TITLE: Reconstructing the Evolutionary History of Nitrogenases: Evidence for Ancestral Molybdenum-Cofactor Utilization

SHORT TITLE: Reconstructing the evolutionary history of nitrogenases

AUTHORS: Amanda K. Garcia[1], Hanon McShea[2,a], Bryan Kolaczkowski[3], Betül Kaçar[1,4]

AFFILIATIONS:

[1]Department of Molecular and Cellular Biology, University of Arizona, Tucson, Arizona

[2]Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts

[3]Department of Microbiology and Cell Science, University of Florida, Gainesville, Florida

[4]Department of Astronomy and Steward Observatory, University of Arizona, Tucson, Arizona

[a]Present address: Department of Earth System Science, Stanford University, Stanford, California

CORRESPONDING AUTHOR INFORMATION:

Betül Kaçar

1007 E. Lowell St., Tucson, AZ 85721

betul@arizona.edu

ABSTRACT

The nitrogenase metalloenzyme family, essential for supplying fixed nitrogen to the biosphere, is one of life's key biogeochemical innovations. The three isozymes of nitrogenase differ in their metal dependence, each binding either a FeMo-, FeV-, or FeFe-cofactor where the reduction of dinitrogen takes place. The history of nitrogenase metal dependence has been of particular interest due to the possible implication that ancient marine metal availabilities have significantly constrained nitrogenase evolution over geologic time. Here, we reconstructed the evolutionary history of nitrogenases, and combined phylogenetic reconstruction, ancestral sequence inference, and structural homology modeling to evaluate the potential metal dependence of ancient nitrogenases. We find that active-site sequence features can reliably distinguish extant Mo-nitrogenases from V- and Fe-nitrogenases, and that inferred ancestral sequences at the deepest nodes of the phylogeny suggest these ancient proteins most resemble modern Mo-nitrogenases. Taxa representing early-branching nitrogenase lineages lack one or more biosynthetic *nifE* and *nifN* genes that both contribute to the assembly of the FeMo-cofactor in studied organisms, suggesting that early Mo-nitrogenases may have utilized an alternate and/or simplified pathway for cofactor biosynthesis. Our results underscore the profound impacts that protein-level innovations likely had on shaping global biogeochemical cycles throughout the Precambrian, in contrast to organism-level innovations that characterize the Phanerozoic Eon.

KEYWORDS

ancestral sequence reconstruction, cofactor pocket, metalloenzyme, nitrogenase, nitrogen fixation

1 INTRODUCTION

All known life requires nitrogen for the synthesis of essential biomolecules, including nucleotides and amino acids. However, though the atmosphere contains nearly 80% N_2 by volume, most organisms are not able to assimilate N_2. Select bacteria and archaea called diazotrophs accomplish biological nitrogen fixation by nitrogenase metalloenzymes (E.C. 1.18.6.1) that catalyze the reduction of N_2 to bioavailable NH_3. Both phylogenetic and geochemical evidence suggest that nitrogenases are an ancient family of enzymes; time-calibrated phylogenies place the evolution of nitrogenases at ~1.5–2.2 billion years ago (Ga) (Boyd, Anbar, et al., 2011), and the oldest potential isotopic biosignatures of nitrogenase activity date even further back to ~3.2 Ga (Stueken, Buick, Guy, & Koehler, 2015). Because nitrogen has been suggested to be an important limiting nutrient on geologic timescales (Falkowski, 1997), nitrogenases have likely played a key role in the expansion of the biosphere for much of Earth's history.

The nitrogenase family consists of three homologous isozymes (Boyd, Hamilton, & Peters, 2011; Raymond, Siefert, Staples, & Blankenship, 2004) named after the differential metal content of the active-site cofactor: Mo-nitrogenase (Nif, encoded by *nif*), V-nitrogenase (Vnf, encoded by *vnf*), and Fe-nitrogenase (Anf, encoded by *anf*) (Bulen & LeComte, 1966; Eady, 1996; Joerger & Bishop, 1988; Mus, Alleman, Pence, Seefeldt, & Peters, 2018) (Figure 1). Of the three isozymes, Mo-nitrogenases are the most common and widely studied; V- and Fe-nitrogenases are comparatively rarer and only known in taxa that also possess Mo-nitrogenase (Boyd, Hamilton, et al., 2011; Dos Santos, Fang, Mason, Setubal, & Dixon, 2012). All three nitrogenase isozymes are structurally and functionally similar, each containing two protein components: the electron delivery component (NifH, VnfH, or AnfH) is a homodimer and the catalytic component is either

an $\alpha_2\beta_2$ heterotetramer (MoFe protein, NifDK) or an $\alpha_2\beta_2\gamma_2$ heterohexamer (VFe protein, VnfDGK or FeFe protein, AnfDGK) (Figure 1a) (Bulen & LeComte, 1966; Hales, Case, Morningstar, Dzeda, & Mauterer, 1986; Schmid et al., 2002; D. Sippel & Einsle, 2017). During catalysis, the electron delivery component transiently associates with and delivers electrons to the catalytic component (Hageman & Burris, 1978). Electrons accumulate at the active site for N_2 binding and reduction (Hoffman, Lukoyanov, Yang, Dean, & Seefeldt, 2014), which houses a homocitrate-metallocluster cofactor unique to each isozyme: the FeMo-cofactor in Nif (Figure 1b), FeV-cofactor in Vnf (Figure 1c), and FeFe-cofactor in Anf (Figure 1d) (Eady, 1996; Harris, Lukoyanov, et al., 2018; Krahn et al., 2002; Mus et al., 2018; D. Sippel & Einsle, 2017; Spatzal et al., 2011). Available spectral evidence suggests that these cofactors are structurally similar, with the main difference being the substitution of a Mo, V, or additional Fe atom (FeV-cofactor is also proposed to incorporate a carbonate ligand in place of one sulfur atom (D. Sippel & Einsle, 2017)) (Eady, 1996; Krahn et al., 2002; Spatzal et al., 2011). Nevertheless, biochemical studies demonstrate variable catalytic properties among the three nitrogenase isozymes, including differential abilities to reduce alternative substrates (Harris, Lukoyanov, et al., 2018; Harris, Yang, Dean, Seefeldt, & Hoffman, 2018; Hu et al., 2018; Hu, Lee, & Ribbe, 2011; Zheng et al., 2018). These catalytic variations likely arise due to a combination of the aforementioned cofactor compositional differences as well as differences in the surrounding protein environment (Fixen et al., 2016; Harris, Yang, et al., 2018; Lee et al., 2018; Rebelein, Lee, Newcomb, Hu, & Ribbe, 2018; Zheng et al., 2018).

Metal cofactor incorporation in nitrogenases is constrained at multiple levels. At the level of single enzyme functionality, nitrogenase biochemical and biophysical properties shape metal-binding

87 specificity. At a higher level, constraints arise from the partner proteins that constitute the
88 biosynthetic mechanism for active-site cofactor assembly and insertion, best studied in Mo-
89 nitrogenases (Curatti et al., 2007; Hu & Ribbe, 2011; Rubio & Ludden, 2008). In the *Azotobacter*
90 *vinelandii* (*A. vinelandii*) model and *in vitro* systems, FeMo-cofactor assembly requires several
91 associated proteins encoded within the *nif* gene cluster (Curatti et al., 2007; Hu, Fay, & Ribbe,
92 2005; Shah, Allen, Spangler, & Ludden, 1994; Shah, Imperial, Ugalde, Ludden, & Brill, 1986;
93 Tal, Chun, Gavini, & Burgess, 1991). Most notably, the biosynthetic *nifB, nifE,* and *nifN* genes are
94 considered, in addition to the catalytic *nifHDK* genes, to be — perhaps minimally — required for
95 FeMo-cofactor assembly and Mo-nitrogenase function (Boyd, Anbar, et al., 2011; Curatti et al.,
96 2007; Dos Santos et al., 2012; Hu et al., 2005; Shah et al., 1994; Shah et al., 1986; Tal et al., 1991).
97 In *A. vinelandii, nifE* and *nifN* loci are located just downstream of the *nifHDK* cluster, whereas the
98 *nifB* locus is located within a separate *nif* region near other regulatory and biosynthetic *nif* genes
99 (Setubal et al., 2009). NifB catalyzes the formation of a Fe-S-C metallocluster, a precursor that
100 forms the core of the mature FeMo-cofactor (Allen, Chatterjee, Ludden, & Shah, 1995; Hu &
101 Ribbe, 2011). This precursor metallocluster is then transferred to a protein heterotetramer
102 composed of NifE and NifN subunits (Allen et al., 1995; Roll, Shah, Dean, & Roberts, 1995),
103 homologous to NifD and NifK, respectively, and likely having arisen by gene duplication (Boyd,
104 Anbar, et al., 2011). Within NifEN, the precursor is further modified via the addition of
105 homocitrate and Mo, and the mature cofactor is subsequently transferred to the nitrogenase NifDK
106 catalytic protein component (Roll et al., 1995). Unlike that for the FeMo-cofactor, the biosynthetic
107 pathways for the formation of the FeV- and FeFe-cofactors are relatively unknown. However,
108 transcriptional profiling of the three nitrogenase systems in *A. vinelandii* suggests that FeV- and
109 FeFe-cofactor synthesis relies on several *nif* genes in addition to *vnf* and *anf* genes, respectively

(Hamilton et al., 2011; Joerger & Bishop, 1988; Kennedy & Dean, 1992). These include *nifBEN* for most alternative gene clusters, with the exception of certain taxa (including *A. vinelandii*) that possess *vnfEN* homologs of *nifEN* that likely perform a similar biosynthetic function (Boyd, Anbar, et al., 2011; Boyd & Peters, 2013; Hamilton et al., 2011).

Paleobiological interest in nitrogenases has primarily centered on the coevolution of nitrogenase metal usage and the geochemical environment, with the possible implication that marine metal availabilities have significantly constrained nitrogenase evolution over geologic time (Anbar & Knoll, 2002; Boyd, Hamilton, et al., 2011; Canfield, Glazer, & Falkowski, 2010; Raymond et al., 2004). Inferences of ancient nitrogenase metal usage have relied on isotopic biosignatures (Stueken et al., 2015) and metal abundances (Anbar & Knoll, 2002) evidenced by the geologic record, as well as on phylogenetic reconstructions of both catalytic and cofactor biosynthesis proteins (Boyd, Anbar, et al., 2011; Boyd, Hamilton, et al., 2011; Raymond et al., 2004). High marine Fe concentrations and potential Mo scarcity prior to increased atmospheric oxygenation surrounding the ~2.3-2.5 Ga Great Oxidation Event (Anbar et al., 2007; Lyons, Reinhard, & Planavsky, 2014) has led to the hypothesis that Fe- or V- nitrogenases may have been dominant in early oceans (Anbar & Knoll, 2002; Canfield et al., 2010) and possibly predate Mo-nitrogenases (Raymond et al., 2004). More recent phylogenetic reconstructions have instead suggested that the evolution of Mo-nitrogenases, dated by time-calibrated phylogenies of Nif/Vnf/AnfDKEN sequences to ~1.5–2.2 Ga (Boyd, Anbar, et al., 2011), preceded that of V- and Fe-nitrogenases (Boyd, Hamilton, et al., 2011). These phylogenetic inferences are also consistent with the observation that *vnf* and *anf* genes are only present in organisms that also harbor *nif*, and that V-/Fe-nitrogenase assembly relies on *nif* biosynthetic genes (Hamilton et al., 2011; Joerger & Bishop,

1988; Kennedy & Dean, 1992). However, ~3.2-Ga isotopic signatures of biological nitrogen fixation suggest an earlier origin of nitrogenase (Stueken et al., 2015), and, even though isotopically consistent with Mo-dependent nitrogen fixation, predate age estimates of both Mo-nitrogenase (Boyd, Anbar, et al., 2011) and earliest marine Mo availability (Anbar et al., 2007; Anbar & Knoll, 2002; Lyons et al., 2014). Thus, the evolutionary trajectory of nitrogenase metal usage — and by extension the link between nitrogenase evolution and marine metal availabilities over geologic time — is not yet known.

Here, we explored the indicators of nitrogenase metal usage history by a combined method relying on ancestral sequence reconstruction, an evolutionary approach by which inferred, historical protein sequence information can be linked to functional inferences of molecular properties evidenced by computational analyses or laboratory experiments (Aadland, Pugh, & Kolaczkowski, 2019; Benner, Sassi, & Gaucher, 2007; Thornton, 2004). These paleogenetic approaches have been increasingly applied in biogeochemically relevant molecular studies to offer insights into the coevolution of life and Earth (Garcia & Kacar, 2019; Gomez-Fernandez et al., 2018; Kacar, Hanson-Smith, Adam, & Boekelheide, 2017). We reconstructed the phylogenetic history of Mo-, V-, and Fe-nitrogenases in order to resurrect ancestral nitrogenases *in silico*, as well as to map the taxonomic distribution of cofactor biosynthetic components considered necessary for cofactor assembly (Boyd, Anbar, et al., 2011; Curatti et al., 2007; Dos Santos et al., 2012; Hu et al., 2005; Shah et al., 1994; Shah et al., 1986; Tal et al., 1991). Through this combined approach, we access ancestral enzyme sequence features suggestive of ancient Mo-dependence. We further present phylogenetic data that suggests that the ancient Mo-utilization was potentially achieved via a unique pathway for the synthesis of the FeMo- or similar cofactor that may be present in poorly

156 characterized, basal lineages. Integration of protein evolution and paleobiology is a unique
157 melding of disparate data sets; this approach may provide hypotheses that address interactions
158 ranging from the external environment to the cellular environment, and from the cellular
159 environment to that maintained around the interacting protein. The exchange of materials across
160 these different scales necessitates constraints on the flow and availability of substrates that make
161 such exchanges possible. It is the specific nature of these constraints and how they may change in
162 response to external perturbations that enable us to develop completely new experimentally
163 testable hypotheses that connect geochemical reservoirs with biological metabolisms — those that
164 cannot be constructed from macroevolutionary or geological frameworks alone.
165

166 **2 MATERIALS AND METHODS**

167

168 **2.1 Ancestral reconstruction of nitrogenase protein sequences**

169 An initial dataset of extant nitrogenase Nif /Vnf/AnfHDK homologs was constructed by retrieving
170 amino acid sequences from the National Center for Biotechnology Information non-redundant
171 protein database, accessed September 2018 (O'Leary et al., 2016). Potential homologs were
172 identified by BLASTp (Camacho et al., 2009) using query sequences from *A. vinelandii* (NifH:
173 Avin_01380, NifD: Avin_01390, NifK: Avin_01400) and an Expect value cutoff of <1e-5. The
174 dataset was then manually curated to remove partial and distantly related sequences. Additional
175 nitrogenase sequences were manually retrieved from the Joint Genome Institute Integrated
176 Microbial Genomes and Microbiomes database, accessed September 2018 (Chen et al., 2019). The
177 nitrogenase sequence dataset was finalized to include NifHDK sequences from 256 taxa, AnfHDK
178 sequences from 14 taxa, VnfHDK sequences from 14 taxa, and outgroup light-independent

179 protochlorophyllide oxidoreductase (Bch/ChlLNB) sequences — sharing distant homology with
180 nitrogenases (Boyd, Anbar, et al., 2011; Hu & Ribbe, 2015; Raymond et al., 2004) — from 10 taxa
181 (Appendix S1; additional analyses were performed with an expanded outgroup, Appendix S2).
182 Only one Nif/Anf/VnfHDK sequence set was retained per genus to broaden taxonomic sampling.
183 Equal sequence sampling for Anf and Vnf was made to remove the potential for oversampling bias
184 in ancestral sequence inference. H-, D-, and K-subunit sequences corresponding to each taxon
185 were manually checked for synteny of their encoding genes. AnfHDK and VnfHDK sequences
186 were identified by the proximity of each gene locus to *anfG* or *vnfG*, which encodes the additional
187 G-subunit present in the VeFe or FeFe protein, respectively, but not present in the MoFe protein
188 (Eady, 1996). Finally, the presence of the cofactor biosynthetic *nifBEN* genes was investigated for
189 all taxa represented in our dataset by BLASTp, as well as by manually inspecting the *nif* genome
190 region.
191
192 Reconstruction of ancestral nitrogenase sequences was performed by PhyloBot (Hanson-Smith &
193 Johnson, 2016) (www.phylobot.com), which automates multiple sequence alignment,
194 phylogenetic reconstruction, and ancestral sequence inference methods. The concatenated 294-
195 sequence dataset of Nif/Anf/VnfHDK homologs (including 10 Bch/ChlLNB outgroup sequences)
196 was aligned by MSAProbs v0.9 5r1 (Liu, Schmidt, & Maskell, 2010) and MUSCLE v3.8.31
197 (Edgar, 2004). Both alignment outputs were then used to perform phylogenetic reconstruction by
198 RAxML v8.1.15 (Stamatakis, 2014) under 6 different combinations of amino acid substitution and
199 rate heterogeneity models. Branch support was evaluated by the approximate likelihood ratio test
200 (aLRT) (Anisimova & Gascuel, 2006), which assesses the gain in overall likelihood against a null
201 hypothesis of branch length = 0. Additional phylogenetic reconstructions with an expanded

202 outgroup were performed outside of Phylobot to resolve root positioning, but were not used in
203 subsequent ancestral sequence inference (Appendix S2). Ancestral sequences were inferred by
204 joint maximum likelihood using CODEML v4.2 (Yang, 2007) at all nodes within the 12 Phylobot-
205 constructed phylogenies (Tree-1 – Tree-12), with gaps inferred by parsimony. To assess ancestral
206 sequence robustness to phylogenetic uncertainty (Hanson-Smith, Kolaczkowski, & Thornton,
207 2010), ancestors inferred from the top five phylogenies ranked by log likelihood scores were
208 selected for further analysis (Table 1, Appendix S2). Finally, to evaluate the effects of ambiguously
209 reconstructed sites on subsequent structural analyses, Bayesian sampled ancestors were inferred
210 from the maximum likelihood site posterior probabilities calculated by CODEML (Aadland et al.,
211 2019). 100 random Bayesian sequences were generated for each of five ancestral nodes of interest
212 across the top five phylogenies. Thus, 25 maximum likelihood and 2,500 Bayesian-sampled
213 ancestral sequences were analyzed in total. All maximum likelihood reconstructed trees and
214 ancestral sequences are available for view and download at http://phylobot.com/613282215/.
215

216 **2.2 Probabilistic model for metal-dependence classification of nitrogenase sequences**

217 Extant nitrogenase sequences of known metal dependence were labelled as binding either the
218 FeMo- ("Nif"), FeV- ("Vnf"), or FeFe-cofactor ("Anf") according to their phylogenetic clustering,
219 and were used to train a support vector classifier that identifies cofactor specificity, based on the
220 residues in the active-site cofactor-binding pocket. The 30 active-site residues were identified as
221 those residing within 5 Å of any atom in either the FeMo-cofactor of the *A. vinelandii* NifD protein
222 (PDB 3U7Q (Spatzal et al., 2011)) or the FeV-cofactor of the *A. vinelandii* VnfD protein (PDB

223 Table 1. Alignment and evolutionary model parameter combinations for the top five phylogenies, ranked by log likelihood scores.

Phylogeny	Alignment method	Evolutionary model	Log likelihood	Maximum likelihood ancestors
Tree-1	MSAProbs	CAT + LG	-300069.08	AncA-1 – AncE-1
Tree-2	MSAProbs	CAT + WAG	-303296.08	AncA-2 – AncE-2
Tree-3	MSAProbs	Γ + LG	-303951.21	AncA-3 – AncE-3
Tree-4	MUSCLE	CAT + LG	-304457.52	AncA-4 – AncE-4
Tree-5	MSAProbs	Γ + WAG	-305229.49	AncA-5 – AncE-5

230 References: (Le & Gascuel, 2008; Quang, Gascuel, & Lartillot, 2008; Whelan & Goldman, 2001; Yang, 1993)

231

5N6Y (D. Sippel & Einsle, 2017)) (see Section 3.3 for further discussion of selected residues). The corresponding residues in other extant homologs and inferred ancestral sequences were identified by multiple sequence alignment using MAFFT (Katoh & Standley, 2013). For each sequence in the labelled dataset and residue in the active-site, the probability distribution over all 20 possible amino acids was inferred by averaging (1) the observed distribution for that sequence (which places probability 1.0 on the observed residue, (2) the probability distribution that would be expected over a short evolutionary distance (0.01 substitutions/site), starting from the observed residue and assuming the JTT substitution model (Jones, Taylor, & Thornton, 1992), and (3) a diffuse Dirichlet mixture regularizer that places relatively low probabilities across all possible amino acids (Sjolander et al., 1996). The labelled dataset was randomly divided into 60% training data for training the support vector classifier and 40% out-of-sample testing. The support vector classifier used a radial basis kernel, with the kernel coefficient estimated as 1 / (# of predicted features × Var(features)). Data samples were weighted based on the frequencies of Nif, Vnf, and Anf labels in the training data: the weight assigned to samples labelled as cofactor I (c_i) = # of samples / (3 × Freq(c_i)). The regularization parameter was optimized using 5-fold cross-validation, and classification accuracy was assessed by out-of-sample testing. The procedure of randomly splitting data into training/testing, training the classifier, and assessing out-of-sample accuracy was replicated 10 times.

Extant sequences and maximum-likelihood ancestral-reconstructed sequences were classified as Nif, Vnf, or Anf using one-vs-rest multi-class classification, based on the observed amino-acid residues in their active sites. In addition, ancestral sequences were also classified based on the inferred posterior-probability distributions over all 20 possible amino acids at each position in the

active site. Classification support is reported as the distance of each data sample from the support-vector hyperplane defining the inferred class.

2.3 Structural homology modeling of extant and ancestral nitrogenase D-subunits

Structural homology modeling of extant and ancestral (25 maximum likelihood and 2,500 Bayesian-sampled) nitrogenase D-subunit proteins was performed by Modeller v9.2 (Sali & Blundell, 1993). Extant nitrogenase sequences, broadly sampled from the reconstructed nitrogenase phylogeny, were modeled to provide comparisons with ancestral models. D-subunit sequences from extant and ancestral nitrogenases were aligned to 38 NifD and 2 VnfD structural templates retrieved from the Protein Data Bank (Berman et al., 2000), accessed November 2018 (Appendix S3; published AnfD models not available at time of analysis). Information from all 40 templates was used to model each structure. All models were generated by specifying the inclusion of the FeMo-cofactor of the 3U7Q NifD structure (Spatzal et al., 2011), selected as the highest resolution Mo-nitrogenase template. To assess the effect of the template cofactor type on the generated structure, additional models were constructed by specifying the inclusion of the FeV-cofactor of the 56NY VnfD template (D. Sippel & Einsle, 2017) (Appendix S3). Modeling replicates (100 in total) were performed per sequence and assessed by averaging over the scaled Modeller objective function, Discrete Optimized Protein Energy, and high resolution Discrete Optimized Protein Energy scores, as previously described (Aadland et al., 2019). The ten best modeling replicates per extant sequence, ten best replicates per maximum likelihood ancestral sequence, and the single best replicate per Bayesian-sampled variant sequence were selected for further analysis, totaling 3,080 models with the FeMo-cofactor specified.

278 **2.4 Active-site pocket volume calculation of extant and ancestral D-subunit models**

279 Volumes of the modeled ancestral and extant D-subunit active-site cofactor pockets were
280 calculated by POVME v2.0 (Durrant, Votapka, Sorensen, & Amaro, 2014). Spatial coordinates
281 and the inclusion region for volume calculation were specified manually. Pocket volumes were
282 calculated with a grid spacing of 0.5 Å and a 1.09 Å distance cutoff from any receptor atom's van
283 der Waals radius. Volume outside of the modeled convex hull of the cofactor pocket as well as
284 noncontiguous volume were removed. Statistical analysis of ancestral and extant pocket volume
285 data was performed in R (R Core Team, 2014).

286

287 **3 RESULTS**

288

289 **3.1 V- and Fe-nitrogenases diversified after Mo-nitrogenases**

290 We reconstructed the phylogenetic history of Mo-, V-, and Fe-nitrogenases to infer ancestral
291 nitrogenase sequences and associated indicators of nitrogenase metal dependence.
292 Nif/Anf/VnfHDK protein homologs curated from the National Center for Biotechnology
293 Information and Joint Genome Institute databases represent 20 bacterial and archaeal phyla, 11 of
294 which are known from experimental investigations to include diazotrophic taxa (Dos Santos et al.,
295 2012; Ormeño-Orrillo, Hungria, & Martinez-Romero, 2013) (Appendix S1). The five most
296 represented phyla in our dataset — Bacteroidetes, Cyanobacteria, Firmicutes, Proteobacteria, and
297 Euryarchaeota — encompass ~80% of the curated sequences. Our genomic dataset also suggests
298 the presence of nitrogen fixation within the Acidobacteria, Actinobacteria, Aquificae, Chlorobi,
299 Chloroflexi, Chrysiogenetes, Deferribacteres, Elusimicrobia, Fusobacteria, Lentisphaerae,

300 *Candidatus* Margulisbacteria, Nitrospirae, Planctomycetes, Spirochaetes, and Verrucomicrobia
301 clades.
302
303 Our maximum likelihood nitrogenase phylogeny segregates Nif/Anf/VnfHDK sequences into two
304 major lineages similarly observed in previous studies (Boyd, Anbar, et al., 2011; Raymond et al.,
305 2004) (Tree-1; Figure 2): the first comprises Nif-I and Nif-II Mo-nitrogenases (primarily
306 aerobic/facultative bacteria and anaerobic/facultative bacteria/archaea, respectively; highlighted in
307 blue) (Boyd, Costas, Hamilton, Mus, & Peters, 2015; Raymond et al., 2004), and the second
308 comprises V- and Fe-only-nitrogenases (Vnf and Anf, respectively; highlighted in red) as well as
309 three clades of "uncharacterized" nitrogenases that lack extensive experimental characterization
310 with regard to metal dependence (Mb-Mc, highlighted in purple; F-Mc, highlighted in green; Clfx,
311 highlighted in yellow) (Boyd, Hamilton, et al., 2011; Dekas, Poretsky, & Orphan, 2009; Dos
312 Santos et al., 2012; Mehta & Baross, 2006). We additionally analyzed four alternate phylogenies
313 (ranked by log likelihood scores), reconstructed by varying alignment methods, amino acid
314 substitution models, and rate heterogeneity models (Tree-2–Tree-5; Appendix S2). The
315 aforementioned major nitrogenase clades are present across all alternate phylogenetic topologies.
316 Branch ordering of these major clades is similarly consistent, with the exception of root position:
317 in Tree-3 and Tree-5, the root is placed between Nif-I and Nif-II/uncharacterized/Anf/Vnf, rather
318 than between Nif-I/Nif-II and uncharacterized/Anf/Vnf as in Tree-1, Tree-2, and Tree-4.
319 Nevertheless, additional phylogenetic analyses incorporating an expanded outgroup provide
320 stronger support for the root position of Tree-1, Tree-2, and Tree-4 (Appendix S2).
321

322 In all analyzed nitrogenase phylogenies, Vnf and Anf sequences (highlighted in red) form
323 reciprocally monophyletic clades that branch immediately distal to Mb-Mc nitrogenases
324 (highlighted in purple) (Figure 2, Appendix S2). The reciprocal monophyly of Vnf and Anf, as
325 well as the clustering of Mb-Mc, Vnf, and Anf, is well-supported across all phylogenetic
326 topologies (aLRT > 10^6). The Mb-Mc clade is composed of archaeal hydrogenotrophic
327 methanogens within classes Methanobacteria and Methanococci. Because all Mb-Mc taxa possess
328 *nifBEN* genes considered to be minimally required for the synthesis of the FeMo-cofactor (Boyd,
329 Anbar, et al., 2011; Curatti et al., 2007; Dos Santos et al., 2012; Hu et al., 2005; Shah et al., 1994;
330 Shah et al., 1986; Tal et al., 1991), it is likely that Mb-Mc nitrogenases are Mo-dependent (Boyd,
331 Hamilton, et al., 2011). Thus, the phylogenetic positioning of Vnf and Anf is consistent with
332 previous suggestions that V- and Fe-nitrogenases diversified after Mo-nitrogenases (Boyd,
333 Hamilton, et al., 2011).
334

335 **3.2 Basal uncharacterized nitrogenases lack associated genes for FeMo-cofactor synthesis**

336 In addition to investigating the phylogenetic relationships between Mo-, V-, and Fe-nitrogenase
337 isozymes, we mapped the presence of the biosynthetic *nifB*, *nifE*, and *nifN* genes — considered to
338 be necessary for FeMo-cofactor assembly (Boyd, Anbar, et al., 2011; Curatti et al., 2007; Dos
339 Santos et al., 2012; Hu et al., 2005; Shah et al., 1994; Shah et al., 1986; Tal et al., 1991) — among
340 taxa represented in our dataset. All analyzed taxa possess the full complement of *nifBEN*
341 biosynthetic genes, with the exception of two uncharacterized clades: Clfx (highlighted in yellow)
342 and F-Mc (highlighted in green) (Figure 2). Within the lineage containing Vnf and Anf
343 nitrogenases, Clfx and F-Mc clades are most basal. These branching positions of Clfx and F-Mc
344 clades within the Vnf/Anf lineage are consistently observed across all phylogenetic topologies

345 (Trees-1–5; Appendix S2) and are well-supported (aLRT > 10^3 for Clfx, aLRT > 10^{18} for F-Mc).
346 In Tree-1-, 2-, and -4, as well as in trees reconstructed with an expanded outgroup (Appendix S2),
347 Clfx and F-Mc clades also branch immediately distal to the root.
348
349 Both basal Clfx and F-Mc homologs are found primarily in thermophilic taxa and lack one or more
350 associated *nifEN* cofactor biosynthesis genes, as has been noted previously (Boyd, Hamilton, et
351 al., 2011; Dos Santos et al., 2012; Shock & Boyd, 2015). Clfx homologs, constituting the most
352 basal uncharacterized clade within the uncharacterized/Vnf/Anf lineage, represent in the present
353 dataset three mesophilic or thermophilic Chloroflexi species (Figure 2, highlighted in yellow). The
354 *nif* clusters of Clfx taxa lack both *nifE* and *nifN* genes, and the typically continuous *nifHDK* genes
355 observed in other taxa are instead interrupted by *nifB* (arranged *nifHBDK*) (Setubal et al., 2009).
356 F-Mc homologs branch immediately distal to the Clfx clade and represent eight thermophilic
357 Firmicutes and archaeal methanogen (class Methanococci) taxa (Figure 2, highlighted in green).
358 F-Mc species possess biosynthetic *nifB* and *nifE* genes, but not *nifN*, with the exception of
359 *Methanothermococcus thermolithotrophicus* that retains *nifN*. A previous study found that
360 sequence features and modeled structural features of F-Mc nitrogenases closely resemble those of
361 Mo-nitrogenases (McGlynn, Boyd, Peters, & Orphan, 2012). However, the absence of *nifN* in most
362 F-Mc taxa suggests that these strains may not be capable of synthesizing the FeMo-cofactor, given
363 previous *in vitro* and *in vivo* experiments demonstrating the requirement of both *nifEN* for FeMo-
364 cofactor biosynthesis (Curatti et al., 2007; Hu et al., 2005; Shah et al., 1994; Shah et al., 1986; Tal
365 et al., 1991). The lack of one or more *nifEN* genes in Clfx and F-Mc taxa might additionally suggest
366 that such strains cannot express functional nitrogenases. Nitrogen fixation among Clfx taxa has
367 not been definitively demonstrated; the Clfx species *Oscillochloris trichoides* has only been

368 reported to reduce acetylene, an alternate substrate of nitrogenase (Keppen, Baulina, &
369 Kondratieva, 1994; Keppen, Lebedeva, Troshina, & Rodionov, 1989). However, the F-Mc species
370 *Methanocaldococcus sp.* FS406 (lacking *nifN*), as well as an uncharacterized, anaerobic methane-
371 oxidizing archaeon not included in the present study, have been shown to assimilate isotopically
372 labeled nitrogen (Dekas et al., 2009; Mehta & Baross, 2006). The ability of early-branching F-Mc
373 nitrogenases to fix nitrogen in the absence of the full complement of *nifEN* genes may indicate an
374 alternative and/or simplified pathway for cofactor assembly not used for other Mo-, V-, and Fe-
375 nitrogenases.
376

377 **3.3 High statistical support for ancestral nitrogenase active-site residues**

378 We inferred ancestral sequences for each of the H-, D-, and K-subunits that constitute the
379 nitrogenase enzyme complex (Figure 1) across five phylogenetic topologies (Tree-1–5; Appendix
380 S2). Ancestral nitrogenase sequences were inferred for five well-supported internal nodes along a
381 phylogenetic transect between Mo- (highlighted in blue) and V-/Fe-nitrogenases (highlighted in
382 red) (Figure 2). The five targeted nodes are: AncA (ancestral to Anf and Vnf), AncB (ancestral to
383 AncA and Mb-Mc), AncC (ancestral to AncB and F-Mc), AncD (ancestral to AncC and Clfx), and
384 AncE (ancestral to Nif-I and Nif-II). Thus, AncA–D are nested, whereas AncE lies along a
385 divergent lineage toward Nif-I and Nif-II Mo-nitrogenases. For further analyses, we selected the
386 maximum likelihood ancestral sequence per target node from each of the five phylogenies (Tree-
387 1–5), totaling 25 sequences. Due to differences in root position, identical AncE nodes were not
388 present across all topologies and analogous nodes were instead selected (Appendix S2). Ancestral
389 sequences are hereafter labeled with the tree likelihood rank from which they were inferred (e.g.,

390 AncA from Tree-1 is labeled AncA-1). All tree and ancestral sequence information can be found
391 at http://phylobot.com/613282215/.
392
393 Mean site posterior probabilities for ancestral nitrogenase HDK sequences across all phylogenies
394 range between ~0.83 and 0.91, and for the highest-likelihood phylogeny (Tree-1), between ~0.84
395 and 0.90 (Appendix S4). Ancestral sequence support generally decreases with increasing
396 phylogenetic node age. For example, within the uncharacterized/V-/Fe-nitrogenase linage, AncA-
397 1 has the highest mean posterior probability (0.90 ± 0.18) and AncD-1 has the lowest mean
398 posterior probability (0.84 ± 0.22). Mean ancestral sequence probability for each node also does
399 not deviate by more than ~0.02 across each of the five phylogenetic topologies (Tree-1–5;
400 Appendix S4). These observations suggest that sequence support for ancestral nitrogenases is more
401 sensitive to ancestral node position than to topological differences between the analyzed trees.
402
403 In addition to surveying total ancestral HDK sequence support, we analyzed support for 30 active-
404 site residues, defined as those residing within 5 Å of any atom in either the FeMo-cofactor of the
405 *A. vinelandii* NifD protein (PDB 3U7Q (Spatzal et al., 2011)) or the FeV-cofactor of the *A.*
406 *vinelandii* VnfD protein (PDB 5N6Y (D. Sippel & Einsle, 2017)) (Figure 3). These active-site
407 residues are not contiguous but are instead scattered throughout the D-subunit sequence. Mean
408 posterior probabilities of ancestral active-site residues, which range between 0.92 to 0.98 across
409 all phylogenies, are consistently greater than those of entire reconstructed nitrogenase HDK
410 sequences (0.83–0.91) (Appendix S4). Of the 30 active-site residues, only five sites have, in one
411 or more ancestral sequences, plausible alternative reconstructions with posterior probabilities
412 >0.30: sites 59, 69, 358, 360, 425, 441 (site numbering both here and hereafter based on *A.*

413 *vinelandii* NifD). No ancestral sequences have more than three active-site residues with such
414 plausible alternative reconstructions. Ten active-site residues are conserved across all analyzed
415 extant nitrogenases: Val-70, Gln-191, His-195, Cys-275, Arg-277, Ser-278, Gly-356, Phe-381,
416 Gly-424, His-442. These conserved residues are thus reconstructed in all ancestral nitrogenases
417 unambiguously (site posterior probability = 1.00). Statistical support for ancestral active-site
418 residues (greater than 0.92) underpins subsequent analyses of ancestral active-site properties that
419 may inform inferences of nitrogenase metal dependence.
420

421 **3.4 Oldest ancestral nitrogenase active-site sequences resemble extant Mo-nitrogenases**

422 We analyzed sequence features of both ancestral and extant nitrogenases to identify those
423 correlated with metal dependence. In particular, we focused on 30 nitrogenase active-site residues
424 (Figure 3; see Section 3.3. for further details) for three reasons: (1) active-site residues are known
425 to affect catalytic efficiency and substrate specificity (Brigle et al., 1987; Christiansen, Cash,
426 Seefeldt, & Dean, 2000; Fixen et al., 2016; Kim, Newton, & Dean, 1995; Sarma et al., 2010; Yang,
427 Moure, Dean, & Seefeldt, 2012) and thus may be tuned to nitrogenase metal dependence (2) active-
428 site sequence features have previously been used, in part, to classify the metal dependence of extant
429 uncharacterized nitrogenases (McGlynn et al., 2012) (3) statistical support for ancestral active-site
430 residues are higher than the mean support across entire HDK sequences (see Section 3.3).
431

432 We first assessed the sensitivity of ancestral sequence variation to phylogenetic uncertainty and
433 ancestral statistical support. Overall, mean identities for ancestral sequences compared across
434 different nodes range from ~55 to 90%. Ancestral sequences inferred from the same node across
435 alternate phylogenies (Tree-1–5) have relatively high mean identities, ranging from ~93 to 95%

436 across the total HDK sequence and from ~96 to 100% within the active site. These high mean
437 identities suggest that topological differences among the alternate phylogenies used for ancestral
438 sequence inference do not contribute to a high degree of ancestral sequence variation. Identities
439 among sequences inferred from the same node also do not appear to be correlated with statistical
440 support. For example, though full AncA HDK sequences are reconstructed with the highest mean
441 statistical support (~0.89–0.91), they exhibit neither the lowest nor the highest mean identities as
442 compared with sequences inferred from other nodes (Appendix S4).
443
444 We next identified specific residues within the nitrogenase active site that are unique to particular
445 isozymes of known metal dependence in order to survey their occurrence in ancestral sequences.
446 We found that three active-site residues are unique to Mo-nitrogenases (Ala-65, Arg-96, Gln-440),
447 six are unique to V-nitrogenases (Leu-69, Trp-300, Thr-355, Pro-358, Pro-425, Val-427), five are
448 unique to Fe-nitrogenases (His-69, Pro-355, Lys-359, Pro-427, Ala-441), and six are unique to V-
449 and Fe-nitrogenases (Cys-65, Lys-96, Leu-360, Lys-380, Arg-426, Asn-440) (Appendix S5).
450 Though we are unaware of any experimental mutagenesis studies identifying a role for these
451 uniquely conserved residues, we hypothesize that these residues may contribute to the metal
452 specificity of the active-site environment. Surprisingly, most nitrogenase ancestors exhibit
453 comparable numbers of residues unique to either V-/Fe-nitrogenases or Mo-nitrogenases, and thus
454 their occurrence does not appear informative for inferring ancestral metal dependence (e.g.,
455 ancestral AncC sequences contain two residues unique to V-/Fe-nitrogenases and two residues
456 unique to Mo-nitrogenases). An exception is AncA sequences, which contain a preponderance of
457 active-site residues that are unique to extant V- and Fe-nitrogenases. One of the residues unique
458 to only V-nitrogenases, Thr-355, has recently been suggested to interact directly with a proposed

FeV-cofactor carbonate ligand not present in the FeMo-cofactor (D. Sippel & Einsle, 2017). This carbonate ligand lies within a secondary-structure loop that also contains Pro-358, unique to V-nitrogenases, and Leu-360, unique to V- and Fe-nitrogenases. These Thr-355, Pro-358, and Leu-360 residues are observed across all AncA sequences.

In addition to analyzing specific active-site residues, we compared the total active-site sequence composition of 25 ancestral nitrogenases (inferred across Tree-1–5) and all 284 extant nitrogenase sequences used for phylogenetic reconstruction. Our analysis shows clear distinctions between the active-site compositions of extant Mo- versus V-/Fe-nitrogenases (Figure 4a). Specifically, the mean identity between Vnf/Anf and Nif-I/Nif-II is ~50%, as compared with the mean identity between Vnf and Anf (~71%) and between Nif-I and Nif-II (~77%). The difference between Vnf/Anf and Nif-I/Nif-II sequences is not seen as distinctly when comparing total HDK sequences (Figure 4b), suggesting that the active-site sequence differences, in particular, may be better indicators of metal dependence than whole sequence differences due to greater catalytic tuning toward the cofactor.

Because we observed active-site sequence distinctions between extant Vnf/Anf and Nif-I/Nif-II nitrogenases, we compared active-site sequences of ancestral versus extant nitrogenases to provide clues regarding ancestral metal dependence. Nearly all ancestral nitrogenases, including those inferred for the oldest ancestral nodes, share greater active-site identity with Mo-nitrogenases than with V-/Fe-only-nitrogenases (Figure 4a). Mean active-site sequence identities between AncB–AncE and Anf/Vnf nitrogenases range between ~50 and 63%, whereas those between AncB–AncE and Nif-I/Nif-II nitrogenases range between ~69 and 87%. An exception is AncA (ancestral to Vnf

482 and Anf), which has higher mean identity to Anf/Vnf nitrogenases (~85%) than to Nif-I/Nif-II
483 nitrogenases (~51%). Because active-site sequence identity can reliably differentiate extant Mo-
484 from V-/Fe-nitrogenases, the resemblance of most ancestral active sites to those of Mo-
485 nitrogenases is suggestive of Mo-dependence. These results are consistent with previous sequence-
486 based analyses that found that basal F-Mc nitrogenase sequences most resemble those of Mo-
487 nitrogenases (McGlynn et al., 2012).
488
489 Finally, in order to incorporate the posterior probability distributions of inferred ancestral
490 nitrogenase sequences into our sequence-based interpretations, we built a probabilistic model from
491 the active-site amino acid probability distributions of 268 modern nitrogenases of known metal
492 dependence (i.e., excepting 16 Clfx, F-Mc, and Mb-Mc homologs). Amino acid probability
493 distributions were determined for active-site residues and used to train a support-vector classifier
494 (see Section 2.2 for details). Surprisingly, the classifier was able to achieve 100% accurate
495 classification on out-of-sample test data. However, principle-component analysis of active-site
496 residue composition suggests that Nif sequences are completely separable from Vnf and Anf
497 sequences using only one principle component, and Vnf and Anf sequences also separable using
498 only two dimensions (Figure S6-1). The trained classifier was used to classify uncharacterized
499 Clfx, F-Mc, and Mb-Mc sequences, as well as 25 maximum-likelihood ancestral nitrogenase
500 sequences. All uncharacterized and ancestral sequences were classified as Nif-like, with the
501 exception of all AncA ancestors, which were classified as Vnf-like (Table S6-1). Classification
502 support (reported as distance of each data sample from the support-vector hyperplane defining the
503 inferred class) for ancestral sequences is nearly identical across different ancestors inferred for the
504 same phylogenetic node (Table S6-1). There is little difference in classification support between

ancestors inferred from a MUSCLE alignment as from a MSAProbs alignment, with the exception between AncA-4 (inferred from MUSCLE alignment) and other AncA sequences (inferred from MSAProbs alignment). Support for AncA-4 classification is -0.0884 (for Nif), 2.1059 (for Vnf), and 0.9685 (for Anf), as opposed to mean support for other AncA sequences, 0.9774 ± 0.0023 (for Nif), 2.0626 ± 0.0056 (for Vnf), and -0.0455 ± 0.0046 (for Anf) (greater distance indicates greater support for the inferred classification). These probability-distribution-based classification results are consistent with the sequence-identity based comparisons, the latter of which similarly suggest higher identity between most uncharacterized/ancestral (AncB-E) sequences and extant Mo-nitrogenases than V-/Fe-nitrogenases. It is perhaps not surprising that the inclusion of the inferred ancestral probability distributions does not substantially change classifications of metal dependence, given that the support for active-site residues is high (>0.92), and that, of these, only five sites are reconstructed with plausible alternate states (alternate residues with probability >0.30) (See Section 3.3 for further details).

3.5 Active-site structural features are uninformative for inferring ancestral metal dependence

To investigate metal-specific features of ancestral nitrogenase structures, we generated homology models of both extant and ancestral nitrogenase D-subunits that house the active site (Figure 5a). First, we modeled 33 broadly sampled extant nitrogenase NifD, VnfD, and AnfD sequences to benchmark classifications of ancestral nitrogenase models. Second, we calculated structural models of 25 nitrogenase ancestors inferred by maximum likelihood and of 2,500 ancestors inferred by random Bayesian sampling of maximum likelihood site posterior probabilities (100 Bayesian samples per maximum likelihood ancestor). We generated ten model replicates per extant

528 sequence and maximum likelihood sequence, and one model per Bayesian-sampled sequence. All
529 structures were modeled with the FeMo-cofactor included (additional modeling runs were
530 executed with the FeV-cofactor included; Appendix S3). In total, 3,080 models were generated
531 with the FeMo-cofactor.
532
533 For each of the 3,080 extant and ancestral D-subunit nitrogenase models, we calculated the volume
534 of the active-site pocket (Figure 5b), a parameter previously used, in part, to classify the metal
535 dependence of extant uncharacterized nitrogenases (McGlynn et al., 2012). These pocket volume
536 values are plotted in Figure 6. Among modeled extant nitrogenases, mean pocket volumes are
537 1175.12 ± 51.93 $Å^3$ for Mo-nitrogenases, 1121.86 ± 36.36 $Å^3$ for V-nitrogenases, and 963.39 ±
538 75.80 $Å^3$ for Fe-nitrogenases.
539
540 We observe less difference between mean pocket volumes of extant V-nitrogenases (1121.86 ±
541 36.36 $Å^3$) and Nif-II Mo-nitrogenases (1141.13 ± 46.30 $Å^3$) than between Nif-I (1209.11 ± 30.79
542 $Å^3$) and Nif-II Mo-nitrogenases. A statistical nonparametric test of volume median differences also
543 suggests greater similarity between V- and Nif-II Mo-nitrogenases than between Nif-I and Nif-II
544 nitrogenases (Appendix S3). All V-nitrogenase and Nif-II Mo-nitrogenase volume values range
545 between 1002.13 and 1220.63 $Å^3$, which problematically overlap with the volume ranges of AncB
546 (917.38–1256.25 $Å^3$), AncC (933.75–1269.50 $Å^3$), AncD (968.50–1272.75 $Å^3$), and AncE
547 (966.75–1225.25 $Å^3$) ancestral models, as well as ranges for uncharacterized Clfx (988.25–
548 1218.13 $Å^3$) and F-Mc (908.00–1277.75 $Å^3$) structures. The volume range of AncA models
549 (919.00-1191.13 $Å^3$) lies between the ranges of both V- (1020.75–1197.63 $Å^3$) and Fe-
550 nitrogenases (821.38–1177.00 $Å^3$). These same overall patterns are observed when comparing only

maximum likelihood ancestral models or Bayesian-sampled ancestral models, as well as alternate modeling runs with the FeV-cofactor (Appendix S3). Greater similarities between V- and Nif-II Mo-nitrogenases than between Nif-I and Nif-II Mo-nitrogenases suggests that, contrary to findings from previous analyses (McGlynn et al., 2012), modeled pocket volume is not informative for classifying modern or ancestral nitrogenases of unknown metal dependence. At the very least, the overlap in volume ranges between ancestors and extant isozymes of varying metal dependence in our analyses preclude the unambiguous classification of ancestral metal dependence by these structural features.

4 DISCUSSION

Nitrogenase mediates the reduction of N_2 to NH_3, a key step in nitrogen fixation (Hoffman et al., 2014; Seefeldt, Hoffman, & Dean, 2009). The metal dependence of nitrogenase, which impacts both catalytic properties (Eady, 1996; Harris, Yang, et al., 2018; Hu et al., 2011; Lee et al., 2018; Rebelein et al., 2018; Zheng et al., 2018) and ecological distribution (McRose, Zhang, Kraepiel, & Morel, 2017; Zhang et al., 2016), suggests a potential role for geochemical constraints on its evolution (Anbar & Knoll, 2002; Boyd, Hamilton, et al., 2011; Canfield et al., 2010; Raymond et al., 2004). Thus, understanding ancestral nitrogenase metal dependence can help resolve the early history of biological nitrogen fixation, and, in a broader sense, the impact that ancient metal availabilities have had on the evolution of biologically essential metabolisms over Earth history (Anbar & Knoll, 2002; Moore, Jelen, Giovannelli, Raanan, & Falkowski, 2017). Previous phylogenetic work has established that Mo-, V-, and Fe-nitrogenases, though genetically distinct, are homologous (Boyd, Anbar, et al., 2011; Boyd, Hamilton, et al., 2011; Raymond et al., 2004). Most recent phylogenetic analyses also indicate that V- and Fe-nitrogenases are derived from Mo-

574 nitrogenases, the latter having originated subsequent to the gene duplication event that produced
575 *nifE* and *nifN* (Boyd, Anbar, et al., 2011). However, the precise trajectory of metal-binding
576 evolution in the nitrogenase family is not completely known, and discrepancies between current
577 phylogenetics-based models and the geochemical record of nitrogen fixation (e.g., (Stueken et al.,
578 2015)) remain.
579
580 We used phylogenetic reconstruction, ancestral sequence inference, and structural homology
581 modeling to explore outstanding questions of early nitrogenase evolution and metal dependence.
582 Though the tree reconstructions presented here are largely in congruence with previous
583 phylogenetic analyses, certain topological differences, particularly with regard to basal
584 uncharacterized homologs lacking one or more associated *nifEN* genes, suggest important
585 deviations from previous narratives of early metal dependence (Boyd, Anbar, et al., 2011; Boyd,
586 Hamilton, et al., 2011; Boyd & Peters, 2013; Raymond et al., 2004). The reconstruction of
587 ancestral nitrogenase sequences *in silico* provides the means to directly infer ancient metal-binding
588 properties from molecular information.
589
590 **4.1 Inferred metal dependence of ancestral nitrogenases**
591 The nitrogenase active site is known, in addition to the metal content of the cofactor, to contribute
592 to the variable catalytic properties of different nitrogenase isozymes (Brigle et al., 1987;
593 Christiansen et al., 2000; Fixen et al., 2016; Kim et al., 1995; Sarma et al., 2010; Yang et al., 2012).
594 We therefore explored active-site features of ancestral nitrogenases for correlations with metal
595 dependence.

We find that modeled active-site structural features are not informative for the inference of ancestral metal dependence. Pocket volumes of modeled extant nitrogenases do not appear to be strongly correlated with metal cofactor content. Problematically, the predicted volume ranges of oldest ancestors (as well as those of uncharacterized nitrogenases) overlap with both V- and Mo-nitrogenases (Figure 6). We acknowledge that these homology modeling results may not precisely reflect true biological differences (for which a comprehensive analysis is not yet possible due the limited availability of V- and Fe-nitrogenase structures). This is illustrated by the difference between pocket volumes of published structures and those of homology models (e.g., *A. vinelandii* PDB 3U7Q NifD structure pocket volume ≈ 994.750 and mean *A. vinelandii* NifD modeled structure pocket volume ≈ 1186.412). However, it is not surprising that active-site pocket volume does not predict metal dependence, given the probable similar sizes and structures of the FeMo-, FeV-, and FeFe-cofactors (Eady, 1996; Krahn et al., 2002; Daniel Sippel et al., 2018; Spatzal et al., 2011). These findings contrast with a previous study by McGlynn and coworkers (2012) in which pocket volume comparisons were used, in part, to classify the Mo-dependence of uncharacterized nitrogenases. In this previous study, the volume means for Mo-nitrogenases were sufficiently distinct from V- and Fe-nitrogenases as to provide unambiguous classification of Mo-dependence. Our analyses differ from this previous study in several ways: (1) we incorporated V-nitrogenase structural templates (D. Sippel & Einsle, 2017; Daniel Sippel et al., 2018) for homology modeling, not available at the time of the previous study, (2) we modeled 20 extant sequences of known metal dependence rather than 12, (3) we used consistent, explicit parameters for both homology modeling and pocket volume calculation across all sequences , (4) we included modeling replicates for pocket volume calculation. In summary, our expanded modeling analyses

618 reveals a reduced efficacy of the pocket volume parameter for inference of ancestral (and
619 uncharacterized) nitrogenase metal dependence.
620
621 Unlike structure-based analyses, we find that active-site sequence features do reliably differentiate
622 Mo-nitrogenases and V-/Fe-nitrogenases, as was also previously noted by McGlynn and coauthors
623 (2012). We find that both sequence-based and probability-distribution-based analyses are thus
624 useful for classifying nitrogenases of unknown metal dependence, and are robust to the underlying
625 uncertainty of ancestral sequence inference (Section 3.4; Appendix S6). Regarding AncA
626 (ancestral to V- and Fe-nitrogenases), we identified specific active-site residues that have
627 previously been suggested to interact with a proposed carbonate ligand not present in the FeMo-
628 cofactor. These residues form a ^{355}TGGPRL^{360} loop conserved only among V-nitrogenases and
629 homologous to ^{355}IGGLRP^{360} in *A. vinelandii* NifD (D. Sippel & Einsle, 2017). This substitution
630 of Thr-355 for Ile-355, as well as the exchange of Leu and Pro positions may permit the inclusion
631 of the FeV-cofactor carbonate ligand by VnfD that is not possible by the NifD protein (D. Sippel
632 & Einsle, 2017). Thr-355 and Pro-358 are unique to V-nitrogenases, and Leu-360 is unique to V-
633 and Fe-nitrogenases. All AncA sequences conserve the ^{355}TGGPRL^{360} residue loop capable of
634 accommodating the FeV-cofactor. Furthermore, AncA sequences generally exhibit greater
635 numbers of residues unique to V-nitrogenases than those unique to Fe-nitrogenases, the mean
636 identity of AncA active-site sequences is highest for V-nitrogenases (Figure 4a), and the classifier
637 analysis based on active-site amino acid probability distributions suggests that AncA sequences
638 most resemble modern V-nitrogenases (Section 3.4; Appendix S6). Together, these observations
639 suggest that AncA is specific toward the VFe-cofactor.
640

641 Ancestral and extant sequence comparisons indicate that the active-sites of oldest nitrogenase
642 ancestors (AncB–AncE) resemble those of extant Mo-nitrogenases more than V- or Fe-
643 nitrogenases (Figure 4a; Appendix S6). This observation is particularly significant given that these
644 same patterns are not observed across the total HDK sequence (Figure 4b). Specifically, this
645 discrepancy supports the notion that the nitrogenase active-site has adapted to the catalytic
646 properties of each metal cofactor over its evolutionary history (Harris, Yang, et al., 2018), and that
647 this adaptation has manifested in active-site sequence differences that stand apart from baseline
648 phylogenetic distance. Though we are not able to identify specific residues or motifs that may
649 functionally be attributed to metal dependence (as with AncA), the resemblance of the early
650 ancestral nitrogenase active site to those of Mo-nitrogenases is highly suggestive of an ancient role
651 for an active-site cluster resembling the FeMo-cofactor.
652

653 **4.2 A proposed model for the evolution of nitrogenase metal dependence over geologic time**

654 Despite the lack of information provided by structural analyses, the active-site sequence features
655 of oldest ancestral sequences (i.e., AncB–E) support the inference that these early nitrogenase
656 incorporated the FeMo-cofactor or similar cluster. However, the observation that ancestral AncC
657 and AncD active sites in particular most resemble those of extant Mo-nitrogenases is at odds with
658 the phylogenetic distribution of *nifE* and *nifN* genes, which suggest that early-branching
659 uncharacterized Clfx and F-Mc taxa (for which AncC and AncD are ancestral) do not possess the
660 full biosynthetic complement in the canonical FeMo-cofactor assembly pathway.
661

662 The placement of Clfx and F-Mc clades in our analyses differs from previous phylogenetic
663 reconstructions. For example, the phylogenetic tree presented by Boyd and coauthors nests Clfx

664 and F-Mc clades within Mo-nitrogenases, which notably branch more recently than V-, Fe-, and
665 Mb-Mc Mo-nitrogenases (Boyd, Hamilton, et al., 2011). Such a topology would be consistent with
666 the secondary loss of *nifEN* genes in Clfx and F-Mc clades, and the absence of such genes would
667 not likely represent an ancestral state for the cofactor biosynthetic pathway. By contrast, we find
668 that Clfx and F-Mc clades are early-branching, and that the presence of *nifE* and *nifN* genes
669 decreases stepwise with divergence age within the uncharacterized/V-/Fe-nitrogenase lineage:
670 Mb-Mc taxa, most recently branched, have both *nifEN,* most F-Mc taxa only have *nifE*, and Clfx
671 taxa, earliest branched, have neither (*vnf* and *anf* gene clusters, resulting only from the duplication
672 of *nif* structural genes (Boyd, Hamilton, et al., 2011), typically lack dedicated scaffolding genes
673 and instead co-opt *nifEN;* (Hamilton et al., 2011; Mus et al., 2018)). One may thus parsimoniously
674 conclude that uncharacterized AncC–D ancestors similarly lacked one or more *nifEN* genes. A
675 more recent, previously published topology is more similar to the tree presented here, though
676 lacking in Clfx sequences and thus missing relevant information regarding the lack of associated
677 biosynthetic components for early-diverged homologs (Boyd & Peters, 2013). It is possible that
678 the larger and more broadly sampled sequence dataset used here has refined the placement of these
679 uncharacterized clades, which is supported by our analyses with an expanded outgroup that
680 maintains the positions of Clfx and F-Mc sequences (Appendix S2).
681
682 At least two possible models might resolve the discrepancy between inferred early Mo-dependence
683 and suggested lack of cofactor biosynthesis genes: (1) The last common nitrogenase ancestor
684 represented in our tree was in fact hosted by an organism that possessed *nifEN* genes and was able
685 to synthesize the FeMo-cofactor via the full, canonical biosynthetic pathway. Further, one or more
686 *nifEN* genes were lost in both Clfx and F-Mc taxa, but were retained in all ancestors (AncA–E) as

687 well as in Nif-I, Nif-II, and Mb-Mc taxa. The resemblance of oldest ancestral active sites (AncB–
688 E) to those of Mo-nitrogenases indicate ancient FeMo-cofactor dependence, and the same
689 resemblance of Clfx and F-Mc nitrogenases to Mo-nitrogenases may be inherited from these
690 ancestors. (2) The last common nitrogenase ancestor was hosted by an organism that did not
691 possess *nifEN* genes, and achieved cofactor biosynthesis by alternative and/or simplified means.
692 Such an early-evolved and perhaps inefficient pathway for cofactor synthesis may still be present
693 in F-Mc taxa, and possibly Clfx taxa as well. *nifEN* genes then evolved through the
694 uncharacterized/V-/Fe-nitrogenase lineage via gene duplication of *nifDK* and subsequent
695 subfunctionalization (Boyd, Anbar, et al., 2011*)*, and was selected for once the canonical FeMo-
696 cofactor pathway was established prior to the AncB nitrogenase at the earliest, but at least in Mb-
697 Mc nitrogenases (*nifEN* genes may have been transferred between ancestors of Mb-Mc, Nif-I, and
698 Nif-II taxa).
699
700 We prefer the second model for several reasons:
701
702 (1) It is evolutionarily unlikely for a portion of the FeMo-cofactor biosynthetic pathway to be lost
703 in Clfx and F-Mc taxa. To our knowledge, no other Mo-nitrogenases are known to lack associated
704 *nifEN* genes, suggesting that the canonical biosynthetic pathway for the FeMo-cofactor is
705 selectively advantageous (*anf* and *vnf* gene clusters typically lack dedicated scaffolding genes and
706 co-opt *nifEN*; (Hamilton et al., 2011; Mus et al., 2018)). Rather, it is more parsimonious to infer
707 that these early-branching uncharacterized clades diverged prior to the *nifDK* duplication that
708 produced *nifEN*, as is evidenced by the nesting of *nifEN* genes within *nifDK* clades in previous
709 phylogenetic reconstructions (Boyd, Anbar, et al., 2011).

710

711 (2) It has previously been proposed that early nitrogenases may have been capable of reducing
712 nitrogen prior to the development of the canonical FeMo-cofactor biosynthetic pathway (Boyd,
713 Hamilton, et al., 2011; Boyd & Peters, 2013; Mus, Colman, Peters, & Boyd, 2019; Soboh, Boyd,
714 Zhao, Peters, & Rubio, 2010). Further, an ancient Mo-independent nitrogenase may have been
715 capable of — albeit inefficiently — reducing nitrogen by a cofactor resembling the Fe-S-C cluster
716 assembled by NifB, which constitutes the biosynthetic precursor to the FeMo-cofactor (Boyd &
717 Peters, 2013; Mus et al., 2019; Soboh et al., 2010). Though our sequence analyses cannot fully
718 assess ancestral nitrogenase dependence for a NifB-cofactor, it is likely that the NifB-cofactor
719 resembles the structure and composition of the FeFe-cofactor, excepting homocitrate (Corbett et
720 al., 2006; Guo et al., 2016; Harris, Lukoyanov, et al., 2018). However, the greater similarity of
721 AncC–D active sites to those of Mo-nitrogenases than Fe-nitrogenases likely suggests ancestral
722 dependence on a cofactor incorporating Mo rather than only Fe, as would be the case for the NifB-
723 cofactor. Instead, it is possible that, lacking *nifEN*, an alternative and/or simplified pathway for
724 FeMo- or similar cofactor synthesis may have acted as a transition state between such an Mo-
725 independent stage (i.e., binding a cluster resembling the NifB-cofactor; (Boyd & Peters, 2013;
726 Mus et al., 2019)) and the development of the canonical FeMo-cofactor biosynthetic pathway. It
727 is thus reasonable to speculate that this transition to Mo-usage may be exhibited by AncC–D
728 ancestors.

729

730 (3) Mo-nitrogenases are far more efficient at reducing nitrogen than other isozymes (Eady, 1996;
731 Harris et al., 2019; Harris, Yang, et al., 2018), and the majority of all extant nitrogenases are Mo-
732 dependent across both anoxic and oxic environments (Boyd et al., 2015; Mus et al., 2019;

Raymond et al., 2004). Even those organisms that have additional V- or Fe-nitrogenases still retain and preferentially express Mo-nitrogenases (Boyd, Anbar, et al., 2011; Boyd, Hamilton, et al., 2011; Dos Santos et al., 2012; Hamilton et al., 2011; Raymond et al., 2004). The presence of a pathway, albeit not fully established, for FeMo-cofactor or similar cofactor synthesis would be selectively advantageous even if Mo was transiently available prior to ~2.3–2.5 Ga Earth surface oxygenation (Anbar et al., 2007; Anbar & Knoll, 2002; Canfield et al., 2010; Lyons et al., 2014; Raymond et al., 2004). In fact, previous studies suggest an early generalist behavior for ancient nitrogenases with regard to their metal usage (Raymond et al., 2004). Such limited and possibly non-specific Mo-incorporation into an active-site cluster early in nitrogenase history could pave the way for increased Mo-selectivity over time via the addition of NifEN biosynthetic components, building from an earlier Mo-independent state incorporating something resembling the modern NifB-cofactor (Boyd & Peters, 2013; Mus et al., 2019). An early nitrogenase that was effective at incorporating Mo would have favored the selection of other ancillary genes that later gave rise to increased Mo-specificity.

(4) ~3.2-Ga nitrogen isotopic signatures have been interpreted to reflect ancient Mo-dependent nitrogenase activity, due to similarities in isotopic fractionation by extant Mo-nitrogenases (Stueken et al., 2015). However, this interpretation conflicts with age estimates of both *nifEN* (inferred to also reflect the age of the FeMo-cofactor) (Boyd, Anbar, et al., 2011) and of earliest marine Mo availability (Anbar et al., 2007; Lyons et al., 2014). It is possible that an alternative pathway for FeMo- of similar cofactor synthesis in early nitrogenases may provide an explanation for these so-far unresolved, early isotopic signatures.

The proposed model of an ancestral, alternative pathway for Mo-cofactor assembly, mapped to our phylogenetic reconstruction and inferred ancestors, is illustrated in Figure 7. In the first stage, represented by AncD and extant Clfx homologs, ancient enzymes use an alternative and/or simplified pathway to incorporate an Mo-containing cluster, likely resembling or possibly identical to the canonical FeMo-cofactor in the absence of NifEN. So as to remain agnostic toward the particular cofactor structure, we term this a "proto-Mo-cofactor" (Figure 7a). Such enzymes may not be capable of reducing nitrogen, as this has not been demonstrated definitively for Clfx homologs. In the absence of both NifEN, cofactor assembly may be inefficient and subject to transient marine Mo availabilities prior to ~2.3 – 2.5 Ga Earth surface oxygenation, as evidenced by geochemical data (Anbar et al., 2007). In the second stage, represented by AncC and extant F-Mc nitrogenases, gene duplication of NifD forms NifE, which may be able to form a homodimer or homotetramer for the scaffolding of the FeMo-cofactor (Figure 7b). In the third stage, represented by AncB, AncE, and extant Mb-Mc nitrogenases, full gene duplication of NifDK additionally forms NifN (Figure 7c). Together with NifB, NifE and NifN complete the canonical FeMo-cofactor biosynthetic pathway, resulting in more refined nitrogenase Mo-dependence. In the fourth stage, gene duplication of *nif* results in, first, *vnf*, followed by *anf* (Figure 7d). These V- and Fe-dependent enzymes rely on NifBEN biosynthetic components (though some V-nitrogenases, as in *A. vinelandii*, rely on more recently evolved VnfEN proteins that derive from a duplication of NifEN; (Boyd, Anbar, et al., 2011)). Though the greater efficiency of Mo-nitrogenase results in widespread diversification, V- and Fe-nitrogenases provide selective advantage in microenvironments deficient in Mo. This model, built from the phylogenetic and ancestral sequence inferences provided here, as well as from decades of previous geobiological investigations of nitrogenase evolution, helps resolve outstanding questions regarding ancient

779 metal dependence, and, importantly, provides testable hypotheses for future investigations. Such
780 investigations may, for example, seek to clarify the nitrogen capability and metal dependence of
781 uncharacterized Clfx and F-Mc nitrogenases, as well as experimentally resurrect and characterize
782 ancestral nitrogenases in the laboratory.
783

784 **CONCLUSION**

785 We reconstructed the phylogenetic history of nitrogenase proteins, as well as inferred and
786 modelled ancestral nitrogenase sequences, in order to explore the evolutionary trajectory of
787 nitrogenase metal dependence. We find that, whereas modeled structural features of ancestral
788 nitrogenases do not offer conclusive indications of ancient metal usage, active-site sequence
789 features of oldest ancestors most resemble those of extant Mo-nitrogenases. The absence of
790 associated cofactor biosynthesis proteins that function within the canonical FeMo-cofactor
791 biosynthetic pathway in several early-branching, uncharacterized homologs evidences a possible
792 alternative and/or simplified pathway for Mo-cofactor assembly. We speculate that this alternative
793 pathway may today be present in extant uncharacterized taxa, and we propose a model wherein
794 canonical Mo-cofactor assembly evolved via the stepwise introduction of FeMo-cofactor
795 biosynthetic components following the divergence of more basal uncharacterized lineages. V-
796 nitrogenases subsequently diversified, followed by Fe-nitrogenases, in agreement with previous
797 phylogenetic inferences that Mo-dependence evolved first (Boyd, Hamilton, et al., 2011). This
798 model helps to reconcile phylogenetic and geobiological explanations of nitrogenase evolution
799 (Anbar & Knoll, 2002; Boyd, Anbar, et al., 2011; Boyd, Hamilton, et al., 2011). Future molecular
800 paleobiology studies, particularly those that integrate experimental assessments of laboratory-

resurrected ancestral nitrogenases, may continue to refine our understanding of nitrogenase and environmental coevolution.

ACKNOWLEDGEMENTS

This work was supported by a NASA Astrobiology Postdoctoral Fellowship (Garcia), the Harvard Origins of Life Initiative (Kacar, McShea), the National Science Foundation (Kacar, #1724090), the John Templeton Foundation (Kacar, #61239), a NASA Early Career Fellowship (Kacar, 80NSSC19K1617), and the University of Arizona Foundation Small Grants Program (Kacar, Garcia). We thank the Kacar Lab, Scott Edwards, Jennifer Glass, Andrew Knoll, Sebastian Kopf, Lance Seefeldt, and Xinning Zhang for invaluable discussions, as well as three anonymous reviewers for constructive feedback.

FIGURES

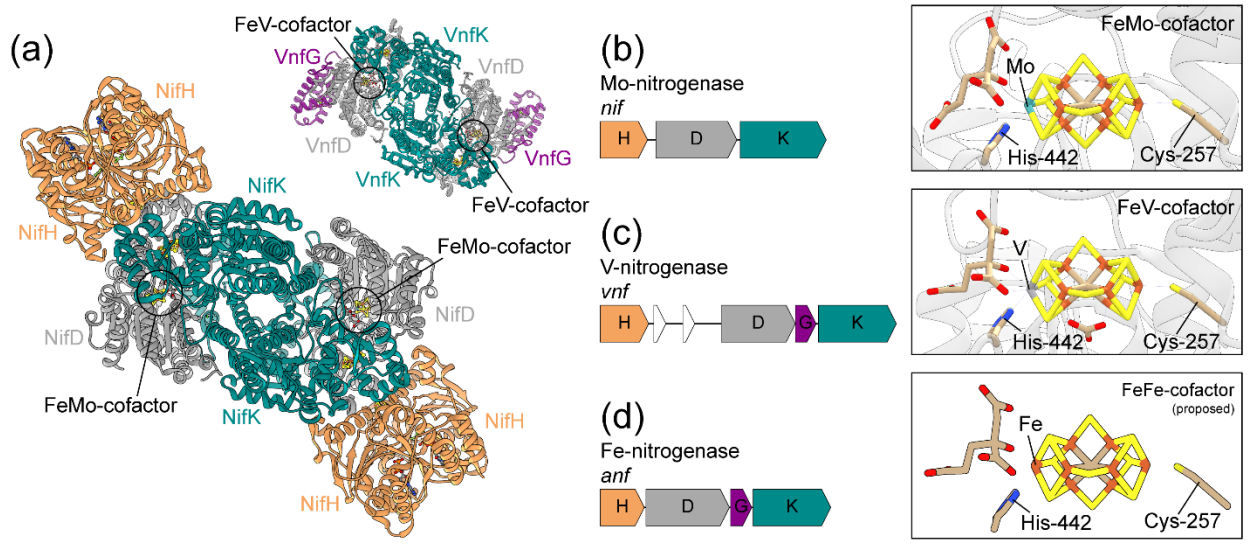
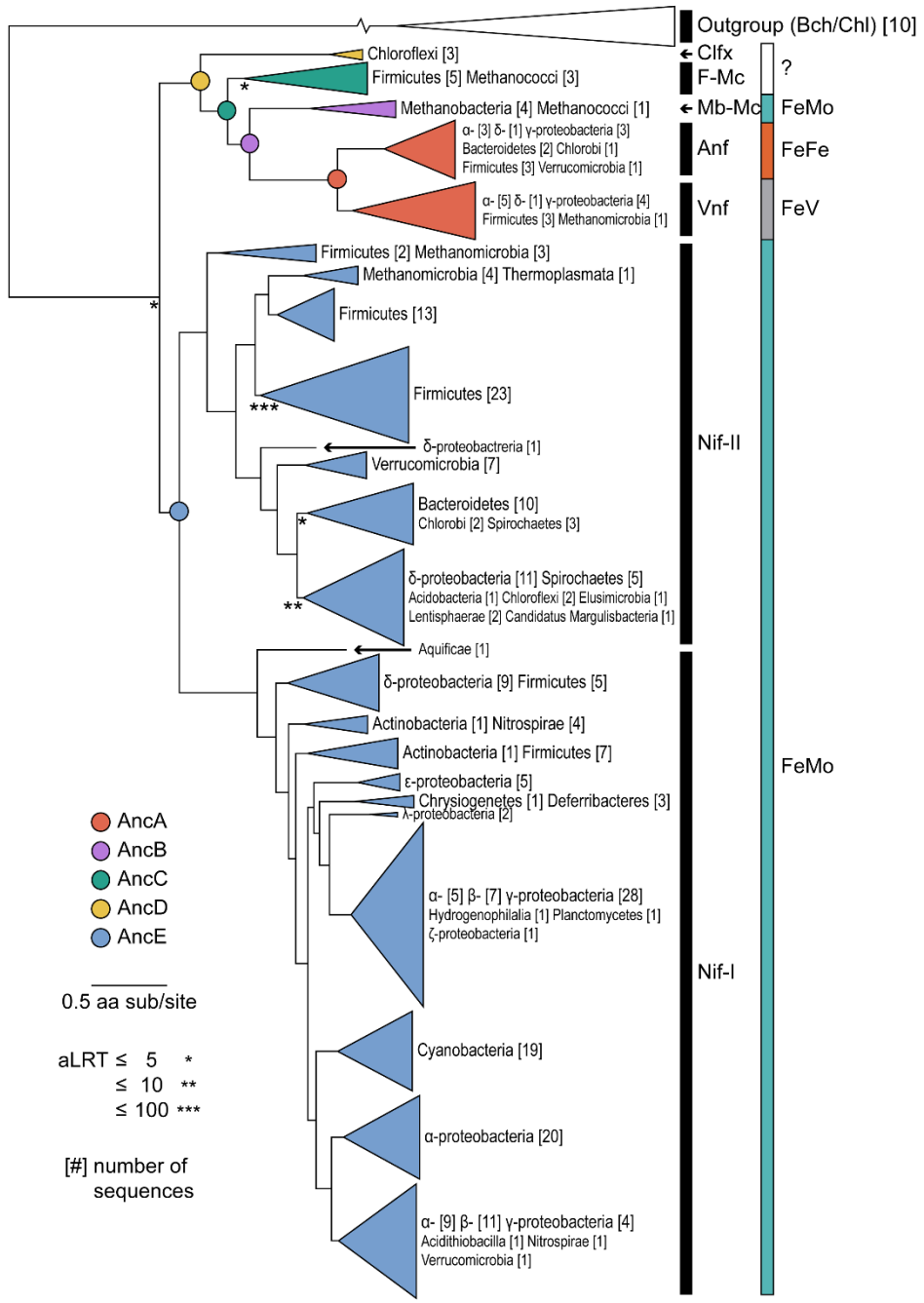


Figure 1. Structure and genetics of the three nitrogenase isozymes. (a) Structure of the *A. vinelandii* Mo-nitrogenase enzyme complex (NifHDK; PDB 1M34 (Schmid et al., 2002)) and V-nitrogenase VFe-protein component (VnfDGK; PDB 5N6Y (D. Sippel & Einsle, 2017)) (Fe-nitrogenase structure not previously published). The active-site FeMo-cofactor of Mo-nitrogenase and FeV-cofactor of V-nitrogenase are circled. (b-d) Catalytic genes and cofactor structures of *A. vinelandii* Mo-nitrogenase (b; PDB 3U7Q (Spatzal et al., 2011)), V-nitrogenase (c; PDB 5N6Y (D. Sippel & Einsle, 2017)), and Fe-nitrogenase (d; proposed structure (Harris, Lukoyanov, et al., 2018)). Residue numbering from aligned *A. vinelandii* NifD. Cofactor atom coloring is as follows: C, tan; Fe, rust; Mo, cyan; N, blue; O, red; S, yellow.



826

827 Figure 2. Maximum likelihood phylogeny of concatenated Nif/Anf/VnfHDK nitrogenase and
828 Bch/ChlLNB outgroup protein sequences (Tree-1; see Table 1). Ancestral nodes analyzed in this
829 study are labeled AncA–AncE. Known active-site cofactor metal content is listed on the right.
830 Branch support is derived from the approximate likelihood ratio test (aLRT). Branch length scale
831 is in units of amino acid substitutions per site. Outgroup branch break used to conserve space; true

branch length = 5.578 substitutions per site. Phylogeny coloring is as follows: Clfx, yellow; F-Mc, green; Mb-Mc, purple; Anf/Vnf, red; Nif-I/-II, blue.

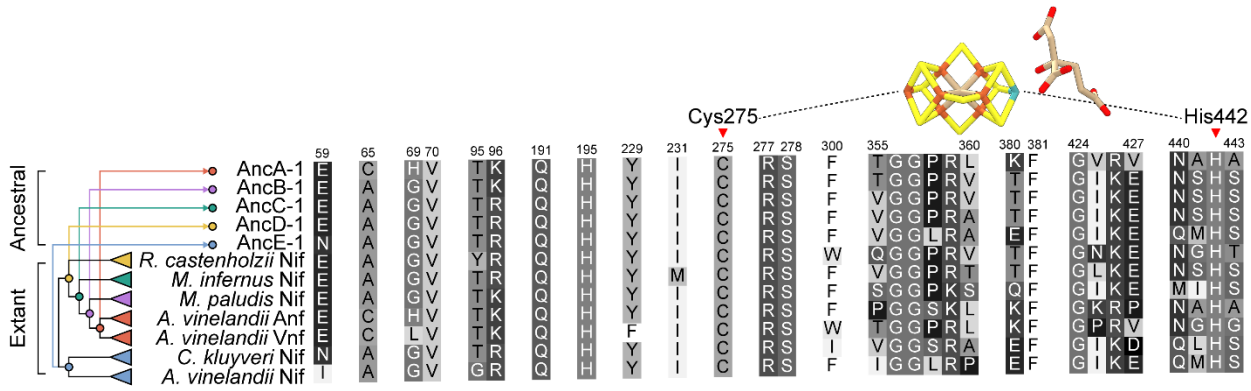
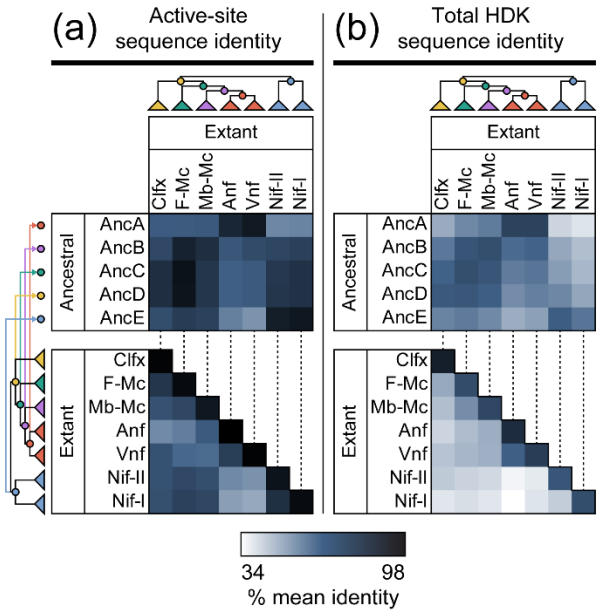


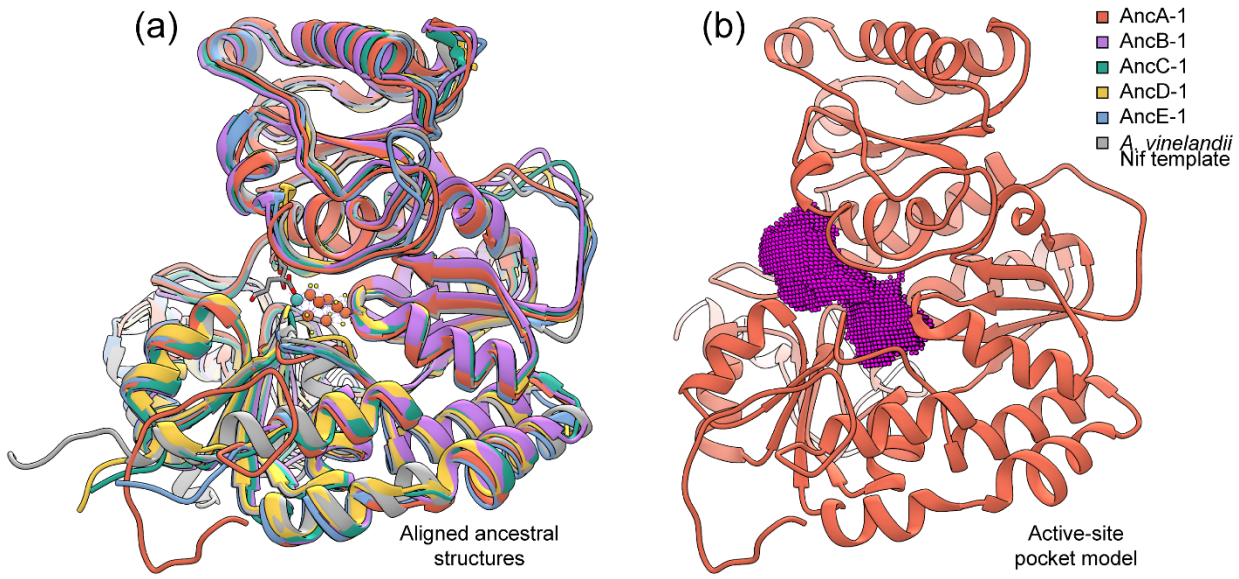
Figure 3. Active-site protein environment of representative ancestral and extant nitrogenases. All residues located within 5 Å of any atom in either the FeMo-cofactor of the *A. vinelandii* NifD protein (PDB 3U7Q (Spatzal et al., 2011)) or the FeV-cofactor of the *A. vinelandii* VnfD protein (5N6Y (D. Sippel & Einsle, 2017)). Residue numbering from aligned *A. vinelandii* NifD. Cys-275 and His-442 residues that coordinate the cofactor are indicated by red arrows. Phylogeny coloring is as follows: Clfx, yellow; F-Mc, green; Mb-Mc, purple; Anf/Vnf, red; Nif-I/-II, blue.



843

844 Figure 4. Active-site and full HDK sequence comparisons between extant and ancestral
845 nitrogenases. (a) Active-site sequence identities of ancestral and extant nitrogenases. Active-site
846 residues include 30 amino acids positioned within 5 Å of the active-site cofactor. (b) Total HDK
847 sequence identities of ancestral and extant nitrogenases. Percentage identity values have been
848 averaged within each field of comparison. All 25 maximum likelihood ancestors and 284 extant
849 sequences included in this study were used for sequence identity calculation. Phylogeny coloring
850 is as follows: Clfx, yellow; F-Mc, green; Mb-Mc, purple; Anf/Vnf, red; Nif-I/-II, blue.

851



852

853 Figure 5. Structural and active-site pocket modelling of ancestral nitrogenases (a) Modeled D-
854 subunit protein structures of ancestral nitrogenases inferred from the highest-likelihood phylogeny
855 (Tree-1; Figure 2) aligned to an *A. vinelandii* Nif structural template (PDB 3U7Q (Spatzal et al.,
856 2011)). (b) Example of a modeled active-site pocket for ancestral nitrogenase AncA-1. The 0.5-
857 Å-resolution point field generated for pocket volume calculation is shown in pink.

858

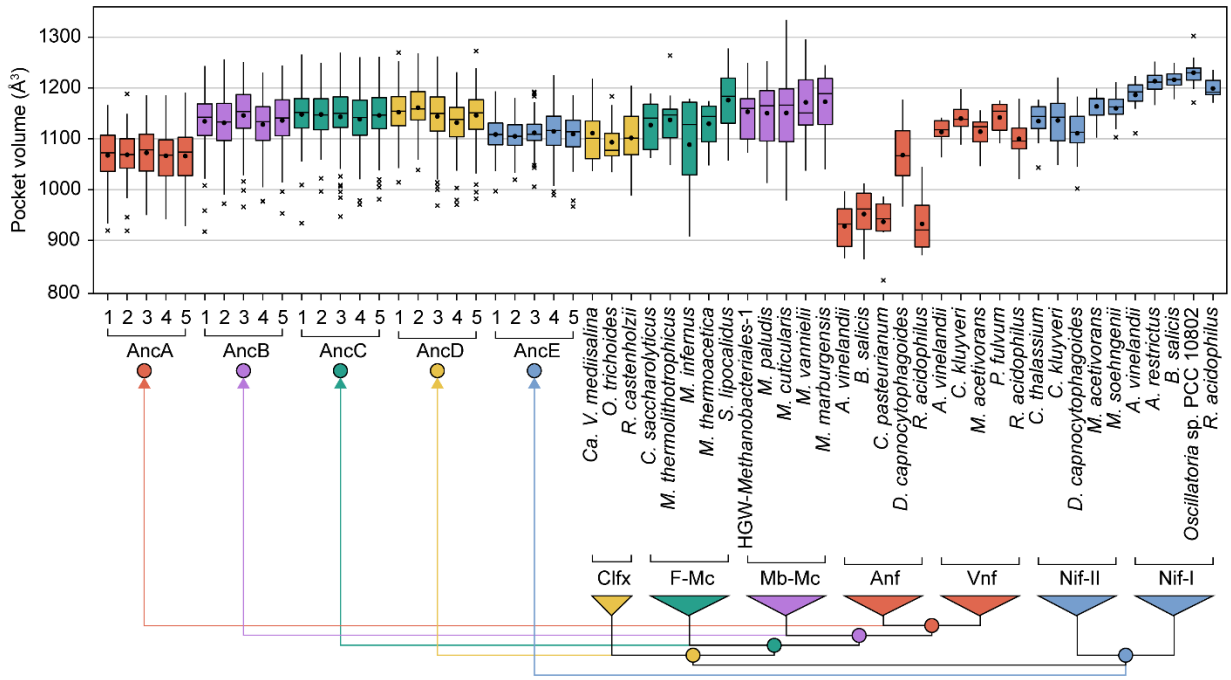
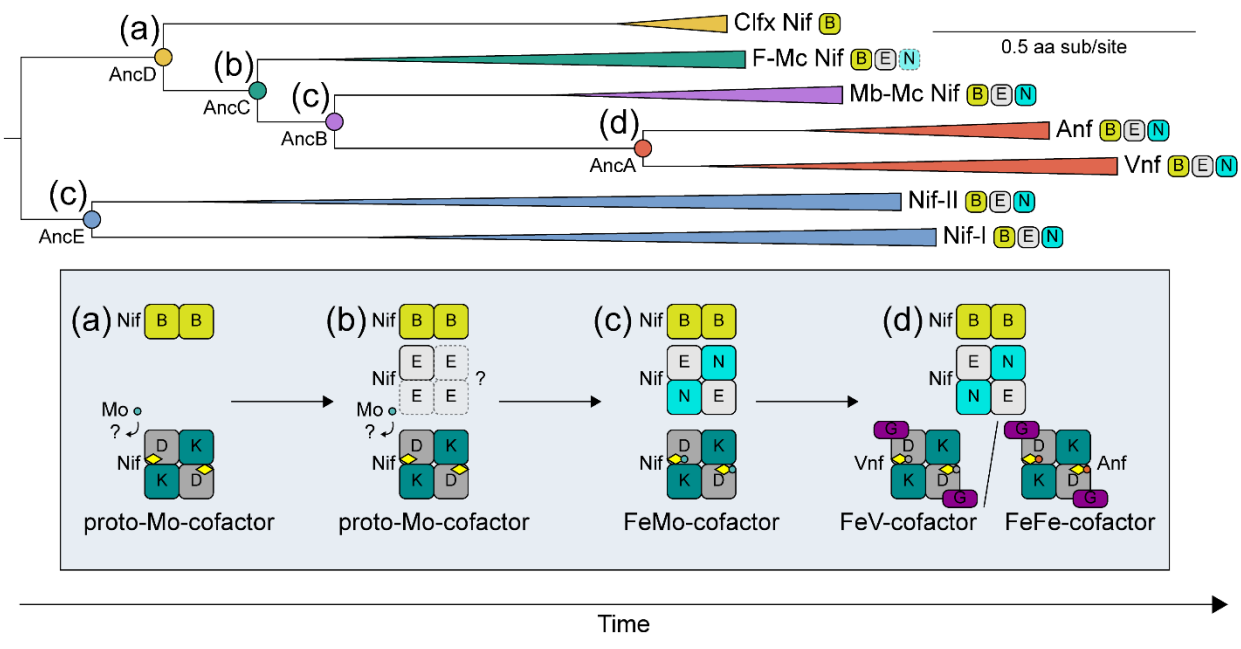


Figure 6. Extant and ancestral nitrogenase active-site pocket volumes. Pocket volumes calculated for ancestral and representative extant nitrogenase D-subunit structures modeled with the FeMo-cofactor. Each ancestral plot contains 110 volume calculations (ten model replicates per maximum likelihood sequence plus one model for each of 100 Bayesian-sampled sequences) and each extant plot contains 10 volume calculations (10 model replicates per extant sequence). Median values are indicated by bars, mean values by points, the range (excluding outliers) by whiskers, and outliers by crosses. Phylogeny coloring is as follows: Clfx, yellow; F-Mc, green; Mb-Mc, purple; Anf/Vnf, red; Nif-I/-II, blue.



869

870 Figure 7. Proposed model for the evolution of nitrogenase metal specificity from an ancestral,
871 alternative and/or simplified pathway for Mo-cofactor incorporation. Full description of stages
872 (a)–(d) in the evolution of nitrogenase metal dependence is provided in the main text (Section 4.2).
873 Possible alternative pathway for nitrogenase Mo-cofactor synthesis indicated in stages (a)–(b).
874 Phase (a)–(d) are also mapped to analyzed ancestors within the nitrogenase phylogeny. Presence
875 of NifB, NifE, and NifN in represented taxa indicated next to each clade of the phylogeny (F-Mc
876 clade has only one species that harbors NifN).

877

878

879 **REFERENCES**

880 Aadland, K., Pugh, C., & Kolaczkowski, B. (2019). High-Throughput Reconstruction of
881 Ancestral Protein Sequence, Structure, and Molecular Function. *Methods Mol Biol, 1851*,
882 135-170. doi:10.1007/978-1-4939-8736-8_8
883 Allen, R. M., Chatterjee, R., Ludden, P. W., & Shah, V. K. (1995). Incorporation of iron and
884 sulfur from NifB cofactor into the iron-molybdenum cofactor of dinitrogenase. *J Biol*
885 *Chem, 270*(45), 26890-26896. doi:10.1074/jbc.270.45.26890

Anbar, A. D., Duan, Y., Lyons, T. W., Arnold, G. L., Kendall, B., Creaser, R. A., . . . Buick, R. (2007). A whiff of oxygen before the great oxidation event? *Science, 317*(5846), 1903-1906. doi:10.1126/science.1140325

Anbar, A. D., & Knoll, A. H. (2002). Proterozoic ocean chemistry and evolution: a bioinorganic bridge? *Science, 297*(5584), 1137-1142. doi:10.1126/science.1069651

Anisimova, M., & Gascuel, O. (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol, 55*(4), 539-552. doi:10.1080/10635150600755453

Benner, S. A., Sassi, S. O., & Gaucher, E. A. (2007). Molecular paleoscience: systems biology from the past. *Adv Enzymol Relat Areas Mol Biol, 75*, 1-132, xi.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., . . . Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Res, 28*(1), 235-242.

Boyd, E. S., Anbar, A. D., Miller, S., Hamilton, T. L., Lavin, M., & Peters, J. W. (2011). A late methanogen origin for molybdenum-dependent nitrogenase. *Geobiology, 9*(3), 221-232. doi:10.1111/j.1472-4669.2011.00278.x

Boyd, E. S., Costas, A. M., Hamilton, T. L., Mus, F., & Peters, J. W. (2015). Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. *J Bacteriol, 197*(9), 1690-1699. doi:10.1128/JB.02611-14

Boyd, E. S., Hamilton, T. L., & Peters, J. W. (2011). An alternative path for the evolution of biological nitrogen fixation. *Front Microbiol, 2*, 205. doi:10.3389/fmicb.2011.00205

Boyd, E. S., & Peters, J. W. (2013). New insights into the evolutionary history of biological nitrogen fixation. *Front Microbiol, 4*, 201. doi:10.3389/fmicb.2013.00201

Brigle, K. E., Setterquist, R. A., Dean, D. R., Cantwell, J. S., Weiss, M. C., & Newton, W. E. (1987). Site-directed mutagenesis of the nitrogenase MoFe protein of Azotobacter vinelandii. *Proc Natl Acad Sci U S A, 84*(20), 7066-7069.

Bulen, W. A., & LeComte, J. R. (1966). The nitrogenase system from Azotobacter: two-enzyme requirement for N2 reduction, ATP-dependent H2 evolution, and ATP hydrolysis. *Proc Natl Acad Sci U S A, 56*(3), 979-986. doi:10.1073/pnas.56.3.979

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: architecture and applications. *BMC Bioinformatics, 10*, 421. doi:10.1186/1471-2105-10-421

Canfield, D. E., Glazer, A. N., & Falkowski, P. G. (2010). The evolution and future of Earth's nitrogen cycle. *Science, 330*(6001), 192-196. doi:10.1126/science.1186120

Chen, I. A., Chu, K., Palaniappan, K., Pillay, M., Ratner, A., Huang, J., . . . Kyrpides, N. C. (2019). IMG/M v.5.0: an integrated data management and comparative analysis system for microbial genomes and microbiomes. *Nucleic Acids Res, 47*(D1), D666-D677. doi:10.1093/nar/gky901

Christiansen, J., Cash, V. L., Seefeldt, L. C., & Dean, D. R. (2000). Isolation and characterization of an acetylene-resistant nitrogenase. *J Biol Chem, 275*(15), 11459-11464.

Corbett, M. C., Hu, Y., Fay, A. W., Ribbe, M. W., Hedman, B., & Hodgson, K. O. (2006). Structural insights into a protein-bound iron-molybdenum cofactor precursor. *Proc Natl Acad Sci U S A, 103*(5), 1238-1243. doi:10.1073/pnas.0507853103

Curatti, L., Hernandez, J. A., Igarashi, R. Y., Soboh, B., Zhao, D., & Rubio, L. M. (2007). In vitro synthesis of the iron-molybdenum cofactor of nitrogenase from iron, sulfur,

molybdenum, and homocitrate using purified proteins. *Proc Natl Acad Sci U S A, 104*(45), 17626-17631. doi:10.1073/pnas.0703050104

Dekas, A. E., Poretsky, R. S., & Orphan, V. J. (2009). Deep-sea archaea fix and share nitrogen in methane-consuming microbial consortia. *Science, 326*(5951), 422-426. doi:10.1126/science.1178223

Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C., & Dixon, R. (2012). Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics, 13*, 162. doi:10.1186/1471-2164-13-162

Durrant, J. D., Votapka, L., Sorensen, J., & Amaro, R. E. (2014). POVME 2.0: An Enhanced Tool for Determining Pocket Shape and Volume Characteristics. *J Chem Theory Comput, 10*(11), 5047-5056. doi:10.1021/ct500381c

Eady, R. R. (1996). Structure-function relationships of alternative nitrogenases. *Chemical Reviews, 96*(7), 3013-3030. doi:10.1021/cr950057h

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res, 32*(5), 1792-1797. doi:10.1093/nar/gkh340

Falkowski, P. G. (1997). Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature, 387*(6630), 272-275. doi:10.1038/387272a0

Fixen, K. R., Zheng, Y., Harris, D. F., Shaw, S., Yang, Z. Y., Dean, D. R., . . . Harwood, C. S. (2016). Light-driven carbon dioxide reduction to methane by nitrogenase in a photosynthetic bacterium. *Proc Natl Acad Sci U S A, 113*(36), 10163-10167. doi:10.1073/pnas.1611043113

Garcia, A. K., & Kacar, B. (2019). How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radic Biol Med*. doi:10.1016/j.freeradbiomed.2019.03.033

Gomez-Fernandez, B. J., Garcia-Ruiz, E., Martin-Diaz, J., Gomez de Santos, P., Santos-Moriano, P., Plou, F. J., . . . Alcalde, M. (2018). Directed -in vitro- evolution of Precambrian and extant Rubiscos. *Sci Rep, 8*(1), 5532. doi:10.1038/s41598-018-23869-3

Guo, Y., Echavarri-Erasun, C., Demuez, M., Jimenez-Vicente, E., Bominaar, E. L., & Rubio, L. M. (2016). The Nitrogenase FeMo-Cofactor Precursor Formed by NifB Protein: A Diamagnetic Cluster Containing Eight Iron Atoms. *Angew Chem Int Ed Engl, 55*(41), 12764-12767. doi:10.1002/anie.201606447

Hageman, R. V., & Burris, R. H. (1978). Nitrogenase and nitrogenase reductase associate and dissociate with each catalytic cycle. *Proc Natl Acad Sci U S A, 75*(6), 2699-2702. doi:10.1073/pnas.75.6.2699

Hales, B. J., Case, E. E., Morningstar, J. E., Dzeda, M. F., & Mauterer, L. A. (1986). Isolation of a new vanadium-containing nitrogenase from Azotobacter vinelandii. *Biochemistry, 25*(23), 7251-7255. doi:10.1021/bi00371a001

Hamilton, T. L., Ludwig, M., Dixon, R., Boyd, E. S., Dos Santos, P. C., Setubal, J. C., . . . Peters, J. W. (2011). Transcriptional profiling of nitrogen fixation in Azotobacter vinelandii. *J Bacteriol, 193*(17), 4477-4486. doi:10.1128/JB.05099-11

Hanson-Smith, V., & Johnson, A. (2016). PhyloBot: A Web Portal for Automated Phylogenetics, Ancestral Sequence Reconstruction, and Exploration of Mutational Trajectories. *PLoS Comput Biol, 12*(7), e1004976. doi:10.1371/journal.pcbi.1004976

Hanson-Smith, V., Kolaczkowski, B., & Thornton, J. W. (2010). Robustness of ancestral sequence reconstruction to phylogenetic uncertainty. *Mol Biol Evol, 27*(9), 1988-1999. doi:10.1093/molbev/msq081

Harris, D. F., Lukoyanov, D., Kallas, H., Trncik, C., Yang, Z., Compton, P., . . . Seefeldt, L. C. (2019). Mo-, V-, and Fe-nitrogenases use a universal eight-electron reductive-elimination mechanism to achieve N_2 reduction. *Biochemistry*.

Harris, D. F., Lukoyanov, D. A., Shaw, S., Compton, P., Tokmina-Lukaszewska, M., Bothner, B., . . . Seefeldt, L. C. (2018). Mechanism of N2 Reduction Catalyzed by Fe-Nitrogenase Involves Reductive Elimination of H2. *Biochemistry, 57*(5), 701-710. doi:10.1021/acs.biochem.7b01142

Harris, D. F., Yang, Z. Y., Dean, D. R., Seefeldt, L. C., & Hoffman, B. M. (2018). Kinetic Understanding of N2 Reduction versus H2 Evolution at the E4(4H) Janus State in the Three Nitrogenases. *Biochemistry, 57*(39), 5706-5714. doi:10.1021/acs.biochem.8b00784

Hoffman, B. M., Lukoyanov, D., Yang, Z. Y., Dean, D. R., & Seefeldt, L. C. (2014). Mechanism of nitrogen fixation by nitrogenase: the next stage. *Chem Rev, 114*(8), 4041-4062. doi:10.1021/cr400641x

Hu, B., Harris, D. F., Dean, D. R., Liu, T. L., Yang, Z. Y., & Seefeldt, L. C. (2018). Electrocatalytic CO_2 reduction catalyzed by nitrogenase MoFe and FeFe proteins. *Bioelectrochemistry, 120*, 104-109. doi:10.1016/j.bioelechem.2017.12.002

Hu, Y., Fay, A. W., & Ribbe, M. W. (2005). Identification of a nitrogenase FeMo cofactor precursor on NifEN complex. *Proc Natl Acad Sci U S A, 102*(9), 3236-3241. doi:10.1073/pnas.0409201102

Hu, Y., Lee, C. C., & Ribbe, M. W. (2011). Extending the Carbon Chain: Hydrocarbon Formation Catalyzed by Vanadium/Molybdenum Nitrogenases. *Science, 333*(6043), 753-755. doi:10.1126/science.1206883

Hu, Y., & Ribbe, M. W. (2011). Biosynthesis of Nitrogenase FeMoco. *Coord Chem Rev, 255*(9-10), 1218-1224. doi:10.1016/j.ccr.2010.11.018

Hu, Y., & Ribbe, M. W. (2015). Nitrogenase and homologs. *J Biol Inorg Chem, 20*(2), 435-445. doi:10.1007/s00775-014-1225-3

Joerger, R. D., & Bishop, P. E. (1988). Bacterial alternative nitrogen fixation systems. *Crit Rev Microbiol, 16*(1), 1-14. doi:10.3109/10408418809104465

Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci, 8*(3), 275-282.

Kacar, B., Hanson-Smith, V., Adam, Z. R., & Boekelheide, N. (2017). Constraining the timing of the Great Oxidation Event within the Rubisco phylogenetic tree. *Geobiology, 15*(5), 628-640. doi:10.1111/gbi.12243

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

Kennedy, C., & Dean, D. (1992). The nifU, nifS and nifV gene products are required for activity of all three nitrogenases of Azotobacter vinelandii. *Mol Gen Genet, 231*(3), 494-498.

Keppen, O. I., Baulina, O. I., & Kondratieva, E. N. (1994). Oscillochloris trichoides neotype strain DG-6. *Photosynth Res, 41*(1), 29-33. doi:10.1007/BF02184143

Keppen, O. I., Lebedeva, N. V., Troshina, O. Y., & Rodionov, Y. V. (1989). The nitrogenase activity of filamentous phototrophic green bacterium. *Mikrobiologiya, 58*, 520-521.

Kim, C. H., Newton, W. E., & Dean, D. R. (1995). Role of the MoFe protein alpha-subunit histidine-195 residue in FeMo-cofactor binding and nitrogenase catalysis. *Biochemistry, 34*(9), 2798-2808.

Krahn, E., Weiss, R., Krockel, M., Groppe, J., Henkel, G., Cramer, P., . . . Muller, A. (2002). The Fe-only nitrogenase from Rhodobacter capsulatus: identification of the cofactor, an unusual, high-nuclearity iron-sulfur cluster, by Fe K-edge EXAFS and 57Fe Mossbauer spectroscopy. *J Biol Inorg Chem, 7*(1-2), 37-45. doi:10.1007/s007750100263

Le, S. Q., & Gascuel, O. (2008). An improved general amino acid replacement matrix. *Mol Biol Evol, 25*(7), 1307-1320. doi:10.1093/molbev/msn067

Lee, C. C., Tanifuji, K., Newcomb, M., Liedtke, J., Hu, Y., & Ribbe, M. W. (2018). A Comparative Analysis of the CO-Reducing Activities of MoFe Proteins Containing Mo- and V-Nitrogenase Cofactors. *Chembiochem, 19*(7), 649-653. doi:10.1002/cbic.201800035

Liu, Y., Schmidt, B., & Maskell, D. L. (2010). MSAProbs: multiple sequence alignment based on pair hidden Markov models and partition function posterior probabilities. *Bioinformatics, 26*(16), 1958-1964. doi:10.1093/bioinformatics/btq338

Lyons, T. W., Reinhard, C. T., & Planavsky, N. J. (2014). The rise of oxygen in Earth's early ocean and atmosphere. *Nature, 506*(7488), 307-315. doi:10.1038/nature13068

McGlynn, S. E., Boyd, E. S., Peters, J. W., & Orphan, V. J. (2012). Classifying the metal dependence of uncharacterized nitrogenases. *Front Microbiol, 3*, 419. doi:10.3389/fmicb.2012.00419

McRose, D. L., Zhang, X., Kraepiel, A. M., & Morel, F. M. (2017). Diversity and Activity of Alternative Nitrogenases in Sequenced Genomes and Coastal Environments. *Front Microbiol, 8*, 267. doi:10.3389/fmicb.2017.00267

Mehta, M. P., & Baross, J. A. (2006). Nitrogen fixation at 92 degrees C by a hydrothermal vent archaeon. *Science, 314*(5806), 1783-1786. doi:10.1126/science.1134772

Moore, E. K., Jelen, B. I., Giovannelli, D., Raanan, H., & Falkowski, P. G. (2017). Metal availability and the expanding network of microbial metabolisms in the Archaean eon. *Nature Geoscience, 10*(9), 629-636. doi:10.1038/ngeo3006

Mus, F., Alleman, A. B., Pence, N., Seefeldt, L. C., & Peters, J. W. (2018). Exploring the alternatives of biological nitrogen fixation. *Metallomics, 10*(4), 523-538. doi:10.1039/c8mt00038g

Mus, F., Colman, D. R., Peters, J. W., & Boyd, E. S. (2019). Geobiological feedbacks, oxygen, and the evolution of nitrogenase. *Free Radic Biol Med*. doi:10.1016/j.freeradbiomed.2019.01.050

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., . . . Pruitt, K. D. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res, 44*(D1), D733-745. doi:10.1093/nar/gkv1189

Ormeño-Orrillo, E., Hungria, M., & Martinez-Romero, E. (2013). Dinitrogen-Fixing Prokaryotes. In *The Prokaryotes* (pp. 427-451).

Quang, L. S., Gascuel, O., & Lartillot, N. (2008). Empirical profile mixture models for phylogenetic reconstruction. *Bioinformatics, 24*(20), 2317-2323. doi:10.1093/bioinformatics/btn445

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raymond, J., Siefert, J. L., Staples, C. R., & Blankenship, R. E. (2004). The natural history of nitrogen fixation. *Mol Biol Evol, 21*(3), 541-554. doi:10.1093/molbev/msh047

Rebelein, J. G., Lee, C. C., Newcomb, M., Hu, Y., & Ribbe, M. W. (2018). Characterization of an M-Cluster-Substituted Nitrogenase VFe Protein. *MBio, 9*(2). doi:10.1128/mBio.00310-18

Roll, J. T., Shah, V. K., Dean, D. R., & Roberts, G. P. (1995). Characteristics of NIFNE in Azotobacter vinelandii strains. Implications for the synthesis of the iron-molybdenum cofactor of dinitrogenase. *J Biol Chem, 270*(9), 4432-4437. doi:10.1074/jbc.270.9.4432

Rubio, L. M., & Ludden, P. W. (2008). Biosynthesis of the iron-molybdenum cofactor of nitrogenase. *Annu Rev Microbiol, 62*, 93-111. doi:10.1146/annurev.micro.62.081307.162737

Sali, A., & Blundell, T. L. (1993). Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol, 234*(3), 779-815. doi:10.1006/jmbi.1993.1626

Sarma, R., Barney, B. M., Keable, S., Dean, D. R., Seefeldt, L. C., & Peters, J. W. (2010). Insights into substrate binding at FeMo-cofactor in nitrogenase from the structure of an alpha-70(Ile) MoFe protein variant. *J Inorg Biochem, 104*(4), 385-389. doi:10.1016/j.jinorgbio.2009.11.009

Schmid, B., Einsle, O., Chiu, H. J., Willing, A., Yoshida, M., Howard, J. B., & Rees, D. C. (2002). Biochemical and structural characterization of the cross-linked complex of nitrogenase: comparison to the ADP-AlF4(-)-stabilized structure. *Biochemistry, 41*(52), 15557-15565.

Seefeldt, L. C., Hoffman, B. M., & Dean, D. R. (2009). Mechanism of Mo-dependent nitrogenase. *Annu Rev Biochem, 78*, 701-722. doi:10.1146/annurev.biochem.78.070907.103812

Setubal, J. C., dos Santos, P., Goldman, B. S., Ertesvag, H., Espin, G., Rubio, L. M., . . . Wood, D. (2009). Genome sequence of Azotobacter vinelandii, an obligate aerobe specialized to support diverse anaerobic metabolic processes. *J Bacteriol, 191*(14), 4534-4545. doi:10.1128/JB.00504-09

Shah, V. K., Allen, J. R., Spangler, N. J., & Ludden, P. W. (1994). In vitro synthesis of the iron-molybdenum cofactor of nitrogenase. Purification and characterization of NifB cofactor, the product of NIFB protein. *J Biol Chem, 269*(2), 1154-1158.

Shah, V. K., Imperial, J., Ugalde, R. A., Ludden, P. W., & Brill, W. J. (1986). In vitro synthesis of the iron-molybdenum cofactor of nitrogenase. *Proc Natl Acad Sci U S A, 83*(6), 1636-1640. doi:10.1073/pnas.83.6.1636

Shock, E. L., & Boyd, E. S. (2015). Principles of Geobiochemistry. *Elements, 11*(6), 395-401. doi:10.2113/gselements.11.6.395

Sippel, D., & Einsle, O. (2017). The structure of vanadium nitrogenase reveals an unusual bridging ligand. *Nat Chem Biol, 13*(9), 956-960. doi:10.1038/nchembio.2428

Sippel, D., Rohde, M., Netzer, J., Trncik, C., Gies, J., Grunau, K., . . . Einsle, O. (2018). A bound reaction intermediate sheds light on the mechanism of nitrogenase. *Science, 359*(6383), 1484-1489. doi:10.1126/science.aar2765

Sjolander, K., Karplus, K., Brown, M., Hughey, R., Krogh, A., Mian, I. S., & Haussler, D. (1996). Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci, 12*(4), 327-345. doi:10.1093/bioinformatics/12.4.327

Soboh, B., Boyd, E. S., Zhao, D., Peters, J. W., & Rubio, L. M. (2010). Substrate specificity and evolutionary implications of a NifDK enzyme carrying NifB-co at its active site. *FEBS Lett, 584*(8), 1487-1492. doi:10.1016/j.febslet.2010.02.064

Spatzal, T., Aksoyoglu, M., Zhang, L., Andrade, S. L., Schleicher, E., Weber, S., . . . Einsle, O. (2011). Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science, 334*(6058), 940. doi:10.1126/science.1214025

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics, 30*(9), 1312-1313. doi:10.1093/bioinformatics/btu033

Stueken, E. E., Buick, R., Guy, B. M., & Koehler, M. C. (2015). Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature, 520*(7549), 666-669. doi:10.1038/nature14180

Tal, S., Chun, T. W., Gavini, N., & Burgess, B. K. (1991). The delta nifB (or delta nifE) FeMo cofactor-deficient MoFe protein is different from the delta nifH protein. *J Biol Chem, 266*(16), 10654-10657.

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

Whelan, S., & Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol, 18*(5), 691-699. doi:10.1093/oxfordjournals.molbev.a003851

Yang, Z. (1993). Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol Biol Evol, 10*(6), 1396-1401. doi:10.1093/oxfordjournals.molbev.a040082

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol, 24*(8), 1586-1591. doi:10.1093/molbev/msm088

Yang, Z. Y., Moure, V. R., Dean, D. R., & Seefeldt, L. C. (2012). Carbon dioxide reduction to methane and coupling with acetylene to form propylene catalyzed by remodeled nitrogenase. *Proc Natl Acad Sci U S A, 109*(48), 19644-19648. doi:10.1073/pnas.1213159109

Zhang, X., McRose, D. L., Darnajoux, R., Bellenger, J. P., Morel, F. M. M., & Kraepiel, A. M. L. (2016). Alternative nitrogenase activity in the environment and nitrogen cycle implications. *Biogeochemistry, 127*(2-3), 189-198. doi:10.1007/s10533-016-0188-6

Zheng, Y., Harris, D. F., Yu, Z., Fu, Y., Poudel, S., Ledbetter, R. N., . . . Harwood, C. S. (2018). A pathway for biological methane production using bacterial iron-only nitrogenase. *Nat Microbiol, 3*(3), 281-286. doi:10.1038/s41564-017-0091-5