clades. These distance values were then normalized to yield a value between -1 and +1 indicating the relative similarity of an ancestral protein site to a homologous extant nitrogenase site, here called the "D-score" (i.e., similarity to the nitrogenase D-subunit):

$$\text{D-score} = (\text{dist}_E - \text{dist}_D) / (\text{dist}_E + \text{dist}_D)$$

where dist_D is the J-S distance between ancestral and maturase E-subunit sites, and dist_E is the J-S distance between ancestral and nitrogenase D-subunit sites. D-scores were analyzed on a site-wise basis as well as averaged across the length of each ancestral sequence for all constructed phylogenies. All data and scripts related to the prediction of functionally divergent sites prediction and D-score calculations can be found at https://github.com/kacarlab/maturase2021.

ACKNOWLEDGEMENTS

This work was supported by the National Aeronautics and Space Administration Postdoctoral Program to A.K.G., the National Science Foundation Division of Molecular and Cellular Biosciences (1817942 to B. Kolaczkowski), the National Science Foundation Emerging Frontiers Program (1724090 to B. Kaçar), the National Aeronautics and Space Administration Early Career Faculty Award (80NSSC19K1617 to B. Kaçar), and the Metal Utilization and Selection across Eons (MUSE) Interdisciplinary Consortium for Astrobiology Research, sponsored by the National Aeronautics and Space Administration Science Mission Directorate (19-ICAR19_2-0007 to B. Kaçar). We thank Eric Boyd and Jennifer Glass for the helpful discussions.

FIGURES AND FIGURE LEGENDS

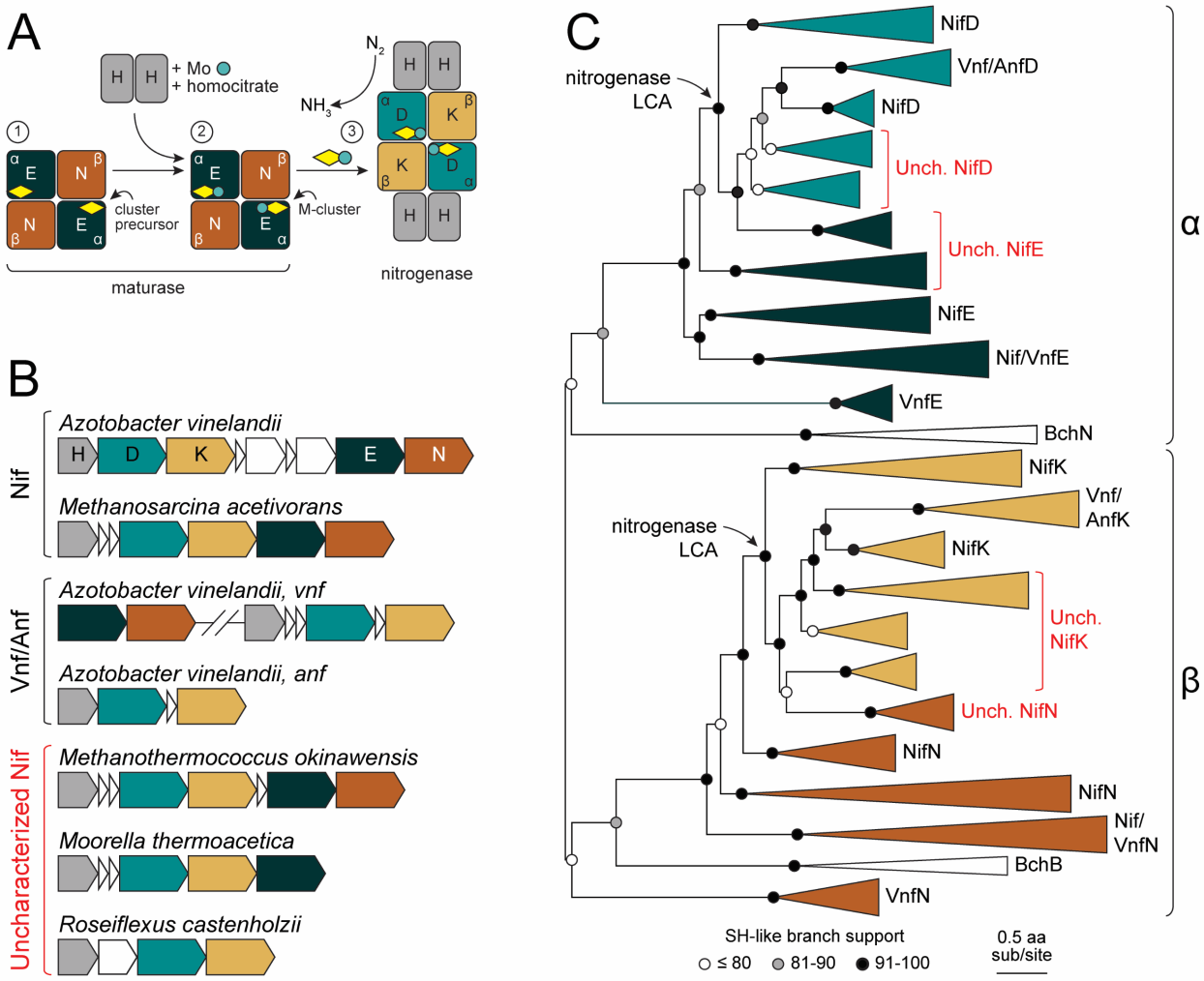
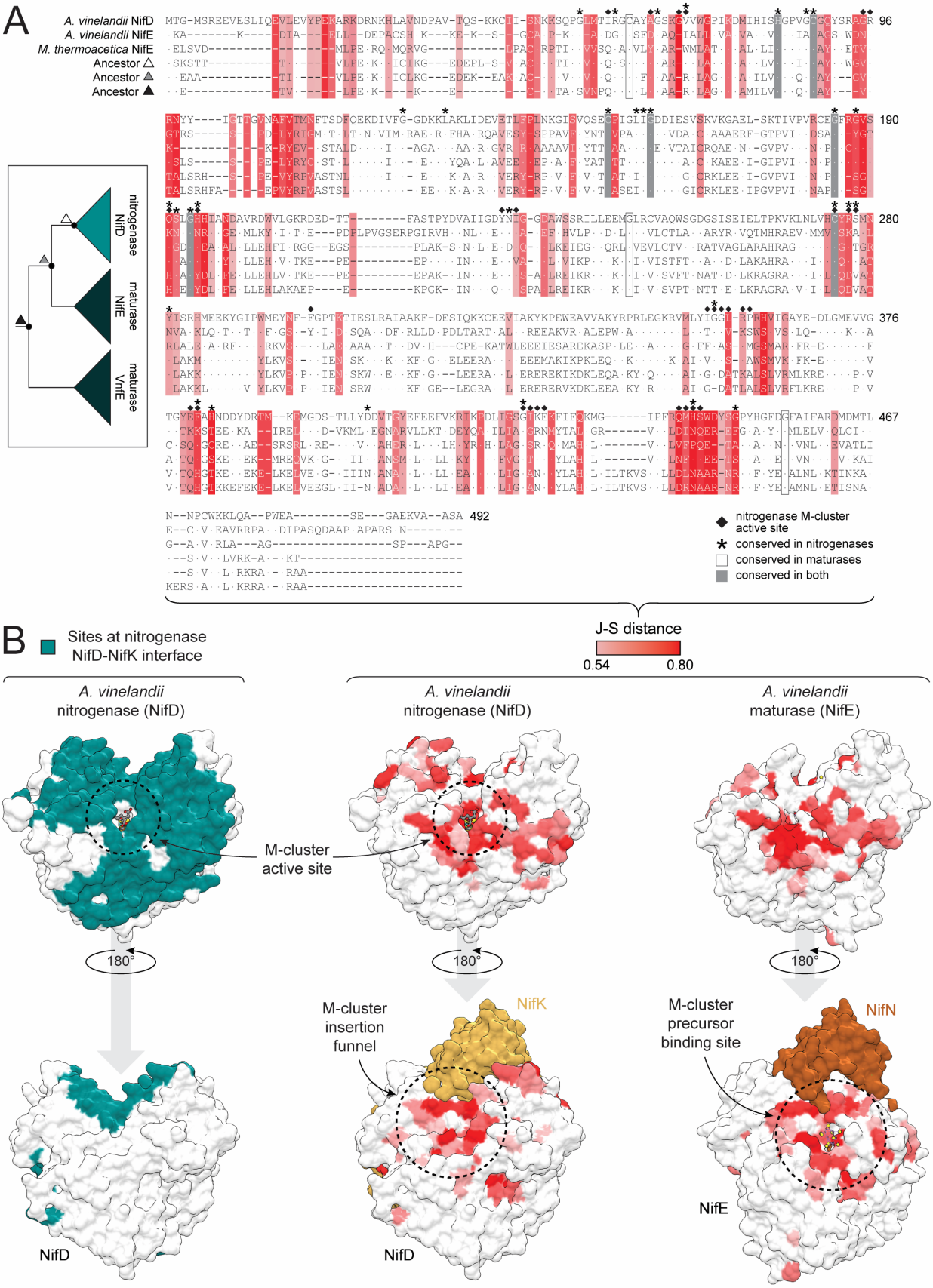


Fig. 1. Nitrogenase and maturase functionality, genetic organization, and evolutionary history. (*A*) Simplified biosynthetic pathway for nitrogenase active-site cluster maturation and incorporation (molybdenum-dependent nitrogenase system shown). Pathway steps are indicated by circled numbers. Step 1). Maturase NifEN proteins (NifE, dark green; NifN, brown), which are $\alpha_2\beta_2$ heterotetramer homologs to nitrogenase NifDK proteins (NifD, teal; NifK, yellow), are scaffolds for maturation of nitrogenase iron-sulfur cluster precursors (yellow diamond). Step 2) Cluster precursors are matured to M-clusters (yellow diamond with teal circle) by incorporation of molybdenum and homocitrate, delivered to the maturase complex by NifH proteins. Step 3) Mature M-clusters are incorporated into the nitrogenase complex, where they serve as the active sites for N_2 reduction to NH_3. During enzyme turnover, NifH proteins transiently interface with the nitrogenase NifDK complex to deliver electrons to the M-cluster active site. (*B*) Representative gene locus structures for molybdenum-dependent (Nif), vanadium-dependent (Vnf), and iron-dependent (Anf) nitrogenase systems, as well as for uncharacterized nitrogenase homologs (see

590 text for discussion). Gene and intergenic region lengths are approximate. Hash marks indicate
591 significant distance between represented genes. (*C*) Maximum likelihood phylogenetic tree (Tree-
592 1; see text for details) built from nitrogenase and maturase protein sequences. α- or β-subunit
593 designations for protein sequences are indicated on the right. Homologs from uncharacterized
594 taxa are highlighted in red. Clade widths are not to scale. “LCA”: Last common ancestor.
595


A
A. vinelandii NifD
A. vinelandii NifE
M. thermoacetica NifE
Ancestor
Ancestor
Ancestor
nitrogenase NifD
maturase NifE
maturase VnfE
nitrogenase M-cluster active site
conserved in nitrogenases
conserved in maturases
conserved in both
J-S distance
0.54
0.80
B
Sites at nitrogenase NifD-NifK interface
A. vinelandii nitrogenase (NifD)
A. vinelandii nitrogenase (NifD)
A. vinelandii maturase (NifE)
M-cluster active site
180°
180°
180°
M-cluster insertion funnel
NifK
M-cluster precursor binding site
NifN
NifD
NifD
NifE

596

597

598 **Fig. 2.** Structure and sequence maps of putative functionally divergent protein sites between
599 nitrogenases and maturases. (*A*-*B*) Putative functionally divergent sites are defined as those
600 above the 75th percentile J-S distance between nitrogenase NifD and maturase NifE proteins (see
601 text for details). The J-S distance scale applies to both (*A*) and (*B*). (*A*) Maximum likelihood
602 ancestors (inferred from Tree-1; fig. 1c) aligned to representative extant nitrogenase (NifD) and
603 maturase (NifE) sequences. *Moorella thermoacetica (M. thermoacetica*) NifE is an
604 uncharacterized, putative maturase sequence. Ancestor triangle symbols correspond to labeled
605 nodes in the simplified phylogeny (*left*) and match those in fig. 3. Nitrogenase M-cluster active-
606 site residues are defined as those within 5 Å of the M-cluster. Dots within the alignment indicate
607 residue identity to *Azotobacter vinelandii* (*A. vinelandii*) NifD. Site numbering based on *A.*
608 *vinelandii* NifD. (*B*) Divergent sites mapped to aligned nitrogenase (*center*; *A. vinelandii* NifD, PDB
609 3U7Q) and maturase (*right*; *A. vinelandii* NifE, PDB 3PDI) subunit structures. Protein sites at the
610 nitrogenase NifD-NifK interface (*left*) are defined as those within 5 Å of the NifK subunit.
611

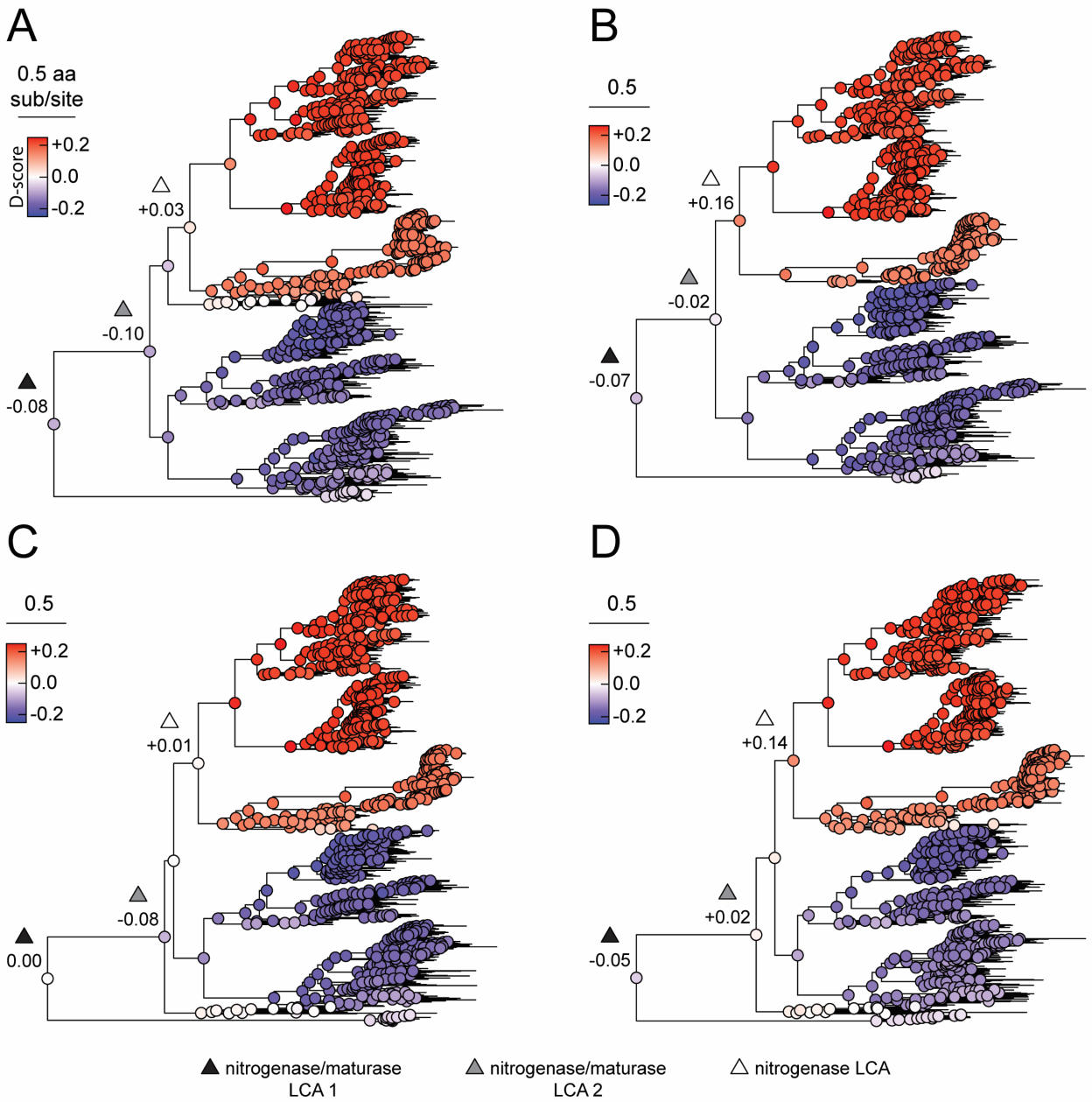
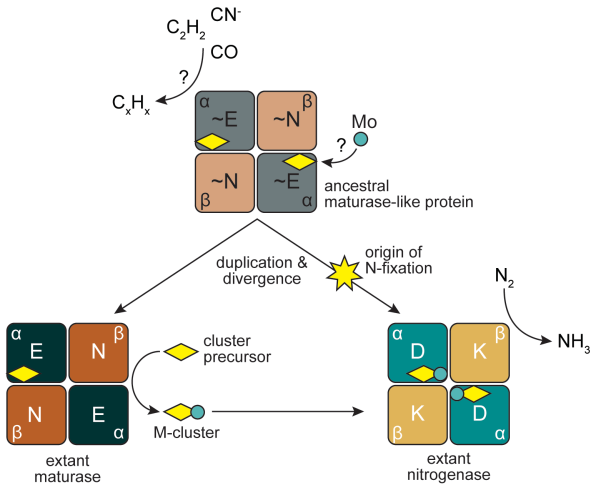


Fig. 3. Ancestral sequence similarity to extant nitrogenases or maturases, mapped across four alternate phylogenies: (*A*) Tree-1, (*B*) Tree-2, (*C*) Tree-3, and (*D*) Tree-4. Similarity is expressed as the "D-score" parameter, where a positive D-score (red) indicates greater similarity to extant nitrogenases, and a negative D-score (blue) indicates greater similarity to extant maturases (see Materials and Methods). D-scores were averaged across putative functionally divergent sites for each ancestral node. Nodes are labeled nitrogenase/maturase last common ancestor ("LCA") 1—including early diverged VnfE homologs (black triangle; see text for discussion); nitrogenase

621 /maturase LCA 2—excluding early diverged VnfE homologs (grey triangle); and nitrogenase LCA
622 (white triangle), along with their mean D-scores.
623



624
625

626 **Fig. 4.** Proposed model for the origins and functional divergence of maturase and nitrogenase
627 proteins. An ancestral maturase-like protein (~NifEN, grey and light brown), incapable of reducing
628 N_2, may have otherwise reduced various carbon-containing substrates and/or played a role in
629 cluster (yellow diamond) biosynthesis. The ancestor may have been capable of incorporating
630 molybdenum (teal circle) into the cluster. Duplication of the encoding ancestral genes and
631 functional divergence would then have yielded canonical maturase (NifEN, dark green and brown)
632 and nitrogenase (NifDK, teal and yellow) proteins. Maturases would have specialized to provide
633 a scaffold for the maturation of the nitrogenase cluster precursor (yellow diamond) to the
634 nitrogenase active-site M-cluster (yellow diamond with teal circle). In parallel, tuning of the
635 ancestral peptide environment along a divergent lineage would have spurred the origin of N_2
636 reduction and specialization of nitrogenases for a solely catalytic role in nitrogen fixation. Protein
637 components of the $\alpha_2\beta_2$ heterotetrameric nitrogenase and maturase structures are labeled.
638
639
640 **TABLES**
641
642 **Table 1.** Nitrogenase and maturase phylogenies built in this study.
643

Tree	Alignment method	Sequence Dataset

Tree-1	MAFFT	Nif/Vnf/AnfDK, Nif/VnfEN, BchNB
Tree-2	MAFFT	Tree-1 dataset without uncharacterized homologs
Tree-3	MAFFT	Tree-1 dataset without β-subunit sequences (Nif/Vnf/AnfK, Nif/VnfN, BchB)
Tree-4	MUSCLE	Same as Tree-1 dataset

644

645

646 **DATA AVAILABILITY STATEMENT**

647

648 All phylogenetic datasets and scripts are available at https://github.com/kacarlab/maturase2021.

649

650

651 **REFERENCES**

652

653 Aadland K, Pugh C, Kolaczkowski B. 2019. High-Throughput Reconstruction of Ancestral
654 Protein Sequence, Structure, and Molecular Function. *Methods Mol Biol*. 1851:135-170. doi:
655 10.1007/978-1-4939-8736-8_8.

656 Adam ZR, Fahrenbach AC, Kacar B, Aono M. 2018. Prebiotic Geochemical Automata at the
657 Intersection of Radiolytic Chemistry, Physical Complexity, and Systems Biology. *Complexity*.
658 2018:1-21. doi: 10.1155/2018/9376183.

659 Allen JF, Thake B, Martin WF. 2019. Nitrogenase Inhibition Limited Oxygenation of Earth's
660 Proterozoic Atmosphere. *Trends Plant Sci*. 24:1022-1031. doi: 10.1016/j.tplants.2019.07.007.

661 Anbar AD. 2008. Oceans. Elements and evolution. *Science*. 322:1481-1483. doi:
662 10.1126/science.1163100.

663 Anbar AD, Knoll AH. 2002. Proterozoic ocean chemistry and evolution: a bioinorganic bridge?
664 *Science*. 297:1137-1142. doi: 10.1126/science.1069651.

665 Benner S. 2017. Uniting natural history with the molecular sciences. The ultimate
666 multidisciplinarity. *Acc Chem Res*. 50:498-502. doi: 10.1021/acs.accounts.6b00496.

667 Benner SA, Sassi SO, Gaucher EA. 2007. Molecular paleoscience: systems biology from the
668 past. *Adv Enzymol Relat Areas Mol Biol*. 75:1-132, xi.

669 Blankenship RE, Hartman H. 1998. The origin and evolution of oxygenic photosynthesis. *Trends*
670 *Biochem Sci*. 23:94-97. doi: 10.1016/S0968-0004(98)01186-4.

671 Blount ZD, Lenski RE, Losos JB. 2018. Contingency and determinism in evolution: Replaying
672 life's tape. *Science*. 362:eaam5979. doi: 10.1126/science.aam5979.

673 Boyd ES, Anbar AD, Miller S, Hamilton TL, Lavin M, Peters JW. 2011a. A late methanogen
674 origin for molybdenum-dependent nitrogenase. *Geobiology*. 9:221-232. doi: 10.1111/j.1472-
675 4669.2011.00278.x.

676 Boyd ES, Hamilton TL, Peters JW. 2011b. An alternative path for the evolution of biological
677 nitrogen fixation. *Front Microbiol*. 2:205. doi: 10.3389/fmicb.2011.00205.

678 Boyd ES, Peters JW. 2013. New insights into the evolutionary history of biological nitrogen
679 fixation. *Front Microbiol*. 4:201. doi: 10.3389/fmicb.2013.00201.

680 Buren S, Jimenez-Vicente E, Echavarri-Erasun C, Rubio LM. 2020. Biosynthesis of Nitrogenase
681 Cofactors. *Chem Rev*. 120:4921-4968. doi: 10.1021/acs.chemrev.9b00489.

682 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009.
683 BLAST+: architecture and applications. *BMC Bioinformatics*. 10:421. doi: 10.1186/1471-2105-
684 10-421.

685 Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated
686 alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*. 25:1972-1973. doi:
687 10.1093/bioinformatics/btp348.

688 Chen Y, Nishihara A, Haruta S. 2021. Nitrogen-fixing Ability and Nitrogen Fixation-related
689 Genes of Thermophilic Fermentative Bacteria in the Genus *Caldicellulosiruptor*. *Microbes*
690 *Environ*. 36:ME21018. doi: 10.1264/jsme2.ME21018.

691 Copley SD. 2015. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem*
692 *Sci*. 40:72-78. doi: 10.1016/j.tibs.2014.12.004.

693 Copley SD. 2021. Setting the stage for evolution of a new enzyme. *Curr Opin Struct Biol*. 69:41-
694 49. doi: 10.1016/j.sbi.2021.03.001.

695 De Tarafder A, Parajuli NP, Majumdar S, Kacar B, Sanyal S. 2021. Kinetic Analysis Suggests
696 Evolution of Ribosome Specificity in Modern Elongation Factor-Tus from 'Generalist' Ancestors.
697 *Mol Biol Evol*.msab114. doi: 10.1093/molbev/msab114.

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

700 Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high
701 throughput. *Nucleic Acids Res*. 32:1792-1797. doi: 10.1093/nar/gkh340.

702 Einsle O, Rees DC. 2020. Structural Enzymology of Nitrogenase Enzymes. *Chem Rev*.
703 120:4969-5004. doi: 10.1021/acs.chemrev.0c00067.

704 Falkowski PG. 1997. Evolution of the nitrogen cycle and its influence on the biological
705 sequestration of CO_2 in the ocean. *Nature*. 387:272-275. doi: 10.1038/387272a0.

706 Fani R, Gallo R, Lio P. 2000. Molecular evolution of nitrogen fixation: the evolutionary history of
707 the *nifD*, *nifK*, *nifE*, and *nifN* genes. *J Mol Evol*. 51:1-11. doi: 10.1007/s002390010061.

708 Fay AW, Blank MA, Rebelein JG, Lee CC, Ribbe MW, Hedman B, Hodgson KO, Hu Y. 2016.
709 Assembly scaffold NifEN: A structural and functional homolog of the nitrogenase catalytic
710 component. *Proc Natl Acad Sci USA*. 113:9504-9508. doi: 10.1073/pnas.1609574113.

711 Fox GE. 2010. Origin and Evolution of the Ribosome. *CSH Perspect Biol*. 2:a003483-a003483.
712 doi: 10.1101/cshperspect.a003483.

713 Fujita Y, Bauer CE. 2000. Reconstitution of Light-independent Protochlorophyllide Reductase
714 from Purified Bchl and BchN-BchB Subunits. *J Biol Chem*. 275:23583-23588. doi:
715 10.1074/jbc.M002904200.

716 Garcia AK, Kacar B. 2019. How to resurrect ancestral proteins as proxies for ancient
717 biogeochemistry. *Free Radic Biol Med*. 140:260-269. doi: 10.1016/j.freeradbiomed.2019.03.033.

718 Garcia AK, McShea H, Kolaczkowski B, Kacar B. 2020. Reconstructing the evolutionary history
719 of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization. *Geobiology*. 18:394-
720 411. doi: 10.1111/gbi.12381.

721 Gerlt JA, Babbitt PC. 2001. Divergent Evolution of Enzymatic Function: Mechanistically Diverse
722 Superfamilies and Functionally Distinct Suprafamilies. *Annu Rev Biochem*. 70:209-246. doi:
723 10.1146/annurev.biochem.70.1.209.

724 Ghebreamlak SM, Mansoorabadi SO. 2020. Divergent Members of the Nitrogenase
725 Superfamily: Tetrapyrrole Biosynthesis and Beyond. *ChemBioChem*. 21:1723-1728. doi:
726 10.1002/cbic.201900782.

727 Glass JB, Wolfe-Simon F, Anbar AD. 2009. Coevolution of metal availability and nitrogen
728 assimilation in cyanobacteria and algae. *Geobiology*. 7:100-123. doi: 10.1111/j.1472-
729 4669.2009.00190.x.

730 Gould SJ. 1989. Wonderful Life: The Burgess Shale and the Nature of History. New York, NY:
731 W. W. Norton & Company, Inc.

732 Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms
733 and methods to estimate maximum-likelihood phylogenies: assessing the performance of
734 PhyML 3.0. *Syst Biol*. 59:307-321. doi: 10.1093/sysbio/syq010.

735 Hoffman BM, Lukoyanov D, Yang ZY, Dean DR, Seefeldt LC. 2014. Mechanism of nitrogen
736 fixation by nitrogenase: the next stage. *Chem Rev*. 114:4041-4062. doi: 10.1021/cr400641x.

737 Hu Y, Fay AW, Lee CC, Wiig JA, Ribbe MW. 2010. Dual functions of NifEN: insights into the
738 evolution and mechanism of nitrogenase. *Dalton T*. 39. doi: 10.1039/b922555b.

739 Hu Y, Fay AW, Lee CC, Yoshizawa J, Ribbe MW. 2008. Assembly of Nitrogenase MoFe
740 Protein. *Biochemistry*. 47:3973-3981. doi: 10.1021/bi7025003.

741 Hu Y, Fay AW, Schmid B, Makar B, Ribbe MW. 2006. Molecular Insights into Nitrogenase
742 FeMoco Insertion. *J Biol Chem*. 281:30534-30541. doi: 10.1074/jbc.M605527200.

743 Jimenez-Vicente E, Yang ZY, Ray WK, Echavarri-Erasun C, Cash VL, Rubio LM, Seefeldt LC,
744 Dean DR. 2018. Sequential and differential interaction of assembly factors during nitrogenase
745 MoFe protein maturation. *J Biol Chem*. 293:9812-9823. doi: 10.1074/jbc.RA118.002994.

746 Kacar B, Garcia AK, Anbar AD. 2020. Evolutionary History of Bioessential Elements Can Guide
747 the Search for Life in the Universe. *ChemBioChem*. doi: 10.1002/cbic.202000500.

748 Kaçar B, Gaucher EA. 2012. Towards the recapitulation of ancient history in the laboratory:
749 Combining synthetic biology with experimental evolution. *Artif Life*. 13:11-18. doi: 10.1162/978-
750 0-262-31050-5-ch002.

751 Kaiser JT, Hu Y, Wiig JA, Rees DC, Ribbe MW. 2011. Structure of Precursor-Bound NifEN: A
752 Nitrogenase FeMo Cofactor Maturase/Insertase. *Science*. 331:91-94. doi:
753 10.1126/science.1196954.

754 Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast
755 model selection for accurate phylogenetic estimates. *Nature Methods*. 14:587-589. doi:
756 10.1038/nmeth.4285.

757 Kanehisa M, Sato Y, Morishima K. 2016. BlastKOALA and GhostKOALA: KEGG Tools for
758 Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol*. 428:726-731.
759 doi: 10.1016/j.jmb.2015.11.006.

760 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
761 improvements in performance and usability. *Mol Biol Evol*. 30:772-780. doi:
762 10.1093/molbev/mst010.

763 Kim CH, Newton WE, Dean DR. 1995. Role of the MoFe protein alpha-subunit histidine-195
764 residue in FeMo-cofactor binding and nitrogenase catalysis. *Biochemistry*. 34:2798-2808.

765 Kim J, Rees D. 1992. Structural models for the metal centers in the nitrogenase molybdenum-
766 iron protein. *Science*. 257:1677-1682. doi: 10.1126/science.1529354.

767 Lee H-I, Thrasher KS, Dean DR, Newton WE, Hoffman BM. 1998. ^{14}N Electron Spin−Echo
768 Envelope Modulation of the S = 3/2 Spin System of the *Azotobacter vinelandii* Nitrogenase
769 Iron−Molybdenum Cofactor. *Biochemistry*. 37:13370-13378. doi: 10.1021/bi980956a.

770 Li H. 2002. *Azotobacter vinelandii* Nitrogenase: Multiple Substrate-Reduction Sites and Effects
771 of pH on Substrate Reduction and CO Inhibition. [Dissertation]. [Blacksburg, VA]: Virginia
772 Polytechnic Institute and State University.

773 Luo G, Junium CK, Izon G, Ono S, Beukes NJ, Algeo TJ, Cui Y, Xie S, Summons RE. 2018.
774 Nitrogen fixation sustained productivity in the wake of the Palaeoproterozoic Great Oxygenation
775 Event. *Nat Commun*. 9:978. doi: 10.1038/s41467-018-03361-2.

776 McGlynn SE, Boyd ES, Peters JW, Orphan VJ. 2012. Classifying the metal dependence of
777 uncharacterized nitrogenases. *Front Microbiol*. 3:419. doi: 10.3389/fmicb.2012.00419.

778 Mehta MP, Baross JA. 2006. Nitrogen fixation at 92 degrees C by a hydrothermal vent
779 archaeon. *Science*. 314:1783-1786. doi: 10.1126/science.1134772.

780 Moore EK, Jelen BI, Giovannelli D, Raanan H, Falkowski PG. 2017. Metal availability and the
781 expanding network of microbial metabolisms in the Archaean eon. *Nat Geosci*. 10:629-636. doi:
782 10.1038/ngeo3006.

783 Moser J, Brocker MJ. 2011. Enzymatic Systems with Homology to Nitrogenase. In: Ribbe M,
784 editor. Nitrogen Fixation. Methods in Molecular Biology (Methods and Protocols): Humana
785 Press. p. 67-77.

786 Mus F, Alleman AB, Pence N, Seefeldt LC, Peters JW. 2018. Exploring the alternatives of
787 biological nitrogen fixation. *Metallomics*. 10:523-538. doi: 10.1039/c8mt00038g.

788 Mus F, Colman DR, Peters JW, Boyd ES. 2019. Geobiological feedbacks, oxygen, and the
789 evolution of nitrogenase. *Free Radic Biol Med*. doi: 10.1016/j.freeradbiomed.2019.01.050.

790 Navarro-Gonzalez R, McKay CP, Mvondo DN. 2001. A possible nitrogen crisis for Archaean life
791 due to reduced nitrogen fixation by lightning. *Nature*. 412:61-64. doi: Doi 10.1038/35083537.

792 Noda-Garcia L, Liebermeister W, Tawfik DS. 2018. Metabolite-Enzyme Coevolution: From
793 Single Enzymes to Metabolic Pathways and Networks. *Annu Rev Biochem*. 87:187-216. doi:
794 10.1146/annurev-biochem-062917-012023.

795 North JA, Narrowe AB, Xiong W, Byerly KM, Zhao G, Young SJ, Murali S, Wildenthal JA,
796 Cannon WR, Wrighton KC, et al. 2020. A nitrogenase-like enzyme system catalyzes
797 methionine, ethylene, and methane biogenesis. *Science*. 369:1094-1098. doi:
798 10.1126/science.abb6310.

799 Ohno S. 1970. Evolution by gene duplication. Berlin: Springer-Verlag.

800 Parsons C, Stüeken EE, Rosen CJ, Mateos K, Anderson RE. 2020. Radiation of nitrogen-
801 metabolizing enzymes across the tree of life tracks environmental transitions in Earth history.
802 *Geobiology*. 19:18-34. doi: 10.1111/gbi.12419.

803 Perez-Gonzalez A, Jimenez-Vicente E, Gies-Elterlein J, Salinero-Lanzarote A, Yang ZY, Einsle
804 O, Seefeldt LC, Dean DR. 2021. Specificity of NifEN and VnfEN for the Assembly of
805 Nitrogenase Active Site Cofactors in Azotobacter vinelandii. *mBio*. 12:e0156821. doi:
806 10.1128/mBio.01568-21.

807 Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for
808 large alignments. *PLoS One*. 5:e9490. doi: 10.1371/journal.pone.0009490.

809 Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The natural history of nitrogen
810 fixation. *Mol Biol Evol*. 21:541-554. doi: 10.1093/molbev/msh047.

811 Seefeldt LC, Yang ZY, Lukoyanov DA, Harris DF, Dean DR, Raugei S, Hoffman BM. 2020.
812 Reduction of Substrates by Nitrogenases. *Chem Rev*. 120:5082-5106. doi:
813 10.1021/acs.chemrev.9b00556.

814 Sickerman NS, Tanifuji K, Hu Y, Ribbe MW. 2017. Synthetic Analogues of Nitrogenase
815 Metallocofactors: Challenges and Developments. *Chemistry*. 23:12425-12432. doi:
816 10.1002/chem.201702496.

817 Silver WS, Postgate JR. 1973. Evolution of asymbiotic nitrogen fixation. *J Theor Biol*. 40:1-10.
818 doi: 10.1016/0022-5193(73)90160-4.

819 Smethurst DGJ, Shcherbik N. 2021. Interchangeable utilization of metals: New perspectives on
820 the impacts of metal ions employed in ancient and extant biomolecules. *J Biol Chem*.
821 297:101374. doi: 10.1016/j.jbc.2021.101374.

822 Solomon JB, Lee CC, Jasniewski AJ, Rasekh MF, Ribbe MW, Hu Y. 2020. Heterologous
823 Expression and Engineering of the Nitrogenase Cofactor Biosynthesis Scaffold NifEN. *Angew*
824 *Chem Int Ed Engl*. 59:6887-6893. doi: 10.1002/anie.201916598.

825 Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of
826 large phylogenies. *Bioinformatics*. 30:1312-1313. doi: 10.1093/bioinformatics/btu033.

827 Staples CR, Lahiri S, Raymond J, Von Herbulis L, Mukhophadhyay B, Blankenship RE. 2007.
828 Expression and association of group IV nitrogenase NifD and NifH homologs in the non-
829 nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *J Bacteriol*. 189:7392-7398. doi:
830 10.1128/JB.00876-07.

Stueken EE, Buick R, Guy BM, Koehler MC. 2015. Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature*. 520:666-669. doi: 10.1038/nature14180.

Stüeken EE, Kipp MA, Koehler MC, Buick R. 2016. The evolution of Earth's biogeochemical nitrogen cycle. *Earth-Sci Rev*. 160:220-239. doi: 10.1016/j.earscirev.2016.07.007.

Tanifuji K, Lee CC, Ohki Y, Tatsumi K, Hu Y, Ribbe MW. 2015. Combining a Nitrogenase Scaffold and a Synthetic Compound into an Artificial Enzyme. *Angew Chem Int Ed Engl*. 127:14228-14231. doi: 10.1002/ange.201507646.

Thornton JW. 2004. Resurrecting ancient genes: experimental analysis of extinct molecules. *Nat Rev Genet*. 5:366-375. doi: 10.1038/nrg1324.

Vermeij GJ. 2006. Historical contingency and the purported uniqueness of evolutionary innovations. *Proc Natl Acad Sci USA*. 103:1804-1809. doi: 10.1073/pnas.0508724103.

Vialle RA, Tamuri AU, Goldman N. 2018. Alignment Modulates Ancestral Sequence Reconstruction Accuracy. *Mol Biol Evol*. 35:1783-1797. doi: 10.1093/molbev/msy055.

Williams RJP, da Silva JRRF. 1996. The natural selection of the chemical elements: the environment and life's chemistry: Clarendon Press.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586-1591. doi: 10.1093/molbev/msm088.

Zheng K, Ngo PD, Owens VL, Yang X-p, Mansoorabadi SO. 2016. The biosynthetic pathway of coenzyme F430 in methanogenic and methanotrophic archaea. *Science*. 354:339-342. doi: 10.1126/science.aag2947.