INTRODUCTION

Nitrogen is a critical element to life on Earth, important as an essential building block in the synthesis of biological molecules, and for its role in redox reactions for microbial energy metabolism. It is often a limiting nutrient in marine and terrestrial environments and likely had a significant influence on the evolutionary trajectory of the biosphere over Earth’s history. The nitrogen cycle is largely controlled by a variety of microorganisms that enzymatically catalyze the reduction and oxidation of nitrogen at various redox states (Figure 1).
Reconstructing the genetic proliferation of these enzymes across the tree of life through gene birth, duplication, loss, and horizontal gene transfer can therefore provide novel insights into the evolution of the biosphere and its productivity over time.

The most important steps in Earth’s nitrogen cycle are largely catalyzed by microbes, including the first crucial step of reducing molecular nitrogen to bioavailable forms (Kuypers et al., 2018; Zerkle & Mikhail, 2017). Nitrogen fixation is catalyzed by nitrogenase, of which there are three varieties, distinguished by the metal in their associated active site cofactors: Nif (Fe-Mo), Vnf (Fe-V), and Anf (Fe-Fe) (Joerger et al., 1988; Miller & Eady, 1988) (Figure 1). The ability to fix nitrogen is spread across a wide range of archaeal and bacterial lineages, but does not occur in eukaryotes (Dos Santos et al., 2012; Gaby & Buckley, 2014). Importantly, nitrogenase is strongly inhibited by oxygen, forcing nitrogen fixers to develop various means to reduce their intracellular oxygen concentrations or to confine themselves to suboxic environments (Gallon, 1981). Ammonium produced from nitrogen fixation or ammonification is converted to organic forms of fixed nitrogen by a variety of enzymes or, in the presence of oxygen, oxidized to nitrite ($\text{NO}_2^-$) or nitrate ($\text{NO}_3^-$) through the chemoautotrophic nitrification pathway via the enzymes Amo and Hao (Figure 1). In environments with insufficient $\text{O}_2$ concentrations to support aerobic respiration, nitrate and nitrite can be utilized as alternative terminal electron acceptors through the denitrification pathway; consequently, denitrification rates are highest in suboxic conditions, including swamps and marine oxygen minimum zones (Canfield et al., 2005; Lősch et al., 2012; Voss et al., 2013). Reduction of nitrate to nitrite, nitric oxide (NO), nitrous oxide ($\text{N}_2\text{O}$), and dinitrogen ($\text{N}_2$) is catalyzed by the Nas, Nar, Nap, Nir, Nor, and Nos enzymes, respectively. Dissimilatory nitrate reduction to ammonium (DNRA) also reduces nitrate via the enzymes Nar, Nap, Nir, and Nrf (Figure 1), but this pathway retains fixed nitrogen as ammonium and may therefore have been a critical metabolism in nutrient-starved ecosystems. DNRA, while less understood than denitrification, has been shown to be a major nitrate sink in a variety of aquatic systems, especially warm intertidal zones (Giblin et al., 2013) and it may be dominant under ferruginous conditions, as suggested by modern analogue studies (Michiels et al., 2017).

Given the importance of nitrogen as a building block of life, as an energy source for microbes, and as the most abundant element in the Earth’s atmosphere, better constraining the evolutionary history of the nitrogen cycle is important for understanding, among other things, variation in global primary productivity and atmospheric pressure over time. The bioavailability and cycling of important limiting nutrients through Earth’s history, including nitrogen, would have been important for biological productivity and the rise of early eukaryotic algae (Anbar & Knoll, 2002; Isson et al., 2018; Sánchez-Baracaldo et al., 2014). The relative abundance of nitrogenous gases in the atmosphere could also have had important implications for atmospheric pressure as well as planetary climate during the Archean. Potential changes in atmospheric pressure during the Archean may have resulted from biological $\text{N}_2$ drawdown (Som et al., 2016), whereas the greenhouse gas nitrous oxide ($\text{N}_2\text{O}$), produced as part of the nitrogen cycle via nitrification/denitrification, may have contributed to planetary warming when the Sun was younger and fainter (Buick, 2007; Roberson et al., 2011; Stanton et al., 2018). Finally, one of the most important questions for the early evolution of life is understanding when fixed nitrogen first became widely available, and through what means. Experimental data suggest that fixed nitrogen can be produced during lightning reactions and under hydrothermal conditions (e.g., Brandes et al., 1998; Navarro-González et al., 2001), and either or both of these sources were likely pivotal for the origin of life. However, the invention of biological $\text{N}_2$ fixation would have made Earth’s biosphere less dependent on abiotic reactions and likely spurred primary productivity.

The question of how these metabolisms unfolded over Earth’s history has previously been addressed with both geochemical and phylogenetics-based approaches. Geochemical approaches, relying on the reconstruction of metabolisms based on the rock record, have suggested that biological nitrogen fixation emerged early (Koehler et al., 2018; Ossa Ossa et al., 2019; Stüeken et al., 2016) and that the nitrogen cycle expanded considerably during the Neoarchean (2.8–2.5 Ga) and Paleoproterozoic (2.5–1.8 Ga) (Garvin et al., 2009a; Zerkle, Poulton, et al., 2017). Phylogenetics studies, relying on sequence data, reconstruct the evolutionary history of genes of interest (e.g., Boyd et al., 2011; Garcia et al., 2020; Jones et al., 2020; Neuhoff et al., 2019; Parsons et al., 2013; Stanton et al., 2018).
et al., 2008), and some molecular clock studies have yielded conservative estimates for the approximate timing of an enzyme’s origin (Boyd et al., 2011; Boyd & Peters, 2013; Raymond et al., 2004). While each approach provides valuable insights, they both have weaknesses. Geochemical data cannot reliably distinguish between all enzymatic pathways, because the isotopic effects of some reactions (e.g., denitrification, DNRA, and ANAMMOX) are too similar to each other. Furthermore, geochemical data, which are typically collected from bulk rock samples, only preserve evidence of the most dominant metabolisms and may therefore not capture the origin of new enzymes until they gain ecological significance. Conversely, the phylogenetic approach of dating the antiquity of enzymes does not show when these enzymes gained ecological dominance. Furthermore, most phylogenetic studies have so far focused on nitrogenase, leaving the evolutionary history of most nitrogen-cycling enzymes poorly constrained. Additional work is therefore needed to address key questions about the dynamics of the nitrogen cycle on the early Earth and its evolution over time. The proliferation of sequencing data over the past decade has made vast amounts of genomic data available, which can provide novel insights into the evolution of nitrogen-cycling genes over time.

As a new approach to these questions, we track the timing of birth, speciation, duplication, loss, and horizontal gene transfer events for genes involved in each step of the nitrogen cycle, which can provide contextual information for the rise and spread of key nitrogen-cycling genes across the tree of life. We place particular focus on the acquisition of new genes via horizontal gene transfer (HGT), which is common in microbial lineages and is a crucial evolutionary mechanism by which a clade of organisms can develop new and useful phenotypes without the evolutionary cost associated with independently evolving genes (Beiko et al., 2005; Gogarten & Townsend, 2005). Studies of gene gain and loss have revealed a history of widespread HGT throughout the microbial tree of life, which we have attempted to leverage as a means to attribute trends in microbial evolution to events in Earth history (Koonin et al., 2001; Mirkin et al., 2003). Many of the major genes in the nitrogen cycle have been shown to have experienced extensive HGT, presumably due to their modularity and their strong dependence on oxygen availability (Jones et al., 2008; Kechris et al., 2006; Stolz & Basu, 2002). We tracked birth, speciation, duplication, loss, and HGT of genes in the nitrogen cycle over time by comparing the phylogenies for specific nitrogen-metabolizing genes to a time-calibrated tree of life to demonstrate when these genes first arose and then spread across the tree of life on the early Earth.

2 | MATERIALS AND METHODS

2.1 | Genome selection and compilation

The construction of both the gene and species trees for this study was based upon the manual curation of a genome database containing 308 genomes (including 254 bacterial and archaeal genomes that served as the basis for the species tree and was subsequently searched to find genes related to nitrogen metabolism. Assembled genomes were downloaded from ggKBase (Hug et al., 2016) and the NCBI assembly database (Kitts et al., 2016). Additionally, six genomes were collected from a recent study identifying novel nitrogen fixers (Delmont et al., 2018). In constructing the tree, we included at least one genome from each bacterial or archaeal phylum represented in the most recent comprehensive tree of life (Hug et al., 2016) in order to create a tree fully representative of our current understanding of microbial diversity. It also includes a set of genomes associated with a database of nifH genes (Gaby & Buckley, 2014). Some eukaryotic genomes were included for construction of the species tree, but for this study we focused only on archaeal and bacterial genomes for identification of nitrogen-cycling genes. Relative to archaea and bacteria, eukaryotes play a more minor role in the nitrogen cycle—while some species of fungi and protists reduce nitrate or nitrite to more reduced forms of nitrogen, there are no known eukaryotes that mediate nitrogen fixation, nitrification, DNRA, or anammox (Stein & Klotz, 2016).

2.2 | Species tree and chronogram construction

To create the species tree, all bacterial and archaeal genomes in our database were mined for single-copy ribosomal protein sequences L2, L3, L4, L5, L6, L14, L15, L16, L18, L22, L24, S3, S8, S10, S17, and S19 using Phylosift (Darling et al., 2014), with the isolate and best hit command line flags. These 16 ribosomal proteins represent the same proteins used to create a recent comprehensive tree of life (Hug et al., 2016) and, in the case of eukaryotes, ribosomal sequences were directly drawn from their dataset. All genomes included in the dataset contained fewer than 50% gaps in the alignment. These 16 single-copy ribosomal proteins were concatenated to create a final alignment of 2,897 characters for phylogenetic reconstruction and molecular clock evaluation.

Alignments for the species trees were made using the Phylosift pipeline (Darling et al., 2014) and curated to only include the target ribosomal proteins listed above. The phylogeny was constructed using RAxML v.8.2.9 with 100 rapid bootstraps (Stamatakis, 2014). The CAT model, which calculates site-specific evolutionary rates, was used with an LG substitution matrix to construct the species tree based on reference marker genes (the species tree). The root for the species tree was placed in the Bacterial domain (Fournier & Gogarten, 2010).

2.3 | Chronogram construction

The species tree was converted into chronograms using PhyloBayes (Lartillot et al., 2009) using different clock models and calibration points in order to test the sensitivity of our results to variation in Phylobayes parameters. The root age was set via a normally distributed gamma root prior according to the liberal or conservative
calibration points set in Table 1 and Table S1, with standard deviation set to 200 in accordance with previous studies (Magnabosco et al., 2018). We tested two separate sets of calibration points, one liberal (which represents the earliest date for which there is any evidence of a given event based on the current scientific literature) and one conservative (which represents the earliest date for which there is the most consensus for a given event based on the current scientific literature), to test the sensitivity of the methodology (Table 1 and Table S1). For each set of internal calibration points, the ages in the calibration files were set as the hard lower bound for the analysis. The liberal calibration points shown in Table S1 yielded unrealistic root ages (>4.5 Ga) and therefore were not used for further analyses.

We created chronograms using both the uncorrelated gamma (UGAM) model (Drummond et al., 2006) and the autocorrelated CIR relaxed clock model (Lepage et al., 2007) to compare the effects of clock model type on the results. Two chains were run in parallel for each set of parameters, so that the two concurrent runs could be compared to one another as a test of convergence. Convergence of the MCMC chains was checked visually by plotting the summary statistics and quantitatively by comparing the posterior distributions of two parallel chains using the tracecomp and bpcomp programs in PhyloBayes. We required an effective size >100 and a maximum difference between chains of <0.3. Simultaneous chains were run for approximately 36,000 cycles. Chronograms were generated using the readiv function in Phylobayes 4.1, with approximately 20% of initial cycles discarded as burn-in. Chronograms were visualized using the phytools package in R (Revell, 2012). In order to test for the influence of the priors, we generated additional chronograms in the absence of sequence data using the -prior flag in Phylobayes. These chronograms displayed substantially different node timings, demonstrating that the priors did not overly influence the inferred dates and the sequence data informed the chronogram. Results from molecular clock analyses should be interpreted with caution, given the limitations associated with these analyses, including but not limited to changing generation times, the influence of natural selection, and variation in mutational rates across species (Ayala, 1999; Bromham et al., 2018; Schwartz & Maresca, 2006). Here, we attempted to ameliorate some of these challenges by using methods allowing for variation of rates between and across lineages, and by comparing results produced by different clock models. Finally, our goal with this analysis was not to pinpoint exact dates for many of the transitions discussed here, but rather to compare the relative timing on broad evolutionary scales.

### 2.4 | Identification of nitrogen-cycling genes and construction of gene trees

Curated gene database queries for each of the nitrogen-cycling genes, we investigated were generated based on KEGG orthologies (Ogata et al., 1999) and downloaded from the UniProt database (The Uniprot Consortium, 2017). Nitrogen-cycling amino acid sequences for the gene trees were identified by conducting BLASTP (Altschul et al., 1990) searches of the open reading frames (ORFs) of every genome from the collection of 254 genomes. All ORFs were identified using Prodigal (Hyatt et al., 2010). The maximum e-value cutoff for BLAST hits was $10^{-12}$ and matches were excluded if the length of the local alignment was <50% of the length of the query sequence. BLAST results were compared with results from AnnoTree (Mendler et al., 2019) to verify gene distributions. Given that $nifK$ and $nifD$ have a shared evolutionary history (Fani et al., 2000), an e-value cutoff of $10^{-30}$ was selected based on close examination of BLASTP hit results and KEGG annotations, which was sufficient to distinguish these subunits. For nitrogenase subunits $nifH$, $vnfD$, $vnfK$, $anfD$, and $anfK$, we manually curated alignments by identifying key residues that were crucial for enzyme structure and function, as determined through literature searches and visualization using PyMol (Briegle et al., 1987; Howard et al., 2013; Kaiser et al., 2011; Keable et al., 2018; McGlynn et al., 2012). Jalview (Waterhouse et al., 2009) was used for visualization of key residues in alignments. All genes that did not include the key residues were removed from the alignment. Finally, to ensure that only the genes of interest were included in alignments and to verify annotations, all genes identified in the

| Calibration event | Date (Mya) | Refs |
|-------------------|-----------|------|
| LUCA (set as root prior) | 3,800, 200 S.D. | Schidlovski et al. (1983); Schidlovski (1988); Mozgis et al. (1996); Rosing (1999); Czaja et al. (2013); Nutman et al. (2016) |
| Origin of Methanogenesis | >2,700 | Eigenbrode and Freeman (2006) |
| Origin of Cyanobacteria | >2,450 | Bekker et al. (2004) |
| Origin of Eukaryotes | >1,700 | Pang et al. (2013) |
| Origin of plastids/ Rhodophytes diverge | >1,050 | Gibson et al. (2018) |
| Akinetes diverge from cyanobacteria lacking cell differentiation | >1,000 | Pang et al. (2018) |

**TABLE 1** Fossil calibration points used in Phylobayes runs. Calibration points were set as the hard constraint indicating the latest date by which a specific clade split. The selected time points reflect the dates for which there is the most consensus.
BLAST search were compared to the KEGG database using Kofam Koala (Aramaki et al., 2019), which assigns KEGG Ortholog numbers to each gene by using a homology search against a database of profile hidden Markov models. Only genes verified to be the gene of interest according to Kofam Koala were retained for downstream analysis. It is important to note that gene identification is necessarily limited by the search tools and databases used for annotation, and the methods used here were chosen to be conservative so as to ensure the removal of non-target genes from the analysis.

Alignments for the gene trees were created using MUSCLE (Edgar, 2004) and trimmed with TrimAl (Capella-Gutierrez et al., 2009) using the -automated1 option. The model of evolution was selected using Model Selection as implemented in IQ-TREE (Kalyaanamoorthy et al., 2017) using the default parameters. Trees were generated with RAxML-NG (Kozlov et al., 2019) using the model of evolution identified in IQ-TREE. Trees were run with at least 1,000 bootstraps or until the diagnostic statistic based on the MRE-based bootstrapping test as implemented in RAxML-NG dropped below a cutoff of 0.03.

2.5 | Gene tree and species chronogram reconciliation

Gene trees were reconciled with species chronograms using the Analyzer of Gene and Species Trees (AnGST) (David & Alm, 2011). AnGST compares the topology of the gene tree with that of the species tree, rather than rely solely on presence and absence patterns, in order to identify gene birth, transfer, duplication, and loss events. Event penalties were set to hgt: 3, dup: 2, los: 1, and spc: 0. Ultrametric was set to True in order to constrain events temporally. Each run was conducted with 100 gene tree bootstraps in order to
nirB

nrfA

napA

narG

nasA

nirS

nirK

norB

nosZ

- nitrogen fixation
- denitrification from nitrate
- denitrification from nitrite
- nitrification
- DNRA
increase accuracy (David & Alm, 2011). Individual event timings were defined as the midpoint of the temporal region during which a given event could occur.

3 | RESULTS

3.1 | Generation of a species tree and fossil-calibrated chronogram

We compared all gene trees to a species tree constructed from an alignment of concatenated sequences of 16 single-copy universal proteins from 308 organisms (Figure 2; genome list deposited to FigShare at https://doi.org/10.6084/m9.figshare.12791747.v1, see below). Monophyly was preserved for most major phyla, with three exceptions: (a) Tenericutes is contained within Firmicutes, (b) the PVC superphylum contains Omnitrophica, and (c) Lentisphaerae is nested within Verrucomicrobia. Additionally, in contrast to a recently published comprehensive tree of life (Hug et al., 2016), our tree does not place the recently discovered Candidate Phyla Radiation (CPR) bacteria as the deepest-rooted bacterial clade; however, the CPR is placed as a sister group to the Cyanobacteria and Melainabacteria, which is consistent with that phylogeny (Hug et al., 2016). For the purposes of this study, placement of these groups should not greatly affect our results due to the high number of duplication/loss/transfer events we inferred overall across all groups. The species tree used for this analysis is a three-domain tree, in contrast to recent studies which have shown the addition of the Asgard Archaea to the tree of life to cause Eukaryotes to group within Archaea (Zaremba-Niedzwiedzka et al., 2017). However, we are agnostic as to the placement of the Asgard superphylum and the eukaryotes on the tree of life, as this was not our focus; the relationship of the three domains to one another should not substantially affect the results shown here due to the relative infrequency of interdomain HGT relative to interdomain HGT, and the exclusion of eukaryotic nitrogen-cycling genes.

We constructed four different chronograms from our species tree using two different clock models (UGAM and CIR) as well as liberal (representing the earliest date for which there is any evidence of a given event based on the current scientific literature) and conservative (representing the earliest date for which there is the most consensus for a given event based on the current scientific literature) fossil calibration points (see Methods). The liberal calibration points yielded an unreasonable root age (>4.5 Ga) and so were not further used for analysis. The UGAM clock model yielded a greater spread in estimated ages for node divergences, with an earlier root age (approximately 4,044.68 ± 143.093 Mya compared to 3,982 ± 131.218 Mya for the CIR clock model). The results shown in Figure 3 and Table 4 derive from the CIR clock model as this has been previously shown to outperform uncorrelated models (Lepage et al., 2007), but the results from the UGAM clock model are shown in Figure S3 and Table S2. All genome lists, alignments, and Newick files have been deposited in FigShare at: https://figshare.com/projects/Radiation_of_nitrogen_cycling_genes_across_the_tree_of_life/87461

3.2 | Identification of duplication, speciation, loss, and horizontal gene transfer events for nitrogen-cycling genes

We identified 18 different nitrogen-cycling genes for analysis of horizontal gene transfer, speciation, duplication, and loss events (Table 2). This methodology is best suited to genes that have many representatives from a diverse suite of taxa. Very few genes from...
the hao and amo gene families were identified in these genomes or passed our stringent filtration tests, and thus, they did not yield sufficient data for robust conclusions and were not included in further analyses. We generated maximum likelihood gene trees for each of the remaining genes (see Data Availability Statement below) and compared these gene trees to the fossil-calibrated chronogram in order to identify and infer the timing of birth, duplication, loss, and horizontal gene transfer events for each of these genes. The number of speciation events inferred for each gene in this study was relatively high compared to the number of inferred loss and HGT events (Table 3), with very few duplication events observed. Inferred loss events generally skewed younger than inferred horizontal gene transfer and speciation events. Our results indicate that the iron-molybdenum nitrogenases (nifH/nifD/nifK) were among the oldest genes, originating during the Archean era and beginning to spread across the tree of life via horizontal gene transfer fairly early in Earth history (Figure 3). In contrast, the alternative nitrogenases anf and vnf were inferred to have arisen and radiated across the tree of life much later (Table 4, Table S2, Figure S4 and S5), but very few anf and vnf genes were identified among our sample set and, therefore, these results should be treated with caution (Table 2). The denitrifying genes nrfA and nirB, which are involved in the DNRA process, are inferred to have arisen by approximately 2.7–2.2 Ga, and began to increasingly proliferate widely across the tree of life into the Mesoproterozoic. Genes involved in nitrate reduction, including narG, nasA, and napA, were inferred to have arisen relatively early (approximately 2.8 Ga for narG and napA; approximately 2.3 Ga for nasA) (Table 4), and we inferred a few speciation and horizontal gene transfer events for these genes between 2–2.5 Ga, but we did not observe a rise in HGT events for these genes until much later, at approximately 1.5 Ga (Figure 3). We observed a similar trend for nxxA, a gene involved in nitrite oxidation, which was inferred to have arisen around 2.8 Ga but did not exhibit a rise in speciation, duplication, or transfer events until approximately 1.5 Ga (Figure 3, Table 4).

Taken together, we observed that genes related to the fixation of nitrogen from dinitrogen gas to ammonium arose early and proliferated across the tree of life relatively quickly, while genes related to nitrate reduction and nitrite oxidation also arose early but did not begin to proliferate across the tree of life until later. Genes related to denitrification, particularly downstream from nitrite, arose much later (Figure 3). Our results regarding the timing of the birth, duplication, speciation, loss and HGT events for specific nitrogen-cycling genes showed a few differences between the CIR (Figure 3, Table 4) and UGAM clock models (Figure S3, Table S2). However, the overall patterns in the relative timing for the birth and spread of specific genes in the nitrogen cycle were similar regardless of the type of clock models used.

### 4 | DISCUSSION

We have focused here on tracking gene birth, duplication, speciation, and particularly horizontal gene transfer events across deep
TABLE 4  Inferred birth dates for nitrogen-cycling genes based on the chronogram generated using the CIR clock model with conservative calibration points

| Gene  | Upper node date | 95% confidence interval | Lower node date | 95% confidence interval | Midpoint between nodes | Geologic era       |
|-------|-----------------|-------------------------|-----------------|-------------------------|------------------------|------------------|
| vnfD  | 1445.24         | 1052.37, 1735.83        | 482.23          | 210.11, 778.444         | 963.73                 | Neoproterozoic   |
| vnfK  | 1445.24         | 1052.37, 1735.83        | 482.23          | 210.11, 778.444         | 963.73                 | Neoproterozoic   |
| nosZ  | 1325.57         | 1048.93, 1588.52        | 1139.65         | 851.174, 1408.81        | 1232.61                | Mesoproterozoic  |
| nirK  | 1452.66         | 1193.07, 1706.53        | 1343.08         | 1094.21, 1585.92        | 1397.87                | Mesoproterozoic  |
| nirS  | 1913.87         | 1694.13, 2144.31        | 1423.17         | 1153.32, 1677.23        | 1668.52                | Paleoproterozoic |
| norB  | 1913.87         | 1694.13, 2144.31        | 1423.17         | 1153.32, 1677.23        | 1668.52                | Paleoproterozoic |
| anfD  | 2349.58         | 2153.39, 2583.93        | 1745.53         | 1498.39, 1999.54        | 2047.55                | Paleoproterozoic |
| anfK  | 2349.58         | 2153.39, 2583.93        | 1745.53         | 1498.39, 1999.54        | 2047.55                | Paleoproterozoic |
| nasA  | 2540.50         | 2351.4, 2778.03         | 2063.23         | 1848.31, 2293.91        | 2301.87                | Paleoproterozoic |
| nrfA  | 2454.68         | 2178.51, 2737.37        | 2267.54         | 1918.96, 2574.71        | 2361.11                | Paleoproterozoic |
| nirB  | 2748.37         | 2545.88, 2989.62        | 2603.25         | 2384.99, 2850.17        | 2675.81                | Neoarchean        |
| nifK  | 2926.54         | 2734.92, 3180.11        | 2629.22         | 2415.05, 2877.89        | 2777.88                | Neoarchean        |
| nrxA  | 2816.89         | 2630.27, 3063.26        | 2787.46         | 2603.09, 3029.43        | 2802.17                | Mesoarchean       |
| napA  | 2816.89         | 2630.27, 3063.26        | 2787.46         | 2603.09, 3029.43        | 2802.17                | Mesoarchean       |
| narG  | 2816.89         | 2630.27, 3063.26        | 2787.46         | 2603.09, 3029.43        | 2802.17                | Mesoarchean       |
| nifD  | 2934.72         | 2744.91, 3187.67        | 2869.82         | 2682.33, 3118.74        | 2902.27                | Mesoarchean       |
| nifH  | 3060.98         | 2863.64, 3325.02        | 3026.38         | 2832.29, 3286.79        | 3043.68                | Mesoarchean       |

All gene birth events are inferred to occur between nodes on the species chronogram, and, therefore our methods do not allow us to infer specific dates for gene birth events. Here, we list the inferred timing for the earliest possible timing (upper node) and latest possible timing (lower node) with 95% confidence intervals as reported by PhyloBayes, as well as the calculated midpoint between the two node times. Times are reported in millions of years ago. The geological era reported in the right column corresponds to the midpoint between the upper and lower node. Given the uncertainty inherent in the estimation of node timings and inference of events, these are provided for geological reference and should not be taken as definitive timings. The equivalent table, generated using the UGAM clock model, is depicted as Table S2.
et al., 2018; Luo et al., 2018; Zerkle, Poulton, et al., 2017). Our results suggest that genes involved in nitrate reduction arose by approximately 2.8 Ga, indicating that nitrate was present and used as a metabolic substrate at that time. However, our data suggest that denitrification may not have been a reliable energy source until about 1.5 Ga and therefore used by a smaller diversity of clades. From the mid-Proterozoic onwards, nitrate may have been sufficiently bio-available to make it a more widely used substrate. In the following, we discuss the implications of our results for the evolution of the nitrogen cycle and its relationship to the redox state of the Earth.

4.1 Nitrogen fixation

Our data indicate that nitrogen fixation through the use of molybdenum nitrogenase (Nif) is an ancient process, arising by approximately 3.1–2.7 Ga (Table 4). Our results are consistent with previous work from geochemical analyses suggesting that nitrogen fixation must have arisen early in order to support an expanding biosphere (Koehler et al., 2019; Ossa et al., 2019; Stüeken et al., 2016). In contrast, a previous analysis based on the evolutionary rate of nitrogenase genes suggested that functional Mo-nitrogenase arose relatively late (approximately 2.2–1.5 Ga) (Boyd et al., 2011), which may also be consistent with the hypothesis that modern planktonic nitrogen fixers did not become abundant in global oceans until the Neoproterozoic (Sánchez-Baracaldo et al., 2014). Our results, which are based on reconciliation of gene trees with a chronogram inferred from universally conserved, single-copy genes, are instead consistent with recent nitrogen isotope evidence for biological nitrogen fixation back to at least 3.2 Ga (Stüeken et al., 2015). It is also consistent with phylogenetics work suggesting that molybdenum nitrogenase arose early in the evolution of life on Earth (Raymond et al., 2004). Similarly, phylogenetic reconstructions showing that Mo-nitrogenases arose before V- and Fe-nitrogenases are consistent with this conclusion (Garcia et al., 2020). It has been argued that nitrogenase was present in LUCA (Weiss et al., 2016), though that has been disputed (Berkemer & McGlynn, 2020; Boyd et al., 2011; Mus et al., 2019) and our results do not resolve this issue. However, the identification of horizontal gene transfer events for nitrogenase subunits during the Archean (Figure 3) suggest that this metabolism may have been abundant and beneficial enough during this time period to have been successfully transferred and retained in microbial genomes.

Our results support the argument that abiotic sources of fixed nitrogen were unlikely to be significant enough to sustain the early biosphere (Canfield et al., 2010; Raymond et al., 2004). It has been proposed that early life received its nitrogen from the lightning-catalyzed reaction between N₂ and CO₂ as a source of NO₃⁻ and that a steady reduction in atmospheric CO₂ levels reduced this flux, eventually leading to a nitrogen crisis around 2.2 Ga that would have favored the rise and spread of biological nitrogen fixation (Navarro-González et al., 2001). In contrast, several studies have concluded that abiotic nitrogen fixation on the early Earth would produce 50- to 5000-fold lower rates of fixed nitrogen than what is contained in the modern ocean, making nitrogen highly limiting to the early biosphere, even at higher CO₂ levels (Canfield et al., 2010). Our data suggest that biological nitrogen fixation arose and proliferated early and that acquisition of the nitrogenase enzyme provided enough of a selective advantage to be successfully transferred across lineages during the early Archean, supporting the notion that fixed nitrogen was not widely available.

Moreover, if the radiation of biological N₂ uptake dates back to the early Archean, it may have had a significant impact on atmospheric pressure. Today, nitrogen makes up approximately 78% of the Earth's modern atmosphere by volume, predominantly as N₂ gas. The atmosphere of the early Earth was probably also N₂-rich, but while some models suggest a 2–4 times larger atmospheric N₂ reservoir in the Archean (Johnson & Goldblatt, 2018), proxy evidence ranges from near modern values (Avic et al., 2018; Marty et al., 2013) to less than half of today's reservoir (Som et al., 2016). If atmospheric N₂ pressure changed over time, it is likely that biological activity would have played a major role in driving these changes by enhancing nitrogen burial in sediments and by accelerating oxidative weathering of nitrogen from continental crust (Stüeken et al., 2016; Zerke & Mikhail, 2017). If nitrogen was fixed by nitrogenase early in the Archean, as indicated by our results, then biological nitrogen burial began long before the onset of oxidative weathering in the Neoproterozoic (Stüeken et al., 2012) which makes it possible that atmospheric N₂ decreased over the course of the Archean. Although further paleobarometric proxies are needed to verify this trend, our results provide an important anchor point for the onset of biological N₂ drawdown.

Lastly, though we identified very few iron and vanadium nitrogenases (anf and vnf, respectively) in our datasets, our results tentatively suggest that these nitrogenases radiated across the tree of life later than the iron-molybdenum nitrogenases (nif), which supports conclusions from other phylogenetics studies (Garcia et al., 2020). The early rise of Mo-dependent nitrogenase would have required a source of Mo to act as a cofactor for Mo-containing nitrogenases, implying that nanomolar levels of dissolved Mo (Reinhard et al., 2013; Scott et al., 2008), presumably derived from anoxic weathering or hydrothermal sources of molybdenum to the early ocean, were sufficient for the development of Nif and the proliferation of nitrogen fixers. Oxidative weathering was therefore evidently not required for the onset of biological nitrogen fixation (cf., Boyd et al., 2011). We speculate that the evolutionary pressure for the diversification of alternative nitrogenases arose in the Neoproterozoic with the rise of eukaryotic algae (Brocks et al., 2017), which may have substantially increased the nitrogen demand in the global ocean. Vanadium and iron-based nitrogenases are less efficient at fixing nitrogen than Nif (Mus et al., 2018), which perhaps lowers the probability of successful gene transfer for Vnf and Anf. However, when nitrogen demand increased in the environment, it is conceivable that more of these genes were shared successfully. Another hypothesis to explain our data is that the gene transfer of alternative nitrogenases was affected by global climate. Vnf becomes more efficient than Nif at cold temperatures (Miller & Eady, 1988) and the activity of Nif-using
cyanobacteria decreases at high latitudes (Brauer et al., 2013). Thus, the cold climate of the Cryogenian ‘Snowball Earth’ period (720–635 Ma) (Hoffman et al., 2017) may have increased the fitness of organisms possessing Vnf. However, these hypotheses remain speculative until more genomic data for Vnf and Anf has been acquired. In any case, it is important to keep in mind that the origin of alternative nitrogenases likely occurred long before the Neoproterozoic. Vanadium nitrogenase in Azotobacter vinelandii has been shown to reduce carbon monoxide (CO) to the hydrocarbons ethylene (C₂H₄), ethane (C₂H₆), and propane (C₃H₈) (Lee et al., 2010), which supports the origin of V-nitrogenase in the Archean eon when CO was more abundant in the atmosphere (Anbar & Knoll, 2002). Additionally, V was bioavailable in Archean ocean waters under slightly acidic conditions (Moore et al., 2020). The genomes of more V-nitrogenase and Fe-nitrogenase organisms must be sequenced to better understand the evolutionary history of these alternative nitrogenases.

4.2 Nitrogen cycling prior to the oxygenation of the oceans and atmosphere

Although our results suggest that nitrogen fixation was an important process on the early Earth, microbial metabolisms making use of more oxidized forms of nitrogen, particularly through the process of denitriﬁcation downstream from nitrite, do not appear to have arisen and spread across the tree of life until much later. Some early steps in the nitrogen cycle, such as nitrate reduction to nitrite via nasA, narG, or napA, appear to have arisen relatively early (in the late Archean or early Proterozoic), consistent with geochemical record of denitriﬁcation at that time (Godfrey & Falkowski, 2009; Koepler et al., 2019). However, we did not identify many horizontal gene transfer events early in the evolutionary history of these genes, and their proliferation across the tree of life only began in earnest much later, approximately 1.5 Ga. Our results therefore suggest that nitrate availability was restricted in space and/or time. Indeed, thermodynamic constraints and environmentally resolved geochemical datasets show that the anoxic deep ocean of the Precambrian contained ammonium while nitrate was restricted to surface waters (Stüeken et al., 2016; Yang et al., 2019). It is therefore likely that the nitrate-reducing genes appeared in locally oxygenated regions of the Archean surface ocean, which have been inferred from isotopic studies back to 3.0 Ga (Olson et al., 2013; Planavsky et al., 2014). In such oxygen oases, oxygen concentrations may have reached micromolar concentrations (Olson et al., 2013), which is high enough for the complete oxidation of ammonium to nitrite and nitrate (Lipschultz et al., 1990). Alternatively, a small ﬂux of nitrogen oxides from lightning and volcanism may have existed (Mather et al., 2004; Navarro-González et al., 2001) and maintained populations of nitrate-reducing denitrifiers. Denitriﬁcation is a highly energy-yielding metabolism (Schoepp-Cothenet et al., 2013), and therefore even a small source ﬂux of nitrate to the early ocean would probably have been exploited. However, we stress that denitriﬁcation left its geochemical mark in only a few Archean basins (Garvin et al., 2009b; Godfrey & Falkowski, 2009; Koepler et al., 2019), while other localities show no geochemical evidence of nitrate reduction (Ossa Ossa et al., 2019). This observation may argue against a signiﬁcant contribution of lightning-derived nitrogen oxides, which we would expect to have been more uniformly distributed.

Our results suggest that genes for nitrite-metabolizing enzymes were not frequently transferred during the early Archean, which may imply that nitrite-metabolizing genes were not widespread in Archean ecosystems. It is possible that only speciﬁc taxa were responsible for the conversion of nitrite into nitrate or ammonium. Alternatively, the conversion of nitrite to ammonium may have happened abiotically. Ferrous iron (Fe²⁺), which is thought to have been abundant in the early anoxic ocean, may have readily reduced nitrite to ammonium or nitrogen gas without biological intervention (Canfield et al., 2010; Summers & Chang, 1993). Nitrite levels during this time period were therefore likely too low to cause substantial proliferation of genes for enzymes which take nitrite as a substrate, while nitrate and ammonium concentration were high enough for selection to favor biological metabolism of these molecules by diverse microorganisms.

4.3 Effects of increased oxygen levels on the nitrogen cycle

Multiple lines of evidence point to increasing oxygenation of surface environments from 2.75 Ga onwards, culminating in the Great Oxidation Event (GOE) at 2.3 Ga (Crowe et al., 2013; Lyons et al., 2014; Noffke et al., 2007; Planavsky et al., 2014). Our data suggest that the genes nrfA and nirB, which reduce a variety of oxidized forms of nitrogen to ammonium through the DNRA pathway, arose in the Neoarchean or Paleoproterozoic, but the relative number of HGT events for this gene began to rise approximately around the time of the GOE. Unlike denitriﬁcation to N₂ gas, DNRA retains the reduced nitrogen product in the system as aqueous NH₄⁺, which may have prevented a potential ‘nitrogen crisis’ during the wake of the GOE (Falkowski & Godfrey, 2008). The isotopic ﬁngerprint of DNRA can so far not be distinguished from denitriﬁcation in the geochemical record, and thus, our results provide the ﬁrst indication that this metabolism may have been of greater ecological importance than previously proposed. DNRA rather than denitriﬁcation may therefore explain geochemical evidence of biological nitrate reduction in the Neoarchean and early Proterozoic (Garvin et al., 2009b; Godfrey & Falkowski, 2009; Kipp et al., 2018; Zerkle et al., 2017), that is, long before our observed radiation in denitriﬁcation genes. This observation may suggest that these early nitrate-reducing ecosystems were primarily performing DNRA, consistent with iron-rich conditions in the Archean and Paleoproterozoic ocean (Michiels et al., 2017).

After approximately 1.5 Ga, the relative number of HGT events for enzymes involved in the modern denitriﬁcation pathway began to rise. Additionally, our data suggest that the gene nxr, which oxidizes nitrite into nitrate, arose relatively early (approximately 2.8 Ga) but did not begin to spread across the tree of life until ~1.5 Ga. One possibility is that increased copper availability facilitated the spread of
that archaeal ammonia oxidizers most likely originated during the Archean (see above), may have slowed down with the deepening of the chemocline after the GOE.

Our results show that biological nitrogen fixation appears to have arisen and proliferated early in the Archaean, and that genes in the denitrification pathway and genes related to the consumption of nitrite and its downstream products began to proliferate across the tree of life following the oxygenation of Earth’s atmosphere and oceans. Moreover, our results support the hypothesis that the molybdenum-based variety of nitrogenase diversified much earlier than alternative nitrogenases, which may only have become important with the increasing nitrogen demand of algae in the Neoproterozoic, implying that molybdenum was not a limiting resource in the Archean ocean. Furthermore, our data provide the first indirect evidence for a small source of nitrate to the early ocean, possibly as a result of lightning and volcanism or from localized oxygen oases. Some of this nitrate appears to have been used for nitrate reduction to ammonium (DNRA), which may have helped overcome nitrogen limitation by retaining fixed nitrogen in the system. DNRA rather than denitrification may explain some of the isotopic records of biological nitrate utilization in the Archean and early Proterozoic. We cannot confirm the hypothesized suppression of N2O-metabolizing enzymes in the mid-Proterozoic, but our results are consistent with vigorous nitrification and denitrification after the GOE, in particular from 1.5 Ga onwards, which may have led to leakage of N2O into the atmosphere and the stabilization of global climate. The proliferation of key nitrogen-metabolizing genes across the tree of life at different points in Earth’s redox history provides important insights into the evolution and radiation of microbial metabolisms on the early Earth in response to major environmental transitions and supports the notion that increasing nitrate levels in the mid-Proterozoic may have contributed to the rise of eukaryotic life.

ACKNOWLEDGEMENTS

We thank three anonymous reviewers for extremely detailed and helpful comments that greatly improved the manuscript. We would like to thank Rou-Jia Sung for assistance with structural predictions, as well as Casey Bryce, Colin Goldblatt, Ben Johnson, John Baross, Victoria Meadows, Jaclyn Saunders, and Cara Magnabosco for support and useful discussions. This work was performed as part of NASA’s Virtual Planetary Laboratory, supported by the NASA Astrobiology Program under grant 80NSSC18K0829 as part of the Nexus for Exoplanet System Science (NExSS) research coordination network. CWP received a Towsley Fellowship from Carleton College, CR was supported by the Virtual Planetary Laboratory, KM was supported by a Student Research Partnership grant from Carleton College, and REA was supported in part by a NASA Postdoctoral Fellowship from the NASA Astrobiology Institute.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in FigShare at the following DOIs: https://doi.org/10.6084/m9.figshare.12791747.v1 (list of genomes), https://doi.org/10.6084/m9.figshare.12791750.v1 (UGAM chronogram), https://doi.org/10.6084/m9.figshare.12791756.v1 (CIR chronogram), https://doi.org/10.6084/m9.figshare.12791759.v1 (gene alignments), https://doi.org/10.6084/m9.figshare.12791765.v1 (gene trees), https://doi.org/10.6084/m9.figshare.12791966.v1 (species tree and alignment).
Zhang, K., Zhu, X., Wood, R. A., Shi, Y., Gao, Z., & Poulton, S. W. (2018). Oxygenation of the Mesoproterozoic ocean and the evolution of complex eukaryotes. *Nature Geoscience*, 11, 345–350. https://doi.org/10.1038/s41561-018-0111-y

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Parsons C, Stüeken EE, Rosen CJ, Mateos K, Anderson RE. Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history. *Geobiology*. 2020;00:1–17. https://doi.org/10.1111/gbi.12419