Supplementary Materials

Quantifying taxonomy-function associations across hierarchical scales of bacterial nitrogen cycling

Mingli Jiang[a], Yanni Huang[a], Zhiming Wu[e], Qian Zhu[a], Qian Li[a], Kaihua Pan[a], Mingliang Zhang[a], Liang Shi[d], Jiguo Qiu[a], Pengfa Li[c], Xin Yan[a], Yiyong Zhu[b], Qing Hong[a*]

[a] Department of Microbiology, College of Life Sciences, Nanjing Agricultural University, Key Laboratory of Agricultural and Environmental Microbiology, Ministry of Agriculture and Rural Affairs, Nanjing 210095, China.

[b] College of Resources and Environment Sciences, Nanjing Agricultural University, Nanjing 210095, China.

[c] College of Resources and Environment, Fujian Agriculture and Forestry University, Fuzhou 350002, China.

[d] College of Life Sciences, Nanjing Agricultural University, Nanjing 210095, China.

[e] Nanchang Key Laboratory of Microbial Resources Exploitation & Utilization from Poyang Lake Wetland, College of Life Sciences, Jiangxi Normal University, Nanchang 330022, China

Corresponding author: Qing Hong

E-mail address: hongqing@njau.edu.cn

This PDF file includes:

Supplementary Extended Methods

Supplementary EM1: Functional gene assignment criteria and exclusion rationale

Given our objective to quantify taxonomy-function associations, high-confidence functional assignments are essential to minimize false positives that would obscure genuine taxonomic patterns. Our dual-requirement approach (core genes plus auxiliary genes for most pathways) ensures detection of complete functional pathways rather than fragmentary gene presence.

Functional potential for six nitrogen transformation pathways was determined using pathway-specific core and auxiliary genes (Supplementary Table S1). Core genes included: nitrate reductase (*nasA*) and nitrite reductase (*nirB*) for assimilatory nitrate reduction (ANRA); periplasmic nitrate reductase (*napA*) or membrane-bound nitrate reductase (*narG*) for dissimilatory nitrate reduction to nitrite (DNN); nitrite reductases (*nirS* or *nirK*) for denitrification (DNF), with downstream genes (*cnorB*, *qnor*, *nosZ*) determining gaseous products (Graf et al., 2014); nitrogenase reductase (*nifH*) for nitrogen fixation (NF); ammonia monooxygenase (*amoA*) and hydroxylamine oxidoreductase (*hao*) for nitrification (NIT); cytochrome c nitrite reductase (*nrfA*) for dissimilatory nitrite reduction to ammonium (DNRA).

Alternative gene variants were excluded based on combined considerations of prevalence and annotation reliability. Nitrite oxidoreductase (*nxr*) subunits share substantial sequence similarity with nitrate reductase subunits, making reliable differentiation challenging in homology-based annotation pipelines. NCycDB database documentation explicitly notes that "*nxrAB* and *narGH* were difficult to distinguish from each other due to their close genetic relationships" (Tu et al., 2019). Additionally, *nxr* genes exhibit extremely narrow phylogenetic distribution, being restricted to nitrite-oxidizing bacterial lineages including Nitrospira, Nitrobacter, and related taxa (Daims et al., 2016; Lücker et al., 2010), combined with severe underrepresentation in current genomic databases. These annotation reliability concerns motivated *nxr* exclusion to prevent false positive inflation in nitrate

reduction pathway detection. Ferredoxin-nitrite reductase (*nirA*) was detected in 4.4% of genomes but predominantly co-occurred with the canonical ANRA pathway (*nasA* + *nirB*); among genomes containing both *nirA* and *nirB* (0.5%), 95.7% also contained *nasA*, indicating functional redundancy. Only 0.5% of genomes contained *nirA* without *nasA*, representing *nirA*-exclusive assimilatory reduction primarily in cyanobacteria with distinct evolutionary origin (Guerrero and Lara, 1987). Alternative nitrogenases (*vnfH*, *anfH*) exhibited very low prevalence (*vnfH*: 0.003%; *anfH*: 0.36%) compared to canonical molybdenum-dependent nitrogen fixation (*nifH*: 8.7%), representing specialized metal-limited systems rather than the widespread Mo-dependent pathway (Kuypers et al., 2018). These exclusions ensure that functional assignments reflect well-characterized pathways with adequate representation and minimal annotation ambiguity.

Supplementary EM2: Sensitivity analysis for pathway annotation strategies

To assess whether observed taxonomy-function associations reflect genuine biological patterns or artifacts of pathway definition choices, we tested three annotation strategies representing different stringency levels. All 73,472 genomes were re-annotated under each strategy and genus-level Information Gain was recalculated for all six nitrogen cycling pathways using identical computational approaches. Spearman rank correlation quantified pattern consistency across strategies.

Strict strategy employed annotation criteria detailed in Supplementary EM1, requiring all pathway-specific core genes plus at least one auxiliary gene for most pathways (Supplementary Table S1). Denitrification was assigned based solely on nitrite reductase (*nirS* or *nirK*) presence, as this core enzyme is sufficient to indicate denitrification capability, with downstream genes (*cnorB*, *qnor*, *nosZ*) determining terminal products rather than pathway presence. Moderate strategy required only core genes without auxiliary requirements: ANRA (*nasA* AND *nirB*), DNN (*napA* OR *narG*), DNF (*nirS* OR *nirK*), DNRA (*nrfA*), NF (*nifH*), NIT (*amoA* AND *hao*).

Relaxed strategy assigned pathway presence if any gene listed in Supplementary Table S1 for that pathway was detected.

Strict and Moderate strategies yielded highly correlated genus-level Information Gain values (Spearman $\rho = 0.943$, $P = 0.005$) with identical pathway rankings preserved. Relaxed strategy showed reduced correlation with Strict ($\rho = 0.657$, $P = 0.156$), attributable primarily to ANRA, whose positive genome proportion increased from 15.2% under Strict criteria to 21.5% under Relaxed criteria—a larger expansion than observed for any other pathway. The higher ANRA Information Gain under Relaxed criteria therefore indicates that Strict criteria provide conservative rather than inflated association estimates for ANRA. Pathway-specific results and correlation analyses are provided in Supplementary Figure S4.

Supplementary EM3: Environmental classification procedures and methodological considerations

Environmental sources were subdivided into 22 detailed subtypes based on refined metadata extraction. Within the Aquatic category, genomes were classified into Marine, Freshwater, Estuarine, and other aquatic subcategories. Within the Terrestrial category, genomes were classified into Agricultural soils, Forest soils, Grassland soils, and other terrestrial subtypes. The Other category comprised 15.1% of genomes with ambiguous environmental descriptors, which were retained in global analyses but excluded from environment-specific comparisons.

Supplementary EM4: Environmental modulation analysis

Environmental modulation was quantified using adaptation indices calculated as Euclidean distances between environment-specific and global functional profiles for classes present in multiple environments (≥10 genomes per environment). Function-environment associations were assessed using Fisher's exact tests with Benjamini-Hochberg correction (Benjamini and Hochberg, 1995).

Supplementary EM5: Phylogenetic reconstruction, signal analysis, and functional heterogeneity quantification

Genus-level phylogenetic trees were constructed using 16 ribosomal proteins (Parks et al., 2018). Representative genomes were selected from each genus based on highest completeness and lowest contamination. Ribosomal protein sequences were extracted using DIAMOND searches. Single-copy orthologs were identified and aligned using MAFFT (Katoh and Standley, 2013). Conserved regions were extracted using TrimAl with gap threshold 0.2 and conservation threshold 50% (Capella-Gutiérrez et al., 2009). Maximum-likelihood phylogenetic trees were constructed using IQ-TREE with automated model selection and 1000 bootstrap replicates (Nguyen et al., 2015).

Phylogenetic signal analysis employed Moran's I statistics. For binary ecological strategy variables, phylogenetic signal was calculated using Abouheif's proximity matrix method in adephylo (Jombart et al., 2010). For continuous variables, standard Moran's I was computed using phylogenetic distance matrices with row-standardized weights in ape (Paradis and Schliep, 2019). Statistical significance was assessed through permutation testing (n=1000) with false discovery rate correction.

Environmental associations were analyzed using Fisher's exact tests. Functional heterogeneity within genera was quantified using five metrics: Shannon Functional Combination Diversity (adapted from Shannon entropy (Shannon, 1948)), Coefficient of Variation of Functional Frequencies (CV_FFV), Functional Dominance, Combination Richness, and Functional Evenness. Shannon FCD was calculated as $H = -\Sigma(pi \times \ln(pi))$, where pi represents the proportion of genomes exhibiting each unique combination of six nitrogen cycling functions within a genus. Differences among ecological strategy types were assessed using Kruskal-Wallis rank-sum tests (Kruskal and Wallis, 1952) followed by pairwise Dunn's tests (Dunn, 1964) with Holm correction (Holm, 1979). Effect sizes were quantified using epsilon-squared (ε^2) (Tomczak and Tomczak, 2014).

Supplementary EM6: Molecular evolutionary analysis methods and quality control

For nucleotide diversity (π) calculations, gap positions and ambiguous nucleotides were excluded from calculations. Software versions used: MAFFT v7.475 (Katoh and Standley, 2013), boot package v1.3-28 (Canty and Ripley, 2017).

For dN/dS analyses, nucleotide sequences were aligned using MUSCLE v3.8.31 (Edgar, 2004), and alignments were refined using TrimAl v1.4 (Capella-Gutiérrez et al., 2009) with moderate filtering parameters (-gt 0.7 -st 0.01 -cons 60) to remove poorly aligned regions while preserving codon integrity. Codon alignment integrity was ensured by removing internal stop codons, trimming alignments to multiples of three nucleotides, and excluding alignment columns containing stop codons. Phylogenetic trees for each genus-level alignment were constructed using IQ-TREE v2.0.3 (Minh et al., 2020) under the GTR+G nucleotide substitution model.

The Single-Likelihood Ancestor Counting (SLAC) method implemented in HyPhy v2.5.29 (Kosakovsky Pond and Frost, 2005) employs maximum likelihood to estimate synonymous (dS) and non-synonymous (dN) substitution rates across the phylogeny. Median ω values with interquartile ranges were reported to summarize selection pressures across genera, as dN/dS distributions are typically right-skewed due to occasional lineages under relaxed or positive selection.

Supplementary EM7: Data visualization tools and parameters

Data visualization was performed using ggplot2 v3.4.0 (Gómez-Rubio, 2017) for scatter plots, bar charts, and box plots, with color palettes selected from viridis v0.6.2 (Garnier et al., 2021) to ensure accessibility for color-vision deficiencies. Heatmaps were generated using pheatmap v1.0.12 (Kolde, 2019), with hierarchical clustering (Euclidean distance, complete linkage) applied where appropriate. Correlation matrices were visualized using corrplot v0.92 (Wei et al., 2017). Phylogenetic trees with metadata annotations were created using ggtree v3.6.2 (Yu et al., 2017). Final figure assembly and annotation were performed in Adobe Illustrator 2024.

Supplementary Notes

Supplementary Note 1: Extended functional distribution data and genome source patterns analysis

Among functionally active genomes, single-pathway specialization dominated, with the five most abundant single-function strategies closely reflecting the overall functional hierarchy: DNN-only (6,646 genomes), ANRA-only (5,104), NF-only (3,470), DNF-only (2,571), and DNRA-only (2,495), collectively representing 60.4% of functionally active genomes (Figure S8). Multi-functional genomes exhibited exponential frequency decline (dual-function: 13.4%, triple-function: 4.1%), with higher-order combinations showing extreme rarity: quadruple-function genomes comprised only 431 genomes (0.6%), quintuple-function genomes 27 genomes (0.04%), and complete six-function genomes were virtually absent (Figure 1B). This distribution follows an exponential decay pattern that supports metabolic cost-benefit models where each additional pathway imposes increasing regulatory burdens (Lynch, 2007). Among dual-function combinations, the top five were DNF+DNN (2,920 genomes), ANRA+DNN (2,579), DNN+DNRA (1,853), ANRA+DNF (596), and DNN+NF (490). Triple-function combinations were dominated by ANRA+DNF+DNN (1,193 genomes, 39.4% of all triple-function genomes), indicating modular integration of assimilatory and respiratory nitrogen cycling strategies (Graf et al., 2014).

The unexpected ANRA-DNN pairing exhibits dual-level dominance in bacterial nitrogen cycling: both pathways show high individual prevalence and strong combinatorial association, establishing them as complementary metabolic strategies. Organisms possessing both ANRA and DNN can flexibly allocate nitrate between biosynthetic assimilation (supporting growth) and respiratory dissimilation (supporting energy generation), potentially providing competitive advantages in environments experiencing variable nitrogen availability and oxygen tension. In contrast, DNN-DNRA represents a committed respiratory strategy that channels nitrite exclusively toward ammonium production rather than allowing biosynthetic

flexibility. Recent studies have demonstrated that in facultative anaerobes, assimilatory and dissimilatory nitrate reduction pathways can coexist with coordinated expression even under aerobic conditions, allowing flexible allocation of nitrate between biosynthetic and respiratory functions depending on nitrogen availability (Ahn et al., 2025).

Major environmental categories showed distinct quantitative associations with nitrogen cycling functions (Figure 1D, S8A). Aquatic isolates contributed 32.1% of DNF genomes and 29.2% of DNN genomes despite comprising only 27.0% of total genomes, consistent with the enrichment of anaerobic respiratory processes in oxygen-depleted aquatic environments (Seitzinger et al., 2006). Terrestrial isolates exhibited pronounced enrichment for assimilatory processes, contributing 27.9% of ANRA genomes while representing just 14.1% of the dataset, consistent with nitrogen limitation in soil ecosystems where assimilatory nitrogen uptake is essential for microbial growth (Schimel and Bennett, 2004). Host-associated sources displayed a distinctive functional signature: despite having the highest proportion of genomes lacking detectable nitrogen cycling functions (72.4%), they exhibited the strongest enrichment for nitrogen fixation among functionally active genomes, with NF representing 6.4% of host-associated genomes—the highest proportion across all environmental categories.

Fine-scale analysis across environmental subtypes revealed additional functional specialization patterns (Figure S9B). Among aquatic sources, marine environments showed stronger denitrification enrichment (DNF: 34.2% of genomes vs 27.0% expected) than freshwater environments (DNF: 29.8% vs 27.0% expected), potentially reflecting differences in organic carbon availability and oxygen depletion patterns. Within terrestrial environments, forest soils showed the highest ANRA enrichment (31.4% of ANRA genomes while representing 8.2% of the dataset) compared to agricultural soils (24.1% vs 6.3% expected), consistent with different nitrogen availability regimes in natural versus managed systems (Templer et al., 2012). Plant-associated genomes within terrestrial sources demonstrated intermediate nitrogen

fixation patterns with rhizosphere isolates (7.8% NF) exceeding phyllosphere isolates (3.1%). Host-associated environments revealed heterogeneous functional patterns across host types: ruminant-associated genomes showed higher nitrogen fixation frequencies (8.9% of genomes) than monogastric-associated genomes (4.2%). These associations are interpreted as descriptive characterizations of functional enrichment patterns across genome source categories. Observed patterns are consistent with established biogeochemical principles, providing contextual support for the genome annotation quality and environmental classification scheme used in this study. Statistical analysis across 22 environmental subtypes with ≥100 genomes each revealed 82 significant function-environment associations among 132 tested combinations (62.1%, $q < 0.05$, Benjamini-Hochberg FDR correction); these patterns document the co-distribution of bacterial functional potential and environmental context within the current database, without implying causal environmental control of functional organization.

Supplementary Note 2: Extended class-level functional archetype analysis

Bootstrap resampling (n = 1,000) yielded mean ARI = 0.553 ± 0.265; the wide 95% CI (0.003–1.000) is attributable to stochastic variation across numerically unequal clusters (n = 25, 34, 17), with the dominant cluster structure remaining reproducible, as indicated by 60.5% of resamples exceeding ARI = 0.5. Kruskal–Wallis tests across the three multi-class archetypes confirmed significant inter-archetype differences for all functional metrics (Supplementary Table S4), with η^2 ranging from 0.242 (ANRA) to 0.789 (participation rate). Cluster number validation and individual class silhouette values are shown in Figure S3 and Figure S11, respectively; archetype mean silhouettes were: Functionally Inactive, 0.625; Functionally Moderate, 0.371; N-Retention Dominant, 0.155; Nitrification Specialist, 0 (by convention for single-member clusters in the R cluster package).

Among Functionally Inactive classes, Bacilli_A (948 genomes, PR = 0.040) showed only residual DNRA (3.4%), while Coriobacteriia (945 genomes, PR = 0.099) retained limited DNN (5.8%) and DNRA (4.9%), consistent with its host-associated

niche. Acidimicrobiia (203 genomes, PR = 0.074) showed modest DNN (6.4%). At the archetype boundary, Fusobacteriia (92 genomes, PR = 0.141) and Rhodothermia (50 genomes, PR = 0.140) approach but do not cross the Jenks threshold. Thirteen classes showed PR = 0 (including Vampirovibrionia, Chlamydiia, Brachyspirae), indicating complete genomic absence of nitrogen cycling genes.

Within Functionally Moderate classes, Gammaproteobacteria (9,466 genomes, PR = 0.710) showed balanced representation across DNN (48.6%), ANRA (31.0%), DNF (26.2%), and DNRA (10.1%). Actinomycetes (6,752 genomes, PR = 0.600) emphasised DNN (40.6%) and ANRA (33.2%), reflecting aerobic soil metabolism. Bacteroidia (6,816 genomes, PR = 0.291) maintained 70.9% functional absence while active genomes showed DNRA (13.0%) and DNF (9.3%). Alphaproteobacteria (6,094 genomes, PR = 0.550) exhibited distributed representation across DNN (29.9%), ANRA (24.7%), and DNF (22.7%). Clostridia (5,690 genomes, PR = 0.209) showed NF (18.2%) as its primary pathway despite low overall participation, while Cyanobacteriia (523 genomes, PR = 0.317) showed NF (28.1%) as its dominant function.

N-Retention Dominant classes are enriched for obligately or facultatively anaerobic lineages. Desulfovibrionia (175 genomes, PR = 0.886, NF = 54.3%, DNRA = 66.3%, loss = 0), Desulfuromonadia (81 genomes, PR = 0.963, NF = 88.9%, DNRA = 92.6%), and Desulfobulbia (75 genomes, PR = 0.987, NF = 72.0%, DNRA = 72.0%) confirm NF + DNRA co-dominance as a shared constraint across sulfate-reducing Desulfobacterota. Chlorobiia (86 genomes, PR = 0.977) showed near-universal NF (97.7%) as its sole function, while Brocadiia (22 genomes, PR = 0.955) combined DNRA (54.5%), DNF (54.5%), and DNN (86.4%), consistent with the anammox pathway (Strous et al., 1999). Syntrophia (27 genomes, PR = 0.556, DNRA = 44.4%) represents the lower PR boundary of this archetype. NF and DNRA frequencies were significantly higher in N-Retention Dominant than in Functionally Moderate classes (Mann–Whitney U: $p = 0.002$ and $p < 0.001$, respectively), while nitrogen loss was significantly lower ($p = 0.012$). The lower mean silhouette (0.155;

Figure S3) reflects a PR gradient within this archetype rather than misclassification, as all 17 classes share the NF + DNRA dominance and low nitrogen loss profile that define this archetype.

Nitrospiria's silhouette of 0 (Figure S3; set to 0 by convention for single-member clusters in the R cluster package) reflects its functional isolation: no comparable functional profile exists among the remaining 76 classes across k = 2–8.

Genome source association analysis provides quantitative validation of the intrinsic stability of taxonomy-based functional organization, examining whether class-level functional archetypes remain consistent across genomes sampled from different environmental categories. This analysis operates along two complementary dimensions: strategy-level stability quantified by source association indices (Figure S5), and functional frequency stability quantified by inter-class ranking consistency across environments (Figure S10).

Source association indices ranged 0.003–0.813 (mean 0.234 ± 0.158). Major generalist lineages exhibited the lowest indices (Gammaproteobacteria = 0.066, Actinomycetes = 0.063, Negativicutes = 0.019), indicating that their functional strategy profiles are intrinsic genomic properties insensitive to sampling provenance. Classes with higher indices (e.g., Terriglobia = 0.313, Clostridia = 0.267) tend to be functionally restricted specialists with low participation rates, for which sparse or environmentally skewed sampling produces greater apparent strategy deviation. The negative correlation between source association indices and class size ($\rho = -0.334$, $p < 0.001$) confirms that this variation is itself taxonomically structured rather than reflecting genuine environmental reprogramming of functional repertoires.

At the functional frequency level, inter-class rankings across six environmental categories were highly conserved: Spearman correlations of class-level rankings between environment pairs averaged $\rho = 0.978$ (range 0.937–1.000) for DNN, $\rho = 0.947$ (range 0.883–1.000) for ANRA, and $\rho = 0.925$ (range 0.709–1.000) for DNF. Within individual classes, absolute frequencies showed limited variation: Gammaproteobacteria DNN frequencies ranged 44–58% across environments (CV =

8.5%), representing stable functional identity, while higher CV values observed for lower-prevalence functions (e.g., DNRA CV = 53.6%) reflect stochastic sampling effects on rare capabilities rather than environmentally driven reprogramming. The 23 of 57 classes (40.4%) showing CV > 0.3 in participation rates (Figure S5) are disproportionately low-prevalence specialists, consistent with this interpretation. Taken together, these two dimensions of analysis confirm that class-level functional archetypes represent stable intrinsic properties of bacterial lineages. Within-class frequency variation across environments is real but occurs within a highly conserved inter-class ranking structure, supporting the use of taxonomic identity as a reliable organizational framework for nitrogen cycling functional inference across diverse ecological contexts.

Supplementary Note 3: Extended genus-level ecological strategy analysis

Specialist strategies maintain distinct ecological niches within structured communities, while generalist strategies provide competitive advantages through metabolic flexibility across variable nitrogen regimes. The moderate significance rate (53.3%) observed across genome source-strategy combinations reflects the complexity of these relationships, indicating that while systematic associations exist, they are modulated by multiple ecological and evolutionary factors. Detailed mechanistic analysis of environment-strategy associations reveals that aquatic environments present variable redox conditions characteristic of fluctuating oxygen availability, where metabolic flexibility in nitrogen transformations may provide advantages (Mosley et al., 2022). In host-associated environments, genome streamlining principles operate where microbes reduce investments in metabolically expensive pathways when substrates are readily available through host provisioning (Morris et al., 2012).

Moran's I analysis revealed the following detailed values: Multifunctional (I = 0.4394, p = 0.001), Non-N-Cycling (I = 0.4248, p = 0.001), N-Retention (I = 0.3217, p = 0.001), Intermediate (I = 0.3075, p = 0.001), and N-Loss (I = 0.2067, p = 0.001). For comparison, genome source distribution patterns across five major categories

exhibited Moran's I values of 0.032, while nitrogen cycling function frequencies displayed mean signals of 0.033, yielding an overall mean of 0.338 for ecological strategies. The evolutionary basis for this disparity likely involves horizontal gene transfer of small gene cassettes enabling individual gene mobility with relative ease (Arnold et al., 2022), while complete multi-gene functional modules require coordinated inheritance. This pattern aligns with observations that complex traits encoded by multiple genes tend to be more phylogenetically conserved than simpler traits (Isobe et al., 2019; Martiny et al., 2013). Phylogenetic clustering analysis revealed the following distributions: Non-N-Cycling genera within Clostridia and Bacilli (50.7% combined) and Bacteroidia (12.1%); Multifunctional genera within Gammaproteobacteria (35.9%) and Alphaproteobacteria (27.4%), with additional representation in Bacteroidia (12.9%).

Functional profile analysis across the five ecological strategy types revealed characteristic pathway combinations within each strategy. N-Retention genera (n = 468, mean ANRA 15.3%, NF 18.9%, DNRA 16.0%) ranged from single-pathway specialists to dual-retention lineages: *Geomonas* (23 genomes, NF 100%, DNRA 100%) and *Centipeda* (16 genomes, NF 100%, DNRA 100%) maintained maximum frequencies across both retention pathways simultaneously, while *Campylobacter* (251 genomes, DNRA 55.8%) and *Chlorobium* (68 genomes, NF 97.1%) exemplified single-pathway dominance at high frequency. N-Loss genera (n = 77, mean DNF 40.9%, NIT 1.7%) were dominated by denitrification, with *Nitrosomonas* (84 genomes, DNF 92.9%, NIT 51.2%) and *Nitrosospira* (24 genomes, DNF 87.5%, NIT 79.2%) carrying dual nitrification-denitrification capacity at high frequency. Intermediate genera (n = 64, mean DNN 35.4%) showed wide internal variation despite uniform absence of committed downstream pathways, with *Staphylococcus* (181 genomes, DNN 84.5%) and *Proteus* (23 genomes, DNN 95.7%) at the high end of DNN frequency. Multifunctional genera (n = 387, mean ANRA 30.7%, DNN 41.9%, DNF 34.9%, NF 9.5%, DNRA 9.1%) exhibited the broadest internal heterogeneity, with 51 genera simultaneously exceeding 50% retention and 50% loss

scores. Among these, *Bradyrhizobium* (268 genomes, NF 63.4%, DNF 56.7%), *Azospirillum* (61 genomes, NF 95.1%, DNF 72.1%), and *Methylobacter* (50 genomes, ANRA 66.0%, NF 66.0%, DNF 80.0%) represented lineages with substantial bidirectional nitrogen transformation capacity. As noted in the main text, *Methylomonas* (31 genomes, NF 90.3%, DNF 93.5%) exemplifies functionally paradoxical combinations in which nitrogen fixation and denitrification co-occur at high frequency. Additional genera exhibiting this pattern included *Basfia* (13 genomes, DNRA 100%, DNF 100%), *Aromatoleum* (18 genomes, ANRA 66.7%, NF 55.6%, DNF 100%), and *Thauera* (25 genomes, ANRA 56.0%, DNF 100%), collectively demonstrating that simultaneous high-frequency retention and loss capacities are not restricted to a single lineage but distributed across phylogenetically distinct genera within the Multifunctional strategy type.

These genus-level functional profiles provide reference baselines for community-scale nitrogen cycling inference and targeted cultivation. For 16S rRNA gene amplicon studies, community-level functional potential is estimated through abundance-weighted summation across all detected genera using values from Supplementary Table S2. For example, if 16S sequencing reveals *Pseudomonas* at 15% relative abundance, querying Supplementary Table S2 shows this genus (based on 1,020 representative genomes) exhibits DNN at 39.1% frequency, ANRA at 35.9%, and DNF at 28.9%, yielding genus-specific contributions to community functional potential of 5.9% for DNN, 5.4% for ANRA, and 4.3% for DNF. Summing such contributions across all detected genera provides comprehensive community-level functional estimates for each nitrogen transformation pathway. For targeted cultivation, genera combining substantial genome representation with elevated functional frequencies serve as efficient screening targets: *Neisseria* (105 genomes, DNF 92.4%) and *Nitrosomonas* for denitrification isolation; *Campylobacter* and *Citrobacter* (41 genomes, DNRA 97.6%) for DNRA-encoding bacteria; *Nostoc* (68 genomes, NF 100%) and *Chlorobium* for nitrogen fixation. These genomic potential estimates require validation through activity-based measurements including

quantitative PCR of functional genes, metatranscriptomics, or biogeochemical rate determinations, as environmental modulation, horizontal gene transfer, and post-transcriptional regulation introduce uncertainty in translating gene presence to metabolic activity.

Functional heterogeneity analysis across ecological strategy types revealed systematic differences in intra-genus functional diversity patterns (Figure S12A). Kruskal-Wallis tests demonstrated highly significant variation across all five heterogeneity indices: Shannon Functional Combination Diversity ($\chi^2 = 913.96$, $p < 0.001$), Coefficient of Variation of Functional Frequencies ($\chi^2 = 971.11$, $p < 0.001$), Functional Dominance ($\chi^2 = 828.51$, $p < 0.001$), Combination Richness ($\chi^2 = 977.75$, $p < 0.001$), and Functional Evenness ($\chi^2 = 677.22$, $p < 0.001$). Among nitrogen-cycling genera, Multifunctional genera exhibited the highest functional heterogeneity levels (mean Shannon FCD = 1.380), which likely reflects the complexity of coordinating both nitrogen retention and loss capabilities within genera. Retention genera showed intermediate heterogeneity (mean Shannon FCD = 0.732), while Intermediate and N-Loss genera displayed more constrained functional variation patterns. These differences indicate that coordination of multiple nitrogen transformation pathways within genera creates greater functional complexity than specialized single-pathway strategies (Shade and Gilbert, 2015).

Genome source distribution breadth analysis demonstrated a positive relationship between genus-level genome source distribution diversity and functional heterogeneity across ecological strategy types (Figure S12B). Integration of genome source plasticity and functional specialization patterns across qualified genera revealed four distinct categories through standardized axis analysis (Figure S13). Quadrant 2 (Phylogenetic Specialists) contained the largest proportion of genera (478 genera, 37.3%), characterized by low genome source plasticity and high functional specialization, dominated by Non-N-Cycling genera. Quadrant 3 (Phylogenetic Generalists) included 413 genera (32.2%) with low genome source plasticity and low functional specialization, primarily comprising Multifunctional and Retention genera.

Quadrant 4 (Genome Source Generalists) encompassed 227 genera (17.7%) exhibiting high genome source plasticity and low functional specialization, with Multifunctional genera representing the majority. Quadrant 1 (Genome Source Specialists) contained 163 genera (12.7%) displaying high genome source plasticity and high functional specialization.

Supplementary Note 4: Extended molecular evolutionary analysis

Class-level analysis of *nrfA* revealed systematic phylogenetic structuring in contrast to its environmental stability: Alphaproteobacteria exhibited the lowest diversity ($\pi = 0.390$), while Clostridia and Coriobacteriia displayed elevated values ($\pi = 0.447–0.463$), and Gammaproteobacteria showed intermediate-to-high values ($\pi = 0.453$) (Figure 5C). The combination of environmental stability and phylogenetic structuring indicates that the bimodal dN/dS distribution reflects lineage-specific rather than environment-driven evolutionary dynamics. The nine genera exhibiting *nrfA* dN/dS ≥ 1.0 span seven phylogenetically distinct bacterial classes: Gammaproteobacteria (*Shewanella*, *Actinobacillus*), Actinomycetes (*Arachnia*, *Actinomyces*), Myxococcia (*Myxococcus*), Bacteroidia (*Parabacteroides*), Blastocatellia (*OLB17*), Bacilli_A (*Bulleidia*), and Coriobacteriia (*Adlercreutzia*), with ω values ranging from 1.591 to 4.190. Despite exhibiting relaxed or positive selection at the sequence level, these nine genera maintain high DNRA functional frequencies (mean 70.1%, range 40.3-95.8% across genera), with elevated *nrfA* gene presence rates observed in sequenced genomes (e.g., *Shewanella* 94%, *Actinobacillus* 96%, *Myxococcus* 82%). This functional retention distinguishes the high-ω pattern from simple pathway degradation, where declining functional prevalence would be expected to accompany relaxed selection. Given that DNRA is deployed across diverse regulatory contexts (Durand and Guillier, 2021) and electron donor regimes (Liu et al., 2021; Yoon et al., 2015) in different bacterial lineages, the phylogenetically structured selection patterns likely reflect lineage-specific optimization of *nrfA* for integration within distinct metabolic networks, where

regulatory and energetic contexts of each bacterial clade necessitate particular sequence configurations while preserving the fundamental nitrite reduction chemistry.

Phylogenetic analysis of *napA* conservation across 20 bacterial classes revealed pronounced lineage-specific variation. Deinococci exhibited exceptionally stringent purifying selection ($\pi = 0.116$)—the lowest value observed for any gene-class combination—while Campylobacteria ($\pi = 0.313$) and Gammaproteobacteria ($\pi = 0.331$) showed substantially elevated diversity, representing a nearly 3-fold range in sequence conservation (Figure 5C). Environmental deployment contexts revealed additional variation: aquatic ($\pi = 0.399$) and engineered system ($\pi = 0.386$) environments occupied intermediate positions between the terrestrial and extreme environment extremes (Figure 5D). Unlike *narG*'s function during strict anaerobic respiration, *napA* enables periplasmic nitrate scavenging under microoxic-to-oxic conditions (Potter et al., 2001; Sparacino-Watkins et al., 2014), where maintenance of electron transfer efficiency across fluctuating redox states likely imposes stringent structural requirements. The correlation between *napA*'s high genus-level conservation and its environmental responsiveness demonstrates that ecological deployment contexts can impose selection pressures that modulate sequence evolution independently of pathway-level taxonomy-function association strength.

Detailed structural analysis of denitrification genes revealed mechanistic bases for within-pathway heterogeneity. The exceptional conservation of cnorB reflects structural constraints imposed by the binuclear heme b_3-FeB active site, where three conserved histidine residues and a glutamate residue coordinate the metal centers required for NO reduction (Murali et al., 2024; Schurig-Briccio et al., 2013). The evolutionarily unrelated nitrite reductases *nirS* and *nirK* showed significant differences reflecting their divergent structural architectures—cytochrome cd_1 (*nirS*) versus copper-containing (*nirK*) metalloforms. The divergent environmental sensitivities of *nirS* and *nirK* reflect their structural architectures and functional autonomy differences: *nirK* can function independently while *nirS* requires additional gene products for nitrite reduction (Ming et al., 2024), potentially accounting for

nirS's greater responsiveness to environmental variation. These patterns reveal that taxonomy-function association strength, while shaped by pathway-level associations, is further modulated by gene-specific structural and catalytic properties that contribute comparably to evolutionary rate variation.

Detailed analysis of nucleotide diversity patterns across 13 nitrogen cycling genes revealed substantial variation in evolutionary constraints beyond those discussed extensively in the main text (Figure 5). Genes not analyzed in detail in the main text include nitrous oxide reductase *nosZ* ($\pi = 0.20 \pm 0.073$), ammonia monooxygenase *amoA* ($\pi = 0.20 \pm 0.071$), quinol-dependent nitric oxide reductase *qnor* ($\pi = 0.24 \pm 0.089$), and hydroxylamine oxidoreductase *hao* ($\pi = 0.27 \pm 0.13$). Notably, *hao* displayed the highest standard deviation among all genes analyzed, suggesting substantial variation in evolutionary constraints across nitrifying lineages. Comparative analysis of functional gene groups revealed that assimilatory genes (*nas*, *nirB*; mean $\pi = 0.285$) exhibited significantly higher diversity than dissimilatory respiratory genes (*napA*, *narG*, *nrfA*; mean $\pi = 0.207$), indicating stronger purifying selection on respiratory pathway components. Within denitrification, evolutionary rates varied systematically, with *cnorB* and *nosZ* showing the highest conservation ($\pi = 0.18$-0.20), followed by *nirS* ($\pi = 0.22$), *qnor* ($\pi = 0.24$), and *nirK* ($\pi = 0.25$), reflecting structural complexity hierarchies from binuclear metal centers to copper-dependent catalytic sites.

Class-level molecular evolution analysis across 20 major bacterial classes demonstrated systematic phylogenetic structuring of nitrogen cycling gene evolution (Figure 5C). Beyond the range extremes noted in the main text (Cyanobacteriia mean $\pi = 0.29$, Clostridia mean $\pi = 0.45$), intermediate conservation patterns were observed across multiple classes. Deinococci showed elevated molecular conservation across analyzed genes, with particularly low diversity in *napA* ($\pi = 0.116$), while Negativicutes displayed consistently elevated conservation across multiple nitrogen cycling functions. Classes exhibiting more relaxed molecular evolution included Spirochaetia and Bacilli_A, with mean π values approaching those observed in

Clostridia. This phylogenetic variation demonstrates that evolutionary rate variation operates at multiple hierarchical levels, with class-level phylogenetic background contributing substantially to molecular evolutionary rates independently of gene function. Hierarchical clustering revealed organizational structures reflecting both functional modularity and phylogenetic relationships, with gene clustering separating nitrogen cycling functions into distinct modules while class clustering grouped phylogenetically related lineages regardless of their functional capabilities.

Environmental sensitivity analysis revealed substantial gene-specific variation in environmental responsiveness across all 13 nitrogen cycling genes, with environmental sensitivity quantified as the coefficient of variation in π values across six environmental categories (Figure 5D-E). Beyond genes discussed extensively in the main text, comprehensive analysis confirmed that environmental sensitivity does not correspond to pathway membership or functional categories. Within denitrification, genes exhibited divergent environmental sensitivities despite belonging to the same functional pathway, and within dissimilatory nitrate reduction pathways, *nrfA* displayed exceptional environmental stability (CV = 0.023) while *napA* showed high environmental responsiveness (CV = 0.11), demonstrating that genes within the same metabolic context can respond independently to ecological gradients. Detailed examination of environment-specific π patterns (Figure 5D) revealed that extreme environments consistently induced elevated diversity across multiple genes, while terrestrial environments showed reduced diversity particularly for nitrite reductases and periplasmic reductases, confirming that environmental contexts impose gene-specific rather than pathway-general selective pressures.

Genus-level diversity analysis across 891 genera representing four ecological strategy types revealed systematic relationships between functional organization and molecular evolution (Figure S14). Intermediate genera exhibited the lowest nucleotide diversity (mean $\pi = 0.19 \pm 0.085$), followed by N-Loss (mean $\pi = 0.22 \pm 0.076$), Multifunctional (mean $\pi = 0.23 \pm 0.061$), and Retention (mean $\pi = 0.26 \pm 0.088$). Statistical analysis confirmed significant differences among ecological types

(Kruskal-Wallis $\chi^2 = 54.89$, df = 3, $p < 0.001$). These patterns suggest systematic differences between functional strategy and molecular evolutionary constraints, though substantial variation within strategy types indicates complex interplay between functional repertoire, taxonomic position, and environmental adaptation.

Supplementary Figure

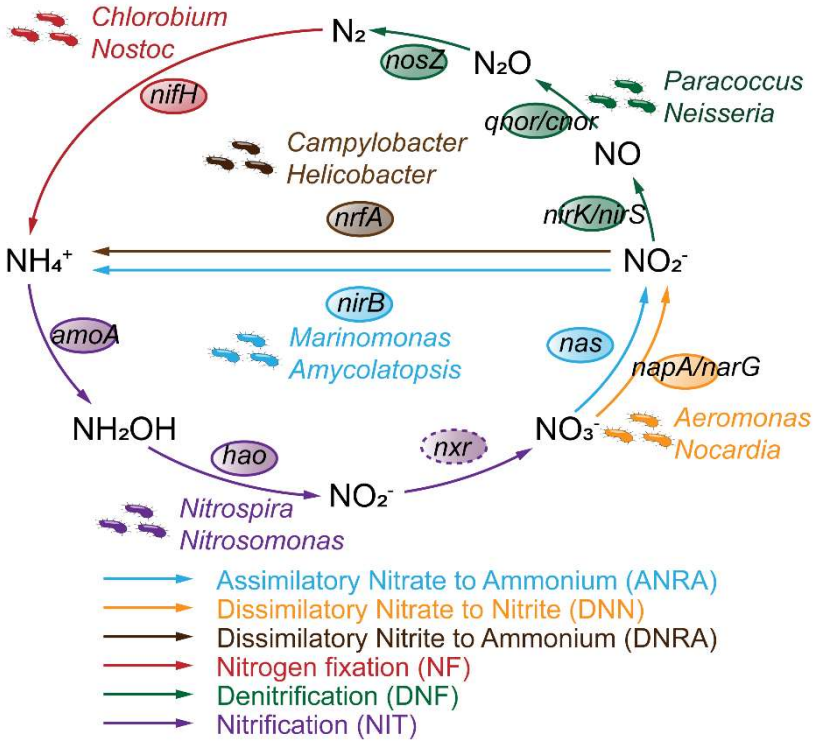


Figure S1. Conceptual framework of bacterial nitrogen cycling pathways and gene detection strategy. Six major nitrogen transformation processes: Assimilatory Nitrate to Ammonium reduction (ANRA, blue), Dissimilatory Nitrate to Nitrite reduction (DNN, orange), Dissimilatory Nitrite to Ammonium reduction (DNRA, brown), Nitrogen fixation (NF, red), Denitrification (DNF, green), and Nitrification (NIT, purple). Representative key genes are shown in gray ovals with solid lines; the *nxr* gene (dashed oval) was not included in the analysis due to annotation database limitations. The complete list of core and auxiliary genes for each pathway is provided in Supplementary Table S1. Functional capability was assigned using a dual-detection strategy requiring all core genes plus at least one auxiliary gene per pathway. Representative bacterial genera known to carry these pathways are labeled for each transformation process.

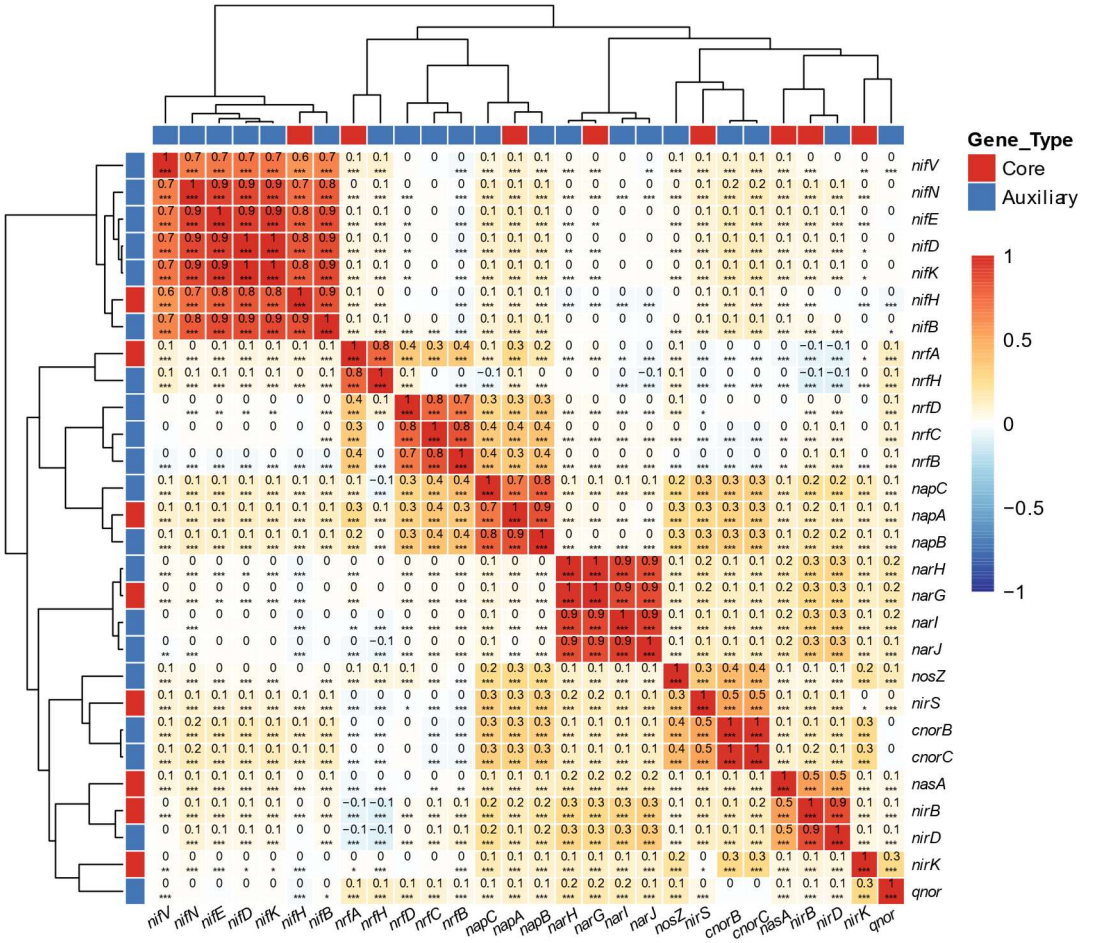


Figure S2. Core-auxiliary gene correlation analysis for nitrogen cycling pathways. Spearman correlation matrix of nitrogen cycling genes with frequency ≥5% across 73,472 GTDB representative genomes. Gene categories: Core (red), Auxiliary (blue), Regulatory (light blue), Other (gray). Correlation coefficients shown with statistical significance: ***P<0.001, **P<0.01, *P<0.05 (FDR-corrected). Hierarchical clustering reveals pathway-specific correlation patterns between core and auxiliary genes.

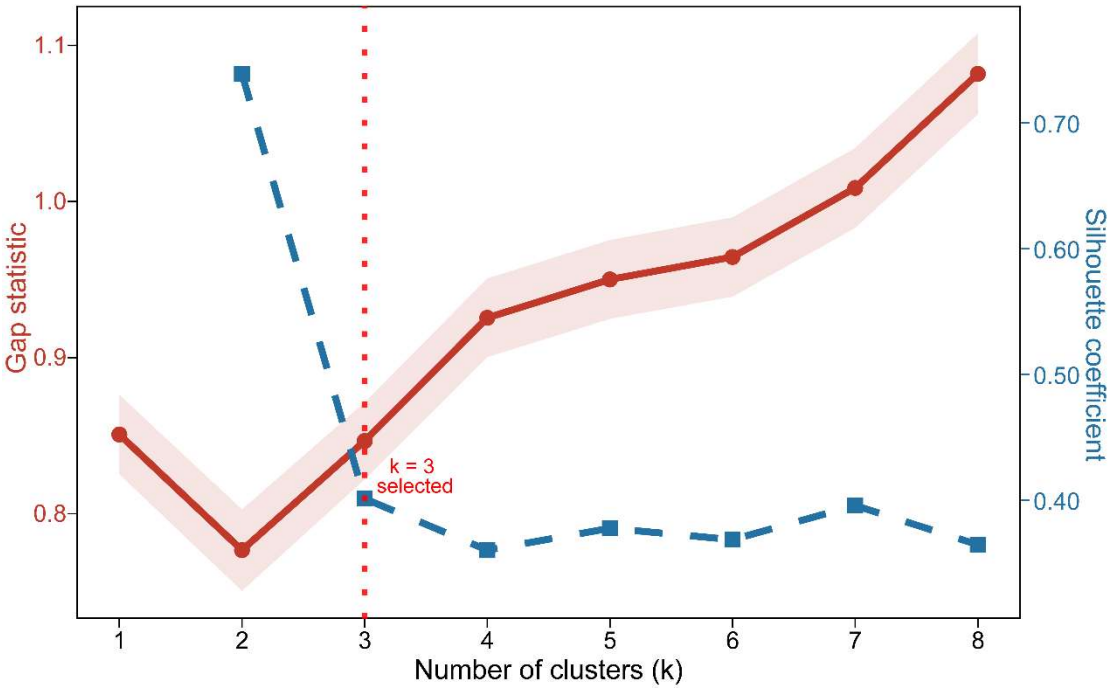


Figure S3. Cluster number validation for Ward D2 hierarchical clustering of 77 bacterial classes. Gap statistic (red solid line, left y-axis) and mean silhouette coefficient (blue dashed line, right y-axis) across k = 1–8. Vertical dotted line indicates selected k = 3. Point size represents genome count per class.

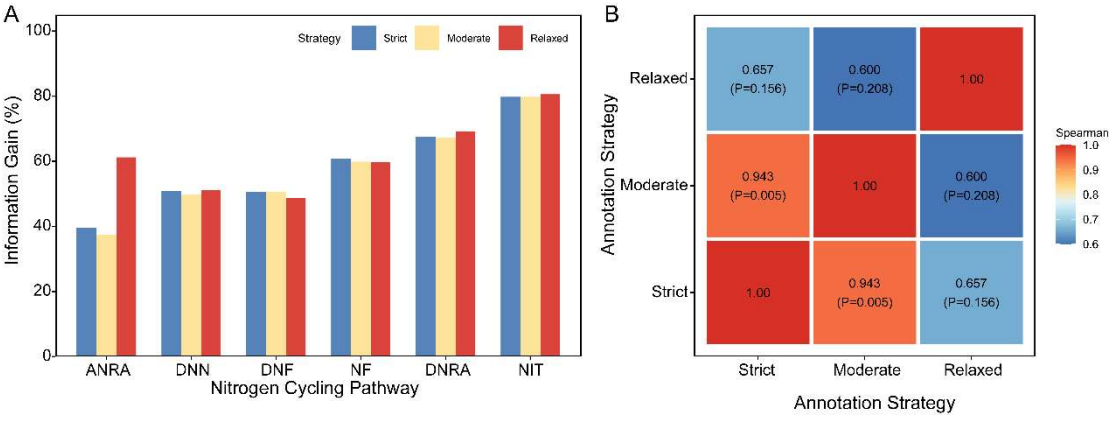


Figure S4. Sensitivity analysis of genus-level Information Gain across annotation strategies. (A) Comparison of Information Gain values for six nitrogen cycling pathways under Strict (core plus auxiliary genes), Moderate (core genes only), and Relaxed (any pathway-related gene) annotation strategies. (B) Pairwise Spearman rank correlations between annotation strategies across six pathways.

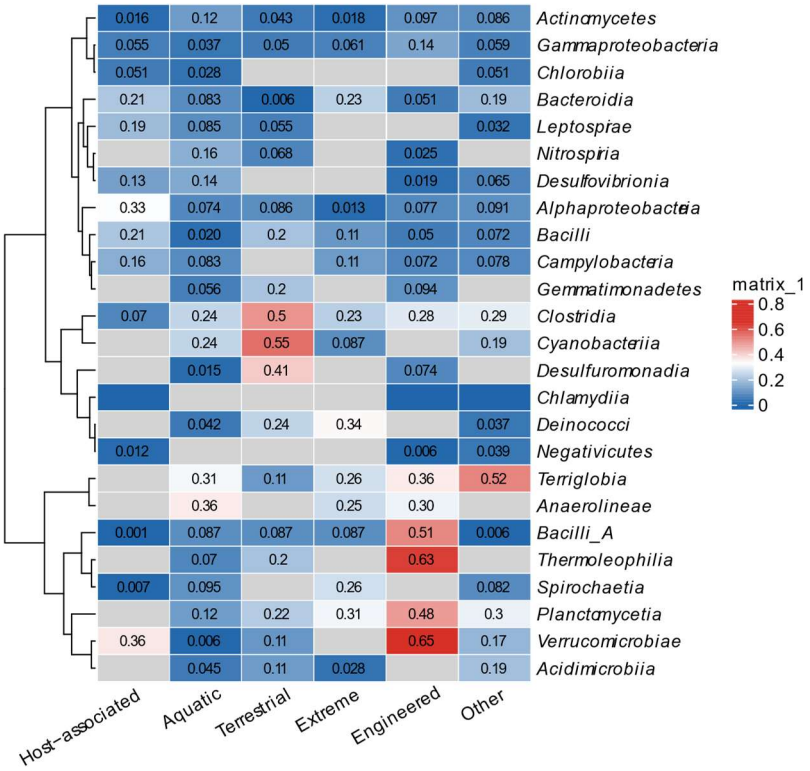


Figure S5. Environmental adaptation patterns across bacterial classes. Heatmap displaying adaptation strength calculated as Euclidean distance between class genome source strategy and average strategy. Higher values indicate greater genome source plasticity

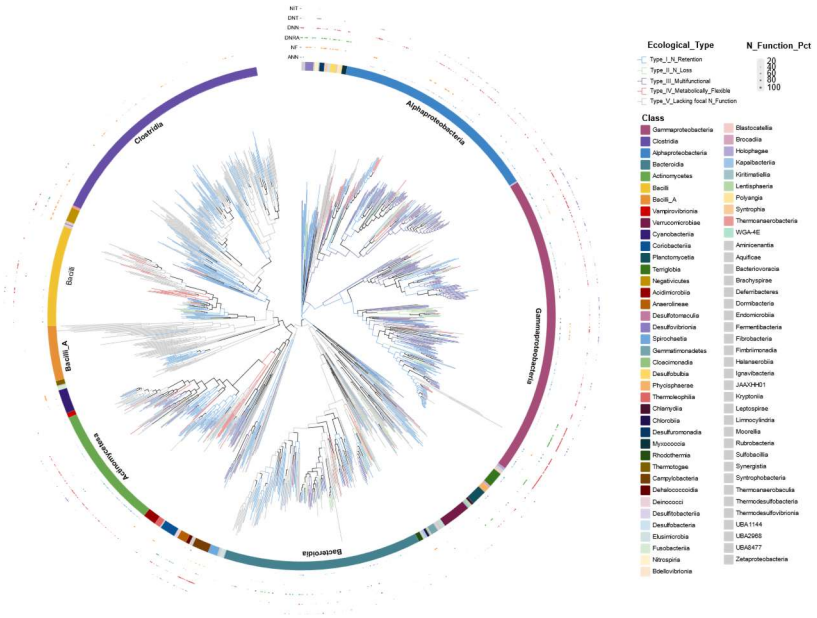


Figure S6. Phylogenetic distribution of ecological strategies across bacterial genera. Maximum-likelihood phylogenetic tree of 1,281 qualified genera with branches colored by ecological strategy types. Outer rings show prevalence patterns of six nitrogen cycling functions within each genus. Tree construction based on concatenated alignment of 16 ribosomal proteins.

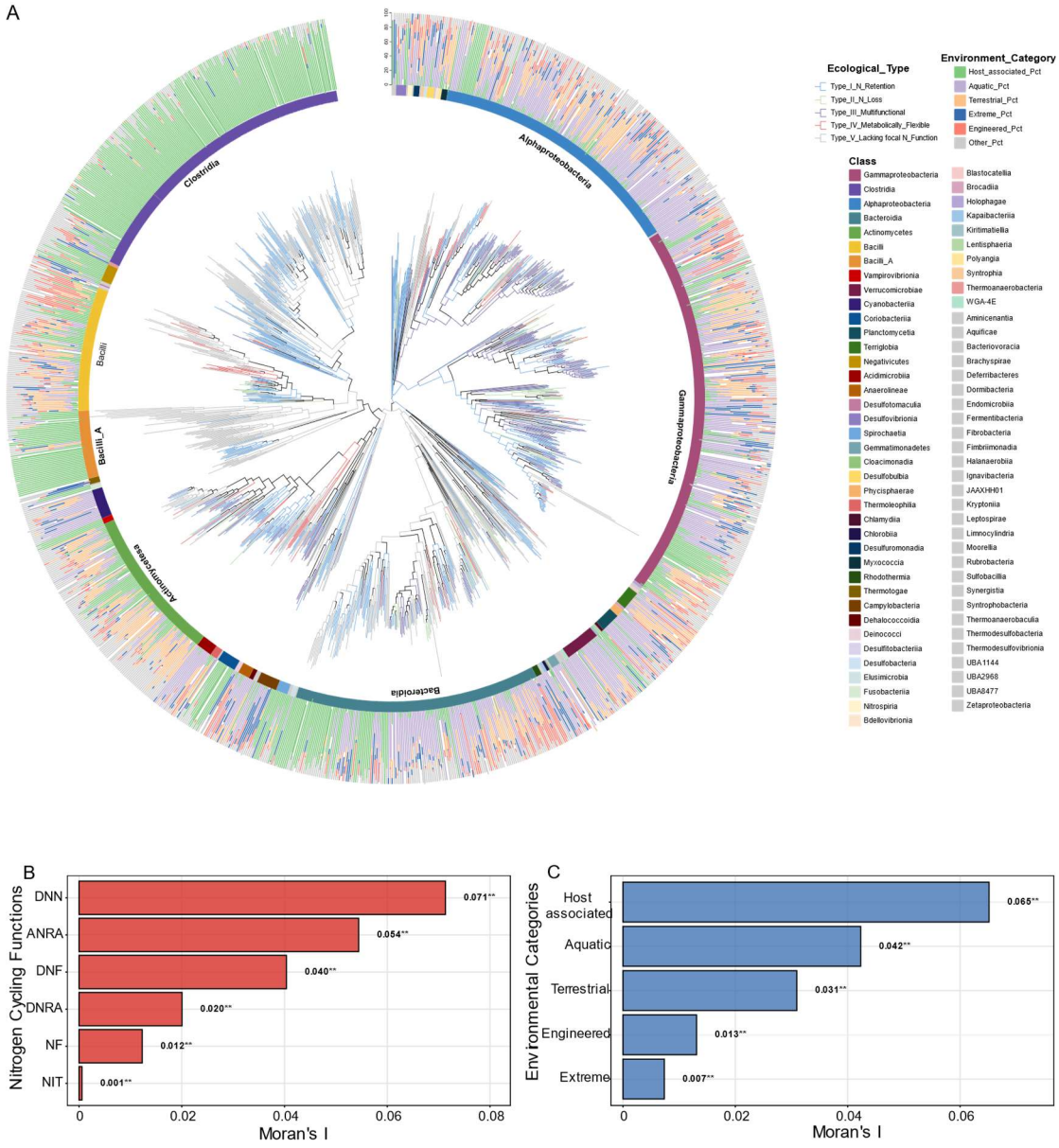


Figure S7. Multi-scale phylogenetic constraints in nitrogen cycling functions and genome source preferences. (A) Phylogenetic distribution of ecological strategies and genome source associations. Maximum-likelihood phylogenetic tree of 1,281 genera with tree branches colored by ecological strategy types. Inner ring displays strategy classifications; outer ring shows genome source category distributions. (B) Phylogenetic signal in genome source distributions. Moran's I values for six genome source categories across 1,281 genera showing weak phylogenetic signals. (C) Phylogenetic signal in nitrogen cycling gene frequencies. Moran's I values for six nitrogen cycling functions across 1,281 genera ordered by signal strength. Statistical significance: **P < 0.01.

Supplementary Note Figure

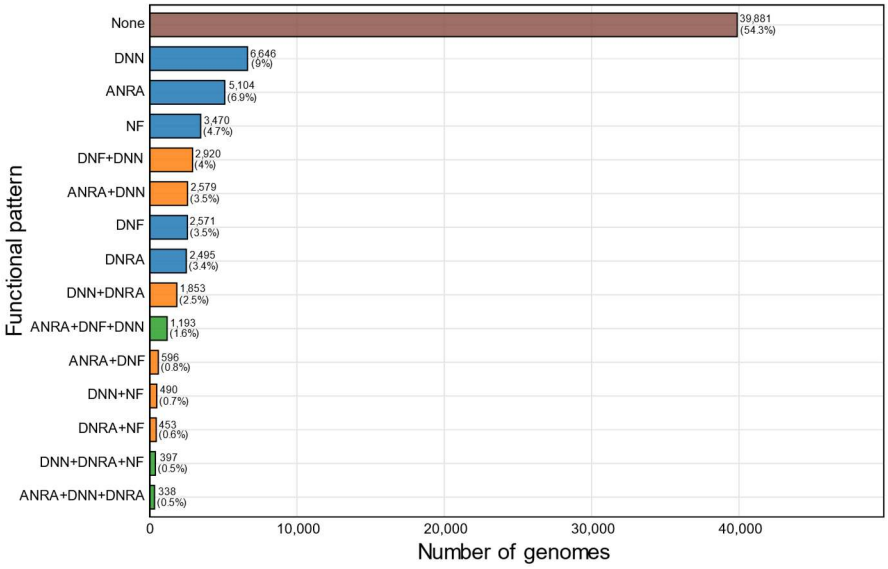


Figure S8. Distribution of nitrogen cycling functional patterns across bacterial genomes. The 15 most common nitrogen cycling function combinations among 73,472 GTDB genomes. "None" indicates no detectable function markers.

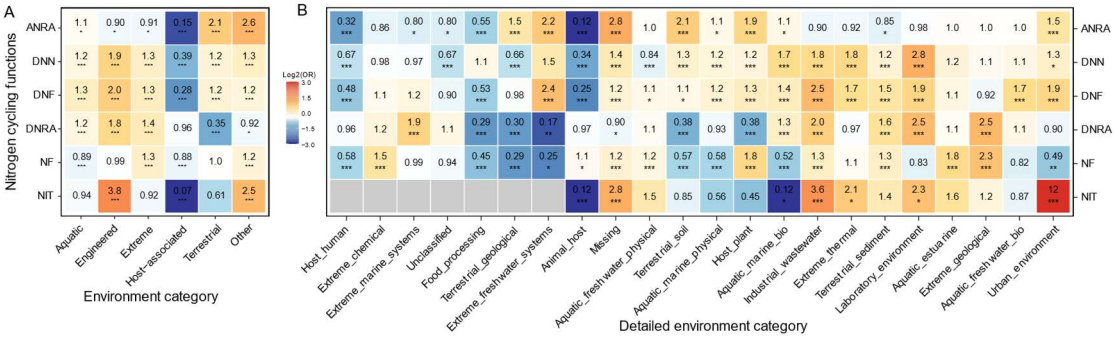


Figure S9. Environmental associations of nitrogen cycling functions. (A) Function enrichment across major genome source categories. (B) Enrichment patterns across detailed source subtypes. Red: enrichment (OR>1.5), blue: depletion (OR<0.67), grey: function absent (no genomes with the function detected in that environment). Statistical significance: ***P<0.001, **P<0.01, *P<0.05.

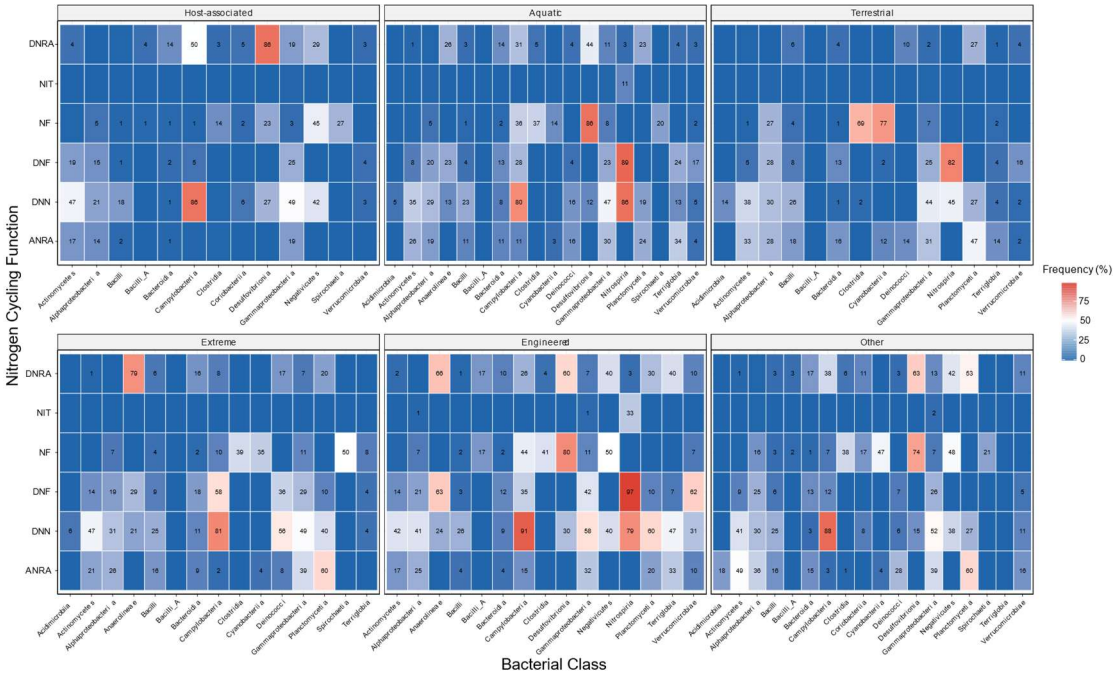


Figure S10. Class-level functional profiles across major bacterial lineages. Stacked horizontal bars showing distribution of individual nitrogen cycling function frequencies within major bacterial classes (≥500 genomes).

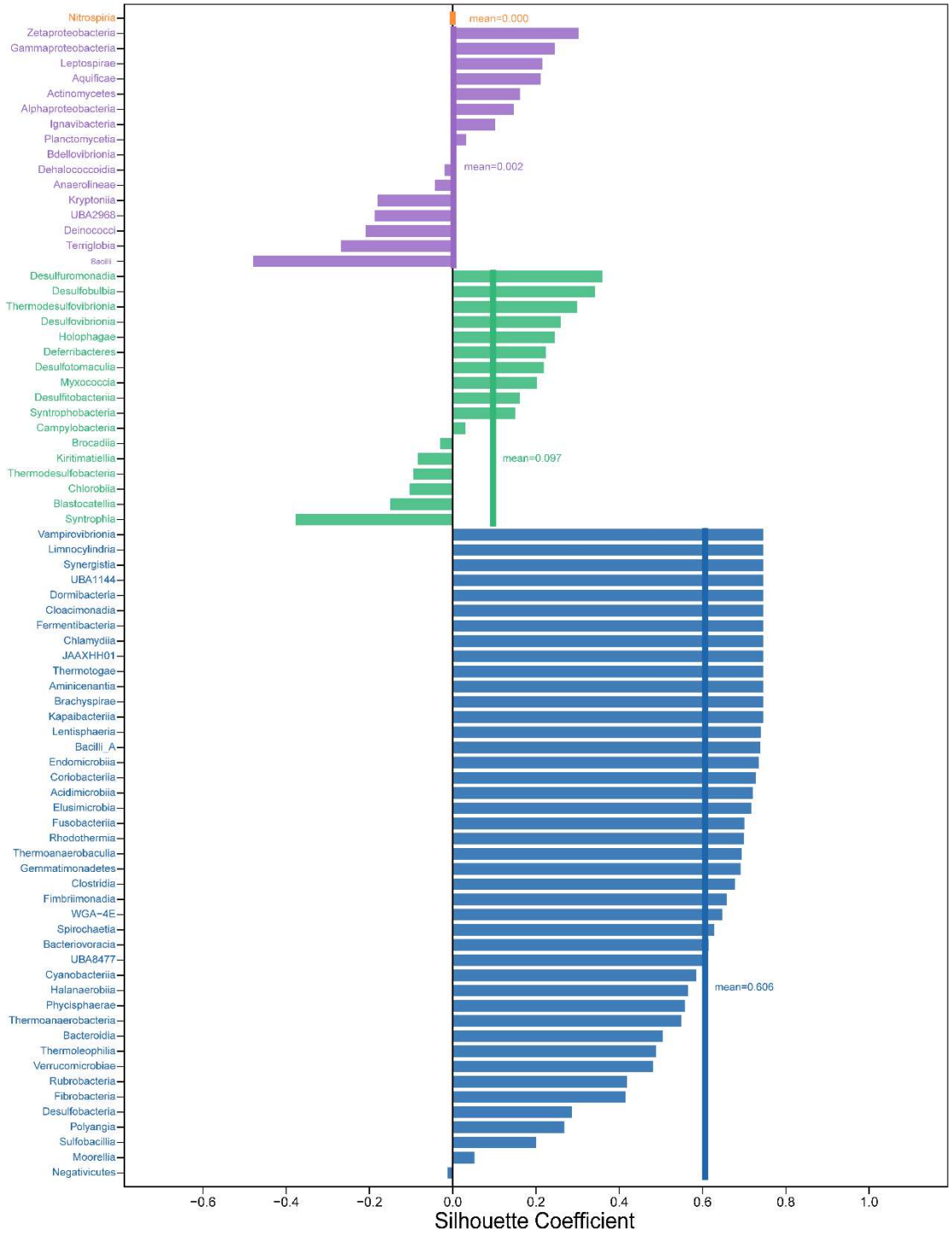


Figure S11. Silhouette analysis of Ward D2 hierarchical clustering (k = 3) for 77 bacterial classes. Each bar represents one class, ordered by archetype then silhouette value. Vertical lines indicate archetype mean silhouette values. Nitrification Specialist silhouette set to 0 by convention for single-member clusters.

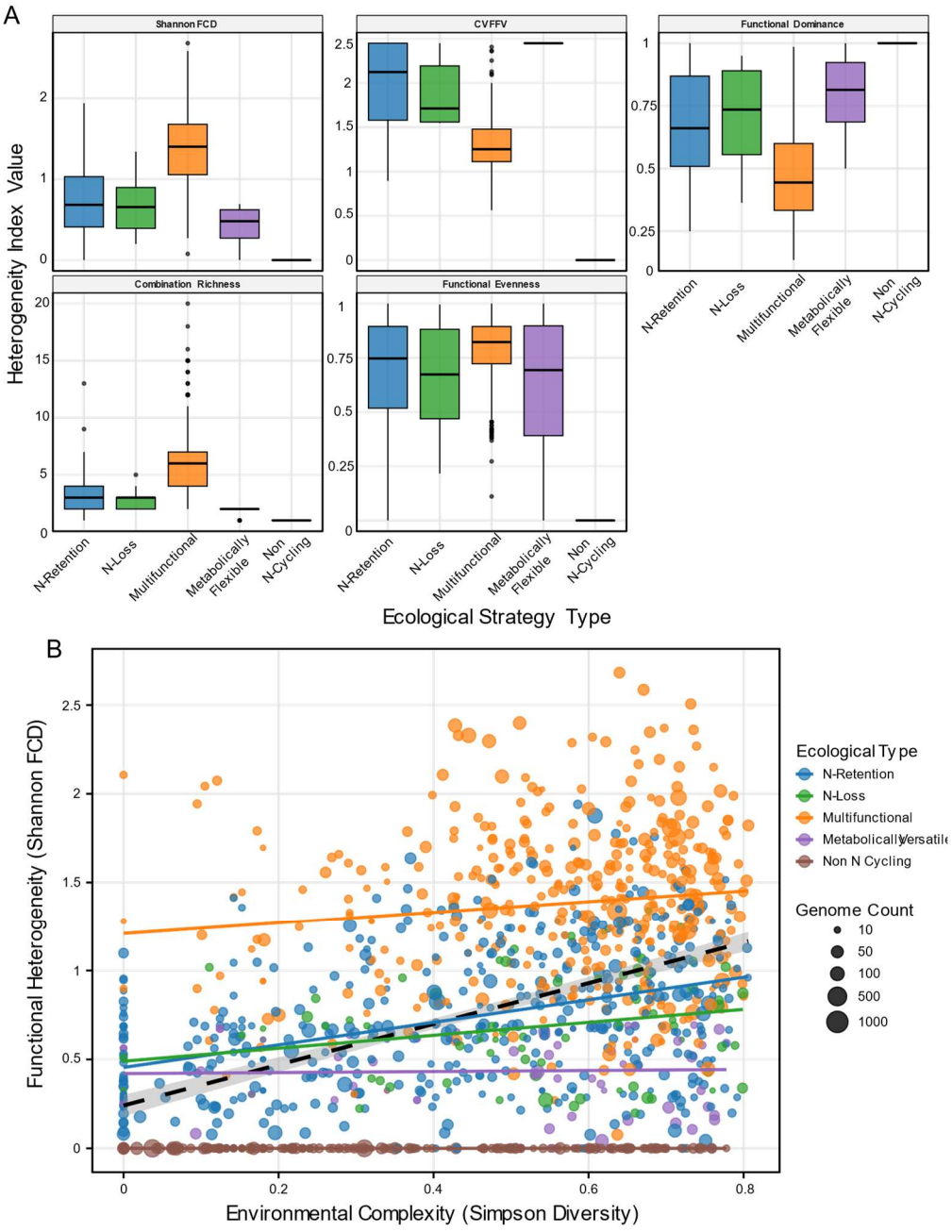
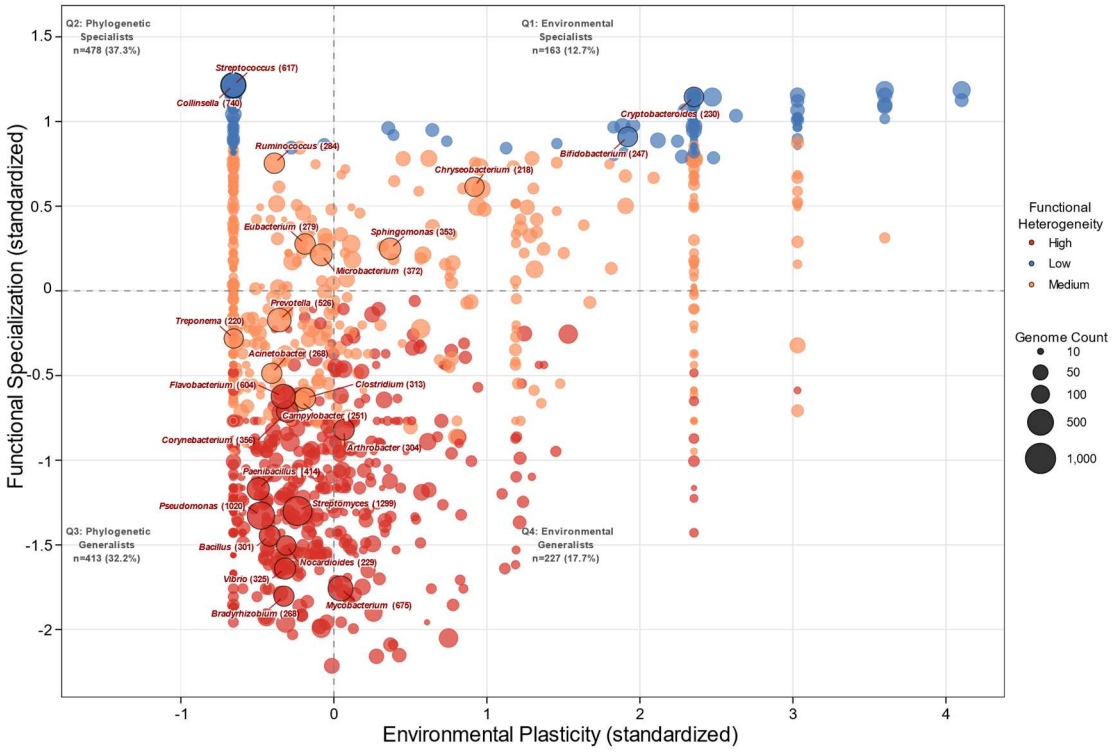
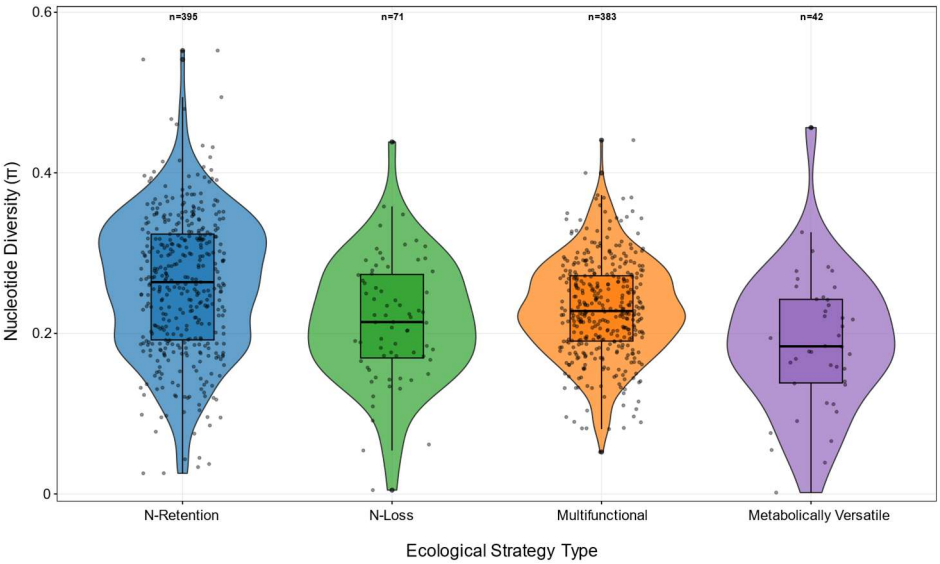


Figure S12. Functional heterogeneity patterns across ecological strategy types. (A) Functional heterogeneity indices by ecological strategy type. Five-panel boxplot displays Shannon Functional Combination Diversity, Coefficient of Variation, Functional Dominance, Combination Richness, and Functional Evenness across qualified genera. Kruskal-Wallis tests demonstrated significant variation across all indices. (B) Genome source distribution breadth versus functional heterogeneity across ecological strategy types. Point colors indicate strategy types. Linear regression: $R^2 = 0.68$, $p < 0.001$.



636

637 Figure S13. Four-quadrant integration framework for evolutionary strategies. Scatterplot
638 positioning qualified genera along Genome Source Plasticity (x-axis) and Functional
639 Specialization (y-axis). Point colors represent functional heterogeneity levels; quadrant analysis
640 reveals four distinct adaptation strategies.



641

642 Figure S14. Conservation indices across ecological strategy types. Violin plots showing
643 conservation index distributions across ecological strategy types (n = 814 genera). Box plots
644 indicate quartiles. Kruskal-Wallis test: $\chi^2 = 54.96$, $P < 0.001$.

Supplementary Tables

Supplementary Table S1. Core and auxiliary genes used in nitrogen transformation pathway identification

Pathway	Gene Symbol	KEGG Annotation	RAST Annotation	Core Gene
Assimilatory Nitrate Reduction (ANRA)	*nasA*	Assimilatory nitrate reductase electron transfer subunit (NasB)	Assimilatory nitrate reductase large subunit (EC 1.7.99.4)	Yes
	nirB	Nitrite reductase (NADH) large subunit (NirB)	Nitrite reductase [NAD(P)H] large subunit (EC 1.7.1.4)	Yes
	nirD	Nitrite reductase (NADH) small subunit (NirD)	Nitrite reductase [NAD(P)H] small subunit (EC 1.7.1.4)	No
Dissimilatory Nitrate to Nitrite (DNN)	*napA*	Periplasmic nitrate reductase (NapA)	Periplasmic nitrate reductase (EC 1.7.99.4)	Yes
	napB	Cytochrome c550-type protein (NapB)	Nitrate reductase cytochrome c550-type subunit	No
	napC	Cytochrome c-type protein NapC	Cytochrome c-type protein NapC	No
	narG	Nitrate reductase delta subunit (NarG)	Respiratory nitrate reductase alpha chain (EC 1.7.99.4)	Yes
	narH	Nitrate reductase gamma subunit (NarH)	Respiratory nitrate reductase beta chain (EC 1.7.99.4)	No
	narI	Nitrate reductase alpha subunit (NarI)	Respiratory nitrate reductase gamma chain (EC 1.7.99.4)	No
	narJ	Nitrate reductase delta subunit (NarJ)	Respiratory nitrate reductase delta chain (EC 1.7.99.4)	No
	narC	Cytochrome b-561 (NarC)	Respiratory nitrate reductase subunit, conjectural (EC 1.7.99.4)	No
Dissimilatory Nitrite to Ammonia (DNRA)	*nrfA*	Nitrite reductase cytochrome c-552 (NrfA)	Cytochrome c552 precursor (EC 1.7.2.2)	Yes
	nrfH	Protein NrfH (NrfH)	Cytochrome c nitrite reductase, small subunit NrfH	No
	nrfB	Cytochrome c-type protein (NrfB)	Cytochrome c-type protein NrfB precursor	No

	nrfC	Nitrite reductase protein NrfC (NrfC)	NrfC protein	No
	nrfD	Protein nrfD (NrfD)	NrfD protein	No
Denitrification (DNF)	*nirK*	Copper-containing nitrite reductase (NirK)	Copper-containing nitrite reductase (EC 1.7.2.1)	Yes
	nirS	Cytochrome cd1 nitrite reductase (NirS)	Cytochrome cd1 nitrite reductase (EC 1.7.2.1)	Yes
	qnor	Nitric-oxide reductase quinol-dependent (qNor)	Nitric-oxide reductase (EC 1.7.99.7), quinol-dependent	No
	cnorB	Nitric-oxide reductase subunit B (cNor-B)	Nitric-oxide reductase subunit B (EC 1.7.99.7)	No
	cnorC	Nitric-oxide reductase subunit C (cNor-C)	Nitric-oxide reductase subunit C (EC 1.7.99.7)	No
	nosZ	Nitrous-oxide reductase (NosZ)	Nitrous-oxide reductase (EC 1.7.99.6)	No
Nitrogen Fixation(NF)	*nifH*	Nitrogenase reductase (NifH)	Nitrogenase (molybdenum-iron) reductase and maturation protein NifH	Yes
	nifD	Nitrogenase molybdenum-iron protein subunit alpha (NifD)	Nitrogenase (molybdenum-iron) alpha chain (EC 1.18.6.1)	No
	nifK	Nitrogenase molybdenum-iron protein subunit beta (NifK)	Nitrogenase (molybdenum-iron) beta chain (EC 1.18.6.1)	No
	nifB	Nitrogenase molybdenum-iron cofactor biosynthesis protein (NifN)	Nitrogenase FeMo-cofactor synthesis FeS core scaffold and assembly protein NifB	No
	nifE	Nitrogenase molybdenum-cofactor biosynthesis protein (NifE)	Nitrogenase FeMo-cofactor scaffold and assembly protein NifE	No
	nifN	Nitrogenase iron protein 2 (NifH2)	Nitrogenase FeMo-cofactor scaffold and assembly protein NifN	No
	nifV	Homocitrate synthase (NifV)	Homocitrate synthase (EC 2.3.3.14)	No
Nitrification (NIT)	*amoA*	Ammonia monooxygenase (Amo)	Ammonia monooxygenase	Yes

	hao	Hydroxylamine oxidase (Hao)	Hydroxylamine oxidoreductase precursor (EC 1.7.3.4)	Yes

Supplementary Table S3. dN/dS ratios across nitrogen cycling genes (≥20 sequences per genus)

Gene	**N[a]**	**Median ω**	**IQR (Q25-Q75)**	**Mean±SD**	**% ω<0.3[b]**	**% ω≥1.0[c]**
napA	46	0.057	0.050-0.087	0.15±0.43	93.5	2.2
cnorB	36	0.089	0.045-0.13	0.12±0.21	97.2	2.8
amoA	8	0.083	0.049-0.11	0.08±0.03	100	0
nosZ	52	0.091	0.067-0.15	0.81±1.66	78.8	19.2
narG	96	0.096	0.070-0.31	0.45±0.94	75	12.5
nifH	62	0.081	0.062-0.11	0.14±0.22	91.9	1.6
nirS	26	0.14	0.096-1.40	0.82±1.17	69.2	30.8
qnorB	49	0.1	0.073-0.21	0.37±0.68	79.6	12.2
nrfA	47	0.11	0.075-0.49	0.57±0.95	74.5	19.1
nirK	62	0.17	0.11-0.70	0.58±0.92	62.9	17.7
nirB	228	0.29	0.097-0.84	0.54±0.64	51.8	18.4
hao	5	0.089	0.076-0.12	0.16±0.15	80	0
nasA	98	0.17	0.14-0.67	0.53±0.70	63.3	18.4

[a] Number of genera with ≥20 sequences

[b] Percentage of genera under strong purifying selection (ω < 0.3)

[c] Percentage of genera under relaxed constraint or positive selection (ω ≥ 1.0)

dN/dS ratios (ω) were estimated using HyPhy SLAC. Median values and interquartile ranges (IQR) are reported due to the highly skewed distribution of ω across genera. Genes are ordered by increasing nucleotide diversity (π) as shown in Figure 5A.

Supplementary Table S4. Kruskal–Wallis test results for inter-archetype differences across functional metrics.

Metric	H	p	η^{2a}
Participation rate	60.59	< 0.001	0.789
Nitrogen retention index	58.97	< 0.001	0.767
DNRA frequency	42.05	< 0.001	0.535
DNN frequency	31.77	< 0.001	0.394
NF frequency	25.70	< 0.001	0.311
DNF frequency	24.58	< 0.001	0.296
Nitrogen loss index	24.58	< 0.001	0.296
ANRA frequency	20.67	< 0.001	0.242

[a] $\eta^2 = (H - k + 1)/(n - k)$; calculated across Functionally Inactive (n = 25), Functionally Moderate (n = 34), and N-Retention Dominant (n = 17). Nitrification Specialist (n = 1) excluded from Kruskal–Wallis analysis.

Supplementary References:

Ahn, S., Cho, M., Sadowsky, M.J., Jang, J. (2025) Dissimilatory nitrate reductions in soil Neobacillusand Bacillusstrains under aerobic condition. Journal of Microbiology 63, e2411019.

Arnold, B.J., Huang, I.T., Hanage, W.P. (2022) Horizontal gene transfer and adaptive evolution in bacteria. Nature Reviews Microbiology 20, 206-218.

Benjamini, Y., Hochberg, Y. (1995) CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING. Journal of the Royal Statistical Society Series B-Statistical Methodology 57, 289-300.

Canty, A., Ripley, B. (2017) boot: Bootstrap R (S-Plus) Functions. R package version 1.3-20. CRAN R Project.

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T. (2009) trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972-1973.

Daims, H., Lücker, S., Wagner, M. (2016) A new perspective on microbes formerly known as nitrite-oxidizing bacteria. Trends in microbiology 24, 699-712.

Dunn, O.J. (1964) Multiple comparisons using rank sums. Technometrics 6, 241-252.

Durand, S., Guillier, M. (2021) Transcriptional and post-transcriptional control of the nitrate respiration in bacteria. Frontiers in molecular biosciences 8, 667758.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32, 1792-1797.

Garnier, S., Ross, N., Rudis, R., Camargo, P.A., Sciaini, M., Scherer, C. (2021) viridis-Colorblind-friendly color maps for R. R package version 0.6 2.

Gómez-Rubio, V. (2017) ggplot2-elegant graphics for data analysis. Journal of Statistical Software 77, 1-3.

Graf, D.R.H., Jones, C.M., Hallin, S. (2014) Intergenomic comparisons highlight modularity of the denitrification pathway and underpin the importance of community structure for N2O emissions. PLoS One 9, e114118.

Guerrero, M.G., Lara, C. (1987) Assimilation of inorganic nitrogen.

Holm, S. (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics, 65-70.

Isobe, K., Allison, S.D., Khalili, B., Martiny, A.C., Martiny, J.B.H. (2019) Phylogenetic conservation of bacterial responses to soil nitrogen addition across continents. Nature Communications 10.

Jombart, T., Balloux, F., Dray, S. (2010) adephylo: new tools for investigating the phylogenetic signal in biological traits. Bioinformatics 26, 1907-1909.

Katoh, K., Standley, D.M. (2013) MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol 30, 772-780.

Kolde, R., (2019) pheatmap: Pretty Heatmaps. R package version 1.0. 12.

Kosakovsky Pond, S.L., Frost, S.D.W. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22, 1208-1222.

Kruskal, W.H., Wallis, W.A. (1952) Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 47, 583-621.

Kuypers, M.M.M., Marchant, H.K., Kartal, B. (2018) The microbial nitrogen-cycling network. Nature Reviews Microbiology 16, 263-276.

Liu, S., Dai, J., Wei, H., Li, S., Wang, P., Zhu, T., Zhou, J., Qiu, D. (2021) Dissimilatory nitrate reduction to ammonium (DNRA) and denitrification pathways are leveraged by cyclic AMP receptor protein (CRP) paralogues based on electron donor/acceptor limitation in Shewanella loihica PV-4. Applied and Environmental Microbiology 87, e01964-01920.

Lücker, S., Wagner, M., Maixner, F., Pelletier, E., Koch, H., Vacherie, B., Rattei, T., Damsté, J.S.S., Spieck, E., Le Paslier, D. (2010) A Nitrospira metagenome illuminates the physiology and evolution of globally important nitrite-oxidizing bacteria. Proceedings of the National Academy of Sciences 107, 13479-13484.

Lynch, M. (2007) The evolution of genetic networks by non-adaptive processes. Nature Reviews Genetics 8, 803-813.

Martiny, A.C., Treseder, K., Pusch, G. (2013) Phylogenetic conservatism of functional traits in microorganisms. Isme Journal 7, 830-838.

Ming, Y., Al, M.A., Zhang, D., Zhu, W., Liu, H., Cai, L., Yu, X., Wu, K., Niu, M., Zeng, Q. (2024) Insights into the evolutionary and ecological adaption strategies of nirS- and nirK-type denitrifying communities. Molecular Ecology 33, 13.

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., Von Haeseler, A., Lanfear, R. (2020) IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37, 1530-1534.

Morris, J.J., Lenski, R.E., Zinser, E.R. (2012) The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. Mbio 3.

Mosley, O.E., Gios, E., Close, M., Weaver, L., Daughney, C., Handley, K.M. (2022) Nitrogen cycling and microbial cooperation in the terrestrial subsurface. Isme Journal 16, 2561-2573.

Murali, R., Pace, L.A., Sanford, R.A., Ward, L.M., Lynes, M.M., Hatzenpichler, R., Lingappa, U.F., Fischer, W.W., Gennis, R.B., Hemp, J. (2024) Diversity and evolution of nitric oxide reduction in bacteria and archaea. Proceedings of the National Academy of Sciences 121, e2316422121.

Nguyen, L.T., Schmidt, H.A., von Haeseler, A., Minh, B.Q. (2015) IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Mol Biol Evol 32, 268-274.

Paradis, E., Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35, 526-528.

Parks, D.H., Chuvochina, M., Waite, D.W., Rinke, C., Skarshewski, A., Chaumeil, P.A., Hugenholtz, P. (2018) A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. Nature Biotechnology 36, 996-+.

Potter, L., Angove, H., Richardson, D., Cole, J. (2001) Nitrate reduction in the periplasm of gram-negative bacteria.

Schimel, J.P., Bennett, J. (2004) Nitrogen mineralization: challenges of a changing paradigm. Ecology 85, 591-602.

Schurig-Briccio, L.A., Venkatakrishnan, P., Hemp, J., Bricio, C., Berenguer, J., Gennis, R.B. (2013) Characterization of the nitric oxide reductase from Thermus thermophilus. Proceedings of the National Academy of Sciences 110, 12613-12618.

Seitzinger, S., Harrison, J.A., Böhlke, J.K., Bouwman, A.F., Lowrance, R., Peterson, B., Tobias, C., Drecht, G.V. (2006) Denitrification across landscapes and waterscapes: a synthesis. Ecological applications 16, 2064-2090.

Shade, A., Gilbert, J.A. (2015) Temporal patterns of rarity provide a more complete view of microbial diversity. Trends in microbiology 23, 335-340.

751 Shannon, C.E. (1948) A mathematical theory of communication. The Bell system technical journal 27,
752 379-423.
753 Sparacino-Watkins, C., Stolz, J.F., Basu, P. (2014) Nitrate and periplasmic nitrate reductases. Chemical
754 Society Reviews 43, 676-706.
755 Templer, P.H., Mack, M.C., Iii, F.S.C., Christenson, L.M., Compton, J.E., Crook, H.D., Currie, W.S.,
756 Curtis, C.J., Dail, D.B., D'Antonio, C.M. (2012) Sinks for nitrogen inputs in terrestrial ecosystems: a
757 meta-analysis of 15N tracer field studies. Ecology 93, 1816-1829.
758 Tomczak, M., Tomczak, E. (2014) The need to report effect size estimates revisited. An overview of some
759 recommended measures of effect size.
760 Tu, Q.C., Lin, L., Cheng, L., Deng, Y., He, Z.L. (2019) NCycDB: a curated integrative database for fast
761 and accurate metagenomic profiling of nitrogen cycling genes. Bioinformatics 35, 1040-1048.
762 Wei, T., Simko, V., Levy, M., Xie, Y., Jin, Y., Zemla, J. (2017) Package 'corrplot'. Statistician 56, e24.
763 Yoon, S., Cruz-García, C., Sanford, R., Ritalahti, K.M., Löffler, F.E. (2015) Denitrification versus
764 respiratory ammonification: environmental controls of two competing dissimilatory NO3−/NO2−
765 reduction pathways in Shewanella loihica strain PV-4. The ISME journal 9, 1093-1104.
766 Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.Y. (2017) ggtree: an R package for visualization and
767 annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and
768 Evolution 8, 28-36.
769