AMERICAN SOCIETY FOR MICROBIOLOGY | Applied and Environmental Microbiology®

EVOLUTIONARY AND GENOMIC MICROBIOLOGY

mobileOG-db: a Manually Curated Database of Protein Families Mediating the Life Cycle of Bacterial Mobile Genetic Elements

Connor L. Brown,[a] James Mullet,[b] Fadi Hindi,[b] James E. Stoll,[c] Suraj Gupta,[a] Minyoung Choi,[b] Ishi Keenum,[b] Peter Vikesland,[b] Amy Pruden,[b] Liqing Zhang[d]

[a]Department of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, Virginia, USA
[b]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA
[c]Fralin Life Science Institute, Blacksburg, Virginia, USA
[d]Department of Computer Science, Virginia Tech, Blacksburg, Virginia, USA

ABSTRACT Bacterial mobile genetic elements (MGEs) encode functional modules that perform both core and accessory functions for the element, the latter of which are often only transiently associated with the element. The presence of these accessory genes, which are often close homologs to primarily immobile genes, incur high rates of false positives and, therefore, limits the usability of these databases for MGE annotation. To overcome this limitation, we analyzed 10,776,849 protein sequences derived from eight MGE databases to compile a comprehensive set of 6,140 manually curated protein families that are linked to the "life cycle" (integration/excision, replication/recombination/repair, transfer, stability/transfer/defense, and phage-specific processes) of plasmids, phages, integrative, transposable, and conjugative elements. We overlay experimental information where available to create a tiered annotation scheme of high-quality annotations and annotations inferred exclusively through bioinformatic evidence. We additionally provide an MGE-class label for each entry (e.g., plasmid or integrative element), and assign to each entry a major and minor category. The resulting database, mobileOG-db (for mobile orthologous groups), comprises over 700,000 deduplicated sequences encompassing five major mobileOG categories and more than 50 minor categories, providing a structured language and interpretable basis for an array of MGE-centered analyses. mobileOG-db can be accessed at mobileogdb.flsi.cloud.vt.edu/, where users can select, refine, and analyze custom subsets of the dynamic mobilome.

IMPORTANCE The analysis of bacterial mobile genetic elements (MGEs) in genomic data is a critical step toward profiling the root causes of antibiotic resistance, phenotypic or metabolic diversity, and the evolution of bacterial genera. Existing methods for MGE annotation pose high barriers of biological and computational expertise to properly harness. To bridge this gap, we systematically analyzed 10,776,849 proteins derived from eight databases of MGEs to identify 6,140 MGE protein families that can serve as candidate hallmarks, i.e., proteins that can be used as "signatures" of MGEs to aid annotation. The resulting resource, mobileOG-db, provides a multilevel classification scheme that encompasses plasmid, phage, integrative, and transposable element protein families categorized into five major mobileOG categories and more than 50 minor categories. mobileOG-db thus provides a rich resource for simple and intuitive element annotation that can be integrated seamlessly into existing MGE detection pipelines and colocalization analyses.

KEYWORDS antibiotic resistance, bacteriophages, insertion sequence, integrative elements, metagenomics, mobile genetic elements, mobilome, plasmids, transposons

Bacterial mobile genetic elements (MGEs) are of broad interest across multiple research communities. MGEs are critical drivers of horizontal gene transfer (HGT; i.e., the movement of genetic material between nonparental lineages of bacteria) (1, 2). MGEs are especially

Editor Hideaki Nojiri, The University of Tokyo

Address correspondence to Amy Pruden, apruden@vt.edu, or Liqing Zhang, lqzhang@vt.edu.

The authors declare no conflict of interest.

Received 17 June 2022
Accepted 27 July 2022
Published 29 August 2022

a concern as a fundamental driver of the spread of antibiotic resistance. Antibiotic resistance is a growing global threat to public health (3), empowering bacterial infections to survive antibiotic treatment and rendering these life-saving drugs ineffective. For instance, the multidrug resistance plasmid NR1 (R100), first recovered from a *Shigella flexneri* isolate in the 1950s (4), was found to confer resistance through its carriage of a mobile transposon (Tn*21*) harboring multiple antibiotic resistance genes (ARGs) (2, 5). Antibiotic resistance is now pandemic in many clinically relevant bacteria and pathogens enriched with ARG-harboring MGEs ranked among the most urgent and serious threats identified in the 2019 U.S. Centers for Disease Control and Prevention's Antibiotic Threats report (6). Thus, surveilling the occurrence of HGT in genomic data is a key tool in the fight to abate the global spread of antibiotic resistance. Unfortunately, a major hinderance to research centered on MGEs is that currently available MGE databases are disparate, redundant, and contain extraneous genes that encode accessory functions for the host bacterium.

Next generation sequencing is commonly applied to profile MGEs in environmental or clinical samples (e.g., shotgun metagenomics or whole-genome sequencing) using bioinformatic analysis to identify molecular MGE signatures, such as genes encoding recombinases (7–11), sequence features (12), or even the full nucleotide sequence of MGEs (13, 14). However, unlike ARGs, which have benefitted from widespread efforts to curate unified databases (e.g., the Comprehensive Antibiotic Resistance Databases [CARD] [15] or the Structured Antibiotic Resistance Gene Database [SARG] [16]), collections of MGEs have thus far been compiled in disparate, independent databases. This situation has led to redundant entries, inconsistent nomenclature, and the presence of extraneous genes. Furthermore, there is no centralized resource for MGE hallmark genes that could serve as the basis for annotating diverse classes of MGEs. Instead, decentralized databases exist for phages, including pVOG (17) and the GutPhage Database (GPD) (18); insertion sequences, ISfinder (19); integrative genomic elements (IGEs), ICEberg (11), immedb (20); or plasmids, COMPASS (21), NCBI Plasmid RefSeq (22). While ACLAME (23) combines multiple element types, there has not been a substantial update to this database since 2010 (23, 24).

Current databases of MGEs also contain accessory genes, such as ARGs (5, 25), metal resistance genes (5, 26), or virulence factors (1, 27–29), which are irrelevant for element classification and may themselves be separate targets in genomic analyses (30–32) (Fig. 1A). For instance, Slizovskiy et al. found that overlap in MGE and ARG databases resulted in ambiguous or even erroneous annotations of mobile ARGs (10) (Fig. 1A and B). Well-documented accessory genes such as ARGs can be removed using existing databases (Fig. 1A); however, many cargo genes with tenuous relevance to mobility will remain. Thus, the presence of these cargo genes in MGE databases leads to the frequent occurrence of false-positive matches that confound and complicate MGE annotation (10) (Fig. 1B). To combat this, other studies have used custom databases of MGE-marker genes created using keyword searches against public databases such as UniProt or GenBank (e.g., integrase, recombinase, or transposase) (7–9, 25, 33, 34). While providing a solution to the presence of accessory genes, such usage of large-scale public repositories may introduce poorly characterized or moonlighting (i.e., multifunctional) members of MGE protein families (35–37). For instance, Smyshlyaev et al. demonstrated that the largest phylogenetic subgroup of tyrosine recombinases, a large protein family including phage and integron integrases, included the *xerC/xerD* family of simple tyrosine recombinases, which are infrequently associated with MGEs (38) (Fig. 1C). Further, there is increasing recognition of the need to classify MGEs, especially plasmids, using the entire backbone of the element rather than on traditional markers such as replicons and mobilases (11, 21, 39, 40). However, which genes might suitably comprise the "backbone," or essential and conserved regions of MGEs, remains uncertain. Thus, there is a need for a structured resource and classification system that limits false positives, provides reliable matches to well-characterized MGEs, and centralizes available knowledge pertaining to diverse MGE classes.

To facilitate MGE annotation, we developed the mobile orthologous groups database (mobileOG-db), an interactive resource compiling a comprehensive variety of proteins

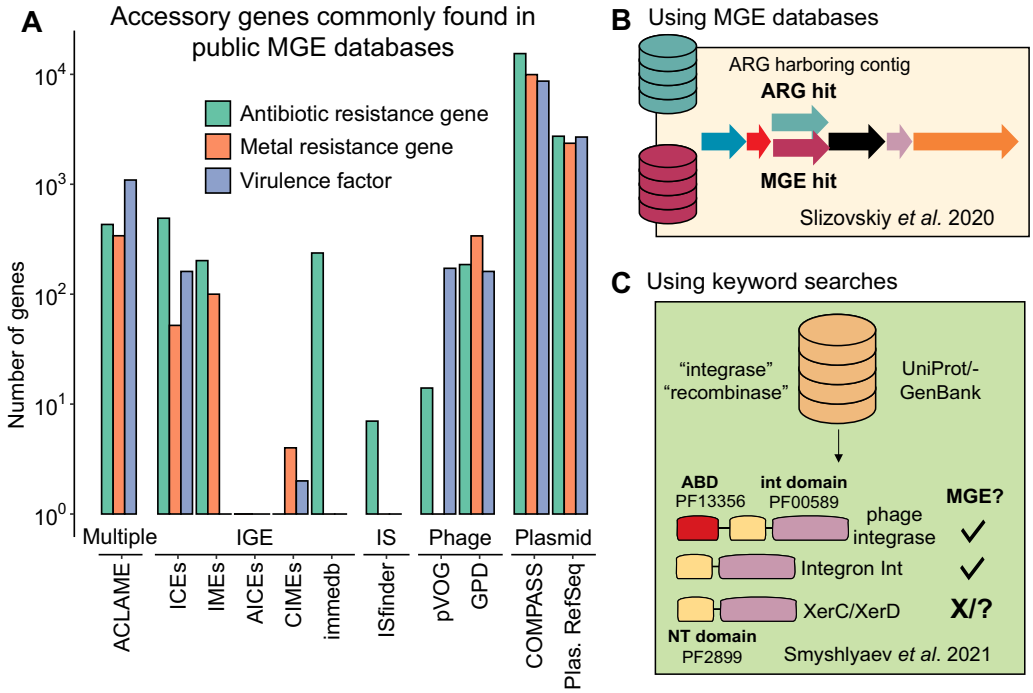

FIG 1 Challenges associated with existing databases for MGE colocalization analysis. (A) Example accessory genes found within MGE databases, with the databases grouped by the element class that they are intended to contain. Annotation criteria and databases are provided in Supplementary Methods. ACLAME is a database of multiple element types, thus is labeled "multiple." (B) Synthesis of Slizovskiy et al. (10). Overlap in MGE/ARG databases produces ambiguous annotations of mobile ARGs. (C) Synthesis of Smyshlyaev et al. (38). The largest phylogenetic subgroup of tyrosine recombinases includes the *xerC/xerD* family of simple tyrosine recombinases, meaning studies using keyword search-based databases are likely to inadvertently include false-positive sequences. IGE, integrative genomic elements; IS, insertion sequence.

that mediate the essential functions or "life cycle" of bacterial MGEs. Here, we define the essential functions of MGEs to include: (i) integration and excision (IE) from one genetic locus to another; (ii) replication, recombination, or nucleic acid repair (RRR); (iii) interorganism transfer (T); (iv) element stability, transfer, or defense (STD); and (v) phage (P) specific biological processes (e.g., genome packaging, or lysis and lysogeny).

The motivation of this work was to provide a structured and central resource for MGE annotation that is rooted in a biologically defensible classification of MGE protein function. In support of this effort, we analyzed 10,776,849 proteins from eight databases containing MGEs to: (i) remove accessory or poorly characterized genes; (ii) classify each entry as belonging to a phage, plasmid, insertion sequence, or IGE; and (iii) assign each entry a major mobileOG category (i.e., IE, RRR, T, STD, P) and one of over 100 minor mobileOG categories. We posit that these proteins are suitable candidate MGE hallmarks because of the key functions they perform, and thus they can be used for accurately detecting and characterizing MGEs in various genomic data sets.

RESULTS

Here, we analyzed 10,776,849 proteins derived from eight databases containing bacterial MGEs to create a novel database, mobileOG-db, that directly addresses limitations imposed by current resources. Through an iterative annotation process, we eliminated 8,728,599 protein sequences from this set that were uncharacterized or indefensible as candidate MGE hallmarks from a biological standpoint. Because each sequence is directly derived from an existing MGE database, we reduce the probability of false positives incurred because of multifunctional or moonlighting proteins present in databases like UniProt. In addition, we provide a classification system consisting of:

- Major mobileOG categories: IE, STD, P, RRR, T (Fig. 2).

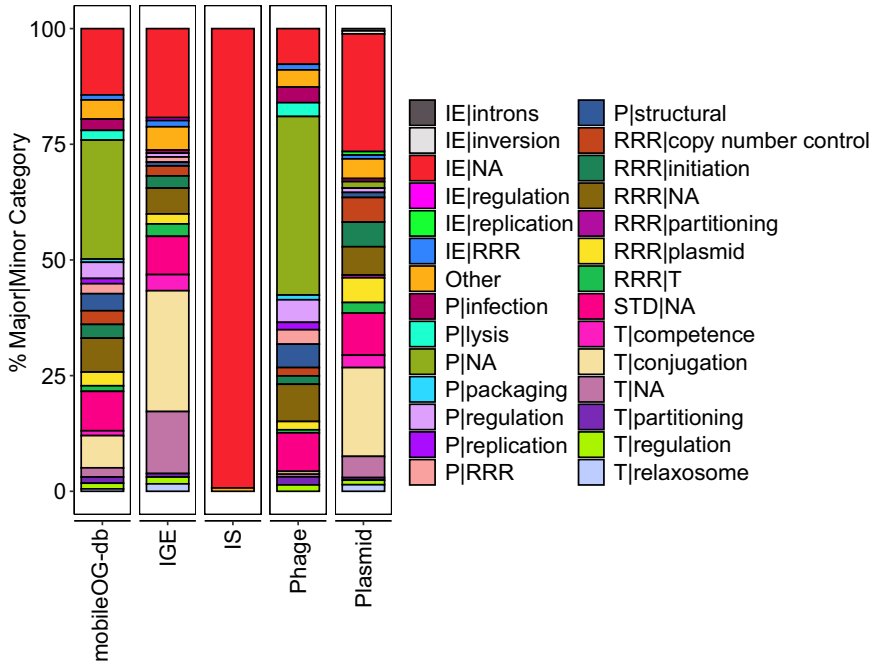

FIG 2 mobileOG-db comprises a diverse repertoire of proteins that orchestrate the unique life cycles of integrative genomic elements (IGEs), insertion sequences (IS), bacteriophages, and plasmids. Manually ascribed major and minor categories of entries derived from phage, plasmid, IS, or IGE sequences are displayed. The key displayed has the form Major mobileOG category|Minor mobileOG Category. Only categories with >1% occurrence in a given element type were included for visualization.

- More than 50 minor mobileOG Categories (Table S4), including: initiation (a subcategory of RRR referring to the Rep family proteins), structural or infection (subcategories of P), shufflon, inversion, or intron (subcategories of IE), conjugation, competence, or phage receptor (subcategories of T), among others (Fig. 2; Table S4).
- Element class labels derived from the source databases of phage, plasmid, insertion sequence, IGE, or multiple.

Altogether, this first release (beatrix) of mobileOG-db comprises 775,257 deduplicated proteins, including 29,721 derived directly from manually curated entries; 6,346 protein clusters or families (defined as greater than 40% identical over 50% of the subject and query length; see Materials and Methods); 2,444 unique manual annotations, and 1,393 references. The mobileOG-db web interface (Fig. 3) provides a simple and intuitive way to work with the database. Through this interface, users are offered the opportunity to identify overlap between element class gene contents, download custom databases, or select only manually curated or homologous entries. The complete database (comprising keyword search results, manually curated, and homology-based entries) is available as a standalone download on the website, while the web interface exclusively hosts the manually curated sequences and their homologs (Fig. 3).

In addition to aggregating and classifying existing databases, the curation of diverse MGE protein sequences allowed for the reannotation of more dated public database entries in light of recent discoveries. For example, we found components of several distinct MGE-defense systems on diverse element types (Table S5). These include the Bacteriophage Exclusion (BREX) (41) system genes *brxB*, *brxF*, *pglX*, *brxL*, among others (Table S5). Surprisingly, *pglX* gene homologs as well as the cognate antitoxin were found across several element types, including phages in the Gut Phage Database (Table S5), while components of the cyclic oligonucleotide-based antiphage signaling system (CBASS) (42) were exclusively found within ICEs and plasmids (Table S5). Further, we note the presence of several homologs of CRISPR-system components encoded on ICEs, plasmids, and even bacteriophages in the GutPhage database (Table S5). While the

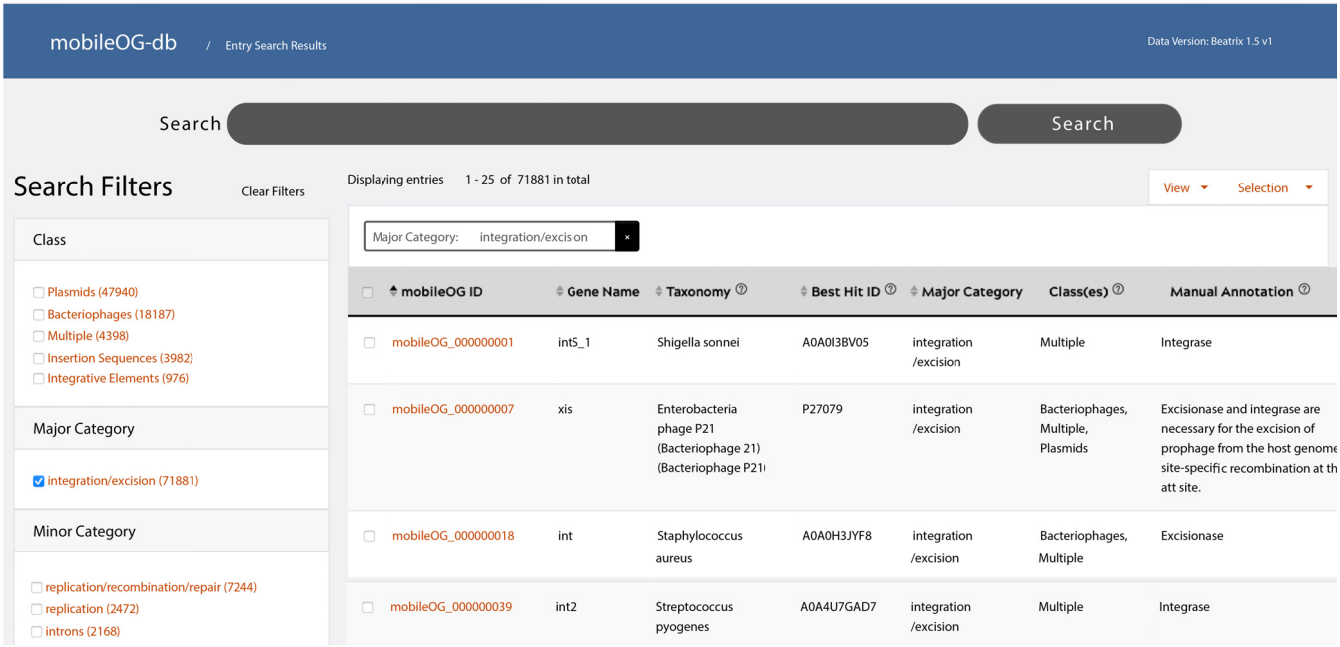

FIG 3 The mobileOG-db web-interface provides an interactive format for parsing mobileOG-db entries. mobileOG entries, representing deduplicated sequences, can be manually selected by category or by element type (e.g., plasmid, bacteriophage, or insertion sequence).

functionality of these proteins is unknown, the presence of defense system components highlights the potential for yet unexplored MGE cross talk that may contribute to inter-element or interelement class (e.g., plasmid versus phage) competition. The accessions of these sequences are available as standalone downloads in Table S5 and are incorporated into mobileOG-db under the STD major category.

Usage recommendations and examples. For detecting and classifying elements from long genomic segments (e.g., long reads or assembled short reads), it is recommended that multiple colocalized hits in close proximity should be incorporated into the annotation criteria, similar to the pattern-based colocalization approach leveraged by ICEBerg (11) for IGE detection. Likewise, because plasmids and phages frequently encode homologs of RRR machinery that are also present in exclusively cellular DNA, it is noted hits solely to RRR modules are not necessarily indicative of an MGE (43). An additional caveat is hits to type 4 secretion systems may not be indicative of an MGE; paralogues of these proteins are also virulence determinants in some organisms (44). A simple annotation pipeline, mobileOGs.pl-kyanite (https://github.com/clb21565/mobileOG-db/tree/main/mobileOG-pl), has been developed (Supplementary Methods; Table S5) to allow for automated element annotation (Supplementary Methods; Fig. S2) that generates, annotates, and identifies orfs with homology to mobileOGs. Then, using a python script, it produces a summary table for the number of mobileOGs of specific element classes (e.g., phage or plasmid) were found for the contig. Using mobileOGs.pl-kyanite, we evaluated the ability of mobileOG-db annotations to correctly label sequences from COMPASS or pVOG as plasmids and phages, respectively (Table S2). This pipeline enabled successful classification of up to 98.2% and 99.7% of the plasmids and phages, respectively (https://doi.org/10.6084/m9.figshare.15170736; Fig. S2; Table S3). Importantly, we acknowledge that mobileOG-db alone cannot provide a highly granular classification of MGEs, as this would require leveraging additional sequence features, including, for example, recombination or replication initiation sites. However, database-derived annotations display strong agreement with VirSorter on real-world data (Fig. S2 and 3).

DISCUSSION

While the goal of this work was to establish a well-curated MGE database, future research can further refine approaches to harness this information, depending on the

application. In addition, though mobileOG-db addresses many of the limitations of existing resources for MGE annotation, particularly for colocalization applications, it is not a replacement for databases of full-length MGEs, e.g., PLSDB (45), TnCentral (46), and ISfinder (19). The creation of mobileOG-db relied on existing documentation of protein family function in the literature. Thus, proteins lacking experimental characterization in the function of MGE biology are likely absent from mobileOG-db. On the other hand, mobileOG-db also offers the distinct advantage over databases of individual elements in that it provides a multiclass labeling scheme derived from the databases of origin (Fig. 2). This structure allows for analysis of MGEs as hierarchical entities which would otherwise be impossible when using databases of specific MGE classes. To ensure long-term utility of the database, mobileOG-db will be updated approximately once a year with major changes, bug fixes, and new annotations. Minor grammatical or technical errors will be fixed on a semiregular basis as discovered or reported through the GitHub page (https://github.com/clb21565/mobileOG-db) or through the "Contact us" feature of the mobileOG-db website.

The rapidly increasing volume of sequence data has outpaced the ability to annotate and classify MGEs, in part due to a lack of a suitable knowledgebase (39). Thus, mobileOG-db was created specifically to address this critical gap. mobileOG-db and its web interface provide the ability to test a targeted hypothesis with reliable, customized databases tailored to the research objective. For instance, users interested in detecting mobile ARGs could select insertion sequence proteins, transfer/conjugation proteins, and plasmid-associated proteins (e.g., RepA, FinO), thus allowing for the detection of conjugative plasmid bearing insertion sequences, a key mechanism of ARG HGT (47). In the future, we expect the structured language provided by mobileOG-db will enable novel analyses that leverage the diverse and compositional nature of MGEs to enhance and refine the profiling of the dynamic mobilome. Such efforts will advance a diverse array of critical research fronts, particularly efforts aiming at understanding and tracking the spread of antibiotic resistance.

MATERIALS AND METHODS

Filtering accessory and uncharacterized MGE proteins from public databases. Our curation efforts aimed to remove proteins with accessory or poorly understood functions from the contents of ICEBerg 2.0 (11), COMPASS (21), NCBI Plasmid RefSeq (22), GutPhage Database (18), pVOG (17), ISfinder (19), ACLAME (23), and immedb (20). To focus curation efforts exclusively on those referenced in the MGE literature, we created and queried a database of MGE abstracts from PubMed. To build this database, we gathered a set of descriptive keywords related to MGEs (Table S1) referencing *Mobile DNA III* (48) to identify important keywords. These terms were searched against NCBI PubMed using entrez to extract article metadata and abstracts. To exclude unrelated abstracts, the initial set was filtered to remove those with less than three keyword matches. The final resulting abstract set was then queried in subsequent searches and manually curations.

A pan-mobilome, i.e., an extensive collection of sequences comprising diverse MGEs, was created by merging the contents of eight publicly available MGE databases into a single database of protein sequences. The genomes comprising pVOG, COMPASS, and immedb, all nucleotide sequence databases, were processed with prodigal (v2.6.3) to generate open reading frames using the -p meta setting. This setting was chosen because the alternative, -p single, requires 20,000 bp of training data which would not be possible for many of the MGEs (49). These databases were selected as they comprise a representative sample of MGE diversity. Future iterations of mobileOG-db will expand to other databases to improve the coverage of MGE protein functional diversity. The final aggregated data set included 10,776,849 sequences (Fig. 4). The 10,776,849 proteins were searched against UniProt (release-2020_06) using diamond (50) blastp, with minimum identity 90% and minimum query coverage of 80%. Of the 10,776,749 sequences, 8,460,321 had matches to UniProt that passed filtering criteria, likely due to erroneous or truncated open reading frames generated by prodigal. Gene names for the 8,460,321 protein sequences were then extracted from a merged Bacterial, Archaeal, and viral UniProt knowledge base (.dat file downloaded using wget from the UniProt ftp server) with a custom script (available on the project GitHub page, https://github.com/clb21565/mobileOG-db/tree/main/scripts). This process produced 110,234 gene names corresponding to the 8,460,321 sequences; 20,979 of the 110,234 gene names were unique. Many of the resulting gene names were likely to create spurious hits (e.g., antiholin gene *S*, UniProt entry P03705), or contained additional characters (e.g., *traG_2*, UniProt entry G9G740), and so the names were processed to produce suitable queries for searching against the abstract database. Queries were produced from the 20,979 gene names in the following way: Gene names with three or fewer characters were prefixed with "protein" or "gene" to reduce spurious hits. Gene names with an underscore or special character were split on either side (e.g., *tnpA_2* would become two queries: *tnpA* and *2*) and both sides were searched against the MGE-abstract database. Altogether this process produced 41,303 unique queries (the complete list of search terms is

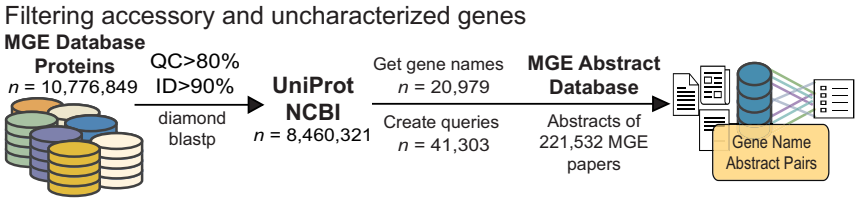

FIG 4 Filtering accessory genes using a database of published scientific abstracts relating to bacterial MGEs. The contents of the eight databases (n = 10,776,849) were mapped to UniProt entries and assigned gene names and annotations. The resulting gene names (n = 20,979) were converted into queries and searched against the MGE abstract database to create gene name abstract set pairs.

available on the figshare project site: https://doi.org/10.6084/m9.figshare.15170736). These queries were then searched against the MGE abstract database, resulting in 7,012 gene name-abstract set pairs that were then passed to manual curation.

Manual curation and classification of MGE protein family function. Each of the 7,012 gene name-abstract set pairs were scanned by at least two researchers to determine whether the gene name might encode a protein that performed one of the target functions (IE, T, P, RRR, or STD) (Fig. 5). This process resulted in the discarding of 4,222 gene names that corresponded to proteins with accessory or unknown function, leaving 2,790 gene names for annotation refinement. These 2,790 gene names corresponded to 98,465 sequences, with many sequences sharing the same name (see Supplementary Methods for examples). To reduce the extent of curation required, these proteins were grouped into protein clusters, or families defined as being 40% identical over 50% of the reference and query sequence using mmseqs2 (mmseqs easy-clust -c 0) (51). This produced 13,441 clusters associated with the 2,790 names. This coverage criterion was selected to reduce the incidence of singlet clusters induced by fragmented open reading frames. The 40% identity cut-off is used by the enzyme commission to determine whether a protein fold retains the same function as a reference fold (52). One named representative from each cluster (i.e., a cluster representative bearing one of the 2,790 gene names) was selected and passed to annotation refinement (Fig. 6). To refine annotations and assign mobileOG categories, the cluster representative's putative function was compared with that described by the abstracts. If it was inconsistent with the attributing abstract(s) (see Supplementary Methods), the sequence was reannotated following a review of the literature recovered by searching for the gene name and putative function in PubMed. Finally, curated cluster representatives were assigned a major and minor mobileOG category. Minor mobileOG categories were assigned as secondary functional labels under the major mobileOG category by considering what putative function the family was associated with. Homologs of the manually curated sequences were assigned a category using the manually curated representative(s) (Fig. 6).

To improve coverage, keyword matches in the fasta headers with a table of MGE protein keywords (Table S2) were used as evidence for inclusion in mobileOG-db. The evidence used to determine inclusion (manual curation, homology, or keyword searches) is recorded in mobileOG-db. Examples of the rationales applied are provided in Supplementary Methods. Sequences with matches to Swiss-Prot entries, which are manually curated UniProt entries, were considered a special case and were manually curated regardless of whether they were returned during the abstract analysis. Final mobileOG-db entries are deduplicated at 100% identity using mmseqs2 (–min-seq-id 1 -c 1 –cov-mode 0), and the occurrence of identical sequences across different databases and element classes is recorded (Fig. 6). Orthology assignments from EggNOG (53) as well as associated pfam (54) domain labels are provided through the web interface and as standalone downloads. The gene names, queries, and the abstract database, are available at the Figshare project (https://doi.org/10.6084/m9.figshare.15170736).

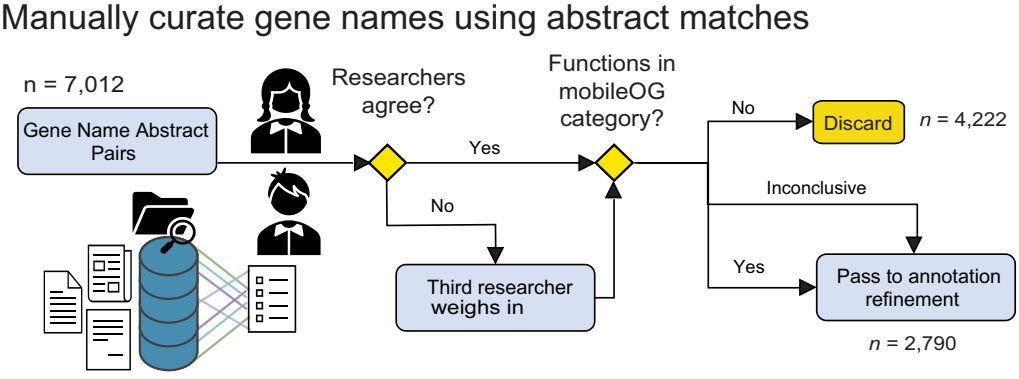

FIG 5 Preliminary gene name curation. A total of 7,012 gene abstract-set pairs were manually inspected by at least two researchers to determine whether the gene name could belong to a protein that performed one of the target processes. This curation resulted in 4,222 discarded entries (identified as likely cargo, or off-target matches) and 2,790 gene name-abstract set pairs to have refined functional annotations.

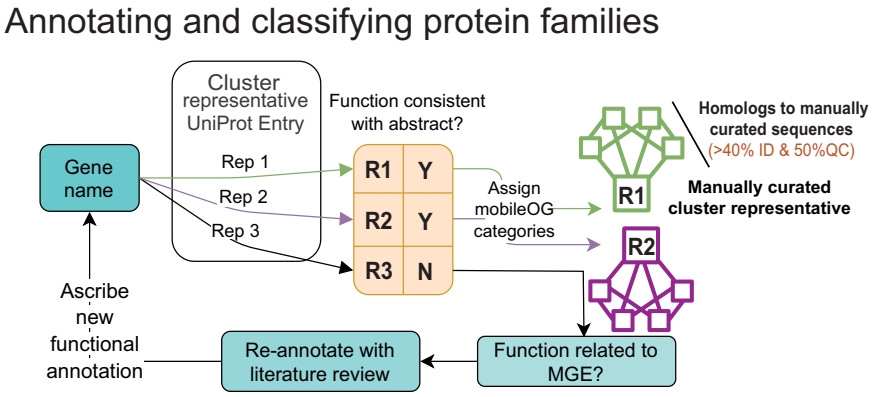

FIG 6 Annotation refinement and classification of MGE protein sequences. Each of the 2,790 gene names passed to refinement typically belonged to multiple protein clusters or families. For each cluster, one named representative was selected, and its putative function was compared with the literature-derived descriptions recovered from the abstract analysis. If the UniProt/NCBI entry did not support a link between the gene name and function, the protein was annotated via literature review by one of two researchers, with uncertainties and disagreements settled by discussion. Protein families that were ultimately confirmed to perform a target function were assigned a major and minor mobileOG category.

Data availability. mobileOG-db is available at mobileogdb.flsi.cloud.vt.edu/, where users can browse, filter, search, and download customized data sets and references. Scripts used in the text mining analysis are available at https://github.com/clb21565/mobileOG-db/tree/main/scripts, while the preliminary annotation pipeline, mobileOG-pl. kyanite can be accessed at https://github.com/clb21565/mobileOG-db/tree/main/mobileOG-pl.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 0.4 MB.

SUPPLEMENTAL FILE 2, XLSX file, 0.02 MB.

SUPPLEMENTAL FILE 3, XLSX file, 0.01 MB.

SUPPLEMENTAL FILE 4, XLSX file, 0.9 MB.

ACKNOWLEDGMENTS

We would like to directly express our appreciation for the work and expertise that went into designing the databases making up mobileOG-db. The authors acknowledge the Advanced Research Computing at Virginia Tech for providing computational resources.

This study was supported by NSF PIRE (PI Vikesland) Award 1545756, NSF CI4WARS (PI Zhang) Award 2004751, NSF NRT Award 2125798 (PI Pruden), USDA National Institute of Food and Agriculture competitive Grant 2017-68003-26498, Water Research Foundation Project 4961, the Genetics, Bioinformatics, and Computational Biology Interdisciplinary Graduate Education Program (IGEP), the Virginia Tech Sustainable NanoTechnology IGEP, NanoEarth, Fralin Life Sciences Institute, the Virginia Tech Open Access Support Fund, and the Virginia Tech ICTAS Center for Science and Engineering of the Exposome.

The authors report no conflicts of interest.

REFERENCES

1. Rankin DJ, Rocha EPC, Brown SP. 2011. What traits are carried on mobile genetic elements, and why. Heredity (Edinb) 106:1–10. https://doi.org/10.1038/hdy.2010.24.
2. Davies J. 1996. Origins and evolution of antibiotic resistance. Microbiologia https://doi.org/10.1128/mmbr.00016-10.
3. General Assembly of the United Nations. 2016. High-level meeting on antimicrobial resistance general assembly of the United Nations. https://www.un.org/pga/71/event-latest/high-level-meeting-on-antimicrobial-resistance/. Retrieved 26 October 2021.
4. Nakaya R, Nakamura A, Murata Y. 1960. Resistance transfer agents in Shigella. Biochem Biophys Res Commun 3:654–659. https://doi.org/10.1016/0006-291x(60)90081-4.
5. Liebert CA, Hall RM, Summers AO. 1999. Transposon Tn 21, flagship of the floating genome. Microbiol Mol Biol Rev 63:507–522. https://doi.org/10.1128/MMBR.63.3.507-522.1999.
6. CDC. 2019. Biggest threats and data. Antibiotic/antimicrobial resistance. CDC. https://www.cdc.gov/drugresistance/biggest-threats.html. Retrieved 26 October 2021.
7. Pärnänen KMM, Narciso-da-Rocha C, Kneis D, Berendonk TU, Cacace D, Do TT, Elpers C, Fatta-Kassinos D, Henriques I, Jaeger T, Karkman A, Martinez JL, Michael SG, Michael-Kordatou I, O'Sullivan K, Rodriguez-Mozaz S, Schwartz T, Sheng H, Sørum H, Stedtfeld RD, Tiedje JM, Giustina SVD, Walsh F, Vaz-Moreira I, Virta M, Manaia CM. 2019. Antibiotic resistance in European wastewater treatment plants mirrors the pattern of

clinical antibiotic resistance prevalence. Sci Adv 5:eaau9124. https://doi.org/10.1126/sciadv.aau9124.
8. Oh M, Pruden A, Chen C, Heath LS, Xia K, Zhang L. 2018. Meta compare: a computational pipeline for prioritizing environmental resistome risk. FEMS Microbiol Ecol 94. https://doi.org/10.1093/femsec/fiy079.
9. Arango-Argoty GA, Dai D, Pruden A, Vikesland P, Heath LS, Zhang L. 2019. NanoARG: a web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes. Microbiome 7. https://doi.org/10.1186/s40168-019-0703-9.
10. Slizovskiy IB, Mukherjee K, Dean CJ, Boucher C, Noyes NR. 2020. Mobilization of antibiotic resistance: are current approaches for colocalizing resistomes and mobilomes useful? Front Microbiol 11:1376. https://doi.org/10.3389/fmicb.2020.01376.
11. Liu M, Li X, Xie Y, Bi D, Sun J, Li J, Tai C, Deng Z, Ou HY. 2019. ICEberg 2.0: an updated database of bacterial integrative and conjugative elements. Nucleic Acids Res 47:D660–D665. https://doi.org/10.1093/nar/gky1123.
12. Krawczyk PS, Lipinski L, Dziembowski A. 2018. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res 46:e35. https://doi.org/10.1093/nar/gkx1321.
13. Che Y, Xia Y, Liu L, Li AD, Yang Y, Zhang T. 2019. Mobile antibiotic resistome in wastewater treatment plants revealed by Nanopore metagenomic sequencing. Microbiome 7. https://doi.org/10.1186/s40168-019-0663-0.
14. Calderón-Franco D, van Loosdrecht MCM, Abeel T, Weissbrodt DG. 2021. Free-floating extracellular DNA: systematic profiling of mobile genetic elements and antibiotic resistance from wastewater. Water Res 189:116592. https://doi.org/10.1016/j.watres.2020.116592.
15. Alcock BP, Raphenya AR, Lau TTY, Tsang KK, Bouchard M, Edalatmand A, Huynh W, Nguyen A-LV, Cheng AA, Liu S, Min SY, Miroshnichenko A, Tran H-K, Werfalli RE, Nasir JA, Oloni M, Speicher DJ, Florescu A, Singh B, Faltyn M, Hernandez-Koutoucheva A, Sharma AN, Bordeleau E, Pawlowski AC, Zubyk HL, Dooley D, Griffiths E, Maguire F, Winsor GL, Beiko RG, Brinkman FSL, Hsiao WWL, Domselaar GV, McArthur AG. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 48:D517–D525.
16. Yin X, Jiang XT, Chai B, Li L, Yang Y, Cole JR, Tiedje JM, Zhang T. 2018. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics 34:2263–2270. https://doi.org/10.1093/bioinformatics/bty053.
17. Grazziotin AL, Koonin EV, Kristensen DM. 2017. Prokaryotic virus orthologous groups (pVOGs): a resource for comparative genomics and protein family annotation. Nucleic Acids Res 45:D491–D498. https://doi.org/10.1093/nar/gkw975.
18. Camarillo-Guerrero LF, Almeida A, Rangel-Pineros G, Finn RD, Lawley TD. 2021. Massive expansion of human gut bacteriophage diversity. Cell 184:1098–1109.e9. https://doi.org/10.1016/j.cell.2021.01.029.
19. Siguier P, Perochon J, Lestrade L, Mahillon J, Chandler M. 2006. ISfinder: the reference centre for bacterial insertion sequences. Nucleic Acids Res 34. https://doi.org/10.1093/nar/gkj014.
20. Jiang X, Hall AB, Xavier RJ, Alm EJ. 2019. Comprehensive analysis of chromosomal mobile genetic elements in the gut microbiome reveals phylum-level niche-adaptive gene pools. PLoS One 14:e0223680. https://doi.org/10.1371/journal.pone.0223680.
21. Douarre PE, Mallet L, Radomski N, Felten A, Mistou MY. 2020. Analysis of COMPASS, a new comprehensive plasmid database revealed prevalence of multireplicon and extensive diversity of IncF plasmids. Front Microbiol 11:483. https://doi.org/10.3389/fmicb.2020.00483.
22. Pruitt KD, Tatusova T, Maglott DR. 2007. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35:D61–D65. https://doi.org/10.1093/nar/gkl842.
23. Leplae R, Lima-Mendez G, Toussaint A. 2009. ACLAME: a classification of mobile genetic elements, update 2010. Nucleic Acids Res 38.
24. Durrant MG, Li MM, Siranosian BA, Montgomery SB, Bhatt AS. 2020. A bioinformatic analysis of integrative mobile genetic elements highlights their role in bacterial adaptation. Cell Host Microbe 27:140–153.e9. https://doi.org/10.1016/j.chom.2019.10.022.
25. Ellabaan MMH, Munck C, Porse A, Imamovic L, Sommer MOA. 2021. Forecasting the dissemination of antibiotic resistance genes across bacterial genomes. Nat Commun 12:1–10. https://doi.org/10.1038/s41467-021-22757-1.
26. Li L-G, Xia Y, Zhang T. 2017. Co-occurrence of antibiotic and metal resistance genes revealed in complete genome collection. ISME J 11:651–662. https://doi.org/10.1038/ismej.2016.155.
27. Siguier P, Gourbeyre E, Chandler M. 2014. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev 38:865–891. https://doi.org/10.1111/1574-6976.12067.
28. Escudero JA, Loot C, Nivina A, Mazel D. 2015. The integron: adaptation on demand. Microbiol Spectr 3. https://doi.org/10.1128/microbiolspec.MDNA3-0019-2014.
29. Craig NL. 2015. A moveable feast: an introduction to mobile DNA, p 3–39. *In* Mobile DNA III. Wiley.
30. Veress A, Nagy T, Wilk T, Kömüves J, Olasz F, Kiss J. 2020. Abundance of mobile genetic elements in an *Acinetobacter lwoffii* strain isolated from Transylvanian honey sample. Sci Rep 10. https://doi.org/10.1038/s41598-020-59938-9.
31. Chen J, Li J, Zhang H, Shi W, Liu Y. 2019. Bacterial heavy-metal and antibiotic resistance genes in a copper tailing dam area in northern China. Front Microbiol 10:1916. https://doi.org/10.3389/fmicb.2019.01916.
32. Yoo K, Yoo H, Lee J, Choi EJ, Park J. 2020. Exploring the antibiotic resistome in activated sludge and anaerobic digestion sludge in an urban wastewater treatment plant via metagenomic analysis. J Microbiol 58:123–130. https://doi.org/10.1007/s12275-020-9309-y.
33. Lanza VF, Baquero F, Martínez JL, Ramos-Ruíz R, González-Zorn B, Andremont A, Sánchez-Valenzuela A, Ehrlich SD, Kennedy S, Ruppé E, van Schaik W, Willems RJ, de la Cruz F, Coque TM. 2018. In-depth resistome analysis by targeted metagenomics. Microbiome 6:11–14. https://doi.org/10.1186/s40168-017-0387-y.
34. Pfeifer E, Moura De Sousa JA, Touchon M, Rocha EPC. 2021. Bacteria have numerous distinctive groups of phage-plasmids with conserved phage and variable plasmid gene repertoires. Nucleic Acids Res 49:2655–2673. https://doi.org/10.1093/nar/gkab064.
35. Lu Y, Zeng J, Wu B, Shunmei E, Wang L, Cai R, Zhang N, Li Y, Huang X, Huang B, Chen C. 2017. Quorum sensing N-acyl homoserine lactones-SdiA suppresses Escherichia coli-Pseudomonas aeruginosa conjugation through inhibiting tral expression. Front Cell Infect Microbiol 7:7. https://doi.org/10.3389/fcimb.2017.00007.
36. Mustard JA, Little JW. 2000. Analysis of Escherichia coli RecA interactions with LexA, λ CI, and UmuD by site-directed mutagenesis of recA. J Bacteriol 182:1659–1670. https://doi.org/10.1128/JB.182.6.1659-1670.2000.
37. Cahill J, Young R. 2019. Phage lysis: multiple genes for multiple barriers. Adv Virus Res 103:33–70. https://doi.org/10.1016/bs.aivir.2018.09.003.
38. Smyshlyaev G, Bateman A, Barabas O. 2021. Sequence analysis of tyrosine recombinases allows annotation of mobile genetic elements in prokaryotic genomes. Mol Syst Biol 17:e9880. https://doi.org/10.15252/msb.20209880.
39. Partridge SR, Kwong SM, Firth N, Jensen SO. 2018. Mobile genetic elements associated with antimicrobial resistance. Clin Microbiol Rev 31. https://doi.org/10.1128/CMR.00088-17.
40. Orlek A, Phan H, Sheppard AE, Doumith M, Ellington M, Peto T, Crook D, Walker AS, Woodford N, Anjum MF, Stoesser N. 2017. Ordering the mob: insights into replicon and MOB typing schemes from analysis of a curated dataset of publicly available plasmids. Plasmid 91:42–52. https://doi.org/10.1016/j.plasmid.2017.03.002.
41. Hui W, Zhang W, Kwok L-Y, Zhang H, Kong J, Sun T. 2019. A novel bacteriophage exclusion (BREX) system encoded by the pglX gene in Lactobacillus casei Zhang. Appl Environ Microbiol 85. https://doi.org/10.1128/AEM.01001-19.
42. Millman A, Melamed S, Amitai G, Sorek R. 2020. Diversity and classification of cyclic-oligonucleotide-based anti-phage signalling systems. Nat Microbiol 5:1608–1615. https://doi.org/10.1038/s41564-020-0777-y.
43. Chen SH, Byrne RT, Wood EA, Cox MM. 2015. Escherichia coli radD(yejH) gene: a novel function involved in radiation resistance and double-strand break repair. Mol Microbiol 95:754–768. https://doi.org/10.1111/mmi.12885.
44. Costa TRD, Harb L, Khara P, Zeng L, Hu B, Christie PJ. 2021. Type IV secretion systems: advances in structure, function, and activation. Mol Microbiol 115:436–452. https://doi.org/10.1111/mmi.14670.
45. Schmartz GP, Hartung A, Hirsch P, Kern F, Fehlmann T, Müller R, Keller A. 2022. PLSDB: advancing a comprehensive database of bacterial plasmids. Nucleic Acids Res 50:D273–D278. https://doi.org/10.1093/nar/gkab1111.
46. Ross K, Varani AM, Snesrud E, Huang H, Alvarenga DO, Zhang J, Wu C, McGann P, Chandler M. 2021. TnCentral: a prokaryotic transposable element database and web portal for transposon analysis. mBio 12:e0206021. https://doi.org/10.1128/mBio.02060-21.
47. Che Y, Yang Y, Xu X, Brinda K, Polz MF, Hanage WP, Zhang T. 2021. Conjugative plasmids interact with insertion sequences to shape the horizontal transfer of antimicrobial resistance genes. Proc Natl Acad Sci U S A 118.

48. Chandler M, Gellert M, Lambowitz AM, Rice PA, Sandmeyer SB. 2015. Mobile DNA III. Wiley.
49. Hyatt D, Chen GL, LoCascio PF, Land ML, Larimer FW, Hauser LJ. 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11:119. https://doi.org/10.1186/1471-2105-11-119.
50. Buchfink B, Xie C, Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. Nat Methods 12:59–60. https://doi.org/10.1038/nmeth.3176.
51. Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nat Biotechnol 35:1026–1028. https://doi.org/10.1038/nbt.3988.
52. Todd AE, Orengo CA, Thornton JM. 2001. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307:1113–1143. https://doi.org/10.1006/jmbi.2001.4513.
53. Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P. 2016. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Res 44:D286–D293. https://doi.org/10.1093/nar/gkv1248.
54. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. Nucleic Acids Res 40. https://doi.org/10.1093/nar/gkt1223.