D1020–D1028 *Nucleic Acids Research, 2021, Vol. 49, Database issue* *Published online 3 December 2020*
doi: 10.1093/nar/gkaa1105

RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation

Wenjun Li †, Kathleen R. O'Neill†, Daniel H. Haft, Michael DiCuccio, Vyacheslav Chetvernin, Azat Badretdin, George Coulouris, Farideh Chitsaz, Myra K. Derbyshire, A. Scott Durkin, Noreen R. Gonzales, Marc Gwadz, Christopher J. Lanczycki, James S. Song, Narmada Thanki, Jiyao Wang, Roxanne A. Yamashita, Mingzhang Yang, Chanjuan Zheng, Aron Marchler-Bauer and Françoise Thibaud-Nissen*

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

Received September 15, 2020; Revised October 19, 2020; Editorial Decision October 20, 2020; Accepted November 02, 2020

ABSTRACT

The Reference Sequence (RefSeq) project at the National Center for Biotechnology Information (NCBI) contains nearly 200 000 bacterial and archaeal genomes and 150 million proteins with up-to-date annotation. Changes in the Prokaryotic Genome Annotation Pipeline (PGAP) since 2018 have resulted in a substantial reduction in spurious annotation. The hierarchical collection of protein family models (PFMs) used by PGAP as evidence for structural and functional annotation was expanded to over 35 000 protein profile hidden Markov models (HMMs), 12 300 BlastRules and 36 000 curated CDD architectures. As a result, >122 million or 79% of RefSeq proteins are now named based on a match to a curated PFM. Gene symbols, Enzyme Commission numbers or supporting publication attributes are available on over 40% of the PFMs and are inherited by the proteins and features they name, facilitating multi-genome analyses and connections to the literature. In adherence with the principles of FAIR (findable, accessible, interoperable, reusable), the PFMs are available in the Protein Family Models Entrez database to any user. Finally, the reference and representative genome set, a taxonomically diverse subset of RefSeq prokaryotic genomes, is now recalculated regularly and available for download and homology searches with BLAST. RefSeq is found at https://www.ncbi.nlm.nih.gov/refseq/.

INTRODUCTION

The RefSeq collection for prokaryotes has grown to nearly 200 000 genomes and 150 million non-redundant proteins and, after over a decade, remains a trusted source for microbial genomics. The foundation of RefSeq is the continued effort by researchers around the world to sequence the genomes they collect and to publish them in INSDC databases (GenBank, the European Nucleotide Archive and the DNA Database of Japan). The added value of RefSeq over the archival records originally submitted to INSDC resides in the consistency of structural and functional annotation methods across all genomes, the quality control governing acceptance into the collection, and the continuous modernization of content as the collected knowledge about microbial genomes continues to grow. While the sequence records deposited in GenBank are updated only rarely, RefSeq regularly reannotates genomes with PGAP, the Prokaryotic Genome Annotation Pipeline (1,2), to reflect newly characterized prokaryotic metabolic and regulatory systems published in the literature and in specialized resources (3,4) and taxonomic re-assignment of genomes (5).

In this manuscript, we provide an update on the two components sustaining the RefSeq prokaryotic collection: the annotation pipeline itself and the protein family model (PFM) collection on which PGAP relies. We describe adjustments made in PGAP to improve the quality of gene annotation and to refresh annotation in a timely fashion, as well as a substantial expansion over the past few years of the PFMs applied by PGAP as evidence for structural and functional definition of gene features. Areas of focus have

*To whom correspondence should be addressed. Tel: +1 301 402 5721; Fax: +1 301 402 5721; Email: thibaudf@ncbi.nlm.nih.gov
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Table 1. Rate of growth in prokaryotic assemblies submitted to INSDC and accepted into RefSeq. The total INSDC assemblies include high-volume surveillance projects and metagenome assemblies (MAG) which are not included in RefSeq

Year	INSDC assemblies added	Total INSDC assemblies	RefSeq assemblies added	Total RefSeq assemblies
2009			1	1
2010	2433	2433	2021	2022
2011	1004	3437	720	2742
2012	4176	7613	2824	5566
2013	7339	14 952	6094	11 660
2014	13 229	28 181	12 057	23 717
2015	27 472	55 653	26 515	50 232
2016	29 265	84 918	23 678	73 910
2017	41 727	126 645	29 073	102 983
2018	55 104	181 749	36 207	139 190
2019	299 569	481 318	35 080	174 270
2020 (through August)	233 517	714 835	24 370	198 640

Table 2. Rate of growth of species and genera covered by INSDC and RefSeq

Year	INSDC Species	RefSeq Species	INSDC Genera	RefSeq Genera
2010	1 053	979	606	562
2011	1475	1402	769	735
2012	2109	1905	986	917
2013	3291	2829	1370	1187
2014	4355	3934	1618	1477
2015	5492	5132	1835	1714
2016	7428	6400	2205	1952
2017	8694	8336	2438	2293
2018	10 430	9994	2820	2658
2019	12 067	11 644	3128	2965
2020	12 966	12 568	3259	3096

been the incorporation of expert-curated information including function, publications, Enzyme Commission (EC) numbers and gene symbols from the literature and specialized databases, such as the virulence factor database VFDB (4) and the transporter classification database TCDB (3) into new and existing PFMs, and the development of web pages for PFMs. PFMs are available in the Entrez Protein Family Models database to encourage their re-use outside of the PGAP context and will provide a bridge to other NCBI resources.

GROWTH AND MAINTENANCE OF THE REFSEQ PROKARYOTIC COLLECTION

Table 1 shows the number of prokaryotic assemblies submitted to INSDC and added to RefSeq every year since 2014. As the table indicates, the number of assemblies submitted to INSDC continues to accelerate, while the acceptance into RefSeq which increased steadily until 2018 has stabilized to about 36 000 per year. The disproportionate growth of INSDC compared to RefSeq in the past three years is largely explained by high-volume surveillance projects that generate thousands of highly redundant genomes for a single species and by the increase in submitted assemblies that are derived from metagenomic samples (MAGs), both of which are currently excluded from RefSeq. Note that while MAGs do not currently become part of RefSeq, their annotation by PGAP may be requested by submitters at submission time and displayed on GenBank records.

Table 2 shows the growth of bacterial species and genera represented in the current INSDC and RefSeq bacterial sequence collections. Of the 20 135 bacterial species and 3980 bacterial genera described in NCBI Taxonomy, 12 568 (62%) and 3096 (77%) respectively have at least one genome assembly in RefSeq. Since 2014, RefSeq has steadily added coverage for ∼1400 new species per year. The growth in coverage over the past decade has been facilitated by the recent emphasis in deposition of assembled genomes for bacterial type specimens: ∼1200 type assemblies for new species have been deposited per year for the last seven years. Type assemblies have allowed the development of the average nucleotide identity process (ANI) for the verification of the organism name assigned by submitters to sequenced assemblies (5) and allows reliable characterization of assemblies that flow into RefSeq.

The scope of RefSeq has been modified to better represent plasmid sequences. A thorough understanding of the biology of plasmids has major implications to public health. As mobile genetic elements carrying genes outside of the typical arsenal of bacterial chromosomes, plasmids provide their hosts the ability to survive in challenging environments and play a key role in the spread of anti-bacterial resistance by disseminating resistance genes among clinical pathogens (6). Prokaryotic plasmids were historically handled inconsistently by PGAP. Those sequences submitted to INSDC as part of a full genome assembly were annotated by PGAP and included in RefSeq, while those submitted independently of an assembly (stand-alone plasmids) had not been annotated since 2017 or earlier. To provide consistent high-quality annotation of plasmids, HMM and BlastRule evidence which hit proteins found disproportionately on plasmids, were reviewed and improved. A new workflow identifies, and schedules for annotation and addition to RefSeq, plasmids submitted to INSDC that are not part of assemblies and that were sequenced from archaeal or bacterial samples. As of 10 August 2020, 1780 stand-alone plasmids that were added to RefSeq more than three years ago and whose annotation had not been kept current were re-annotated by PGAP, while 4295 were annotated for the first time. A total of 26 older plasmids were subsequently suppressed because they contain >90% vector contamination.

The RefSeq annotation of stand-alone plasmids is available at: ftp://ftp.ncbi.nlm.nih.gov/refseq/release/plasmid/.

In the past, all RefSeq genome assemblies were reannotated once every few years to ensure that older genomes benefit from the latest improvements in PGAP. As shown above, the RefSeq prokaryotic genome collection will pass the 200 000-genome-assembly milestone in 2020. This large volume makes re-annotation of the entire set in a short amount of time more difficult as it requires securing at once a significant amount of computing resources. To address this growth, we have adopted a rolling re-annotation model in which every day the 750 oldest live assemblies are re-annotated. As of 16 August 2020, the median annotation age for a RefSeq assembly is 4.5 months and 95% of assemblies had been annotated in the past 12 months.

In addition to this automated rescheduling, we prioritize subsets of the RefSeq corpus for a targeted reannotation if expected to be substantially impacted by changes in evidence and associated data (e.g. creation of novel PFMs for the prediction of proteins in taxonomically restricted biosynthesis pathways). Moreover, it is important to stress that novel assemblies that meet the RefSeq criteria and updated versions of existing assemblies continue to be annotated shortly after submission. The RefSeq prokaryotic genomes are available for download at:

- ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/archaea/
- ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/

IMPROVEMENTS TO THE PGAP ANNOTATION PIPELINE

A series of data and algorithm changes were introduced in PGAP in the past three years that results in the production of higher confidence annotation products and reduces the annotation of poorly-supported /pseudo features: predicted protein-coding region gene that have been disrupted either biologically (pseudogenes), as by truncation or frameshift, or by artifacts caused by sequencing and/or assembly errors.

GeneMarkS-2+ (7) replaced GeneMarkS+ in PGAP in November 2018 for more accurate prediction, including better start site predictions in leaderless protein-coding genes, and more robust prediction of horizontally transferred and plasmid genes. The content of the two protein databases against which each genome is aligned during the PGAP process were reviewed. These two databases are: (i) a small set of reference proteins (proteins annotated by community-recognized experts on 15 reference genomes selected by RefSeq curators), for which alignments are given a high weight in the annotation of the genomes in the same genus and (ii) a set of 10 million proteins representing clusters of evidence proteins (1,2). Spurious or faulty protein sequences in these databases were identified by examination of multiple sequence alignments containing these proteins, similar RefSeq proteins, and GeneMarkS-2 *ab initio* predictions on other genomes in the same genus or order. A total of 3437 low-quality proteins from otherwise highly trusted reference genomes, that are frameshifted, have start sites inconsistent with most related proteins, or are suspect in other ways (7.4% of all reference proteins) were removed from the reference protein database. Similarly, a total of 195 964 protein fragments, frameshifted proteins or proteins in incorrect frames, were removed from the database of cluster representatives based on manual curation, detection of transposase fragments according to transposase reference sequences, multiple sequence alignments of protein homologs, and analyses of proteins with partial HMM hits. Furthermore, the weight of 2.5 million proteins representative of clusters from rare species was adjusted down to minimize their impact on annotation, and PGAP was modified to trust GeneMarkS-2+ more than homology evidence from rare protein clusters on the presence or absence of a protein-coding feature. Since October 2019, candidate models supported by the alignments of proteins with low combined weight are rejected unless they are also supported by GeneMarkS-2+ *ab initio* model or an HMM hit. The acceptance threshold for short GeneMarkS-2+ predictions lacking support from evidence by homology (HMM hits or protein alignments) has been lowered from 60 to 45 amino acids.

The overall impact from this culling of bad proteins, the downgrading of low-confidence proteins from the pool of PGAP evidence, and the higher weight given to GeneMarkS-2+ predictions has been a drop since January 2018 in the number features PGAP creates with the/pseudo designation. The comparison of the annotation of 3262 genomes from January 2018 versus 2020 shows a 34% drop in the number of /pseudo features, from an average of 218 to 144 per genome, 88% of which is directly attributable to the removal of proteins from the reference or the cluster representative sets. By contrast, we observed only a 0.26% decrease in the number of non-pseudo CDSs in these genomes, suggesting that the changes were effective at increasing the signal to noise ratio in PGAP annotations.

Changes were also made in the prediction of non-coding models. Since July 2019 (PGAP release 4.9), PGAP has been using tRNAscan 2.0.4 (8) for the prediction on tRNAs. In addition, 49 Rfam models for small, non-coding RNA features that are known to be found in prokaryotes, such as riboswitches and attenuators (see list in Supplementary Table S1) are searched in all genomes using Infernal cmsearch (9). Starting in September 2020 (PGAP release 4.13), 23S and 16S rRNAs are detected using Rfam SSU and LSU models for bacteria and archaea (RF00177, RF01959, RF02540 and RF02541) rather than by BLAST against in-house databases of ribosomal RNAs. Note that 5S rRNAs have been detected using Rfam RF00001 since 2016.

The assignment of function to proteins was improved by the addition of the Subfamily Protein Architecture Labeling Engine (SPARCLE(10)) to PGAP for associating curated CDD architectures to proteins. PGAP uses multiple evidence sources organized hierarchically for assigning names to proteins, with more specific rules awarded higher precedence (2). CDD architectures were previously already used for the re-naming of RefSeq proteins matching lower-precedence evidence, such as a BLAST hits, post annotation. In addition to allowing RefSeq proteins to receive higher precedence names earlier, the addition of SPARCLE to PGAP uniformizes the functional assignment across

Protein accession	WP_003401737.1	WP_042507723.1	WP_003900124.1	WP_003401747.1
Former name	hypothetical protein	hypothetical protein	dynamin-like GTPase family protein	dynamin-like GTPase family protein
Former evidence	No evidence	No evidence	CDD architecture 11431116	CDD architecture 11431116
New name	IniB N-terminal domain-containing protein	isoniazid-induced protein IniB	isoniazid-induced protein IniA	isoniazid-induced protein IniC
New evidence	NF038175.1	NBR011041	NBR011040	NBR011042

Figure 1. TCDB-derived evidence (NBR011040, NBR011041, NBR011042 and NF038175) improves the annotation of three proteins (WP_003900124.1, WP_042507723.1, WP_003401747.1 and WP_003401737.1, respectively) from the *iniBAC* operon in *Mycobacterium tuberculosis* CDC1551 genome (NC_002755.2: 408694–414298). The new evidence was built from TCDB and used to name RefSeq proteins.

RefSeq genomes and GenBank genomes for which submitters have requested PGAP annotation.

BUILDING UP THE PROTEIN FAMILY MODEL COLLECTION FOR ANNOTATION

An important focus of the past couple of years has been the growth and expert curation of the hierarchy of PFMs used by PGAP as evidence for both structural and functional annotation (see (2) for more details). The three major types of PFMs used by PGAP are Hidden Markov Models (HMMs), BlastRules and curated CDD architectures (for functional annotation only). The HMMs used by PGAP come from a variety of sources. NCBIFAMs were built at NCBI, and include models for acquired or innate anti-microbial resistance proteins, and models derived from NCBI protein clusters (PRKs (11)). PGAP also uses TIGRFAMs originally developed at the J. Craig Venter Institute (JCVI, previously known as The Institute for Genomic Research, or TIGR) (12), and now owned by NCBI, and Pfams in Release 32.0 (13).

In the past two years, we increased the number of HMMs used by PGAP in structural annotation, functional annotation, or both, from 33 387 to 35 539 by incorporating 1641 HMMs from newer releases of Pfam and by building 510 new NCBIFAMs. We also nearly doubled the number of BlastRules from 6208 to 12 392 and increased curated CDD architectures to ∼36 000. In addition, we have improved existing PFMs based on the current literature by assigning better protein names, gene symbols and other attributes that are transferred to the PGAP-annotated proteins they hit. This curation, described in more details below, was driven by the following objectives: increase the value of Pfams, and improve the representation and specificity of transporter proteins, virulence factors and ribosomally synthesized and post-translationally modified peptide natural products.

A large fraction of the Pfam collection is built to identify and characterize individual domains found within proteins rather than full-length proteins. To make these useful for functional annotation, NCBI curators have assigned product names to 5733 of the 10 675 Pfam models that hit prokaryotic proteins based on (i) names assigned to curated CDD domain architectures that contain the Pfam accession under review, and (ii) publications describing the proteins they hit, thereby making them available for functional annotation. Finding publications based on their mention of proteins belonging Pfam or other PFMs was aided significantly by the use of PaperBLAST (14).

In 2019 and 2020, PFMs were built or revised based on specialized community resources so they could be leveraged by PGAP. The transporter classification database (TCDB (3)) is a comprehensive transporter annotation system, which classifies membrane transporters and some other types of membrane proteins into classes, subclasses, families, and subfamilies based on functional and phylogenetic analyses. Prokaryotic transporters were retrieved from TCDB and reviewed by RefSeq curators to establish 163 BlastRules and 77 NCBIFAM HMMs. In addition, the transporter classification of TCDB was used to improve the protein product name on 67 Pfam HMMs. As of August 2020, TCDB-derived PFMs hit 3 483 172 RefSeq proteins and were used to assign names to 1 839 864 RefSeq proteins.

Figure 1 is an example showing how TCDB-derived evidence improves the annotation of RefSeq proteins. Mutations in the isoniazid-induced proteins IniA, IniB, IniC can lead to adaptive resistance to either isoniazid or ethambutol, two front-line drugs for the treatment of tuberculosis, therefore the correct annotation of these proteins is highly important. Prior to our recent curation, IniA and IniC were annotated correctly, but identically and somewhat generically, as dynamin-like GTPase family proteins, based on the CDD architecture 11431116, while the annotation of isoniazid-induced protein IniB was not yet covered for annotation by any PFM in PGAP. The three TCDB-derived BlastRules (NBR011040, NBR011041, and NBR011042) improved the annotation of 134, 213 and 137 RefSeq proteins, respectively, which were annotated on 6636, 6208 and 6739 RefSeq genomes, respectively. Searches for homologs of IniB, much of which is repetitive and extremely glycine-rich, revealed that the N-terminal 50 amino acids represent a novel homology domain. A new HMM for this domain, NF038175.1, also hits one ad-

ditional protein in *Mycobacterium tuberculosis*, Rv0340, a small protein that is found immediately upstream of IniB (Rv0341) and that is likewise known for adaptive mutations conferring resistance to inhibitors of cell envelope biosynthesis.

Bacterial virulence factors (VF) contributing to disease in humans have been a long-standing annotation priority for many. To improve the annotation of VFs by PGAP, we augmented the coverage of the PFMs used by PGAP by mining the VFDB (Virulence Factor DataBase) set A collection of proteins downloaded June 17, 2019 (4). We improved equivalog-level HMMs from TIGRFAMs for conserved proteins of secretion systems, such as TIGR02499.1 (type III secretion system stator protein SctL), to apply modern nomenclature as reviewed by Portaliou *et al.* (15), and then developed more specific BlastRules for lineage-specific variants catalogued in the literature and in VFDB, e.g. BscL from Bordetella pertussis (NBR011452) and YscL from Yersinia (NBR011453). Currently, 3014 of the 3180 proteins in set A (94.8 %) are hit by at least one PFM. Altogether, 1739 exception-level BlastRules and 575 equivalog-level HMMs covering set A proteins provide highly specific names to 3.5 million RefSeq proteins and 390 CDD architectures with hit to set A proteins name over 5.6 million RefSeq proteins.

The short lengths and extreme diversity of ribosomally synthesized and post-translationally modified peptide natural products (RiPPs) precursor genes can confound *ab initio* and homology-based detection methods, unless a sufficiently sensitive protein family HMM can be constructed. To improve the detection by PGAP of RiPP precursors and of similarly sized but only minimally modified type II bacteriocins, we surveyed the literature for references to RiPP and bacteriocin precursors not yet detectable by any existing HMM, and examined genomic sequences in the immediate vicinity of known classes of RiPP maturation proteins. For each candidate founding member of a new family of RiPP or type II bacteriocin, iterative searches, by PSI-BLAST (16) or by HMMER3 (17), local genomic context, and inspection of multiple sequence alignments were used to add new proteins to each family. A total of 50 NCBI HMMs for RiPP and bacteriocin families or their leader peptide regions (with a median HMM length of 44 aa) complement those already available from Pfam and TIGRFAMs. The families modeled include various lasso peptides, thiazole/ozazole-modified microcins (TOMMs) such as listeriolysin S, radical SAM/SPASM-modified RiPPs such as the darobactins, and leaderless bacteriocins. The full set of HMMs and BlastRules used by PGAP, including TIGRFAMs and Pfam models, is provided in Supplementary Table S2.

Finally, in preparation for the expansion of the scope of RefSeq to stand-alone plasmids described above, 144 HMMs that hit proteins found disproportionately on plasmids were reviewed, and addition or improvement to their product names were made when possible. Most of these are low-precedence HMMs that identify protein functional domains, and have low-information names, so, in addition, 229 more specific BlastRules were created to inform the names of 7000 RefSeq high-interest proteins.

ENTREZ DATABASE OF PROTEIN FAMILY MODELS

As demonstrated above, PFMs are one of the critical components in the PGAP process and an important factor in the quality of the annotation it produces. However, these models can be applied to any protein and nucleotide sequences by commonly used tools (BLAST for BlastRules, RPS-BLAST for CDD architectures (18) and hmmsearch for HMMs (17)). In order to promote their use in contexts beyond PGAP, we have built the Protein Family Models Entrez database, a database that contains HMMs, BlastRules and CDD architectures, available at https://www.ncbi.nlm.nih.gov/protfam. As of 10 September 2020, this database contains 35 540 HMMs and 12 634 BlastRules, 32 669 reviewed and 116 793 provisional CDD architectures (including many that only apply to eukaryotes or viruses).

PFMs can be queried using a variety of terms, including any name known for the model, the full name or a substring of the name of a protein, a gene symbol, a publication, or the type of model (Conserved Domain Architectures, BlastRule or HMM). For example, searching with the term 'IniC' returns PFMs where the term IniC appears in the model name, its description, gene symbol, or a publication title. 'iniC[GeneSymbol]' returns any PFM with the gene symbol iniC, and the query 'isoniazid[Description] AND blastrule[Method]' returns any BlastRule where the text describing the BlastRule contains the term 'isoniazid'. An example database record, NF033727.1 (https://www.ncbi.nlm.nih.gov/genome/annotation_prok/evidence/NF033727/) is shown in Figure 2. Attributes of the model such as product name, family type, gene symbol and publications are displayed on the page, and there is a button to download the statistical model. The list of RefSeq proteins that are hit by the models is presented in the lower part of the page, along with the most basal taxonomic node in which they are found, the number of genome assemblies on which they are annotated, and the strength of the hit of the model to the protein, as represented by the hmmsearch score. The hits are divided into two tabs: RefSeq proteins that inherit the product name and attributes from the model represented, and proteins that are hit by the model, but named by a model of higher precedence than the one described in the page. The table of proteins can be downloaded, proteins can be selected for FASTA or document summary download, or subjected to multiple sequence alignment with COBALT (19). Similarly, BlastRule records contain the general information about the model, and the list of RefSeq proteins that it hits (not shown). In order to promote their reuse, PFMs are also available at:

- HMM models: https://ftp.ncbi.nih.gov/hmm/current
- BlastRules: https://ftp.ncbi.nih.gov/pub/blastrules/
- CDD architectures: https://ftp.ncbi.nlm.nih.gov/pub/mmdb/cdd/SparcleLabel

IMPACT OF THE PROTEIN FAMILY MODEL EXPANSION ON PROKARYOTIC REFSEQ PROTEINS

The impact of the continued curation of the PFMs can be measured by the increase in the percentage of RefSeq prokaryotic proteins with product names derived from cu-

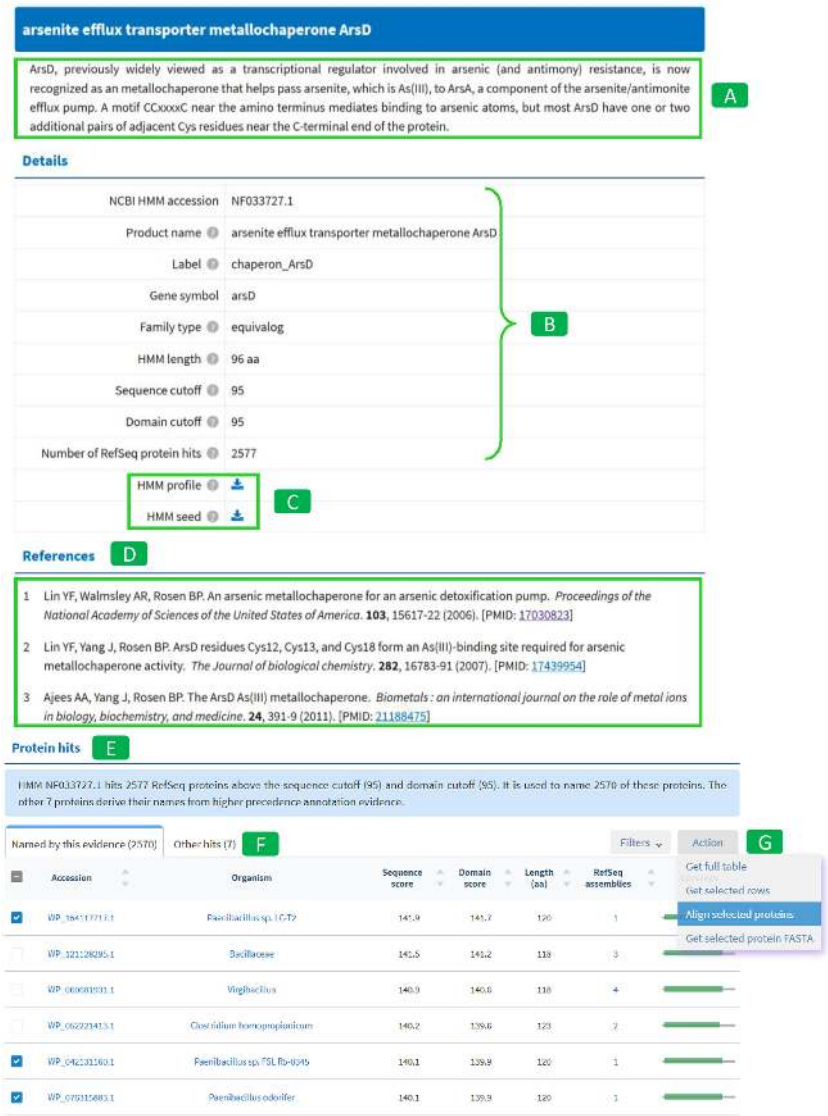

Figure 2. Example record for HMM NF033727.1. (**A**) Description of the function of proteins included in the family defined by the model. (**B**) Attributes and characteristics of the model, including name that is propagated to protein products named by PGAP based on the model. (**C**) Download options for the profile and the seed proteins (**D**) Publications supporting the definition of the model and its functional assignment. (**E**) RefSeq proteins hit by the model. (**F**) One tab lists the hits for which NF033727.1 is the highest precedence evidence (shown), and another tab lists the proteins that are hit by the model but named after a higher-precedence evidence (not shown). (**G**) Menu of possible actions for proteins selected on the left.

rated HMMs, BlastRules or CDD architectures over time. As shown in Figure 3, nearly 80% of prokaryotic RefSeq proteins (121 million proteins, as of August 2020) are named by curated evidence, up from 68% (81 million) in July 2019. The particularly high increase in the number of proteins named after an HMM observed in May 2020 is due to the addition described above of product names to wide-reaching Pfam HMMs. This continued growth in coverage of the RefSeq space by PFMs is particularly noteworthy when considering that new proteins are added to RefSeq at a steady rate of 3.2 million per month due to the taxonomic expansion and growing sequence diversity in prokaryotic assemblies submitted to GenBank and added to RefSeq.

To better expose the connection between the names assigned to RefSeq proteins and PFMs, we added a comment block to the RefSeq protein records in Entrez Protein (Figure 4) that are named by a PFM (records for proteins named based on a BLAST hit and for hypothetical proteins do not contain the new comment). The comment includes the PFM category (HMM, BlastRule or CDD architecture), accession, and source. The content of the comment is indexed, so it can be searched. For example, the query 'Evidence-For-Name-Assignment[Properties]' in Entrez Protein returns all proteins with names based on a PFM (as opposed to a BLAST hit, or no evidence at all). The query 'Evidence Accession=NF033727.1'[Text Word] returns all proteins named by NF033727.1. 'Evidence Source=NCBIFAM'[Text Word] returns all proteins named by an HMM in the NCBIFAM collection.

We recognize that protein attributes beyond product names are valued by RefSeq users. Publications provide the context in which the function of gene was established, and

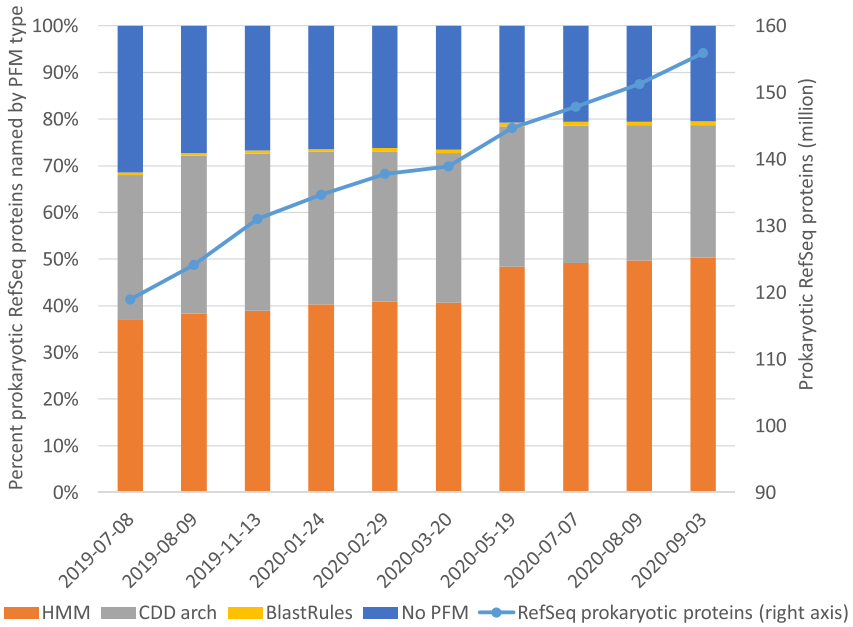

Figure 3. Increase over time of RefSeq non-redundant proteins named after a protein family model: the stacked bars (values on the left axis) indicate the proportion of proteins named by HMMs (orange), CDD architecture (gray), BlastRules (yellow), or Blast hits to cluster representative protein (blue). The blue line (values on the right axis) represents the growth in the total number of prokaryotic RefSeq proteins

```
##Evidence-For-Name-Assignment-START##
Evidence Category   :: HMM
Evidence Accession  :: NF033727.1
Evidence Source     :: NCBIFAM
##Evidence-For-Name-Assignment-END##
```

Figure 4. Example comment block on a non-redundant RefSeq protein named after HMM NF033727.1.

gene symbols and EC number are essential for comparative genomics. In 2019, we enabled the addition of EC numbers, gene symbols, and publication attributes to newly created and older PFMs. As of August 2020, 8021 HMMs and 8198 BlastRules with product names also have a gene symbol, and 6385 HMMs and 1211 BlastRules have an EC number. We also modified the PGAP naming process so that these attributes are inherited by gene and CDS features on annotated genomes (gene symbols and EC numbers only) and by RefSeq proteins named by these models. For some biologically important organisms the number of genes with gene symbols annotated by PGAP has increased 5-fold since June 2018 (Supplemental Figure S1).

UPDATES IN THE REPRESENTATIVE AND REFERENCE GENOME COLLECTION

While the RefSeq collection of prokaryotic genomes provides an expansive view of the variation in the landscape of sequenced microbes, it is large and consists of an uneven representation of species, with genome assemblies for human pathogens being particularly abundant. This poses a challenge for some uses, such as sequence homology searching, or *k*-mer indexing (20,21). As a possible solution, NCBI has defined a set of 'reference and representative' genomes. This set is a compact, normalized, and taxonomically diverse view of the RefSeq collection that can be used for the taxonomic identification and characterization of novel sequences. It includes 15 reference genome assemblies (down from 120 prior to April 2020) that were annotated and are updated by the assembly submitters and chosen by the RefSeq curatorial staff based on their quality and importance to the community as anchors for the analysis of other genomes in their taxonomic group. Some reference genomes are selected based on a long history of collaboration and wide recognition as a community standard, such as the reference genome of *Escherichia coli* str. K-12 substr. MG1655 (22). Other reference genomes are selected based on medical importance, sequence and annotation quality, and the availability of experimental support. Gene annotation on these genomes is reviewed and may be modified by RefSeq, but largely reflects the work of the submitters.

For species that do not have a reference genome, a representative genome is selected among PGAP-annotated genomes. The criteria for choosing representatives (listed in https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/#representative_genomes) aim at selecting assemblies that are of the best quality and that are not outliers for their species.

Representative assemblies are updated three times a year to take into account newly added assemblies to RefSeq, changes in the NCBI Taxonomy, modified taxonomic assignments, and recently discovered contamination in existing assemblies. Note that no representative is selected for undefined species (such as '*Vibrio sp.*'). The collection of reference and representative genome assemblies for Bacteria and Archaea released in August 2020 contains 11 735

selected assemblies among the 192 000 assemblies in RefSeq.

The reference and representative set can be searched in Entrez Assembly (23) (https://www.ncbi.nlm.nih.gov/assembly), using the filter (representative_genome[filter] OR reference_genome[filter]) and downloaded.

BLAST databases for reference and representative genomic sequences and proteins are offered on the NCBI website:

- Microbial Nucleotide BLAST database at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&BLAST_SPEC=MicrobialGenomes)
- And the RefSeq Representative Genome Database, in the Database menu at: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastn&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)
- Proteins annotated on representative genomes are in the RefSeq Select proteins databases (refseq_select): https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome

CONCLUSION AND FUTURE WORK

The RefSeq prokaryotic collection continues to grow and to incorporate more species and genera every year. Despite the challenge this increase in scope represents, the quality of the collection has increased as well. Changes in the PGAP process and in the BLAST databases and PFMs on which it relies have resulted in higher confidence and better-named proteins annotated on RefSeq genomes. The number of /pseudo features per annotated genome has decreased on average by 34% in the past 3 years and the percentage of proteins named after curated PFMs is at an all-time high of 79%. Improvements were driven in part by considerable efforts in the expert curation of the PFMs used as evidence for structural and functional annotation. All PFMs used by PGAP are now searchable in Entrez, available for download, and linked to protein records. We hope that this greater transparency will facilitate the reuse of the PFMs, decrease the redundancy across existing PFM collections and focus the curation work, at NCBI and elsewhere, on proteins not covered by any of the current PFM collections. We are exploring using PFMs from more sources, including UniProt's UniRules (24), and fine-tuning together the coverage of PFMs that are linked to a single biological process, in the manner of Genome Properties (12,25) and RAST (26). Our plans also include exploring the addition of GO terms to PFMs and expanding the assignment of gene symbols to more annotated genes to further the potential of the RefSeq prokaryotic collection for comparative genomics.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Intramural Research Program of the National Library of Medicine at National Institutes of Health/DHHS. Funding for open access charge: Intramural Research Program of the National Library of Medicine at the National Institutes of Health/DHHS.

Conflict of Interest Statement. None declared.

REFERENCES

1. 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.
2. Haft,D.H., DiCuccio,M., Badretdin,A., Brover,V., Chetvernin,V., O'Neill,K., Li,W., Chitsaz,F., Derbyshire,M.K., Gonzales,N.R. *et al.* (2018) RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res.*, **46**, D851–D860.
3. Saier,M.H. Jr., Reddy,V.S., Tsu,B.V., Ahmed,M.S., Li,C. and Moreno-Hagelsieb,G. (2016) The Transporter Classification Database (TCDB): recent advances. *Nucleic Acids Res.*, **44**, D372–D379.
4. Liu,B., Zheng,D., Jin,Q., Chen,L. and Yang,J. (2019) VFDB 2019: a comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.*, **47**, D687–D692.
5. Ciufo,S., Kannan,S., Sharma,S., Badretdin,A., Clark,K., Turner,S., Brover,S., Schoch,C.L., Kimchi,A. and DiCuccio,M. (2018) Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int. J. Syst. Evol. Microbiol.*, **68**, 2386–2392.
6. San Millan,A. (2018) Evolution of plasmid-mediated antibiotic resistance in the clinical context. *Trends Microbiol.*, **26**, 978–985.
7. Lomsadze,A., Gemayel,K., Tang,S. and Borodovsky,M. (2018) Modeling leaderless transcription and atypical genes results in more accurate gene prediction in prokaryotes. *Genome Res.*, **28**, 1079–1089.
8. Chan,P.P. and Lowe,T.M. (2019) tRNAscan-SE: searching for tRNA genes in genomic sequences. *Methods Mol. Biol.*, **1962**, 1–14.
9. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.
10. Lu,S., Wang,J., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Marchler,G.H., Song,J.S. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265–D268.
11. 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.
12. 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.
13. El-Gebali,S., Mistry,J., Bateman,A., Eddy,S.R., Luciani,A., Potter,S.C., Qureshi,M., Richardson,L.J., Salazar,G.A., Smart,A. *et al.* (2019) The Pfam protein families database in 2019. *Nucleic Acids Res.*, **47**, D427–D432.
14. Price,M.N. and Arkin,A.P. (2017) PaperBLAST: text mining papers for information about homologs. *mSystems*, **2**, e00039-17.
15. Portaliou,A.G., Tsolis,K.C., Loos,M.S., Zorzini,V. and Economou,A. (2016) Type III secretion: building and operating a remarkable nanomachine. *Trends Biochem. Sci.*, **41**, 175–189.
16. Schäffer,A.A., Aravind,L., Madden,T.L., Shavirin,S., Spouge,J.L., Wolf,Y.I., Koonin,E.V. and Altschul,S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.
17. Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.
18. Yang,M., Derbyshire,M.K., Yamashita,R.A. and Marchler-Bauer,A. (2020) NCBI's conserved domain database and tools for protein domain analysis. *Curr. Protoc. Bioinformatics*, **69**, e90.
19. Papadopoulos,J.S. and Agarwala,R. (2007) COBALT: constraint-based alignment tool for multiple protein sequences. *Bioinformatics*, **23**, 1073–1079.
20. Nasko,D.J., Koren,S., Phillippy,A.M. and Treangen,T.J. (2018) RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome Biol.*, **19**, 165.

21. Piro,V.C., Dadi,T.H., Seiler,E., Reinert,K. and Renard,B.Y. (2020) ganon: precise metagenomics classification against large and up-to-date sets of reference sequences. *Bioinformatics*, **36**, i12–i20.
22. Karp,P.D., Ong,W.K., Paley,S., Billington,R., Caspi,R., Fulcher,C., Kothari,A., Krummenacker,M., Latendresse,M., Midford,P.E. *et al.* (2018) The EcoCyc database. *EcoSal Plus*, **8**, ESP-0009-2013.
23. Kitts,P.A., Church,D.M., Thibaud-Nissen,F., Choi,J., Hem,V., Sapojnikov,V., Smith,R.G., Tatusova,T., Xiang,C., Zherikov,A. *et al.* (2016) Assembly: a resource for assembled genomes at NCBI. *Nucleic Acids Res.*, **44**, D73–D80.
24. MacDougall,A., Volynkin,V., Saidi,R., Poggioli,D., Zellner,H., Hatton-Ellis,E., Joshi,V., O'Donovan,C., Orchard,S., Auchincloss,A.H. *et al.* (2020) UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics*. **36**, 4643–4648,
25. Richardson,L.J., Rawlings,N.D., Salazar,G.A., Almeida,A., Haft,D.R., Ducq,G., Sutton,G.G. and Finn,R.D. (2019) Genome properties in 2019: a new companion database to InterPro for the inference of complete functional attributes. *Nucleic Acids Res.*, **47**, D564–D572.
26. Overbeek,R., Olson,R., Pusch,G.D., Olsen,G.J., Davis,J.J., Disz,T., Edwards,R.A., Gerdes,S., Parrello,B., Shukla,M. *et al.* (2014) The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.*, **42**, D206–D214.