Data in Brief 63 (2025) 112135

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

NEMESISdb: A full length 16S rRNA gene dataset for the detection of human, fish, and crustacean potentially pathogenic bacteria

Son-Hoang Tran[a,b,d,#], Claudia Ximena Restrepo-Ortiz[c,#], Dinh Quang Vu[b], Marc Troussellier[a], Yvan Bettarel[a], Thierry Bouvier[a], Van Ngoc Bui[d,f], Nguyen Hieu Minh[e], Trung Du Hoang[e], Quang Huy Nguyen[b], Jean-Christophe Auguet[a,*]

[a] UMR MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Montpellier, France
[b] University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Hanoi, Vietnam
[c] UMR MARBEC, Univ Montpellier, CNRS, Ifremer, IRD, Sète, France
[d] Institute of Biology (IB), Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Hanoi, Vietnam
[e] Institute of Oceanography, Vietnam Academy of Science and Technology (VAST), 01, Cau Da, Nha Trang, Khanh Hoa, Vietnam
[f] Graduate University of Science and Technology (GUST), Vietnam Academy of Science and Technology (VAST), 18 Hoang Quoc Viet, Hanoï, Vietnam

ARTICLE INFO

Article history:
Received 14 July 2025
Revised 2 September 2025
Accepted 29 September 2025
Available online 6 October 2025

Dataset link: NEMESISdb (Original data)

Keywords:
One Health
Dataset
Pathogenic bacteria
Marine ecosystems, Human, Animal

ABSTRACT

NEMESISdb is a 16S rRNA full length sequence curated dataset designed to enable the identification and tracking of potentially pathogenic bacteria (PPB) for human, fish, and crustacean hosts. It addresses the limited focus on marine and coastal environments as key reservoirs for PPB, where bacteria from diverse sources—terrestrial, marine, and animal—can coexist. Leveraging recent advances in high-throughput sequencing, NEMESISdb provides a robust resource for the detection of PPB in 16S rRNA gene metabarcoding or metagenomic data. The database comprises three datasets corresponding to human, fish, and crustacean hosts, containing 1703, 222, and 64 PPB species, respectively, with

* Corresponding author.
E-mail address: jean-christophe.auguet@cnrs.fr (J.-C. Auguet).
Social media: @jcauguet (J.-C. Auguet)
# Authors contributed equally to the manuscript.

https://doi.org/10.1016/j.dib.2025.112135

a total of over 150,000 16S rRNA full length sequences curated for accuracy. This resource was constructed by extracting sequences from the SILVA 138.2 SSU Ref NR99 database, refining them through a rigorous curation pipeline to ensure taxonomic consistency and eliminate misclassifications. The resulting datasets are optimized for use with popular tools such as BLAST and classifier software, enabling rapid and accurate detection of PPB in metabarcoding and metagenomic data. NEMESISdb supports diverse applications, including pathogen surveillance in aquatic ecosystems, studies on environmental factors influencing PPB dynamics, and the development of targeted strategies for mitigating pathogen impacts in aquaculture. Additionally, it facilitates research within the One Health framework by linking the circulation of PPB across environmental, animal, and human compartments.

Specifications Table

Subject	Microbiology
Specific subject area	Full length 16S rRNA gene sequences pathogenic bacteria dataset
Type of data	Information table, Pathogens lists, Filtered fasta files, Python scripts
Data collection	We constructed a list of pathogenic bacteria for humans, fishes, and crustaceans from various studies and pathogen detection pipeline such as 16SPIP, FAPROTAX, MPD and MBPD. Afterward, full length 16S rRNA gene sequences of each of the pathogenic bacteria of the list was downloaded from the SILVA 138.2 SSU Ref NR99 bacterial database in order to obtain three pathogenic reference datasets for humans, fishes, and crustaceans, respectively. Lastly, each dataset was curated with homemade scripts to remove all sequences wrongly assigned at the species taxonomic level in SILVA 138.2 SSU Ref NR99.
Data source location	Raw data for the construction of the pathogen list came from Zhang et al. [1], Zhang et al. [2] Wardeh et al. [3], Blauwkamp et al. [4], Urban et al. [5], Louca et al. [6], Miao et al. [7] and Xinrun et al. [8]. Full length 16S rRNA gene sequences of each of the pathogenic bacteria of the list were downloaded from the SILVA 138.2 SSU Ref NR99 bacterial database [9].
Data accessibility	Repository name: Zenodo Data identification number: 10.5281/zenodo.16992968 Direct URL to data: https://doi.org/10.5281/zenodo.16992968
Related research article	None

1. Value of the Data

- NEMESISdb is a set of three curated 16S rRNA full length sequence datasets enabling the identification and tracking of potentially pathogenic bacteria (PPB) across human, fish and crustacean hosts and helping reveal factors that influence their dynamics.
- NEMESISdb can be directly and easily used in blast or in classifier softwares for fast detection of PPB in 16S rRNA gene metabarcoding or metagenomic data.
- NEMESISdb could benefit a wide range of stakeholders involved in diseases outbreak prevention and food security (e.g. health agencies, aquaculture and fisheries industries), biodiversity conservation and pathoecology (e.g. researchers and environmental monitoring organizations) and coastal management (e.g. policy makers).

- These datasets can be utilized and reused in several ways to provide further insights in pathogen surveillance by monitoring the dynamics and hotspot of PPB in aquatic environments, in comparative studies aiming to investigate how environmental factors influence pathogen diversity and abundance, in targeted interventions and mitigation strategies by guiding aquaculture management practices, to reduce pathogen impact and in the framework of One Health studies by facilitating the identification of PPB circulating within the environmental, animal and human compartments.

2. Background

Most research on infection diseases has focused on inland systems with comparatively little efforts directed towards marine habitats. However, marine and particularly coastal environments can function as transmission foci for potentially pathogenic bacteria (PPB) because of the concentrated aggregations of bacteria from different sources, both marine and terrestrial, where environmental, human, and/or animal related bacteria can coexist [10,11]. Comprehensive pathogen monitoring in water is difficult to achieve using commonly applied approaches, such as culture-based techniques or quantitative polymerase chain reaction (qPCR), due to their limited throughput [12]. Recent breakthroughs in high-throughput sequencing technologies now allow for the detection of PPB on an unprecedented scale using 16S rRNA gene sequencing [13–18]. The accuracy and breadth of pathogen detection through 16S sequencing largely depend on the reference pathogen database used [8]. However, the datasets needed to precisely identify PPB circulating among the human, marine environment and marine animal compartments accordingly to a One Health framework remain largely underdeveloped. Here, we constructed NEMESISdb, a set of three curated 16S rRNA full-length sequence datasets, allowing the use of both long-read and short-read sequencing across different 16S rRNA gene variable regions to accurately detect PPB. NEMESISdb is a convenient tool for the rapid identification of human, fish, and crustacean PPB in next generation sequencing (NGS) data, supporting key areas such as food safety, epidemic prevention in both livestock and humans, disease detection, and environmental surveillance.

3. Data Description

NEMESISdb [9], available with the following DOI 10.5281/zenodo.16992968, is composed of 14 files and one folder. These include three fasta files containing the full-length 16S rRNA gene sequences of human, fish, and crustacean datasets; three tab-separated text files listing the genus–species pairs of PPB used to construct each dataset; one Excel file providing information on the sources used; and, for each group, an Excel file giving the taxonomic synonyms identified as well as another Excel file listing the species that compose the curated datasets together with their corresponding synonyms. In addition, a GitHub repository is provided containing the PathoLens Python package used to create and curate the datasets. Finally, we provide also in the zip file named “PPB_not_dereplicated.zip” three additional fasta files containing the full-length 16S rRNA gene sequences of human, fish, and crustacean datasets resulting from the application of the PathoLens Python package on the SILVA 138.2 SSU Ref database.

The three files, Human_Pathogen_DB.fasta, Fish_Pathogen_DB.fasta, Crustacean_Pathogen_DB.fasta contain the full-length 16S rRNA gene sequence of PPB for humans, fishes and crustacean respectively. Headers of each sequence within the fasta files correspond to the ACC number followed by the SILVA 138.2 SSU Ref NR99 taxonomy of the sequence from the kingdom to the species level. The datasets contain 8 795, 20 849 and 50 973 16S rRNA gene sequences with an average length of 1479.1 bp, 1491.3 bp, and 1499.4 bp, respectively for crustaceans, fishes and human (Table 1). This number of sequences encompasses 64, 222 and 1703 species of PPB for crustaceans, fishes and human, respectively.

Overall, PPB sequences from the three datasets mainly belonged to the same two phyla namely *Bacillota* and *Pseudomonadota*, which represented on average 50.83 % and 42.66 % of

Table 1
Summary of dataset's properties for PPB retrieved from the SILVA 138.2 SSU Ref NR99 database.

	Crustacean	Fish	Human
Species	64	222	1703
Sequences	8795	20,849	50,973
Length's mean (bp)	1479.1	1491.3	1499.4
Length's sd (bp)	89.5	84.1	79.4

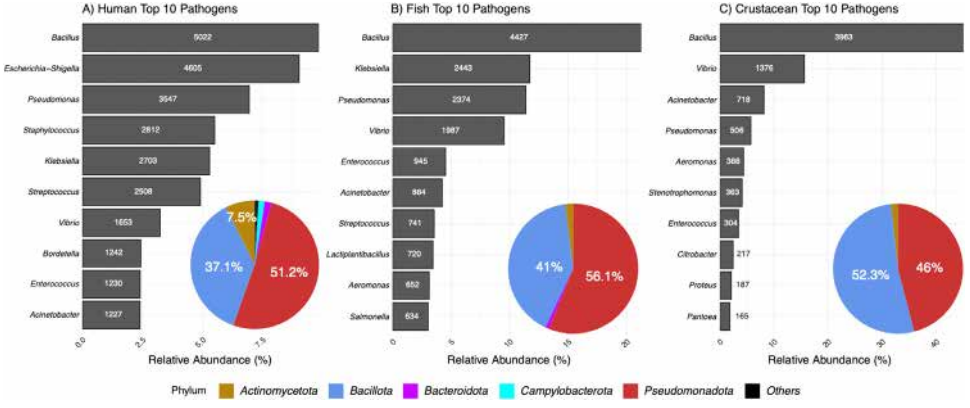

Fig. 1. Taxonomic composition of the three PPB datasets. Barplot represents the contribution of the top ten genera in each dataset. The number of full length 16S rRNA gene sequences in each genus is indicated. Pie chart represents the taxonomic composition of each dataset at the phylum level (The percentage of Phyla higher than 5 % is indicated).

the PPB dataset (Fig. 1). The diversity of PPB sequences was greater in humans, with twelve phyla represented, compared to four and three phyla observed in fishes and crustaceans, respectively. *Bacillus* was the most represented genus in the three datasets and represented up to 45 % of all the PPB sequences in the crustacean dataset. Similarly to *Bacillus*, other genera such as *Pseudomonas, Vibrio, Enterococcus* and *Acinetobacter* were common to the three datasets. As expected, we observed also some differences of composition among the 10 most represented genera of each dataset with notably the presence of *Aeromonas* only in fishes and crustacean datasets while the genera *Escherichia-Shigella, Staphylococcus* and *Bordella* were only present in the human dataset.

The "Pathogen_dataset_sources.xls" file contains 2 sheets indicating the different sources where the PPB derived from (sheet 1) and the list of PPB species extract from each source (sheet 2).

The initial list of PPB species used to generate extended list and extract the full length 16S rRNA gene sequences from the (non redundant) SILVA 138.2 SSU Ref NR99 database is contained in three tab-separated text files containing the genus-species pairs of PPB for each host group: Crustacean_sp_pathogens_list.txt (70 species), Fish_sp_pathogens_list.txt (240 species), and Human_sp_pathogens_list.txt (1942 species)

The Zenodo repository contains a GitHub repository of the PathoLens package, a Python tool designed to filter and curate taxonomic databases. It includes various modules and functions for validating records, which were used in the creation of the three PPB datasets.

Since overrepresentation of sequences in reference databases can impact the accuracy and precision of taxonomy assignment in rRNA studies [19], the three files, Human_Pathogen_DB.fasta, Fish_Pathogen_DB.fasta, Crustacean_Pathogen_DB.fasta have been obtained by applying our PathoLens Python package on the non redundant SILVA 138.2 SSU Ref NR99. However, we also provide an additional zip file in the Zenodo repository, containing the datasets obtained on the complete SILVA 138.2 SSU Ref database. This is important because cer-

Table 2
Summary of dataset's properties for PPB retrieved from the complete SILVA 138.2 SSU Ref database.

	Crustacean	Fish	Human
Species	65	223	1757
Sequences	34,481	80,761	196,770
Unique sequences	26,670	57,663	115,991
Length's mean (bp)	1150.7	1151.8	1158.8
Length's sd ((bp)	63.6	66.3	65.7

tain PPB species or strains share >99 % identity across the full length of their 16S rRNA gene. Such strains are eliminated during dereplication in the SILVA 138.2 SSU Ref NR99 database, which can artificially reduce the apparent richness of the PPB community (Table 2). Moreover, when using classifiers on the non-redundant dataset, these species or strains may be subject to over-classification.

4. Experimental Design, Materials and Methods

4.1. Data acquisition and cleaning

To support the tracking and identification of potentially pathogenic bacteria (PPB) across different hosts, we developed PathoLens v0.1 [20], a custom Python 3.10.9 package tailored for this study. PathoLens integrates modular scripts and functions for automated data retrieval, processing, and curation of reference sequences. The package includes configuration files that define all required dependencies, ensuring reproducibility and ease of use.

The primary focus of this work was to build a curated set of 16S rRNA datasets enabling the tracking of potentially pathogenic bacteria (PPB) across hosts and their rapid detection using BLAST [21] or classifier software. The human PPB list was constructed using a list of pathogenic bacteria for humans from various studies [2–5] and pathogen detection pipeline such as 16SPIP [7], FAPROTAX [6], MPD [1] and MBPD [8] (See "Pathogen_dataset_sources.xls" file for details). The fish and crustacean PPB lists were derived from the study of Wardeh et al. [3]. Crustacean PPB were not explicitly listed in the Wardeh dataset but were grouped under arthropods. To isolate crustacean pathogens, we used the script "ensembl_crustacea.py", included in the PathoLens package. This script queries the Ensembl REST API [22], a comprehensive genome browser that provides various tools such as BLAST, BLAT [23], BioMart [24], and the Variant Effect Predictor (VEP) for all supported species. The script was designed to check if a given species belongs to the Crustacea class, by querying the Ensembl database for taxonomic information and determines whether the species falls under the "Crustacea" class. If it does, the species is labeled as a crustacean in the output. The script reads the input CSV file [Dataset] "SpeciesInteractions_EID.csv", which contains information on host-pathogen interactions [3,25]. Once the list of PPB for humans, fish, and crustaceans was obtained, three tab-separated text files containing the genus-species pairs of PPB for each host group: "Crustacean_sp_pathogens_list.txt", "Fish_sp_pathogens_list.txt" and "Human_sp_pathogens_list.txt" were prepared for further analysis.

Given the dynamic nature of bacterial taxonomy and the fact that databases such as SILVA are not updated synchronously with taxonomic databases like NCBI Taxonomy [26], we performed a thorough synonym search for each genus-species pair in these intermediate lists to maximize sequence recovery. This was done using the script get_sp_synonyms.py, which queries the NCBI Taxonomy database via Bio.Entrez package from Biopython [27]. For each species name, the script retrieves its currently accepted scientific name along with all known synonyms. In cases where no taxonomic record was found, the script performs a secondary search in the general NCBI database to obtain an accession number—provided the entry is valid and not associated with uncultured or unknown organisms—and uses it to retrieve

the correct taxonomic ID and associated name. This process yields an expanded taxonomy that includes all known naming variants for each species. The script generates an Excel file per host group (CRUSTACEAN_Pathogen_TaxSyn_List.xlsx, FISH_Pathogen_TaxSyn_List.xlsx, HUMAN_Pathogen_TaxSyn_List.xlsx) that lists all taxonomic variants (synonyms, basionyms and 'included' names) identified for each pathogenic species. From this, an intermediate file is created with the extended species list including all nomenclatural variants for further query of the SILVA 138.2 SSU Ref NR99 database (CRUSTACEAN_sp_pathogens_list-EXT.txt, FISH_sp_pathogens_list-EXT.txt, HUMAN_sp_pathogens_list-EXT.txt), and a curated list of pathogenic species containing only the currently accepted scientific names, which serves as the final reference for each host group.

4.2. Generate SILVA reference pathogens dataset

To generate the SILVA reference pathogen dataset, the *database_builder* module ("1_run_database_builder.py") from the PathoLens package was implemented. The process began by filtering the SILVA 138.2 SSU Ref NR99 database to retain only entries corresponding to the taxon Bacteria. At this step, 15.53 % (79 329 sequences) of the initial sequences and 32.67 % (39 118 taxonomies) of the unique taxonomies (i.e.; identical taxonomy from the kingdom to the species level) were excluded. Next, all sequences labeled as "uncultured," "unidentified," "unclassified," "uncultivated," "unculturable," or "unicellular" were systematically removed to ensure the quality and relevance of the data. At this step, 59.61 % (257 059 sequences) of the Bacteria sequences and 20.36 % (16 417 taxonomies) of the unique taxonomies were excluded. After cleaning, the "Bacteria_filtered.fasta" dataset was created and used to extract species matches from the extended PPB species list generated in prior steps. These matches were cross-referenced with the Bacteria dataset for each host group, ensuring that only relevant pathogens were included. Finally, a custom pathogen dataset was generated for each host group (CRUSTACEAN_Pathogen_DB_Unfiltered.fasta, FISH_Pathogen_DB_Unfiltered.fasta and HUMAN_Pathogen_DB_Unfiltered.fasta), which will serve as the basis for the subsequent steps in the analysis pipeline. Most filtering occurred during the removal of unidentified or uncultured entries, resulting in the exclusion of over 257,000 sequences and 16,000 taxonomies, (Fig. 2).

4.3. Data curation

After extracting the sequences from the SILVA 138.2 SSU Ref NR99 database, a comprehensive curation process was applied to each FASTA dataset to ensure the quality of the taxonomy annotations. This step was critical for removing any sequences with taxonomic discrepancies, misclassifications, or incomplete annotations that could negatively impact the correct identification of PPB. The curation process is divided into three key steps, each implemented through specific functions in the *database_filter* module ("2_run_db_filters.py"):

Genus-Species Correspondence Check- This is the most important step in the curation process. When importing the sequences coming from repositories such as NCBI, SILVA curators verify their correct taxonomical assignment. If discrepancies are observed between the original taxonomy and the phylogenetic assignment in the SILVA tree, SILVA curators correct the taxonomy until the genus level but conserve the original genus-species pair at the species level (see examples in Table 2). This would result in wrongly affiliated PPB during the detection process or even worse in false positive PPB. This curation step ensures that not only instances of these discrepancies observed in the MBPD [8] database are now systematically corrected, but also their sequences are accurately aligned to the pathogenic sequences before being presented in the NEMESISdb dataset.

Hence, the first step of our curation process involved the identification of discrepancies when the genus in the taxonomy did not match the genus derived at the species level. The input for

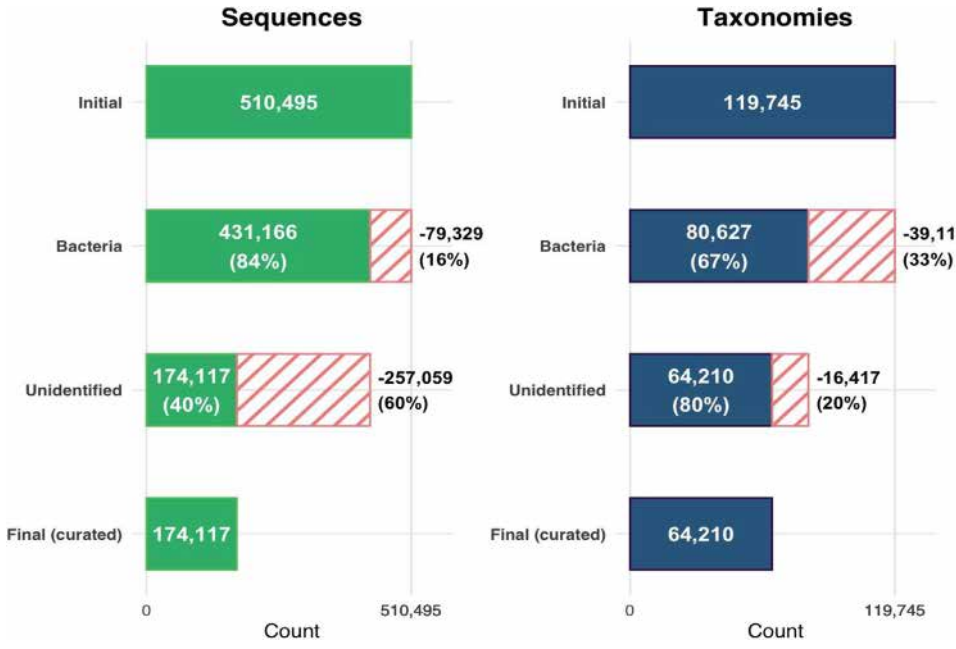

Fig. 2. Overview of sequence and taxonomy retention across data cleaning steps in the construction of the SILVA reference pathogen dataset. Bars represent the total number of entries retained (solid) and removed (striped) at each stage of the pipeline: Initial, Bacteria filtering, Unidentified/uncultured removal, and Final (curated). For each step, the number and percentage of retained and removed entries are indicated. The left panel shows the evolution of sequence entries, and the right panel displays unique taxonomies.

Table 3
Example of discrepancies between the Genus and Species level within the SILVA 138.2 SSU Ref NR99 taxonomy. The correct taxonomy goes until the genus level indicating that the sequence belongs to the bacillus genus but the genus-species pair at the species level is incorrect.

Acc number	Kingdom	Phylum	Class	Order	Family	Genus	Species
EU146061.1.1484	*Bacteria*	*Firmicutes*	*Bacilli*	*Bacillales*	*Bacillaceae*	*Bacillus*	*Streptomyces clavuligerus*

this step consisted of the FASTA files produced from the *database_builder* analysis. Discrepancies and unique taxonomies with mismatches are flagged (i.e.; marked for further revision) in two Python lists for the subsequent curation step.

Multiple-Genera Check - The second curation step assessed multiple genera mentioned within a single taxonomic description. For example, taxonomies that included multiple genera, such as *Hafnia-Obesumbacterium or Shigella-Escherichia*, were reviewed (Tables 3 and Table 4). If one of the genera of the genus level matched with the genus at the species level, the taxonomy was retained; otherwise, an Excel file, "Tax_to_manual-review_{group}.xlsx"**,** was generated with sequences flagged for further manual review due to ambiguous or missing genera.

Manual Review - A manual review process was conducted to validate the flagged discrepancies from the ambiguous or missing genera list. This review was essential for finalizing the list of sequences to be removed from the database. Following this manual review, the final set of sequences marked for deletion was established, and these sequences were subsequently removed from the dataset. The input for this stage was the file "Tax_to_manual-review_{group}.xlsx", and the output was **"**Tax_reviewed_{group}.xlsx", which included the "Retained" column with values of "Yes" or "No" to indicate whether the associated taxonomy (and all sequences with the same taxonomies) would be retained or deleted from the dataset.

Table 4
Example of multiple genera within the genus level of the taxonomy.

Acc number	Kingdom	Phylum	Class	Order	Family	Genus	Species	Decision
JMPC01000305.1.1285	*Bacteria*	*Proteobacteria*	*Gamma-proteobacteria*	*Enterobacterales*	*Entero bacteriaceae*	*Escherichia-Shigella*	*Acinetobacter baumannii 42,057_5*	Flagged
HG738867.2611898.2613439	*Bacteria*	*Proteobacteria*	*Gamma-proteobacteria*	*Enterobacterales*	*Entero bacteriaceae*	*Escherichia-Shigella*	*Escherichia coli str. K-12 substr. MC4100*	retained

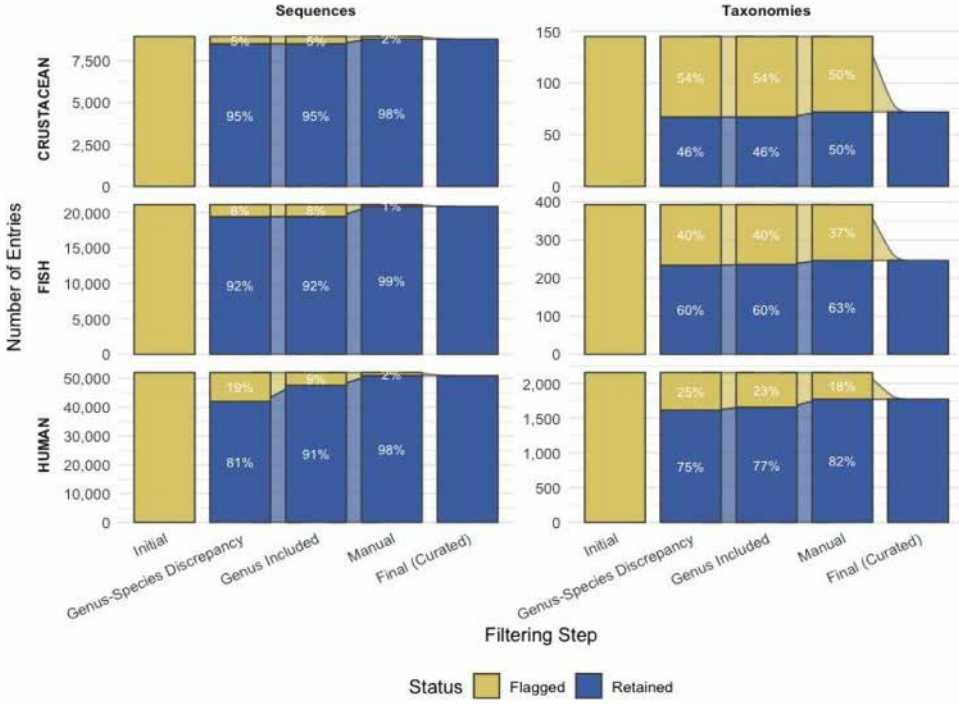

Fig. 3. Retention and flagging of sequences and taxonomies across curation filters for each host group. Each alluvial plot shows the evolution of the number of entries (sequences or taxonomies) that were retained or flagged during the successive data curation steps. The top panels display results for sequence entries, while the bottom panels show taxonomic entries. Rows correspond to different host groups (CRUSTACEAN, FISH, HUMAN), and Y-axis scales are adapted to each case.

4.4. The final curated FASTA dataset

To generate the final curated and validated FASTA datasets, the database_curation module ("3_run_db_curation.py") was implemented. The process begins by reading the input Excel file Tax_reviewed_{group}.xlsx, which indicates which taxonomic entries should be excluded. For each taxonomy marked as "No", a function retrieves the corresponding sequences from the unfiltered FASTA files produced by the *database_builder* module (CRUSTACEAN_Pathogen_DB_Unfiltered.fasta, FISH_Pathogen_DB_Unfiltered.fasta, and HUMAN_Pathogen_DB_Unfiltered.fasta) to identify and remove the corresponding sequences.

As a result, the script outputs the final curated FASTA files—CRUSTACEAN_Pathogen_DB.fasta, FISH_Pathogen_DB.fasta, and HUMAN_Pathogen_DB.fasta—which include only the sequences retained after the curation process.

Additionally, at the end of this module, a species-level summary is generated for each group. An Excel file is created (Species_match_CRUSTACEAN.xlsx, Species_match_FISH.xlsx, Species_match_HUMAN.xlsx) listing the currently accepted scientific names along with all synonyms or variant names found in the database that correspond to each accepted species. This provides a reliable reference for analyzing the species composition of the curated dataset.

Throughout the entire curation process of the datasets, the number of sequences and unique taxonomies that passed through each filter was meticulously recorded. This tracking allowed for a comprehensive understanding of the sequences and taxonomies to be eliminated for each host dataset (Fig. 3). Overall, this plot highlights how the filtering process progressively reduces the pool of sequences and taxonomies marked for elimination, leaving only a small set of sequences

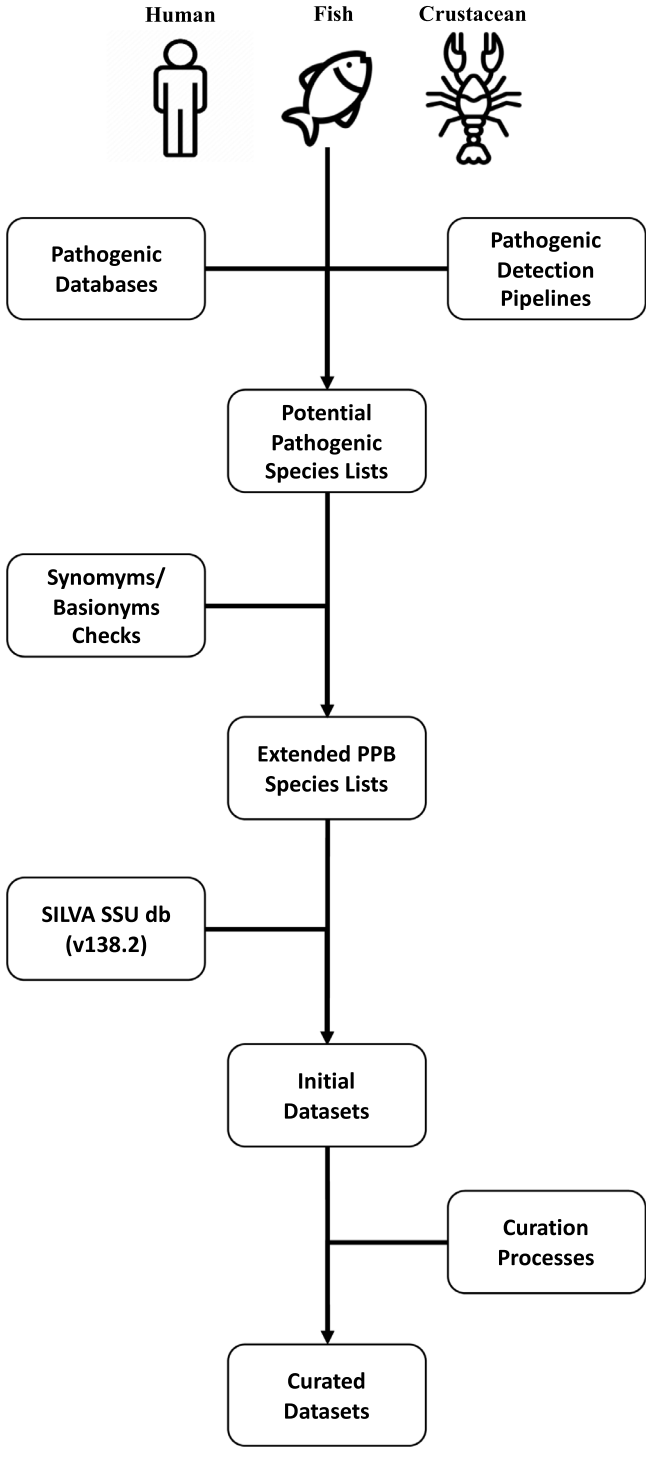

Fig. 4. Pipeline of the creation and curation of the dataset.

(i.e.; 161, 269 and 1098 respectively for Crustacean, Fish and Human) and unique taxonomies (i.e.; 73, 147 and 385 respectively for Crustacean, Fish and Human) to be removed after the final "Manual Review". Overall, the pipeline for creating and curating the dataset is briefly described in Fig. 4.

Limitations

While amplicon and metagenome sequencing have been used to analyse the composition and risk of pathogen contamination [14–16], establishing the definitive pathogenicity of a bacteria still demands additional experimental validations.

Ethics Statement

Authors have read and follow the ethical requirements for publication in Data in Brief. Authors confirm that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

Data Availability

NEMESISdb (Original data) (Zenodo)

CRediT Author Statement

Son-Hoang Tran: Formal analysis, Investigation, Data curation, Writing – original draft; **Claudia Ximena Restrepo-Ortiz:** Methodology, Software, Data curation, Validation, Writing – review & editing; **Dinh Quang Vu:** Data curation; **Marc Troussellier:** Conceptualization, Writing – review & editing; **Yvan Bettarel:** Writing – review & editing; **Thierry Bouvier:** Writing – review & editing; **Van Ngoc Bui:** Writing – review & editing; **Nguyen Hieu Minh:** Data curation; **Trung Du Hoang:** Writing – review & editing; **Quang Huy Nguyen:** Conceptualization, Writing – review & editing; **Jean-Christophe Auguet:** Conceptualization, Methodology, Data curation, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

This project was supported by (1) the Montpellier University of Excellence (KIM MUSE) project “Unravelling the biogeography of the marine pathobiome for health and food management (NYMPHE)”, (2) the Montpellier University EXPOSUM project “HealTH and food security Risks associated to marinE Aquaculture practices and their influence on the circulation of anTibiotic microbial reSistances and pathogens in floating farm socio-ecosystems” (THREATS) (3) the NEMESIS project (2021-EST-149) funded by the French Agency for Food, Environmental and Occupational Health & Safety (ANSES) and (4) the grant project from IO and VAST (Grant project No. TĐĐTMT.01/24–26).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] T. Zhang, J. Miao, N. Han, Y. Qiang, W. Zhang, MPD: a pathogen genome and metagenome database, Database 2018 (2018), doi:10.1093/database/bay055.

[2] A.N. Zhang, J.M. Gaston, C.L. Dai, S. Zhao, M. Poyet, M. Groussin, X. Yin, L.G. Li, M.C.M. van Loosdrecht, E. Topp, M.R. Gillings, W.P. Hanage, J.M. Tiedje, K. Moniz, E.J. Alm, T. Zhang, An omics-based framework for assessing the health risk of antimicrobial resistance genes, Nat. Commun. 12 (2021), doi:10.1038/s41467-021-25096-3.

[3] M. Wardeh, C. Risley, M.K. Mcintyre, C. Setzkorn, M. Baylis, Database of host-pathogen and related species interactions, and their global distribution, Sci. Data 2 (2015) 1–11, doi:10.1038/sdata.2015.49.

[4] T.A. Blauwkamp, S. Thair, M.J. Rosen, L. Blair, M.S. Lindner, I.D. Vilfan, T. Kawli, F.C. Christians, S. Venkatasubrahmanyam, G.D. Wall, A. Cheung, Z.N. Rogers, G. Meshulam-Simon, L. Huijse, S. Balakrishnan, J.V. Quinn, D. Hollemon, D.K. Hong, M.L. Vaughn, M. Kertesz, S. Bercovici, J.C. Wilber, S. Yang, Analytical and clinical validation of a microbial cell-free DNA sequencing test for infectious disease, Nat. Microbiol. 4 (2019) 663–674, doi:10.1038/s41564-018-0349-6.

[5] M. Urban, A.G. Irvine, A. Cuzick, K.E. Hammond-Kosack, Using the pathogen-host interactions database (PHI-base) to investigate plant pathogen genomes and genes implicated in virulence, Front. Plant Sci. 6 (2015), doi:10.3389/fpls.2015.00605.

[6] S. Louca, L.W. Parfrey, M. Doebeli, Decoupling function and taxonomy in the global ocean microbiome, Science (1979) 353 (2016) (1979) 1272–1277, doi:10.1126/science.aaf4507.

[7] J. Miao, N. Han, Y. Qiang, T. Zhang, X. Li, W. Zhang, 16SPIP: a comprehensive analysis pipeline for rapid pathogen detection in clinical samples based on 16S metagenomic sequencing, BMC. Bioinformatics. 18 (2017), doi:10.1186/s12859-017-1975-3.

[8] X. Yang, G. Jiang, Y. Zhang, N. Wang, Y. Zhang, X. Wang, F.J. Zhao, Y. Xu, Q. Shen, Z. Wei, MBPD: a multiple bacterial pathogen detection pipeline for One Health practices, Imeta 2 (2023), doi:10.1002/imt2.82.

[9] C. Quast, E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, F.O. Glockner, The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, Nucleic. Acids. Res. 41 (2013) D590–D596, doi:10.1093/nar/gks1219.

[10] H. Ferchichi, A. St-Hilaire, T.B.M.J. Ouarda, B. Lévesque, Impact of the future coastal water temperature scenarios on the risk of potential growth of pathogenic Vibrio marine bacteria, Estuar. Coast. Shelf. Sci. 250 (2021) 107094, doi:10.1016/j.ecss.2020.107094.

[11] P.J. Landrigan, J.J. Stegeman, L.E. Fleming, D. Allemand, D.M. Anderson, L.C. Backer, F. Brucker-Davis, N. Chevalier, L. Corra, D. Czerucka, M.-Y.D. Bottein, B. Demeneix, M. Depledge, D.D. Deheyn, C.J. Dorman, P. Fénichel, S. Fisher, F. Gaill, F. Galgani, W.H. Gaze, L. Giuliano, P. Grandjean, M.E. Hahn, A. Hamdoun, P. Hess, B. Judson, A. Laborde, J. McGlade, J. Mu, A. Mustapha, M. Neira, R.T. Noble, M.L. Pedrotti, C. Reddy, J. Rocklöv, U.M. Scharler, H. Shanmugam, G. Taghian, J.A.J.M. Van de Water, L. Vezzulli, P. Weihe, A. Zeka, H. Raps, .P.. Rampal, Human health and ocean pollution, Ann. Glob. Health 86 (2020) 151, doi:10.5334/aogh.2831.

[12] T.G. Aw, J.B. Rose, Detection of pathogens in water: from phylochips to qPCR to pyrosequencing, Curr. Opin. Biotechnol. 23 (2012) 422–430, doi:10.1016/j.copbio.2011.11.016.

[13] Q. Cui, T. Fang, Y. Huang, P. Dong, H. Wang, Evaluation of bacterial pathogen diversity, abundance and health risks in urban recreational water by amplicon next-generation sequencing and quantitative PCR, Journal of Environmental Sciences 57 (2017) 137–149, doi:10.1016/j.jes.2016.11.008.

[14] J. Naudet, E.R. d'Orbcastel, T. Bouvier, J.-C. Auguet, Plastic-associated pathogens in marine environments: a meta-analysis, Mar. Pollut. Bull. 219 (2025) 118266, doi:10.1016/j.marpolbul.2025.118266.

[15] J. Naudet, J.-C. Auguet, T. Bouvier, R. Rakotovao, T. Motte, L. Gaumez, T. Crucitti, F. Rieuvilleneuve, E. Roque d'Orbcastel, Polymers and immersion time shape bacterial pathogen and antibiotic resistance profiles in aquaculture facilities, FEMS. Microbiol. Ecol. 101 (2025), doi:10.1093/femsec/fiaf076.

[16] J. Naudet, E.R. d'Orbcastel, T. Bouvier, S. Godreuil, S. Dyall, S. Bouvy, F. Rieuvilleneuve, C.X. Restrepo-Ortiz, Y. Bettarel, J.-C. Auguet, Identifying macroplastic pathobiomes and antibiotic resistance in a subtropical fish farm, Mar. Pollut. Bull. 194 (2023) 115267, doi:10.1016/j.marpolbul.2023.115267.

[17] E. Garner, B.C. Davis, E. Milligan, M.F. Blair, I. Keenum, A. Maile-Moskowitz, J. Pan, M. Gnegy, K. Liguori, S. Gupta, A.J. Prussin, L.C. Marr, L.S. Heath, P.J. Vikesland, L. Zhang, A. Pruden, Next generation sequencing approaches to evaluate water and wastewater quality, Water. Res. 194 (2021) 116907, doi:10.1016/j.watres.2021.116907.

[18] E. Lewis, J.A. Hudson, N. Cook, J.D. Barnes, E. Haynes, Next-generation sequencing as a screening tool for foodborne pathogens in fresh produce, J. Microbiol. Methods 171 (2020) 105840, doi:10.1016/j.mimet.2020.105840.

[19] S.D. Chorlton, Ten common issues with reference sequence databases and how to mitigate them, Front. Bioinform. 4 (2024), doi:10.3389/fbinf.2024.1278228.

[20] C.X. Restrepo-Ortiz, PathoLens [software], 2025. Https://Doi.Org/10.5281/Zenodo.15298262.

[21] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Bio. 215 (1990) 403–410 <Go to ISI>://A1990ED16700008.

[22] A. Yates, K. Beal, S. Keenan, W. McLaren, M. Pignatelli, G.R.S. Ritchie, M. Ruffier, K. Taylor, A. Vullo, P. Flicek, The Ensembl REST API: ensembl data for any language, Bioinformatics. 31 (2015) 143–145, doi:10.1093/bioinformatics/btu613.

[23] W.J. Kent, BLAT —The BLAST-like alignment tool, Genome Res. 12 (2002) 656–664, doi:10.1101/gr.229202.

[24] D. Smedley, S. Haider, B. Ballester, R. Holland, D. London, G. Thorisson, A. Kasprzyk, BioMart – biological queries made easy, BMC. Genomics. 10 (2009) 22, doi:10.1186/1471-2164-10-22.

[25] M. Wardeh, C. Risley, M. McIntyre, C. Setzkorn, M. Baylis, SpeciesInteractions_EID2. figshare., (2015). https://doi.org/10.6084/m9.figshare.1381853.v5.

[26] C.L. Schoch, S. Ciufo, M. Domrachev, C.L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J.P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI Taxonomy: a comprehensive update on curation, resources and tools, , Database 2020 (2020), doi:10.1093/database/baaa062.
[27] P.J.A. Cock, T. Antao, J.T. Chang, B.A. Chapman, C.J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, M.J.L. de Hoon, Biopython: freely available Python tools for computational molecular biology and bioinformatics, Bioinformatics. 25 (2009) 1422–1423, doi:10.1093/bioinformatics/btp163.