NAR Genomics and Bioinformatics, 2024, **6**, lqae063
https://doi.org/10.1093/nargab/lqae063
Advance access publication date: 6 June 2024
Standard Article

NF*ix*DB (Nitrogen Fixation DataBase)—a comprehensive integrated database for robust 'omics analysis of diazotrophs

Madeline Bellanger[1,2], Jose L. Figueroa III[1,2], Lisa Tiemann[3], Maren L. Friesen[4] and Richard Allen White III[1,2,*]

[1]North Carolina Research Campus (NCRC), Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, 150 N Research Campus Dr, Kannapolis, NC 28081, USA
[2]Computational Intelligence to Predict Health and Environmental Risks (CIPHER), Department of Bioinformatics and Genomics, The University of North Carolina at Charlotte, 9201 University City Blvd, Charlotte, NC 28223, USA
[3]Department of Plant, Soil and Microbial Sciences, Michigan State University, Plant and Soil Sciences Building, 1066 Bogue St Room A286, East Lansing, MI 48824, USA
[4]Department of Plant Pathology, Washington State University, Clark Hall, 2040 Ellis Way, Pullman, WA 99163, USA

[*]To whom correspondence should be addressed. Tel: +1 704 687 8792; Email: rwhit101@charlotte.edu

Abstract

Biological nitrogen fixation is a fundamental biogeochemical process that transforms molecular nitrogen into biologically available nitrogen via diazotrophic microbes. Diazotrophs anaerobically fix nitrogen using the nitrogenase enzyme which is arranged in three different gene clusters: (i) molybdenum nitrogenase (*nifHDK*) is the most abundant, followed by it's alternatives, (ii) vanadium nitrogenase (*vnfHDK*) and (iii) iron nitrogenase (*anfHDK*). Multiple databases have been constructed as resources for diazotrophic 'omics analysis; however, an integrated database based on whole genome references does not exist. Here, we present NF*ix*DB (Nitrogen Fixation DataBase), a comprehensive integrated whole genome based database for diazotrophs, which includes all nitrogenases (*nifHDK*, *vnfHDK*, *anfHDK*) and nitrogenase-like enzymes (e.g. *nflHD*) linked to ribosomal RNA operons (16S–5S–23S). NF*ix*DB was computed using Hidden Markov Models (HMMs) against the entire whole genome based Genome Taxonomy Database (GTDB R214), providing searchable reference HMMs for all nitrogenase and nitrogenase-like genes, complete ribosomal RNA operons, both GTDB and NCBI/RefSeq taxonomy, and an SQL database for querying matches. We compared NFixDB to *nifH* databases from Buckley, Zehr, Mise and FunGene finding extensive evidence of *nifH*, in addition to *vnfH* and *nflH*. NF*ix*DB contains >4000 verified *nifHDK* sequences contained on 50 unique phyla of bacteria and archaea. NF*ix*DB provides the first comprehensive nitrogenase database available to researchers unlocking diazotrophic microbial potential.

Introduction

Biological nitrogen fixation (BNF) is an ancient biogeochemical process on Earth, which is the conversion of atmospheric nitrogen (N_2) into fixed biologically available nitrogen as ammonium (NH_3), completed by specialized microbes known as diazotrophs (1). Nitrogen is essential to all life on the planet, and required for amino acid and nucleic acid synthesis, yet prior to the emergence of the enzyme nitrogenase, elemental nitrogen on the early Earth could only be fixed by lighting (2). Prior to industrialization, the bioavailable nitrogen supplied by crop ratios required to support the ecosystem productivity was produced almost solely via biological nitrogen fixation by diazotrophs (3).

For many decades it was thought that only symbiotic BNF, in which bacteria colonize specialized plant structures, provided significant amounts of ecosystem nitrogen, since high energy demands limit BNF to circumstances with adequate supplies of carbon (3). In recent years, there has been a growing realization that free living nitrogen fixation (FLNF) can provide fixed nitrogen at rates equal to or greater than symbiotic nitrogen fixation and may be the dominant source of new nitrogen inputs to many terrestrial ecosystems (4).

The nitrogenase enzyme as three variations which include different complex metal clusters that are oxygenic-sensitive metalloenzymes: (i) molybdenum-dependent nitrogenase (*nifHDK*), which is the most abundant, followed by (ii) vanadium-dependent nitrogenase (*vnfHDK*) and (iii) iron nitrogenase (*anfHDK*). All nitrogenases contain both catalytic and biosynthetic genes within the nitrogenase gene cluster. The catalytic genes all contain a dinitrogenase reductase, the *H* gene (e.g. *nifH*), which functions as an ATP-dependent electron donor, and a metalloenzyme heterotetramer of *D* and *K* genes, which are metal dependent iron protein alpha chain and iron protein beta chain (5). The biosynthetic gene cluster includes *nifB*, *nifE,* and *nifN*, which are required for FeMo-co biosynthesis, with *nifB*, *nifU*, *nifS*, *nifV* and *nifM* are also required by the alternative nitrogenases (5). A subset of diazotrophs also contain alternative nitrogenases, *vnf* or *anf*. The *vnf* gene cluster contains the iron-vanadium cofactor, while the *anf* gene cluster contains the iron-iron cofactor (6). It is still widely debated which nitrogenase emerged first (7). Nitrogenase-like protoenzymes evolved first in methanogens for F_{430} cofactor biosynthesis and are known as Ni-sirohydrochlorin- a,c- diamide reductive cyclase

Received: March 11, 2024. **Revised:** May 16, 2024. **Editorial Decision:** May 22, 2024. **Accepted:** May 23, 2024

(*nfl*) (8,9). The *nfl* enzymes are ubiquitous in diazotrophic prokaryotes (8,9). Bacteriochlorophyll (*Bch*) and Chlorophyll (*Chl*) biosynthesis (gene light-independent protochlorophyllide reductase) evolved from *nfl*, and are also nitrogenase-like homologs (related to cobalamin biosynthesis, F_{430} cofactor biosynthesis, and biosynthesis of chlorophyll and bacteriochlorophyll) (10).

A major gap in the current literature is the lack of a comprehensive database of nitrogenase enzymes that is rooted in whole-genomes, and thus, our ability to define diazotroph phylogeny or infer metabolic properties of genomes that are able to function effectively as diazotrophs is severely limited. The original nitrogenase databases were based on amplicon sequences that were not complete, due to high sequencing costs for whole genomes and lack of reference genomes (11,12). The most current *nifD* and *nifH* database from FunGene contains 19,514 *nifH* and 10,482 *nifD* sequences and alternative nitrogenases are not well defined (13). FunGene is currently no longer available as of 2022, as the website is no longer functional (13). The Buckley and Zehr lab groups have *nifH* specific databases publicly available on their groups' websites, however they have not been updated since 2012 and 2017, respectively (11,12). Mise *et al.* classified *nifH* sequences and compiled this information into a database, but no information on alternative nitrogenases is available (14). To address this, a novel database, NF*ix*DB ('Nitrogen Fixation DataBase'), was created. The inclusion of the *nifDK* genes and the alternative nitrogenases in a new database would be the first comprehensive collection of this data. Compiling sequences with connecting rRNA marker (16S–5S–23S) databases to nitrogenase genes via complete genomes will provide an extensive database that can be the foundation for the current and future studies of nitrogen fixation.

Materials and methods

Genome curation and HMM construction

An overview of the methods can be seen in Figure 1. Initial seed sequences for *nifHDK*, *anfHDK*, *vnfHDK*, *nflHD* and *ChlBIN* were manually curated. Genomes for the initial seed sequences were selected with the following rules: (i) high quality genomes, (ii) free of contamination via checkM, and/or (iii) physiological validation (15). All *nifHDK* and/or alternative nitrogenase gene clusters (i.e. *anf* and *vnfHDK*) had to be physically arranged together as a clustered block of genes not far away as 'pseudo-*nifH*' described by Mise *et al.* (14). The sequences were locally aligned using MAFFT (16). An HMM of each seed sequence was created using HMMER's hmmbuild (17), then combined together to make a concatenated file of the *nifHDK*, *anfHDK*, and *vnfHDK* HMMs and a concatenated file of the *nflHD* and *ChlBIN* HMMs. Using HMMER's hmmsearch (17), each genome in the release 214 of the Genome Taxonomy Database (GTDB) (18) was examined for the presence of nitrogen fixation enzymes. Over 80,000 representative genomes were analyzed to create NF*ix*DB.

Taxonomic assignment linkage of GTDB-NCBI to NFixDB

The taxonomy of each protein sequence was identified using both GTDB and the National Center for Biotechnology and Information (NCBI)'s taxonomic classifications found within GTDB's metadata files, with the NCBI taxonomy ID also being included (19). Additionally, an *E*-value cutoff was established at $<9.9 \times 10^{-10}$ for HMM searching with HMMER. Any entry with an *E*-value lower than the cutoff of 9.9×10^{-10} is placed into the evalue_taxonomy TSV file (Zenodo). From this TSV, the best result for each protein sequence was found and placed into the tophits TSV file (Supplementary Table S1). Additionally, another *E*-value cutoff of $<9.9 \times 10^{-15}$ and a bitscore cutoff value of >50 was established for updated seed sequence generation. These seed sequences also had an alignment length requirement of >125.

We applied a top hit approach against our HMMs for classification of *nif*/*anf*/*vnfHDK*, *nflDH*, or *chl*/*bchBLN*. For a hit to be considered for its classification it had to: (i) be the top hit via *E*-value and bitscore, (ii) be unique to that genome accession, (iii) have local genome proximity near its assigned cluster (e.g. *nifH* near a *nifDK*) and (iv) all genes within the cluster had to be present to be considered valid (e.g. *HDK* had to be present for *nif*/*anf*/*vnf*). For instance, a genome would need to contain *nifH*, *nifD*, and *nifK* to pass this qualification. Hits that were not near proximity of other genes (e.g. an *H* without proximity to a *DK*) or if one or more genes were missing (e.g. having an *H* but no *DK*) were considered pseudo-nitrogenases and removed. The sequences that fit within these parameters were placed into the topfasta TSV file (Supplementary Table S1). Both the tophits TSV and the topfasta TSV were examined to find genomes that contained all of the genes considered in the gene cluster. The tophits genomes that passed were placed into the filteredhits TSV and the topfasta genomes that passed were placed into the filteredfasta TSV (Supplementary Table S1). The filteredfasta TSV was used to create an updated fasta file for each seed sequence. Multiple iterations were found to be unnecessary and led to bias towards similar genes (i.e. *nifH* mistaken for *vnfH*).

Ribosomal RNA operon database construction

The ribosomal RNA operons (16S–5S–23S) linked to each genome were identified using barrnap (20).

Database availability

The final database can be found in the NF*ix*DB TSV or as an SQL database, both on Zenodo, with each TSV mentioned being included as an SQL table. All genomes that were identified as containing nitrogenase, alternative nitrogenase, or pseudo-nitrogenase can be found on Zenodo (DOI: 10.5281/zenodo.10950414). The code to generate the data is on GitHub (github.com/raw-lab/NFixDB).

Cross database comparisons

Other *nifH* databases from Buckley (11), Zehr (12), FunGene (13) and Mise (14) were clustered using CD-HIT (21) at 100%, 99% and 97% similarity. The external databases' representative sequences obtained from CD-HIT clustering at 97% similarity were used to analyze the NF*ix*DB HMMs and test for validity. Cutoffs for *E*-value $<9.9 \times 10^{-10}$ and an alignment length >150 amino acids were put in place, with results being stored in the oDB class TSV (Supplementary Table S2). The best result for each sequence ID was identified and stored in the oDB hits TSV for all databases, along with a separate TSV for each individual database (Supplementary Table S2). As a secondary confirmation, global alignments were performed using SWORD (22), with the representative sequences for clusters at 97% similarity as the queries and the final seed

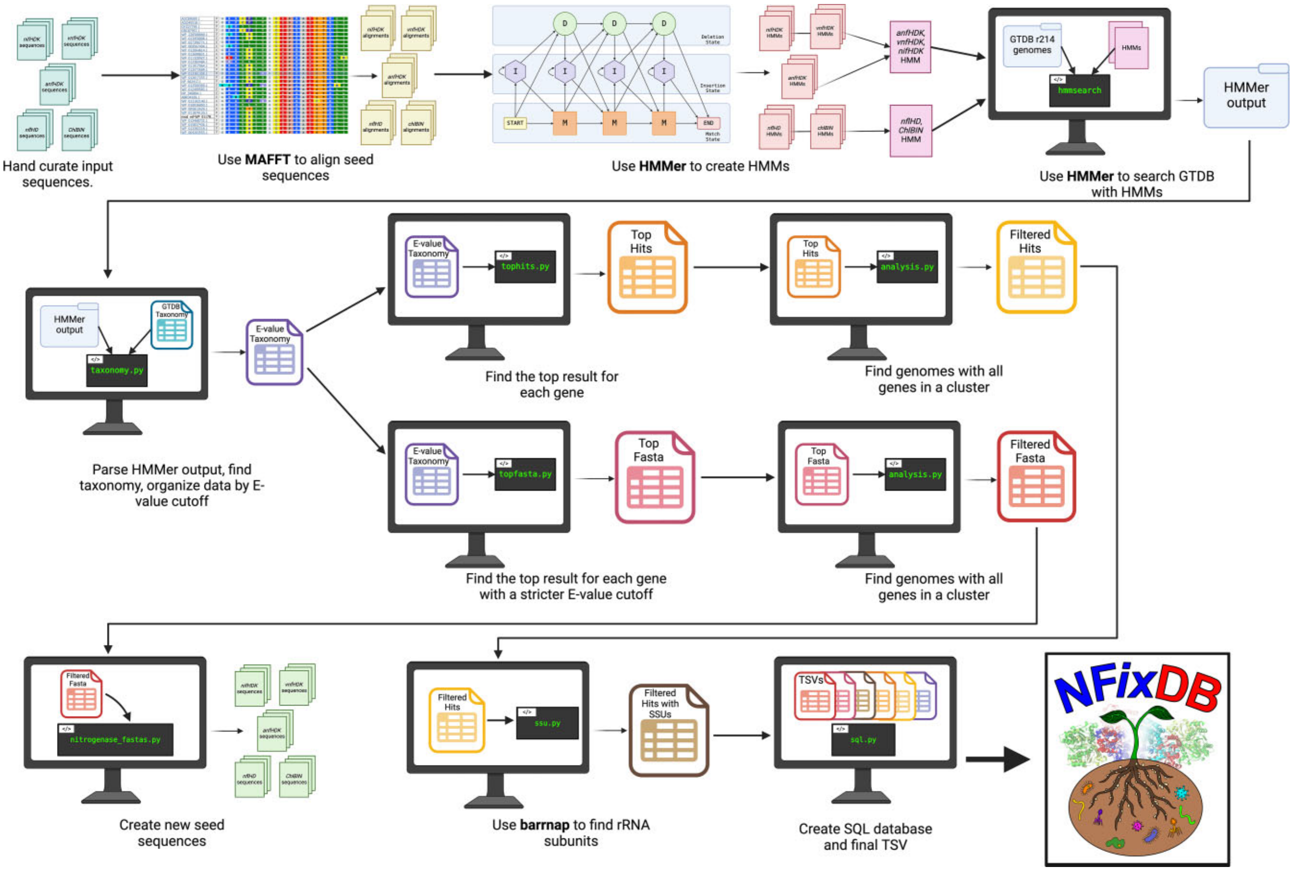

Figure 1. Flow graph of NF*ix*DB. Initial sequences were manually curated. Each gene was aligned using MAFFT. Then, an HMM was created with HMMer to search through all of the genomes in GTDB r214. The output was searched to find the best result for both the gene and the sequence ID. The sequences that occurred with all genes in a gene cluster were compiled together to make the new seed sequences.

sequences from NF*ix*DB as the databases. This ensures that the final seed sequences produced from NF*ix*DB are accurate. Results that had an alignment length $\geq$220 amino acids, $\geq$50% identity and E-value $\leq 9.9 \times 10^{-15}$ were analyzed to find the best result for each sequence ID. These entries were stored in separate TSVs for each database (Supplementary Table S3).

A significant increase in the number of sequences found in each gene cluster was observed from the initial seeds to the final seeds (Supplementary Table S4, Wilcoxon Rank Sum, $P < 0.01$). The seed sequences initially used were hand curated, making an uneven amount within each gene cluster. The production of the final seed sequences ensured an equal amount of sequences within each gene cluster (i.e. *nifH*, *nifD* and *nifK* all have the same amount of sequences), and thus an equal amount in each new seed sequence FASTA. We recommend using our pre-built HMMER HMMs with the following cut-offs: alignment length $\geq$220 amino acids, $\geq$50% identity, and E-value $\leq 9.9 \times 10^{-15}$ for functional annotation of nitrogenases, alternative nitrogenases, and pseudo-nitrogenases. All files are available for the seed and final sequences for standard alignment based tools (i.e. non-HMM based) (Supplementary Table S5).

Results

Overall, NF*ix*DB resulted in the identification of >4,000 *nifHDK* genes, with an average length of 271 amino acids for *nifH*, 474 amino acids for *nifD* and 472 amino acids for *nifK* (Supplementary Figure S1-S4). Of all the sequences found to be from Proteobacteria, >50% were the *nifHDK* genes (Figure 2). Alphaproteobacteria and Deltaproteobacteria were the most common classes found in the *nifHDK* genes (Figure 2). In addition to the *nifHDK* genes identified, more than 250 *anfHDK* genes were found, with an average length of 270 amino acids for *anfH*, 505 amino acids for *anfD*, and 452 amino acids for *anfK* (Supplementary Figure S1). As with the *nifHDK* genes, the most common phyla found among the *anfHDK* genes were Proteobacteria (Figure 2). The most common classes in the *anfHDK* genes were Alphaproteobacteria and Gammaproteobactiera (Figure 2). Among the approximately 60 *vnfHDK* genes accurately identified, only three phyla were found (in order of most to least common): Proteobacteria, Cyanobacteria, and Firmicutes (Figure 2). Within those phyla, four classes were identified (in order of most to least common): Gammaproteobacteria, Alphaproteobacteria, Betaproteobacteria, and Bacilli (Figure 2). The lengths of *vnfH* averaged 270 amino acids, *vnfD* averaged 447 amino acids, and *vnfK* averaged 459 amino acids (Supplementary Figure S1).

Of the 181 phyla analyzed, only 16% were found to have all the catalytic nitrogenase genes (*nifHDK* and/or the alternatives) (Figure 2). The majority of phyla with diazotrophs were Proteobacteria at more than 50%. Cyanobacteria held roughly 15% of organisms with diazotrophic activity (Figure 2).

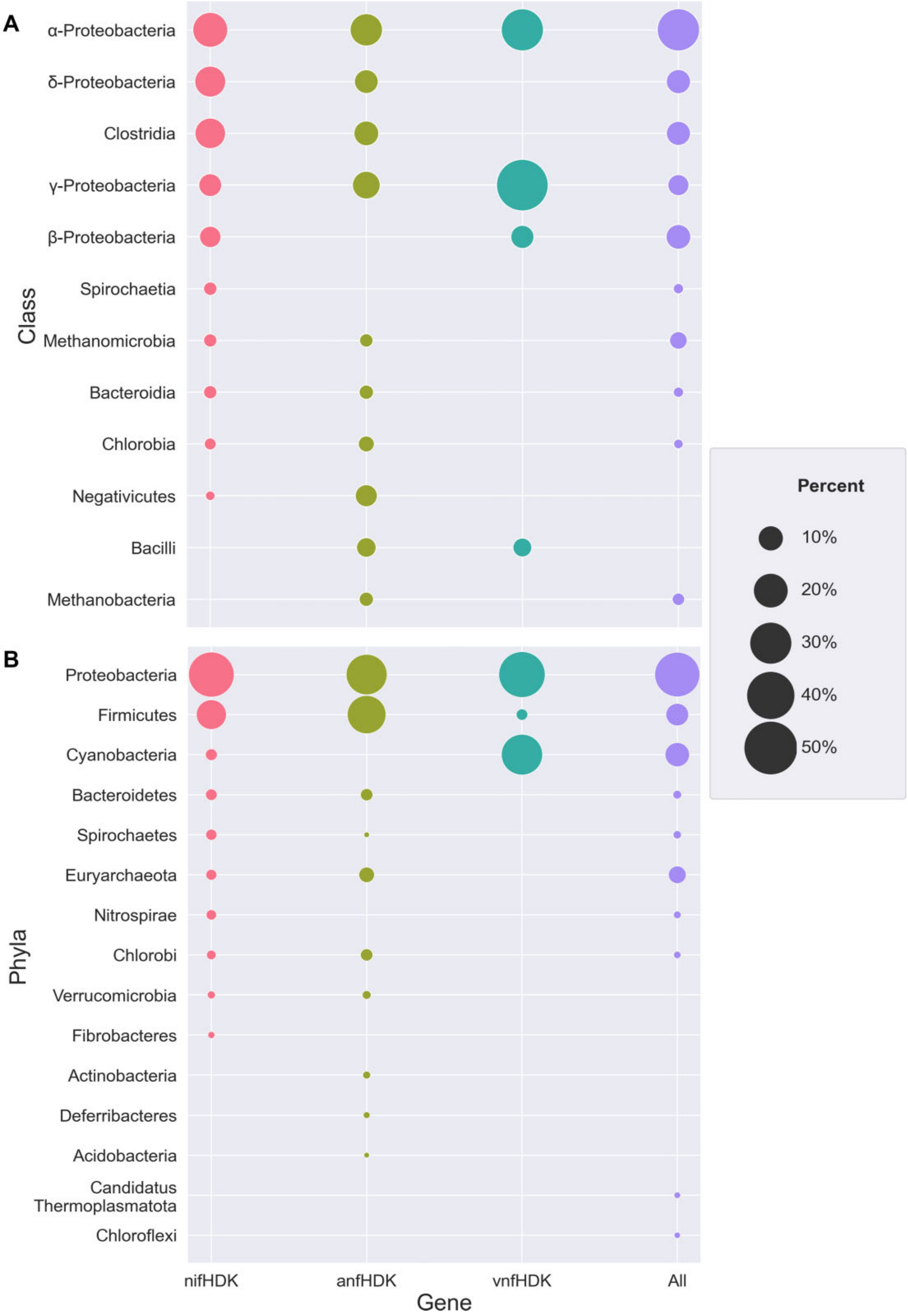

Figure 2. NF*ix*DB taxonomic classification at phyla and class level. The classes or phyla identified with the *nifHDK* genes are shown in pink. The classes or phyla identified with the *anfHDK* genes are shown in green. The classes or phyla identified with the *vnfHDK* genes are shown in blue. The top 10 overall classes or phyla are shown in purple. The size of each dot represents the percentage of occurrences among all of the classes or phyla identified. (**A**) Top 10 classes for each nitrogenase gene group. (**B**) Top 10 phyla for each nitrogenase gene group.

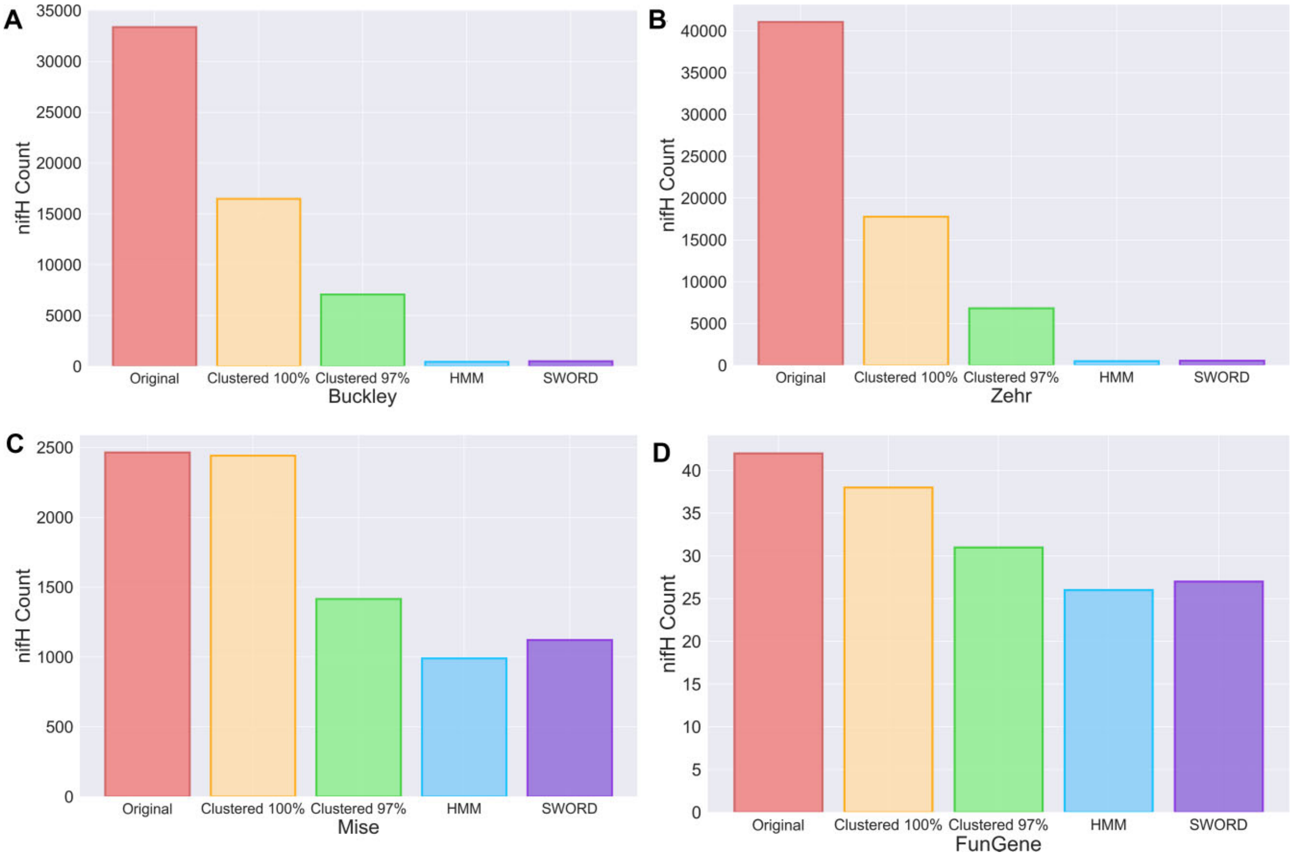

Figure 3. NF*ix*DB database comparison against Buckley, Zehr, Mise and FunGene databases. The original count is shown in pink. Clustering at 100% similarity is shown in yellow and 97% similarity is shown in green. The count after classifying with our HMMs is shown in blue. The count after classifying with SWORD is shown in purple. (**A**) Counts of sequences found in the Buckley database at each step. (**B**) Counts of sequences found in the Zehr database at each step. (**C**) Counts of sequences found in the Mise database at each step. (**D**) Counts of sequences found in the FunGene database at each step.

We compared other *nifH* databases against NF*ix*DB, which include Buckley, Zehr, Mise, and FunGene. Prior to comparison we clustered to remove duplicates using CD-HIT (100%, 99% and 97% similarity) then compared with HMMER3/HMMs and SWORD via global alignment (Figure 4, Supplementary Figures S5 and S6). Clustering at 100% similarity resulted in a more than 50% decrease in the amount of sequences in both the Buckley and Zehr databases (Figure 3, Supplementary Tables S1–S7). The Mise and FunGene databases both had a less than 10% decrease per gene (Figure 3, Supplementary Tables S6 and S7). Clustering down to 99% similarity resulted in a ∼60% decrease in the Buckley and Zehr databases (Figure 3, Supplementary Tables S6 and S7). In the Mise and FunGene databases, the clusters resulted in a ∼20% decrease per gene (Figure 3, Supplementary Tables S6 and S7). When clustering at 97%, the Buckley and Zehr databases decreased by ∼80%, leaving roughly 7000 sequences in each database (Figure 3, Supplementary Tables S6 and S7, Supplementary Tables S5 and S6). The Mise and FunGene databases also saw decreases, ranging from 25% to 45% per gene (Figure 3, Supplementary Tables S6 and S7).

When analyzing with both our HMMs and SWORD, these databases were found to have similar estimations of *nifH* sequences to what was originally estimated after clustering at 97% similarity (Figure 3, Table 1). The Mise database is split into three classifications of *nifH*. All three classifications were analyzed and 97% clustering resulted in 1416 sequences total (Supplementary Tables S6, S8, Supplementary Figures S5 and S6). Ten of those sequences did not pass our filtering, with an *E*-value $<9.9 \times 10^{-15}$ and an alignment length >150 amino acids. Our HMMs revealed that the majority of sequences were found to be *nifH*. Roughly 30% of sequences analyzed were found to be *vnfH* or *nflH*. The SWORD analysis showed similar results, except fewer *vnfH* sequences and more *nifH* were identified (Figure 3, Table 1).

The Zehr database clustered down to 6831 sequences (Supplementary Tables S6 and S8). Only 974 of those sequences passed our filters for HMM analysis, with an *E*-value $<9.9 \times 10^{-15}$ and an alignment length >150 amino acids. Our HMMs revealed that more than 50% of the sequences analyzed were classified as *nifH*. Roughly 17% of the sequences were classified as *anfH* and *vnfH*. Nearly 30% of sequences were identified as *nflH* and a small subset of sequences were identified as *ChlI* (<1%). The SWORD analysis again showed similar results, with more *nifH* and less *nflH* and *vnfH* being identified over 896 sequences (Figure 3, Table 1).

The Buckley database was clustered down to 7,062 *nifH* sequences (Supplementary Tables S6 and S8). Only 627 of those sequences passed our filters with an *E*-value $<9.9 \times 10^{-15}$ and an alignment length >150 amino acids. More than 70% of those were classified as *nifH* in our analysis. Nearly 15% of sequences were classified as *vnfH*. The rest of the sequences were found to be *anfH* (∼5%), *nflH* (∼8%) and *ChlI* (∼0.3%).

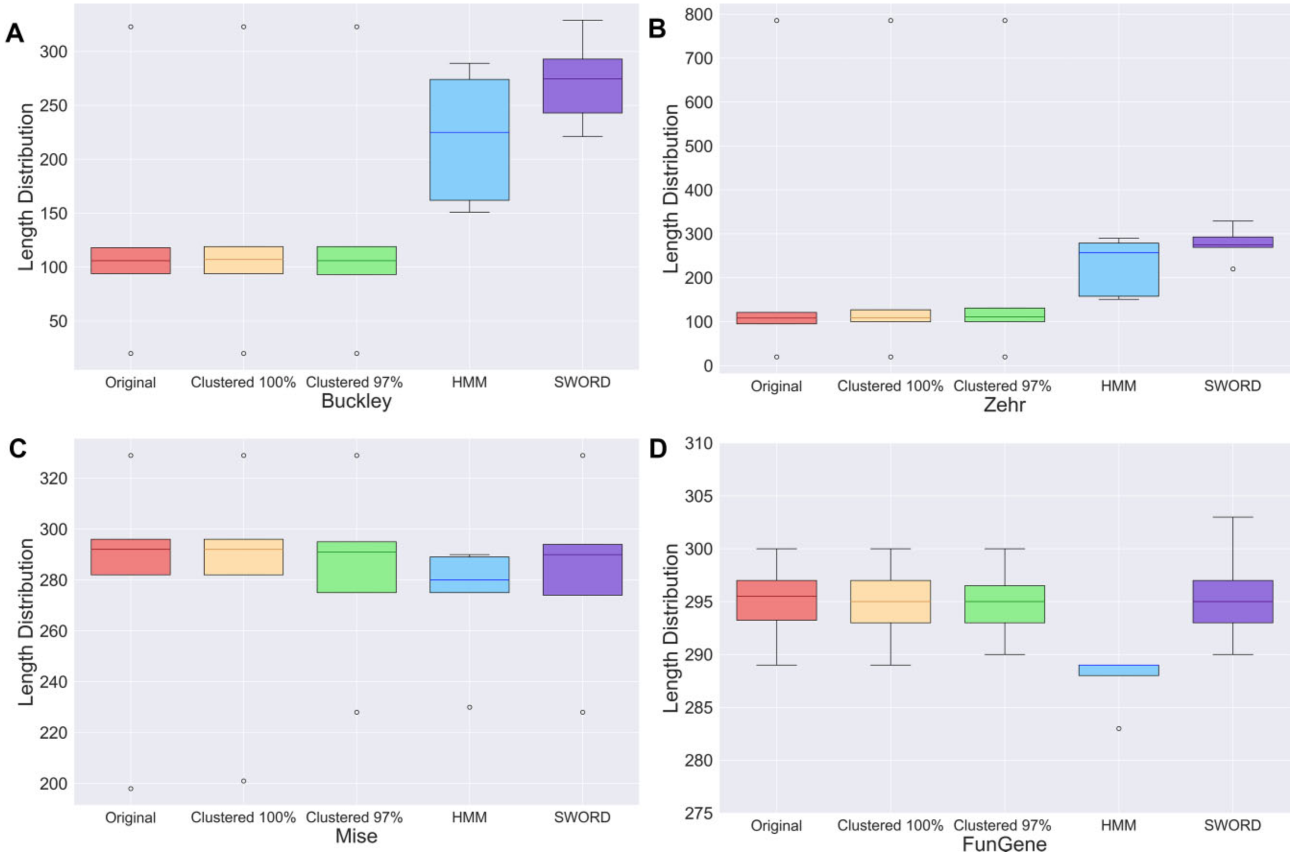

Figure 4. NF*ix*DB length distribution comparison against Buckley, Zehr, Mise, and FunGene databases. The box length corresponds to the sequence lengths that fall within the first quartile to third quartile range. The whisker length corresponds to the largest/smallest sequence length within 1.5× the interquartile range from the box. The original length distribution is shown in pink. Clustering at 100% similarity is shown in yellow and 97% similarity is shown in green. The length distribution after classifying with our HMMs is shown in blue. The length distribution after classifying with SWORD is shown in purple. (**A**) Length distribution of sequences found in the Buckley database at each step. (**B**) Length distribution of sequences found in the Zehr database at each step. (**C**) Length distribution of sequences found in the Mise database at each step. (**D**) Length distribution of sequences found in the FunGene database at each step.

The SWORD analysis resulted in 606 sequence hits after filtering. Of those, >80% were identified as *nifH*. Nearly 5% were identified as *vnfH*. The rest were classified similar to the HMM results (Figure 3, Table 1).

FunGene clustered at 97% similarity led to 31 *nifH* seed sequences, 190 *nifD* seed sequences and 20 *vnfD* seed sequences (Supplementary Tables S6 and S8). It is important to note that FunGene does contain many more sequences for these genes. For the purpose of this study, only the seed sequences identified by FunGene were analyzed. After our analysis with both HMMs and SWORD, it was found that all of the *vnfD* sequences in FunGene were accurately classified. The majority of FunGene's *nifD* sequences were classified as *nifD* using our HMM and SWORD analyses, with ∼1.5% of them classified as *vnfD* when using HMMs. The *nifH* sequence results were similar using both HMMs and SWORD, with one more sequence being identified as *nifH* over *vnfH* when using SWORD. There were roughly 85% of sequences correctly identified as *nifH* and roughly 15% of sequences identified as *vnfH* (Figure 3, Table 1).

Discussion

Generally, the measurement of *nifH* has been the gold standard for quantifying the diversity, abundance, presence, and potential activity of diazotrophs. In the era of highly cost effective next generation sequencing and high throughput quantitative PCR measurements, understanding the quality of the resulting *nifH* databases is critical to agriculture, food security and bioenergy applications, which are all limited by nitrogen. Our analysis revealed that whole genome based curation provides a framework to further exploration of diazotrophs, highlighting the need to include alternative nitrogenases and nitrogenase-like genes.

Beginning with hand curated seed sequences, an alignment and an HMM was made for each gene (Supplementary Figure S7). Every genome in GTDB was searched with each HMM to identify any potential nitrogenase genes. Further processing was done for each hit to ensure that only the top result for each accession number was kept. After finding genomes with all genes in a cluster present (i.e. *nifH*, *nifD* and *nifK* present), new seed sequences were gathered to create NF*ix*DB.

The misclassification of *nifH* has had a large impact on the understanding of nitrogenase. In the past, alternative nitrogenases and pseudo-nitrogenases have not been classified well, despite *nifH*, *vnfH* and *nflH* being very closely related. Throughout our database curation, *nifH* and *vnfH* were revealed to be continually mislabeled. For instance, *Azotobacter vinelandii DJ* (GCF_000021045.1) contains three genes labeled as *nifH*. One of these three genes is the true

Table 1. NF*ix*DB database comparison summary table

Database	HMM analysis		Sword analysis	
Original Gene Detected Gene	Counts	% of Total	Counts	% of Total
Buckley Total	**627**		**606**	
nifH	**627**		**606**	
anfH	34	5.42%	33	5.45%
ChlI	2	0.32%	2	0.33%
nflH	54	8.61%	41	6.77%
nifH	444	70.81%	500	82.51%
vnfH	93	14.83%	30	4.95%
FunGene Total	**241**		**241**	
nifD	**190**		**190**	
nifD	187	98.42%	190	100.00%
vnfD	3	1.58%	0	N/A
nifH	**31**		**31**	
nifH	26	83.87%	27	87.10%
vnfH	5	16.13%	4	12.90%
vnfD	**20**		**20**	
vnfD	20	100.00%	20	100.00%
Mise Total	**1406**		**1406**	
nifH	**1406**		**1406**	
nflH	113	8.04%	111	7.89%
nifH	990	70.41%	1122	79.80%
vnfH	303	21.55%	173	12.30%
Zehr Total	**974**		**896**	
nifH	**974**		**896**	
anfH	59	6.06%	58	6.47%
ChlI	2	0.21%	7	0.78%
nflH	285	29.26%	224	25.00%
nifH	517	53.08%	557	62.17%
vnfH	111	11.40%	50	5.58%

Counts of the number of genes found in all of the outside databases after clustering at 97% similarity and classifying using both SWORD and HMMs. The gene that is shown in bold is the original gene identification from the corresponding database. All genes below that are genes that were identified through our classification.

nifH (WP_012698831.1). The other two genes are *anfH* (WP_012703362.1) and *vnfH* (WP_012698955.1). Within *A. vinelandii DJ*, there are no annotated duplicates of a true *nifH*. Issues like this have led to *nifH* databases containing large amounts of *vnfH* and *nflH* (>10%).

Additionally, HMMs have a difficult time distinguishing between extremely closely related genes, like *nifH* and *vnfH*, however, a global alignment provides validation of the HMM results, when these methods are combined together. Recent advances in machine and deep learning such as convolutional neural networks (CNNs) could be applied to enhance detection of nitrogenases within genomes or metagenomic data.

There has been a lack of an all encompassing database for diazotrophs that contains alternative nitrogenases, which could lead to misinterpretation of nitrogenase diversity, presence, and activity within ecosystems. NF*ix*DB provides the first comprehensive whole genome based database for nitrogenase, alternative nitrogenases, and nitrogenase-like genes. Through NF*ix*DB, we provide a fundamental framework to unravel the diversity, presence, and potential activity of diazotrophs across the tree of life.

Data availability

NF*ix*DB scripts are written in Python and distributed under a BSD license. The source code and database of NFixDB is freely available at https://github.com/raw-lab/NFixDB. Scripts, data, and SQL database are available on Zenodo (DOI: 10.5281/zenodo.10950414) and GitHub (https://github.com/raw-lab/NFixDB).

Contributing to NF*ix*DB and Fungene

NF*ix*DB as a community resource has recently acquired Fungene (13). We welcome contributions of other experts, expanding annotation of all domains of life (viruses, bacteria, archaea, eukaryotes). Please send us an issue on our NF*ix*DB GitHub (https://github.com/raw-lab/NFixDB/issues). We will fully annotate your genome, add suggested pathways/metabolisms of interest, and make custom HMMs to be added to NF*ix*DB and FunGene. Also, NF*ix*DB is available within the metaomics tool MetaCerberus (https://github.com/raw-lab/MetaCerberus) (23). A reference tree is present for NF*ix*DB for the initial seeds (Supplementary Figure S7).

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

We acknowledge the support of the following units of the University of North Carolina at Charlotte: the College of Computing and Informatics, the Bioinformatics Research Center, the Department of Bioinformatics and Genomics, Research and Economic Development, Academic Affairs, University Research Computing, and North Carolina Research Campus.

Funding

Bellanger M, Figueroa III, J.L, and White III, R.A. are supported by a UNC Charlotte start-up package. All authors were further supported by United States Department of Agriculture (USDA) Agriculture and Food Research Initiative (AFRI) project 1030783.

Conflict of interest statement

The authors declare that there are no conflicts of interest. RAWIII is the CEO of RAW Molecular Systems (RAW), LLC, but no financial, IP, or others from RAW LLC were used or contributed to the study.

References

1. Garcia,A.K., McShea,H., Kolaczkowski,B. and Kaçar,B. (2020) Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization. *Geobiology*, **18**, 394–411.
2. Mancinelli,R.L. and McKay,C.P. (1988) The evolution of nitrogen cycling. *Orig. Life Evol. Biosph.*, **18**, 311–325.
3. Goyal,R.K., Schmidt,M.A. and Hynes,M.F. (2021) Molecular biology in the improvement of biological nitrogen fixation by Rhizobia and extending the scope to cereals. *Microorganisms*, **9**, 25.
4. Van Langenhove,L., Depaepe,T., Vicca,S., van den Berge,J., Stahl,C., Courtois,E., Weedon,J., Urbina,I., Grau,O., *et al.* (2020) Regulation of nitrogen fixation from free-living organisms in soil and leaf litter of two tropical forests of the Guiana shield. *Plant Soil*, **450**, 93–110.

5. Burén,S., Jiménez-Vicente,E., Echavarri-Erasun,C. and Rubio,L.M. (2020) Biosynthesis of nitrogenase cofactors. *Chem. Rev.*, **120**, 4921–4968.
6. Schwartz,S.L., Garcia,A.K., Kaçar,B. and Fournier,G.P. (2022) Early nitrogenase ancestors encompassed novel active site diversity. *Mol. Biol. Evol.*, **39**, msac226.
7. Mus,F., Colman,D.R., Peters,J.W. and Boyd,E.S. (2019) Geobiological feedbacks, oxygen, and the evolution of nitrogenase. *Free Radic. Biol. Med.*, **140**, 250–259.
8. Boyd,E.S., Anbar,A.D., Miller,S., Hamilton,T.L., Lavin,M. and Peters,J.W. (2011) A late methanogen origin for molybdenum-dependent nitrogenase. *Geobiology*, **9**, 221–232.
9. Boyd,E.S., Costas,A.M., Hamilton,T.L., Mus,F. and Peters,J.W. (2015) Evolution of molybdenum nitrogenase during the transition from anaerobic to aerobic metabolism. *J. Bacteriol.*, **197**, 1690–1699.
10. Staples,C.R., Lahiri,S., Raymond,J., Von Herbulis,L., Mukhophadhyay,B. and Blankenship,R.E. (2007) Expression and association of group IV nitrogenase *NifD* and *NifH* homologs in the non-nitrogen-fixing archaeon *Methanocaldococcus jannaschii*. *J. Bacteriol.*, **189**, 7392–7398.
11. Gaby,J.C. and Buckley,D.H. (2014) A comprehensive aligned *nifH* gene database: A multipurpose tool for studies of nitrogen-fixing bacteria. *Database (Oxford)*, **2014**, bau001.
12. Heller,P., Tripp,H.J., Turk-Kubo,K. and Zehr,J.P. (2014) ARBitrator: A software pipeline for on-demand retrieval of auto-curated nifH sequences from GenBank. *Bioinformatics*, **30**, 2883–2890.
13. Fish,J.A., Chai,B., Wang,Q., Sun,Y., Brown,C.T., Tiedje,J.M. and Cole,J.R. (2013) FunGene: the functional gene pipeline and repository. *Front. Microbiol.*, **4**, 291.
14. Mise,K., Masuda,Y., Senoo,K. and Itoh,H. (2021) Undervalued Pseudo-*nifH* Sequences in Public Databases Distort Metagenomic Insights into Biological Nitrogen Fixers. *mSphere*, **6**, e0078521.
15. Parks,D.H., Imelfort,M., Skennerton,C.T., Hugenholtz,P. and Tyson,G.W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.*, **25**, 1043–1055.
16. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
17. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
18. Parks,D.H., Chuvochina,M., Rinke,C., Mussig,A.J., Chaumei,P.-A. and Hugenholtz,P. (2022) GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic. Acids. Res.*, **50**, D785–D794.
19. Sayers,E.W., Bolton,E.E., Brister,J.R., Canese,K., Chan,J., Comeau,D.C., Connor,R., Funk,K., Kelly,C., Kim,S., *et al.* (2022) Database resources of the national center for biotechnology information. *Nucleic. Acids. Res.*, **50**, D20–D26.
20. Seemann,T. (2018) In: *barrnap : Rapid ribosomal RNA prediction (0.9)*. https://github.com/tseemann/barrnap (31 January 2024, date last accessed).
21. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
22. Vaser,R., Pavlović,D. and Šikić,M. (2016) SWORD-a highly efficient protein database search. *Bioinformatics*, **32**, i680–i684.
23. Figueroa,J.L., Dhungel,E., Bellanger,M., Brouwer,C.R. and White III,R.A. (2024) MetaCerberus: distributed highly parallelized HMM-based processing for robust functional annotation across the tree of life. *Bioinformatics*, **40**, btae119.

Received: March 11, 2024. **Revised:** May 16, 2024. **Editorial Decision:** May 22, 2024. **Accepted:** May 23, 2024