LA-UR-18-21897 (Accepted Manuscript)

Comparative Genomics of Nitrogen Cycling Pathways in Bacteria and Archaea

Albright, Michaeline BN
Timalsina, Bibek
Martiny, Jennifer BH
Dunbar, John

Provided by the author(s) and the Los Alamos National Laboratory (2019-08-14).

To be published in: Microbial Ecology

DOI to publisher's version: 10.1007/s00248-018-1239-4

Permalink to record: http://permalink.lanl.gov/object/view?what=info:lanl-repo/lareport/LA-UR-18-21897

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by Triad National Security, LLC for the National Nuclear Security Administration of U.S. Department of Energy under contract 89233218CNA000001. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Comparative genomics of nitrogen cycling pathways in bacteria and archaea

Michaeline B.N. Albright[a#], Bibek Timalsina[a*], Jennifer B.H. Martiny[b], John Dunbar[a#]

Author affiliations:

[a]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM

[b]Department of Ecology and Evolutionary Biology, University of California, Irvine, CA

Address correspondence to:

Michaeline B.N. Albright

mbnalbright@lanl.gov

ORCID ID: 0000-0002-4528-0312

Los Alamos Natl. Lab

P.O. Box 1663

Bioscience Division, Mail Stop M888

Los Alamos, NM 87545

*Present Address: Bibek Timalsina, Department of Biochemistry, Mississippi State University, MS

Acknowledgements

We thank Cheryl Kuske and Renee Johansen for comments on previous versions of this manuscript. This work was supported by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Division, under award number F260LANL2018, and by an Office of Science Graduate Student Research (SCGSR) Fellowship to MA.

Abstract

Despite the explosion of metagenomic sequencing data, using -omics data to predict environmental biogeochemistry remains a challenge. One or a few genes (referred to as marker genes) in a metabolic pathway of interest in meta-omic data are typically used to represent the prevalence of a biogeochemical reaction. This approach often fails to demonstrate a consistent relationship between gene abundance and an ecosystem process rate. One reason this may occur is if a marker gene is not a good representative of a complete pathway. Here, we map the presence of eleven nitrogen (N)-cycling pathways in over 6000 complete bacterial and archaeal genomes using the Integrated Microbial Genomes database. Incomplete N-cycling pathways occurred in 39% of surveyed archaeal and bacterial species revealing a weakness in current marker-gene analyses. Furthermore, we found that most organisms have limited ability to utilize inorganic N in multiple oxidation states. This suggests that inter-organism exchange of inorganic N compounds is common, highlighting the importance of both community composition and spatial structure in determining the extent of recycling versus loss in an ecosystem.

Keywords: nitrogen cycle, comparative genomics, bacteria, archaea

Introduction

46 The genomic revolution created the tantalizing possibility that the metabolic pathways
47 encoded in microbial genomes can be used to predict processes at the ecosystem and global-
48 scale. The basic rationale is that the abundance and/or genetic diversity of a relevant metabolic
49 pathway(s) in a microbial community reflects the magnitude or state of an ecosystem process.
50 The conventional approach to link a metabolic pathway to an ecosystem process is to measure
51 the abundance of a specific marker gene (or corresponding mRNA) as a proxy for a metabolic
52 function. For example, the *nifH* gene is often used to infer nitrogen fixation. The marker gene
53 approach often fails to demonstrate a positive relationship between the abundance of a gene of
54 interest and the corresponding ecosystem process rate [1-3]. Inaccurate inference of a
55 functioning pathway based on marker gene presence may contribute to failure.
56 While enzymatic reactions can be carried out by one gene, they often rely on a cohort of
57 genes (modules) that function together within a single organism to transform a reactant to a final
58 product. When multiple genes are required for a reaction to occur, there is a potential for some
59 organisms to harbor incomplete metabolic pathways. For example, numerous organisms contain
60 *nifH,* but lack other genes in the nitrogen fixation pathway and thus cannot perform this function
61 [4]. Furthermore, there are multiple forms of *nifH* and many have not been tested or verified to
62 be involved in this function [5, 6]. One solution to the problem of incomplete pathways is to use
63 more than one marker gene when assessing the presence of a pathway in an environmental
64 sample [4]. A second solution is to use every gene in a pathway. Summing or averaging the
65 abundance of all relevant genes in a pathway can potentially reduce the impact of false positives
66 [7, 8]. An additional refinement may be to use a type weighted analysis of relevant genes to
67 account for the uncertainty of pathway completeness, especially if the likelihood of false
68 positives varies among genes, pathways, or taxa. This is analogous to the routine approach of

69 providing confidence scores when assigning taxonomy to sequence data [9, 10]. Weighted
70 analysis of pathways requires *a priori* quantification of the distribution of genes and pathways of
71 interest among many organisms. Although this approach of weighting functional genes has not
72 been implemented, it now may be feasible due to the accumulation of sequenced genomes in
73 public databases.

74 Microbial nitrogen cycling pathways are of particular interest in ecosystem studies
75 because nitrogen cycling is central to ecosystem performance. The potential for nitrogen (N)
76 cycling in the environment is commonly inferred through marker gene approaches [11-13].
77 Microbes are primarily responsible for many of the N compound transformations [14, 15]. N
78 occurs in many oxidation states, and organisms can use N in assimilatory (biomass producing) or
79 dissimilatory (energy producing) processes. Genetic markers for N cycling pathways have been
80 derived from studies of N transformations and genetic regulation in model organisms [16, 17].
81 New sequencing technologies and advances in isolation techniques have revealed the complexity
82 and horizontal mobility of some N-cycling genes across the phylogenetic tree [18]. While N is
83 found in many oxidation states connected through a number of reactions (N cycle pathways),
84 individual N cycle pathways are most commonly studied in isolation. Indeed, the distribution of
85 multiple N cycle pathways across bacteria and archaea has not been assessed. The increasing
86 number of sequenced genomes and their standardized annotation provide an opportunity to
87 investigate gene presence and abundance across thousands of organisms [19].

88 In this study, the distribution of eleven nitrogen metabolism (N-cycling) pathways (Fig.
89 1) was quantified among over 6000 complete prokaryotic genomes. Nitrogen metabolism (N-
90 cycling) genes based on KEGG database annotations [20] were obtained from 6384 bacterial and
91 252 archaeal finished genomes in the Integrated Microbial genomes (IMG) database [21]. The

92 eleven pathways included the seven canonical N-cycling pathways—nitrification, dissimilatory
93 nitrate reduction, denitrification, dissimilatory nitrite reduction, assimilatory nitrate reduction,
94 assimilatory nitrite reduction and nitrogen fixation (Fig. 1). Four additional pathways were
95 defined as partial versions of the canonical, multi-step denitrification and nitrification pathways
96 because it is conceivable for an organism to perform only segments of each canonical pathway
97 [22, 23]. For denitrification, partial pathways were classified as partial NO (NO_2^- to NO), partial
98 N_2O (NO to N_2O), or partial N_2 (N_2O to N_2) if organisms contained all genes to complete one of
99 the three steps. Furthermore, for nitrification, the ammonium to hydroxylamine reaction was
100 classified as a partial process (NH_4^+ to H_2NOH). Lastly, the nitrate/nitrite conversion step was
101 included as a separate reaction instead of within the complete denitrification or nitrification
102 pathways (Fig. 1). It is important to note that some of these pathways have reactions that can
103 potentially occur in both directions. Furthermore, although "pathway" implies a multi-step
104 process, some of the N-cycling pathways are single-step transformations in which an inorganic
105 nitrogen substrate is converted to a product by a multi-domain enzyme, with separate genes
106 encoding the enzymatic domains (e.g. nirBD, nitrite-reductase small and large subunits).
107 For each of the eleven N-cycling pathways, we quantified the percentage of genomes
108 with complete, incomplete, or absent pathways. A pathway was classified as "complete" in a
109 genome if it contained at least one copy of every gene in the pathway (as defined by KEGG N
110 metabolism pathways)[20]. An incomplete pathway could occur if for example, a gene in a
111 pathway was lost or the gene present in the incomplete pathway is used for alternate metabolic
112 functions. Alternatively, a pathway may be incomplete because of a failure to annotate one or
113 more genes (this is a false positive incomplete pathway). We hypothesized that: 1) incomplete
114 versions of multi-gene N-cycle pathways are common; 2) the likelihood of incomplete N cycle

115 pathways depends on both the pathway and phylogeny, and 3) N cycle pathways will frequently
116 co-occur in organisms, because capture of growth-limiting N in multiple forms is conceptually
117 advantageous.
118
119 **Materials and Methods**
120 *Nitrogen cycling genes in microbial genomes*
121 Nitrogen metabolism (N-cycling) genes based on KEGG database annotations [20] were
122 obtained from 6384 bacterial and 252 archaeal finished genomes in the Integrated Microbial
123 genomes (IMG) database [21] in June 2017. While this set of taxa is biased towards commonly
124 studied organisms, it is representative of our current knowledge of microorganisms and a
125 database that is commonly used to annotate genomic data from environmental samples. IMG is
126 highly curated and has been optimized for genome annotation, for example with tools to identify
127 pseudogenes. Metadata including genome size and taxonomy for each genome was also
128 downloaded.
129
130 *Annotation quality check*
131 To assess potential for false negatives (missing genes) in incomplete pathways, we used
132 the IMG chromosome viewer to spot-check a subset of genomes for annotation errors.
133 Specifically, we attempted to find one or more examples for each of the seven canonical N cycle
134 pathways in which a missing pathway gene was actually present in a genome but had not been
135 annotated as an open reading frame. First, we examined the gene neighborhood(s) of complete
136 pathways in model organism genomes to confirm that pathway genes were clustered and that
137 gene arrangement was generally conserved among genomes. Then, we visually inspected the top

138 15-25 related organism gene regions, confirming conservation of gene organization. Second, we
139 performed the same procedure to examine the gene neighborhood of an organism with an
140 incomplete pathway, again inspecting the top 15-25 related regions. With this procedure, we
141 screened about 5% (320) of the genomes harboring an incomplete pathway. We did not identify
142 any false negatives through this process, however, we acknowledge that false negatives may
143 exist within the more extensive dataset.
144
145 *Phylogenetic distribution of N cycle pathways*
146 The phylogenetic distribution of potential nitrogen cycling in bacteria and archaea was
147 visualized on a precomputed phylogenetic tree of isolate genomes downloaded from IMG on
148 October 29, 2017. The abundance of ammonia assimilation genes and inorganic nitrogen
149 pathways were plotted on the tree using iTOL v3 [24]. The percentage of genomes with
150 complete, incomplete, or not present pathways was then calculated at the phyla level. Next, we
151 aggregated genomes into phyla and for any phyla with ≥ 10 genomes we assessed the relative
152 abundance of each pathway within each phyla.
153
154 *Co-occurrence of N pathways*
155 N occurs in many different forms, and thus organisms can potentially catalyze multiple
156 reactions between N oxidation states. Accordingly, organisms have the potential to vary widely
157 in their metabolic range. We mapped the distribution of multiple N pathway occurrences across
158 the genomes. Pearson's correlations were used to determine which, if any, of the N pathways
159 were commonly found co-occurring within the same organisms.
160

Results

Prevalence of complete and incomplete N cycle pathways

Our data set included 6636 genomes representing 2481 unique species. The most abundant complete pathways among genomes were ammonia assimilation, dissimilatory nitrate reduction, dissimilatory nitrite reduction, and denitrification (Table 1). The least abundant pathways were ammonium to hydroxylamine and nitrification (Table 1), likely due to poor representation in the genomes dataset. Many genomes did not contain any inorganic nitrogen cycling pathways (Online Resource, Fig. S1). The likelihood of a pathway being incomplete ranged from 0 to 20% and was highest for assimilatory nitrate reduction. Of the species that contained at least one incomplete pathway, this was manifested in three ways as follows: every strain of a species contained the same incomplete pathway (92.2%), only some strains had an incomplete pathway (7%), every strain had an incomplete pathway, but the pathways differed (0.8%). Thus, in almost all cases, every strain of a species contained the same incomplete pathway, suggesting that incompletes are not a result of false negatives.

Comparing across the N cycling pathways, across the surveyed genomes, the eleven individual pathways varied in the likelihood of being incomplete versus complete (Online Resource, Fig. S1). When at least one gene within a pathway was present (i.e. excluding non-present pathways, which were most common across genomes), complete pathways were abundantly present in a number of pathways, including dissimilatory nitrate reduction (97%), dissimilatory nitrite reduction (88%) and nitrogen fixation (82%) were often complete. By contrast, incomplete pathways were often identified for assimilatory nitrate reduction (84%) and denitrification (48%). In the denitrification pathway, it was more common for organisms to contain a partial pathway (75%) than the complete pathway (25%).

184

185 *Phylogenetic distribution of complete and incomplete pathways*

186 As expected, the ammonia assimilation pathway (including glutamine synthetase,
187 glutamate synthase, glutamate dehydrogenase genes) was nearly ubiquitous (93% of genomes).
188 The species without genes in the ammonia assimilation pathway were predominantly in two
189 genera: *Mycoplasma* and *Chlamydia*, known to often live as obligate parasites [25]. The
190 abundance of ammonia assimilation genes within genomes varied, but on average each genome
191 contained four and a half ammonia assimilation genes (Online Resource, Fig. S2). *Pseudomonas*
192 *putida* (Gammaproteobacteria) and *Clostridium caroxidivorans* (Firmicutes) contained the most
193 genes (n=23) in the ammonia assimilation pathway. In the Actinobacteria phylum, eighteen taxa
194 contained greater than sixteen genes in this pathway. In soil environments, Actinobacteria
195 abundance has been shown to decline with increasing N deposition [26, 27]. The high abundance
196 of ammonia assimilation genes in these organisms may make them more competitive in low N
197 environments. The number of ammonia assimilation genes was correlated with genome size
198 (R^2=0.5, $p<0.001$, Online Resource, Fig. S3a), however, the number of complete inorganic
199 nitrogen cycling pathways was highest in organisms with intermediate genome sizes (Online
200 Resource, Fig. S3b).

201 In contrast to the common ammonia assimilation pathway, other inorganic nitrogen
202 cycling pathways were completely absent in approximately one third (36%) of the genomes
203 surveyed (Fig. 3, Online resource, Fig. S1, Fig. S4). The nitrogen cycling pathways that were
204 present, were broadly distributed across the phylogenetic tree (Fig. 3, Online Resource, Fig. S4).
205 The percentage of genomes with absent or incomplete pathways varied among phyla (Fig. 3).
206 The denitrification pathway was absent or incomplete in the Actinobacteria, Cyanobacteria,

207 Firmicutes, Planctomycetes, and Chrenarchaeota. In contrast, the pathway was mostly complete
208 in Alphaproteobacteria (91%, n=102 genomes total). The dissimilatory nitrite reduction pathway
209 was absent or incomplete in all Archaea, Chloroflexi, and Fusobacteria. In contrast, when
210 dissimilatory nitrite reduction genes were present in Proteobacteria, the pathway was complete
211 87% (n=1991) of the time. Nitrogen fixation was more often incomplete than complete in two
212 phyla, Fusobacteria (36% incomplete: 4% complete) and Aquificiae (13% incomplete: 7%
213 complete).
214
215 *Co-occurrence of N pathways*
216 Of the organisms that contained at least one N-cycle pathway, 79% of genomes had the
217 genetic machinery for one or two N cycling pathways (excluding ammonia assimilation which is
218 common to most organisms), while the remaining 21% of genomes had the potential to transform
219 N through three or more N pathways (Fig. 4). Furthermore, only two pathways, dissimilatory
220 nitrate reduction and dissimilatory nitrite reduction were correlated within genomes (R^2=0.6,
221 $p<0.001$). Assimilatory nitrite reduction and nitrogen fixation, the two pathways (other than
222 ammonia assimilation) that lead to the integration of inorganic N into biomass occurred in only
223 13% of genomes. Assimilatory nitrite reduction was more commonly found alone (51% of the
224 time) or coupled with dissimilatory nitrate reduction (33% of the time), than with assimilatory
225 nitrate reduction (16% of the time).
226 Taxa with the most inorganic nitrogen cycling pathways present were predominantly in
227 the Rhizobiales order. Out of the eleven complete and partial pathways only two organisms had
228 six pathways present including *Bradyrhizobium icense* (Rhizobiales) and *Methylocella silvestris*
229 (Rhizobiales). Moving to a broader level of phylogenetic classification, the number of complete

230 pathways in each phylum varied from zero (Chlamydiae, Tenericutes or Thermotogae) to eleven
231 pathways (Gammaproteobacteria). The relative abundance of each pathway also differed among
232 phylum (Fig. 5). For example, the Gammaproteobacteria had eleven pathways, but the
233 dissimilatory nitrate reduction and dissimilatory nitrite reduction pathways were the most
234 common (>90%). In contrast, Deinococcus-Thermus had five pathways with a more even
235 distribution.

236 Additionally, a number of organisms contained multiple copies of a single pathway
237 (Online Resource, Fig. S4, Table S1). The N fixation pathway had the largest variation in copy
238 number, ranging from 0 to 8, and showed substantial variation within genera. In particular,
239 strains of *Rhodopseudomonas palustris* (n=7), *Clostridium* (n= 68), *Paenibacillus* (n=43), and
240 *Methanosarcina* (n=23) contained a wide range of copies of the N fixation pathway (1-8, 0-8, 0-
241 7, and 0-6 respectively).

242

243 **Discussion**

244 Attempts to link microbial communities to biogeochemical processes often use marker
245 genes to represent relevant biochemical pathways. Our large-scale comparative genomic
246 analysis of eleven bacterial and archaeal N-cycling pathways suggests that these approaches may
247 not be robust, because the eleven pathways are often incomplete in microbial genomes (Fig. 2,
248 Fig. 3). We found incomplete N-cycling pathways in ~39% of 2481 species surveyed, consistent
249 with our hypothesis that substantial false positives can occur when the presence of a single gene
250 is interpreted as a fully functional pathway. This is likely a conservative estimate because our
251 analysis did not include regulatory genes or other auxiliary genes that may be required for the
252 expression and effective integration of a pathway into cellular metabolism.

253 The prevalence of incomplete pathways in sequenced genomes varied among individual
254 N cycle pathways, consistent with our second hypothesis (Fig. 2, Fig 3). At one end of the
255 spectrum, the nitrification pathway (conversion of NH_4 to H_2NOH to NO_2^-) was complete in all
256 organisms surveyed. Nitrification occurs in few organisms and is phylogenetically constricted,
257 suggesting lack of horizontal gene transfer. In contrast, all other multi-gene N pathways
258 contained some incomplete pathways. In particular, assimilatory nitrate reduction (20%),
259 dissimilatory nitrite reduction (7%) and denitrification (4%). A number of factors could drive
260 differences in the occurrence of incomplete pathways. Pathways that have a greater number of
261 genes or non-clustered genes (weak genetic linkage) may be more likely to be incomplete [28,
262 29]. For example, the assimilatory nitrate reduction pathway genes are non-clustered in some
263 organisms [30]. An incomplete assimilatory nitrate reduction pathway was found in 1339 of
264 6384 genomes. Another factor that may lead to incomplete pathways is gene recruitment for
265 alternate metabolic functions [31]. Incomplete denitrification may represent this phenomenon.
266 For denitrification, incomplete pathways were driven primarily by the absence of a gene in the
267 cytochrome bc complex (NorBC). Two types of bacterial nitric oxide reductases have been
268 characterized (cNOR) and (qNOR). The latter lacks cytochrome c and is hypothesized to be used
269 primarily for detoxification (an alternate function), not denitrification [32, 33].
270 The prevalence of individual pathways also varied among phyla (Fig. 3). For example,
271 the denitrification pathway was absent or was incomplete in a number of phyla, including
272 Actinobacteria and Chrenarchaeota, but was predominantly complete in Alphaproteobacteria.
273 This type of finding could be routinely applied in microbial ecology/environmental genomics to
274 improve taxonomy-to-function prediction tools. For instance, community taxonomic profiles
275 based on 16S rRNA sequence data are the most common type of microbial community data

276 generated, and profiles exist for more than 25,000 community samples collected globally [34].
277 Taxonomic profiles can be used to infer the functional architecture of a community if the
278 probabilities of metabolic pathways being present in phylogenetic groups are known. Inference
279 tools like PICRUST [35] and Tax4Fun [36]) map 16S rRNA gene profiles from environmental
280 samples to the most closely related sequenced genomes to infer function. A recent comparison of
281 inferred (PICRUST) versus observed (metagenomic functional sequence counts) functions in soil
282 samples showed vastly different estimates of the abundance of N cycle process genes [37]. Such
283 discrepancies may arise if N cycle gene presence in a reference genome does not accurately
284 represent organisms in an environmental sample. In such cases, interpretation of environmental
285 sequence data could be strengthened by weighting a functional inference based on the potential
286 for false positive as a function of gene identity and taxonomic/phylogenetic origin. For example,
287 if currently a study relies on functional gene abundance as an indicator of the magnitude of a
288 process, a weighted analysis could either reduce the observed abundance or amplify the error
289 estimate associated with the abundance to reflect the uncertainty.
290 Although incomplete or absent pathways may often indicate absence of a metabolic
291 function, it is possible that the function exists but is encoded by novel, unidentified genes.
292 Absence of function is the most likely explanation for Mycoplasma (Tenericutes), which
293 contained neither inorganic N pathways, nor ammonia assimilation genes. Mycoplasma, often a
294 parasitic organism, has a greatly reduced genome, and is known for its loss of genes coding for
295 ammonia assimilation [25]. However, an incomplete or absent pathway may indicate gaps in our
296 knowledge of the genes involved in N cycle pathways, as indicated by ongoing discoveries of
297 new N-cycle pathways and genes [38-40]. Based on the phylogenetic prevalence of incomplete
298 pathways identified in this study, we propose a "Most Wanted" list (Table 2), providing a

299 starting point to explore the possible existence of cryptic pathways and to test the impact of
300 weighted pathway analyses.

301 The diversity of complete N cycling pathways within individual organisms provides
302 insight into the likelihood for inter-organism transfer of inorganic nitrogen compounds. At one
303 extreme, every N-cycle pathway could conceivably occur in an organism, enabling the organism
304 to exploit and sequester any available form of inorganic nitrogen for production of biomass or
305 energy. At the other extreme, a single pathway can occur in an organism, providing limited
306 access to inorganic nitrogen and increasingly the likelihood of shuttling nitrogenous compounds
307 among organisms. Contrary to our third hypothesis, we found that co-occurrence of different N
308 cycle pathways in individual organisms was infrequent; instead limited access to inorganic
309 nitrogen appears typical. Most organisms that use inorganic N were specialists, harboring only
310 one or two pathways (Fig. 4). Furthermore, the decoupling of modular pathways—that is
311 pathways with more than one enzymatic step—was common. For denitrification, 75% of the
312 organisms contained only a partial pathway (i.e., only one step was present in the transformation
313 of NO_2- to NO to N_2O to N_2), rather than the whole pathway (NO_2- to N_2) (Table 1). These data
314 suggest extensive shuttling of inorganic N among organisms that use the compounds
315 predominantly for energy production, which creates the possibility for a leaky N cycle in natural
316 ecosystems. If tight spatial coupling of organisms does not occur, inefficient N transfers may
317 result in N loss from the immediate ecosystem. This suggests a direct link between N loss and
318 ecosystem stability. Disturbances that disrupt spatial structure likely enhance N losses from an
319 ecosystem [41, 42].

320 Inter-organism transfers of metabolic products may be a general phenomenon in
321 microbial communities, not limited to the N cycle. A study of subsurface microbial communities

322 found that few microorganisms were capable of conducting three or more sequential redox
323 transformations [43]. Other studies have used species co-occurrence patterns and modeling
324 approaches to highlight cooperative metabolic interactions in microbial communities [44, 45].
325 These observations are consistent with our comparative genomic study. There are few organisms
326 that can single-handedly process N in multiple oxidation states, thus use of N in multiple
327 oxidation states requires transfers away from organisms. The independent lines of evidence,
328 evidence from this study and others mentioned above, suggest that metabolic "specialism" in
329 microbial communities is common.

330 Sequenced genomes have the potential both to improve interpretation of complex
331 environmental sequence data [19, 21, 46] and to provide fundamental insights into the
332 organization and evolution of biological systems [47, 48]. Here, using a large-scale multi-
333 pathway comparative genomic analysis, we demonstrate the potential for incomplete pathways to
334 distort traditional marker-gene approaches to quantify N-cycling potential. Furthermore, the
335 preponderance of genomes without any inorganic N cycle genes highlights possible targets for
336 discovery of cryptic genes and pathways. The concentration of organic N often exceeds that of
337 inorganic N in environments such as soils [49, 50]. However, when studying N in ecosystems,
338 measurement techniques rely on inorganic N processes, largely due to the lack of tools for
339 assessing organic N transformations [51]. Omics studies offer one promising approach to focus
340 on the interactions between the inorganic and organic N cycle, and the coupling of N and C
341 cycles. Successful application will require further work in defining and categorizing genes
342 involved in organic N transformations. Overall, given that N cycling can profoundly impact
343 ecosystem behavior and meta-omics techniques are increasingly being used to measure N
344 transformations, accurate assessment of N-cycling potential is critical.

345

346 **Data availability**

347 All data used in this study is publicly available through the Integrated Microbial Genomes (IMG)

348 database.

349

350 **Conflict of interest**

351 The authors declare no conflict of interest.

352

353 **References**

354

355 1. Di HJ, Cameron KC, Podolyan A, Robinson A (2014) Effect of soil moisture status and a

356 nitrification inhibitor, dicyandiamide, on ammonia oxidizer and denitrifier growth and

357 nitrous oxide emissions in a grassland soil. Soil Biol Biochem 73: 59-68. doi:

358 10.1016/j.soilbio.2014.02.011

359 2. Morales SE, Jha N, Saggar S (2015) Biogeography and biophysicochemical traits link

360 N2O emissions, N2O emission potential and microbial communities across New Zealand

361 pasture soils. Soil Biol Biochem 82: 87-98. doi: 10.1016/j.soilbio.2014.12.018

362 3. Rocca JD, Hall EK, Lennon JT, Evans SE, Waldrop MP, Cotner JB, Nemergut DR,

363 Graham EB, Wallenstein MD (2015) Relationships between protein-encoding gene

364 abundance and corresponding process are commonly assumed yet rarely observed. Isme J

365 9: 1693-1699. doi: 10.1038/ismej.2014.252

366 4. Dos Santos PC, Fang Z, Mason SW, Setubal JC, Dixon R (2012) Distribution of nitrogen
367 fixation and nitrogenase-like sequences amongst microbial genomes. BMC Genomics 13:
368 162. doi: 10.1186/1471-2164-13-162
369 5. Zehr JP, Jenkins BD, Short SM, Steward GF (2003) Nitrogenase gene diversity and
370 microbial community structure: a cross-system comparison. Environ Microbiol 5: 539-
371 554. doi: 10.1046/j.1462-2920.2003.00451.x
372 6. Raymond J, Siefert JL, Staples CR, Blankenship RE (2004) The natural history of
373 nitrogen fixation. Mol Biol Evol 21: 541-554. doi: 10.1093/molbev.msh047
374 7. Nelson MB, Berlemont R, Martiny AC, Martiny JBH (2015) Nitrogen Cycling Potential
375 of a Grassland Litter Microbial Community. Appl Environ Microb 81: 7012-7022. doi:
376 10.1128/Aem.02222-15
377 8. Nelson MB, Martiny AC, Martiny JBH (2016) Global biogeography of microbial
378 nitrogen-cycling traits in soil. P Natl Acad Sci USA 113: 8033-8040. doi:
379 10.1073/pnas.1601070113
380 9. Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases,
381 parallelization and more. Nat Methods 8: 367-367. doi: DOI 10.1038/nmeth0511-367
382 10. Cole JR, Chai B, Farris RJ, Wang Q, Kulam SA, McGarrell DM, Garrity GM, Tiedje JM
383 (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-
384 throughput rRNA analysis. Nucleic Acids Res 33: D294-D296. doi: 10.1093/nar/gki038
385 11. Mason OU, Scott NM, Gonzalez A, Robbins-Pianka A, Baelum J, Kimbrel J, Bouskill
386 NJ, Prestat E, Borglin S, Joyner DC, Fortney JL, Jurelevicius D, Stringfellow WT,
387 Alvarez-Cohen L, Hazen TC, Knight R, Gilbert JA, Jansson JK (2014) Metagenomics

388 reveals sediment microbial community response to Deepwater Horizon oil spill. Isme J 8:
389 1464-1475. doi: 10.1038/ismej.2013.254
390 12. Morales SE, Cosart T, Holben WE (2010) Bacterial gene abundances as indicators of
391 greenhouse gas emission in soils. Isme J 4: 799-808. doi: 10.1038/ismej.2010.8
392 13. Tu QC, He ZL, Wu LY, Xue K, Xie G, Chain P, Reich PB, Hobbie SE, Zhou JZ (2017)
393 Metagenomic reconstruction of nitrogen cycling pathways in a CO2- enriched grassland
394 ecosystem. Soil Biol Biochem 106: 99-108. doi: 10.1016/j.soilbio.2016.12.017
395 14. Galloway JN, Townsend AR, Erisman JW, Bekunda M, Cai ZC, Freney JR, Martinelli
396 LA, Seitzinger SP, Sutton MA (2008) Transformation of the nitrogen cycle: Recent
397 trends, questions, and potential solutions. Science 320: 889-892. doi:
398 10.1126/science.1136674
399 15. Vitousek PM, Aber JD, Howarth RW, Likens GE, Matson PA, Schindler DW,
400 Schlesinger WH, Tilman D (1997) Human alteration of the global nitrogen cycle:
401 Sources and consequences. Ecol Appl 7: 737-750. doi: Doi 10.2307/2269431
402 16. Geisseler D, Horwath WR, Joergensen RG, Ludwig B (2010) Pathways of nitrogen
403 utilization by soil microorganisms - A review. Soil Biol Biochem 42: 2058-2067. doi:
404 10.1016/j.soilbio.2010.08.021
405 17. Zehr JP, Ward BB (2002) Nitrogen cycling in the ocean: New perspectives on processes
406 and paradigms. Appl Environ Microb 68: 1015-1024. doi: 10.1128/Aem.68.3.1015-
407 1024.2002
408 18. Isobe K, Ohte N (2014) Ecological Perspectives on Microbes Involved in N-Cycling.
409 Microbes Environ 29: 4-16. doi: 10.1264/jsme2.ME13159

410 19. Berlemont R, Martiny AC (2015) Genomic Potential for Polysaccharide Deconstruction
411 in Bacteria. Appl Environ Microb 81: 1513-1519. doi: 10.1128/Aem.03718-14
412 20. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M (2016) KEGG as a
413 reference resource for gene and protein annotation. Nucleic Acids Res 44: D457-D462.
414 doi: 10.1093/nar/gkv1070
415 21. Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Pillay M, Ratner A, Huang
416 JH, Woyke T, Huntemann M, Anderson I, Billis K, Varghese N, Mavromatis K, Pati A,
417 Ivanova NN, Kyrpides NC (2014) IMG 4 version of the integrated microbial genomes
418 comparative analysis system. Nucleic Acids Res 42: D560-D567. doi:
419 10.1093/nar/gkt963
420 22. Carlson CA, Ingraham JL (1983) Comparison of Denitrification by Pseudomonas-
421 Stutzeri, Pseudomonas-Aeruginosa, and Paracoccus-Denitrificans. Appl Environ Microb
422 45: 1247-1253.
423 23. Ye RW, Thomas SM (2001) Microbial nitrogen cycles: physiology, genomics and
424 applications. Curr Opin Microbiol 4: 307-312.
425 24. Letunic I, Bork P (2016) Interactive tree of life (iTOL) v3: an online tool for the display
426 and annotation of phylogenetic and other trees. Nucleic Acids Res 44: W242-W245. doi:
427 10.1093/nar/gkw290
428 25. Amon J, Titgemeyer F, Burkovski A (2010) Common patterns - unique features: nitrogen
429 metabolism and regulation in Gram-positive bacteria. Fems Microbiol Rev 34: 588-605.
430 doi: 10.1111/j.1574-6976.2010.00216.x

431 26. Eisenlord SD, Zak DR (2010) Simulated Atmospheric Nitrogen Deposition Alters
432 Actinobacterial Community Composition in Forest Soils. Soil Sci Soc Am J 74: 1157-
433 1166. doi: 10.2136/sssaj2009.0240
434 27. Treseder KK (2008) Nitrogen additions and microbial biomass: a meta-analysis of
435 ecosystem studies. Ecol Lett 11: 1111-1120. doi: 10.1111/j.1461-0248.2008.01230.x
436 28. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of
437 transcriptional regulation in Escherichia coli. Genome Biol 9: R4. doi: 10.1186/gb-2008-
438 9-1-r4
439 29. Boto L (2010) Horizontal gene transfer in evolution: facts and challenges. Proc Biol Sci
440 277: 819-827. doi: 10.1098/rspb.2009.1679
441 30. Eberl L, Ammendola A, Rothballer MH, Givskov M, Sternberg C, Kilstrup M, Schleifer
442 KH, Molin S (2000) Inactivation of gltB abolishes expression of the assimilatory nitrate
443 reductase gene (nasB) in Pseudomonas putida KT2442. J Bacteriol 182: 3368-3376. doi:
444 Doi 10.1128/Jb.182.12.3368-3376.2000
445 31. Popescu L, Yona G (2005) Automation of gene assignments to metabolic pathways using
446 high-throughput expression data. Bmc Bioinformatics 6. doi: Artn 217
447 10.1186/1471-2105-6-217
448 32. Hendriks J, Oubrie A, Castresana J, Urbani A, Gemeinhardt S, Saraste M (2000) Nitric
449 oxide reductases in bacteria. Bba-Bioenergetics 1459: 266-273. doi: Doi 10.1016/S0005-
450 2728(00)00161-4
451 33. de Vries S, Schroder I (2002) Comparison between the nitric oxide reductase family and
452 its aerobic relatives, the cytochrome oxidases. Biochem Soc Trans 30: 662-667. doi:
453 10.1042/

454 34. Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, Prill RJ, Tripathi
455 A, Gibbons SM, Ackermann G, Navas-Molina JA, Janssen S, Kopylova E, Vazquez-
456 Baeza Y, Gonzalez A, Morton JT, Mirarab S, Zech Xu Z, Jiang L, Haroon MF, Kanbar J,
457 Zhu Q, Jin Song S, Kosciolek T, Bokulich NA, Lefler J, Brislawn CJ, Humphrey G,
458 Owens SM, Hampton-Marcell J, Berg-Lyons D, McKenzie V, Fierer N, Fuhrman JA,
459 Clauset A, Stevens RL, Shade A, Pollard KS, Goodwin KD, Jansson JK, Gilbert JA,
460 Knight R, Earth Microbiome Project C (2017) A communal catalogue reveals Earth's
461 multiscale microbial diversity. Nature 551: 457-463. doi: 10.1038/nature24621
462 35. Langille MGI, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente
463 JC, Burkepile DE, Thurber RLV, Knight R, Beiko RG, Huttenhower C (2013) Predictive
464 functional profiling of microbial communities using 16S rRNA marker gene sequences.
465 Nat Biotechnol 31: 814-+. doi: 10.1038/nbt.2676
466 36. Asshauer KP, Wemheuer B, Daniel R, Meinicke P (2015) Tax4Fun: predicting functional
467 profiles from metagenomic 16S rRNA data. Bioinformatics 31: 2882-2884. doi:
468 10.1093/bioinformatics/btv287
469 37. Hartman WH, Ye RZ, Horwath WR, Tringe SG (2017) A genomic perspective on
470 stoichiometric regulation of soil carbon cycling. Isme J 11: 2652-2665. doi:
471 10.1038/ismej.2017.115
472 38. Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA (2005)
473 Isolation of an autotrophic ammonia-oxidizing marine archaeon. Nature 437: 543-546.
474 doi: 10.1038/nature03911
475 39. Sanford RA, Wagner DD, Wu Q, Chee-Sanford JC, Thomas SH, Cruz-Garcia C,
476 Rodriguez G, Massol-Deya A, Krishnani KK, Ritalahti KM, Nissen S, Konstantinidis

477 KT, Loffler FE (2012) Unexpected nondenitrifier nitrous oxide reductase gene diversity
478 and abundance in soils. Proc Natl Acad Sci U S A 109: 19709-19714. doi:
479 10.1073/pnas.1211238109
480 40. van Kessel MA, Speth DR, Albertsen M, Nielsen PH, Op den Camp HJ, Kartal B, Jetten
481 MS, Lucker S (2015) Complete nitrification by a single microorganism. Nature 528: 555-
482 559. doi: 10.1038/nature16459
483 41. Cardenas E, Kranabetter JM, Hope G, Maas KR, Hallam S, Mohn WW (2015) Forest
484 harvesting reduces the soil metagenomic potential for biomass decomposition. Isme J 9:
485 2465-2476. doi: 10.1038/ismej.2015.57
486 42. Smith CR, Blair PL, Boyd C, Cody B, Hazel A, Hedrick A, Kathuria H, Khurana P,
487 Kramer B, Muterspaw K, Peck C, Sells E, Skinner J, Tegeler C, Wolfe Z (2016)
488 Microbial community responses to soil tillage and crop rotation in a corn/soybean
489 agroecosystem. Ecol Evol 6: 8075-8084. doi: 10.1002/ece3.2553
490 43. Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, Thomas BC,
491 Singh A, Wilkins MJ, Karaoz U, Brodie EL, Williams KH, Hubbard SS, Banfield JF
492 (2016) Thousands of microbial genomes shed light on interconnected biogeochemical
493 processes in an aquifer system. Nat Commun 7. doi: ARTN 13219
494 10.1038/ncomms13219
495 44. Embree M, Liu JK, Al-Bassam MM, Zengler K (2015) Networks of energetic and
496 metabolic interactions define dynamics in microbial communities. P Natl Acad Sci USA
497 112: 15450-15455. doi: 10.1073/pnas.1506034112

498 45. Zelezniak A, Andrejev S, Ponomarova O, Mende DR, Bork P, Patil KR (2015) Metabolic
499 dependencies drive species co-occurrence in diverse microbial communities (vol 112, pg
500 6449, 2015). P Natl Acad Sci USA 112: E7156-E7156. doi: 10.1073/pnas.1522642113
501 46. Chase AB, Karaoz U, Brodie EL, Gomez-Lunar Z, Martiny AC, Martiny JBH (2017)
502 Microdiversity of an Abundant Terrestrial Bacterium Encompasses Extensive Variation
503 in Ecologically Relevant Traits. Mbio 8. doi: ARTN e01809-17
504 10.1128/mBio.01809-17
505 47. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN,
506 Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM,
507 Amundson R, Thomas BC, Banfield JF (2016) A new view of the tree of life. Nat
508 Microbiol 1. doi: Artn 16048
509 10.1038/Nmicrobiol.2016.48
510 48. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on
511 our understanding of evolution. Cell 101: 573-576.
512 49. Bardgett RD, Streeter TC, Bol R (2003) Soil microbes compete effectively with plants
513 for organic-nitrogen inputs to temperate grasslands. Ecology 84: 1277-1287. doi: Doi
514 10.1890/0012-9658(2003)084[1277:Smcewp]2.0.Co;2
515 50. Schulten HR, Schnitzer M (1997) The chemistry of soil organic nitrogen: a review. Biol
516 Fert Soils 26: 1-15. doi: DOI 10.1007/s003740050335
517 51. Schimel JP, Bennett J (2004) Nitrogen mineralization: Challenges of a changing
518 paradigm. Ecology 85: 591-602. doi: Doi 10.1890/03-8002
519

520 **Figure Legends**

521

522 **Fig. 1** Nitrogen pathways explored in this study and associated genes mapped to these pathways
523 as defined by the KEGG database nitrogen metabolism module

524

525 **Fig. 2** Phylogenetic distribution of N pathways across all genomes. A precomputed phylogenetic
526 tree containing bacterial and archaeal genomes was downloaded from IMG. Taxonomy (phyla
527 level) is indicated by colors in inner circle. The black bars represent all genomes analyzed in this
528 study. Outer circle symbols correspond to individual N cycle pathways, where incomplete (I)
529 pathways are shown with open symbols and complete (C) pathways are shown with closed
530 circles.

531

532 **Fig. 3** Distribution across phyla of complete (blue), incomplete (green) or not present (white)
533 pathways in each of the eleven individual inorganic N cycle pathways

534

535 **Fig. 4** Histogram of pathway multiplicity within genomes. The bars show the number of
536 genomes containing zero, one, two, three, four, five, or six different complete inorganic N
537 cycling pathways.

538

539 **Fig. 5** The distribution of the genomes with inorganic N cycle pathways found in each phyla.
540 The total number of genomes in each phyla is shown by the number (n) next to the phylum. The
541 number of pathways found in each phyla is shown at the top of the panel

542