Published online 3 November 2017 *Nucleic Acids Research, 2018, Vol. 46, Database issue* ***D851–D860***
doi: 10.1093/nar/gkx1068

RefSeq: an update on prokaryotic genome annotation and curation

Daniel H. Haft*, Michael DiCuccio, Azat Badretdin, Vyacheslav Brover, Vyacheslav Chetvernin, Kathleen O'Neill, Wenjun Li, Farideh Chitsaz, Myra K. Derbyshire, Noreen R. Gonzales, Marc Gwadz, Fu Lu, Gabriele H. Marchler, James S. Song, Narmada Thanki, Roxanne A. Yamashita, Chanjuan Zheng, Françoise Thibaud-Nissen, Lewis Y. Geer, Aron Marchler-Bauer and Kim D. Pruitt

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 45 Center Drive, Bethesda, MD 20892-6511, USA

Received September 15, 2017; Revised October 17, 2017; Editorial Decision October 18, 2017; Accepted October 25, 2017

ABSTRACT

The Reference Sequence (RefSeq) project at the National Center for Biotechnology Information (NCBI) provides annotation for over 95 000 prokaryotic genomes that meet standards for sequence quality, completeness, and freedom from contamination. Genomes are annotated by a single Prokaryotic Genome Annotation Pipeline (PGAP) to provide users with a resource that is as consistent and accurate as possible. Notable recent changes include the development of a hierarchical evidence scheme, a new focus on curating annotation evidence sources, the addition and curation of protein profile hidden Markov models (HMMs), release of an updated pipeline (PGAP-4), and comprehensive re-annotation of RefSeq prokaryotic genomes. Antimicrobial resistance proteins have been reannotated comprehensively, improved structural annotation of insertion sequence transposases and selenoproteins is provided, curated complex domain architectures have given upgraded names to millions of multidomain proteins, and we introduce a new kind of annotation rule—BlastRules. Continual curation of supporting evidence, and propagation of improved names onto RefSeq proteins ensures that the functional annotation of genomes is kept current. An increasing share of our annotation now derives from HMMs and other sets of annotation rules that are portable by nature, and available for download and for reuse by other investigators. RefSeq is found at https://www.ncbi.nlm.nih.gov/refseq/.

INTRODUCTION

The International Nucleotide Sequence Database Collaboration (INSDC), consisting of GenBank, the European Nucleotide Archive, and the DNA Data Bank of Japan, houses a vast archive of submitted sequence data (1). INSDC provides an indispensable complement to the similarly vast published literature on genomes, genes, and proteins. Its inherently archival nature, however, leaves a role for a companion resource in which consistent methods of structural annotation are used to facilitate comparative genomics, updated functional annotation is provided to reflect new characterizations that appear in the literature, misattributions of taxonomy are corrected, and genomes with poor sequencing and assembly quality or with significant contamination are screened out. In April 1999, NCBI introduced the Reference Sequences (RefSeq) project (2) to provide users with a resource that ensures assembly quality, is updated continuously with new information, assigns informative names to genes, provides some annotation for every gene found in each genome it analyzes, and supports comparative studies by using consistent structural and functional annotation methods. Since the first data release of 3439 human transcript and protein records, RefSeq has grown to encompass over 71 000 organisms, over 19 million transcripts, and over 88 million proteins (July 2017 RefSeq release 83).

RefSeq uses tailored data models and process flows to provide reference collections for eukaryotes, viruses and prokaryotes. For bacterial and archaeal genomes, the focus of this article, RefSeq uses a single annotation pipeline, PGAP (Prokaryotic Genome Annotation Pipeline) (3) and generates structural and functional annotation for all genomes that meet quality requirements. Because the identical protein sequence can be predicted by translating similar coding regions from thousands of bacterial genome assemblies, RefSeq in 2013 introduced a set of non-redundant

*To whom correspondence should be addressed. Tel: +1 301 594 7689; Email: haftdh@ncbi.nlm.nih.gov protein sequences, each with a single name and each potentially representing a large number of identical protein sequences encoded by many genomes (4). The RefSeq group has been steadily working to further improve the quality of structural annotation on prokaryotic genome assemblies and functional annotation for the non-redundant protein set. As PGAP is also provided as an annotation service to GenBank submitters, many bacterial genome submissions to GenBank are getting similarly improved annotation.

Improvements to RefSeq prokaryotic genome annotation system are ongoing. This article discusses our goals, significant changes to the PGAP pipeline through its upgrade to version 4.x (currently 4.2), a comprehensive reannotation of RefSeq prokaryotic genomes, the evidence system we have been developing, and our progress to date in improving protein annotations through use of additional evidence sources and a newly instituted evidence hierarchy scheme. We highlight here a significant change in curator focus from reviewing and correcting sets of proteins to instead building and refining an evidence layer that comes from a growing collection of expert-curated annotation rules.

RefSeq PROKARYOTIC PROTEIN DATA MODEL—A SINGLE SEQUENCE AND NAME FOR ALL OCCURRENCES OF AN EXACT PROTEIN SEQUENCE

Central to RefSeq's mission for prokaryotic genomes is the non-redundant protein data model (3,4). An accession number that begins with 'WP_' signifies one RefSeq non-redundant prokaryotic protein sequence, corresponding to identical translations from at least one, but sometimes thousands of coding sequence regions (CDS) annotated on RefSeq genomes. Each RefSeq non-redundant protein represents an **Identical Protein Group** (IPG) that includes INSDC proteins as well as translations from complete or whole genome shotgun sequencing (WGS) RefSeq assemblies. When a single protein is found on multiple genomes, the DNA of these coding regions may be identical, or may differ by synonymous codon changes. RefSeq accession number WP_000059093.1, 'chromosomal replication initiator protein DnaA,' represents the translation of the *dnaA* gene of *Salmonella enterica* subsp. *enterica serovar Typhimurium str. LT2*, for example, whose sequence was reported in 2001 (PMID: 11677609). But the IPG (https://www.ncbi.nlm.nih.gov/ipg/WP_000059093.1) also points to CDS feature translations from over 6000 additional annotated genome assemblies (counting both RefSeq and INSDC versions).

NCBI assesses the taxonomic placements of member proteins of an identical protein group, in order to assign a taxon for its non-redundant protein representative. A single WP protein may represent proteins from numerous species because of horizontal gene transfer. Plasmids encoding WP_004201164.1, the 'subclass B1 metallo-beta-lactamase NDM-1', for example, have spread so broadly (reaching *Acinetobacter baumannii, Enterobacter cloacae, Escherichia coli, Klebsiella pneumoniae* etc.) that RefSeq assigns the suitably broad taxonomic assignment 'Bacteria' to this protein and to the Identical Protein Group it represents.

If a new characterization is ready to be published for a previously known protein with an existing Identical Protein Group, it is appropriate to cite the 'WP_' accession. Using an accession number to represent a protein and all genes that encode it exactly provides greater clarity than merely describing the protein or naming the gene. Purely descriptive information is much harder to link to actual sequences and often turns out to be ambiguous.

A critical advantage of the non-redundant protein data model is that it allows on-going biocuration and automated updates to functional annotation after the initial run of the PGAP annotation pipeline. As more information about the function of a protein is discovered, RefSeq biocurators can add or improve on protein names that are associated with supporting homology evidence. Updated names then propagate onto matching records. For example, the protein from the locus named *Rv0693* of the genome of *Mycobacterium tuberculosis* H37Rv, sequenced in 1998 (5), was annotated as 'probable coenzyme PQQ synthesis protein E', a reasonable guess at the time. More recent work now suggests the name 'mycofactocin radical SAM maturase' for it (6), and for identical gene products from over 9000 annotated genomes. The identical protein group report at https://www.ncbi.nlm.nih.gov/ipg/WP_003403490.1 is over 200 pages long, with RefSeq annotations first, all identical and correct. In contrast, INSDC records seen in the last 100 pages show the heterogeneities, outdated information, and occasional errors expected in archival data.

RefSeq annotations of protein names, while designed to be terse, are also intended to inform as much as possible, and not merely to identify a protein. In our view, WP_000180747.1 from *Helicobacter pylori* should not be named 'cytotoxicity-associated immunodominant antigen,' which is ambiguous and outdated, nor 'exotoxin CagA,' which is specific but only mildly informative. RefSeq chose the name 'type IV secretion system oncogenic effector CagA.' This construction of the protein name alerts users to its biomedical importance, its involvement in a complex set of host-pathogen interactions, and its dependence on a whole system of additional proteins for its delivery and function (7). We designed the name for this protein as if we were building a hierarchy of protein names in which similarities in names reflect similarities in important characteristics such as molecular function, biological process, and/or evolutionary origin. The annotation that resulted should help those searching for all type IV secretion system effectors, or for all known oncogenic bacterial toxins (CagA is considered the first), and not only those users who already know what CagA is and how it functions.

Outside a very small number of highly studied model organisms, most proteins from prokaryotic genomes lack direct experimental characterization. Their molecular function and/or biological role must be inferred by homology. Frequently, only a very general type of functional name can be applied. RefSeq applies a name to every predicted protein in a genome using hierarchical rules (see below), even if the name is relatively low in information content. A growing class of researchers encounters genes and their protein translations as genomic features first, and only later will see the annotations they carry. Their methods may include basic genomics: inspecting a gene neighborhood, discovering a genomic island by pan-genome comparisons, finding sequence relatedness through BLAST searches, or simply browsing the predicted proteins from a newly sequenced species. Or the methods may involve high-throughput experimentation: RNA-Seq to follow gene expression levels, mass spectroscopy to find sites of post-translational modification, TnSeq to generate lists of genes essential for growth under various conditions, etc. For such users, low-information names are better than none. When viewed together in the context of gene neighborhoods or other groupings, names that convey relatively little information individually may still provide important clues that assist researchers aiming to characterize new systems of genes and their corresponding biological processes (8).

RefSeq has been working to make its functional annotation style as consistent as possible with that used by SwissProt/UniProt and preferred by GenBank, agreeing on actual names where possible, and on the guidance for how protein names should be structured in our respective databases. A jointly produced guide to recommended usages in protein naming remains in draft form at this writing, but should be made publicly available by both groups within a few months. This document will explain elements of protein naming as conducted at the respective databases and suggest a style that submitters of annotation to INSDC may use. The actual names that RefSeq applies are tightly linked to the forms of evidence we use to make the name assignments. RefSeq's use of a hierarchy of homology evidence to annotation rules for naming proteins is described in later sections.

UPDATES TO PGAP: IMPROVEMENTS TO STRUCTURAL ANNOTATION

In July 2015, NCBI announced the release of the PGAP-3.x series of its annotation pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/release_notes/) (3), used for both GenBank submissions and for processing of RefSeq bacterial assemblies. The release of PGAP-3.x was followed by a global reannotation of all RefSeq assemblies, with the intention of standardizing annotation across all genomes: all assemblies available were annotated by the same software methods. This global reannotation was a watershed moment for RefSeq, since we could declare with confidence a consistent level of annotation quality across all RefSeq genomes.

PGAP uses homology methods to predict protein coding regions, aided by the an *ab initio* gene-finding tool GeneMarkS+ (9) for regions where homology evidence to help select a reading frame is lacking. This approach gives consistency in structural annotation within a species, and is indispensable for the task of detecting genes that are disrupted by mutation or by sequencing/assembly errors, but it suffers several flaws inherent to a feed-forward system. Past errors of missing a gene entirely, choosing an incorrect start site, treating gene fragments as complete genes, or making false-positive gene predictions can be replicated automatically on novel genomes or during periodic reannotation of existing assemblies. To reduce errors of this type and resolve other issues, we have modified the gene finding approach in the PGAP-4.x series.

PGAP-3.x relied on alignments to lineage-specific core protein clusters, and on *ab initio* predictions from GeneMarkS+ outside of core gene regions, to produce an initial set of gene predictions, which were then used to find additional evidence to refine structural predictions. PGAP-4.x eliminates this first pass. The new approach focuses on identifying all ORFs with evidence by homology that they represent real protein sequence. We now take all possible ORFs, translated from stop-to-stop, to allow maximal exposure of each ORF to known lines of evidence, even for genes that have been disrupted. Each identified ORF is tested against a collection of libraries of protein profile hidden Markov models (HMMs) and then against the protein clusters. For computational efficiency, BLAST searches are skipped for any ORF that is fully contained in the HMM hit region to a different ORF in one of the other five reading frames. We include HMMs both for full-length proteins and for recognizable homology domains. Currently, NCBI uses HMM libraries imported from TIGRFAMs (10) and Pfam (11); a new library created from our previously described PRK cluster set (12); and additional HMMs, called NCBIfams (https://ftp.ncbi.nlm.nih.gov/hmm/), custom-built to identify high-value protein families, including proteins involved in antimicrobial resistance. Our current HMM collection includes over 33 000 HMMs, covering a wide variety of protein families and known protein domains. This collection ensures proper structural annotation of hard-to-find small proteins such as bacteriocin precursors, leader peptides, PqqA, phenol-soluble modulins, methanobactins, ribosomal protein L36, etc. It helps steer gene-calling to the correct reading frame in highly GC-rich DNA, where purely *ab initio* gene-finding methods may be error-prone. Regular import of updated HMM libraries such as Pfam, plus in-house development of new NCBIfam HMMs, continually improves our structural annotation. PGAP makes no direct use of proteomics or transcriptomics data, but those (including us) who build new protein profile hidden Markov models consider evidence from many types of studies: proteomics, transcriptomics, comparative genomics, mutant phenotypes, crystallography, and direct functional assays of purified proteins.

Identified ORF HMM hits that meet our quality criteria are mapped to the genome. Lineage-specific high-quality reference proteins, and proteins identified by ORF hits to the protein cluster set, are (re)aligned to the genome, using ProSplign (13). PGAP examines the evidence for selection of a start site, or to spot genes with frameshifts or other disruptions that should be marked as such. It also detects known exceptions to the normal rules of translation, such as programmed frameshifts in the genes for many transposases or in the *prfB* gene for release factor 2, or translation of a UGA codon as selenocysteine to form a known selenoprotein. Coding regions that would overlap CRISPR repeats or tRNA or rRNA genes excessively (by more than 0, 50 or 15 bp, respectively) are disallowed. GeneMark S+ makes *ab initio* coding region predictions for genomic regions that lack HMM or protein evidence, and selects start sites for ORFs whose evidence comes from HMMs.

Overall, our workflow for structural annotation has been simplified. We run GeneMark S+ and BLAST just once during the PGAP-4.x revamped structural annotation pathway, in contrast to its use at two different points in the prior PGAP-3.x workflow. The new workflow can be seen in the flowchart for PGAP-4.x, shown in Figure 1. As with PGAP-3.x, our annotation of deeply sequenced species and genera is heavily weighted toward homology-based evidence, and toward selection of start sites by a consensus from homology evidence rather than *ab initio* inference. Adding evidence from HMM hits to evidence from BLAST matches *vs.* the protein cluster set provides a better guarantee that structural annotation will assign genes to the correct reading frames. For less-well sequenced taxa, this improvement relative to PGAP-3.x is especially important. However, some problems still persist in identifying the most likely start sites, telling which genes are disrupted by mutations or by sequencing or assembly problems, and correctly assessing which reading frames have homology evidence that is more reliable than *ab initio* gene-finder predictions.

RefSeq prokaryotic reannotation 2017

In 2017, following the release of PGAP-4.1 and incorporation of multiple new lines of evidence, the RefSeq team undertook a global reannotation of all prokaryotic genome assemblies. Of the 89 732 prokaryotic genome assemblies available as of February 2017, 86 079 were successfully annotated and met criteria for inclusion into RefSeq. Subsequently, additional assemblies have been added to the collection, bringing the current total to 95,336 as of this writing (September 2017). Table 1 shows the growth in recent years of RefSeq prokaryotic genomes and annotated nonredundant protein sequence records. The twenty most abundant species among RefSeq prokaryotic genomes, all of which are pathogenic bacteria, are shown in Supplementary Table S1, in supplementary materials. RefSeq prokaryotic genome annotation can be downloaded as individual genomes or batches of genomes from NCBI's Assembly resource after doing a query (https://www.ncbi.nlm.nih.gov/assembly/) or by ftp (https://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/-or/archaea/) using the file 'assembly_summary.txt' as a guide.

Biocuration efforts leading up to this reannotation event focused heavily on three specific aspects of evidence recovery: removal of spurious or fragmentary proteins from our clustered protein set; identification of the highest quality reference genome annotation, with exclusion of proteins lacking significant support; and generation of new lines of evidence covering protein families not represented in our protein clusters or HMMs. Specific effort was taken in several gene families. NCBI placed emphasis on identifying a high-quality set of reference transposase proteins. Transposases present specific challenges in annotation due to their frequent frameshifts, degradation into pseudogenes, and occurrence as repeats that disrupt assemblies. As a result, our protein cluster set was crowded with a number of short fragmentary proteins representing parts of transposases. Removing these fragments in favor of full-length reference transposase sequences results in a much more consistent annotation of transposases in all organisms.

Many genomes contain large numbers of transposase gene fragments, or encode transposases whose translation is inferred to span a programmed frameshift, or both. We have added a large number of full-length insertion sequence transposase reference sequences, from ISFinder (https://www-is.biotoul.fr/) (14), to our protein cluster set, to assist in their proper structural and functional annotation. Consequently, we have gained the ability to identify degraded transposase genes (with truncations, nonsense mutations, or frameshift mutations) and mark them as genomic features without presenting them as proteins. For some genomes, this change has increased considerably the numbers of features that we mark 'pseudo', with a concomitant reduction in the number of predicted protein products. This reduction alarms some users, but most deleted proteins seem to be very short, and highly dubious, and likely have no active biological role. For a subset of transposases, our structural reference sequence reflects an inferred programmed frameshift, and is marked as such. For homologs to this group, the frameshift is detected and is reported in the markup for the coding region feature, and a proper protein translation is produced. Supplementary Figure S1 in supplementary materials shows examples of transposase regions that were treated incorrectly in the last PGAP-3.x reannotation but are handled correctly now. The view is provided by visualization tool Genome Workbench (https://www.ncbi.nlm.nih.gov/tools/gbench/).

Similarly, our protein cluster set was found to be deficient in specialty proteins such as selenoproteins. As a result, selenoproteins were misannotated uniformly in PGAP-3.x. With PGAP-4.x, we modified our interpretations of alignments to correctly handle matching selenocysteine against genomic sequences, and added a high-quality reference set of exemplar selenoproteins from fifty of the most abundant prokaryotic selenoprotein families, including the selenophosphate synthase SelD itself, a required part of the selenocysteine incorporation system (15,16). With PGAP-4.x, NCBI now produces consistent high-quality annotation of most selenoproteins across applicable genera. The pipeline now finds all three formate dehydrogenase isoforms in *E. coli* K-12, all six selenoproteins expected in *Clostridioides difficile* CD196, and all eight expected in *Geobacter sulfurreducens* KN400 (16), plus many other selenoproteins from other species, apparently without a false-positive. It presents each selenoprotein as one continuous sequence that reads through a UGA codon and represents the selenocysteine as U, as in WP_095452254.1. However, the rarest and the most recently described selenoprotein families are not yet detected by PGAP.

Using hierarchical evidence to automate rule-based protein annotation

Over the last couple of years, RefSeq has been phasing in the use of hierarchical homology evidence to perform functional annotation in prokaryotes, and to assist in structural annotation, while slowly phasing out use of the legacy protein cluster set. The evidence system provides a means to update annotations on existing RefSeq proteins, which now include over 75 million live (and 19 million deprecated) distinct protein sequences, and to record the provenance for every change to annotation. It also allows PGAP to better annotate new proteins entering RefSeq and to better prepare genomes for submission to GenBank.

NCBI has imported libraries of HMMs to use in annotation, including release 15 of TIGRFAMs (4488 mod-

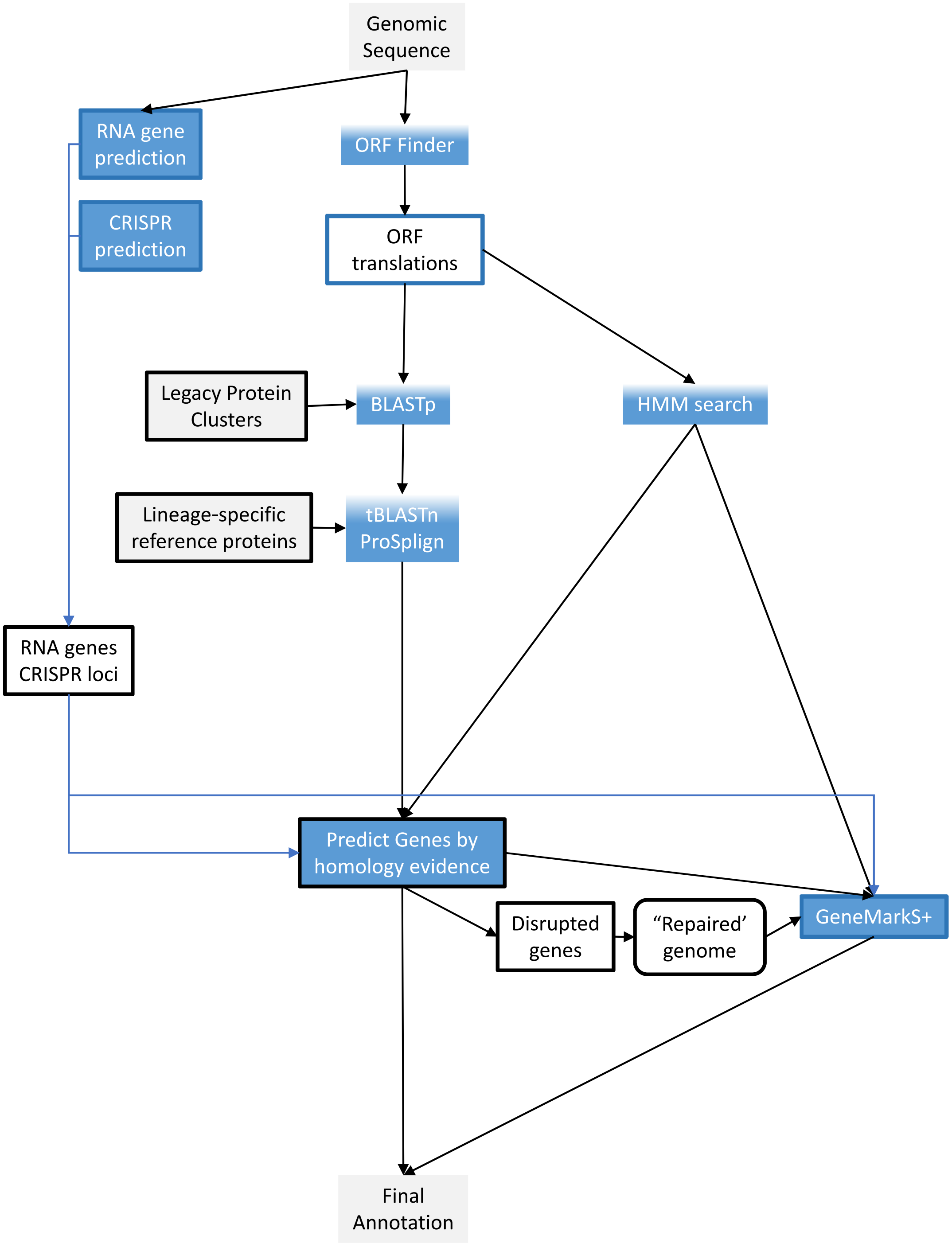

Figure 1. New workflow for structural annotation by the PGAP 4.x series pipeline. Computational processes are shown in blue, data in white or gray. GeneMarkS+ provides *ab initio* prediction of protein-coding genes, but in the context of hints from homology-based evidence, including HMM evidence for the first time. The use of ORFfinder to produce every stop-to-stop translations, and HMM searching to find every translation with an HMM hit, are steps first introduced in the PGAP-4.1 release. The pipeline detects both disrupted genes (e.g. pseudogenes) and exceptional reading frames (e.g. selenoproteins).

Table 1. Growth of RefSeq genomes and RefSeq proteins. Twenty pathogenic bacterial species account for more than half of the prokaryotic genomes included in RefSeq (54 663), and or a substantial share of incoming genomes. Consequently, the number of nonredundant RefSeq proteins is growing somewhat more slowly than the number of genomes

Date	Number of Genomes	Number of Non-redundant Proteins
1 January 2013	7 503	0
1 January 2014	14 762	16 829 357
1 January 2015	28 146	26 883 513
1 January 2016	52 571	41 555 561
1 January 2017	77 292	55 802 502
12 September 2017	95 336	75 878 570

els in use) and release 30 of Pfam (16 306 models in use). RefSeq curators have reviewed protein names assigned by 3433 TIGRFAMs models so far, and in the process revised the names to 1638 of them. This process included a review of nearly every equivalog-level model in TIGRFAMs. The list of name changes to TIGRFAMs models can be found at https://ftp.ncbi.nlm.nih.gov/hmm/ in the file TIGRFAMs.tsv. Similarly, biocurators determined product names for 4052 Pfam models, in 4025 cases different from the name imported with the HMM. This is expected, because Pfam often provides a name for a domain within a protein, while PGAP needs to assign a name to a protein. Some examples of names PGAP applies that differ from the original Pfam names are PF00890.22 (‘FAD-binding protein’ instead of ‘FAD binding domain’), PF03249.11 (‘type-specific antigen TSA56’ instead of ‘Type specific antigen’, and PF01694.20 (‘rhomboid family intramembrane serine protease’ instead of ‘Rhomboid family’).

In addition to importing libraries, NCBI has done extensive work to generate new HMMs. Starting with 7471 established PRK clusters (12), NCBI built and calibrated a new collection of PRK-derived HMMs, named NCBIfam-PRK. Clusters were subdivided if necessary to simply the process of estimating cutoffs, leading to 11 498 new HMMs in all. In cases where a single PRK cluster generated multiple HMMs, cutoffs were set in such a way that several of the resulting HMMs may match a single member protein. We also built new models, designated NCBIfams. Of these, 580 HMM models, named NCBIfam-AMR, were built to find and annotate antimicrobial resistance (AMR) proteins, described in more detail below. Another 142 non-AMR models were built to address various issues in structural or functional annotation (NCBIfam-gen). As with the imported TIGRFAM and Pfam libraries, NCBIfams are built with, and compatible with, HMMER3 (17). NCBIfam data is available at https://ftp.ncbi.nlm.nih.gov/hmm/.

RefSeq staff undertook an intensive effort to identify and reannotate acquired and intrinsic antimicrobial resistance (AMR) proteins, including the beta-lactamases. Resources used during this biocuration effort included ResFams (18), CARD (19), ResFinder (20), among others. A single amino acid change separating two different beta-lactamase alleles can matter both for clinical practice and for epidemiological studies, so a rich nomenclature has emerged. There are specific allele names for most families that have been mobilized by plasmids or transposons. NCBI now maintains the registry of beta-lactamase allele names that previously was provided through the Lahey Clinic resource (21). We have developed a large collection of overlapping HMMs for functional AMR proteins, some broad in their scope and some highly specific. An exact match to a defined allele is now our most specific form of evidence. Currently, we use allele-level naming only for AMR proteins.

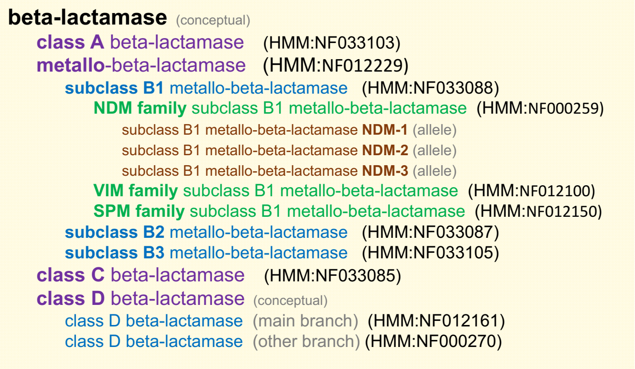
beta-lactamase (conceptual)
class A beta-lactamase (HMM:NF033103)
metallo-beta-lactamase (HMM:NF012229)
subclass B1 metallo-beta-lactamase (HMM:NF033088)
NDM family subclass B1 metallo-beta-lactamase (HMM:NF000259)
subclass B1 metallo-beta-lactamase **NDM-1** (allele)
subclass B1 metallo-beta-lactamase **NDM-2** (allele)
subclass B1 metallo-beta-lactamase **NDM-3** (allele)
VIM family subclass B1 metallo-beta-lactamase (HMM:NF012100)
SPM family subclass B1 metallo-beta-lactamase (HMM:NF012150)
subclass B2 metallo-beta-lactamase (HMM:NF033087)
subclass B3 metallo-beta-lactamase (HMM:NF033105)
class C beta-lactamase (HMM:NF033085)
class D beta-lactamase (conceptual)
class D beta-lactamase (main branch) (HMM:NF012161)
class D beta-lactamase (other branch) (HMM:NF000270)

Figure 2. A partially expanded view of the homology evidence and protein naming hierarchy used in RefSeq and PGAP annotation. Four families of beta-lactamases are shown (A, metallo, C, and D), each of which is more similar to various hydrolases of other substrates, such as RNA, than to any members of the other beta-lactamase classes. For each class, a protein profile HMM identifies members and suggests a protein product name, but further expansion of the hierarchy can reveal multiple child families, each identified by a more specific HMM that receives a higher precedence during annotation. The hierarchy of evidence largely follows an implicit hierarchy of protein names, with exceptions necessary occasionally, as when unrelated proteins perform closely related functions.

The hierarchical approach we use for protein naming, and the close correspondence between the new annotation rules we develop and the families of proteins they represent, is illustrated Figure 2. Similar names do not necessarily indicate homologous proteins; the Ambler class B beta-lactamases (metallo-beta-lactamases) are unrelated to the other Ambler classes (22). Therefore, Figure 2 does not show any HMM that can find all beta-lactamases and separate them from other proteins. Further down the hierarchy, however, names reflect progressively narrower homology groupings from the literature. Each more specific name is supported by a defining HMM, until stopping at the level of individual beta-lactamase alleles. Thus, ‘subclass B1 metallo-beta-lactamase NDM-1’ (an allele) belongs to a broader classification ‘NDM family subclass B1 metallo-beta-lactamase’ (HMM NF000259), which belongs to ‘subclass B1 metallo-beta-lactamase’ (HMM NF033088), and so on.

In addition to HMMs, RefSeq has added another major source of functional annotation. This is CDD-SPARCLE

(23), a recently described resource that leverages the Conserved Domain Database (CDD) (24). CDD has aggregated a set of protein family definitions from Pfam and TIGRFAMs HMM libraries, PRK clusters, COGs (25), CDD-curated models which are organized into sub-family hierarchies, as well as various other sources, and generated position-specific scoring matrices (PSSMs), the RPS-BLAST analog of HMMs. Using their extensive library of domains, CDD identified all unique domain architectures, considering domain order but ignoring repeat numbers for tandem domains, and then began curating protein names to assign per architecture, rather than according to just one of several identified domains. So far, over 14 000 distinct architectures have been curated. These architectures represent more than 30 million IPGs with bacterial RefSeq records, and were used to reannotate over 19 million for which they represented our highest-ranked evidence.

CDD-SPARCLE biocurators focused primarily on the most abundant architectures that were not well-covered by high-ranking evidence from other sources, so the overall impact on RefSeq protein annotation has been large. A notable set of proteins annotated by CDD-SPARCLE architectures was response regulator proteins. Some response regulator proteins contain DNA-binding regions and function as transcriptional regulators. Others lack DNA-binding, and process information for the cell in a different way. Most response regulator proteins partner with a separate histidine kinase, but many are hybrid proteins with both functionalities in a single polypeptide. SPARCLE architectures have provided extensive name cleanup, and enhancement of the information that protein names convey, for response regulator proteins that had no higher-ranking evidence to use.

At present, CDD-SPARCLE architecture determination is available on-line through CDD's domain annotation service (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Architecture characterizations and definitions, as well as off-line domain annotation tools are distributed via CDD's FTP site (https://ftp.ncbi.nih.gov/pub/mmdb/cdd/).

Introducing BlastRules

A limitation of HMMs used in annotation, at NCBI and probably elsewhere, is that the shared collection of models in HMM libraries available for download and distribution, including Pfam, TIGRFAMs and the new NCBIfam-PRK set, have tended to focus on the largest, most-widespread families. The long, thin tail of less abundant protein families that have been discussed in the literature are rather poorly represented in all collections of protein family definitions. These less common, but often high interest, proteins include virulence proteins known from pathogenicity islands in *Salmonella enterica*, or individually named members of a large paralogous family of transporters in *Mycobacterium tuberculosis*, or surface antigens (and vaccine candidates) in *Bordetella pertussis*, or O-antigen flippases and polymerases that predict serology in *E. coli*. First examples of rare enzymes, natural product biosynthesis proteins, adhesins, and many other proteins of specialization, rather than universality, are still poorly covered as defined protein families and poorly handled by automated annotation pipelines. The compounding problem is that building HMMs can be so labor-intensive.

To improve the annotation of highly specialized, high interest and relatively rare proteins at relatively low cost, NCBI has started a new collection, BlastRules. A BlastRule requires only (i) a name to apply during annotation, (ii) one or more reference proteins to use in BLAST searches and (iii) parameters to determine if a BLAST match to a reference protein is sufficient. Additional annotations, including gene symbols, PubMed identifiers, and written remarks about the protein, can also be added. So far, NCBI has generated over 1100 BlastRules and added them to the annotation pipeline. Release notes and a collection of BlastRules, consisting of rule identifiers, rule types, reference protein accessions, thresholds, annotation names, gene symbols, and PubMed identifiers is available at https://ftp.ncbi.nlm.nih.gov/pub/blastrules/.

One of the benefits of BlastRules is that it allows high-speed creation of an early stage annotation rule that biocurators in the future may revisit and expand. The process of building an HMM to define an entire family of homologous proteins with conserved function, such that false-positives and false-negatives are both rare, requires considerable effort and considerable expertise, so barriers to such HMM construction can be high. Often, published characterizations and currently available genomes fail to provide enough information to complete the task of building a high-quality HMM. BlastRules provides a means to build prototype versions for large numbers of new annotation rules, while acknowledging by design that these rules do not yet hit all true-positive homologs. This approach greatly lowers the barrier to working through much of the biocuration backlog, and turning knowledge from published papers into improved genome annotation. Sensitivity is sacrificed, initially, for the sake of speed. RefSeq curators hope to build large number of BlastRules in the next few years, and to work collaboratively with outside experts who would like to spur improvement to our global annotation of their favorite sets of proteins.

Interleaving annotation rules of different precedence from multiple sources

To provide correct priorities for HMMs from different sources, we assign each to a 'family type', which from highest rank to lowest are '*exception*', '*equivalog*', '*subfamily*' and '*domain*'. ***Equivalogs*** are different proteins that perform the same specific function because function was conserved since their last common ancestor; equivalog-level models make predictions that their member proteins indeed share one function. The definition comes from TIGRFAMs, where *equivalog*-level models are marked as such (26). Note that *equivalogs* are not necessarily orthologs (proteins that diverged through speciation events only), since lateral gene transfer may have occurred. Meanwhile, orthologs are not necessarily *equivalogs*, since orthology says nothing about conserved protein function. An *equivalog* HMM typically covers most of the length of any protein it hits. That length may include any number of recognizable homology domains that a resource such as Pfam may

Table 2. Coverage of RefSeq non-redundant proteins hierarchical evidence rules. A single protein may be supported by evidence of multiple types; this table shows counts of proteins having the highest precedence evidence (not counting those proteins a second time if they also have a lower precedence evidence). The precedence scores shown represent an arbitrary scale, but show how additional forms of evidence could be interleaved if an appropriate precedence is chosen. 75 878 570 proteins were analyzed; some evidence types with small protein counts are not shown. Once RefSeq or Conserved Domain Database biocurators construct and approve a protein product name, the HMM, CDD-SPARCLE, or other evidence becomes the basis of a fully automated rule for RefSeq annotation. Annotation improves over time as new rules are added that reach more proteins, or as rules capable of highly specific annotation overrule prior annotations based on less specific, lower-ranked rules

Evidence type	Relative precedence	Count of RefSeq proteins where the evidence is selected	Evidence level description
Allele	100	2 426	An annotation valid for exactly one protein sequence. Used only for antimicrobial resistance (AMR).
exception-level BlastRule	95	22 857	Very close full-length homologs of reference sequences, typically > = 94% identity. Used mostly for virulence factors.
equivalog HMM	70	16 529 578	Proteins with conserved specific function—a mature annotation rule with a curated product name
equivalog-level BlastRule	69	40 763	Proteins with conserved specific function—early stage annotation rule
CDD-SPARCLE domain architectures	60	19 107 223	Proteins with an exact combination of domains, recognized by Conserved Domain Database's RPS-BLAST tools rather than by HMMs.
subfamily HMM	55	1 047 045	Typically full-length homologs, somewhat variable in function, that deserve naming more specific than domain content provides
domain HMM	30	3 347 261	Proteins containing an HMM-defined domain—generally an independently folding region shared by proteins of various functions.
Pending evidence		22 972 109	An HMM or other classifier exists, but the annotation rule is not complete because the name to apply has not been curated.
No evidence	none	12 775 989	Proteins with no HMM, CDD-SPARCLE architecture, BlastRule, etc.

identify. Most *equivalog* HMMs hit no more than once per genome for the vast majority of genomes.

An ***exception*** family HMM conveys even more information than an *equivalog* HMM. It may identify, for example, a distinctive isozyme for some enzyme. *Exception*-level HMMs have the highest precedence among HMMs, but they have the narrowest scope, hitting relatively few proteins.

Just below the *equivalog* HMM in precedence is the ***subfamily*** HMM. *Subfamil*y (as the term is used here) models are broader in scope than *equivalog*, and thus less specific and lower in rank, although narrow compared to the set of all proteins that share at least one region of homology. Like an *equivalog* model, the *subfamily* model typically covers most of the length of any protein it hits, but it may routinely hit multiple proteins per genome. Below *subfamily* HMMs are ***domain*** HMMs, which may hit not only multiple proteins per genome, but a variety of proteins that exhibit very different architectures and functions. *Domain* HMMs, mostly taken from Pfam, become active for annotation in our system only after we construct a name that would be valid for all proteins containing the domain.

RefSeq's use of annotation rules interleaves HMM evidence with CDD-SPARCLE evidence and BlastRules evidence. Table 2 shows the hierarchical order of precedence for the major forms of evidence used, with a larger number indicating higher precedence, and the number of RefSeq proteins covered by that evidence type as of this writing. The scores shown for the relative precedence represent an arbitrary scale, but show how additional forms of evidence, such as combinations of HMMs, could be interleaved into the hierarchy in the future. For example, a fusion of two or more HMMs, where at least one is equivalog-level, might be assigned a precedence level higher than any of its component HMMs but lower than an exception-level BlastRules.

For new RefSeq proteins that the PGAP process cannot name by any other method, the fallback annotation method is blastp versus the proteins of the protein cluster set described above (see Figure 1). Although this cluster set is fine-grained and was seeded with many informative annotations, it also contains some outdated or erroneous annotations with lost provenance. PGAP uses fairly high stringency for BLAST matches to this cluster set, so new proteins that are below 40% identity to clusters, and not otherwise covered by another annotation rule, are named 'hypothetical protein'. RefSeq does not use this legacy cluster set to reannotate pre-existing protein records.

The current state of RefSeq annotation is that annotation rules with defined cutoffs and a portable nature are rapidly supplanting our earlier reliance on the protein cluster set. We are working continuously to complete the process ending reliance on this legacy cluster set for protein naming, in favor of evidence rules with clear provenance. Note that counting non-redundant protein sequences, rather than all (redundant) proteins encoded on all assemblies, overstates the fraction of proteins that lack evidence. A single non-redundant protein with evidence may be found on 1000 more different genome assemblies, while a protein lacking evidence is far more likely to be found on a single genome assembly.

FUTURE DIRECTIONS

This update on prokaryotic genome annotation in RefSeq describes progress at a time of rapid transition. Additional improvements to the system are planned but not yet fully implemented. Some of these are enumerated below.

While our goals for 2018 are to continue to provide RefSeq annotation for new prokaryotic genomes that meet our quality criteria, given the continued rapid growth in bacterial genome sequencing, we are exploring changing this policy. As part of this, we plan to further refine our criteria for identifying 'best' genome assemblies to track as exemplars for species (identified as reference or representative genomes in NCBI's Assembly resource). We anticipate engaging the community for feedback on both policy changes and usage trends for RefSeq prokaryotic genome and protein data using various approaches including brief surveys, community feedback through normal email channels, and discussion at scientific forums.

We plan to continue to actively invest in accumulating a high quality curated evidence layer with accompanying informative and standardized protein names in the coming year. We will focus on reviewing Pfam and PRK HMMs, adding more domain architectures, and building upon the new BlastRules initiative. Our goals are to provide a reusable resource that others can use as evidence for genome annotation, to provide informative and accurate protein names, to collaborate with external groups able to share high-quality biocuration data that we can readily absorb, and to reduce the number of proteins currently annotated as 'hypothetical protein'.

Our longer-term plans include associating more metadata to annotation rules, including EC numbers, GO terms, gene symbols, literature references, and explanatory comments. For the rules themselves, including HMMs derived from PRK clusters or built from scratch, we plan to present this content on NCBI web pages, in order to supplement the ftp repository now available. We plan to develop systems that will transfer annotations beyond the protein name from our HMMs and from other rules onto RefSeq protein records and genomes directly, to provide easy access to such content through a link to each rule's web page, or both.

Lastly, we note that the paradigm of our annotation system has changed. Curators work almost exclusively on creating or updating annotation rules that our systems will use for automated annotation, instead of working to manually to reannotate specific proteins. This approach provides RefSeq a long-term solution for each family of proteins covered by a new annotation rule, including member proteins that have not yet been sequenced. It also improves quality for genomes entering GenBank with annotation from PGAP. All groups working with prokaryotic genomic and metagenomic data face limitations in the power of their bioinformatic approaches because of the backlog in the biocuration tasks needed to build the requisite sets of genome annotation rules. Our libraries of HMMs, BlastRules, and domain architecture classifications are freely available. We welcome testing of our rules by others and contributions of suggested corrections or new rules, and we encourage other annotation groups to make sets of highly trusted annotation rules available to us, so groups can work cooperatively to reduce biology's backlog in biocuration.

AVAILABILITY

RefSeq is found at https://www.ncbi.nlm.nih.gov/refseq/. NCBIfam files are available for download at https://ftp.ncbi.nlm.nih.gov/hmm/. BlastRules files are available for download at https://ftp.ncbi.nlm.nih.gov/pub/blastrules/. SPARCLE (the Subfamily Protein Architecture Labeling Engine) can be found at https://www.ncbi.nlm.nih.gov/sparcle.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

FUNDING

Intramural Research Program of the NIH, National Library of Medicine. Funding for open access charge: Intramural Research Program of the NIH, National Library of Medicine.

Conflict of interest statement. None declared.

REFERENCES

1. Cochrane,G., Karsch-Mizrachi,I., Takagi,T. and International Nucleotide Sequence Database, C. (2016) The international nucleotide sequence database collaboration. *Nucleic Acids Res.*, **44**, D48–D50.
2. Maglott,D.R., Katz,K.S., Sicotte,H. and Pruitt,K.D. (2000) NCBI's LocusLink and RefSeq. *Nucleic Acids Res.*, **28**, 126–128.
3. Tatusova,T., DiCuccio,M., Badretdin,A., Chetvernin,V., Nawrocki,E.P., Zaslavsky,L., Lomsadze,A., Pruitt,K.D., Borodovsky,M. and Ostell,J. (2016) NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.*, **44**, 6614–6624.
4. Tatusova,T., Ciufo,S., Fedorov,B., O'Neill,K. and Tolstoy,I. (2014) RefSeq microbial genomes database: new representation and annotation strategy. *Nucleic Acids Res.*, **42**, D553–D559.
5. Cole,S.T., Brosch,R., Parkhill,J., Garnier,T., Churcher,C., Harris,D., Gordon,S.V., Eiglmeier,K., Gas,S., Barry,C.E. 3rd *et al.* (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature*, **393**, 537–544.
6. Khaliullin,B., Ayikpoe,R., Tuttle,M. and Latham,J.A. (2017) Mechanistic elucidation of the mycofactocin-biosynthetic radical S-adenosylmethionine protein, MftC. *J. Biol. Chem.*, **292**, 13022–13033.
7. Nishikawa,H. and Hatakeyama,M. (2017) Sequence polymorphism and intrinsic structural disorder as related to pathobiological performance of the helicobacter pylori CagA oncoprotein. *Toxins (Basel)*, **9**, E136.
8. Haft,D.H. (2015) Using comparative genomics to drive new discoveries in microbiology. *Curr. Opin. Microbiol.*, **23**, 189–196.
9. Borodovsky,M. and Lomsadze,A. (2014) Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Microbiol.*, **32**, doi:10.1002/0471250953.bi0405s35.
10. Haft,D.H., Selengut,J.D., Richter,R.A., Harkins,D., Basu,M.K. and Beck,E. (2013) TIGRFAMs and genome properties in 2013. *Nucleic Acids Res.*, **41**, D387–D395.
11. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A. *et al.* (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
12. Klimke,W., Agarwala,R., Badretdin,A., Chetvernin,S., Ciufo,S., Fedorov,B., Kiryutin,B., O'Neill,K., Resch,W., Resenchuk,S. *et al.* (2009) The national center for biotechnology information's protein clusters database. *Nucleic Acids Res.*, **37**, D216–D223.

13. Sayers,E.W., Barrett,T., Benson,D.A., Bolton,E., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Federhen,S. *et al.* (2011) Database resources of the national center for biotechnology information. *Nucleic Acids Res.*, **39**, D38–D51.
14. Siguier,P., Varani,A., Perochon,J. and Chandler,M. (2012) Exploring bacterial insertion sequences with ISfinder: objectives, uses, and future developments. *Methods Mol. Biol.*, **859**, 91–103.
15. Peng,T., Lin,J., Xu,Y.Z. and Zhang,Y. (2016) Comparative genomics reveals new evolutionary and ecological patterns of selenium utilization in bacteria. *ISME J.*, **10**, 2048–2059.
16. Zhang,Y., Romero,H., Salinas,G. and Gladyshev,V.N. (2006) Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome Biol.*, **7**, R94.
17. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
18. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2015) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 207–216.
19. Jia,B., Raphenya,A.R., Alcock,B., Waglechner,N., Guo,P., Tsang,K.K., Lago,B.A., Dave,B.M., Pereira,S., Sharma,A.N. *et al.* (2017) CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.*, **45**, D566–D573.
20. Kleinheinz,K.A., Joensen,K.G. and Larsen,M.V. (2014) Applying the ResFinder and VirulenceFinder web-services for easy identification of acquired antibiotic resistance and E. coli virulence genes in bacteriophage and prophage nucleotide sequences. *Bacteriophage*, **4**, e27943.
21. Jacoby,G.A. (2006) Beta-lactamase nomenclature. *Antimicrob. Agents Chemother.*, **50**, 1123–1129.
22. Ambler,R.P. (1980) The structure of beta-lactamases. *Philos. Trans. R Soc. Lond. B Biol. Sci.*, **289**, 321–331.
23. Marchler-Bauer,A., Bo,Y., Han,L., He,J., Lanczycki,C.J., Lu,S., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R. *et al.* (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Res.*, **45**, D200–D203.
24. Marchler-Bauer,A., Derbyshire,M.K., Gonzales,N.R., Lu,S., Chitsaz,F., Geer,L.Y., Geer,R.C., He,J., Gwadz,M., Hurwitz,D.I. *et al.* (2015) CDD: NCBI's conserved domain database. *Nucleic Acids Res.*, **43**, D222–226.
25. Galperin,M.Y., Makarova,K.S., Wolf,Y.I. and Koonin,E.V. (2015) Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res.*, **43**, D261–D269.
26. Haft,D.H., Loftus,B.J., Richardson,D.L., Yang,F., Eisen,J.A., Paulsen,I.T. and White,O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res.*, **29**, 41–43.