The evolution and diversity of noncanonical microbial nitrogen metabolisms

by

Sarah L. Schwartz

S.M Science Writing, Massachusetts Institute of Technology, 2015
B.S. Environmental Systems, University of California, San Diego, 2014

Submitted to the Microbiology Graduate Program in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
at the
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

Signature of Author……
Microbiology Graduate Program
August 2022

Certified by……
Gregory P. Fournier
Associate Professor of Geobiology, MIT
Thesis Supervisor

Accepted by……
Jacquin C. Niles
Associate Professor of Biological Engineering
Chair of Microbiology Program

The evolution and diversity of noncanonical microbial nitrogen metabolisms

by

Sarah L. Schwartz

Submitted to the Microbiology Graduate Program in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Microbiology

DOCTOR OF PHILOSOPHY

Abstract

Microbial nitrogen cycling underpins Earth's most basic ecology. The environmental and commercial significance of nitrogen biogeochemistry has prompted substantial study of the modern nitrogen cycle, but much remains unknown about the evolution of nitrogen biochemistry through deep time. This thesis explores the evolution and nature of divergent or noncanonical nitrogen metabolisms. The first section of this work explores divergent enzymes of denitrification. I report a novel cytochrome-type nitrite reductase (*nirS*) structure within members of Phylum Chloroflexi, which includes a cytochrome superfamily domain in addition to functional domains conserved in Proteobacterial *nirS*. Phylogenetic domain mapping reveals that this gene resulted from a chimeric domain fusion in ancestral Chloroflexi. I also identify an underreported variant of nitric oxide reductase (eNOR) within Chloroflexi MAGs, and provide a detailed phylogeny of this enzyme variant, revealing much broader diversity than previously reported. Next, I explore the evolution of microbial cyanide metabolism. Cyanide-degrading enzymes have been extensively studied for bioremediation and biotechnology, but the extant diversity of such enzymes remains underexplored. Additionally, while cyanide is hypothesized to play a central role in prebiotic chemistry, few biological data constrain the age and emergence of cyanide metabolisms. I provide a comprehensive analysis of the distribution and evolution of the Class I nitrilases, a subfamily specialized for hydrogen cyanide (HCN) reduction. Gene trees reveal that cyanide-reducing nitrilases originated in bacteria and were transferred into eukaryotes, refuting earlier eukaryotic origin hypotheses. Molecular clock analyses indicate that this enzyme subfamily shares an ancestor that emerged 1-2 billion years ago in the Paleo- to Mesoproterozic. I also analyze the constrained prokaryotic distributions of other nonhomologous nitrile-reducing enzymes, thiocyanate hydrolases and nitrile hydratases. Finally, I detail the evolution of promiscuous nitrogen metabolism in nitrogenases. Though highly specialized for dinitrogen reduction, molybdenum and vanadium nitrogenases have been shown to reduce off-target nitrogenous substrates, including HCN. Using ancestral sequence reconstruction, I compare the sequence space of predicted ancestral and extant nitrogenase substrate channels. The results indicate that the predicted highest-likelihood ancestral states for key residues are not represented in extant sequence space, and the physicochemical types of these high-likelihood residue combinations are only rarely represented in divergent nitrogenases. These data suggest that ancestral nitrogenases likely had alternative substrate channel compositions, possibly reflecting selection in early Earth environments. Together, these data suggest that the diversity and age of microbial nitrogen metabolisms are currently underestimated, and further study of these pathways should shed light on large-scale patterns of microbial evolution and ecology.

Thesis Supervisor: Gregory Fournier
Title: Associate Professor of Geobiology
Massachusetts Institute of Technology

Acknowledgements

To my advisor, Greg Fournier: Thank you for your patient, expert guidance; your sense of humor and humanity; and for championing creative, responsible research.

Many thanks to my committee members, Tami Lieberman, Mick Follows, and Ant Poole, for your advice, time, support, and leadership.

Thanks to the other faculty members whose instruction, guidance, or kindness has shaped my doctoral work and career: Betül Kaçar, Kris Prather, Otto Cordero, Jacquin Niles, and Martin Polz.

To members of the Fournier Lab, past and present—Danielle, Makayla, Abby, Jack, Chris, Erik, Fatima, Maddie, Cara, Lily: Thank you so much for your camaraderie, inspiration, and support. I owe a special debt of gratitude to my postdoc mentors. Jo—thanks for being a guide, advocate, and friend (and for the bubble tea and recipes). Thiberio—thank you for the instruction, pep talks, problem solving, and for teaching me to think in new ways, without fear of breaking things. This wouldn't have been possible without your efforts.

To the best Micro cohort, Liz, Rachel, Jai, and Anthony: It was an honor to be your classmate, and a pleasure to be your friend. Thanks also to the people who mentored me as a budding microbiologist—Einat, Tim, Fatima, and Manoshi.

To Jacquie, Mira, Brenda, Susan, and Bonnielee: I can't thank you enough for everything you did to make this possible, and for all your kindness and encouragement throughout.

Thanks to the whole Kaçar Lab for welcoming me and sharing their collaborative energy and scientific excitement. Special thanks to Amanda for your expertise and collaboration—I've been so fortunate to work with you.

To my colleagues in EAPS: Thank you for sharing your talent, resilience, and kindness, and for your tireless work to build a better world. Special thanks to Jeemin, Ellen, Kelsey, Athena, Claire, Marjorie, and Julia for making a huge difference with your friendship and/or guidance.

Thanks to the members of the MIT Communication Lab, and especially Diana and Prerna for years of leadership and support. My time as a Comm Fellow was a huge highlight of my time at MIT.

Having the opportunity to pursue a PhD requires the advocacy and guidance of incredible undergraduate mentors. Many thanks again to Matt Bertin, Bill Gerwick, Lin Chao, Camilla Rang, Donna Ferriero, and Martin Mulvihill.

Thanks to Kevin for the years of partnership and adventures.

Christopher, Chris, Diana, Emilie, Jason, Ludovic, Mike—thank you so much for your friendship. Your brightness and kindness always remind me what is most important in life—and you enrich mine daily. I'm so grateful to know you.

Dad: Thanks for helping me recognize my goals, including those that are different than yours, and for all the Q&A along the way.

Rachel: Thanks for being my best friend and role model. Anything is possible when someone brave, moral, and brilliant goes first, and helps you find the path—or carries you, like a football.

Mom: Thank you for making it possible to get to this, through this, and onward. I'm grateful every day for your strength, love, and example.

This work was made possible by a National Defense Science and Engineering Graduate fellowship (Army Research Office), and a grant from the Simons Foundation Collaboration on the Origins of Life.

Thanks to Paul Tol (Netherlands Institute for Space Research), Martin Krzywinski (Michael Smith Genome Sciences Center), and the Jackson Lab (University of Wisconsin, Madison) for providing colorblind-friendly palettes used in this work.

Table of Contents

List of Figures and Tables

Chapter 1. Introduction

1.1 The modern microbial nitrogen cycle

As a required component of protein and nucleic acid synthesis, nitrogen is essential for all life on Earth (Canfield, Glazer and Falkowski, 2010; Thomazo and Papineau, 2013; Kuypers, Marchant and Kartal, 2018). Though dinitrogen gas is the most abundant constituent of Earth's atmosphere, this inert, stable nitrogen reservoir is inaccessible to most living organisms (Canfield, Glazer and Falkowski, 2010; Thomazo and Papineau, 2013; Kuypers, Marchant and Kartal, 2018). The massive nitrogen pool stored in sediments—usually as ammonium within silicate minerals—can slowly become available through erosion, but this input plays a minor role in nitrogen biogeochemistry. Instead, productivity is severely limited by bioavailable or "fixed" forms of nitrogen, which include the inorganic ions nitrate (NO_3^-), nitrite (NO_2^-), and ammonium (NH_4^+) (Canfield, Glazer and Falkowski, 2010; Thomazo and Papineau, 2013). In modern ecosystems, the bulk of redox cycling between nitrogen species is microbially mediated. The canonical nitrogen cycle connects constituent species through six major pathways. Nitrogen fixation, nitrification, and dissimilatory nitrate reduction to ammonium (DRNA) produce fixed inorganic species; denitrification and anaerobic ammonia oxidation (anammox) ultimately produce dinitrogen gas; and ammonification forms a loop between the ammonium pool and organic nitrogen in biomass (Fig. 1.1).

1.1.1 Nitrogen fixation

Nitrogen fixation converts atmospheric dinitrogen to ammonium, a process that is the purview of the highly specialized, anaerobic nitrogenases. With most nitrogen either tied up in biomass or biologically unavailable, fixation is an essential and limiting process in many ecosystems, even with additional fixed nitrogen inputs from abiotic processes such as lightning discharge or photochemical mineralization, and increasing anthropogenic inputs from fertilizer produced through the Haber-Bosch process (Raymond, 2005; Gruber and Galloway, 2008; Doane, 2017). Though essential and conserved in many prokaryotes—including organisms that are not obligate anaerobes—nitrogen fixation is an extremely metabolically expensive process, requiring 16 ATP molecules per molecule of N_2 (Canfield, Glazer and Falkowski, 2010; Sickerman, Hu and Ribbe, 2019).

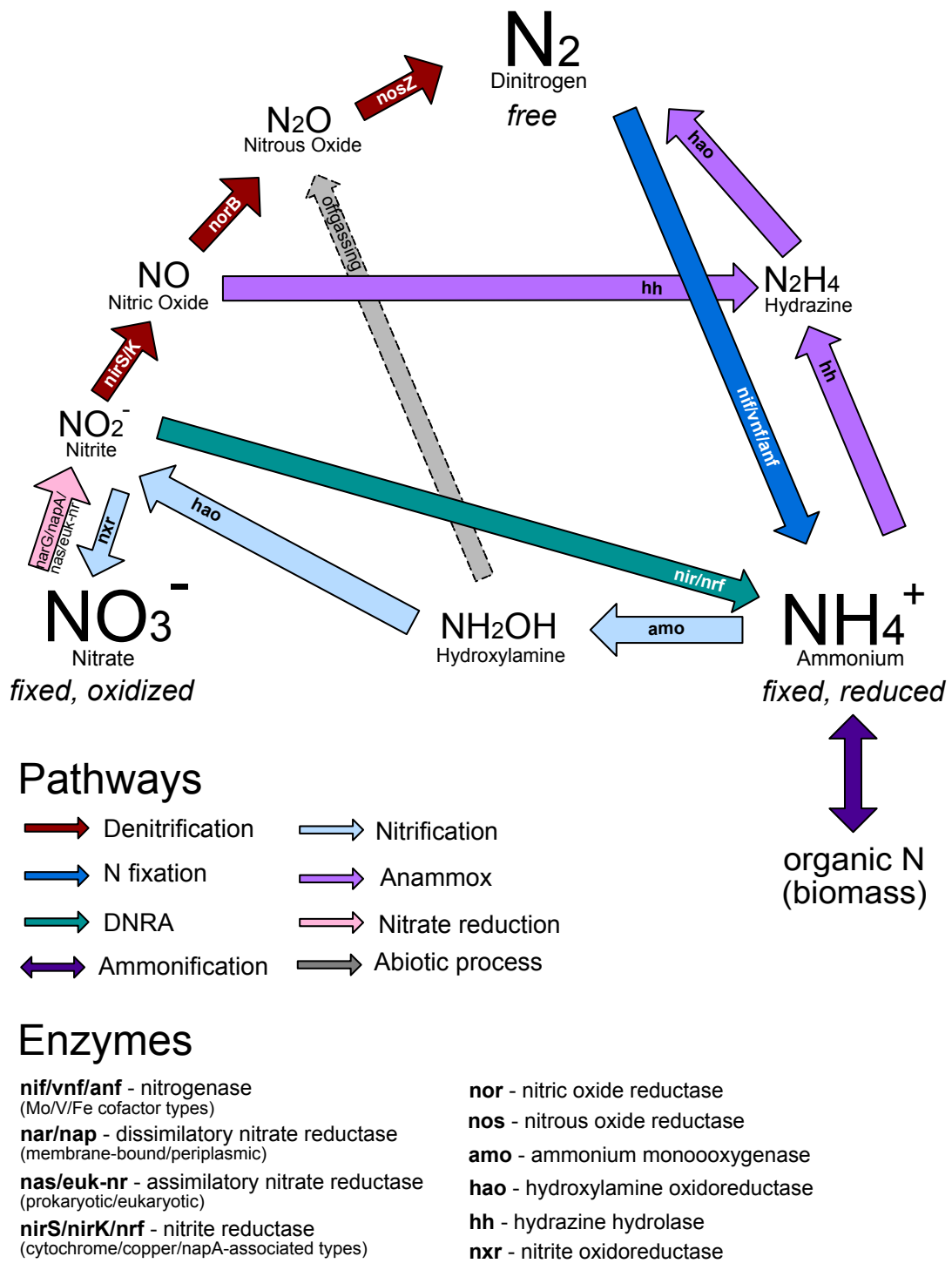


Figure 1.1 The nitrogen cycle
The modern nitrogen cycle consists of six major redox pathways and a multifunctional nitrate reduction step. Denitrification reduces nitrite to dinitrogen; nitrogen fixation reduces dinitrogen to ammonium; dissimilatory reduction of nitrate to ammonium (DNRA) reduces nitrate to ammonium; anaerobic ammonium oxidation (anammox) oxidizes ammonium or reduces nitric oxide to dinitrogen; nitrification oxidizes ammonium to nitrate; ammonification refers to the generalized processes forming a loop between ammonium and the organic nitrogen pool stored in biomass. Figured adapted from Canfield (2010).

1.1.2 Nitrification

Nitrification is a dissimilatory energy metabolism that oxidizes ammonium to nitrate, with hydroxylamine and nitrite as intermediates (Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). This prokaryotic, aerobic process is coupled with carbon chemolithoautotrophy in cosmopolitan Proteobacteria and some members of the Thaumarchaeota (Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). Recently-described *Nitrospira spp.* can also perform complete nitrification (also referred to as "complete ammonia oxidation" or "commamox") (Kuypers, Marchant and Kartal, 2018).

1.1.3 DRNA

DRNA reduces nitrate to nitrite, and then nitrite to ammonium; this latter step is coupled with the oxidation of an electron donor such as carbon, iron, or sulfide, enabling microbial growth (Canfield, Glazer and Falkowski, 2010; Giblin *et al.*, 2013; Kuypers, Marchant and Kartal, 2018; Carlson *et al.*, 2020). Diverse bacteria and some methanotrophic and thermophilic archaea can perform DRNA, as can diatoms and fungi (Giblin *et al.*, 2013; Kuypers, Marchant and Kartal, 2018). DRNA appears to be favored in environments with an excess of carbon or other electron donors, higher salinity, or higher temperatures—otherwise, denitrification is the more prevalent metabolism, resulting in a net loss of organic nitrogen from the system by reduction to N_2 (Giblin *et al.*, 2013; Kuypers, Marchant and Kartal, 2018; Carlson *et al.*, 2020).

1.1.4 Denitrification

Denitrification also couples nitrogen reduction with carbon oxidation, but reduces nitrate or nitrate to dinitrogen gas—essentially reversing the redox directionality of nitrogen fixation (Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). The pathway's intermediates are nitric oxide and nitrous oxide, which are of significant biological and environmental significance: the former is a signaling molecule and cytotoxin, and the latter is an extremely potent, ozone-destroying greenhouse gas (Canfield, Glazer and Falkowski, 2010; Decleyre *et al.*, 2016; Kuypers, Marchant and Kartal, 2018).

Denitrifiers, which represent an extremely diverse assortment of bacteria, several halophilic or hyperthermophilic archaea, and a limited number of unicellular eukaryotes—are most often facultative anaerobes, switching from aerobic respiration to nitrogenous respiration in

oxygen-depleted environments (Zumft, 1997; Priemé, Braker and Tiedje, 2002; Canfield, Glazer and Falkowski, 2010).

1.1.5 Anammox

Like denitrification, anammox produces dinitrogen gas; however, anammox proceeds by reducing nitric oxide using ammonium as an electron donor, in the absence of additional reductants. This produces hydrazine, which is then reduced to N_2 (Kuenen J. G., 2008; Canfield, Glazer and Falkowski, 2010). Anammox is found exclusively in certain members of the Planctomycetes, and occurs within a specialized energy-generating and energy-conserving organelle (the "anammoxosome") (Kuenen J. G., 2008). The pathway has been observed primarily in marine and aqueous habitats, and requires available nitrite (which is subsequently reduced to nitric oxide) and ammonium, which are inputs from other nitrogen cycle pathways (Kuenen J. G., 2008; Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018).

1.1.6 Ammonification

The largest estimated flux of nitrogen in the cycle overall comes through ammonification, which decomposes the organic nitrogen stored in biomass into ammonium (Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). (This process is distinct from "nitrate ammonification," which is an alternative term for DRNA (Bonin, 1996; Strohm *et al.*, 2007).) This pathway forms a redox loop between the inorganic ammonium pool and the abundant organic nitrogen stored as amino groups in biomass (Strock, 2008; Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). Ammonium assimilation often occurs preferentially to nitrate assimilation in marine bacteria and plants (Strock, 2008).

1.1.7 Diversity and distribution of nitrogen cycle pathways

The taxonomic diversity of the core nitrogen metabolisms varies widely. At one extreme, pathways such as denitrification or DNRA can be found conserved within multiple bacterial phyla, as well as a small number of archaeal and eukaryotic groups (Zumft, 1997; Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). At the other extreme, anammox has only been identified within five genera of Planctomycetes bacteria, within Order Candidatus Brocadiales (Kartal and Keltjens, 2016; Kuypers, Marchant and Kartal, 2018). Conceptually,

identifying metabolisms with such limited distribution *in situ* should be quite challenging. The discovery of anammox was aided somewhat by earlier discussions of anaerobic ammonium oxidation as a conspicuously "missing" lithotrophy *in situ* (Broda, 1977; Kuenen J. G., 2008). When inexplicable ammonium loss and dinitrogen production were observed in an anaerobic system years later, the hypothetical anammox metabolism was primed for evaluation (Mulder *et al.*, 1995; Kuenen J. G., 2008). Identifying or isolating other nitrogen metabolisms with a similarly constrained distribution, but a less unique metabolic profile, may prove far more difficult.

Similar challenges will apply to identification of taxonomically limited metabolisms that are not novel metabolisms, but unreported variants of established pathways. In Chapter 2, I present a variant nitrite reductase—a key enzyme in denitrification—that appears to be constrained to members of Order Anaerolineales, within Phylum Chloroflexi. It is possible that this genetic divergence, and any resulting biochemical or metabolic differences, reflect environmental pressures on the strains in question. Chloroflexi at large are cosmopolitan environmental bacteria, representing a variety of lifestyles in diverse ecosystems, but Family Anaerolineaceae are all chemoheterotrophic, filamentous, strict anaerobes usually found in anoxic sediment, subsurface, or sludge biomes (Yamada and Sekiguchi, 2018; Speirs *et al.*, 2019). It is possible, therefore, that the modified denitrification machinery in this order reflects an advantage at the organismal level, or confers specific utility in anoxic, nutrient-limited, light-limited biomes.

Depending on the genes and metabolisms in question, it is also possible that nitrogen metabolisms with limited taxonomic distribution reflect rare evolutionary bottlenecks, or simply observation bias in metagenomic sampling.

1.1.8 Complexities of modern nitrogen cycling

While the number of pathways and enzymes involved in the canonical nitrogen cycle are relatively straightforward, the *in vivo* capacity and *in situ* reality of nitrogen cycling is complicated. Several nitrogen cycle redox steps have enzymatic redundancy. For example, nitrate reduction can be performed by at least four different enzymes, one of which is eukaryotic, some of which perform assimilatory reduction, instead of dissimilatory respiration as seen in denitrification—and which, together, can be found across all domains of life and in a wide variety of environments (Canfield, Glazer and Falkowski, 2010; Kuypers, Marchant and Kartal, 2018). Very often, individual organisms or strains contain only one or some of the enzymes required for a given cycle pathway.

For example, many organisms may be missing one or two of the genes required for "full" denitrification, but are still considered denitrifiers (Decleyre *et al.*, 2016). Or, *Nitrospira* may perform complete nitrification in ammonia-limited conditions, but when oxygen is limiting, might perform only partial nitrification, producing nitrite as the final product (Kuypers, Marchant and Kartal, 2018). Intermediate species are often physically removed at variable points and rates; for example, nitrate and ammonium may be assimilated into biomass, or gaseous nitrous oxide may escape from denitrifying or nitrifying communities, resulting in greenhouse gas "hot spots" where these processes are active (Canfield, Glazer and Falkowski, 2010; Schutte *et al.*, 2015; Decleyre *et al.*, 2016; Luo *et al.*, 2017).

All told, holistically mapping nitrogen cycle dynamics is difficult. While estimates exist for the size of various nitrogen reservoirs and pathway fluxes, the cycle as a whole is not balanced (Galloway *et al.*, 2004; Kuypers, Marchant and Kartal, 2018). The complexity of global nitrogen biogeochemistry is exacerbated by inputs or loss from abiotic nitrogen cycling—such as lightning and photochemical fixation, or thermochemical decomposition of fixed compounds (Galloway *et al.*, 2004; Doane, 2017). The cycle is further complicated by steadily increasing anthropogenic inputs from the industrial Haber-Bosch process (Galloway *et al.*, 2004; Gruber and Galloway, 2008; Kuypers, Marchant and Kartal, 2018). In the coming decades, ongoing research will be needed to not only better-constrain the basics of these interconnected fluxes, but also to understand how these inputs perturb one another on a changing planet.

1.2 The evolution of Earth's nitrogen biogeochemistry

The complex nitrogen cycle has changed markedly over Earth's history, evolving along with abiotic inputs, other organic geochemical cycles, and the evolution and radiation of geobiological metabolisms. The dominance of dinitrogen within the atmosphere originated early, through volcanic and impact outgassing, and thermochemical oxidation of frozen ammonia and simple organic compounds delivered to the accreting Earth (Canfield, Glazer and Falkowski, 2010; Zahnle, Schaefer and Fegley, 2010).

It is unclear what the first nitrogen metabolism was, or if it is one of the pathways currently represented in the modern cycle (Canfield, Glazer and Falkowski, 2010). However, fixed nitrogen would have rapidly become limiting for Earth's earliest organisms, especially after the advent of anoxygenic photosynthesis increased nutrient flux through ecosystems (Canfield, Glazer and

Falkowski, 2010). It is unsurprising, then, that nitrogenases are known to be particularly ancient. Phylogenetic and isotopic analyses have constrained the origin of nitrogen fixation to the Archaean or early Proterozoic (Boyd *et al.*, 2011; Stüeken *et al.*, 2015; Parsons *et al.*, 2021). The iron-dependence and extreme oxygen sensitivity of all nitrogenases also provides convincing support for the origin of these enzymes in the anoxic Archaean environment (Canfield, Glazer and Falkowski, 2010; Garcia *et al.*, 2020).

Denitrification and anammox are also anaerobic pathways, but these processes use oxidized substrates. As such, sustaining the flux of these pathways within the modern nitrogen cycle requires oxygenic nitrification to replenish the nitrate and nitrite pools. This oxygenic process would have become more ubiquitous and relevant after the Great Oxygenation Event (GOE) oxygenated the planet's oceans and atmosphere roughly 2.3 billion years ago (Canfield, Glazer and Falkowski, 2010; Gold *et al.*, 2017; Genming *et al.*, 2022). However, isotopic fractionation of nitrogen and carbon in ancient sediments suggests that nitrification and denitrification may have begun before the GOE, roughly 2.5 billion years ago (Garvin *et al.*, 2009). These data support the existence of ancient nitrogen cycling through oxygen-dependent intermediates, established when transient oxidation of ammonium—provided by ancient nitrogenases, abiotic fixation, or both—could have sustained oxygenic nitrogen metabolisms in an otherwise anoxic world (Garvin *et al.*, 2009).

Despite the apparently early establishment of an N_2-rich atmosphere and extant nitrogen cycle pathways, the environmental context and magnitude of nitrogen pools and fluxes have certainly varied significantly over Earth's deep history. Key nitrogen budget and cycle transitions would have been driven by major shifts in global geobiology, particularly the emergence of key microbial groups and metabolisms. For example, prebiotic Earth is predicted to have sustained a standing stock of hydrogen cyanide (HCN), but this would have increased dramatically during the Archaean, when methanogenic organisms emerged. Atmospheric models show that increasing atmospheric methane concentration increases deposition of HCN, which could be converted to ammonia even in the absence of biotic nitrogen fixation (Tian, Kasting and Zahnle, 2011). The evolution of photosynthetic organisms, in turn, would have rapidly depleted fixed nitrogen and iron reservoirs within photic zones, leading to stratification of ammonium in early oceans—and subsequently exerting strong selective pressure for nitrogen fixation (Canfield, Glazer and Falkowski, 2010). The complete nature of these interconnected biogeochemical processes is still

obscure, and necessarily constrained by an understanding of extant metabolisms and fluxes. It is therefore highly likely that extant nitrogen metabolisms do not represent the full enzymatic or biochemical capacity of ancestral nitrogen metabolisms, especially within the limited number of pathways considered central to the modern nitrogen cycle.

1.3 Molecular evolution approaches for interrogating ancient biogeochemistry

The geochemical record of nitrogen cycle evolution depends upon the quality and availability of preserved nitrogen in the rock record. Most nitrogen is preserved in organic forms or through ammonium replacement of potassium in silicate minerals (Thomazo and Papineau, 2013). ^{15}N isotopic analyses of well-preserved nitrogen compounds can constrain hypotheses about the presence and age of nitrogen metabolisms for which the enrichment of modern pathways is well-established (Garvin *et al.*, 2009; Canfield, Glazer and Falkowski, 2010; Thomazo and Papineau, 2013; Sigman and Fripiat, 2019).

Conversely, direct biological records for ancient prokaryotes and their processes—including nitrogen metabolisms—are sparse. Any rock record of the earliest microorganisms is severely limited by the overall paucity of Archaean or Hadean rocks, which are progressively lost to metamorphic and diagenetic processes (Knoll, Bergmann and Strauss, 2016; Nutman *et al.*, 2016). Existing Archaean and Proterozoic microbial fossils, including preserved microbial mats (stromatolites), are relatively scarce or limited in geographical range (Sugitani *et al.*, 2010; Knoll, Bergmann and Strauss, 2016; Nutman *et al.*, 2016). Some microbial fossils are also morphologically ambiguous or contentious (Martín-González *et al.*, 2009; Sugitani *et al.*, 2010; Knoll, Bergmann and Strauss, 2016).

Analysis of the genetic fossil record, accessed through the sequence information preserved in modern genomes, is therefore invaluable for reconstructing the deep evolutionary history of microbial metabolisms. Various computational analyses and experimental validations can use this genetic information to map the relationships between extant organisms, estimate the age of key groups or metabolisms, and infer the phylogenetic position or biochemical capacity of ancestors. A general pipeline for phylogenetic approaches utilized in this work is shown in Fig. 1.2; the depicted programs and techniques are described in greater detail below.

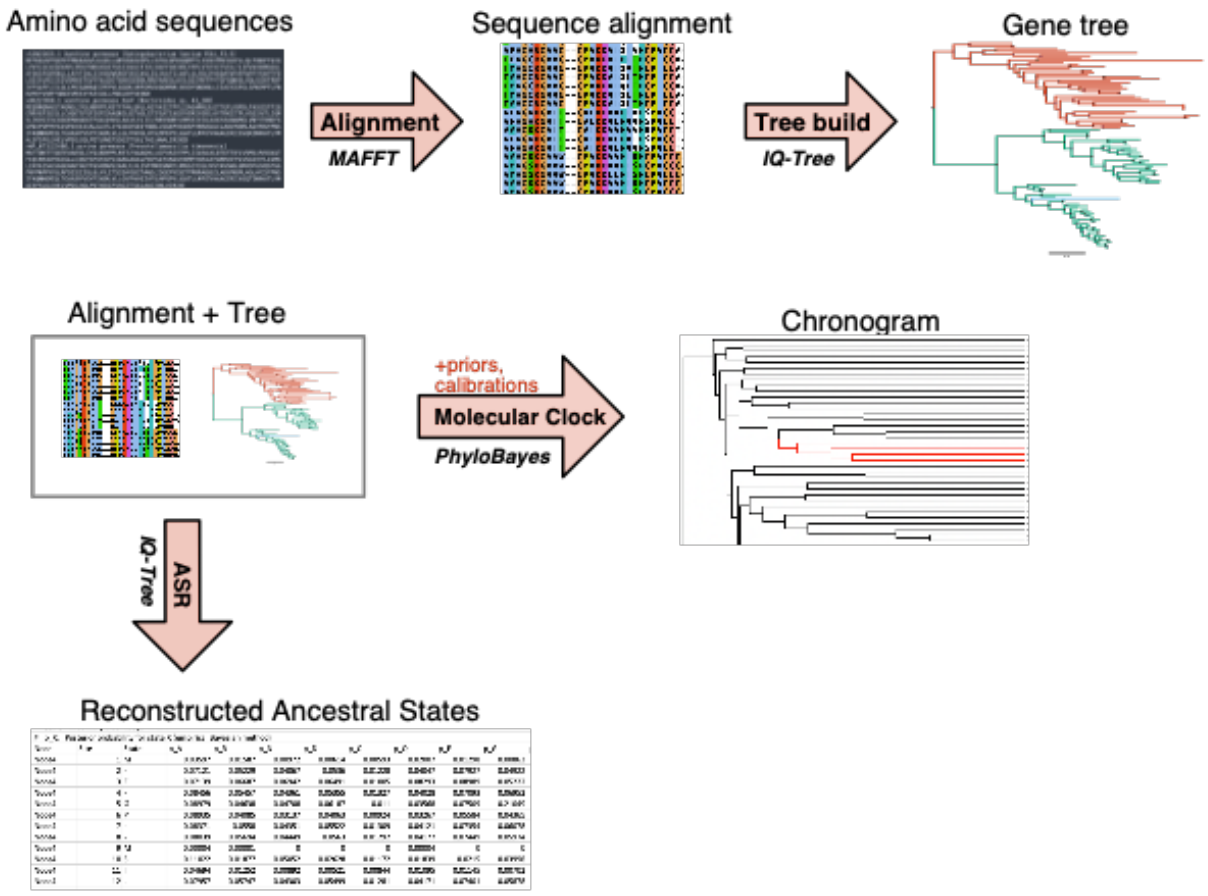


Figure 1.2 Simplified phylogenetics pipeline
The phylogenetic analyses in this work source target amino acid sequences from publicly available databases or sampled metagenomic data; align them using programs such as Muscle (Edgar, 2004) or MAFFT (Katoh *et al.*, 2002; Katoh and Toh, 2008); and construct maximum-likelihood gene trees in IQ-TREE (Nguyen *et al.*, 2015; Minh *et al.*, 2022). Alignments and trees may be used for further analyses, such as ancestral sequence reconstruction (ASR) or molecular clock dating using PhyloBayes (Si Quang, Gascuel and Lartillot, 2008).

1.3.1 Sequence alignment

Successfully mining genomic data depends upon accurate and efficient alignment of sequence data. Alignment provides an initial quantitative assessment of the genetic relationships between orthologs, an essential first step for homolog identification and subsequent tree construction (Edgar and Batzoglou, 2006). Either protein or DNA sequences can be aligned, but 20 amino acids provide increased specificity and diversity, and reduced noise, as compared with four nucleotides (Wernersson and Pedersen, 2003). The goal of sequence alignment is to identify and group true homologous sites between proteins, mapping conserved residues, mutations, gaps,

insertions, or deletions as accurately as possible (Edgar and Batzoglou, 2006). A wide variety of algorithms and programs can be used to generate the most likely alignment between a given set of sequences. Early sequence alignment heuristics such as the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970), Dayhoff estimator (Dayhoff, Schwartz and Orcutt, 1978), and Smith-Waterman algorithm informed various methods still in use today, including the pairwise alignment scoring system that powers the National Center for Biotechnology Information's (NCBI) Basic Local Alignment Search Tool (BLAST), a homology search program relied upon for countless bioinformatic pipelines (Altschul *et al.*, 1990). For alignment of selected orthologs, this work primarily uses two programs: MUSCLE and MAFFT.

MUSCLE uses sequence pairs with matching subsequences to build a distance-based tree, which then guides progressive, pairwise alignment of input sequences (Edgar, 2004). This initial alignment is used to refine a distance matrix and generate an improved guide tree and improved progressive alignment, which are iteratively improved by removing branches, realigning, and reassessing the subtree data (Edgar, 2004; Nuin, Wang and Tillier, 2006).

MAFFT's programming offers several alignment algorithms, but its core Fast Fourier Transform (FFT) strategies are very similar to Muscle's approach. The most common default approach, FFT-NS-2, refines an initial alignment and guide tree with a recalculated matrix and new progressive alignment; the FFT-NS-I strategy further refines this alignment by dividing it into groups and iteratively re-aligning to find a maximal alignment score (Katoh *et al.*, 2002; Nuin, Wang and Tillier, 2006; Katoh and Toh, 2008). These approaches are rapid, and FFT-NS-2 can be used effectively for large datasets (Katoh and Toh, 2008). For smaller-scale alignments, more-accurate iterative refinement techniques such as L-INS-i can be used; this strategy is refined by local pairwise alignments, weighted sequence pairing, and sequence motif scores (Nuin, Wang and Tillier, 2006; Katoh and Toh, 2008). MAFFT allows users to automate algorithm choice, selecting the optimal strategy to balance speed and accuracy based upon the amount of input data (Katoh and Toh, 2008).

Muscle and MAFFT (and especially MAFFT's most computationally expensive automated strategy, L-INS-i) both achieve generally high alignment accuracy with simulated datasets (Nuin, Wang and Tillier, 2006). But despite the accuracy and flexibility afforded by advanced alignment programs, such algorithms are necessarily agnostic to biological context and sometimes prone to overparameterization or inaccuracy (Philippe *et al.*, 2011). Additional programs can help "clean"

or "filter" noisy or gappy alignments (Di Franco *et al.*, 2019)—but such programs can over-curate, removing potentially informative sites. To identify alignment errors, minimize noise without excluding signal, and impose biological insight, manual curation of alignments is an important step before tree construction. Chapter 3 shows one such example of manual alignment curation where automated alignment cleaning programs failed.

1.3.2 Phylogenetic tree construction

Phylogenies are the cornerstone of molecular evolution. Gene trees map the evolutionary relationships between orthologs in different taxa, displaying patterns of protein diversity and inheritance. Species trees map the evolutionary relationships between organisms by concatenating several highly conserved, slowly evolving genes that collectively serve as proxies for the history of cell divisions that constitute a lineage of "vertical inheritance" (Woese, 2002). Where the topologies of gene and species trees are consistent, deep patterns of vertical inheritance appear, further clarifying genome histories across deep time. Where gene and species tree topologies diverge, horizontal gene transfers (HGT) can be inferred. HGT allows the exchange of genetic information between phylogenetically disparate groups, exchanging functionality across millions or billions of years of evolutionary distance (Gogarten and Doolittle, 2007; Fournier, Andam and Gogarten, 2015; Ravenhall *et al.*, 2015). Converting a reliable alignment into an accurate tree requires an appropriate evolutionary rate model and a method of identifying the most likely tree topology.

1.3.2.1 Evolutionary rate models

Several evolutionary rate models exist for estimating the substitution probabilities for individual amino acids in an alignment. These models incorporate data about the physicochemical properties and groupings of amino acids or empirically observed sequence substitutions in large datasets to calculate the likelihood of observing a given change within sequence data (Le and Gascuel, 2008; Pearson, 2013). The complexity and utility of available substitution matrices varies widely. Programs such as ProtTest (Darriba *et al.*, 2011) or IQ-TREE's ModelFinder (Kalyaanamoorthy *et al.*, 2017) can be used to efficiently evaluate various rate models and identify the optimal approach for a given alignment. Model-fitting analyses were performed for all trees calculated in this work, and the WAG (Whelan and Goldman, 2001) or more commonly, LG (Le

and Gascuel, 2008) models were consistently identified as the best fit for the data. These models incorporate approximate likelihood estimates for substitutions, based upon large protein databases; LG is trained upon a larger number and diversity of protein sequences than WAG, and also allows variability of evolutionary rate for different alignment sites (Whelan and Goldman, 2001; Le and Gascuel, 2008).

1.3.2.2 Maximum likelihood approaches

This work uses maximum likelihood (ML) methods for reconstructing tree topologies. An ML approach calculates, for a given tree topology, the likelihood of observing amino acid residues at each aligned site, based on the previously established evolutionary model (with evolutionary rate assumed to be a site-independent Markov process). Tree likelihood is calculated as the product of all site likelihoods. The optimal tree will have a combination of topology and branch lengths that maximizes the overall likelihood, given the observed sequence data (Felsenstein, 2004).

Searching for this optimal tree is a challenging problem. Various methods for rearranging branches have historically been used to gradually identify improved topologies—for example, splitting a tree into subtrees, and then swapping the sister placement of the trees across an internal branch (nearest neighbor interchange or NNI) (Guindon and Gascuel, 2003), or "breaking off" various subtrees and reattaching them in other branches of the tree (subtree pruning and regrafting) (Felsenstein, 2004). These "hill-climbing" approaches are valuable, but vulnerable to getting stuck in local likelihood optima (Nguyen *et al.*, 2015). That is, these subtree swaps only allow identification of incrementally improved topologies, and reject any topological changes that lower likelihood—even if such topological arrangements would be part of a higher global likelihood when combined with other branch rearrangements.

Stochastic algorithms that attempt to escape local maxima by permitting lower-likelihood tree candidates or exchanging possible topologies between multiple populations of trees have been employed in more recent ML tree construction programs (Helaers and Milinkovitch, 2010; Nguyen *et al.*, 2015). Tree construction in this work is primarily performed with the program IQ-TREE, which combines hill-climbing NNI with stochastic NNI to quickly identify high-likelihood topologies while avoiding local likelihood maxima (Fig. 1.3). Initial analyses suggested that this program produces higher-likelihood trees more often than similar ML programs, with similar runtimes (Nguyen *et al.*, 2015).

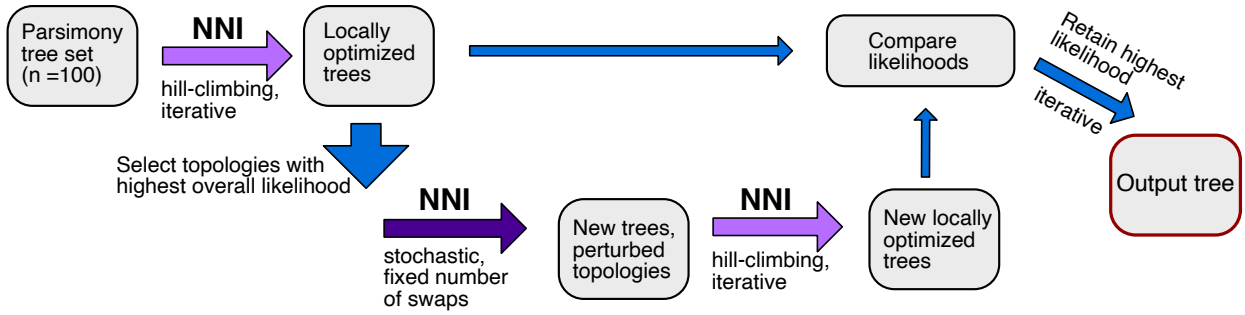


Figure 1.3 Simplified view of IQ-TREE search strategy
The ML tree search program IQ-TREE combines combined hill-climbing and stochastic NNI searches to efficiently identify optimal topologies while escaping potential local likelihood optima (Nguyen *et al.*, 2015). An initial set of trees is constructed with simple parsimony methods; iterative hill-climbing NNI is performed on this set, and the five highest-likelihood topologies retained; these trees are then subjected to a branch-dependent number of random NNIs, and the resulting topologies are used for another round of iterative hill-climbing NNI. The new optimized trees are iteratively compared with the highest-likelihood trees from the initial NNI, with the highest-likelihood trees retained until an optimal output tree is identified.

1.3.2.3 Calculating statistical support

The ML tree search processes described above will necessarily include degrees of uncertainty about the placement of individual bipartitions in a tree. Programs such as IQ-TREE allow for statistical inference of the robustness of ancestral splits using bootstrapping or likelihood ratio testing. Phylogenies in this work are assessed using bootstrap and approximate likelihood ratio test methods.

In general, bootstrapping methods estimate an unknown distribution using empirical subsampling: Individual data points from a dataset are sampled with replacement, and the samples used to represent variation within the true distribution. For phylogenies, standard bootstrapping would involve sampling sites from the original alignment to build bootstrapped alignments, which are then each used to reconstruct a new ML tree (Felsenstein, 2004). A bootstrap value reports the percentage of replicates in which a specific bipartition was observed; so the higher the value, the greater the statistical confidence in that split. Standard bootstrapping is computationally expensive and extremely time consuming, so approximate bootstrap methods are often employed instead. IQ-TREE's ultrafast bootstrap approach retains and resamples the topologies and likelihoods calculated during the initial tree search for the original alignment, instead of reconstructing

bootstrapped alignments *de novo*, and compares calculated likelihoods to identify more-realistic bootstrapped trees (Minh, Nguyen and Von Haeseler, 2013; Hoang *et al.*, 2018).

In addition to ultrafast bootstraps, gene trees in this work were evaluated with Shimodaira-Hasegawa approximate likelihood ratio testing (SH-aLRT) (Guindon *et al.*, 2010). As in ultrafast bootstrapping, SH-aLRT resamples previously calculated likelihoods; however, this test compares the likelihoods of alternative arrangements around a given branch (Anisimova and Gascuel, 2006), indicating how frequently a given bipartition increases the likelihood of the tree. Using both metrics provides additional context for interpreting overall confidence in clade placement within a tree.

Identifying these high support values is most important for inferring either deep splits or the placement of functionally distinct groups. Shallower splits, representing divergences between more similar sequences that diverged more recently, are often more likely to have uncertain placement and low-confidence support values, as fewer substitutions have potentially occurred to recover the bipartition in question. (Fortunately, these splits are also far less likely to represent biologically significant evolutionary hypotheses for the work being presented here.)

1.3.2.4 Rooting

Understanding the position of the deepest split in a phylogenetic tree—the root—is critically important for inferring evolutionary relationships. ML trees are usually calculated as unrooted phylogenies, with the position of the deepest split in the tree unfixed and disconnected from the tree's likelihood calculation (Williams *et al.*, 2015). There is more than one option for inferring a root. Commonly, a group of sequences known to be homologous to the sequences of interest (but more closely related to each other than to any sequence of interest) is included as an outgroup. The split between the outgroup and ingroup then represents the root, the last common ancestor of all ingroup and outgroup sequences (Felsenstein, 2004).

Correctly identifying an outgroup requires some previous knowledge or inference about the monophyly and homology of taxonomic groups (for species trees) or protein subfamilies (for gene trees). In the absence of this information—for example, if ingroup sequences are too sparse or divergent to clearly identify an outgroup—quantitative approaches can be used to estimate a root based upon overall genetic distances within the tree. One approach is to assume that evolutionary rates are consistent across lineages and place the root in a position that makes the

overall genetic change equal between clades (Felsenstein, 2004). This "midpoint" rooting is efficient, but not biologically accurate, as evolutionary rates empirically vary between lineages. In this work, when biologically supported outgroups are unavailable, trees are rooted using Minimal Ancestor Deviation (MAD), which accounts for rate heterogeneity within a tree (Tria, Landan and Dagan, 2017). This method tests each branch in a tree as a possible root and calculates the average absolute deviation between the midpoint of the branch and each tip in the tree. Incorrect rootings will show large deviations, as they force rate-heterogeneous groups together. The evaluated rooting that minimizes overall ancestor rate deviation is chosen as the best root for the tree (Tria, Landan and Dagan, 2017).

1.3.3 Molecular clock dating

1.3.3.1 Bayesian molecular clock approaches

Substitutions along the branches of a tree can be used to not only map evolutionary relationships, but also to mark the passage of time. A molecular clock approach uses evolutionary rate models to correlate genetic and temporal distances. Assuming a clocklike behavior for evolutionary change qualitatively informs the rooting analyses described above. But it can also be used in concert with dated fossils or biomarkers to quantitatively constrain the age of ancestral nodes in a tree. Molecular clocks can extrapolate from extant data and constraints to infer the timing of key geobiological events, such as the emergence or diversification of a biochemically relevant clade, even with limited or no direct geological record of these events. As such, these tools are key for testing hypotheses about ancient biogeochemistry that are unconstrained or indistinguishable by geochemical inference or modeling alone.

Converting a tree of evolutionary distances into a tree of time (a chronogram) requires a reliable phylogeny; a clock model for estimating evolutionary rate distributions; useful calibrations to constrain the distribution of time estimates; and, as with initial ML tree building, a method for identifying the chronograms most likely to generate the observed sequence data. In this work, molecular clocks are calculated with a Bayesian approach. Bayesian likelihood approaches are similar to the ML approaches described above; in fact, ML calculations depend on Bayes' theorem to assess the likelihood of observing empirical data underneath a given hypothesis. The primary difference between Bayesian and strictly ML likelihood approaches is that Bayesian approaches compute a prior distribution of possible trees (Felsenstein, 2004;

Alfaro and Holder, 2006). For molecular clocks, this "best guess" prior distribution incorporates any available information about permitted tree topologies, input calibrations on root and clade ages, and branch rates (Bromham *et al.*, 2018). This prior distribution is then updated using empirical sequence data to build a posterior distribution. However, observed substitutions in the sequence alignment will have different likelihoods under different rates, so sequence data provides additional information about evolutionary processes. Prior distributions are updated using empirical sequence data to build a relaxed molecular clock posterior, which is then sampled to identify the highest-probability chronograms.

Calculating the sum of a full posterior distribution directly is computationally untenable, so posteriors are instead estimated using Markov chain Monte Carlo (MCMC) methods. MCMC randomly samples trees from the posterior distribution and uses probabilities of observing branch or clade placements within this subset to represent probabilities within the full posterior distribution (Felsenstein, 2004; Alfaro and Holder, 2006; Bromham *et al.*, 2018).

In this work, the program PhyloBayes was utilized for relaxed molecular clock construction. PhyloBayes uses an MCMC Metropolis algorithm, which samples the posterior distribution, comparing the probabilities of two possible trees at a time. If a newly considered topology, branch length, or distribution parameter has higher probability than the previous tree, it replaces this tree in the next comparison. If a newly considered tree has lower probability than the previous tree, it is either rejected or accepted stochastically, thereby adding an element of "chance" to avoid getting stuck in local probability maxima (Felsenstein, 2004; Bromham *et al.*, 2018). Multiple chains with different thresholds for accepting new trees are used to sample posterior trees at the same time; when these two chains converge on a similar topology, this is considered the highest-probability solution in the posterior.

1.3.3.2 Root prior

As mentioned above, the prior distribution is constrained by various inputs that must be defined before Bayesian likelihood searching begins. Without prior age constraints, molecular clocks can only provide relative age estimates between groups; assigning concrete ages requires some prior age inputs (Bromham *et al.*, 2018). The first of these is the root prior, which delineates the distribution of ages permitted for the root, the deepest and therefore oldest branch in the tree. This root prior may weight all possible ages in its range equally, or may use a distribution that

weights older, younger, or intermediate ages as more likely (Fig. 1.4) (Bromham *et al.*, 2018); for example, clocks in this work use a normally-distributed root prior.

It is extremely important that the root prior is set as broad as necessary to permit the age of all alternative hypotheses under consideration. If the bounds of the root prior are set too old or too young to accommodate the (unknown) true age, the rate and age estimates for descendants will all be erroneous. Because of this, posterior age estimates should always be considered in the context of the chosen prior; if age estimates closely approach either the older or younger bound, the clock should be rerun with a more permissive prior to ensure that the prior is not constraining output ages (see Chapter 3). More permissive priors generate less precise age estimates, but this is preferable to inaccurate age estimates—especially as additional calibrations (see below) can help compensate for a particularly uninformative root prior.

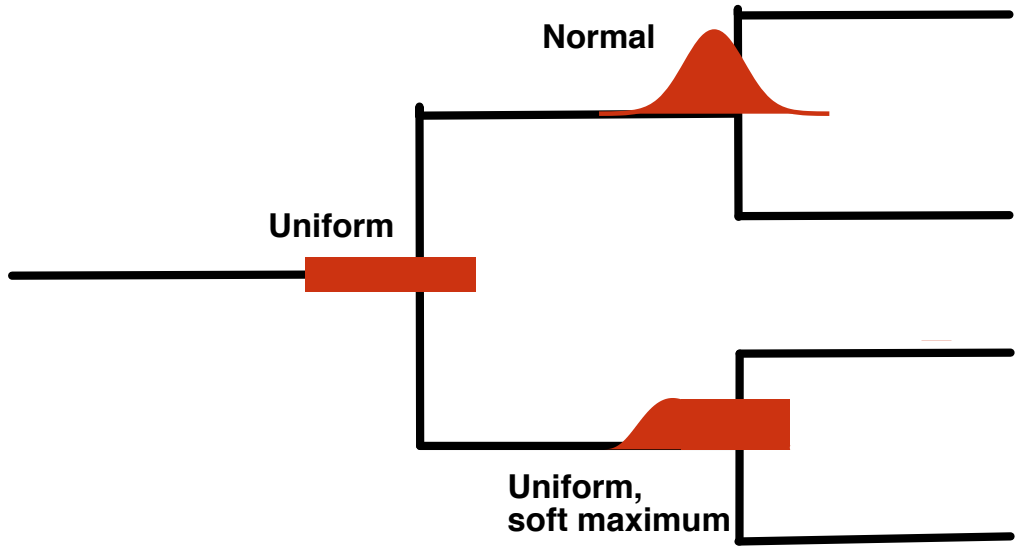


Figure 1.4 Example calibration density distributions
Age calibrations for root or clade priors are applied as probability densities at the constrained node. These distributions may be uniform, representing even likelihood of ages, or may be modeled as a parametric distribution, such as a normal distribution. Older and younger age constraints may independently be represented by fixed hard bounds or distributed soft bounds.

1.3.3.3 Calibrations

Beyond the required root prior, constraining the ages of other bipartitions in the tree further calibrates rate and age estimates, producing more precise and accurate dates for groups of interest. Internal calibrations are usually derived from geological or geochemical data, such as fossil ages. The age of a single fossil would represent a younger age bound for that organism's clade: Evidence of an organism is also evidence that a lineage preceded its preservation. For well-preserved groups, multiple fossils can be used to estimate older and younger age distributions across taxonomic levels (Nascimento, Reis and Yang, 2017). For example, within Fungi, a robust fossil record provides age estimates for the ancestor of Phylum Ascomycota (~680-410 Ma), but also the ancestor of Classes Sordariomycetes, Dothideomycetes, and Leotiomycetes (~350-250 Ma) (Prieto and Wedin, 2013).

Given the limitations of ancient and microbial fossil records (see above), fossil calibrations from younger, well-preserved eukaryotic groups are extremely useful as clock calibrations. Gene trees with transfers between prokaryotic groups and well-characterized plants, animals, or fungi are therefore particularly good candidates for dating. Applying as many diverse calibrations as possible serves to maximize precision, but also to check the influence of each individual calibration on the overall age estimates (see Chapter 3).

1.3.3.4 Clock models

In addition to age range priors, branch rate priors are also assigned for Bayesian clock analyses. There are a variety of these "clock models" that can be applied, varying in which rates may be sampled, and whether the rates of various branches are linked or independent. At one extreme, "strict" clock models enforce a single evolutionary rate on all branches of a tree. Intermediate "discrete" or "local" clocks allow branches or groups to select from a limited number of rate categories. The most flexible "relaxed" clock models allow each branch in the phylogeny to have an independent rate (Bromham *et al.*, 2018). Relaxed models are more computationally intensive and complex than strict or discrete models, but improve accuracy for trees containing diverse evolutionary strategies. This makes relaxed clocks the best choice for microbial phylogenies, in which lineages evolve quickly and over great timescales, and HGT occurs over varying genetic distances.

Molecular clocks in this work were run under three relaxed clock models: uncorrelated gamma (UGAM), lognormal (LN), and Cox-Ingersoll-Ross process (CIR). As its name implies, UGAM (Drummond *et al.*, 2006) is an uncorrelated model, meaning that rates for all branches and groups in the tree are estimated independently of rates in sister groups. LN (Thorne, Kishino and Painter, 1998; Kishino, Thorne and Bruno, 2001) and CIR (Lepage *et al.*, 2006), alternatively, are autocorrelated. Autocorrelated models assume that closely related groups will share more similar evolutionary rates, possibly reflecting patterns of vertical inheritance, population ecology, or environmental selection (Bromham *et al.*, 2018). As a result, LN and CIR rate calculations incorporate terms informed by parent branch rates (Bromham *et al.*, 2018). Both models assume that rate variation is stochastic. LN samples variable rates from a lognormal distribution, and CIR samples rates with a Brownian motion, but incorporates a "correction" factor that drives estimated rate values back towards an average rate calculated using the parent branch (Lepage *et al.*, 2007). This "spring-like" adjustment controls for the risk of stochastically sampling extreme, unrealistic rates (Bromham *et al.*, 2018).

The assumptions underlying all rate models will affect output age estimates differently, and to an increasing degree if data or calibrations are noisy or less informative. Because of this, testing multiple rate models on the same dataset is important to understand which biases, if any, result from model choice (see Chapter 3).

Together, clock model and calibration inputs can be used to generate a probability distribution of ages for each bipartition within the tree, including a "highest probability" age across sampled rates. As discussed in section 1.3.3.1, both prior and posterior age distributions are generated; the former includes only the priors discussed above (topological seed, root prior, calibrations, evolutionary model), while the latter updates this prior distribution using empirical sequence data (Bromham *et al.*, 2018).

1.3.4 Ancestral sequence reconstruction

Molecular clocks allow inference of the age of ancestral nodes in the tree; it is also possible to infer the sequence *state* of these ancestors. Ancestral sequence reconstruction (ASR) uses extant alignment and tree data to estimate the most likely residue at each alignment site for the internal nodes in a tree. The probability of seeing a given amino acid at an internal bipartition can be back-calculated based on the distribution and diversity of extant residues (Merkl and Sterner, 2016;

Selberg, Gaucher and Liberles, 2021). The process of calculating these likelihoods is akin or identical to calculating root or topology likelihood (see above), and previous internal ML calculations can be used for ASR (Felsenstein, 2004; Merkl and Sterner, 2016). For example, if a tree is rooted at a given node, likelihoods are calculated for amino acids at all descendant sites to determine tree likelihood; the most significant residue contributing to overall likelihood will be the most likely state at the ancestral node (Felsenstein, 2004). In this work, ASR was computed in IQ-TREE, which uses empirical Bayesian methods (Yang, Wong and Nielsen, 2005; Minh *et al.*, 2022) to quantify the probability contributions from each possible amino acid at each ancestral node, and estimate the residue with maximum probability (Merkl and Sterner, 2016).

ASR is commonly used as a first step in ancestral enzyme resurrection in the lab (Garcia and Kaçar, 2019; Selberg, Gaucher and Liberles, 2021). It is highly unlikely that the combined ASR across sites will represent the true full sequence of an ancestral gene, for various reasons. ASR evaluates site likelihoods independently, but regions of a gene may co-evolve, making assumptions of independence inaccurate; multiple residues may have similarly high probability estimates, making it difficult to infer a significantly most-likely amino acid; and the ASR probability calculations will have their own inherent uncertainty at all sites. But an ASR sequence can be considered a "consensus sequence" for the overall ancestral likelihood sampled by the alignment and tree (Garcia and Kaçar, 2019). Hypotheses about the state or physicochemistry of individual sites require even fewer assumptions, and may be more tractable; for example, the identity or physicochemistry of sites involved in substrate docking can be directly interrogated, even completely *in silico* (see Chapter 4).

1.4 Thesis overview

The projects within this thesis combine the phylogenetic approaches described above to interrogate the evolutionary history of microbial nitrogen metabolisms that are not directly considered part of the modern canonical nitrogen cycle, but which might have played an important role in the evolutionary history of the nitrogen cycle and life on Earth (Fig. 1.1).

Chapter 2 explores the distribution and transfer history of unusual denitrification genes, especially within Chloroflexi. While denitrification is a core pathway in the modern nitrogen cycle (see section 1.1), this work explores novel or understudied enzyme variants that have not been considered within the broader framework of *in situ* denitrification. I show that some Chloroflexi

contain a previously undescribed nitrite reductase (*nirS*) gene; domain trees reveal that this *nirS* variant is the result of a chimeric fusion of a horizontally transferred cytochrome. I also provide the first comprehensive phylogeny of the rare nitric oxide reductase subtype *eNOR,* revealing a surprisingly broad prokaryotic distribution.

Chapter 3 considers the question of noncanonical metabolisms beyond the confines of the modern nitrogen cycle, specifically focusing on the evolution of microbial cyanide reduction. This process is not considered to be a key player in modern nitrogen biogeochemistry but may have played a role in ancient nitrogen cycling, as hydrogen cyanide was more prominent on the early Earth and is a potential feedstock for early biochemistry. I use gene trees to map the distribution of the Class I nitrilases, an enzyme subfamily specializing in free cyanide metabolism. The trees reveal that these enzymes originated in prokaryotes and were subsequently transferred into plants, animals, and fungi. I also use molecular clocks to constrain the emergence of these enzymes to the Meso- or Paleoproterozoic, suggesting that this nitrogen biochemistry was present and could have played a biogeochemical role in the context of a different, ancestral nitrogen cycle.

Chapter 4 considers off-target metabolism in nitrogenases as a modern signature of historical nitrogen cycling. Extant N_2-fixing nitrogenases are central to the canonical nitrogen cycle, but these enzymes can promiscuously reduce other linear nitrogenous substrates—including hydrogen cyanide. As nitrogenases are particularly ancient enzymes, and hydrogen cyanide was enriched on the early Earth, this promiscuity could reflect ancestral nitrogen biogeochemistry. To explore the possibility of this hypothesis, I use ancestral sequence reconstruction to infer the most-likely states of key sites in ancestral nitrogenase substrate channels. The data reveal that combinations of the most probable ancestral residues are rarely sampled in extant nitrogenase sequence space. The results suggest that ancestral nitrogenases did have alternative active site geometry and indicate no obvious substitutions that would preclude the promiscuous cyanide reduction observed in modern enzymes.

Chapter 2. Divergent denitrification genes in Phylum Chloroflexi

The contents of this chapter were originally published as: Schwartz, S.L.; Momper, L.; Rangel, L.T.; Magnabosco, C.; Amend, J.P.; Fournier, G.P. "Novel nitrite reductase domain structure suggests a chimeric denitrification repertoire in Phylum Chloroflexi." MicrobiologyOpen, 11(1): e1258 (2022).

2.1 Abstract

Denitrification plays a central role in the global nitrogen cycle, reducing and removing nitrogen from marine and terrestrial ecosystems. The flux of nitrogen species through this pathway has widespread impact, affecting ecological carrying capacity, agriculture, and climate. Nitrite reductase (Nir) and nitric oxide reductase (NOR) are the two central enzymes in this pathway. Here we present a previously unreported Nir domain architecture in members of Phylum Chloroflexi. Phylogenetic analyses of protein domains within Nir indicate that an ancestral horizontal transfer and fusion event produced this chimeric domain architecture. We also report an expanded genomic diversity of a rarely reported nitric oxide reductase subtype, eNOR. Together, these results suggest a greater diversity of denitrification enzyme arrangements exist than have been previously reported.

2.2 Introduction

Microbial denitrification is a key pathway in global nitrogen cycling and has been studied extensively for its role in fixed nitrogen loss and as a source of potent greenhouse gases (Zumft, 1997; Decleyre *et al.*, 2016). Diverse bacteria are capable of denitrification, often facultatively using nitrate or nitrite as an alternative electron acceptor in oxygen-limited zones. Several diverse microorganisms have the genomic capacity to perform complete denitrification (Fig. 2.1), reducing nitrate to dinitrogen gas (Philippot, 2002; Canfield, Glazer and Falkowski, 2010).

Denitrification has been widely reported in various taxa (Zumft, 1997; Philippot, 2002), and the utility of the pathway is underscored by the diversity of key constituent enzymes. The canonical denitrification enzyme is dissimilatory nitrite reductase, Nir, which reduces nitrite to nitric oxide.

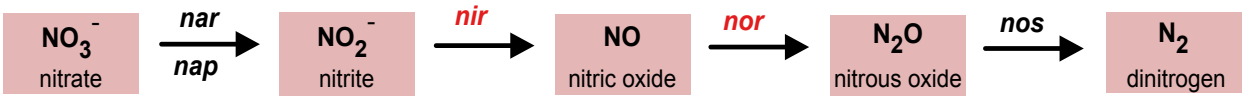


Figure 2.1: Denitrification
Complete denitrification transforms nitrate to dinitrogen gas. Respiratory nitrate reductase gene (*nar*) and periplasmic nitrate reductase (*nap*) genes are distributed in non-denitrifying organisms. Nitrite reductase (Nir) is considered the canonical first enzyme of denitrification (Graf, Jones and Hallin, 2014), followed by nitric oxide reductase (NOR) and nitrous oxide reductase (Nos).

Nir functionality is found in two distinct enzymes—the copper-based nitrite reductase NirK, and the cytochrome-type reductase NirS (Braker *et al.*, 2000; Priemé, Braker and Tiedje, 2002; Decleyre *et al.*, 2016). NirS reduces nitrite via cytochrome *cd1*, a dimer of heme *c* and heme *d1* subunits; in canonical denitrification, as observed in *Pseudomonas aeruginosa*, cytochrome *cd1* catalyzes oxidation of a colocalized cytochrome c551 to reduce nitrite to nitric oxide at the heme *d1* site (Zumft, 1997; Philippot, 2002). The *nirS* gene has been reported as a constituent of a larger gene cluster containing genes such as *nirM* and *nirF,* which encode biosynthetic proteins for the cytochrome *c*551 and heme d_1, respectively, and the nitrite transporter *nirC* (Kawasaki *et al.*, 1997; Philippot, 2002).

The next step in the pathway—the reduction of nitric oxide to nitrous oxide—is catalyzed by nitric oxide reductases (NORs). Most bacterial NORs are homologous and closely related to one another, and to oxygen reductases in the heme-copper oxygen reductase superfamily (Hemp and Gennis, 2008). The most widely studied NOR enzymes are cNOR and qNOR (Hendriks *et al.*, 2000; Hemp and Gennis, 2008; Graf, Jones and Hallin, 2014), distinguished by their respective electron donors. Enzymes in the cNOR subfamily have two subunits—one catalytic site and one heme-containing electron shuttle that accepts electrons from cytochrome *c*—while qNOR family enzymes' single, fused subunit accepts electrons from membrane-bound quinol groups (Hemp and Gennis, 2008). Rarer, alternative NOR enzymes, including sNOR, gNOR and eNOR, have been more recently identified and characterized in limited members of the Proteobacteria, Firmicutes, Archaea and Chloroflexi (Stein *et al.*, 2007; Hemp and Gennis, 2008; Sievert *et al.*, 2008; Hemp *et al.*, 2015). Like cNOR, these enzymes are predicted to have a two-subunit structure, but the second subunit in these NORs contains a cupredoxin instead of heme *c* fold (Hemp and Gennis, 2008).

Many bacteria contain genes encoding only one or a partial subset of the four denitrification steps. Such organisms may perform partial denitrification, while others may use one of these enzymes for non-denitrifying functions (Hendriks *et al.*, 2000; Sanford *et al.*, 2012; Graf, Jones and Hallin, 2014; Roco *et al.*, 2017). In partial denitrifers, the co-occurrence of denitrification pathway genes appears to vary across different taxa and environments (Graf, Jones and Hallin, 2014). Some of this variation may be constrained by the chemistry of certain intermediates. For example, nitric oxide (NO), the product of NirS and NirK, is highly cytotoxic. Both Nir types are periplasmic, and so cells require a means of effluxing or detoxifying nitric oxide before it accumulates to lethal levels. Denitrifiers are thought to immediately reduce nitric oxide to nitrous oxide to avoid injury, using membrane-bound NOR enzymes (Hendriks *et al.*, 2000). Perhaps for this reason, it is rare to find genomes that contain *nir* but not *nor*, while organisms showing the inverse—the presence of a *nor* gene but not a *nir* gene—are far more common (Hendriks *et al.*, 2000; Graf, Jones and Hallin, 2014). While cNORs are only found in denitrifying microbes, other types of NOR—for example, quinol-dependent qNOR—are found in non-denitrifiers and can presumably detoxify environmental nitric oxides (Hendriks *et al.*, 2000). Beyond NORs, alternative pathways to nitric oxide detoxification are possible, including alternative enzymes such as cytochrome c oxidase (Blomberg and Ädelroth, 2018) or oxidoreductase (Gardner, Helmick and Gardner, 2002), flavorubredoxin (Gardner, Helmick and Gardner, 2002), or flavohemoglobins (Sánchez *et al.*, 2011).

While denitrification has been most widely studied and observed in Proteobacteria, the process has also been identified in other phyla, including Chloroflexi. Chloroflexi are ecologically and physiologically diverse, and often key players in oxygen-, nutrient-, and light-limited environments, including anaerobic sludge and subsurface sediments (Hug *et al.*, 2013; Ward *et al.*, 2018). Previous surveys have indicated the presence of diverse nitrite reductases in Chloroflexi; members of Order Anaerolineales and Classes Chloroflexia and Thermomicrobia may have the capacity for nitrite reduction via the copper-type NirK (Hug *et al.*, 2013; Wei *et al.*, 2015; Decleyre *et al.*, 2016). However, recent studies indicate that certain Chloroflexi—including members of Anaerolineales—may possess *nirS* instead of *nirK* (Hemp *et al.*, 2015; Ward, McGlynn and Fischer, 2018), and may also harbor a divergent variant of *nor* previously reported in members of Archaea (Hemp and Gennis, 2008; Hemp *et al.*, 2015). These findings suggest that the evolution

and/or biochemistry of denitrification may be unusual for this subset of bacteria, and informative for a broader understanding of microbial denitrification metabolisms and their origin.

2.3 Methods

2.3.1 Genome sampling and assembly

Collection of all fluid samples and total genomic DNA extractions from those fluids, as well as corresponding physical and geochemical data have been described previously (Lau *et al.*, 2014, 2016; Osburn *et al.*, 2014; Magnabosco *et al.*, 2016; Heard *et al.*, 2017; Momper, Jungbluth, *et al.*, 2017; Momper, Kiel Reese, *et al.*, 2017). All MAGs from North America and Africa were reconstructed according to the methods used in (Momper, Jungbluth, *et al.*, 2017). MAG identifiers and sources are listed in Table 2.S.2. Completeness was calculated using the composite values from five widely accepted core essential gene metrics. Duplicate copies of any of these single copy marker genes was interpreted as a measure of contamination (Creevey *et al.*, 2011; Dupont *et al.*, 2012; Wu and Scott, 2012; Campbell *et al.*, 2013; Alneberg *et al.*, 2014). Individual genomes were then submitted for gene calling and annotations through the DOE Joint Genome Institute IMG-ER (Integrated Microbial Genomes expert review) pipeline (Markowitz *et al.*, 2008; Huntemann *et al.*, 2015). For quality control purposes, the genes flanking every denitrification gene presented in this study were individually searched on the NCBI RefSeq database using the BLASTp algorithm, confirming that top hits for all flanking genes were also to Chloroflexi. This step ensured that the nitrogen transforming genes of interest presented here were not simply on scaffolds that were incorrectly binned into a putative Chloroflexi genome.

2.3.2 Genetic database construction and sequence sampling

Sequences for *nirS* and *eNOR* genes from SURF MAG 42 (Table 2.S.1) were used as queries to BLAST (Camacho *et al.*, 2009) three genomic repositories:

1) Genome databases constructed for 21 Chloroflexi genomes assembled from deep-subsurface MAG data (Jungbluth, Amend and Rappé, 2017; Momper, Jungbluth, *et al.*, 2017) (Table 2.S.2)
2) Genome databases constructed for 86 genomes from recent MAG assembled sludge bioreactor genomes (Parks *et al.*, 2017) (Table 2.S.3)

3) The full NCBI non-redundant protein database as of 25 September, 2019 (Agarwala *et al.*, 2018)

Additionally, putative environmental homologs were evaluated using protein sequence data from SURF MAG 42 to query NCBI's non-redundant environmental metagenomic sequence database (env-nr) as of June 2020 (Agarwala *et al.*, 2018).

Hits from all databases were combined and assessed for quality; hits with e-value ≤ $1x10^{-10}$ were included for initial analyses. To capture diversity while limiting imprecision and biased sampling of overrepresented groups (e.g., Proteobacteria), hits were subsampled to the genus level, except for members of the Chloroflexi (to fully capture the taxonomic distribution of the novel gene variant). One additional, divergent multispecies hit was allowed per genus. The genus-level filter was also removed for C1, where non-Chloroflexi hits were severely limited. Duplicate sequences (from strains with multiple genome entries or in multiple databases surveyed) were removed.

2.3.3 Sequence Alignment

Putative homologous protein sequences were aligned with MAFFT, using auto-parameterization (Nakamura *et al.*, 2018), and visualized in Jalview (Waterhouse *et al.*, 2009). Alignments were manually curated; partial sequences with substantial missing regions or anomalous insertions in conserved regions of the protein were removed to avoid confounding phylogenetic analyses and evolutionary model selection. Protein sequence alignments were trimmed to the length of individual domains identified by NCBI's conserved domains database. Each domain was then re-aligned.

2.3.3.1 eNOR

A preliminary alignment for the *eNOR* gene showed a poorly conserved region near the C-terminal end of the ORF; to improve accuracy and avoid misalignment, this region was manually removed, and the remaining sequences were realigned prior to tree construction. Two sequences (Actinobacteria bacterium RBG_16_68_12, OFW73639.1, and *Thermus* WP_015717644.1) with missing N-terminal regions and three sequences (Chloroflexi bacterium, RME47896.1; Rhodocyclaceae bacterium UTPRO2, OQY7467.1; and *Rhodothermus profundi*, WP_072715415.1) with missing C-terminal regions were included in the final alignment; the

placement of these sequences is therefore based upon fewer alignment sites than other taxa. All retain key active site residues and show no clear evidence of long branch attraction artifacts in the tree.

2.3.3.2 C1

An initial alignment for the C1 domain showed a poorly conserved N-terminal region. To improve accuracy, this region was manually removed, and the remaining sequences realigned prior to tree construction.

2.3.3.3 NirS

Because the *nirS* domain had a C-terminal placement in the ORF across hits, C-terminal sites extending beyond the identified *nirS* domain were included in the trimmed alignment.

2.3.4. Rooting and outgroup identification

2.3.4.1 eNOR

Ingroup *eNOR* subunit I sequences were identified by the presence of a conserved Gln residue in alignment position 333. This site distinguishes eNOR not only from other nitric oxide reductases, but also from members of the oxygen reductase superfamily, which have a conserved Tyr in this site that plays a role in cofactor crosslinking (Hemp and Gennis, 2008) (Table 2.S.5). Outgroup sequences (oxygen reductase superfamily or other divergent nitric oxide reductases) were subsampled to a single taxon representative per major subgroup observed in a preliminary tree. Retained outgroup sequences CCQ74688.1, WP_100277903.1, WP_097280063.1, WP_089728124.1, RLC59399.1, and WP_083704903.1 are annotated as uncharacterized domains. Remaining sequences were re-aligned before tree construction, and manually rooted on the branch leading to the outgroup.

2.3.4.2 C1

Due to a paucity of initial hits (17 total genera), the genus-level filter was removed for all phyla to increase resolution of the domain phylogeny. A preliminary tree (Fig. 2.S.6) was expanded to identify outgroup sequences by including hits with $E \leq 10^{-4}$. The resulting tree was rooted using minimal ancestor deviation (MAD) rooting (Tria, Landan and Dagan, 2017).

2.3.4.3 C2 and NirS

Sequences were rooted using minimal ancestor deviation (MAD) rooting (Tria, Landan and Dagan, 2017).

2.3.5 Tree construction

Maximum-likelihood trees were constructed using IQ-TREE (Nguyen *et al.*, 2015), under the optimal model defined by the ModelFinder (-MFP) command (Kalyaanamoorthy *et al.*, 2017) (Table 2.S.4). Ultrafast bootstraps and approximate likelihood ratio tests were performed using IQ-Tree's ultrafast bootstrap and Sh-aLRT parameters (Minh, Nguyen and Von Haeseler, 2013; Hoang *et al.*, 2018).

2.3.6 Gene and enzyme structural analysis

FIND (Murali *et al.*, 2019) was used to identify structural features and conserved denitrification pathway genes in deep subsurface genomes. Putative domains within denitrification gene ORFs were identified and compared across genomes using BLAST and NCBI's Conserved Domains Database (CDD) (Marchler-Bauer *et al.*, 2015; Lu *et al.*, 2020) and EMBL InterPro (Mitchell *et al.*, 2019). C1 from SURF MAG 42 was classified by CDD as COG4654 (e-value = 6.4×10^{-4}) and by hmmscan (Potter *et al.*, 2018) as cytochrome c superfamily (accession 46626) hit (e = 1.6×10^{-5}). C1 did not show a strong pfam match in hmmscan; the closest match was PF13442.8 (independent e = 0.18). C2 from SURF MAG 42 was classified by CDD as COG2010 (e-value=5.31×10^{-9}), and included an annotated region classified as pfam 13442 (e-value = 4.06×10^{-7}); C2 was identified in hmmscan as a cytochrome c superfamily (accession 46626) hit (e = 1.2×10^{-17}), with a pfam match to PF13442.8 (independent e-value = 2.8×10^{-10}). While clearly homologous, the C1 and C2 families appear distantly related and do not appear within the other's dataset of closely related sequences (see above).

Gene neighborhoods were visualized using Gene Graphics (Harrison, Crécy-Lagard and Zallot, 2018), using a 20,000 base pair region. Gene and domains identified in each neighborhood were sourced and cross-referenced with NCBI's RefSeq and CDD (Marchler-Bauer *et al.*, 2015; O'Leary *et al.*, 2016).

Existing enzyme structures for canonical denitrification genes were downloaded from the RCSB Protein Data Bank (PDB) (Berman *et al.*, 2000). Anaerolineales-type enzyme structures were predicted using SWISS-MODEL (Waterhouse *et al.*, 2018). All enzyme structures were visualized and analyzed in PyMOL (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) SWISS-MODEL outputs are available as supplementary datafiles.

2.4 Results

2.4.1 Apparent novel domain architecture in Chloroflexi NirS

To investigate divergent denitrification genes in Chloroflexi, we performed a comprehensive analysis of denitrification homologs in over 100 recently sequenced Chloroflexi metagenome-assembled genomes (MAGs), as well as previously available genomes and metagenomes from the National Center for Biotechnology Information's (NCBI) protein databases. Domains of interest were initially identified in SURF MAG 42, an Anaerolineales bacterium sampled in the Sanford Underground Research Facility (SURF), a former gold mine in South Dakota (Momper, Jungbluth, *et al.*, 2017). Domain analysis of the Anaerolineales-type nitrite reductase open reading frame (ORF) from SURF MAG 42 indicated three putative functional regions of interest: one cytochrome-type NirS domain and two cytochrome C superfamily domains (Fig. 2.2).

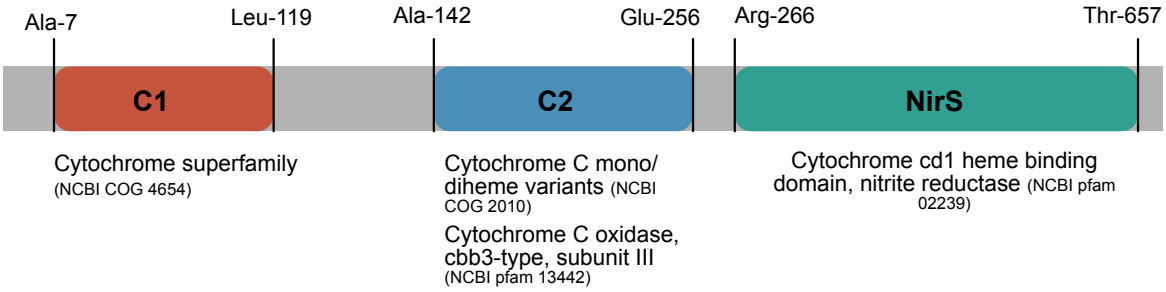


Figure 2.2: Open reading frame domain map for SURF MAG 42 *nirS*
Conserved domain analysis of SURF MAG 42 Chloroflexi nitrite reductase (GenBank RJP53747.1) indicates the presence of two distinct cytochrome superfamily domains and a C-terminal nitrite reductase domain.

The first cytochrome domain (C1) in the Anaerolineales-type ORF was identified by NCBI's Conserved Domains Database (Marchler-Bauer *et al.*, 2015; Lu *et al.*, 2020) as a

cytochrome c551/552. The second cytochrome domain (C2) was predicted with high specificity as a cytochrome C mono- and diheme variant. C2 included a region predicted as a cbb3-type cytochrome c oxidase subunit III; such subunits frequently contain two cytochromes (Bertini, Cavallaro and Rosato, 2006).

2.4.2 NirS gene neighborhood analyses

Gene neighborhood analyses indicated that the MAG-derived nitrite reductase ORF displays a very different local genomic environment as compared with a canonical nitrite reductase neighborhood in *Pseudomonas aeruginosa PAO1* (Fig. 2.2). In *P. aeruginosa*, *nirS* (NCBI reference sequence NP_249210.1)(O'Leary *et al.*, 2016) co-occurs with other genes in the *nir* operon, and is closely adjacent to genes encoding a cytochrome-type nitric oxide reductase (cNOR). This arrangement places nitrite reduction in *cis* with nitric oxide reduction, the next step of canonical denitrification. In the Chloroflexi MAG, no other denitrification genes appear within a 20,000 base pair neighborhood for the C1-C2-NirS nitrite reductase. A similar pattern is observed for a homologous C1-C2-NirS nitrite reductase gene identified in the Chloroflexi *Anaerolinea thermolimosa* (Matsuura *et al.*, 2015); the *A. thermolimosa* neighborhood also shows no evidence of other denitrification genes in the immediate vicinity of the novel NirS, though it does contain some ORFs with predicted functionality similar to those in the SURF MAG 42 neighborhood (Fig. 2.3). Full descriptions of all gene abbreviations are provided in Table 2.1. Analyses of additional selected C1-C2-NirS ORF neighborhoods within other Chloroflexi reveal diverse genetic assemblages also dissimilar to the canonical *Pseudomonas* operon (Fig. 2.4).

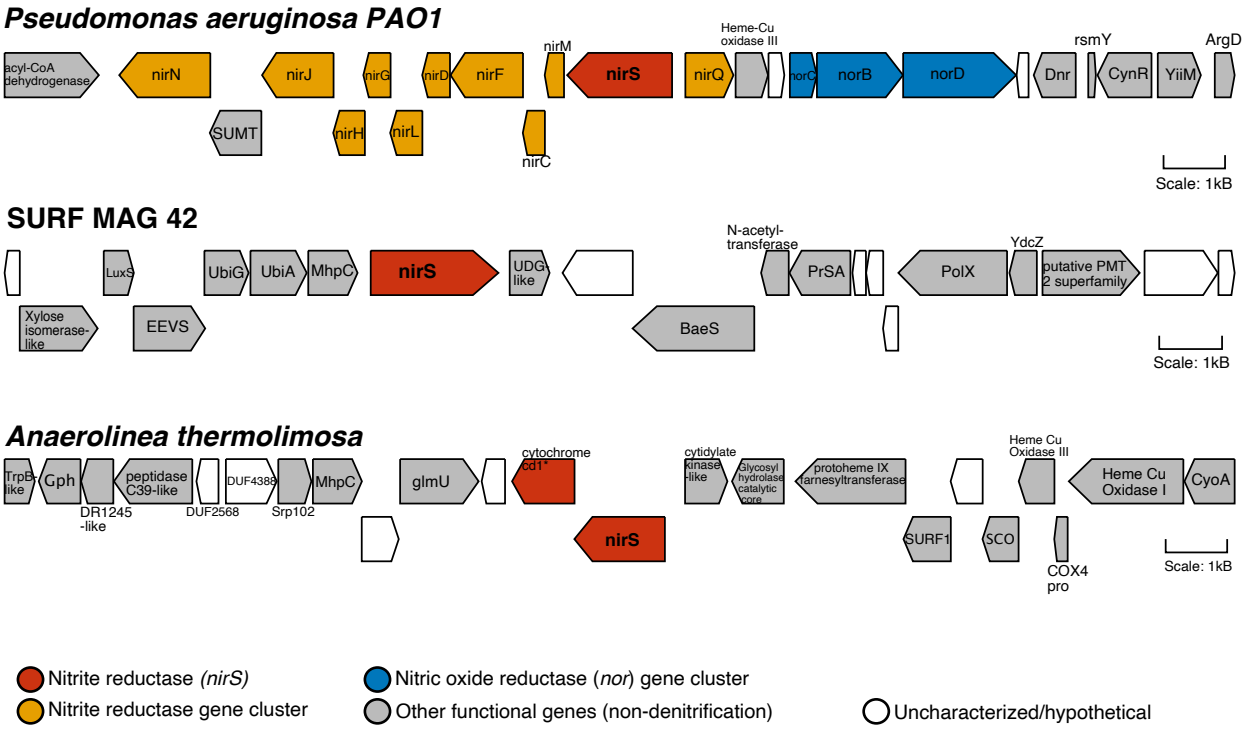


Figure 2.3: *nirS* gene neighborhoods for *Pseudomonas aeruginosa* PAO1, SURF MAG 42, and *Anaerolinea thermolimosa*

Gene neighborhood analyses of the 20,000 base pair region surrounding *nirS* differs markedly between *Pseudomonas aeruginosa PAO1* (GenBank reference sequence NP_249210.1, top) and two Chloroflexi genomes containing the C1-C2-NirS gene: SURF MAG 42 (center) and *Anaerolinea thermolimosa* (bottom). While the *P. aeruginosa* nitrite reductase occurs as part of a larger *nir* operon, and in close proximity to nitric oxide reductase genes, the *nirS* ORF neighborhoods in SURF MAG 42 and *A. thermolimosa* do not appear to contain other denitrification-specific genes.

Table 2.1 Open reading frame classification/description for *P. aeruginosa,* SURF MAG 42, and *A. thermolimosa nirS* neighborhoods

Displayed ORF name	Expanded Description	protein ID	Source MAG/ organism
Xylose isomerase-like	sugar phosphate isomerase/epimerase, xylose isomerase-like TIM barrel	RJP53741.1	SURF MAG 42
LuxS	S-ribosylhomocysteine lyase	RJP53742.1	SURF MAG 42
EEVS	2-epi-5-epi-valiolone synthase/iron-containing alcohol dehydrogenase	RJP53743.1	SURF MAG 42
UbiG	ubiquinone biosynthesis O-methyltransferase/bifunctional 2-polyprenol-6-hydroxyphenol	RJP53744.1	SURF MAG 42
UbiA	ubiA family prenyltransferase	RJP53745.1	SURF MAG 42
MhpC	alpha/beta hydrolase/pimeloyl-ACP methyl ester carboxylesterase	RJP53746.1	SURF MAG 42
UDG-like	uracil-DNA glycosylase	RJP53748.1	SURF MAG 42
BaeS	HAMP domain-containing protein/signal transduction kinase	RJP53750.1	SURF MAG 42
N-acetyltransferase	GNAT family N-acetyltransferase	RJP53751.1	SURF MAG 42
PrSA	ribose-phosphate pyrophosphokinase/phosphoribosylpyrophosphate synthetase	RJP53766.1	SURF MAG 42
PolX	DNA polymerase/3'-5' exonuclease	RJP53755.1	SURF MAG 42
YdcZ	DMT family transporter	RJP53756.1	SURF MAG 42
putative PMT 2 superfamily	Dolichyl-phosphate-mannose-protein mannosyltransferase	RJP53757.1	SURF MAG 42
acetyl-CoA dehydrogenase	probable acyl-CoA dehydrogenase	AAG03897.1	PAO1
nirN	probable c-type cytochrome	AAG03898.1	PAO1
SUMT	probably uroporphyrin-III c-methyltransferase	AAG03899.1	PAO1
nirJ	heme d1 biosynthesis protein NirJ	AAG03900.1	PAO1
nirH	heme d1 biosynthesis protein NirH	AAG03901.1	PAO1
nirG	heme d1 biosynthesis protein NirG/probable transcriptional regulator	AAG03902.1	PAO1
nirL	heme d1 biosynthesis protein NirL	AAG03903.1	PAO1
nirD	heme d1 biosynthesis protein nirD/probable transcriptional regulator	AAG03904.1	PAO1
nirF	heme d1 biosynthesis protein NirF	AAG03905.1	PAO1
nirC	probable c-type cytochrome precursor	AAG03906.1	PAO1
nirM	cytochrome c-551 precursor	AAG03907.1	PAO1
nirQ	regulatory protein NirQ	AAG03909.1	PAO1
Heme-Cu oxidase III	Heme-copper oxidase, subunit III	AAG03911.1	PAO1
norC	nitric-oxide reductase subunit C	AAG03912.1	PAO1

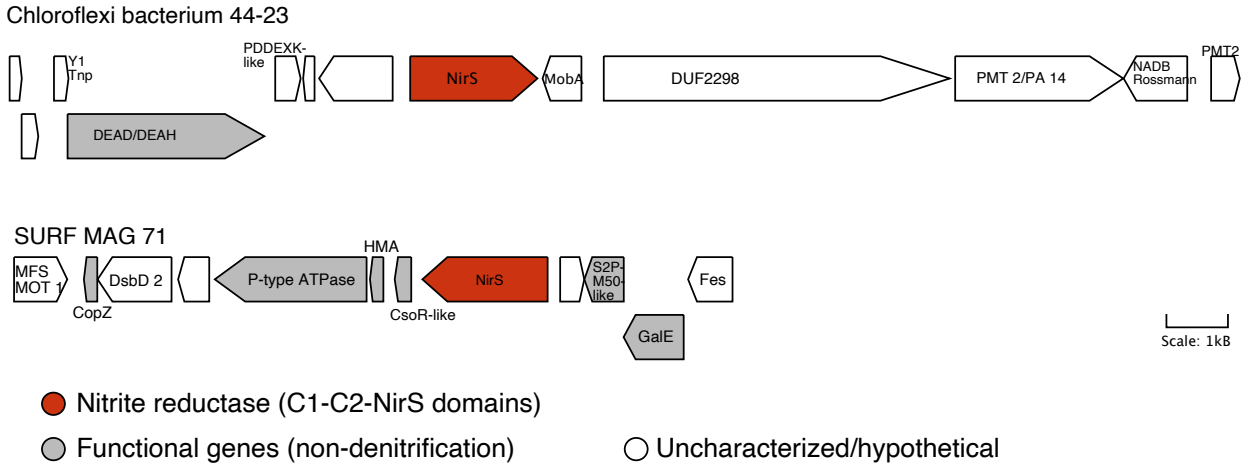


Figure 2.4 Additional Chloroflexi *nirS* gene neighborhoods
Gene neighborhoods showing the 20,000 base pair region adjacent to the putative nitrite reductase genes containing the C1-C2-NirS domain architecture. Neighborhoods are shown for homologs in Chloroflexi bacterium 44-23 (GenBank accession OJX39483.1) and SURF MAG 71 (GenBank accession RJP52528.1). Expanded descriptions of gene classifications are provided in Table 2.2.

Table 2.2 Open reading frame descriptions for Chloroflexi bacterium 44-23 and SURF MAG 71 neighborhoods

Displayed ORF name	Expanded Description	protein ID	Source organism
Y1 Tnp	Hypothetical protein; Y1 Tnp superfamily, transposase IS200-like	OJX39478.1	Chloroflexi bacterium 44-23
DEAD/DEAH	DEAD/DEAH box helicase; type I restriction endonuclease subunit R	OJX39479.1	Chloroflexi bacterium 44-23
PDDEXK-like	Hypothetical protein; PDDEXK nuclease-like superfamily	OJX39480.1	Chloroflexi bacterium 44-23
MobA	Hypothetical molybdenum cofactor guanylyltransferase	OJX39484.1	Chloroflexi bacterium 44-23
DUF2298	Hypothetical protein; uncharacterized membrane protein DUF2298 superfamily; helix-hairpin-helix motif	OJX39485.1	Chloroflexi bacterium 44-23
PMT2/PA14	Hypothetical protein; PA14 domain; Dolichyl-phosphate-mannose-protein mannosyltransferase supefamily	OJX39484.1	Chloroflexi bacterium 44-23
NADB Rossmann	Hypothetical protein; NADB Rossmann superfamily; Rossmann-fold NAD(P)(+)-binding proteins	OJX39487.1	Chloroflexi bacterium 44-23
PMT2	Hypothetical protein; Dolichyl-phosphate-mannose-protein mannosyltransferase	OJX39488.1	Chloroflexi bacterium 44-23
MFS MOT1	Hypothetical protein; putative sulfate/molybdate transporter, MFS superfamily	RJP52521.1	SURF MAG 71
CopZ	Copper chaperone CopZ; heavy-metal-associated domain-containing protein	RJP52522.1	SURF MAG 71
DsbD 2	Hypothetical protein; DsbD 2 cytochrome c biogenesis protein transmembrane region; cupredoxin superfamily-containing protein	RJP52523.1	SURF MAG 71
P-type ATPase	Heavy metal translocating P-type ATPase, Cu-like	RJP52525.1	SURF MAG 71
HMA	Heavy metal transporter, HMA superfamily	RJP52526.1	SURF MAG 71
CsoR-like	Metal-sensitive transcriptional regulator; TthCsoR-like DUF156	RJP52527.1	SURF MAG 71
S2P-M50-like	Site-2 protease family protein; zinc metalloproteases	RJP52530.1	SURF MAG 71
GalE	UDP-glucose 4-epimerase; NADB Rossmann superfamily	RJP52531.1	SURF MAG 71
Fes	Hypothetical protein; Fes superfamily; Enterochelin esterase or related enzyme	RJP52532.1	SURF MAG 71

Conserved domain analysis of NirS homologs in this study suggests that while the C2 and NirS functional domains frequently co-occur in nitrite reductases, the inclusion of C1 in the ORF appears extremely rare and limited to Chloroflexi. A lineage-specific fusion of multiple gene domains could explain this novel C1-C2-NirS arrangement. Different evolutionary histories

among the domain subunits within Chloroflexi would provide evidence for an ancestral horizontal acquisition and fusion event.

2.4.3 Domain phylogenetic trees

To compare the evolutionary histories of each domain in the enzyme ORF, and to determine if the different domains have different ancestry, maximum-likelihood domain trees were reconstructed independently for C1, C2, and the NirS-specific domain (See Methods).

Domain phylogenies indicate similar taxonomic distributions for the C2 and NirS domains (Fig. 2.5). The relative placement of Chloroflexi sequences varies slightly between domain trees: For the C2 domain tree, Chloroflexi sequences are monophyletic within a larger clade comprising polyphyletic sequences including members of the Aquificae, Bacteroidetes, Epsilonproteobacteria, and Spirochaetia; this clade places sister to a large group dominated by Alpha-, Beta-, and Gammaproteobacteria. In the NirS domain tree, Chloroflexi place basally in a clade shared with both polyphyletic sequences and the large radiation of Proteobacterial sequences.

C2 and NirS domain trees reconstructed exclusively from ORFs containing both domains produce similar topologies, but interestingly, the subsampling inverts the placements of the Chloroflexi with respect to the largest Proteobacterial group, as compared with the unsampled trees. This change suggests that sampling and phylogenetic noise are likely responsible for the observed differences in the C2 and NirS domain phylogenies. Additionally, there are notable differences in placement among subclades within the Proteobacteria, and low bipartition support for these subclades for the C2 tree, suggesting that patterns unrelated to Chloroflexi evolution may be polarizing the relative placements of groups in the tree. This lack of robustness caused by alternative sampling, combined with poor support values within the polyphyletic clade or between this clade and the Proteobacteria, suggest that the differences in tree topology may be artifactual, and not reflective of gene reticulation events.

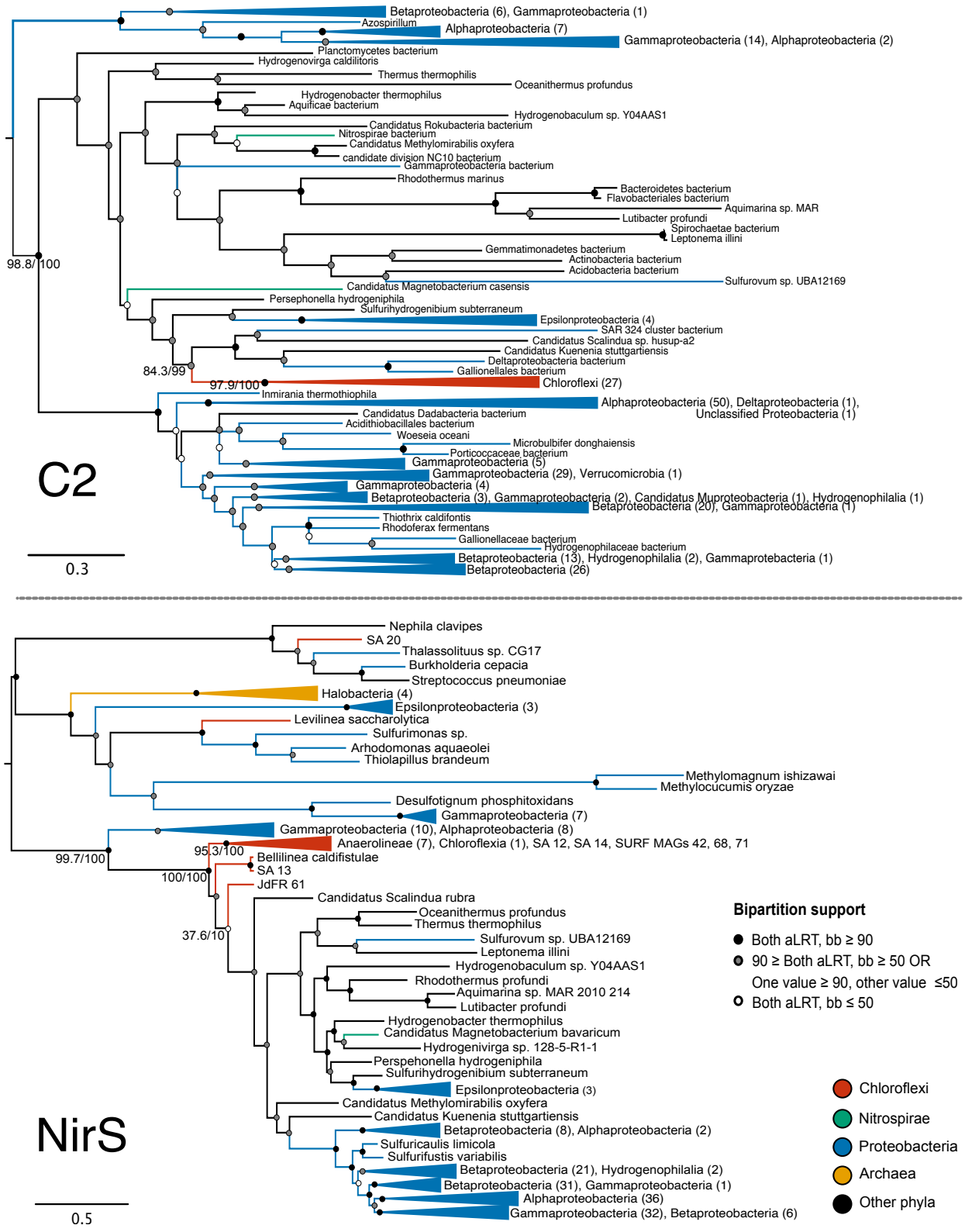


Figure 2.5 Phylogenetic trees for C2 and NirS domains

Phylogenetic analysis of C2 domain homologs (above) places the Chloroflexi within a diverse clade including Epsilonproteobacteria, Aquificae, Bacteroidetes, and Planctomycetes; this clade is sister to a broad radiation of Alpha-

, Beta-, and Gammaproteobacteria. Analysis of NirS domain homologs (below) places the Chloroflexi within the clade dominated by Alpha-, Beta-, and Gammaproteobacteria, which also contains members of the Bacteroidetes and Aquificales. The overall taxonomic representation for the domains is similar. Support values for selected bipartitions are labeled (aLRT/bb). Support for other nodes is indicated with the following color scheme: Strong support with both values ≥ 90 (black); weak support with both values ≤ 50 (white); intermediate support with one or both values between 50 and 90 (grey); conflicting support, with one value ≤ 50 and the other ≥ 90 (grey).

The inferred phylogeny for C1 shows a much different evolutionary history than the other two domains (Fig. 2.6). In contrast to the domain trees for C2 or NirS, the C1 tree shows sequences from Nitrospirae and Nitrospinae grouping together within a large clade of Chloroflexi C1 domains. Additional Chloroflexi sequences group with a small number of more distantly related Proteobacteria. However, the placement and taxonomic representation of Proteobacteria in the C1 tree is different from that seen in the other domain trees.

The majority of ORFs represented in the C1 domain tree contain the C1 domain homolog either as a free cytochrome, or as one of multiple cytochrome-type or cytochrome superfamily domains. In rare or isolated cases, C1 homologs co-occur in ORFs with membrane or structural protein domains (Table 2.3). The ORF containing the C1 homolog was annotated as a nitric oxide reductase in several members of the Nitrospirae and one Geobacteraceae genome; domain analysis of these genes yielded limited additional data, but representative sequences showed detectable sequence similarity to *Pseudomonas norC* genes.

The occurrence of C1 domain homologs within predicted nitrite reductase genes is restricted to the Chloroflexi. The majority of these C1-containing *nir* genes have the cytochrome-type NirS domain; however, a small number of Chloroflexi MAGs contain an ORF pairing the C1 cytochrome with the copper-type NirK domain instead. This NirK ORF also included an N-terminal cupredoxin/plastocyanin domain. As this fusion is only apparent within a small number of MAGs, which are identical across the length of the analyzed ORF, this may represent an assembly artifact. However, several of the *nir* genes that contained a C1 homolog and a cytochrome-type NirS (not copper-type NirK) also contained cupredoxins or other copper-containing domains (Fig. 2.7).

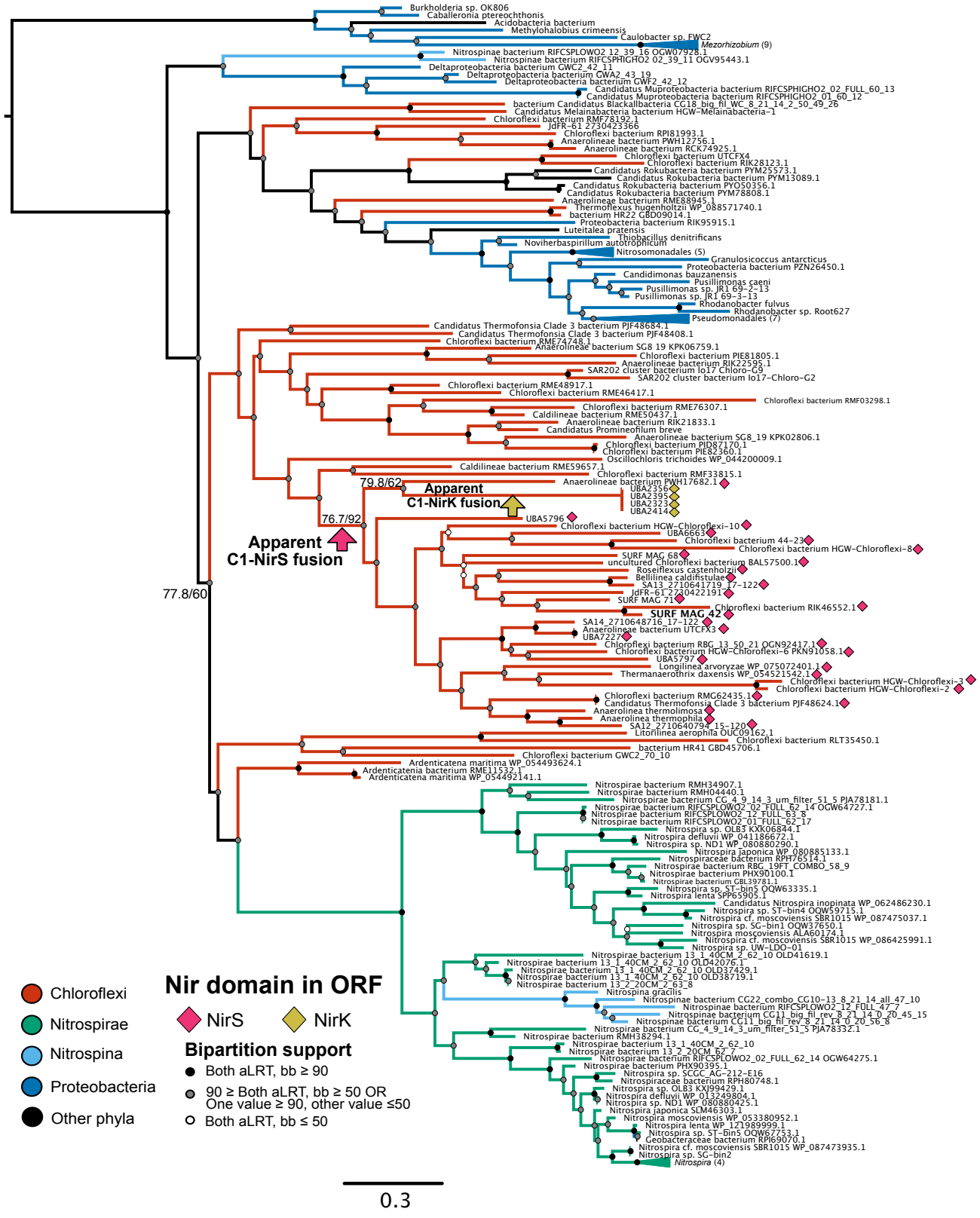


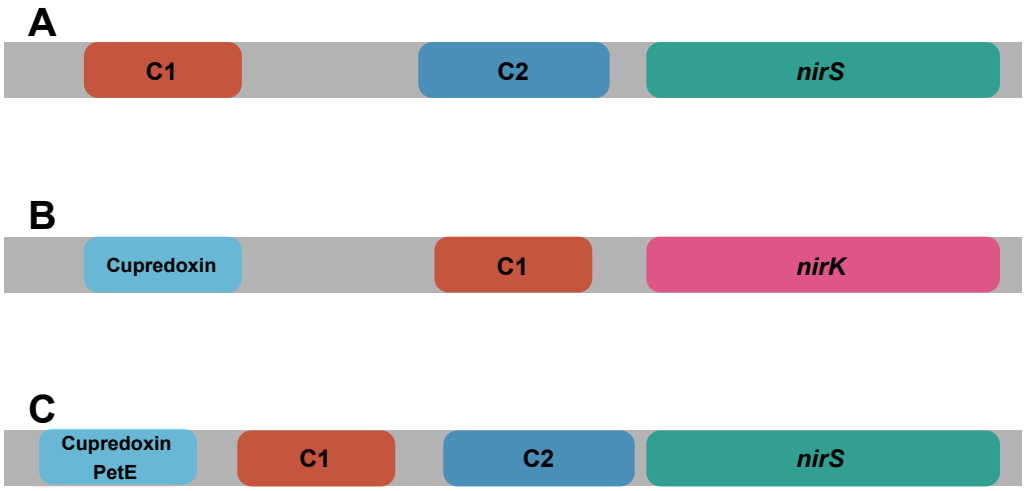
Figure 2.6 Phylogenetic tree for C1 domain

A phylogenetic tree for the C1 domain—with no genus-level filter and inclusion of more distant hits (See Methods)—indicates a limited taxonomic distribution of the domain. The largest group of sequences in Chloroflexi places sister

to domains found in Nitrospirae, Nitrospinae, and Deltaproteobacteria. Within this clade, the branch along which C1 is inferred to have fused into nitrite reductase genes in Chloroflexi is labeled. C1 homologs that co-occur in ORFs with nitrite reductase are indicated with magenta diamonds (NirS) or yellow diamonds (NirK). Support values for selected bipartitions are labeled (aLRT/bb). Support for other nodes is indicated with the following color scheme: Strong support with both values ≥ 90 (black); weak support with both values ≤ 50 (white); intermediate support with one or both values between 50 and 90 (grey); conflicting support, with one value ≤ 50 and the other ≥ 90 (grey).

Table 2.3 Domain families in ORFs with C1 domain homolog but no NirS domain

Domain name	Brief Description	Accession(s)
CyoA	Cytochrome/quinol oxidase	RME74748.1
COX IV	Prokaryotic cytochrome c subunit IV	PKB63614.1; PIQ26724.1
tynA	Primary-amine oxidase	WP_095041976.1
CxxCH_TIGR02603	Putative heme-binding domain	WP_095041976.1; OUC09162.1; OGW07928.1; OGP31126.1; OGV95443.1; OGP47454.1; OGP11994.1; OG149337.1; OYT21294.1; OLD38719.1; OLB21185.1
Cupredoxin	N/A	UBA5757_DIDP01000039.1_40
Caa3_CtaG	Cytochrome c oxidase caa3 assembly factor	PYQ29527.1
PRK10856	Cytoskeleton protein RodZ	OQY877.1
EnvC	Septal ring factor, activator of murein hydrolases	RMF78192.1
PRK03735	Provisional cytochrome b6	RME88945.1
DMSOR_beta-like	DMSO reductase beta subunit	RME88945.1
FlpD	Methyl-viologen-reducing hydrogenase, beta subunit	RME88945.1
TTQ_mauG	tryptophan tryptophylquinone biosynthesis enzyme MauG	RIK95915.1
Heme_Cu_Oxidase_I	Heme-copper oxidase subunit I	WP_073101466.1
PRK14486	Putative bifunctional cbb3-type ytochrome c oxidase subunit II	OYW96265.1
YccC	Uncharacterized membrane protein	OLB04623.1; OAI45763.1; WP_080880425.1; KXJ99429.1
FTR1	Iron permease	WP_053380952.1
EcfT	T component of ECF-type transporters	WP_090902251.1
ArsB_NhaD_permease	Anion permease	OYT18856.1; WP_090747891.1



nir architecture	*Examples*
A	SURF MAG 42 (Ga0104751_10026579); UBA5796 (DIDQ01000096.1); Anaerolinea thermophila (WP_013561145.1); Thermanaerothrix daxensis (WP_054521542.1); Longilinea arvoryzae (WP_075072401.1); Bellilinea caldifistulae (WP_082148881.1); Roseiflexus castenholzii (WP_012121904.1)
B	UBA2323 (DDVB01000085.1_15); UBA2356 (DDTU01000029.1_20); UBA2395 (DDSH01000031.1_54); UBA2414 (DDRO01000006.1_183)
C	UBA7227 (DKTP01000143.1_3); uncultured Chloroflexi bacterium (BAL57500.1); SURF MAG 68 (Ga0104751_100097911)

Figure 2.7 Various nitrite reductase architectures within Chloroflexi
Three different open reading frame types including a C1 cytochrome and a nitrite reductase domain appear in surveyed Chloroflexi. The most commonly seen gene features C1, C2, and cytochrome-dependent *nirS* (A); however, C1 is also seen in ORFs with cupredoxin and copper-dependent nitrite reductase (*nirK*) domains (B). A limited number of Chloroflexi genomes also contain ORFs placing C1, C2, and *nirS* together with an N-terminal cupredoxin, plastocyanin (PetE) or similar copper-containing domain (C).

The distinct phylogeny and taxonomic distribution of C1, as compared with C2 and NirS domains, strongly suggest that the C1-C2-NirS domain structure observed in Chloroflexi is the result of fusion of the C1 domain with a horizontally acquired *nir* gene containing the C2 and NirS domains. Topology and extant taxon sampling of these gene trees does not allow reliable inference of the donor lineage of this transfer. However, the C2-NirS architecture—or similar arrangements of functional domains—is widespread among members of the Alpha-, Beta-, and Gammaproteobacteria; additionally, gene trees for C2 and NirS place the Chloroflexi that also contain the C1 domain within or as sister to Proteobacterial groups, inconsistent with species tree

placements for these phyla. These data suggest the Proteobacteria as a possible donor group for the C2-NirS domains. Additionally, it appears that Chloroflexi may have been the source for an independent transfer of the free C1 domain into Nitrospirae and Nitrospinae.

2.4.4 Metagenome database searches

Though putative homologs exist independently for the constituent C1 and C2-NirS regions, respectively, these hits reflect different cytochrome or cytochrome-type nitrite reductases (largely in Proteobacteria, Nitrospirae, and Nitrospinae). The full C1-C2-NirS architecture appears unique to Chloroflexi and is not observed in other groups. Querying NCBI's non-redundant environmental database (env-nr) with the full ORF from SURF MAG 42 did not identify additional examples of the full gene construct. While several hits were identified that reflected putative homology to the joint C2-NirS domains, none of these included the C1 domain as well. An independent search of the env-nr database using the C1 domain as a query returned few overall hits. While some of these putative C1 homologs were identified in ORFs containing additional cytochrome-type enzyme superfamily domains or subunits, none co-occurred with NirS or NirK domains. These data suggest that there is little to no missing diversity of the Chloroflexi-type chimeric nitrite reductase in existing metagenomes.

2.4.5 Homology modeling of novel Chloroflexi NirS

Attempts to visualize the full enzyme structure using homology modeling (Bienert *et al.*, 2017; Waterhouse *et al.*, 2018) were unsuccessful; structural models were only able to predict a close match for the conserved C2-NirS region of the putative gene. Efforts to independently model the C1 structure could not recover predicted QMEAN scores above -4.50 (Benkert *et al* 2011). The poor scores may reflect the relatively short length of the cytochrome coding region. However, the Chloroflexi *nirS* gene sequence does retain several conserved residues present in the crystal structure of *Pseudomonas aeruginosa* NirS. In *P. aeruginosa* NirS, His51 and Met88 coordinate heme c; His182 coordinates heme d1; and His327 and His369 are believed to stabilize the active site nitrite anion (Rinaldo *et al.*, 2011; Maia and Moura, 2014). Corresponding residues are conserved within the C2 (His65, Met125) and NirS alignments (His46, His239, His300) for the Chloroflexi NirS ORF; interestingly, the residue corresponding to His327 (His239) is not universally conserved, though it is conserved among Chloroflexi with the novel NirS architecture.

2.4.6 Expansion of eNOR diversity

Notably, the majority of genomes with the unique C1-C2-NirS structure do not appear to contain a nitric oxide reductase gene (*nor*). Though the absence of the *nor* gene in genomes with *nirS* is not unprecedented, previous genomic surveys suggest it is relatively uncommon, and the toxicity of the product of Nir (nitric oxide, NO) makes this absence counterintuitive (Hendriks *et al.*, 2000; Graf, Jones and Hallin, 2014). However, analysis of the SURF MAG 42 metagenome—the original assembled genome in which the novel *nirS* ORF was observed—did reveal the presence of an unusual *nor* homolog. Previous studies have identified the established cNOR and qNOR family enzymes (which contain cytochrome c or quinols as electron donors, respectively) in Chloroflexi, as well as a broad distribution of Proteobacteria (Hendriks *et al.*, 2000; Zumft, 2005; Hemp and Gennis, 2008). However, the predicted NOR in SURF MAG 42 included an active site glutamine substitution characteristic of eNOR (Hemp and Gennis, 2008; Hemp *et al.*, 2015) (Table 2.S.5). SURF MAG 42 contains both proposed subunits (Hemp and Gennis, 2008) of eNOR; the diagnostic subfamily substitution is within the heme-copper cytochrome-containing subunit I, and the cupredoxin-containing subunit II is immediately upstream in the MAG (Figure 2.8). eNOR has been previously described in Archaea (Hemp and Gennis, 2008) and at least one isolated Anaerolineales bacterium (Hemp *et al.*, 2015).

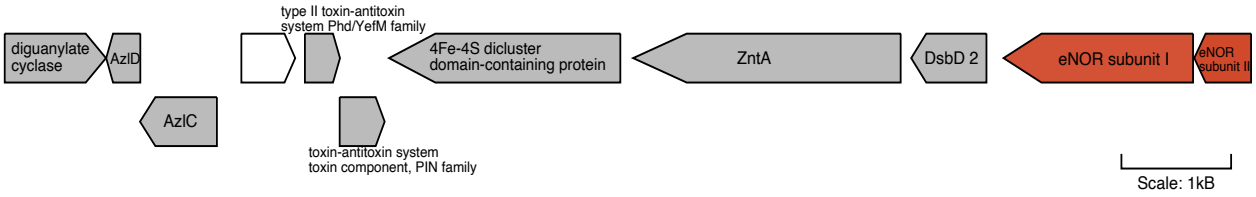


Figure 2.8 eNOR gene neighborhood
A gene neighborhood showing the 20,000 base pair region adjacent to eNOR subunits in SURF MAG 42. Sequence analysis of subunit I (NCBI protein accession RJP50323.1) reflects a conserved glutamine substitution characteristic of the eNOR subfamily. Both proposed subunits of eNOR—the heme-copper cytochrome-containing subunit I, and the cupredoxin-containing subunit II—appear in the MAG. No other denitrification genes appear in the neighborhood.

Phylogenetic analysis indicates the presence of *eNOR* in an expanded diversity of genomes (Fig. 2.9). Previous studies have described eNOR in *Natronomonas*; these data indicate a cluster of *eNOR* genes throughout other Halobacteria as well. Additional putative *eNOR* genes appear in

multiple members of Anaerolineales, as well as other Chloroflexi, and in many members of the Alpha-, Beta-, Gamma-, and Deltaproteobacteria. Many of these putative *eNOR* subunit homologs appear to have been misannotated or mislabeled as cytochrome c oxidase genes, likely because of the structural similarity of the heme-copper cytochrome region (Hemp and Gennis, 2008; Marchler-Bauer *et al.*, 2015).

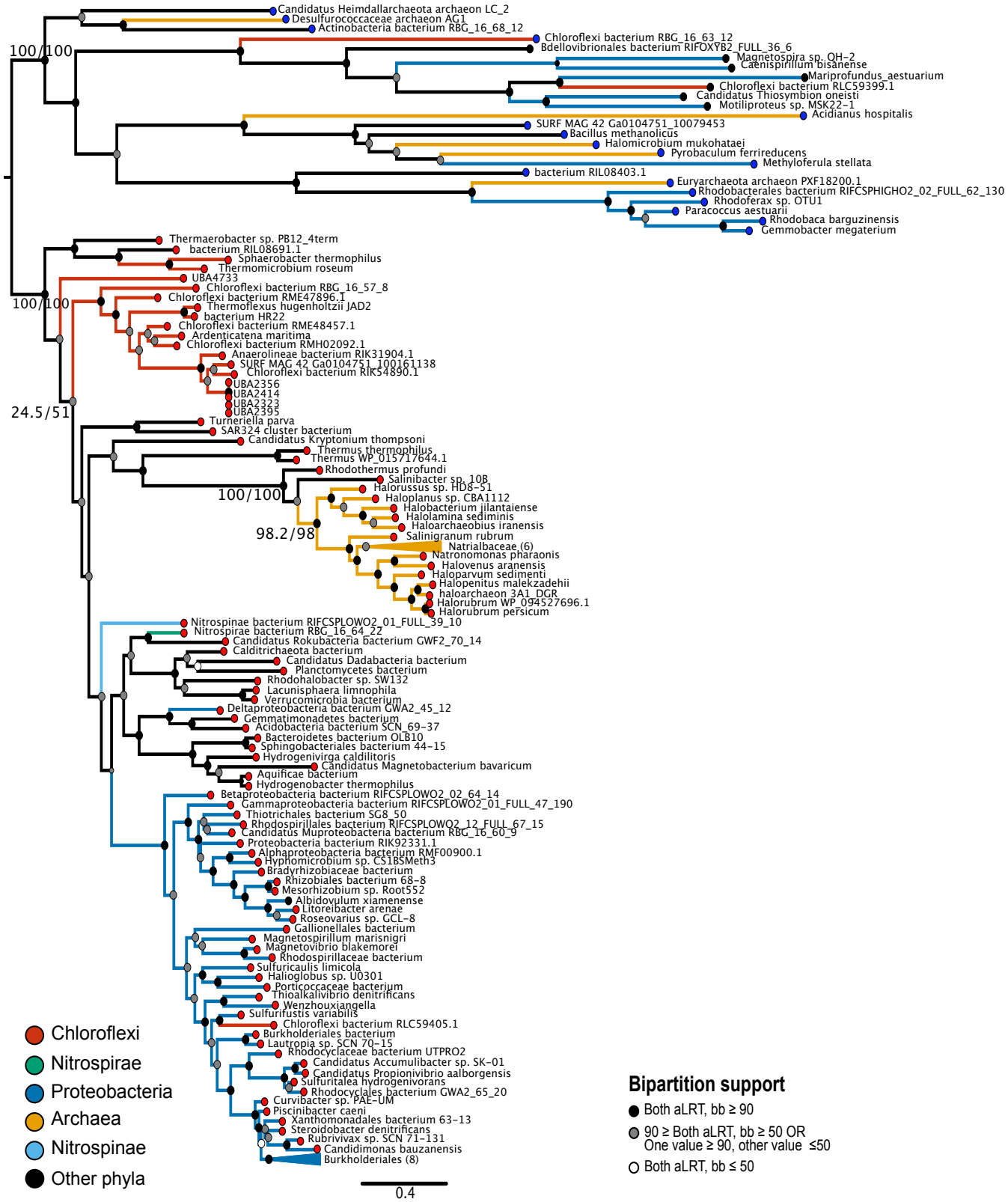


Figure 2.9 eNOR gene tree

A phylogenetic tree of homologs to the nitric oxide reductase from SURF MAG 42 reveals an expanded diversity of putative *eNOR* subunit I homologs in not only Archaea and Chloroflexi, but also Proteobacteria and other diverse

phyla. Putative *eNOR* sequences (red tips) have the characteristic Glu-323 in the alignment; outgroup sequences (blue tips) have Tyr-323 (oxygen reductase superfamily) or other substitutions. Support values for selected bipartitions are labeled (aLRT/bb). Support for other nodes is indicated with the following color scheme: Strong support with both values ≥ 90 (black); weak support with both values ≤ 50 (white); intermediate support with one or both values between 50 and 90 (grey); conflicting support, with one value ≤ 50 and the other ≥ 90 (grey).

2.5 Discussion

2.5.1 Subsurface denitrification in Chloroflexi

The phylogenetic analyses of *nirS* and *eNOR* ORFs in Chloroflexi suggest that subsurface ecosystems may harbor an under-described diversity of denitrification enzymes. The divergence and diversity of these proteins may reflect adaptations to the severe energy and redox limitations present in deep subsurface biomes (Momper, Jungbluth, *et al.*, 2017; Bell *et al.*, 2020). More broadly, a deeper understanding of the ecological extent of microbial denitrification has important implications for basic and applied microbial ecology. The reduction of fixed nitrogen species plays a crucial role in global nitrogen cycling and is also an essential component of smaller-scale systems, such as those associated with agricultural or waste treatment (Butterbach-Bahl and Dannenmann, 2011; Lu, Chandran and Stensel, 2014). The discovery and characterization of novel variants of genes such as *nirS* and *eNOR* may therefore pave the way for future biotechnological applications.

2.5.2 Horizontal transfer and chimeric fusion of C1 in ancestral Chloroflexi

Although the C2 and NirS domains do not have identical evolutionary histories or distributions, the taxonomic representation of these groups is very similar, and the presence of the paired C2-NirS domains in cytochrome-type nitrite reductases appears broadly throughout the Proteobacteria. In contrast, the taxonomic distribution and phylogeny of the C1 domain tree is strikingly different than that of the other domains in the nitrite reductase ORF. Combined with the apparent absence of a full C1-C2-NirS ORF in any taxonomic group other than Chloroflexi, these data suggest that the C1 cytochrome was likely incorporated into *nirS* in a gene fusion event within Chloroflexi, following HGT. As there is no evidence of the C2-NirS ORF in Chloroflexi without the fused C1 domain present, the fusion probably occurred very soon after the acquisition of the C2-NirS region and may be necessary for the function of the gene in Chloroflexi.

2.5.3 Putative function of C1 domain in nitrite reductases

In *Pseudomonas aeruginosa,* cytochrome c551—encoded by *nirM*—is the electron donor for the adjacent cytochrome *cd1 nirS* (Zumft, 1997; Philippot, 2002). Homology searches do not identify any regions within SURF MAG 42 with significant similarity to *Pseudomonas nirM*, but C1 is identified as a putative member of the cytochrome c551/552 family. It is therefore possible that C1, though divergent from Proteobacterial *nirM*, serves a similar redox role for the cytochrome *cd1* now within the same ORF.

Interestingly, putative homologs of C1 cytochrome domains were found in some Chloroflexi genomes in ORFs containing *nirK*, not *nirS* (Fig. 2.6). Though NirS and NirK are functionally equivalent, the two enzymes do not show a shared evolutionary origin, and are often—though not always—mutually exclusive among known denitrifier genomes (Jones *et al.*, 2008; Graf, Jones and Hallin, 2014). Unlike the cytochrome-containing NirS, NirK is a copper-type enzyme. The co-occurrence of cytochrome *c* domains in ORFs with the copper-type *nirK* has been identified in rare instances in Proteobacteria, and noted as surprising, given the cupredoxin-like fold of the NirK enzyme (Bertini, Cavallaro and Rosato, 2006). Similarly surprising is the inverse relationship revealed in the C1 domain tree: Several Chloroflexi ORFs contain a cupredoxin or similar copper-containing domain N-terminal to the C1-C2-NirS architecture (Fig. 2.7). The co-occurrence of C1 with both cytochrome- and copper-dependent Nir domains suggests a general evolutionary trend within Chloroflexi to incorporate this cytochrome into denitrification ORFs. This distribution pattern raises the possibility that the C1-type cytochrome may serve an important but generalized role in nitrite reduction—regardless of the evolutionary history or genetic profile of the nitrite reduction domain itself.

The apparent absence of a *nor* homolog in the majority of genomes with the C1-*nirS* fusion is unexpected. Beyond providing downstream redox capacity, nitric oxide reductase provides an efficient means of reducing and detoxifying nitric oxide, the highly cytotoxic product of NirS. It is not unprecedented for bacterial genomes to harbor a *nir* gene without a *nor* gene, particularly for organisms with *nirK* (Heylen *et al.*, 2007; Graf, Jones and Hallin, 2014). This *nir*-*nor* mismatch is much rarer for putative denitrifiers with *nirS*, representing fewer than 4% of genomes in a recent survey—but a small number of surveyed bacteria do, interestingly, appear to harbor *nirS* without also harboring *cNOR* or *qNOR* (Heylen *et al.*, 2007; Graf, Jones and Ha llin, 2014). To our knowledge, however, *eNOR* has not been included in such analyses of the genomic correlation

between nitrite reductases and nitric oxide reductases. The phylogenetic evidence for diverse *eNOR* homologs suggests likely undocumented or underexplored diversity for divergent nitric oxide reductases. Diversity and function of cytochrome-dependent (cNOR) and quinol-dependent nitric oxide reductases (qNOR) are fairly well-established. However, divergent enzymes such as eNOR and sNOR are less-extensively documented and may not be accurately distinguished from broader oxygen reductase superfamily members in genomic or metagenomic analyses.

Cytochrome *c* proteins function as electron transfer proteins in anaerobic respiration and are often fused to redox enzymes to allow electron passage (Bertini, Cavallaro and Rosato, 2006). It is not surprising, therefore, to find cytochrome c-containing subunits in frame with nitrite reductase. NirS itself is cytochrome-dependent (Bertini, Cavallaro and Rosato, 2006). However, the unusual addition of the upstream cytochrome domain (C1) may reflect additional redox requirement or capacity. It is also possible that the inclusion of this construct could be linked to the conspicuous absence of nitric oxide reductase enzymes in several metagenome-assembled genomes containing a NirS ORF with the C1 fusion. Nitric oxide reduction can be cytochrome-dependent; the well-studied cNOR nitric oxide reductases contain a membrane-anchored cytochrome *c* (Hemp and Gennis, 2008). Further, the C1 domain tree recovers ORFs in the Nitrospirae that contain C1 homologs and are annotated as nitric oxide reductases, with detectable similarity to Proteobacteria nitric oxide reductase subunits. It is therefore possible that the inclusion of a C1 domain in *nir* genes within genomes lacking *eNOR* reflects some generalized NOR-like role in detoxification of the cytotoxic product of NirS.

Additionally, while the presence of NirS suggests an active denitrification pathway, and the NirS domain tree reflects the homology between this domain and NirS from known denitrifying groups, the possibility remains that this group of Chloroflexi do not perform denitrification, and instead use this gene product for a different metabolic function, potentially enabled or constrained by the C1 domain. The genetic diversity observed within C1-C2-NirS gene neighborhoods—varying both from the genetic makeup of a canonical denitrification operon, and between different Chloroflexi MAGs—may reflect such functional flexibility. Shared or convergent function could also be recovered among these diverse neighborhoods. The analyzed gene neighborhoods suggest some conserved functionality between different genomes from varying environmental sources and samples; for example, the presence of heme-copper oxidase subunits, molybdenum cofactor enzymes, and NADB Rossman superfamily proteins (Table 2.1, Table 2.2). However, without

expanded homology analyses and experimental validation, it is impossible to infer whether apparent similarities or differences reflect biologically meaningful patterns of inheritance or function, or are simply artifactual.

2.5.4 Putative function of eNOR homologs

A divergent or generalized role is also possible for the eNOR homologs. Models indicate that eNOR (unlike related nitric oxide reductases cNOR, qNOR, sNOR, and gNOR) has a proton channel, and therefore the capacity for proton pumping (Hemp and Gennis, 2008); therefore, this gene product may serve a key electrogenic role, whether reducing nitric oxide or alternative substrates. Experimental validation would be necessary to determine if the novel Chloroflexi-associated NirS performs differently than canonical NirS *in vivo*, and if NirS and eNOR perform targeted denitrification or nonspecific detoxification or proton pumping. This work, therefore, suggests a promising direction for future investigations.

2.5.5 Conclusions and future directions

The divergent denitrification enzymes described above may or may not reflect different metabolic strategies *in situ*. But the identification of both a novel *nirS* ORF and an expanded diversity of eNOR enzymes suggests that the existing understanding of denitrification may underestimate the genetic diversity and ecological distribution of constituent enzymes. This may be especially true in deep subsurface biomes, such as those from which several Chloroflexi analyzed in this study were isolated. These systems have garnered increasing attention in recent years; extensive evidence supports the existence of dynamic, diverse microbial subsurface ecosystems with the metabolic potential to influence global biogeochemical cycles (Hug *et al.*, 2013; Osburn *et al.*, 2014, 2019; Momper, Jungbluth, *et al.*, 2017; Magnabosco *et al.*, 2018). Chloroflexi are frequently cited as well-represented members of deep sediment and aquifer systems, where they play key roles in carbon cycling dynamics (Hug *et al.*, 2013; Momper, Jungbluth, *et al.*, 2017; Momper, Kiel Reese, *et al.*, 2017; Kadnikov *et al.*, 2020). But Chloroflexi are known to also harbor diverse nitrogen metabolisms (Hemp *et al.*, 2015; Denef *et al.*, 2016; Spieck *et al.*, 2020), and previous studies have linked subsurface Chloroflexi to denitrification pathway genes such as nitrous oxide reductase (*nos*) (Sanford *et al.*, 2012; Hug *et al.*, 2016;

Momper, Jungbluth, *et al.*, 2017). The role of Chloroflexi in subsurface nitrogen cycling—as well as the scope of subsurface microbial nitrogen dynamics at large—requires further investigation.

2.S Supplementary Tables and Figures

Table 2.S.1 Domain query sequences

SURF MAG 42 Domain name	Amino acid sequence
C1	APSEEPQQLEAVELEPQAVVDAFTKGGCGACHAIPGIPGAVGTIGPDLSSIS TVAESHLADSAYTGKAKTAEEYILESVLKPDIFVAPECPTGVCQQGLMPASL GDTLSGTEL
C2	AVDTSSADVSLSDDDFAWAKQTFFDRCAGCHGTMRKGATGPALTPDLTLQ KGTLALSSIIYNGTAKGMPDWGKQGFFTQEETDIMAKYLQNEPPAPPEMSM EQMKATWKVLVEPE
NirS	RNWENYFSVTLRDAGQVAIIDGDTYEVVAKVDTGYAVHISRMSATGRYVYV IGRDGKLALVDLWMEIPDKVAEVQTCYDARSVEVSKYHGELGDFRDEYAIV GCYWPSHFTILDGETLEPFKVVSVRGYTSDTNTYVGDPRVAAIISSEYKPEW IVNVKETGMVWMVNYQDPKNPTIKMINAALYLHDGGWDASKRYFLVAANQS NKIAVVDALNGELTALVDTPEVPHPGRGANWIDPEFGPVWSTPHLSAPSLIAI GTDPEGNPDSAWKVVRDIELPGAGSLFVKTHPNSQWVWVDFVLNSDEKIQR TVCVVAKADPTKVHKCWEAADYGRAVHFEYNQDGTEVWVSIWGSADQPGK TGEIVVYDDATLAEITRIPDLITPTGKFNVYNT
eNOR	MKMTALAPTTPWVAGKKDDYRACPVTTLKVDMHAERLIIANAVMAVVCLLLG GIAALLIALTRWQAIHLLDAKWFYSLLTLHGIDMLVAWIVFFEMAGLHFGSTAL LNARHAAPKLAWFGFILMVIGGLMVNVVVLFDANSIVTFTAYVPLKAHPLFYL SYILFAVGALIQVITFFINIVVAKREKRFEGSLPLAVFGLMTAAILAVYTLLQGAA AIVPAFFWSLGILANYDPGIYRNFFWGFGHPAQQVNLAAMVSIWYMLAQMTV GIRPLNEKFSRLAFILYLFFINLGSAHHLLVDPGLTFAWKAVNTSYAMYLAVLG SMMHAFSIPAAMEVAQRAKGYTKGLFGWLRNAPWGEPGFAGLVMSMVIFG WIGGITGVTIGTEQINMLAHNTLRLPGHFHATVVGGTTTAFMAFTYYVIPLIFR KELKLKKWAQWQPYVFGGGTALVSLGMITSGLQGVARRNWDITFAGVVPGT VDLTLAIFGIGALIAVTGGIMYLTIVLVSVFTGPRMEAGQMLLIGGSENPLLEST KLTDEQMEHPKNQPKGTLVLVFAFLVWFAIYYLSNWWLLGRTWFIR*

Table 2.S.2 Chloroflexi MAG source data from SURF

GENOME	BioProject	BioSample	MAG name	MAG/hit db file name	Publication
1	PRJNA355136	SAMN08499021	SURF MAG 27	NA1	Momper *et al* 2017
2	PRJNA355136	SAMN08499024	SURF MAG 30	NA2	Momper *et al* 2017
3	PRJNA355136	SAMN08499034	SURF MAG 40	NA3	Momper *et al* 2017
4*	PRJNA355136	SAMN08499036	SURF MAG 42	NA4	Momper *et al* 2017
5	PRJNA355136	SAMN08499037	SURF MAG 43	NA5	Momper *et al* 2017
6	PRJNA355136	SAMN08499065	SURF MAG 71	NA6	Momper *et al* 2017
7	PRJNA355136	SAMN08499062	SURF MAG 68	NA7	Momper *et al* 2017
8	PRJNA269163	SAMN06226378	JdFR-61	JdFR61	Jungbluth *et al* 2017
9	PRJNA681409 / PRJNA680468	pending release	----	SA8	----
10	PRJNA681409 / PRJNA680468	pending release	----	SA9	----
11	PRJNA681409 / PRJNA680468	pending release	----	SA10	----
12	PRJNA681409 / PRJNA680468	pending release	----	SA11	----
13	PRJNA681409 / PRJNA680468	pending release	----	SA12	----
14	PRJNA681409 / PRJNA680468	pending release	----	SA13	----
15	PRJNA681409 / PRJNA680468	pending release	----	SA14	----
16	PRJNA681409 / PRJNA680468	pending release	----	SA15	----
17	PRJNA681409 / PRJNA680468	pending release	----	SA16	----
18	PRJNA681409 / PRJNA680468	pending release	----	SA17	----
19	PRJNA681409 / PRJNA680468	pending release	----	SA18	----
20	PRJNA681409 / PRJNA680468	pending release	----	SA19	----
21	PRJNA681409 / PRJNA680468	pending release	----	SA20	----

**SURF MAG 42 is source genome for nirS ORF*

Table 2.S.3 Additional MAG source data from Parks *et al*

	Organism name		Isolate ID	WGS ID
1	Anaerolineaceae	bacterium	UBA1024	DCER00000000
2	Anaerolineaceae	bacterium	UBA2178	DCVW00000000
3	Anaerolineaceae	bacterium	UBA2274	DDWY00000000
4	Anaerolineaceae	bacterium	UBA6073	DIXT00000000
5	Anaerolineaceae	bacterium	UBA7644	DLIK00000000
6	Anaerolineales	bacterium	UBA1429	DCTX00000000
7	Anaerolineales	bacterium	UBA2181	DCVT00000000
8	Anaerolineales	bacterium	UBA2200	DCVA00000000
9	Anaerolineales	bacterium	UBA2317	DDVH00000000
10	Anaerolineales	bacterium	UBA2323	DDVB00000000
11	Anaerolineales	bacterium	UBA2356	DDTU00000000
12	Anaerolineales	bacterium	UBA2395	DDSH00000000
13	Anaerolineales	bacterium	UBA2414	DDRO00000000
14	Anaerolineales	bacterium	UBA2796	DEHQ00000000
15	Anaerolineales	bacterium	UBA6092	DIXA00000000
16	Anaerolineales	bacterium	UBA6663	DKKP00000000
17	Anaerolineales	bacterium	UBA6665	DKKN00000000
18	Anaerolineales	bacterium	UBA7227	DKTP00000000
19	Chloroflexaceae	bacterium	UBA1466	DCSM00000000
20	Chloroflexi	bacterium	UBA2235	DDYL00000000
21	Chloroflexi	bacterium	UBA5177	DHWR00000000
22	Chloroflexi	bacterium	UBA5183	DHWL00000000
23	Chloroflexi	bacterium	UBA6042	DIYY00000000
24	Chloroflexi	bacterium	UBA6077	DIXP00000000
25	Chloroflexi	bacterium	UBA6265	DJVD00000000
26	Chloroflexi	bacterium	UBA6622	DJHK00000000
27	Dehalococcoidia	bacterium	UBA1127	DCAS00000000
28	Dehalococcoidia	bacterium	UBA1151	DBZV00000000
29	Dehalococcoidia	bacterium	UBA1222	DBXC00000000
30	Dehalococcoidia	bacterium	UBA2141	DCXH00000000
31	Dehalococcoidia	bacterium	UBA2158	DCWQ00000000

Table 2.S.4 Best-fit evolutionary models for trees

Gene/ORF	Figure no.	Substitution Model (selected by IQ Tree MF)
C1, no outgroup	A8	WAG + F + R4
C1, with outgroup	5	WAG + F + R6
C2	4, A2	LG + R7
NirS	4, A3	LG + F + R7
eNOR	6	LG+F+R6

Table 2.S.5 Heme-copper superfamily active site classifications (modified from Hemp and Gennis, 2008)

Enzyme	Active-site residue	Predicted # proton channels	Present in
Oxygen reductase	Y	0-2	All Domains; varies by enzyme subfamily
cNOR	E	0	Proteobacteria (Shiro 2012, Hendriks 2000)
qNOR	E	0	Proteobacteria, Cyanobacteria, Firmicutes, Archaea (Shiro 2012, Hemp and Gennis, Heylen *et al* 2007, Hendriks 2000)
sNOR	N	0	Betaproteobacteria, *Silicibacter, Deinococcus*, *Geobacillus* (Stein *et al* 2007)
eNOR	Q	1	Archaea (Hemp and Gennis, this study), Chloroflexi (Ward, this study), Proteobacteria (this study)
gNOR	D	0	*Sulfurimonas*, *Sulfurovum, Persephonella* (Sievert *et al* 2007)
CuANor	N	1	*Bacillus* (Al-Attar and De Vries 2015, Suharti *et al* 2001)

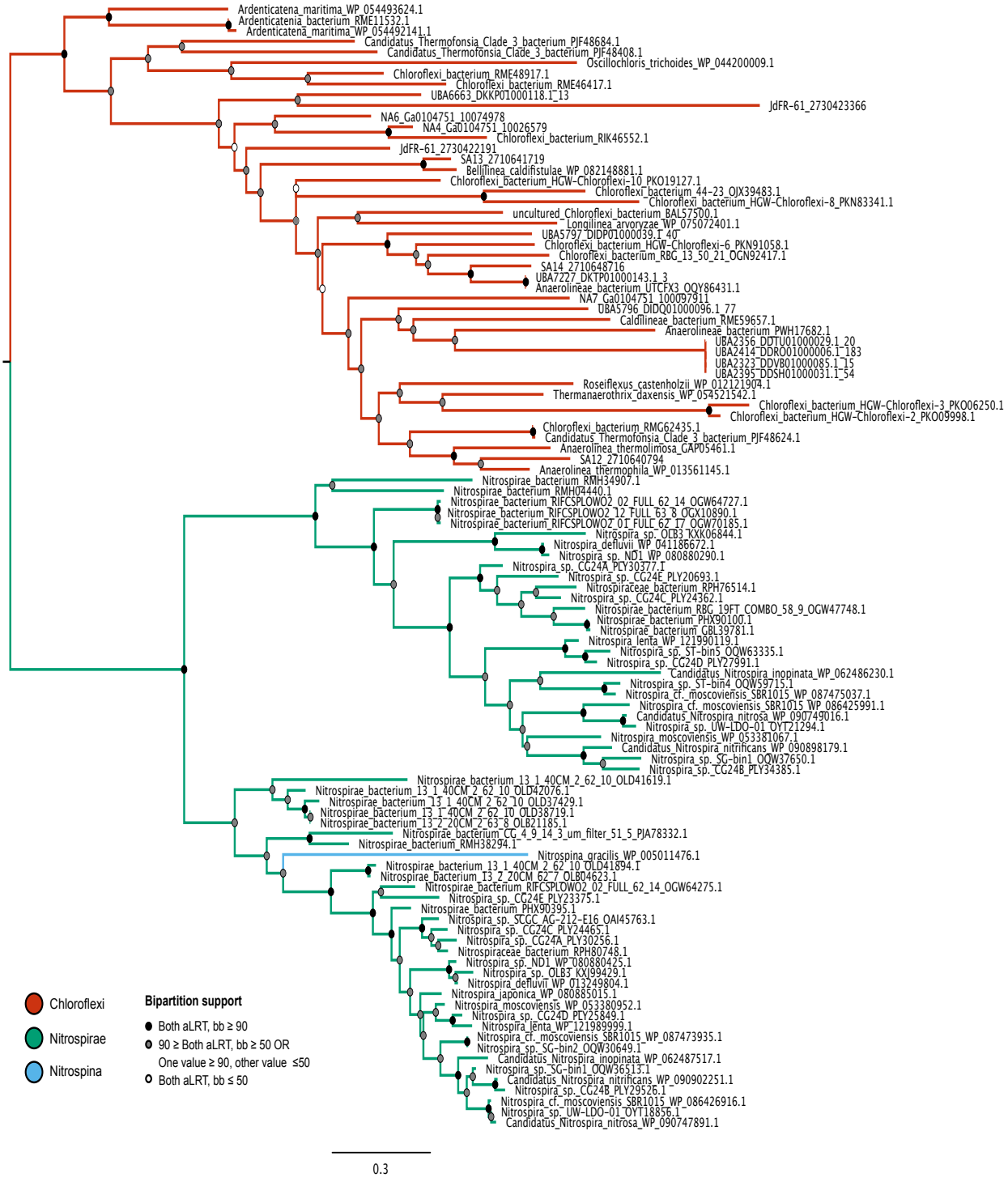


Figure 2.S.6 Conservative initial C1 tree without reduced hit threshold

A preliminary phylogenetic tree for C1, including only sequences with E ≤ 10^{-10}, contains very few overall taxa. The tree contains only members of the Chloroflexi, members of the Nitrospirae, and *Nitrospina gracilis*; the sampling depth does not recover an outgroup for these sister groups. Support for bipartitions is indicated with the following color scheme: Strong support with both values ≥ 90 (black); weak support with both values ≤ 50 (white); intermediate support with one or both values between 50 and 90 (grey); conflicting support, with one value ≤ 50 and the other ≥ 90 (grey).

Chapter 3. Evolution and distribution of cyanide-reducing microbial nitrilases

3.1 Abstract

In addition to its role as an extremely toxic environmental contaminant, cyanide is hypothesized to play a role in prebiotic chemistry and early biogeochemical evolution. Though cyanide-reducing enzymes have been studied and engineered for bioremediation, the extant diversity of these enzymes remains underexplored. Additionally, the age and emergence of cyanide metabolisms is poorly constrained. Here we provide a comprehensive phylogenetic analysis of the distribution and evolution of the Class I nitrilases, revealing a broad distribution of prokaryotic enzymes that could be optimized for cyanide bioremediation. Additionally, we use molecular clock analyses to show that cyanide-reducing nitrilases originated in bacteria in the Paleo- to Mesoproterozoic, and were subsequently horizontally transferred into eukaryotes.

3.2 Introduction

Cyanide is well-known as a globally distributed environmental contaminant. The anion, which is produced as a by-product in several industries, including mining, electroplating, petrochemical refining, and pharmaceutical manufacturing, poses a significant health hazard (Raybuck, 1992; Ebbs, 2004; Alvillo-Rivera *et al.*, 2021). Free cyanide anions and hydrogen cyanide (HCN) inhibit aerobic respiration by binding metalloproteins such as cytochrome *c* oxidase, and are extremely toxic to animals at high concentrations (Raybuck, 1992). Additionally, HCN can disrupt biogeochemical cycling, especially within the nitrogen cycle (Kapoor *et al.*, 2016). However, cyanide is also produced biologically, and diverse plants, fungi, and bacteria have well-established enzymatic pathways for free cyanide metabolism (Raybuck, 1992; Ebbs, 2004; Benedik and Sewell, 2017; Park, Trevor Sewell and Benedik, 2017). As a result, these organisms have been extensively explored and engineered for bioremediation applications in cyanide-contaminated areas (Martínková *et al.*, 2015; Park, Trevor Sewell and Benedik, 2017).

Multiple enzyme families have the capacity to reduce the strong nitrile bond between carbon and nitrogen. For example, the nitrile hydratases (NHases) convert organic nitrile substrates to corresponding amide products (Fig. 3.1), a property that has been utilized in industrial

applications such as the production of acrylamide or nicotinamide (Kobayashi and Shimizu, 1998, 2000; Cheng, Xia and Zhou, 2020). These two-subunit enzymes require iron or cobalt cofactors; cobalt NHases can be further characterized into either high or low molecular weight groups (Kobayashi and Shimizu, 2000). The homologous thiocyanate hydrolase (SCNase) reduces the nitrile bond in thiocyanate (Fig. 3.1) (Kobayashi and Shimizu, 2000; Ebbs, 2004); these enzymes are central players in the bioremediation of thiocyanate-contaminated mining and industrial sites (Kobayashi and Shimizu, 2000; Kantor *et al.*, 2015).

While NHases and SCNases exhibit effective nitrile biochemistry, the enzymes within the Class 1 nitrilase subfamily exhibit more targeted substrate specificity towards HCN. The Class 1 nitrilases are members of the greater nitrilase superfamily, which contains over a dozen other subfamilies with varying distributions and substrate specificities (Pace and Brenner, 2001). The superfamily encompasses diverse functionality ranging from lipoprotein modification to nucleic acid catabolism to protein deamination (Pace and Brenner, 2001). Class 1 enzymes are found in diverse bacteria, but also in animals, fungi, and plants (Pace and Brenner, 2001). In plants, these nitrilases are believed to regulate a growth factor precursor; microbial and fungal Class 1 nitrilases catalyze cyanide hydrolysis (Pace and Brenner, 2001; Benedik and Sewell, 2017). The cyanide-degrading Class 1 nitrilases include both cyanide hydratase (CHT), which is thought to be a fungal enzyme, and cyanide dihydratase (CynD), which is thought to be primarily bacterial in distribution (Ebbs, 2004; Benedik and Sewell, 2017). Although CHT and CynD cannot be distinguished from one another at the sequence level by any diagnostic protein sequence motifs, their biochemistry differs slightly (Fig. 3.1).

CynD exhibits the canonical "nitrilase" activity characteristic of the broader superfamily, reducing a nitrile-containing substrate to the corresponding carboxylic acid and ammonium (Podar, Eads and Richardson, 2005; Martínková *et al.*, 2015; Benedik and Sewell, 2017). CHT, however, performs the same biochemistry as NHases, converting a nitrile (HCN) to an amide (formamide); it is believed that this amide product may be subsequently reduced further *in situ* by another nitrilase superfamily enzyme, amidase (Ebbs, 2004; Benedik and Sewell, 2017). Like NHases, nitrilases have been explored for industrial or bioremediation applications (Pace and Brenner, 2001; DeSantis *et al.*, 2002; Gong *et al.*, 2012; Benedik and Sewell, 2017; Chhiba-Govindjee *et al.*, 2019). Despite these practical similarities, and the biochemical convergence of

CHT and NHase, nitrilases and NHases show no detectable homology (O'Reilly and Turner, 2003).

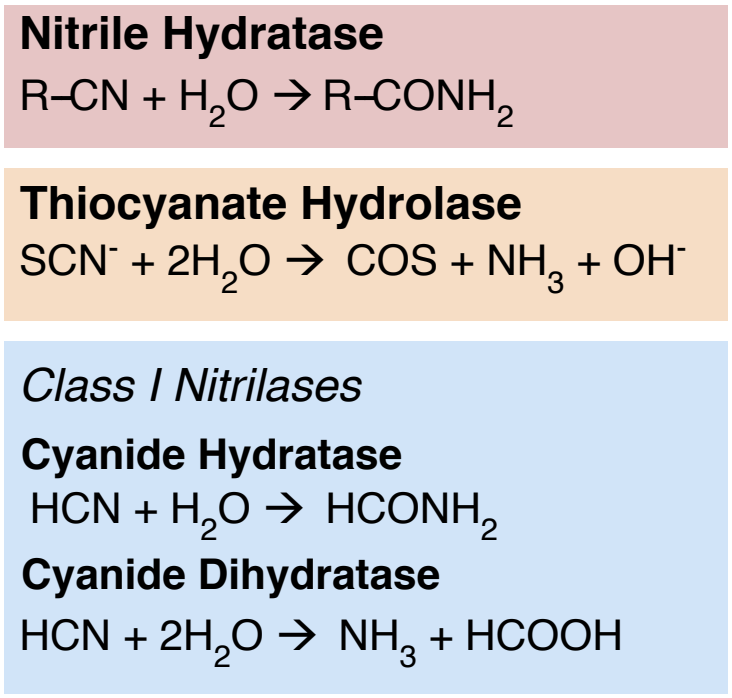


Figure 3.1 Nitrile-hydrolyzing enzymes
Nitrile hydratases (NHases), thiocyanate hydrolases (SCNases), and members of the Class 1 nitrilases all anaerobically hydrolyze nitrile-containing compounds. Nitrilases are nonhomologous to NHases and SCNases but display some convergent function: While cyanide dihydratase exhibits classical nitrilase activity—producing ammonia and a carboxylic acid—cyanide hydratase exhibits NHase-like chemistry, producing an amide product.

The Class 1 nitrilases provide a library of extant enzymes that could be harnessed for cyanide bioremediation or industrial applications. However, this subfamily may also enable exploration of questions about the evolution of early Earth and early microbial life. CHT and CynD are widely distributed across the tree of life and do not require molecular oxygen for catalysis, suggesting they could be reasonably ancient enzymes, and therefore useful targets for molecular clock dating to constrain the age of microbial cyanide metabolism. Such analyses provide independent biological context to constrain the presence of cyanide substrates on early Earth. Beyond improving our understanding of the evolution of the planet's nitrogen cycle, such data

could help constrain hypotheses from prebiotic chemistry that hypothesize a central role for HCN in the origin of life.

HCN has long been viewed as a leading candidate among hypothesized feedstock molecules for abiogenesis. Studies indicate that various abiotic processes, including photochemistry and oligomerization, can transform HCN into a wide variety of nucleic and amino acids including adenine, alanine, glycine, and aspartic acid (Oró, 1961; Oró and Kamat, 1961; Abelson, 1966; Ferris *et al.*, 1978; Draganić *et al.*, 1980). Models suggest that this chemistry could have been supported by atmospheric conditions and a sufficient standing stock of HCN on the early Earth (Zahnle, 1986; Sutherland, 2016). Archaean HCN concentrations are expected to have increased further following the emergence of methanogens, as increased levels of atmospheric methane drive deposition of HCN (Tian, Kasting and Zahnle, 2011). If HCN was indeed a key prebiotic feedstock, and increasingly abundant into the Archaean eon, this would have provided a key substrate for microbial enzyme families that are likely to have well-distributed modern descendants. Reconstructing the evolutionary history of modern HCN-metabolizing enzymes therefore provides a novel approach for constraining the presence and nature of prebiotic chemical and early biochemical systems.

While over six decades of chemical syntheses and planetary modeling has produced extensive data supporting HCN's hypothesized role in the origin or early evolution of life, limited biological rationale exists to constrain these models. Some hypotheses have suggested that HCN biochemistry by nitrilases could actually be quite young, with nitrilases evolving in stem eukaryotes and later undergoing horizontal transfer from plants or animals to sympatric bacteria and archaea (Pace and Brenner, 2001). However, if nitrilases emerged in ancient prokaryotic lineages, this would suggest a much earlier presence and significance for HCN in terrestrial nitrogen biogeochemistry.

3.3 Methods

3.3.1 Sequence sampling and alignments

3.3.1.1 Nitrilases

Representative amino acid sequences were selected for cyanide dihydratase (*Bacillus pumilus,* GenBank AAN77004.1) and cyanide hydratase (*Neurospora crassa,* GenBank

XP_960160.2) (Benedik and Sewell 2012). Each representative sequence was used as a query to search the conservative UniRef90 protein database (Uniprot Consortium, 2021) via NCBI's Basic Local Alignment Search Tool (BLAST)(Camacho *et al.*, 2009). To improve the probability of including homologous sequences, hits were filtered to include only sequences with query sequence identity ≥70%; query sequence coverage ≥95%; and E-value $<10^{-10}$. Sequence lengths were plotted as a Seaborn KDE plot, and only sequences of 200-500 residues in length were included to eliminate outlier skew. Filtered hits were combined, and these sequences aligned using Muscle (Edgar, 2004). The resulting alignment was used to construct a Hidden Markov Model (HMM) in HMMer (Eddy, 2011). The HMMer profile was used to search NCBI's non-redundant protein database (Agarwala *et al.*, 2018) as of January 2019. Resulting hits were filtered for quality; partial hits and hits with e-value $> 10^{-10}$ were removed. Sequence lengths of hits were plotted as a Seaborn KDE plot, and only sequences with length 50-500 residues in length were included to eliminate outlier skew. Resulting hits were subsampled to one per genus. Hits were aligned using Muscle (Edgar, 2004). The resulting alignment was analyzed, and the conserved catalytic triad for nitrilases (Pace and Brenner, 2001) was identified at sites Glu-578, Lys-988, and Cys-1205. Sequences missing data at or surrounding the conserved catalytic sites (29 sequences total) were removed, to improve the likelihood of retaining true nitrilase homologs. Isolated sequences with insertions of ≥ 20 residues (10 sequences total) were also removed to avoid alignment artifacts. The remaining sequences were realigned.

To reduce site-saturation and simplify downstream tree topology for visualization, the initial alignment was further subsampled; the outgroup (identified as the Class 10/Class 13 nitrilases via conserved domain analysis in an initial tree) was subsampled to one member from each clade with class-level monophyly; two members of each clade with class-level monophyly were retained in Metazoa for downstream molecular clock calibrations. The ingroup (identified as Class 1 nitrilases via conserved domain analysis in an initial tree) was subsampled to two members of each clade with order-level monophyly. Two sequences with putative assembly/annotation errors (XP_019577256.1, XP_005975261.1) were removed from the dataset. Remaining sequences were re-aligned in MAFFT (Nakamura *et al.*, 2018), using automatic parameterization. Attempts to reduce noise in the alignment using automated algorithms consistently excluded blocks of subsequence containing well-aligned, likely informative sites, so the alignment was

manually edited to exclude all sites before site 286 and after site 1456. The resulting alignment was used for gene tree construction and visualization.

For molecular clock analyses, Alignment 3 was further subsampled to one representative per clade with phylum-level monophyly, except for groups containing taxa with available fossil calibrations (see below). Sequence XP_023622267.1 was removed due to sequence format incompatibility. The resulting alignment, Alignment 4, was converted into PHYLIP format, with each taxon receiving a unique numerical identifier for Bayesian clock analyses. This alignment was used to construct a maximum-likelihood tree in IQ-TREE, using the optimal substitution model identified under the BIC using the model finder command: LG+F+R7. This tree was used as the starting topology for molecular clock analyses (see below).

3.3.1.2 Nitrile Hydratases (NHases)

Three representative sequences were used for the NHase beta subunit: one iron-type enzyme from *Psuedomonas chlororaphis* (Nishiyama *et al.*, 1991) (GenBank BAA14246.1), one low-molecular weight cobalt-type enzyme from *Rhodococcus rhodocrous* (Kobayashi *et al.*, 1991) (GenBank CAA45711.1), and one high-molecular weight cobalt-type enzyme from *Rhodococcus rhodocrous* (Kobayashi *et al.*, 1991) (GenBank CAA45709.1). Each sequence was used to independently BLAST search the UniRef90 database. Resulting hits were filtered to retain only hits with e-value $< 10^{-10}$, sequence identity ≥ 50%, and query coverage ≥ 75%. The filtered hits were combined and aligned in Muscle. This alignment was used for HMM construction; the resulting HMM was used to search NCBI's non-redundant protein database. Resulting hits were filtered to exclude hits with e-value $>10^{-10}$ and partial sequences. Sequence length of hits was plotted with a Seaborn KDE plot, and sequences <200 and >300 residues in length were excluded from the dataset. The remaining sequences were aligned with Muscle. The alignment was manually checked, and four sequences with apparent misalignment (WP_012455950.1, WP_043079656.1, WP_085127903.1, WP_013808494.1) were removed from the dataset. The remaining sequences were re-aligned, and the resulting alignment used for ML tree construction.

3.3.1.3 Thiocyanate Hydrolases (SCNases)

A representative thiocyanate hydrolase from *Thiobacillus thioparus* (NCBI reference sequence WP_018507189.1) was used as a query to BLAST the conservative UniRef90 protein

database (UniProt). Resulting hits were filtered to retain only hits with e-value < 10^{-10}, sequence identity ≥ 50%, and query coverage ≥ 75%. The resulting hits were aligned in Muscle, and the alignment was used to construct an HMM using HMMer. The HMM was used to search NCBI's non-redundant protein database. Resulting hits were filtered to exclude hits with e-value >10^{-10} and partial sequences. Sequence lengths of hits were plotted using a Seaborn KDE plot, and hits <175 or >300 residues in length were removed from the dataset. Resulting hits were subsampled to one representative per genus. Two sequences with significant missing data (WP_089128740.1, PYJ67070.1) were removed from the dataset. The remaining sequences were aligned with Muscle; the resulting alignment was used for maximum-likelihood (ML) tree construction.

3.3.2 Tree Construction

3.3.2.1 Nitrilases

An initial maximum-likelihood gene tree was constructed for guidance (see above) in IQ-TREE (Nguyen *et al.*, 2015) using the optimal substitution model identified under the Bayes Information Criterion (BIC) with the model finder command: LG+I+G4. A subsampled alignment (see above) was used to construct a maximum-likelihood gene tree (Fig. 3.2) in IQ-TREE using the optimal substitution model identified under the BIC with the model finder command: LG+F+R10. Statistical supports for bipartitions were evaluated using IQ-TREE's approximate-likelihood ratio test (aLRT) and UltraFast Bootstrap (UFBoot) metrics. The resulting gene tree was rooted using Minimal Ancestor Deviation (MAD) rooting (Tria, Landan and Dagan, 2017). This rooting was determined to be largely congruent with a manual rooting between the Class 1 nitrilases and the Class 10/Class 13 nitrilases, as determined by conserved domain analysis (Marchler-Bauer *et al.*, 2015; Lu *et al.*, 2020).

3.3.2.2 NHases and SCNases

The NHase alignment and SCNase alignment were used to construct maximum-likelihood gene trees using IQ-TREE, using the optimal substitution models identified under the BIC by the model finder command: LG+F+R7 (NHase) and WAG+F+R9 (SCNase). Statistical supports for bipartitions were evaluated using IQ-TREE's approximate-likelihood ratio test (aLRT) and UltraFast Bootstrap (UFBoot) metrics. The resulting gene trees were rooted using Minimal Ancestor Deviation (MAD) rooting (Tria, Landan and Dagan, 2017).

3.3.3 Molecular Clock Analyses

3.3.3.1 Constraint optimization

Initial clocks were constructed under a normally-distributed root prior of 3.8-1.0 Ga to evaluate possible diversification ages from the Archean through the mid-Proterozoic. Initial plant, fungal, and animal fossil calibrations were selected from the literature based upon representation of groups in the nitrilase gene tree; this included an initial animal calibration constraining Bilateria with an age of 688-596 Ma (Dos Reis *et al.*, 2015), calibrating the LCA of *Cryptotermes secundus* and *Clupea harengus* in the tree (Fig. 3.S.1). Analysis of the initial clocks revealed minimum predicted root ages just over 1.0 Ga; as a result, subsequent runs relaxed the minimum age on the root prior to 800 Ma, to ensure that posterior root age estimates were not artificially constrained or overdetermined by the selected prior. The initial animal calibration consistently failed to predict accurate age ranges of clades with known divergence times, resulting in unrealistically ancient uncalibrated age estimates for plant and fungal lineages in the tree. Because the Bilateria were under-sampled in some basal groups (Fig. 3.S.1), a more conservative crown Osteichthyes calibration was used for subsequent trees to avoid applying inaccurate biological constraints (see below). Comparative bilaterian-calibrated dates are available as Table 3.S.5.

3.3.3.2 Calibrations

Clocks were run under a permissive, normally distributed root prior of 3.8-1.0 Ga. In addition to the root prior, clocks were run for the prior and posterior distribution with the following additional sets of fossil constraints:

1. No additional calibrations (root prior only)
2. Plant calibration: The Charophyta-Klebsormidiophyta/land plant clade was constrained with an age of 750-590 Ma (Morris *et al.*, 2018), calibrating the last common ancestor (LCA) of *Brassica oleracea* and *Chara braunii* in the tree.
3. Animal calibration: The Osteichthyes (Euteleostomi) were constrained with an age of 444-421 Ma (Dos Reis *et al.*, 2015), calibrating the LCA of *Pelodiscus sinensis* and *Clupea harengus*.
4. Fungal calibrations:

a. Three clades in the tree represented the Sordariomycetes/Dothideomycetes/ Leotiomycetes clade, all constrained with an age of 350-250 Ma (Prieto and Wedin, 2013), calibrating the LCAs of *Sporothrix insectorum* and *Fibularhizoctonia sp. CBS 109695*; *Verruconis gallopava* and *Amorphotheca resinae*; and *Cenococcum* and *Pseudomassariella*, respectively.
b. The Pezizomycetes/Leotiomycetes clade was constrained with an age of 490-400 Ma (Prieto and Wedin, 2013), calibrating the LCA of *Morchella* and *Ustilaginoidea*.
c. The crown Ascomycota were constrained with an age of 680-410 Ma (Prieto and Wedin, 2013), calibrating the LCA of *Hyaloscypha variabilis* and *Pyronema omphalodes CBS 100304*.

5. Plant and Animal calibrations
6. Plant and Fungal calibrations
7. Animal and Fungal calibrations
8. All (Plant, Animal, Fungal) calibrations

3.3.3.2 Models and parameters

To evaluate the effect—and limit the bias—of model selection on age estimates, clocks for all calibration subsets were run under three different clock models: uncorrelated gamma multipliers (Drummond *et al.*, 2006) (UGAM), autocorrelated lognormal (Thorne, Kishino and Painter, 1998) (LN) and the autocorrelated CIR model (Lepage *et al.*, 2006) (CIR). Clocks were run under a fixed C20 empirical profile mixture model (Si Quang, Gascuel and Lartillot, 2008). Clocks were run with two chains, and convergence was assessed as effective size ≥ 50 and variable discrepancies ≤ 0.30 for all parameters.

3.3.3.3 Variance analyses

Variance and population standard deviations were calculated for the three Sodariomycetes/Dothideomycetes/Leotiomycetes calibrations using the following formulae:

$$\mathrm{var_n} = \frac{(\mu - \alpha_\mathbf{n})^2}{3}$$

$$\mathrm{sd_n} = \sqrt{var_n}$$

$$var_m = \sum(var_n)$$
$$sd_m = \sum(sd_n)$$

where

var_n = variance for clade *n*'s mean age estimate, in a given model subset
μ = mean of all clades' mean age estimates for the model subset
α_n = mean age estimate for clade *n*
sd_n = population standard deviation for clade *n* in a given model subset
var_m = summed variance for the model subset
sd_m = summed population standard deviation for the model subset

3.4 Results

3.4.1 Nitrilases

The maximum-likelihood gene tree for nitrilases reveals diverse representatives from all three domains of life (Fig. 3.2). Conserved domain analysis indicates that the outgroup to the Class 1 nitrilases represented in the tree comprises primarily members of the Class 10 and Class 13 nitrilases. This outgroup contains a monophyletic group of metazoans; at the class level, these taxa roughly follow predicted species tree topology for representatives of the Osteichthyes (Euteleostei) and shallower crown bilaterians (Fig. 3.S.1).

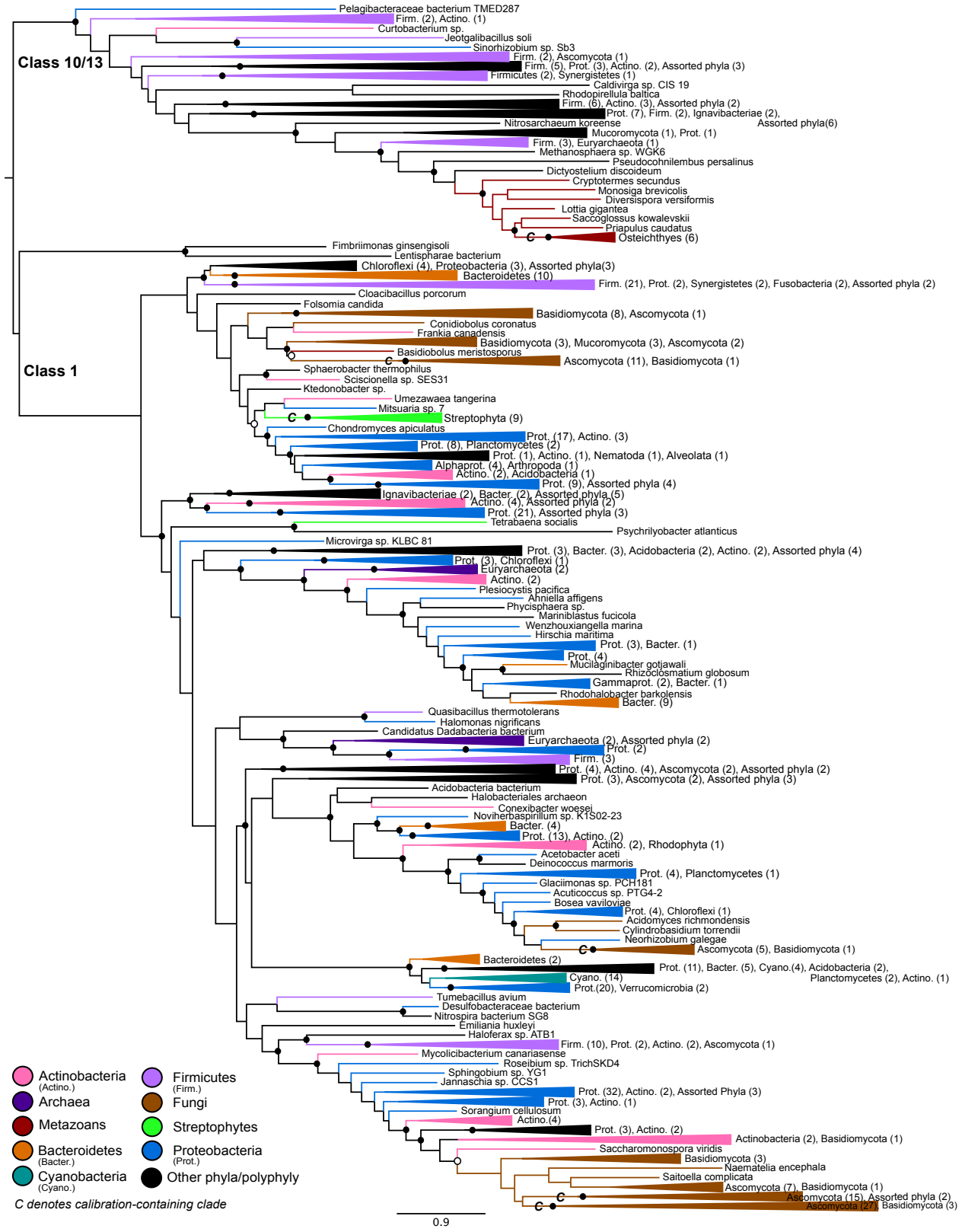


Figure 3.2 Nitrilase gene tree

A maximum-likelihood gene tree for Class 1 nitrilases reveals broad diversity and deep conservation through all three domains of life. Conserved domain analysis indicates the outgroup (above) consists primarily of Class 10 and Class

13 nitrilases. Clades containing over 67% representation of a single group are color-coded as shown; taxon counts and sub-phylum level taxonomies are labeled on the tree. Isolated representatives of given phyla (due to subsampling for visualization) are described as "assorted phyla" within paraphyletic clades. Node supports are colored by the values of both approximate likelihood ratio test (aLRT) and rapid bootstrap (bb) values: Strong support with both aLRT/bb values ≥ 90 (black), weak support with both values ≤ 50 (white). Unlabeled nodes have either intermediate aLRT/bb support (one or both values between 50 and 90) or conflicting support (one value ≥ 90 and the other ≤ 50).

The Class 1 nitrilases represented in the tree are also distributed across domains, though they are particularly enriched within bacteria and fungi, consistent with previously described distributions of cyanide dihydratase and cyanide hydratase. Proteobacteria, Ascomycota, and Basidiomycota are particularly well-represented compared with other phyla.

Several groups in the nitrilase gene tree show phylum- or class-level monophyly, including a clade of plants that is consistent with known species tree topology for Streptophyta (Fig. 3.S.2) (Morris *et al.*, 2018). Domain analysis suggests that this clade contains the Nit6803 nitrilases. Such monophyly suggests a deep history of vertical inheritance of Class 1 nitrilases. However, multiple instances of polyphyly and inconsistency with prokaryotic and eukaryotic species tree topologies suggests extensive horizontal gene transfer as well. Notably, eukaryotic groups do not place together within the tree, but instead are independently nested within larger bacterial groups. This topology is consistent with independent, horizontal transfers of a nitrilase ancestor from bacteria into eukaryotic groups.

Molecular clock analyses including all calibrations predict a Class 1 nitrilase ancestor age between 1.88 and 1.02 Ga (Fig 3.3). The estimated range and mean age for the Class 1 ancestor varies by clock model: LN predicts the oldest age range (1.88-1.45 Ga) and mean (1.63 Ga), CIR predicts the youngest age range (1.02-1.25 Ga) and mean (1.13 Ga), and UGAM offers intermediate estimates for age range (1.57-1.18 Ga) and mean (1.32 Ga) of the Class 1 LCA. Calibration-specific analyses suggest that LN is the most sensitive of the three clock models to calibration choice, and provides the least consistent ages depending on calibration subset (Fig. 3.4). The calibration-specific analyses also indicate that the plant calibration is most-informative for recovering the expected fossil-constrained ages of the other calibrated clades, while the animal calibration often overestimates and the fungal calibration routinely underestimates these expected constrained ages (Table 3.S.3).

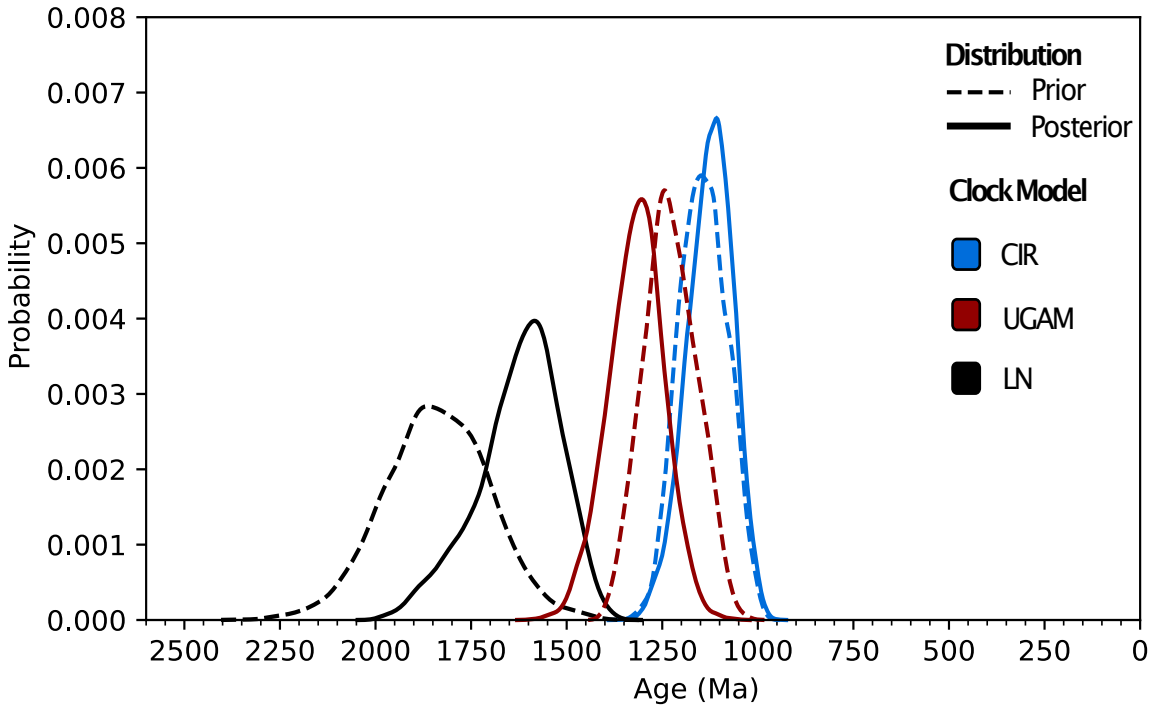


Figure 3.3 Class I nitrilase ancestor age probability densities
Prior and posterior age distributions for the last common ancestor of Class 1 nitrilases are shown for uncorrelated gamma (UGAM), log-normal (LN), and CIR process clock models. All distributions shown are run using all animal, plant, and fungal fossil calibrations (see Methods) and a permissive root prior of 3800-800 Ma.

Variance analyses of the predicted age for the triplicated fungal clades in the tree suggest that the CIR clock model minimizes variance in the majority of calibration subsets (Table 3.1). LN does not minimize variance for these groups under any prior or posterior age distribution.

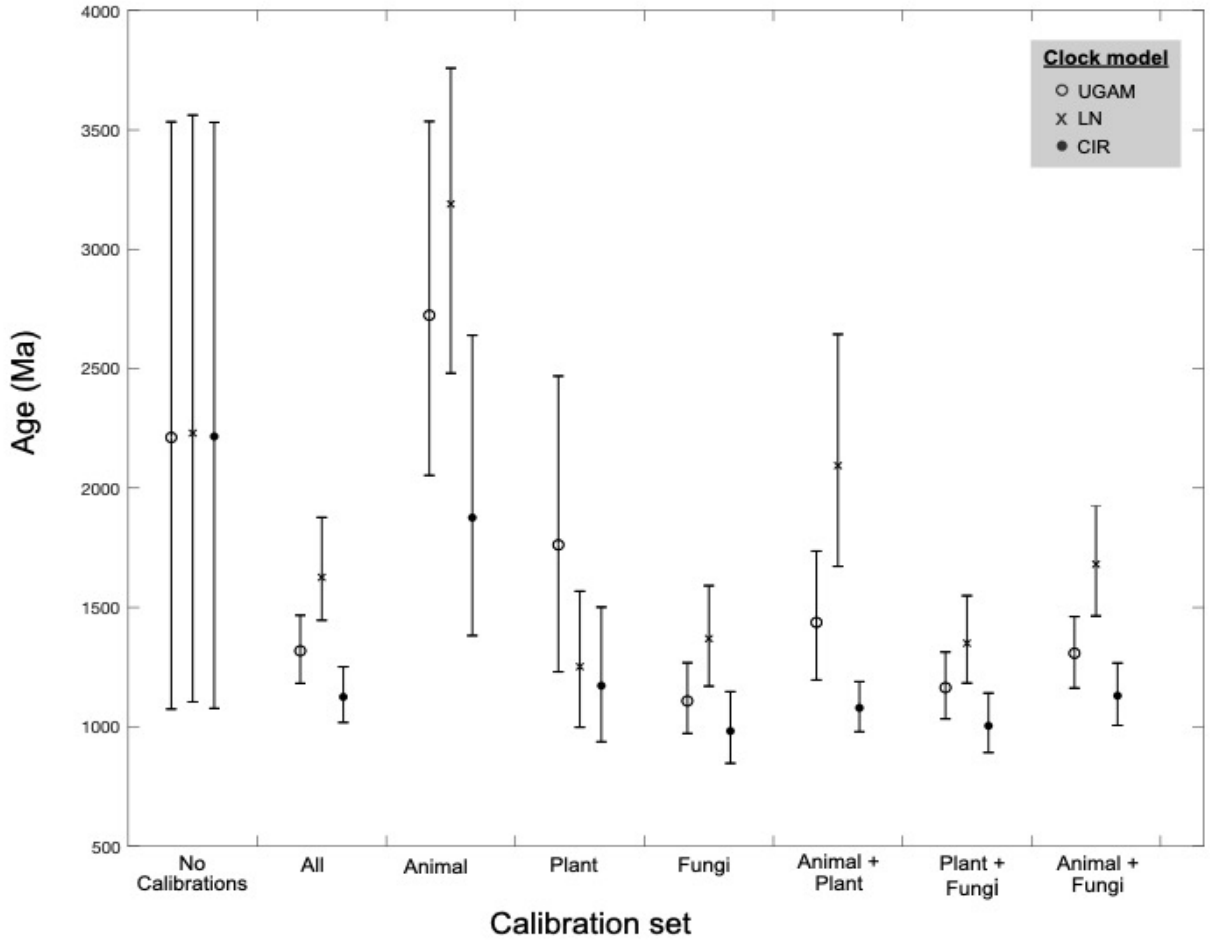


Figure 3.4 Molecular clock age estimates by model and calibration set
Mean age (markers) and age range (bars) estimates are shown for Class I nitrilase ancestor molecular clock posterior distributions by calibration subset. Three clock models were tested for each calibration subset (see marker type).

Table 3.1 Age variance calculated for triplicate fungal clades by clock model

Calibration subset	Prior/Posterior	Variance, UGAM	Variance, LN	Variance, CIR
Animal	Prior	10073.13	19266.38	6959.36
Animal	Posterior	4938.04	24696.60	6125.44
Plant	Prior	4357.46	2038.49	1949.36
Plant	Posterior	2371.88	2861.05	2030.22
Animal+Plant	Prior	7569.54	7141.20	4055.56
Animal+Plant	Posterior	3843.81	9206.00	4797.94
None	Prior	7420.20	10244.72	1183.75
None	Posterior	3681.61	12255.40	288.61

Minimized variance shown in gray for each subset

3.4.2 Nitrile Hydratases and Thiocyanate Hydrolases

Gene trees for NHases (Fig. 3.5) and SCNases (Fig. 3.6) reveal similar taxonomic distributions for these homologous enzymes. Both enzymes appear exclusively prokaryotic in distribution. Most orthologs are within members of the Proteobacteria, with particular enrichment among Alphaproteobacteria; Actinobacteria are also well-represented. Both NHases and SCNases are also found in members of the Cyanobacteria, Firmicutes (specifically Bacilli), and halophilic Archaea. Though both trees contain multiple groups with phylum-level monophyly (including within Bacilli and Archaea), the trees also show substantial polyphyly, especially for Proteobacteria, which are represented at all depths of the tree. This limited sampling of monophyletic groups and an overall paucity of well-established fossil representatives suggest that the NHase and SCNase trees are poor candidates for molecular dating.

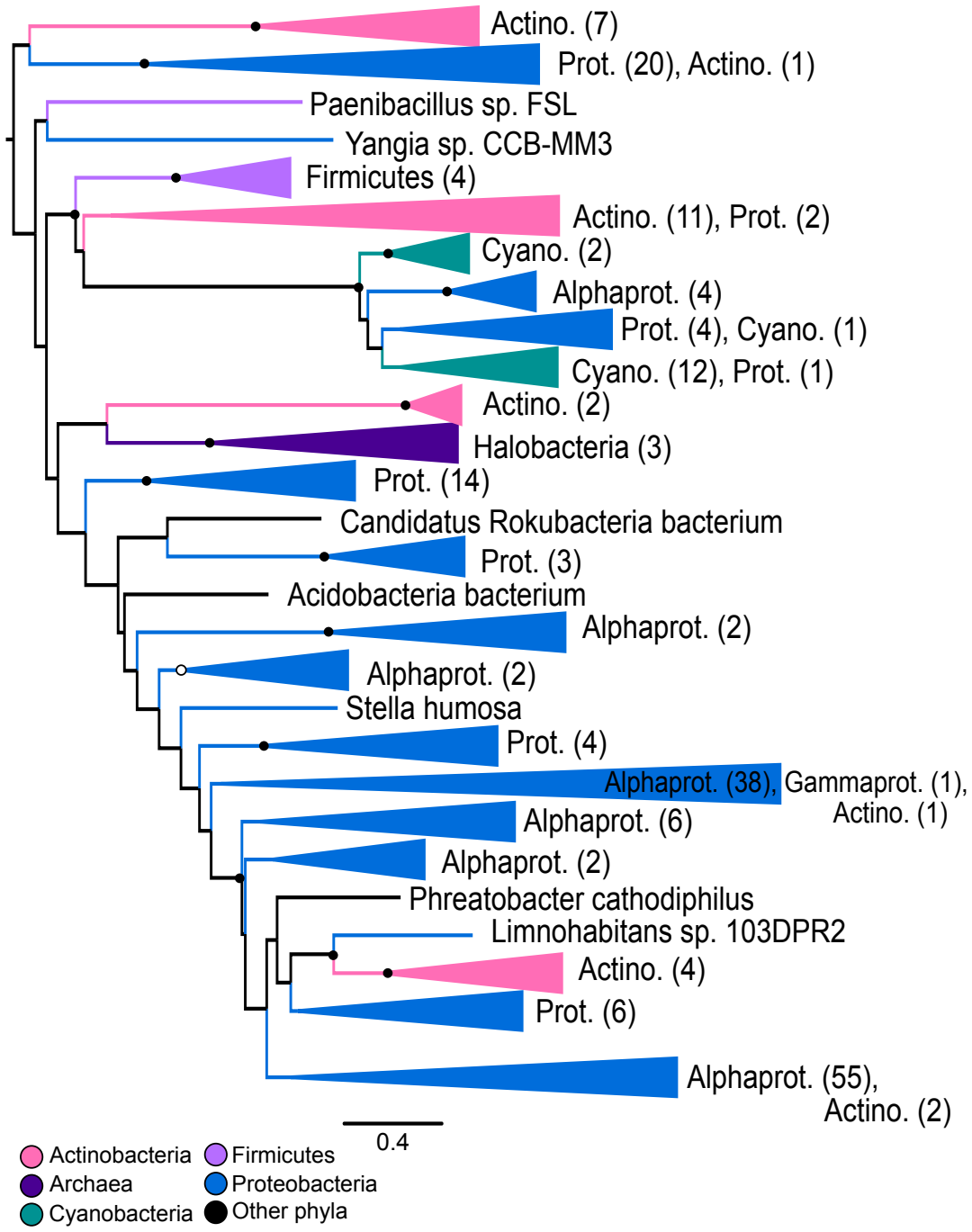


Figure 3.5 Nitrile hydratase gene tree
A maximum-likelihood gene tree for nitrile hydratases reveals an exclusively prokaryotic distribution, particularly enriched in Class Alphaproteobacteria. Clades containing over 67% representation of a single group are color-coded as shown; taxon counts and sub-phylum level taxonomies are labeled on the tree. Node supports are colored by the values of both approximate likelihood ratio test (aLRT) and rapid bootstrap (bb) values: Strong support with both aLRT/bb values ≥ 90 (black), weak support with both values ≤ 50 (white). Unlabeled nodes have either intermediate aLRT/bb support (one or both values between 50 and 90) or conflicting support (one value ≥ 90 and the other ≤ 50).

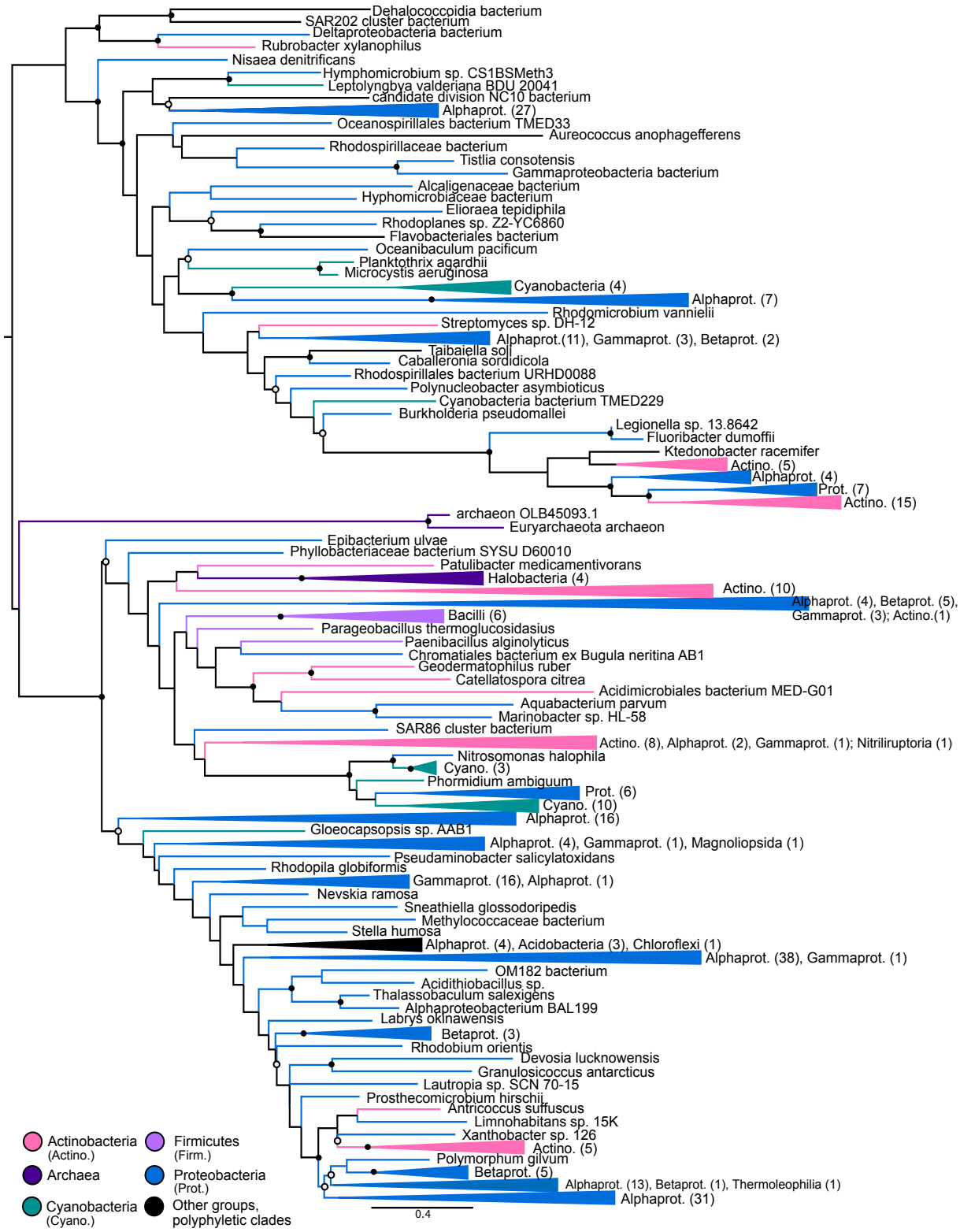


Figure 3.6 Thiocyanate hydrolase gene tree

A maximum-likelihood gene tree for thiocyanate hydrolases reveals an exclusively prokaryotic distribution, particularly enriched in Class Alphaproteobacteria and Phylum Actinobacteria. Clades containing over 67% representation of a single group are color-coded as shown; taxon counts and sub-phylum level taxonomies are labeled

on the tree. Node supports are colored by the values of both approximate likelihood ratio test (aLRT) and rapid bootstrap (bb) values: Strong support with both aLRT/bb values ≥ 90 (black), weak support with both values ≤ 50 (white). Unlabeled nodes have either intermediate aLRT/bb support (one or both values between 50 and 90) or conflicting support (one value ≥ 90 and the other ≤ 50).

3.5 Discussion

Nitrilases, NHases, and SCNases have all been previously identified as enzymes with significant potential in both bioremediation and biotechnology. Comprehensive gene trees provide a top-level view of the extant diversity of these enzymes, which may in turn help identify new candidates of commercial or environmental interest. Previously unstudied enzymes that are closely related to those with established biocatalytic or biosynthetic value may be useful in expanding the available library of enzymes for targeted nitrile metabolism. Perhaps more importantly, divergent enzymes or host strains may identify candidates with previously-unexplored biochemical potential; for example, previous studies have reported nitrilases or NHases that are more thermostable, more pH-stable, more efficient, or more readily inducible than previous enzyme candidates (Gong *et al.*, 2012; Kuhn *et al.*, 2012; Pei *et al.*, 2013; Supreetha *et al.*, 2019; Shen *et al.*, 2021).

The topology of the nitrilase tree is inconsistent with hypotheses suggesting that the nitrilase ancestor originated in eukaryotes and was subsequently transferred into bacteria. While it is not possible to completely rule out multiple lineage-specific losses of an ancestral eukaryotic nitrilase gene, this history would not be parsimonious. The phylogenetic distance observed between eukaryotic groups, nested placement of eukaryotic groups within bacterial-enriched clades, and taxonomic monophyly observed within sampled plants and animals all support a bacterial origin with subsequent horizontal transfer into eukaryotic groups.

Molecular clock dating also supports a prokaryotic origin for the nitrilase ancestor. The CIR clock model predicts the youngest mean Class 1 ancestor age (1.13 Ga) and the youngest-bound age distribution (1.02 Ga) among models run with all fossil calibrations (Figs. 3.3, 3.4). While some of the older mean and age range estimates predicted by other clock models and calibration subsets in this work do, in fact, exceed older-bound age estimates for the last eukaryotic common ancestor at ~1.9 Ga (Penny *et al.*, 2014), this youngest Proterozoic age estimate does not itself preclude a nitrilase origin in ancestral eukaryotes. However, existing hypotheses suggest that ancestral nitrilases diversified in animals and plants *before* horizontal transfer into microbial

lineages (Pace and Brenner, 2001). The last common ancestors of crown Viridiplantae, crown Metazoa, and crown Ascomycota are predicted to be, at maximum, 972 Ma (Morris *et al.*, 2018), 834 Ma (Dos Reis *et al.*, 2015), and 671 Ma (Prieto and Wedin, 2013), respectively—all younger than the lowest bound of the youngest predicted age distribution for nitrilases. The predicted age of these nitrilases is therefore inconsistent with the diversification of these enzymes within eukaryotes. Together with the gene tree topologies, these results support a prokaryotic, Proterozoic origin for cyanide-reducing nitrilases.

Alterations in age estimates based on different calibrations or clock models highlights the importance of rigorous calibration and clock model testing to identify and mitigate possible biases. For the nitrilases, a plant-only calibration appears to be informative in constraining the ages of other fossil-calibrated clades; animal-only calibrations push the overall age estimates older, while fungal-only calibrations push these estimates younger (Fig. 3.4, Table 3.S.3).

Across different combinations and subsets of fossil calibrations, UGAM and CIR clock models offer age estimates that are more consistent with one another than the LN model (Fig. 3.4), suggesting that UGAM and CIR models may provide more reliable dates. This result is further supported through variance analyses between the three fungal clades with the same calibration. The LN model fails to minimize the variance for any relevant substitution subset, for prior or posterior distributions (Table 3.1); this is consistent with the apparent calibration sensitivity and variability observed for LN in the calibration-specific analyses. Together, these data suggest that LN may provide the least-reliable age estimates of the three tested clock models—and CIR may provide the most-reliable estimates.

3.6 Conclusions and Future Directions

The results of this study suggest an early origin for nitrilases, especially as compared with eukaryotic-origin scenarios. But evidence for a Proterozoic emergence for Class 1 nitrilases does not constrain the presence or biochemistry of cyanide for the earliest prokaryotic life during the Archaean. This does not rule out the possibility that cyanide biochemistry was important on the early Earth. The nitrilase ancestor must predate the diversity represented within the chronogram; the estimated ages are for the last common ancestor of Class I nitrilases, and therefore represent a younger bound on the crown age for these extant enzymes. The clocks in this work—whether due

to sampling bias, lineage-specific losses, or other processes—may simply be unable to approach the true ancestral nitrilase age. Another possibility is that other nitrile-reducing enzymes, such as NHase or SCNase, predate Class 1 nitrilases, but are difficult to date due to a limited distribution or a complex history of transfer or loss.

Alternatively, ancient enzymes that are not adapted for HCN or nitrile metabolism in modern organisms may have metabolized HCN on the early Earth. For example, modern nitrogenases—which recent studies suggest are Archaean in origin (Stüeken *et al.*, 2015; Parsons *et al.*, 2021)—are known to promiscuously reduce HCN at their active site (Li, Burgess and Corbin, 1982; Conradson *et al.*, 1989; Lowe *et al.*, 1989). Cyanide metabolism by nitrogenases has been proposed to reflect a "relic" chemistry from early Earth systems, perhaps reflecting environmental pressures in a cyanide-enriched environment (Pickett *et al.*, 2004). Now that it has been established that cyanide reduction was solved reasonably early by prokaryotes, additional work will be required to determine whether a different group of enzymes with this biochemical capacity emerged even earlier than the nitrilases.

The approaches used here to date the nitrilases highlight generalizable principles that can be applied to molecular clock evaluations of other complex microbial gene histories. First, the rigorous calibration testing underscores the benefits of including as many diverse calibrations as are biologically supported. For ancient groups with extensive transfer history, including the nitrilases, maximizing these informative priors is essential for dating deeper, unconstrained lineages within the tree. But as established here, these calibrations vary in their predictive power and precision, and should always be considered independently to determine relative effects and predict biases in a systematic fashion. Similarly, this work highlights the importance of independently evaluating rate models to determine model effect on age estimates. The variance analyses performed for the nitrilase clocks specifically underscores the utility of leveraging duplicate transfers into calibration-constrained clades as a tool for quantifying and optimizing model precision.

3.S Supplementary Tables and Figures

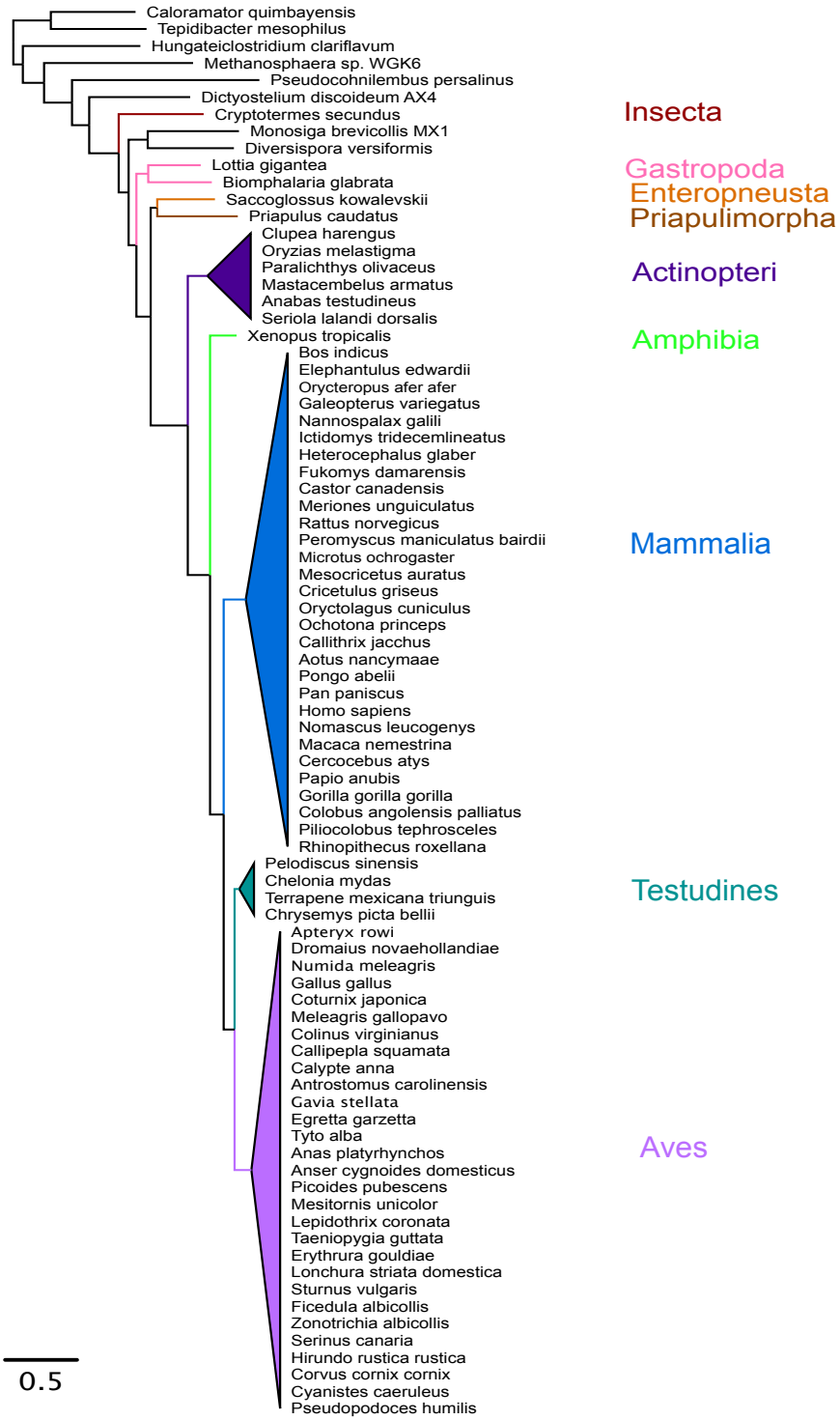


Figure 3.S.1 Pre-subsampling nitrilase subtree for metazoans
An expanded view of the metazoan subtree shows deep sampling in mammals and birds (Aves), with additional orthologs in turtles (Testudines), bony fish (Actinopteri), amphibians, and a small number of invertebrates. Orthologs are categorized as Class 10 nitrilases by conserved domain analysis. The placement of chordate groups in this subtree largely follows species tree organization.

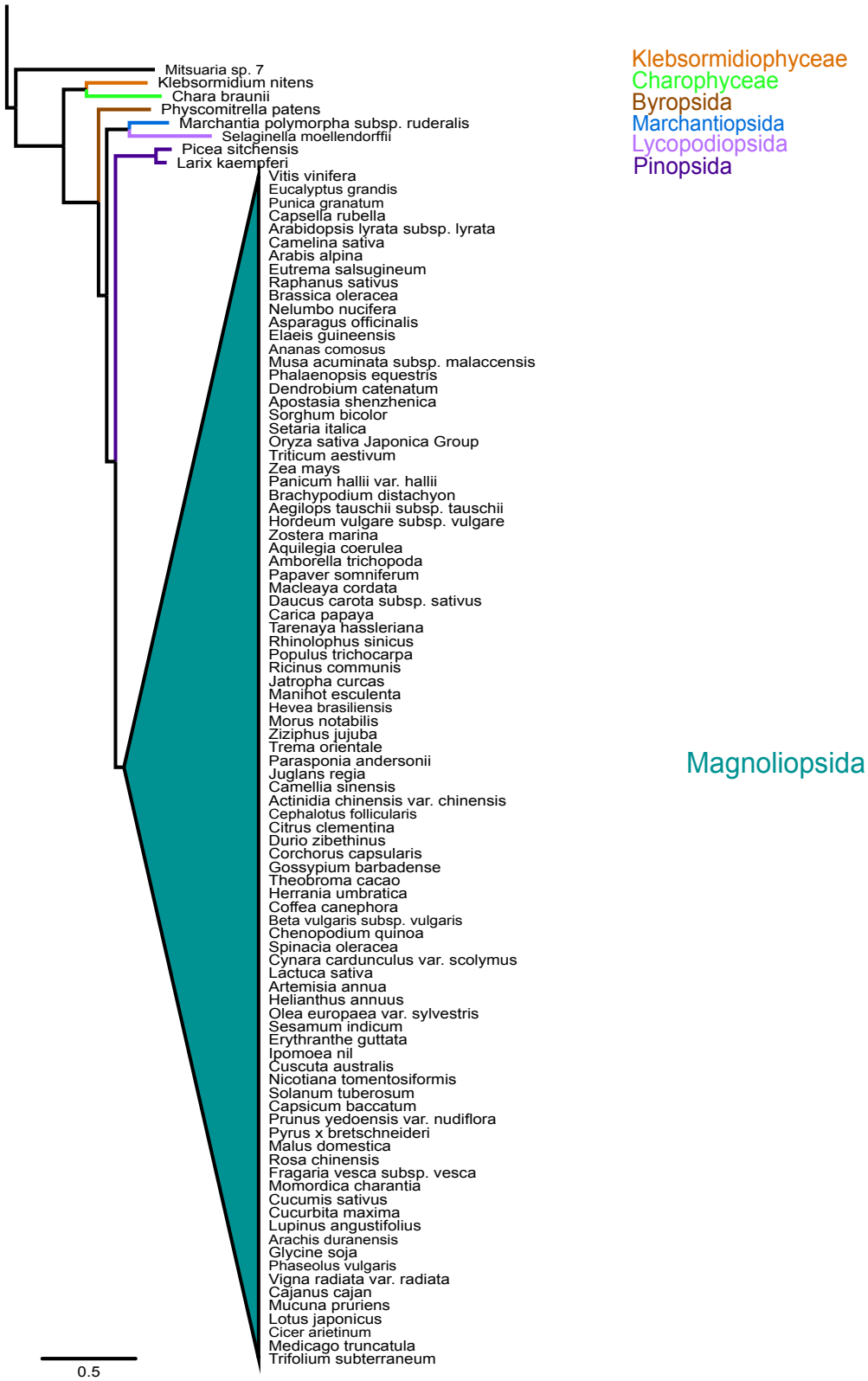


Figure 3.S.2 Pre-subsampling nitrilase subtree for streptophytes
An expanded view of the plant subtree shows deep sampling in Magnoliopsida. The relative placement of basal Viridiplantae roughly follows expected species tree divergence.

Table 3.S.3 Age estimates for fossil-constrained clades by other calibrations

Node/ Calibration	Constraint age input (Ma), range	Age with Osteichthyes calibration (Ma), range/mean Posterior, *prior(pr)*	Age with plant calibration (Ma), range/mean Posterior, *prior(pr)*	Age with fungal calibrations (Ma), range/mean Posterior, *prior(pr)*	Age with Plant+Fungal calibrations (Ma), range/mean Posterior, *prior(pr)*	Age with Plant+Animal calibrations (Ma), range/mean Posterior, *prior(pr)*	Age with Fungal+Animal calibrations (Ma), range/mean Posterior, *prior(pr)*	Age with root-only calibrations (Ma), range/mean Posterior, *prior(pr)*
Osteichthyes	467-445	--	U: 422-165/272 *356-156/ 244(pr)* L: 178-68/114 *173-66/111(pr)* C: 423-192/ 294 *430-211/ 310(pr)*	U: 235-117/171 *220-110/157 (pr)* L: 199-69/128 *220-87/147 (pr)* C: 329-166/244 *330-168/ 243(pr)*	U: 245-124/ 178 *220-115/164(pr)* L: 195-70/120 *206-75/133(pr)* C: 338-166; 249 *337-180; 257(pr)*	--	--	U: 594-154/342 *592-16/ 350 (pr)* L: 437-84/212 *378-90/210(pr)* C: 964-253/ 561 949-249/552
Plantae	750-590	U: 1735-707/1145 *1792-762/ 1216(pr)* L: 2266-1316/1803 *2234-1256/1761(pr)* C: 1640-796/ 1138 *1536-754/ 1093(pr)*	--	U: 660-320/469 *620-318/461(pr)* L: 936-592/760 *1059-616/833(pr)* C: 713-466/ 590 *696-465/ 583(pr)*	--	--	U: 746-366/ 546 *682-360/510(pr)* L: 1110-703/895 *1202-762/966(pr)* C: 785-523/659 *783-519/655 (pr)*	U: 1658-415/ 934 *1732-451; 988 (pr)* L: 2115-615/ *2021-600/ 1233(pr)* C: 2195-640/ 1342 *2162-620/1310 (pr)*
Fungi (1)	350-250	U: 946-470/687 *1035-528/ 752(pr)* L: 746-323/502 *606-272/ 423(pr)* C: 740-301/479 *689-327/483(pr)*	U: 648-278/437 *634-295/440(pr)* L: 293-122/195 *266-109/174(pr)* C: 434-205/314 *480-221/325(pr)*	--	--	U: 763-407/567 *809-434/ 615(pr)* L: 497-199/316 *446-179/ 298(pr)* C: 570-254/ 376 *496-265/ 369 (pr)*	--	U: 924-257/554 *1034-272/614 (pr)* L: 637-149/345 *509-130/284(pr)* C: 1332-377/ 805 *1242-344/ 740 (pr)*
Fungi (2)	350-250	U: 1059-494/737 *1104-543/ 807(pr)* L: 1107-531/782 *911-495/695(pr)* C: 954-413/638 *960-486/673(pr)*	U: 696-289/460 *696-323/471(pr)* L: 400-188/285 *372-181/265(pr)* C: 552-277/ 398 *552-305/424(pr)*	--	--	U: 851-405/601 *895-470/660 (pr)* L: 690-326/483 *658-305/ 468(pr)* C: 674-348/ 506 *652-382/ 513 (pr)*	--	U: 1006-272/584 *1070-301/644 (pr)* L: 966-244/538 *825-231/485 (pr)* C: *1282-359/ 764* *1360-388/822(pr)*
Fungi (3)	350-250	U:1150-606/855 *1104-543/ 807(pr)* L: 1159-614/870 *947-548/735(pr)* C: 934-460/652 *931-456/643(pr)*	U: 790-364/550 *822-415/593 (pr)* L: 430-236/322 *372-181/265(pr)* C: 561-315/ 418 *532-318/411(pr)*	--	--	U: 1208-663/908 *1027-637/ 818(pr)* L: 732-382/543 *658-305/ 468(pr)* C: 725-433/ 535 *609-400/ 493 (pr)*	--	U: 1143-329/695 *1291-354; 810 (pr)* L: 1052-282/606 *830-250/510 (pr)* C: 1291-373/ 783 *1302-380; 795 (pr)*
Fungi (4)	490-400	U: 1482-775/1092 *1638-821/ 1172(pr)* L: 1479-811/1133 *1504-667/ 1074(pr)* C: 1338-626/ 910 *1280-620/ 904 (pr)*	U: 1020-472/700 *966-482/698(pr)* L: 594-307/443 *548-316/425(pr)* C: 740-353/549 *730-448/578(pr)*	--	--	U: 936-536/712 *1265-731/ 968(pr)* L: 898-429/627 *838-423; 612(pr)* C: 875-520/ 697 *830-495/ 651 (pr)*	--	U: 1450-414/ 884 *1554-443/963 (pr)* L: 1402-379/813 *1288-362/762(pr)* C: 1816-512/ 1091 *1776-487/ 1056 (pr)*
Fungi (5)	680-410	U: 1562-854/1169 *1577-942/ 1225(pr)* L: 1481-817; 1105 *1212-742/ 958(pr)* C: 1154-591/815 *1140-587/ 803(pr)*	U: 1064-505/747 *1015-529/733(pr)* L: 531-299/403 *481-253/352(pr)* C: 691-398 / 524 *730-448/578(pr)*	--	--	U: 1262-747/ 977 *1274-774/ 1019(pr)* L: 922-513/690 *856-436/ 631(pr)* C: 824-524/ 656 *748-499/ 612(pr)*	--	U: 1531-452/ 942 *1560-484/997(pr)* L: 1249-357/752 *1047-316/650(pr)* C: 1598-474/983 *1592-476/ 987 (pr)*

Clock Models: U--UGAM; L--LN; C--CIR clock models

Red: estimated age range does NOT intersect calibration age range for node / Black: estimated age range intersects calibration age range, mean age does not; Blue: estimated node age range and mean age intersect calibration age range

Table 3.S.4 Class 1 Nitrilase ancestor age prior and posterior distributions/mean estimates by calibration and model subsets

MODEL	CALIBRATIONS	PRIOR CLASS 1 LCA AGE RANGE (Ma)	PRIOR CLASS 1 LCA MEAN AGE (Ma)	POSTERIOR CLASS 1 LCA AGE RANGE (Ma)	POSTERIOR MEAN CLASS 1 LCA AGE (Ma)
UGAM	Root cal only	3598-1104	2329	3534-1074	2212
UGAM	Fungi	1182-958	1058	1268-972	1108
UGAM	Plant	2306-1226	1689	2469-1230	1762
UGAM	Animal	3652-2137	2808	3536-2053	2724
UGAM	Plant + Animal	2867-1891	2374	2809-1789	2259
UGAM	Plant + Fungi	1189-979	1079	1313-1033	1164
UGAM	Fungi + Animal	1356-1135	1228	1461-1162	1308
UGAM	All calibrations	1356-1098	1229	1468-1182	1318
LN	Root cal only	3542-1114	2228	3562-1104	2230
LN	Fungi	1930-964	1548	1591-1170	1369
LN	Plant	1608-1036	1296	1568-998	1252
LN	Animal	3753-2568	3252	2759-2481	3190
LN	Plant + Animal	2805-1774	2240	2644-1672	2094
LN	Plant + Fungi	1860-1347	1572	1549-1183	1350
LN	Fungi + Animal	2258-1673	1942	1925-1464	1681
LN	All calibrations	2112-1572	1839	1877-1446	1626
CIR	Root cal only	3581-1075	2223	3532-1077	2216
CIR	Fungal	1136-892	1001	1147-847	982
CIR	Plant	1506-967	1195	1501-937	1172
CIR	Animal	2510-1327	1800	2640-1382	1876
CIR	Plant + Animal	1680-1208	1414	1736-1196	1437
CIR	Plant + Fungi	1139-916	1021	1141-892	1004
CIR	Fungi + Animal	1267-1006	1139	1267-1006	1130
CIR	All calibrations	1257-1026	1143	1251-1018	1125

Table 3.S.5 Comparative Bilaterian posterior estimates for Class I ancestor and nitrilase root ages

MODEL	CALIBRATION SET, BILATERIA	POSTERIOR ROOT AGE RANGE, Ma	POSTERIOR MEAN ROOT AGE, Ma	POSTERIOR CLASS 1 LCA AGE RANGE, Ma	POSTERIOR MEAN CLASS 1 LCA AGE, Ma
UGAM	Fungi	1287 – 1018	1114	1283 – 1015	1133
UGAM	Plant	2517 – 1276	1828	2509 – 1274	1822
UGAM	Animal	3135 – 2017	2526	3216 – 2007	2515
UGAM	Plant + Animal	2730 – 1847	2255	2718 – 1842	2246
UGAM	Plant + Fungi	1315 – 1050	1171	1312 – 1047	1167
UGAM	Fungi + Animal	1551 – 1238	1386	1547 – 1233	1382
UGAM	ALL	1574 - 1272	1411	1562 – 1268	1405
LN	Fungi	1640 – 1182	1399	1636 – 1180	1395
LN	Plant	1597 – 1048	1289	1591 – 1044	1284
LN	Animal	3730 – 2477	3135	3718 – 2465	3123
LN	Plant + Animal	2646 – 1706	2122	2634 – 1699	2114
LN	Plant + Fungi	1517 – 1191	1344	1513 – 1188	1340
LN	Fungi + Animal	2108 – 1590	1827	2097 – 1582	1818
LN	ALL	1960 – 1538	1738	1.953 – 1535	1732
CIR	Fungi	1151 – 1002	1047	1146 – 997	1044
CIR	Plant	1531 – 1017	1210	1526 – 1013	1206
CIR	Animal	2013 – 1326	1625	2006 – 1323	1620
CIR	Plant + Animal	1729 – 1220	1443	1723 – 1216	1438
CIR	Plant + Fungi	1165 – 1002	1056	1161 – 998	1053
CIR	Funimal	1354 –1061	1185	1348 – 1055	1179
CIR	ALL	1326 – 1074	1178	1322 – 1068	1173

Chapter 4. Active site diversity and potential for substrate promiscuity in ancestral nitrogenases

The contents of this chapter have been submitted as Schwartz, S.L.; Garcia, A.K.; Kaçar, B.; Fournier, G.P. "Early nitrogenase ancestors encompassed novel active site diversity." (2022)

4.1 Abstract

Ancestral sequence reconstruction (ASR) infers predicted ancestral states for sites within sequences and can constrain the functions and properties of ancestors of extant protein families. Here, ASR is used to explore the evolutionary history of the dinitrogen-fixing nitrogenases. We compare the likely sequences of inferred nitrogenase ancestors to extant nitrogenase sequence diversity and find that the most likely combinations of ancestral states for key substrate channel residues are not represented in extant sequence space, and rarely found within a more broadly defined physiochemical space. The results suggest that the earliest ancestors of extant nitrogenases likely had substrate channels comprising amino acid arrangements not observed in extant sequence diversity. These alternative compositions may reflect selection in ancient environments. For example, modern nitrogenases can promiscuously reduce hydrogen cyanide (HCN); increased HCN availability on the early Earth could have selected for nitrogenases with both HCN and dinitrogen affinity. These results highlight ASR's utility as an *in-silico* tool for constraining hypotheses about deep-time ancestral enzyme function and guiding testable in-vitro and in-vivo experiments.

4.2 Introduction

Ancestral sequence reconstruction (ASR) is widely used for modeling the genetic states and predicting the metabolic capacity of ancient enzyme variants. Maximum likelihood reconstructions combine an optimized substitution model with extant sequence data and tree topology to calculate the likelihood of alignment states for all internal nodes in a tree (Merkl and Sterner, 2016; Selberg, Gaucher and Liberles, 2021). This approach enables exploration of targeted structural or functional hypotheses about ancestral proteins *in silico,* and is also an essential starting point for enzyme resurrection *in vitro* or *in vivo* activity assays (Garcia and Kaçar, 2019;

Liberles *et al.*, 2020). Because most microbial metabolisms have no preserved record in deep time, early evolution must be inferred from modern genomic and biochemical data (Barnabas, Schwartz and Dayhoff, 1982). By modeling protein "fossils," ASR provides a means of gleaning biological data about ancestral enzymes and organisms that can be considered in context with environmental data, such as biosignatures or atmospheric modeling (Merkl and Sterner, 2016; Garcia and Kaçar, 2019; Selberg, Gaucher and Liberles, 2021). This approach is particularly useful when studying ancient, geochemically relevant enzyme families.

Nitrogen-fixing nitrogenases serve as the lynchpin of Earth's biological nitrogen cycle, and have been extensively studied for their essential role in global biogeochemistry (Raymond *et al.*, 2004; Canfield, Glazer and Falkowski, 2010; Sickerman, Hu and Ribbe, 2019). Modern nitrogenases are categorized by their metal cofactors (Raymond *et al.*, 2004; Sickerman, Hu and Ribbe, 2019; Garcia *et al.*, 2020; North *et al.*, 2020). All diazotrophs contain molybdenum nitrogenase (Nif), with its iron-molybdenum cofactor (FeMoCo); some nitrogen fixers also express alternative vanadium nitrogenase (Vnf) or iron-only nitrogenase (Anf) in molybdenum-limited situations (Boyd *et al.*, 2011; Sickerman, Hu and Ribbe, 2019; Garcia *et al.*, 2020). Phylogenetic analyses, molecular clock data, and cofactor-type-specific nitrogen isotope fractionations support the parsimonious inference that Nif is the oldest, ancestral form of nitrogenase, and recent studies suggest that the ancestor of this enzyme may date to the early Archaean eon (Boyd *et al.*, 2011; Stüeken *et al.*, 2015; Garcia *et al.*, 2020; Parsons *et al.*, 2021).

In addition to their unique capacity for dinitrogen (N_2) fixation, modern nitrogenases can promiscuously reduce other linear, triple bond-containing molecules at the FeMoCo active site (Fani, Gallo and Liò, 2000; Seefeldt *et al.*, 2013). One such off-target substrate is hydrogen cyanide (HCN) (Li, Burgess and Corbin, 1982; Conradson *et al.*, 1989). Reduction of HCN by nitrogenase produces methane, acrylamide, and ammonia (Li, Burgess and Corbin, 1982)—thereby providing an alternative means of nitrogen fixation by nitrogenases. Overall, this reduction is a precariously balanced chemical process. While HCN is a substrate for nitrogenases, the cyanide anion (CN^-) alone is an inhibitor; this inhibition can be reversed by synergistic binding of carbon monoxide (CO) to FeMoCo (Li, Burgess and Corbin, 1982; Pickett *et al.*, 2004).

It has been suggested that this cyanide promiscuity reflects "relic" chemistry of nitrogenases—for example, cyanide detoxification in highly-reduced early environments (Fani, Gallo and Liò, 2000; Pickett *et al.*, 2004). As the nitrile triple bond in HCN has marginally lower

dissociation energy than that of N_2 (937 versus 946 kJ/mol) (Benson, 1965; Dean, 1999), HCN reduction could conceptually also have provided a biochemical "stepping-stone" for N_2 reduction, which is one of the most ATP-expensive metabolisms known to biology (Raymond *et al.*, 2004). HCN has also garnered significant attention as a candidate feedstock molecule for abiogenesis, as it can be abiotically converted into several key nucleic and amino acids (see Chapter 3) (Oró, 1961; Oró and Kamat, 1961; Ferris and Orgel, 1966; Ferris *et al.*, 1978; Miyakawa, Cleaves and Miller, 2002; LaRowe and Regnier, 2008; Pearce and Pudritz, 2015; Sutherland, 2016). Studies support the delivery or production and maintenance of a stock of HCN on the early Earth, and modeled prebiotic HCN concentrations would have been sufficient to support abiogenesis (Tian, Kasting and Zahnle, 2011; Ferus *et al.*, 2017; Todd and Öberg, 2020). Throughout the Archaean eon, HCN concentrations are predicted to have been even higher: Increased concentrations of atmospheric methane, bolstered by newly-emerged methanogens, would have driven substantial HCN-producing photochemistry (Zahnle, 1986; Tian, Kasting and Zahnle, 2011). Evidence that early nitrogenases were HCN-metabolizing enzymes would be consistent with observations that certain enzymes had more multifunctional, generalist ancestors (Siddiq, Hochberg and Thornton, 2017), and would support the hypothesized presence and significance of HCN in early ecosystems.

A higher affinity for HCN in ancestral nitrogenases would likely require different amino acid residues in the substrate-binding part of the enzyme (the substrate channel), while preserving catalytic activity. As ancestral sequence reconstructions are inherently probabilistic, individual sites may be inferred to have several possible ancestral states with non-negligible probabilities. Variation across sampled extant diversity allows the combinations within predicted ancestor sequences to explode, producing a far greater number of plausible sequence ancestors than can be experimentally tested. Here we use ASR to reconstruct the amino acid identities of alignment sites directly interacting with the substrate binding channel in the common ancestor of the active site-containing nitrogenase subunit (*nif/vnf/anfD*). A library of plausible sequence ancestors for this collection of sites was generated, and their relative probabilities calculated. We then compared predicted ancestral sequence combinations with the sequence diversity of sampled extant nitrogenases. Each modern nitrogenase sequence serves as a null hypothesis for ancestral sequence variation at the same sites; that is, there is no reason to expect functional differences from a predicted ancestral enzyme with sequence combinations represented in modern sequence space. Conversely, where ancestral sequences diverge from those of modern enzymes, meaningful

differences could be inferred. Comparing ancestral and extant sequence space for key alignment sites can therefore constrain which evolutionary hypotheses and conserved residues, if any, should be targeted for experimental validation such as ancestral enzyme resurrection.

We find that likely ancestral active site sequence and physicochemical state combinations are rarely, if ever, sampled within extant nitrogenase diversity. These results permit the possibility that early nitrogenase ancestors had distinct substrate binding channels from modern enzymes, and therefore may have sampled alternative phenotypes as well.

4.3 Methods

An initial dataset of extant nitrogenase D-subunit homologs was constructed from hits gathered using BLASTp to search NCBI's non-redundant protein database with a query sequence from *Azotobacter vinelandii* (WP_012698832.1). Hits were included for molybdenum (NifD), vanadium (VnfD), and iron-only (AnfD) nitrogenases (385 sequences), as well as outgroup sequences from light-independent protochlorophyllide oxidoreductases (Boyd *et al.*, 2011) (385 sequences). Initial homologs were aligned using MAFFT using auto-parameterization, which selected iterative refinement method L-INS-I (Nakamura *et al.*, 2018). Outgroup sequences were subsampled to retain one sequence per phylum for major monophyletic groups (12 sequences total). A second nested outgroup (5 sequences) of Group IV-A nitrogenase-like homologs (NfaD) (Raymond *et al.*, 2004; North *et al.*, 2020) was profile-aligned to the existing alignment with MAFFT, providing the additional ancestral node required to reconstruct the *nif/anf/vnf* ancestor as an internal node in the tree. The tree was rooted along the split between nitrogenases and other oxidoreductases. Maximum-likelihood tree construction and subsequent ancestral sequence reconstruction were performed in IQ-TREE (Nguyen *et al.*, 2015), under the best-fit model identified by the Bayesian Information Criterion (BIC), LG+R9. In addition to sequence identity, the physicochemical "state" was assessed for each amino acid using the following categories for residues: Acidic (Glu, Asp); Aliphatic (Ile, Leu, Met, Val); Amine (Gln, Asn); Aromatic (Trp, Tyr, Phe); Basic (Lys, Arg); Small Hydroxyl (Ser, Thr); Tiny (Ala, Gly).

4.4 Results

Of 24 sites known to play a role in substrate docking or coordination in modern *Azotobacter vinelandii* Nif (Seefeldt *et al.*, 2013; Smith *et al.*, 2014), five sites showed hypothesized variation (>1 residue with p > 0.10) in the predicted ancestral enzyme (Table 4.1).

Table 4.1 ASR predicted residue for key substrate channel sites

Conserved residue in alignment (*Av*)	**ASR Nif/Anf/Vnf ancestor residue, p(residue)**	**Other residues p > 0.10?**	**Putative role in modern Nif active site**	**Corresponding residue in *Av nifD*** (Seefeldt *et al* 2013, Smith *et al* 2014)
Val-182	Val, 0.99998	No	Modulates substrate access to catalytic core/active site	α-70^{Val}
His-350	His, 0.99999	No	H-bond with FeMoCo, flexible (S2A or S2B)	α-195^{His}
Tyr-497	Tyr, 0.99997	No	Substrate channel formation; gating residue	α-281^{Tyr}
Arg-493	Arg, 0.99991	No	H-bond with FeMoCo in substrate channel (surface "flap" for substrate access)	α-277^{Arg}
His-604	His, 1	No	Substrate channel formation; gating residue	α-383^{His}
Cys-491	Cys, 1	No	Coordinates FeMoCo	α-275^{Cys}
His-745	His, 0.99998	No	Coordinates FeMoCo	α-442^{His}
Arg-224	Arg, 0.99912	No	Coordinates FeMoCo	α-96^{Arg}
Gln-346	Gln, 0.99995	No	Coordinates FeMoCo	α-191^{Gln}
Gly-181	Gly, 0.99934	No	Switch control for α-70^{Val} side chain	α-69^{Gly}
Ser-347	Ser, 0.99888	No	N_2 interaction in substrate channel	α-192^{Ser}
Asn-155	Asn, 0.99972	No	Lines substrate channel	α-49^{Asn}
Gly-178	Gly, 0.99994	No	Lines substrate channel	α-66^{Gly}
Val-183	Val, 0.99824	No	Lines substrate channel	α-71^{Val}
Ser-345	Ser, 0.99918	No	Lines substrate channel	α-190^{Ser}
Leu-348	Ala, 0.29253	**Gln, 0.27519; Leu, 0.16779; Lys, 0.16716**	Lines substrate channel	α-193^{Leu}
His-351	His, 0.99992	No	Lines substrate channel	α-196^{His}
Asn-354	Asn, 0.99946	No	Lines substrate channel	α-199^{Asn}
Ser-494	Ser, 0.99993	No	Lines substrate channel	α-278^{Ser}
Met-495	Ala, 0.74242	**Met, 0.17555**	Lines substrate channel	α-279^{Met}
Asn-496	Thr, 0.79331	**Asn, 0.14693**	Lines substrate channel	α-280^{Asn}
Gly-576	Gly, 0.87529	**Ala, 0.12413**	Lines substrate channel	α-357^{Gly}
Phe-602	Phe, 1	No	Lines substrate channel	α-381^{Phe}
Ala-603	Ala, 0.47988	**Gly, 0.46717**	Lines substrate channel	α-382^{Ala}

For comparison with ASR results, sequence diversity for these five variable sites was evaluated in extant nitrogenases and selected outgroup sequences (Figure 4.1). Extant variation at these sites roughly maps onto the major groups of nitrogenase in the gene tree. Sites 398 and 495 in the alignment highlight the distinction between Group I Nif (comprising well-studied

proteobacterial and cyanobacterial *nifD* sequences, including that of *Azotobacter vinelandii*) and Group II Nif (comprising sequences from Clostridia, Bacteroidetes, and methanogenic archaea, among others) (Raymond *et al.*, 2004). Group I Nif conserves Leu at site 395, while Group II conserves primarily Ala (though fewer cases of Gly, and individual substitutions of Ser and Cys are also observed). Group I Nif contains primarily Met at site 495, while Group II Nif primarily displays Ile.

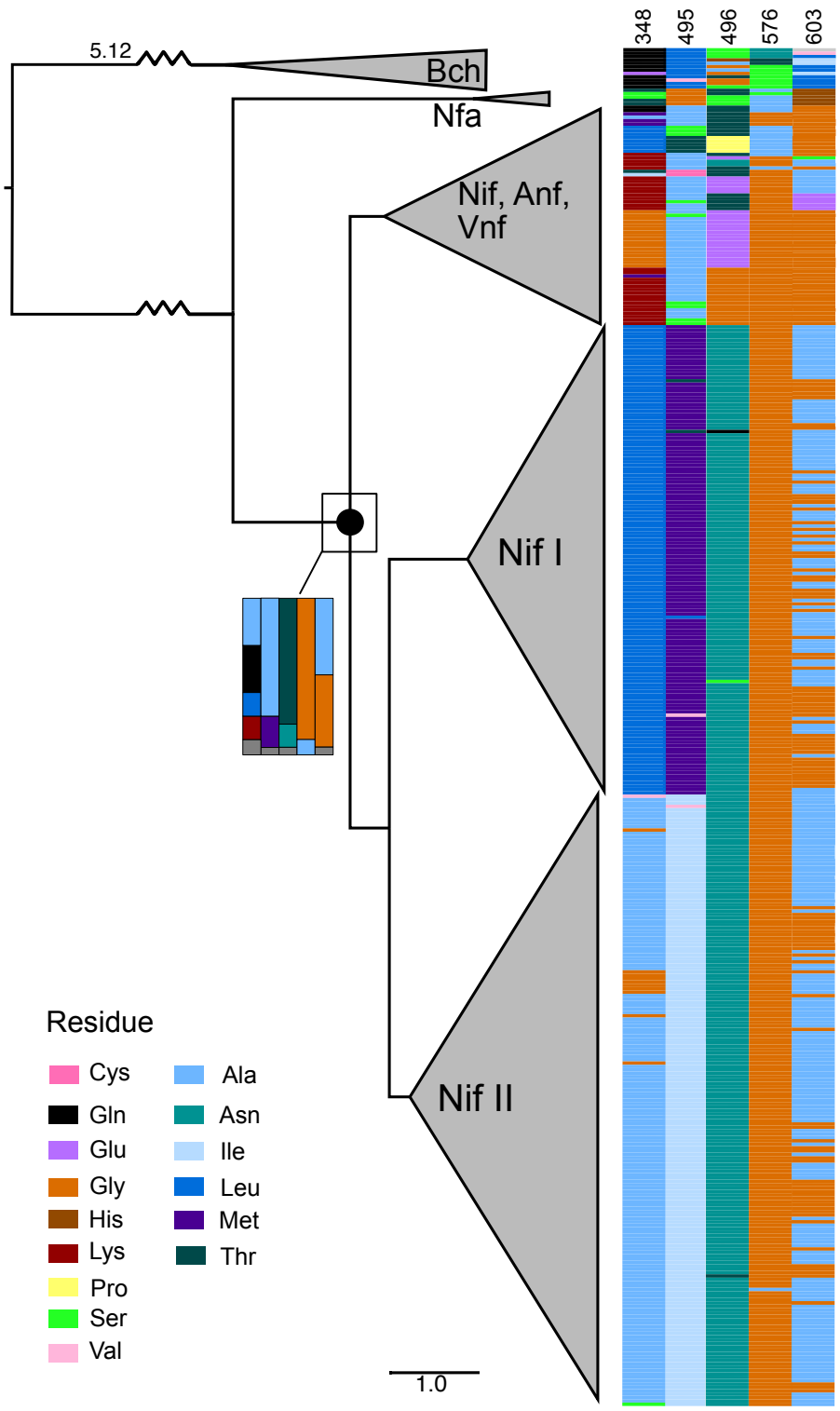


Figure 4.1 Extant nitrogenase sequence diversity
A maximum-likelihood gene tree for the D subunit of extant nitrogenases and outgroup sequences (Nfa: Group IV nitrogenase-like enzymes; Bch: bacteriochlorophyll/protochlorophyllide oxidoreductases), displaying the residue state at each of five variable sites in the *Azotobacter vinelandii* NifD substrate channel. The last common ancestor evaluated for nitrogenases is marked with a black node; reconstructed site likelihoods for all residues $p > 0.10$ are shown by residue color in the inset bar chart.

Variation at site 496, meanwhile, highlights the split between Nif I/II nitrogenases and the clade including alternative (Vnf, Anf) nitrogenases (which also contains several divergent Nif sequences): With very few exceptions, Group I and II nitrogenases conserve Asn at this site, while alternative clade sequences are poorly constrained, and appear to most frequently express Glu, Thr, and Gly at this site.

Predicted combinations of residue states were evaluated in the ASR for the ancestor of Nif, Anf, and Vnf (Fig. 4.1). The 64 evaluated ancestral sequence combinations of five variant sites all had predicted joint likelihood ratios higher than 0.00025% (Fig. 4.3). Only 5/64 (7.8%) predicted ancestral sequence combinations represented a combination of the five variant sites found in extant sequences (Fig. 4.3). Of these five combinations, only one—the second-highest likelihood sequence, AATGG—was in the top-20 most likely predicted ancestral combinations. Four of the five sequences were "rare," observed in no more than two extant sequences; the only non-rare sequence (LMNGG) had the lowest likelihood of these five. Several sequence patterns inferred as high likelihood in reconstructed ancestors were not observed within extant nitrogenases. For example, in sampled extant sequences, Gln-348 does not co-occur with Gly-576; Ala-348 does not co-occur with Met-495; and Leu-348 does not co-occur with Ala-495. Ala-348 and Ala-495 only co-occur in one extant sequence.

It is possible that many of the observed differences between extant sequences and inferred ancestors involve only conservative amino acid substitutions that are not likely to substantially affect substrate binding or phenotype. To evaluate this, amino acids were recoded by general physicochemical properties, to identify radical changes of amino acid type (Fig. 4.2). When reassessed by residues' physicochemical categories, 29/64 (45%) of reconstructed ancestors contained substitutions yielding a combination not represented by extant sequences *or* physicochemical types (Fig. 2). Of the 35 ancestors represented among extant physicochemical types, 28/35 (80%) were rare. The 32 most-likely ancestors all displayed physicochemical types that were rare or nonexistent within extant sequences. The top 10 most-likely ancestors sampled rare physicochemical types only found within the divergent Nif/Anf/Vnf clade, not within Nif I/II.

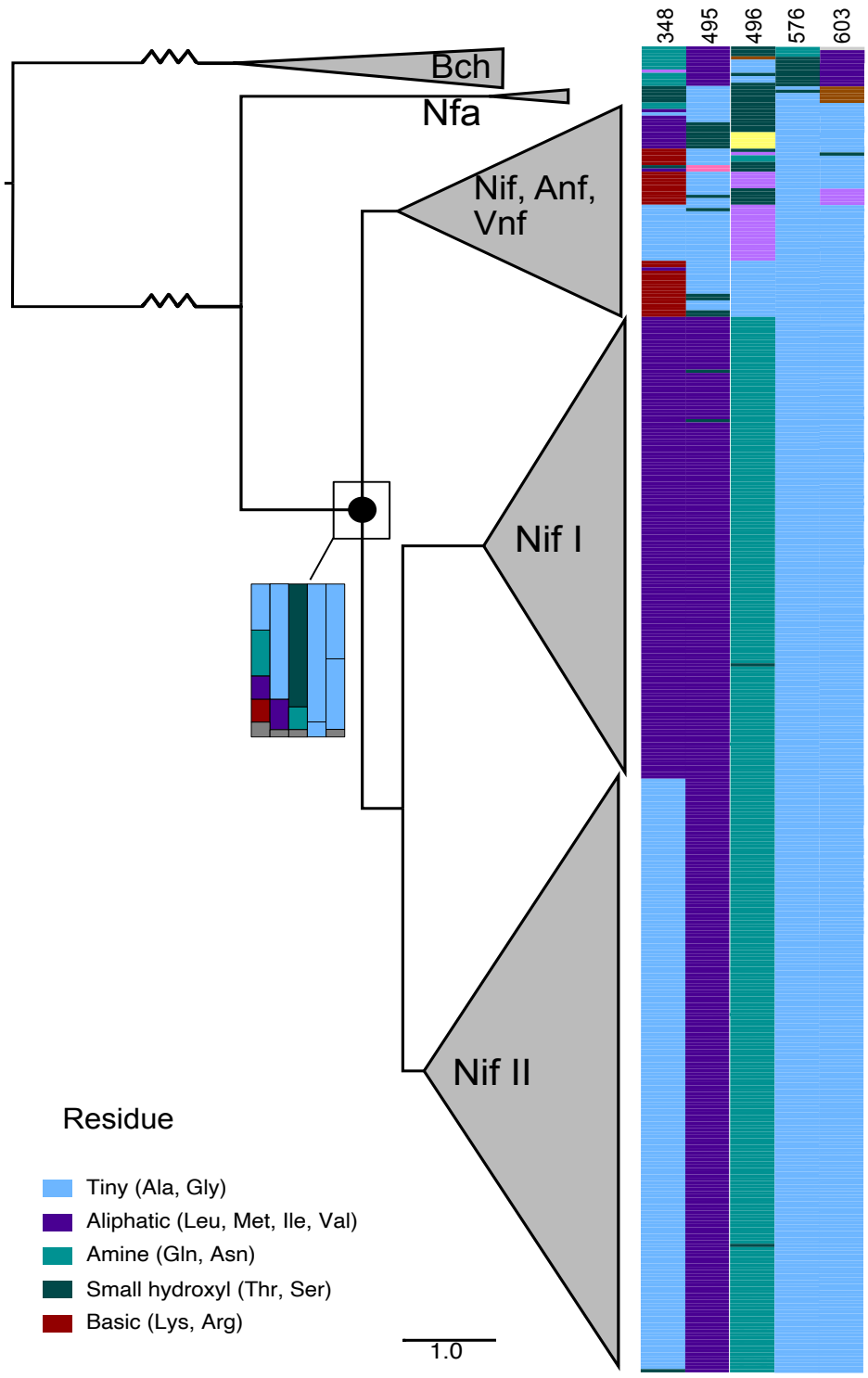


Figure 4.2 Extant nitrogenase diversity by physicochemical group
A maximum-likelihood gene tree for the D subunit of extant nitrogenases and outgroup sequences (Nfa: Group IV nitrogenase-like enzymes; Bch: bacteriochlorophyll/protochlorophyllide oxidoreductases), displaying the physicochemical group of each residue present at five variable sites in the *Azotobacter vinelandii* NifD substrate channel. The last common ancestor evaluated for nitrogenases is marked with a black node; reconstructed site likelihoods for all residues $p > 0.10$ are shown by physicochemical type color in the inset bar chart.

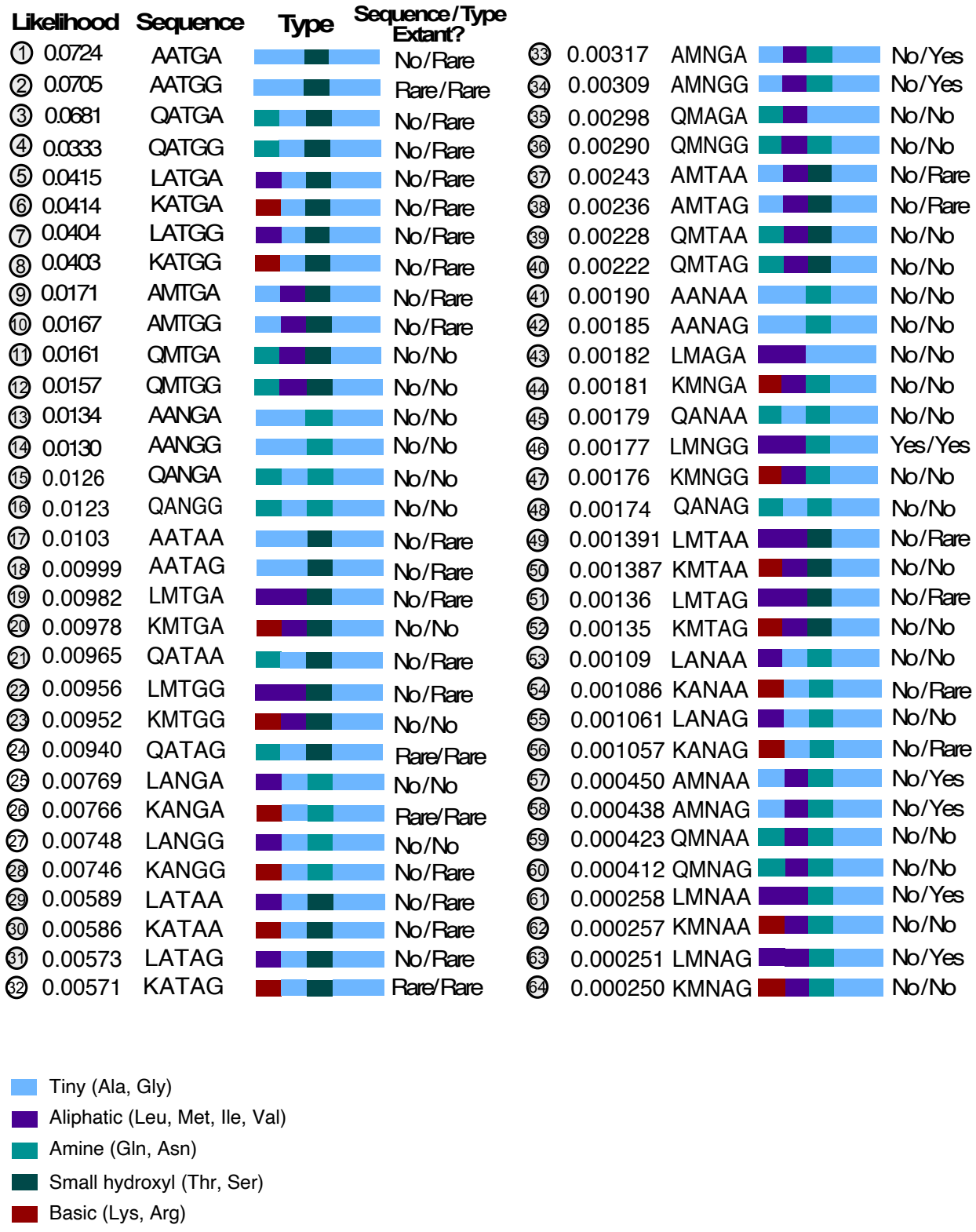


Figure 4.3 Ancestral nitrogenase sequence and state likelihoods
Likelihood ratios, amino acid sequence, and residue physicochemical states are shown for all combinations of the five variable sites (alignment sites 348, 495, 496, 576, and 603) predicted in the nitrogenase substrate channel ASR. "Rare" states occur in no more than two (2) extant sequences.

4.5 Discussion

The ASR results suggest that the inferred ancestor of Nif, Vnf, and Anf nitrogenases most likely contained a substrate channel sequence distinct from that found in extant sequence space. 22 of the 23 highest likelihood sequence combinations of the five variant substrate channel ancestral sites are not observed within known modern nitrogenases; the only represented sequence pattern occurs in only one observed extant enzyme, that of *Chloroflexales bacterium ZM16-3*. The top ten most-likely ASR residue combination physicochemical types were also represented by very few extant sequences, and only within the divergent group of Nif that also contains the alternative nitrogenases. Divergent homologs within this clade may sample ancestral sequence patterns to a greater extent than type I or II nitrogenases. If so, characterizing these extant genes could provide useful analogs for ancestral phenotypes.

The conservative substitutions observed in the highest likelihood predicted ancestors are unlikely to radically influence substrate affinity or binding, at least as compared with the extant enzymes in the Nif/Anf/Vnf group. Because both modern Vnf (Seefeldt *et al.*, 2013) and Nif can reduce HCN, there is no reason to believe that the ASR-predicted ancestors could not as well. Nevertheless, the observed departures from extant sequence space in many likely ancestor sequences opens the possibility of phenotypically different substrate binding in nitrogenase ancestors. Recent work has indicated that even a small number of substitutions, even outside the active site, can elicit different ancestral substrate affinities (Tatsaki *et al.*, 2021). This justifies experimental investigation of HCN binding in reconstructed ancestor candidates, especially those displaying the novel residue type combinations sampled above.

Extant nitrogenases that remain unsampled and excluded from the tree could contain currently missing high-likelihood sequence combinations in the ASR. Additionally, likelihood ratios calculated from only a handful of sites do not represent overall sequence likelihoods, as site likelihoods are calculated independently. Thus, the most likely ancestral site combinations in the substrate channel are unlikely to identify the sequence in the most likely overall ancestor. However, the likelihood calculations for these substrate binding sites may be more informative for inferring ancestral phenotype changes than the overall ancestral sequence of the enzyme. These results could be further contextualized by evaluating more ancient ancestors—for example, the last common ancestor of nitrogenases and maturases (Garcia, Kolaczkowski and Kaçar, 2022)—and less-likely ancestral candidates with methods such as Bayesian sampling and laboratory

reconstruction (Merkl and Sterner, 2016; Garcia and Kaçar, 2019; Selberg, Gaucher and Liberles, 2021). But this likelihood-based, nitrogenase-exclusive approach provides the most conservative evaluation of functional hypotheses based on modern enzyme variants.

4.6 Conclusions

These ASR data suggest that ancestral nitrogenases would have had different substrate channel composition than that sampled by modern enzymes. None of the predicted substitutions would obviously structurally prevent binding of HCN, with its similar geometry to the target N_2 substrate. However, additional work should be performed to confirm this, and to determine whether the predicted ancestral substitutions might in fact enhance affinity for HCN. The results of this ASR could also be used to guide similar investigations for other promiscuously reduced nitrogenase substrates.

Together, the results highlight the value of ASR as a first step in evaluating hypotheses about enzyme evolution. In some cases, ASR itself may prove sufficient to refute certain evolutionary hypotheses; for example, predicted ancestral states may be biologically incompatible with hypothesized substrate interactions or protein folding. Such results would expediently preclude additional experimental work. Alternatively, where ASR fails to identify combinations of key ancestral residues absence from extant diversity, the null hypothesis of uniform phenotype across the reconstructed history should be preferred. Whether functionally-relevant variation is inferred or not, ASR will efficiently identify key sites to target for mutagenesis or analysis in downstream *in vitro* or *in vivo* analysis—such as the novel sequence and type combinations sampled in this work. Additionally, ASR will provide quantitative context for qualitative evolutionary questions, and help frame additional hypotheses for future investigation.

Chapter 5. Conclusions and Future Directions

5.1 Key takeaways

Together, the projects explored in this thesis highlight the complex histories and underexplored diversity of microbial nitrogen metabolisms.

Chapter 2 demonstrates that even within the well-studied core pathways of the modern nitrogen cycle, significant unreported enzyme diversity exists. Novel enzyme domain architectures observed within well-described enzyme families suggest that this diversity could be further explored experimentally to identify functional differences or fitness advantages in specific environments (such as deep subsurface sediments).

Chapter 3 shows the utility in phylogenetically mapping nitrogen metabolisms that are excluded from the core nitrogen cycle. In an applied context, such studies identify extant enzyme variability that could be harnessed for biocatalysis or bioremediation. Geobiologically, when combined with rigorous quantitative modeling and geological constraints, these results provide insight into the age and presence of early Earth metabolisms, including those with no obvious or direct geological or geochemical record.

Chapter 4 explores the role that core enzymes in the modern nitrogen cycle may have played in noncanonical nitrogen metabolisms in the past. This project displays the utility of *in silico* sequence reconstruction in leveraging extant genomic information to evaluate hypotheses of deep-time enzyme evolution. Such likelihood-based approaches are one of the few useful tools for quantitatively evaluating ancient metabolisms, where there is often little or no direct evidence of ancestral biochemistry. More broadly, this chapter highlights the relevance of enzyme promiscuity in expanding biochemical capacity in both extant and ancient ecosystems.

Overall, these combined investigations reveal a larger genetic diversity of extant nitrogen metabolisms (within and external to the canonical nitrogen cycle); an older origin for early microbial nitrogen metabolisms; and greater variability in ancestral nitrogen metabolic sequence space than has been previously hypothesized or shown. These results suggest that despite the extensive, thorough body of research on nitrogen cycle dynamics, the data accumulation curve for nitrogen biogeochemistry is nowhere near saturated. Further study into the distribution and diversity of microbial nitrogen metabolisms should continue to yield new insights with relevance to basic and applied microbial sciences. This is especially true for noncanonical, understudied

metabolisms that are considered to have limited impact in modern ecosystems, but which may have played more significant or different roles on the early Earth.

5.2 Challenges and uncertainties

5.2.1 Approach limitations and assumptions

The computational tools used in this work have various limitations, many of which are discussed in more specific detail in the respective Methods sections above. Considering the challenges inherent to the selected approaches is essential for addressing biases, and appropriately contextualizing all results; some of these considerations are discussed below.

5.2.1.1 Domain-based phylogenies

Because protein domains represent independent evolutionary and structural units, trees based on subgene domains are a well-established method for tracing protein evolution (Yang and Bourne, 2009). However, domains can be relatively short sequences, which limits the amount of phylogenetic information that can be used to detect homology and map evolutionary relationships. Additionally, multifunctional, diverse domains, such as cytochromes, are widely conserved, frequently duplicated, and routinely gained or lost, which makes tracing histories or inferring functionality more difficult.

The domains analyzed in Chapter 2 provide several advantages that can help overcome the effects of uncertainty and limited information. The C1 domain is sufficiently divergent from related cytochromes—including additional cytochromes in the same ORF—to be isolated by simple homology searching. However, the domain is not too divergent to be detected by multiple domain analysis programs as a superfamily motif. Domain regions that do not satisfy these conditions may be less reliable and may not provide useful phylogenetic reconstructions. Despite the short length of the domain query, generating an initial tree that containing fewer taxa than the number of alignment sites helps ensure that site saturation does not significantly bias overall topology; these conservative initial trees can be used to evaluate confidence in the topology of larger, expanded phylogenies as well.

The genomic context of co-occurring domains can also constrain domain distribution and history. For example, domains may be excluded from analysis if they do not occur in an ORF that

contains other domains thought to be necessary for a higher-order metabolic process. However, this does not permit complete, representative analysis of the independent evolutionary histories of each domain. In Chapter 2, this conflict was addressed by generating two sets of domain trees for the C2 and NirS heme domains—one without genomic context, and one subsampled to reflect only taxa where given domains co-occur. These pairs of trees can be compared to determine whether topological differences reflect reliable evolutionary signal or sampling bias.

5.2.1.2 Approaches for tree construction

Both maximum likelihood and Bayesian methods are field standards for gene and species tree construction. These approaches share core methodological similarities; for example, both use odds ratio calculations to assess likelihood of sampled topologies (Felsenstein, 2004; Nascimento, Reis and Yang, 2017). However, while maximum likelihood phylogenetic methods identify a single most likely tree, Bayesian methods produce a most-probable tree that is a consensus of multiple sampled high-likelihood phylogenies (Alfaro and Holder, 2006; Bromham *et al.*, 2018). Bayesian approaches—unlike maximum likelihood approaches—require a prior distribution (Felsenstein, 2004). For phylogenetic trees, priors would include an initial divergence model, such as a uniform or birth-death process; clock priors are discussed in greater detail in Chapter 1.3.3.

Both methods provide robust, statistically supported inference of possible topologies. For trees with low support for the placement of key groups, or for genes with particularly complex evolutionary histories, it can be helpful to construct a tree using both maximum likelihood and (separately) Bayesian approaches; congruent results should increase confidence in topology or biological signal. In practice, this is not always tractable; Bayesian approaches that rely upon MCMC are computationally expensive (see Chapter 1.3.3.1). For trees that are quite large, convergence on an optimal topology can take a prohibitively long time; this becomes especially true for trees which contain several closely related, shallow-branching taxa, as shallow topological changes will have similar probabilities, making identification of the optimal tree increasingly difficult. Almost certainly because of these limitations, attempts to construct a Bayesian nitrilase gene tree (Chapter 3) were unsuccessful, with chains failing to converge after months of runtime, even after additional subsampling of taxa. Fortunately, maximum likelihood trees recovered robust support for key bipartitions for this project.

In cases where maximum likelihood trees do not provide strong topological support, and Bayesian trees fail to converge, iterative subsampling of taxa represented in alignments could be performed to obtain a small, simplified representation of groups for Bayesian analysis. This procedure sacrifices resolution and accuracy as biological signal is increasingly removed; however, it can still serve as a generalized constraint for deep bipartitions and relative clade placement.

5.2.1.3 ASR Inferences

Given that ASRs are calculated in a site-independent manner, it is inaccurate to assume that any reconstructed sequence—or indeed, any combination of ancestral sites—represents a true ancestor or ancestral combination. The ancestor will have been selected at the phenotypic level and may sample less-likely site combinations based upon environmental selection or drift. This limitation confers some advantages to the purely quantitative, *in silico* approach utilized in Chapter 4; site combinations can be considered quickly and in bulk by marginal combined likelihoods, but do not necessarily require that that the combinations represent biological truth (as compared with enzyme resurrections, where secondary structure or higher-order effects may impede protein function *in vitro* or *in vivo)*.

In this work, we attempt to avoid bias towards a nonexistent but high-likelihood ancestor sequence by exploring the sequence, physicochemical space, and marginal likelihood of all 64 site combinations with amino acid probabilities over 0.10. This approach allows us to consider sequence variability and likelihoods overall, including binned by a given likelihood threshold if desired, instead of isolating only the "top" likelihoods as ancestor candidates. But for all ASR work, it is important to temper inferences that rely upon multiple sites without resurrection and experimental validation. For modeling or resurrection studies, site pairings that are not represented anywhere in extant sequence space (as discussed in Chapter 4.4) can be evaluated independently from other variation from wild type sequence. In general, evaluating smaller scale sequence variations should help isolate combinations that drive change or loss of function.

Additionally, it is important to note that ASRs strongly reflect the biases of the alignments used for reconstruction (Vialle, Tamuri and Goldman, 2018). This underscores the importance of using broadly sampled sequences, accurate alignment algorithms, and manual curation of alignments to maximize the accuracy of downstream analyses (see Chapter 1.3.1).

Even with these considerations, sampling shortcomings that exclude homologs from an alignment may significantly impact the inferences of ASR, as well as other downstream phylogenetic analyses.

5.2.2 Sampling limitations and ghost lineages

The completeness and accuracy of phylogenies are inherently limited by the sampling of taxa included in alignments and trees. For environmental microbes and biogeochemically-relevant genes, such as those considered in this thesis, sampling issues begin *in situ*. Genomic diversity from understudied or difficult-to-access ecosystems, such as subsurface sediments, is comparatively underrepresented in genetic repositories. Additionally, culturable or model organisms often have well-sampled strain-level variation, which is relatively limited in environmental or candidate taxa. *In silico* subsampling, such as the genus or class-level subsampling used in this work, can help mitigate taxonomic inequities in environmental sampling. In the future, sourcing and analyzing metagenomes from a greater diversity of ecosystems and improving enrichment techniques for specialized environmental microbes may also slightly improve these limitations. But it is necessary to consider that apparently rare or missing phylogenetic information may reflect unsampled or excluded source taxa, not the absence of a metabolism *in situ*.

The issue of these "ghost lineages"—groups of organisms that are unknown or unsampled—becomes more complex for deep phylogenetic histories. Beyond extant organisms that have not been sampled, most ghost lineages are simply extinct (Tricou, Tannier and de Vienne, 2022); that is, their full crown group is missing in modern ecosystem. These "missing" groups render all phylogenies incomplete to some degree. Ghost lineages are particularly problematic in tracing eukaryotic hybridization and backcrossing events, but play a significant role prokaryotic gene trees as well, and can complicate the inference of HGT donor linages; it is impossible to exclude the possibility that a given gene was not transferred through a ghost lineage, especially if the donor is not topologically clear (Hibbins and Hahn, 2022; Tricou, Tannier and de Vienne, 2022).

For the HGTs considered in this work, inferring conclusive donor lineages was nonessential. The origin of the C1 domain transferred into Chloroflexi does not significantly affect analysis of the resulting fusion, which is present only within Anaerolineales. It is not clear which

bacterial lineage was the source of the nitrilase ancestor, but it is unambiguous from tree topology and age that this group was prokaryotic. Therefore, if the donor group for either transfer was a ghost lineage, it would not significantly change or affect the conclusions of this work. Still, when hypothesizing about the big picture of these ancient domains and genes, it is important to note the possibility that unsampled or extinct lineages served as the original source or donor for the transferred orthologs observed in extant taxa.

Additionally, for phylogenetic reconstructions in which identifying a donor lineage is essential for the hypothesis under evaluation, various studies have established approaches for detecting or considering the effect of ghost lineages (Cavin and Forey, 2007; Recuero *et al.*, 2014; Ward and Shih, 2021). In fact, inferring the presence of these lineages can be powerful for hypotheses of deep-time microbial biogeochemistry. For example, inferring a ghost lineage donor of the Type II reaction center and bacteriochlorophyll complexes can provide a parsimonious hypothesis for the origin and history of photosynthesis (Ward and Shih, 2021).

5.3 Future directions

All three projects described in this thesis underscore the utility of bioinformatic approaches in identifying, contextualizing, and constraining targeted biological hypotheses. Purely *in silico* approaches can sensitively detect genetic variation within metagenomes, rapidly map complex evolutionary relationships, and generate models of extinct ancient sequence space—all with robust, flexible quantitative methods and statistical support. These are methods only feasible using efficient computational frameworks, and they yield evolutionary and genomic results that can independently change our understanding of basic biological and evolutionary processes.

However, bioinformatic and phylogenetic approaches are also natural partners or precursors for targeted experimental work. At the most basic level, while the culturability problem persists, bioinformatic approaches will be required to identify organisms, genes, and metabolic potential in the environment. But more targeted searching within the constructs of specific metabolisms or organisms—for example, identification of novel domain architectures within denitrification genes in the Chloroflexi—should also be increasingly leveraged in the future. As we gather more genomic data, with increasing accuracy, from a greater number of different environments, targeted genetic searches *in silico* will be invaluable for processing and distilling meaning from bulk data.

But as this thesis displays, phylogenetic approaches should also be used increasingly to precede and guide experimentation. Trees, domain maps, and ASRs can directly generate biological hypotheses that can be tested at the bench; various aspects of each chapter in this thesis could be explored further with downstream experimentation. Divergent or modified enzymes such as the novel NirS in Chapter 2 can be constructed and evaluated *in vitro* for divergent function as compared with canonical enzyme variants. Enzymes that are relevant for biosynthesis or bioremediation, such as nitrilases, NHases, or SCNases, can be isolated or modified according to their phylogenetic diversity, and then evaluated *in vitro* or *in vivo* for desired biochemical efficacy. ASRs identify key sites to evaluate in enzyme resurrection, and resurrected enzymes—such as possible nitrogenase ancestors—can be evaluated *in vitro* for metabolic activity, including promiscuous substrate reduction.

Within the context of nitrogen biogeochemistry, the results of this thesis suggest several areas that are ripe for further exploration. Metagenomes from underexplored biomes including the deep subsurface should be examined carefully for divergence in core nutrient cycling genes. Microbial lifestyles in subsurface and sediment ecosystems are poorly mapped, but contain substantial biomass and contribute significantly to nutrient cycling (Osburn *et al.*, 2014; Momper, Jungbluth, *et al.*, 2017); understanding the diversity in geobiological capacity in these systems may reshape our understanding of even well-studied geochemical pathways on the modern Earth.

Noncanonical nitrogen metabolisms should also be carefully characterized and considered for their biogeochemical roles *in situ*, especially when considering models for historical nitrogen cycling. Enzymatic pathways that are excluded from the modern nitrogen cycle may have played a core or significant role in ancestral biogeochemistry. But it is also possible that metabolisms considered to be auxiliary, specialized, or limited play a larger role in modern ecosystems than currently understood, especially if enzymes can promiscuously metabolize off-target substrates.

Substrate promiscuity should also be considered carefully in the context of ancestral biogeochemistry. Trees of enzyme orthologs neither fully represent nor fully record ancestral enzyme diversity. But particularly ancient genes or enzyme families will have co-evolved with changing chemical environments, and may retain signatures of ancestral selection, possibly including substrate affinity. Ancestral enzymes are not always more promiscuous than modern proteins (Siddiq, Hochberg and Thornton, 2017), but enzymes that can metabolize multiple substrates with a similar biochemical "solution"—for example, reduction of a linear, nitrogenous

triple bond—should be investigated further to determine if this flexibility reflects a geochemically-relevant past.

REFERENCES

Abelson, P. H. (1966) 'Chemical Events on the Primitive Earth', *Proceedings of the National Academy of Sciences*, 55(6), pp. 1365–1372. doi: 10.1073/pnas.55.6.1365.

Agarwala, R. *et al.* (2018) 'Database resources of the National Center for Biotechnology Information', *Nucleic Acids Research*. doi: 10.1093/nar/gkx1095.

Alfaro, M. E. and Holder, M. T. (2006) 'The Posterior and the Prior in Bayesian Phylogenetics', *Annual Review of Ecology, Evolution, and Systematics*. Annual Reviews, 37, pp. 19–42. Available at: http://www.jstor.org/stable/30033825.

Alneberg, J. *et al.* (2014) 'Binning metagenomic contigs by coverage and composition', *Nature Methods*. doi: 10.1038/nmeth.3103.

Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: https://doi.org/10.1016/S0022-2836(05)80360-2.

Alvillo-Rivera, A. *et al.* (2021) 'Biological treatment for the degradation of cyanide: A review', *Journal of Materials Research and Technology*, 12, pp. 1418–1433. doi: 10.1016/j.jmrt.2021.03.030.

Anisimova, M. and Gascuel, O. (2006) 'Approximate Likelihood-Ratio Test for Branches: A Fast, Accurate, and Powerful Alternative', *Systematic Biology*, 55(4), pp. 539–552. doi: 10.1080/10635150600755453.

Barnabas, J., Schwartz, R. M. and Dayhoff, M. O. (1982) 'Evolution of major metabolic innovations in the Precambrian', *Origins of life*, 12(1), pp. 81–91. doi: 10.1007/BF00926914.

Bell, E. *et al.* (2020) 'Active sulfur cycling in the terrestrial deep subsurface', *The ISME Journal*, 14(5), pp. 1260–1272. doi: 10.1038/s41396-020-0602-x.

Benedik, M. J. and Sewell, B. T. (2017) 'Cyanide-degrading nitrilases in nature', *The Journal of General and Applied Microbiology*, 93, pp. 90–93. doi: 10.2323/jgam.2017.06.002.

Benson, S. W. (1965) 'III-Bond energies', *Journal of Chemical Education*, 42(9), p. 502.

Berman, H. M. *et al.* (2000) 'The Protein Data Bank', *Nucleic Acids Research*. doi: 10.1093/nar/28.1.235.

Bertini, I., Cavallaro, G. and Rosato, A. (2006) 'Cytochrome c: Occurrence and functions', *Chemical Reviews*, pp. 90–115. doi: 10.1021/cr050241v.

Bienert, S. *et al.* (2017) 'The SWISS-MODEL Repository-new features and functionality', *Nucleic Acids Research*. doi: 10.1093/nar/gkw1132.

Blomberg, M. R. A. and Ädelroth, P. (2018) 'Mechanisms for enzymatic reduction of nitric oxide to nitrous oxide - A comparison between nitric oxide reductase and cytochrome c oxidase', *Biochimica et Biophysica Acta - Bioenergetics*. Elsevier, 1859(11), pp. 1223–1234. doi: 10.1016/j.bbabio.2018.09.368.

Bonin, P. (1996) 'Anaerobic nitrate reduction to ammonium in two strains isolated from coastal marine sediment: A dissimilatory pathway', *FEMS Microbiology Ecology*, 19(1), pp. 27–38. doi: 10.1016/0168-6496(95)00075-5.

Boyd, E. S. *et al.* (2011) 'A late methanogen origin for molybdenum-dependent nitrogenase', *Geobiology*, 9(3), pp. 221–232. doi: 10.1111/j.1472-4669.2011.00278.x.

Braker, G. *et al.* (2000) 'Nitrite reductase genes (nirK and nirS) as functional markers to investigate diversity of denitrifying bacteria in pacific northwest marine sediment communities', *Applied and Environmental Microbiology*, 66(5), pp. 2096–2104. doi: 10.1128/AEM.66.5.2096-2104.2000.

Broda, E. (1977) 'Two kinds of lithotrophs missing in nature', *Zeitschrift für allgemeine*

Mikrobiologie. John Wiley & Sons, Ltd, 17(6), pp. 491–493. doi: https://doi.org/10.1002/jobm.19770170611.
Bromham, L. *et al.* (2018) 'Bayesian molecular dating: opening up the black box', *Biological Reviews*. John Wiley & Sons, Ltd, 93(2), pp. 1165–1191. doi: https://doi.org/10.1111/brv.12390.
Butterbach-Bahl, K. and Dannenmann, M. (2011) 'Denitrification and associated soil N2O emissions due to agricultural activities in a changing climate', *Current Opinion in Environmental Sustainability*. Elsevier B.V., 3(5), pp. 389–395. doi: 10.1016/j.cosust.2011.08.004.
Camacho, C. *et al.* (2009) 'BLAST+: Architecture and applications', *BMC Bioinformatics*. doi: 10.1186/1471-2105-10-421.
Campbell, J. H. *et al.* (2013) 'UGA is an additional glycine codon in uncultured SR1 bacteria from the human microbiota', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1303090110.
Canfield, D. E., Glazer, A. N. and Falkowski, P. G. (2010) 'The evolution and future of earth's nitrogen cycle', *Science*, 330(6001), pp. 192–196. doi: 10.1126/science.1186120.
Carlson, H. K. *et al.* (2020) 'Selective carbon sources influence the end products of microbial nitrate respiration', *ISME Journal*. Springer US, 14(8), pp. 2034–2045. doi: 10.1038/s41396-020-0666-7.
Cavin, L. and Forey, P. L. (2007) 'Using ghost lineages to identify diversification events in the fossil record', *Biology Letters*. Royal Society, 3(2), pp. 201–204. doi: 10.1098/rsbl.2006.0602.
Cheng, Z., Xia, Y. and Zhou, Z. (2020) 'Recent Advances and Promises in Nitrile Hydratase: From Mechanism to Industrial Applications ', *Frontiers in Bioengineering and Biotechnology* . Available at: https://www.frontiersin.org/article/10.3389/fbioe.2020.00352.
Chhiba-Govindjee, V. P. *et al.* (2019) 'Bacterial nitrilases and their regulation', *Applied Microbiology and Biotechnology*, 103(12), pp. 4679–4692. doi: 10.1007/s00253-019-09776-1.
Conradson, S. D. *et al.* (1989) 'Cyanide and methylisocyanide binding to the isolated iron-molybdenum cofactor of nitrogenase.', *The Journal of biological chemistry*. United States, 264(27), pp. 15967–15974.
Creevey, C. J. *et al.* (2011) 'Universally distributed single-copy genes indicate a constant rate of horizontal transfer', *PLoS ONE*, 6(8). doi: 10.1371/journal.pone.0022099.
Darriba, D. *et al.* (2011) 'ProtTest 3: fast selection of best-fit models of protein evolution', *Bioinformatics*, 27(8), pp. 1164–1165. doi: 10.1093/bioinformatics/btr088.
Dayhoff, M., Schwartz, R. and Orcutt, B. (1978) '22 a model of evolutionary change in proteins', *Atlas of protein sequence and structure*. National Biomedical Research Foundation Silver Spring MD, 5, pp. 345–352.
Dean, J. A. (1999) *Lange's Handbook of Chemistry*. 15th edn.
Decleyre, H. *et al.* (2016) 'Erratum to: Highly diverse nirK genes comprise two major clades that harbour ammonium-producing denitrifiers', *BMC Genomics*. BMC Genomics, 17(1), p. 12864. doi: 10.1186/s12864-016-2812-1.
Denef, V. J. *et al.* (2016) 'Chloroflexi CL500-11 populations that predominate deep-lake hypolimnion bacterioplankton rely on nitrogen-rich dissolved organic matter metabolism and C1 compound oxidation', *Applied and Environmental Microbiology*, 82(5), pp. 1423–1432. doi: 10.1128/AEM.03014-15.
DeSantis, G. *et al.* (2002) 'An enzyme library approach to biocatalysis: Development of nitrilases for enantioselective production of carboxylic acid derivatives', *Journal of the American Chemical Society*, 124(31), pp. 9024–9025. doi: 10.1021/ja0259842.
Doane, T. A. (2017) 'The Abiotic Nitrogen Cycle', *ACS Earth and Space Chemistry*. American

Chemical Society, 1(7), pp. 411–421. doi: 10.1021/acsearthspacechem.7b00059.
Draganić, Z. D. *et al.* (1980) 'The radiolysis of aqueous ammonium cyanide: compounds of interest to chemical evolution studies.', *Journal of molecular evolution*. Germany, 15(3), pp. 239–260. doi: 10.1007/BF01732951.
Drummond, A. J. *et al.* (2006) 'Relaxed Phylogenetics and Dating with Confidence', *PLOS Biology*. Public Library of Science, 4(5), p. e88. Available at: https://doi.org/10.1371/journal.pbio.0040088.
Dupont, C. L. *et al.* (2012) 'Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage', *ISME Journal*. Nature Publishing Group, 6(6), pp. 1186–1199. doi: 10.1038/ismej.2011.189.
Ebbs, S. (2004) 'Biological degradation of cyanide compounds', *Current Opinion in Biotechnology*, 15(3), pp. 231–236. doi: 10.1016/j.copbio.2004.03.006.
Eddy, S. R. (2011) 'Accelerated Profile HMM Searches.', *PLoS computational biology*, 7(10), p. e1002195. doi: 10.1371/journal.pcbi.1002195.
Edgar, R. C. (2004) 'MUSCLE: multiple sequence alignment with high accuracy and high throughput', *Nucleic acids research*. Oxford University Press, 32(5), pp. 1792–1797. doi: 10.1093/nar/gkh340.
Edgar, R. C. and Batzoglou, S. (2006) 'Multiple sequence alignment', *Current Opinion in Structural Biology*, 16(3), pp. 368–373. doi: https://doi.org/10.1016/j.sbi.2006.04.004.
Fani, R., Gallo, R. and Liò, P. (2000) 'Molecular evolution of nitrogen fixation: The evolutionary history of the nifD, nifK, nifE, and nifN genes', *Journal of Molecular Evolution*, 51(1), pp. 1–11. doi: 10.1007/s002390010061.
Felsenstein, J. (2004) *Inferring Phylogenies*. New York: Oxford University Press.
Ferris, J. P. *et al.* (1978) 'HCN: A plausible source of purines, pyrimidines and amino acids on the primitive earth', *Journal of Molecular Evolution*, 11(4), pp. 293–311. doi: 10.1007/BF01733839.
Ferris, J. P. and Orgel, L. E. (1966) 'An Unusual Photochemical Rearrangement in the Synthesis of Adenine from Hydrogen Cyanide1', *Journal of the American Chemical Society*, 88(5), p. 1074. doi: 10.1021/ja00957a050.
Ferus, M. *et al.* (2017) 'High Energy Radical Chemistry Formation of HCN-rich Atmospheres on early Earth', *Scientific Reports*, 7(1), p. 6275. doi: 10.1038/s41598-017-06489-1.
Fournier, G. P., Andam, C. P. and Gogarten, J. P. (2015) 'Ancient horizontal gene transfer and the last common ancestors', *BMC Evolutionary Biology*, 15(1), p. 70. doi: 10.1186/s12862-015-0350-0.
Di Franco, A. *et al.* (2019) 'Evaluating the usefulness of alignment filtering methods to reduce the impact of errors on evolutionary inferences', *BMC Evolutionary Biology*, 19(1), p. 21. doi: 10.1186/s12862-019-1350-2.
Galloway, J. N. *et al.* (2004) *Nitrogen cycles: Past, present, and future*, *Biogeochemistry*. doi: 10.1007/s10533-004-0370-0.
Garcia, A. K. *et al.* (2020) 'Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization', *Geobiology*. John Wiley & Sons, Ltd, 18(3), pp. 394–411. doi: https://doi.org/10.1111/gbi.12381.
Garcia, A. K. and Kaçar, B. (2019) 'How to resurrect ancestral proteins as proxies for ancient biogeochemistry', *Free Radical Biology and Medicine*, 140, pp. 260–269. doi: https://doi.org/10.1016/j.freeradbiomed.2019.03.033.
Garcia, A. K., Kolaczkowski, B. and Kaçar, B. (2022) 'Reconstruction of Nitrogenase

Predecessors Suggests Origin from Maturase-Like Proteins.', *Genome biology and evolution*, 14(3). doi: 10.1093/gbe/evac031.
Gardner, A. M., Helmick, R. A. and Gardner, P. R. (2002) 'Flavorubredoxin, an inducible catalyst for nitric oxide reduction and detoxification in Escherichia coli', *Journal of Biological Chemistry*, 277(10), pp. 8172–8177. doi: 10.1074/jbc.M110471200.
Garvin, J. *et al.* (2009) 'Isotopic evidence for an aerobic nitrogen cycle in the latest Archean', *Science*, 323(5917), pp. 1045–1048. doi: 10.1126/science.1165675.
Genming, L. *et al.* (2022) 'Rapid oxygenation of Earth's atmosphere 2.33 billion years ago', *Science Advances*. American Association for the Advancement of Science, 2(5), p. e1600134. doi: 10.1126/sciadv.1600134.
Giblin, A. E. *et al.* (2013) 'The importance of dissimilatory nitrate reduction to ammonium (DNRA) in the nitrogen cycle of coastal ecosystems', *Oceanography*, 26(3), pp. 124–131. doi: 10.5670/oceanog.2013.54.
Gogarten, J. P. and Doolittle, W. F. (2007) 'Prokaryotic Evolution in Light of Gene Transfer', *Molecular Biology*, (C), pp. 2005–2007.
Gold, D. A. *et al.* (2017) 'Paleoproterozoic sterol biosynthesis and the rise of oxygen', *Nature*, 543(7645), pp. 420–423. doi: 10.1038/nature21412.
Gong, J.-S. *et al.* (2012) 'Nitrilases in nitrile biocatalysis: recent progress and forthcoming research', *Microbial Cell Factories*, 11(1), p. 142. doi: 10.1186/1475-2859-11-142.
Graf, D. R. H., Jones, C. M. and Hallin, S. (2014) 'Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community structure for N2O emissions', *PLoS ONE*, 9(12), pp. 1–20. doi: 10.1371/journal.pone.0114118.
Gruber, N. and Galloway, J. N. (2008) 'An Earth-system perspective of the global nitrogen cycle', *Nature*, 451(7176), pp. 293–296. doi: 10.1038/nature06592.
Guindon, S. *et al.* (2010) 'New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0', *Systematic Biology*, 59(3), pp. 307–321. doi: 10.1093/sysbio/syq010.
Guindon, S. and Gascuel, O. (2003) 'A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood', *Systematic Biology*, 52(5), pp. 696–704. doi: 10.1080/10635150390235520.
Harrison, K. J., Crécy-Lagard, V. De and Zallot, R. (2018) 'Gene Graphics: A genomic neighborhood data visualization web application', *Bioinformatics*, 34(8), pp. 1406–1408. doi: 10.1093/bioinformatics/btx793.
Heard, A. W. *et al.* (2017) 'Origins and ages of fracture fluids in the South African Crust', in *Agufm*, pp. H11A-1153.
Helaers, R. and Milinkovitch, M. C. (2010) 'MetaPIGA v2.0: maximum likelihood large phylogeny estimation using the metapopulation genetic algorithm and other stochastic heuristics', *BMC Bioinformatics*, 11(1), p. 379. doi: 10.1186/1471-2105-11-379.
Hemp, J. *et al.* (2015) 'Draft genome sequence of Ardenticatena maritima 110S, a thermophilic nitrate- and iron-reducing member of the Chloroflexi Class Ardenticatenia', *Genome Announcements*, 3(6), p. 2920. doi: 10.1128/genomeA.01347-15.
Hemp, J. and Gennis, R. B. (2008) 'Diversity of the Heme-Copper superfamily in archaea: Insights from genomics and structural modeling', *Results and Problems in Cell Differentiation*, 45(January), pp. 1–31. doi: 10.1007/400_2007_046.
Hendriks, J. *et al.* (2000) 'Nitric oxide reductases in bacteria', *Biochimica et Biophysica Acta - Bioenergetics*, 1459(2–3), pp. 266–273. doi: 10.1016/S0005-2728(00)00161-4.

Heylen, K. *et al.* (2007) 'Nitric oxide reductase (norB) gene sequence analysis reveals discrepancies with nitrite reductase (nir) gene phylogeny in cultivated denitrifiers', *Environmental Microbiology*, 9(4), pp. 1072–1077. doi: 10.1111/j.1462-2920.2006.01194.x.
Hibbins, M. S. and Hahn, M. W. (2022) 'Phylogenomic approaches to detecting and characterizing introgression', *Genetics*, 220(2), p. iyab173. doi: 10.1093/genetics/iyab173.
Hoang, D. T. *et al.* (2018) 'UFBoot2: Improving the ultrafast bootstrap approximation', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msx281.
Hug, L. A. *et al.* (2013) 'Community genomic analyses constrain the distribution of metabolic traits across the Chloroflexi phylum and indicate roles in sediment carbon cycling', *Microbiome*, 1(1), pp. 1–17. doi: 10.1186/2049-2618-1-22.
Hug, L. A. *et al.* (2016) 'Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages', *Environmental Microbiology*, 18(1), pp. 159–173. doi: 10.1111/1462-2920.12930.
Huntemann, M. *et al.* (2015) 'The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4)', *Standards in Genomic Sciences*. doi: 10.1186/s40793-015-0077-y.
Jones, C. M. *et al.* (2008) 'Phylogenetic analysis of nitrite, nitric oxide, and nitrous oxide respiratory enzymes reveal a complex evolutionary history for denitrification', *Molecular Biology and Evolution*, 25(9), pp. 1955–1966. doi: 10.1093/molbev/msn146.
Jungbluth, S. P., Amend, J. P. and Rappé, M. S. (2017) 'Corrigendum: Metagenome sequencing and 98 microbial genomes from Juan de Fuca Ridge flank subsurface fluids', *Scientific data*, 4, p. 170080. doi: 10.1038/sdata.2017.80.
Kadnikov, V. V. *et al.* (2020) 'Microbial Life in the Deep Subsurface Aquifer Illuminated by Metagenomics', *Frontiers in Microbiology*, 11(September), pp. 1–15. doi: 10.3389/fmicb.2020.572252.
Kalyaanamoorthy, S. *et al.* (2017) 'ModelFinder: Fast model selection for accurate phylogenetic estimates', *Nature Methods*. doi: 10.1038/nmeth.4285.
Kantor, R. S. *et al.* (2015) 'Bioreactor microbial ecosystems for thiocyanate and cyanide degradation unravelled with genome-resolved metagenomics', *Environmental Microbiology*, 17(12), pp. 4929–4941. doi: 10.1111/1462-2920.12936.
Kapoor, V. *et al.* (2016) 'Inhibitory effect of cyanide on wastewater nitrification determined using SOUR and RNA-based gene-specific assays', *Letters in Applied Microbiology*, 63(2), pp. 155–161. doi: 10.1111/lam.12603.
Kartal, B. and Keltjens, J. T. (2016) 'Anammox Biochemistry: a Tale of Heme c Proteins', *Trends in Biochemical Sciences*, 41(12), pp. 998–1011. doi: https://doi.org/10.1016/j.tibs.2016.08.015.
Katoh, K. *et al.* (2002) 'MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform', *Nucleic Acids Research*, 30(14), pp. 3059–3066. doi: 10.1093/nar/gkf436.
Katoh, K. and Toh, H. (2008) 'Recent developments in the MAFFT multiple sequence alignment program', *Briefings in Bioinformatics*, 9(4), pp. 286–298. doi: 10.1093/bib/bbn013.
Kawasaki, S. *et al.* (1997) 'Gene cluster for dissimilatory nitrite reductase (nir) from Pseudomonas aeruginosa: Sequencing and identification of a locus for heme d1 biosynthesis', *Journal of Bacteriology*, 179(1), pp. 235–242. doi: 10.1128/jb.179.1.235-242.1997.
Kishino, H., Thorne, J. L. and Bruno, W. J. (2001) 'Performance of a Divergence Time Estimation Method under a Probabilistic Model of Rate Evolution', *Molecular Biology and*

Evolution, 18(3), pp. 352–361. doi: 10.1093/oxfordjournals.molbev.a003811.
Knoll, A. H., Bergmann, K. D. and Strauss, J. V. (2016) 'Life: The first two billion years', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 371(1707). doi: 10.1098/rstb.2015.0493.
Kobayashi, M. *et al.* (1991) 'Cloning, nucleotide sequence and expression in Escherichia coli of two cobalt-containing nitrile hydratase genes from Rhodococcus rhodochrous J1.', *Biochimica et biophysica acta*. Netherlands, 1129(1), pp. 23–33. doi: 10.1016/0167-4781(91)90208-4.
Kobayashi, M. and Shimizu, S. (1998) 'Metalloenzyme nitrile hydratase: Structure, regulation, and application to biotechnology', (lll), pp. 733–736.
Kobayashi, M. and Shimizu, S. (2000) 'Nitrile hydrolases', *Current Opinion in Chemical Biology*, 4, pp. 95–102.
Kuenen J. G. (2008) 'Anammox bacteria: from discovery to application', *Nature Reviews Microbiology*, 6(april), pp. 320–326.
Kuhn, M. L. *et al.* (2012) 'The Fe-type nitrile hydratase from Comamonas testosteroni Ni1 does not require an activator accessory protein for expression in Escherichia coli', *Biochemical and Biophysical Research Communications*, 424(3), pp. 365–370. doi: https://doi.org/10.1016/j.bbrc.2012.06.036.
Kuypers, M. M. M., Marchant, H. K. and Kartal, B. (2018) 'The microbial nitrogen-cycling network', *Nature Reviews Microbiology*. Nature Publishing Group, 16(5), pp. 263–276. doi: 10.1038/nrmicro.2018.9.
LaRowe, D. E. and Regnier, P. (2008) 'Thermodynamic Potential for the Abiotic Synthesis of Adenine, Cytosine, Guanine, Thymine, Uracil, Ribose, and Deoxyribose in Hydrothermal Systems', *Origins of Life and Evolution of Biospheres*, 38(5), p. 383. doi: 10.1007/s11084-008-9137-2.
Lau, M. C. Y. *et al.* (2014) 'of functional genes shared among seven terrestrial subsurface metagenomes reveal N-cycling and microbial evolutionary relationships', *Frontiers in Microbiology*, 5(OCT), pp. 1–17. doi: 10.3389/fmicb.2014.00531.
Lau, M. C. Y. *et al.* (2016) 'An oligotrophic deep-subsurface community dependent on syntrophy is dominated by sulfur-driven autotrophic denitrifiers', *Proceedings of the National Academy of Sciences*, 113(49), pp. E7927–E7936. doi: 10.1073/pnas.1612244113.
Le, S. Q. and Gascuel, O. (2008) 'An Improved General Amino Acid Replacement Matrix', *Molecular Biology and Evolution*, 25(7), pp. 1307–1320. doi: 10.1093/molbev/msn067.
Lepage, T. *et al.* (2006) 'Continuous and tractable models for the variation of evolutionary rates', *Mathematical Biosciences*, 199(2), pp. 216–233. doi: https://doi.org/10.1016/j.mbs.2005.11.002.
Lepage, T. *et al.* (2007) 'A General Comparison of Relaxed Molecular Clock Models', *Molecular Biology and Evolution*, 24(12), pp. 2669–2680. doi: 10.1093/molbev/msm193.
Li, J., Burgess, B. K. and Corbin, J. L. (1982) 'Nitrogenase reactivity: cyanide as substrate and inhibitor', *Biochemistry*. American Chemical Society, 21(18), pp. 4393–4402. doi: 10.1021/bi00261a031.
Liberles, D. A. *et al.* (2020) 'Emerging Frontiers in the Study of Molecular Evolution.', *Journal of molecular evolution*, pp. 211–226. doi: 10.1007/s00239-020-09932-6.
Lowe, D. J. *et al.* (1989) 'Kinetics and mechanism of the reaction of cyanide with molybdenum nitrogenase from Azotobacter vinelandii.', *Biochemistry*. United States, 28(21), pp. 8460–8466. doi: 10.1021/bi00447a028.
Lu, H., Chandran, K. and Stensel, D. (2014) 'Microbial ecology of denitrification in biological wastewater treatment', *Water Research*. Elsevier Ltd, 64, pp. 237–254. doi:

10.1016/j.watres.2014.06.042.
Lu, S. *et al.* (2020) 'CDD/SPARCLE: The conserved domain database in 2020', *Nucleic Acids Research*. doi: 10.1093/nar/gkz991.
Luo, J. *et al.* (2017) 'Chapter Five - Potential Hotspot Areas of Nitrous Oxide Emissions From Grazed Pastoral Dairy Farm Systems', in Sparks, D. L. B. T.-A. in A. (ed.). Academic Press, pp. 205–268. doi: https://doi.org/10.1016/bs.agron.2017.05.006.
Magnabosco, C. *et al.* (2016) 'A metagenomic window into carbon metabolism at 3 km depth in Precambrian continental crust', *ISME Journal*. Nature Publishing Group, 10(3), pp. 730–741. doi: 10.1038/ismej.2015.150.
Magnabosco, C. *et al.* (2018) 'The biomass and biodiversity of the continental subsurface', *Nature Geoscience*, 11(10), pp. 707–717. doi: 10.1038/s41561-018-0221-6.
Maia, L. B. and Moura, J. J. G. (2014) 'How biology handles nitrite', *Chemical Reviews*, 114(10), pp. 5273–5357. doi: 10.1021/cr400518y.
Marchler-Bauer, A. *et al.* (2015) 'CDD: NCBI's conserved domain database', *Nucleic Acids Research*. doi: 10.1093/nar/gku1221.
Markowitz, V. M. *et al.* (2008) 'IMG/M: A data management and analysis system for metagenomes', *Nucleic Acids Research*. doi: 10.1093/nar/gkm869.
Martín-González, A. *et al.* (2009) 'Microbial Cretaceous park: biodiversity of microbial fossils entrapped in amber', *Naturwissenschaften*, 96(5), pp. 551–564. doi: 10.1007/s00114-009-0508-y.
Martínková, L. *et al.* (2015) 'Cyanide hydratases and cyanide dihydratases: emerging tools in the biodegradation and biodetection of cyanide', *Applied Microbiology and Biotechnology*, 99(21), pp. 8875–8882. doi: 10.1007/s00253-015-6899-0.
Matsuura, N. *et al.* (2015) 'Draft Genome Sequences of Anaerolinea thermolimosa IMO-1, Bellilinea caldifistulae GOMI-1, Leptolinea tardivitalis YMTK-2, Levilinea saccharolytica KIBI-1, Longilinea arvoryzae KOME-1, Previously Described as Members of the Class Anaerolineae (Chloroflex', *Genome announcements*, 3(5). doi: 10.1128/genomeA.00975-15.
Merkl, R. and Sterner, R. (2016) 'Ancestral protein reconstruction: techniques and applications', *Biological Chemistry*, 397(1), pp. 1–21. doi: doi:10.1515/hsz-2015-0158.
Minh, B. Q. *et al.* (2022) *IQ-TREE version 2.2.0: Tutorials and manual phylogenomic software by maximum likelihood*. Available at: http:www.iqtree.org.
Minh, B. Q., Nguyen, M. A. T. and Von Haeseler, A. (2013) 'Ultrafast approximation for phylogenetic bootstrap', *Molecular Biology and Evolution*. doi: 10.1093/molbev/mst024.
Mitchell, A. L. *et al.* (2019) 'InterPro in 2019: Improving coverage, classification and access to protein sequence annotations', *Nucleic Acids Research*. doi: 10.1093/nar/gky1100.
Miyakawa, S., Cleaves, H. and Miller, S. L. (2002) 'The Cold Origin of Life : Hydrogen Cyanide and Formamide', 32, pp. 195–208.
Momper, L., Jungbluth, S. P., *et al.* (2017) 'Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community', *ISME Journal*. Nature Publishing Group, 11(10), pp. 2319–2333. doi: 10.1038/ismej.2017.94.
Momper, L., Kiel Reese, B., *et al.* (2017) 'Major phylum-level differences between porefluid and host rock bacterial communities in the terrestrial deep subsurface', *Environmental Microbiology Reports*, 9(5), pp. 501–511. doi: 10.1111/1758-2229.12563.
Morris, J. L. *et al.* (2018) 'The timescale of early land plant evolution', *Proceedings of the National Academy of Sciences of the United States of America*, 115(10), pp. E2274–E2283. doi: 10.1073/pnas.1719588115.

Mulder, A. *et al.* (1995) 'Anaerobic ammonium oxidation discovered in a denitrifying fluidized bed reactor', *FEMS Microbiology Ecology*, 16(3), pp. 177–183. doi: 10.1111/j.1574-6941.1995.tb00281.x.
Murali, R. *et al.* (2019) 'FIND: Identifying Functionally and Structurally Important Features in Protein Sequences with Deep Neural Networks', *bioRxiv*. doi: 10.1101/592808.
Nakamura, T. *et al.* (2018) 'Parallelization of MAFFT for large-scale multiple sequence alignments', *Bioinformatics*. doi: 10.1093/bioinformatics/bty121.
Nascimento, F. F., Reis, M. dos and Yang, Z. (2017) 'A biologist's guide to Bayesian phylogenetic analysis', *Nature Ecology & Evolution*, 1(10), pp. 1446–1454. doi: 10.1038/s41559-017-0280-x.
Needleman, S. B. and Wunsch, C. D. (1970) 'A general method applicable to the search for similarities in the amino acid sequence of two proteins', *Journal of Molecular Biology*, 48(3), pp. 443–453. doi: https://doi.org/10.1016/0022-2836(70)90057-4.
Nguyen, L. T. *et al.* (2015) 'IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies', *Molecular Biology and Evolution*. doi: 10.1093/molbev/msu300.
Nishiyama, M. *et al.* (1991) 'Cloning and characterization of genes responsible for metabolism of nitrile compounds from Pseudomonas chloraraphis B23', *Journal of Bacteriology*, 173(8), pp. 2465–2472. doi: 10.1128/jb.173.8.2465-2472.1991.
North, J. A. *et al.* (2020) 'A nitrogenase-like enzyme system catalyzes methionine, ethylene, and methane biogenesis', *Science*. American Association for the Advancement of Science, 369(6507), pp. 1094–1098. doi: 10.1126/science.abb6310.
Nuin, P. A. S., Wang, Z. and Tillier, E. R. M. (2006) 'The accuracy of several multiple sequence alignment programs for proteins', *BMC Bioinformatics*, 7(1), p. 471. doi: 10.1186/1471-2105-7-471.
Nutman, A. P. *et al.* (2016) 'Rapid emergence of life shown by discovery of 3,700-million-year-old microbial structures', *Nature*, 537(7621), pp. 535–538. doi: 10.1038/nature19355.
O'Leary, N. A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733–D745. doi: 10.1093/nar/gkv1189.
O'Reilly, C. and Turner, P. D. (2003) 'The nitrilase family of CN hydrolysing enzymes - A comparative study', *Journal of Applied Microbiology*, 95(6), pp. 1161–1174. doi: 10.1046/j.1365-2672.2003.02123.x.
Oró, J. (1961) 'Mechanism of synthesis of adenine from HCN under possible primitive earth conditions.', *Nature*, 191(4794), pp. 1193–1194. doi: 10.1038/1911193a0.
Oró, J. and Kamat, S. S. (1961) 'Amino-acid Synthesis from Hydrogen Cyanide under Possible Primitive Earth Conditions', *Nature*, 190, pp. 442–443.
Osburn, M. R. *et al.* (2014) 'Chemolithotrophy in the continental deep subsurface: Sanford underground research facility (SURF), USA', *Frontiers in Microbiology*, 5(NOV), pp. 1–14. doi: 10.3389/fmicb.2014.00610.
Osburn, M. R. *et al.* (2019) 'Establishment of the Deep Mine Microbial Observatory (DeMMO), South Dakota, USA, a Geochemically Stable Portal Into the Deep Subsurface', *Frontiers in Earth Science*, 7(July), pp. 1–17. doi: 10.3389/feart.2019.00196.
Pace, H. C. and Brenner, C. (2001) 'The nitrilase superfamily : Classification, structure and function', (February 2001), pp. 1–9. doi: 10.1186/gb-2001-2-1-reviews0001.
Park, J. M., Trevor Sewell, B. and Benedik, M. J. (2017) 'Cyanide bioremediation: the potential

of engineered nitrilases', *Applied Microbiology and Biotechnology*. Applied Microbiology and Biotechnology, 101(8), pp. 3029–3042. doi: 10.1007/s00253-017-8204-x.
Parks, D. H. *et al.* (2017) 'Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life', *Nature Microbiology*. Springer US, 2(11), pp. 1533–1542. doi: 10.1038/s41564-017-0012-7.
Parsons, C. *et al.* (2021) 'Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history', *Geobiology*. John Wiley & Sons, Ltd, 19(1), pp. 18–34. doi: https://doi.org/10.1111/gbi.12419.
Pearce, B. K. D. and Pudritz, R. E. (2015) 'Seeding the Pregenetic Earth: Meteoritic Abundances of Nucleobases and Potential Reaction Pathways', *Astrophysical Journal*. IOP Publishing, 807(1), p. 85. doi: 10.1088/0004-637X/807/1/85.
Pearson, W. R. (2013) 'Selecting the Right Similarity-Scoring Matrix', *Current Protocols in Bioinformatics*. John Wiley & Sons, Ltd, 43(1), pp. 3.5.1-3.5.9. doi: https://doi.org/10.1002/0471250953.bi0305s43.
Pei, X. *et al.* (2013) 'Efficient cloning and expression of a thermostable nitrile hydratase in Escherichia coli using an auto-induction fed-batch strategy', *Process Biochemistry*, 48(12), pp. 1921–1927. doi: https://doi.org/10.1016/j.procbio.2013.09.004.
Penny, D. *et al.* (2014) 'The Relative Ages of Eukaryotes and Akaryotes', *Journal of Molecular Evolution*, 79(5), pp. 228–239. doi: 10.1007/s00239-014-9643-y.
Philippe, H. *et al.* (2011) 'Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough', *PLOS Biology*. Public Library of Science, 9(3), p. e1000602. Available at: https://doi.org/10.1371/journal.pbio.1000602.
Philippot, L. (2002) 'Philippot 2002 Denitrifying genes in bacterial and Archaeal genomes', 1577, pp. 355–376. doi: 10.1016/S0167-4781(02)00420-7.
Pickett, C. J. *et al.* (2004) 'Synergic binding of carbon monoxide and cyanide to the FeMo cofactor of nitrogenase: Relic chemistry of an ancient enzyme?', *Chemistry - A European Journal*, 10(19), pp. 4770–4776. doi: 10.1002/chem.200400382.
Podar, M., Eads, J. R. and Richardson, T. H. (2005) 'Evolution of a microbial nitrilase gene family: A comparative and environmental genomics study', *BMC Evolutionary Biology*, 5, pp. 1–13. doi: 10.1186/1471-2148-5-42.
Potter, S. C. *et al.* (2018) 'HMMER web server: 2018 update', *Nucleic Acids Research*, 46(W1), pp. W200–W204. doi: 10.1093/nar/gky448.
Priemé, A., Braker, G. and Tiedje, J. M. (2002) 'Diversity of nitrite reductase (nirK and nirS) gene fragments in forested upland and wetland soils', *Applied and Environmental Microbiology*, 68(4), pp. 1893–1900. doi: 10.1128/AEM.68.4.1893-1900.2002.
Prieto, M. and Wedin, M. (2013) 'Dating the Diversification of the Major Lineages of Ascomycota (Fungi)', *PLOS ONE*. Public Library of Science, 8(6), p. e65576. Available at: https://doi.org/10.1371/journal.pone.0065576.
Ravenhall, M. *et al.* (2015) 'Inferring Horizontal Gene Transfer', *PLOS Computational Biology*. Public Library of Science, 11(5), p. e1004095. Available at: https://doi.org/10.1371/journal.pcbi.1004095.
Raybuck, S. A. (1992) 'Microbes and microbial enzymes for cyanide degradation', *Biodegradation*, 3(1), pp. 3–18. doi: 10.1007/BF00189632.
Raymond, J. *et al.* (2004) 'The Natural History of Nitrogen Fixation', *Molecular Biology and Evolution*, 21(3), pp. 541–554. doi: 10.1093/molbev/msh047.
Raymond, J. (2005) 'The Evolution of Biological Carbon and Nitrogen Cycling—a Genomic

Perspective', *Reviews in Mineralogy and Geochemistry*, 59(1), pp. 211–231. doi: 10.2138/rmg.2005.59.9.
Recuero, E. *et al.* (2014) 'Evolutionary history of Ichthyosaura alpestris (Caudata, Salamandridae) inferred from the combined analysis of nuclear and mitochondrial markers', *Molecular Phylogenetics and Evolution*, 81, pp. 207–220. doi: https://doi.org/10.1016/j.ympev.2014.09.014.
Dos Reis, M. *et al.* (2015) 'Uncertainty in the Timing of Origin of Animals and the Limits of Precision in Molecular Timescales', *Current Biology*. The Authors, 25(22), pp. 2939–2950. doi: 10.1016/j.cub.2015.09.066.
Rinaldo, S. *et al.* (2011) 'The catalytic mechanism of Pseudomonas aeruginosa cd1 nitrite reductase', *Biochemical Society Transactions*, 39(1), pp. 195–200. doi: 10.1042/BST0390195.
Roco, C. A. *et al.* (2017) 'Modularity of nitrogen-oxide reducing soil bacteria: linking phenotype to genotype', *Environmental Microbiology*, 19(6), pp. 2507–2519. doi: 10.1111/1462-2920.13250.
Sánchez, C. *et al.* (2011) 'Nitric oxide detoxification in the rhizobia-legume symbiosis', *Biochemical Society Transactions*, 39(1), pp. 184–188. doi: 10.1042/BST0390184.
Sanford, R. A. *et al.* (2012) 'Unexpected nondenitrifier nitrous oxide reductase gene diversity and abundance in soils', *Proceedings of the National Academy of Sciences*, 109(48), pp. 19709–19714. doi: 10.1073/pnas.1211238109.
Schutte, C. A. *et al.* (2015) 'Intense nitrogen cycling in permeable intertidal sediment revealed by a nitrous oxide hot spot', *Global Biogeochemical Cycles*. John Wiley & Sons, Ltd, 29(10), pp. 1584–1598. doi: https://doi.org/10.1002/2014GB005052.
Seefeldt, L. C. *et al.* (2013) 'Nitrogenase reduction of carbon-containing compounds', *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1827(8), pp. 1102–1111. doi: https://doi.org/10.1016/j.bbabio.2013.04.003.
Selberg, A. G. A., Gaucher, E. A. and Liberles, D. A. (2021) 'Ancestral Sequence Reconstruction: From Chemical Paleogenetics to Maximum Likelihood Algorithms and Beyond', *Journal of Molecular Evolution*, 89(3), pp. 157–164. doi: 10.1007/s00239-021-09993-1.
Shen, J.-D. *et al.* (2021) 'Nitrilase: a promising biocatalyst in industrial applications for green chemistry', *Critical Reviews in Biotechnology*. Taylor & Francis, 41(1), pp. 72–93. doi: 10.1080/07388551.2020.1827367.
Si Quang, L., Gascuel, O. and Lartillot, N. (2008) 'Empirical profile mixture models for phylogenetic reconstruction', *Bioinformatics*, 24(20), pp. 2317–2323. doi: 10.1093/bioinformatics/btn445.
Sickerman, N. S., Hu, Y. and Ribbe, M. W. (2019) 'Nitrogenases.', *Methods in molecular biology (Clifton, N.J.)*. United States, 1876, pp. 3–24. doi: 10.1007/978-1-4939-8864-8_1.
Siddiq, M. A., Hochberg, G. K. and Thornton, J. W. (2017) 'Evolution of protein specificity: insights from ancestral protein reconstruction', *Current opinion in structural biology*. 2017/08/23, 47, pp. 113–122. doi: 10.1016/j.sbi.2017.07.003.
Sievert, S. M. *et al.* (2008) 'Genome of the epsilonproteobacterial chemolithoautotroph Sulfurimonas denitrificans', *Applied and Environmental Microbiology*, 74(4), pp. 1145–1156. doi: 10.1128/AEM.01844-07.
Sigman, D. M. and Fripiat, F. (2019) 'Nitrogen Isotopes in the Ocean☆', in Cochran, J. K., Bokuniewicz, H. J., and Yager, P. L. B. T.-E. of O. S. (Third E. (eds). Oxford: Academic Press, pp. 263–278. doi: https://doi.org/10.1016/B978-0-12-409548-9.11605-7.

Smith, D. *et al.* (2014) 'Substrate Channel in Nitrogenase Revealed by a Molecular Dynamics Approach', *Biochemistry*. American Chemical Society, 53(14), pp. 2278–2285. doi: 10.1021/bi401313j.
Speirs, L. B. M. *et al.* (2019) 'The Phylogeny, Biodiversity, and Ecology of the Chloroflexi in Activated Sludge ', *Frontiers in Microbiology* . Available at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.02015.
Spieck, E. *et al.* (2020) 'Extremophilic nitrite-oxidizing Chloroflexi from Yellowstone hot springs', *ISME Journal*. Springer US, 14(2), pp. 364–379. doi: 10.1038/s41396-019-0530-9.
Stein, L. Y. *et al.* (2007) 'Whole-genome analysis of the ammonia-oxidizing bacterium, Nitrosomonas eutropha C91: Implications for niche adaptation', *Environmental Microbiology*, 9(12), pp. 2993–3007. doi: 10.1111/j.1462-2920.2007.01409.x.
Strock, J. S. (2008) 'Ammonification', in Jørgensen, S. E. and Fath, B. D. B. T.-E. of E. (eds). Oxford: Academic Press, pp. 162–165. doi: https://doi.org/10.1016/B978-008045405-4.00256-1.
Strohm, T. O. *et al.* (2007) 'Growth Yields in Bacterial Denitrification and Nitrate Ammonification', *Applied and Environmental Microbiology*. American Society for Microbiology, 73(5), pp. 1420–1424. doi: 10.1128/AEM.02508-06.
Stüeken, E. E. *et al.* (2015) 'Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr', *Nature*, 520(7549), pp. 666–669. doi: 10.1038/nature14180.
Sugitani, K. *et al.* (2010) 'Biogenicity of Morphologically Diverse Carbonaceous Microstructures from the ca. 3400 Ma Strelley Pool Formation, in the Pilbara Craton, Western Australia', *Astrobiology*. Mary Ann Liebert, Inc., publishers, 10(9), pp. 899–920. doi: 10.1089/ast.2010.0513.
Supreetha, K. *et al.* (2019) 'Advances in cloning, structural and bioremediation aspects of nitrile hydratases', *Molecular Biology Reports*, 46(4), pp. 4661–4673. doi: 10.1007/s11033-019-04811-w.
Sutherland, J. D. (2016) 'The Origin of Life - Out of the Blue', *Angewandte Chemie - International Edition*, 55(1), pp. 104–121. doi: 10.1002/anie.201506585.
Tatsaki, E. *et al.* (2021) 'Identification of New Specificity Determinants in Bacterial Purine Nucleobase Transporters based on an Ancestral Sequence Reconstruction Approach', *Journal of Molecular Biology*, 433(24), p. 167329. doi: https://doi.org/10.1016/j.jmb.2021.167329.
'The PyMOL Molecular Graphics System' (no date). Schrödinger, LLC.
Thomazo, C. and Papineau, D. (2013) 'Biogeochemical cycling of nitrogen on the early earth', *Elements*, 9(5), pp. 345–351. doi: 10.2113/gselements.9.5.345.
Thorne, J. L., Kishino, H. and Painter, I. S. (1998) 'Estimating the rate of evolution of the rate of molecular evolution.', *Molecular Biology and Evolution*, 15(12), pp. 1647–1657. doi: 10.1093/oxfordjournals.molbev.a025892.
Tian, F., Kasting, J. F. and Zahnle, K. (2011) 'Revisiting HCN formation in Earth's early atmosphere', *Earth and Planetary Science Letters*, 308(3), pp. 417–423. doi: https://doi.org/10.1016/j.epsl.2011.06.011.
Todd, Z. R. and Öberg, K. I. (2020) 'Cometary Delivery of Hydrogen Cyanide to the Early Earth.', *Astrobiology*. United States, 20(9), pp. 1109–1120. doi: 10.1089/ast.2019.2187.
Tria, F. D. K., Landan, G. and Dagan, T. (2017) 'Phylogenetic rooting using minimal ancestor deviation', *Nature ecology & evolution*. doi: 10.1038/s41559-017-0193.
Tricou, T., Tannier, E. and de Vienne, D. M. (2022) 'Ghost Lineages Highly Influence the Interpretation of Introgression Tests', *Systematic Biology*, 71(5), pp. 1147–1158. doi: 10.1093/sysbio/syac011.

Uniprot Consortium, T. (2021) 'UniProt: the universal protein knowledgebase in 2021', *Nucleic Acids Research*, 49(D1), pp. D480–D489. doi: 10.1093/nar/gkaa1100.
Vialle, R. A., Tamuri, A. U. and Goldman, N. (2018) 'Alignment Modulates Ancestral Sequence Reconstruction Accuracy', *Molecular Biology and Evolution*, 35(7), pp. 1783–1797. doi: 10.1093/molbev/msy055.
Ward, L. M. *et al.* (2018) 'Evolution of phototrophy in the Chloroflexi phylum driven by horizontal gene transfer', *Frontiers in Microbiology*. doi: 10.3389/fmicb.2018.00260.
Ward, L. M., McGlynn, S. E. and Fischer, W. W. (2018) 'Draft genome sequence of a divergent anaerobic member of the Chloroflexi class Ardenticatenia from a sulfidic hot spring', *Genome Announcements*, 6(25), pp. 1–2. doi: 10.1128/genomeA.00571-18.
Ward, L. M. and Shih, P. M. (2021) 'Granick revisited: Synthesizing evolutionary and ecological evidence for the late origin of bacteriochlorophyll via ghost lineages and horizontal gene transfer', *PLOS ONE*. Public Library of Science, 16(1), p. e0239248. Available at: https://doi.org/10.1371/journal.pone.0239248.
Waterhouse, A. *et al.* (2018) 'SWISS-MODEL: Homology modelling of protein structures and complexes', *Nucleic Acids Research*. doi: 10.1093/nar/gky427.
Waterhouse, A. M. *et al.* (2009) 'Jalview Version 2-A multiple sequence alignment editor and analysis workbench', *Bioinformatics*. doi: 10.1093/bioinformatics/btp033.
Wei, W. *et al.* (2015) 'Higher diversity and abundance of denitrifying microorganisms in environments than considered previously', *ISME Journal*. Nature Publishing Group, 9(9), pp. 1954–1965. doi: 10.1038/ismej.2015.9.
Wernersson, R. and Pedersen, A. G. (2003) 'RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences', *Nucleic acids research*. Oxford University Press, 31(13), pp. 3537–3539. doi: 10.1093/nar/gkg609.
Whelan, S. and Goldman, N. (2001) 'A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach', *Molecular Biology and Evolution*, 18(5), pp. 691–699. doi: 10.1093/oxfordjournals.molbev.a003851.
Williams, T. A. *et al.* (2015) 'New substitution models for rooting phylogenetic trees', *Philosophical Transactions of the Royal Society B: Biological Sciences*. Royal Society, 370(1678), p. 20140336. doi: 10.1098/rstb.2014.0336.
Woese, C. R. (2002) 'On the evolution of cells.', *Proceedings of the National Academy of Sciences of the United States of America*, 99(13), pp. 8742–8747. doi: 10.1073/pnas.132266999.
Wu, M. and Scott, A. J. (2012) 'Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2', *Bioinformatics*. doi: 10.1093/bioinformatics/bts079.
Yamada, T. and Sekiguchi, Y. (2018) 'Anaerolineaceae', *Bergey's Manual of Systematics of Archaea and Bacteria*. (Major Reference Works), pp. 1–5. doi: https://doi.org/10.1002/9781118960608.fbm00301.
Yang, S. and Bourne, P. E. (2009) 'The Evolutionary History of Protein Domains Viewed by Species Phylogeny', *PLOS ONE*. Public Library of Science, 4(12), p. e8378. Available at: https://doi.org/10.1371/journal.pone.0008378.
Yang, Z., Wong, W. S. W. and Nielsen, R. (2005) 'Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection', *Molecular Biology and Evolution*, 22(4), pp. 1107–1118. doi: 10.1093/molbev/msi097.
Zahnle, K. J. (1986) 'Photochemistry of methane and the formation of hydrocyanic acid (HCN) in the Earth's early atmosphere.', *Journal of Geophysical Research*, 91(D2), pp. 2819–2834. doi: 10.1029/JD091iD02p02819.

Zahnle, K., Schaefer, L. and Fegley, B. (2010) 'Earth's earliest atmospheres', *Cold Spring Harbor perspectives in biology*. 2010/06/23. Cold Spring Harbor Laboratory Press, 2(10), pp. a004895–a004895. doi: 10.1101/cshperspect.a004895.
Zumft, W. G. (1997) 'Cell biology and molecular basis of denitrification.', *Microbiology and molecular biology reviews : MMBR*, 61(4), pp. 533–616. doi: http://dx.doi.org/10.1016/j.sbspro.2014.08.122.
Zumft, W. G. (2005) 'Nitric oxide reductases of prokaryotes with emphasis on the respiratory, heme-copper oxidase type', *Journal of Inorganic Biochemistry*, 99(1), pp. 194–215. doi: 10.1016/j.jinorgbio.2004.09.024.