Published online 8 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D733–D745***
doi: 10.1093/nar/gkv1189

Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation

Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy and Kim D. Pruitt*

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Building 38A, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received September 01, 2015; Revised October 15, 2015; Accepted October 24, 2015

ABSTRACT

The RefSeq project at the National Center for Biotechnology Information (NCBI) maintains and curates a publicly available database of annotated genomic, transcript, and protein sequence records (http://www.ncbi.nlm.nih.gov/refseq/). The RefSeq project leverages the data submitted to the International Nucleotide Sequence Database Collaboration (INSDC) against a combination of computation, manual curation, and collaboration to produce a standard set of stable, non-redundant reference sequences. The RefSeq project augments these reference sequences with current knowledge including publications, functional features and informative nomenclature. The database currently represents sequences from more than 55 000 organisms (>4800 viruses, >40 000 prokaryotes and >10 000 eukaryotes; RefSeq release 71), ranging from a single record to complete genomes. This paper summarizes the current status of the viral, prokaryotic, and eukaryotic branches of the RefSeq project, reports on improvements to data access and details efforts to further expand the taxonomic representation of the collection. We also highlight diverse functional curation initiatives that support multiple uses of RefSeq data including taxonomic validation, genome annotation, comparative genomics, and clinical testing. We summarize our approach to utilizing available RNA-Seq and other data types in our manual curation process for vertebrate, plant, and other species, and describe a new direction for prokaryotic genomes and protein name management.

INTRODUCTION

For the past 15 years the National Center for Biotechnology Information (NCBI) RefSeq database has served as an essential resource for genomic, genetic and proteomic research. The RefSeq project's provision of curated and stable annotated reference genomes, transcripts, and proteins for selected viruses, microbes, organelles, and eukaryotic organisms, has allowed researchers to focus on the best representative sequence data in contrast to the redundant data in GenBank, and to unambiguously reference specific genetic sequences. The RefSeq collection provides explicitly linked genome, transcript, and protein sequence records that incorporate publications, informative nomenclature, and standardized and expanded feature annotations. Ref-

*To whom correspondence should be addressed. Tel: +1 301 435 5898; Fax: +1 301 435 5898; Email: pruitt@ncbi.nlm.nih.gov

Seq records are integrated into NCBI's resources including the Nucleotide, Protein, and BLAST databases and can be easily identified by the keyword 'RefSeq' and by their distinct accession prefixes that define their type (Table 1). All RefSeq data are subject to quality assurance (QA) checks with some specialized QA tests developed for different taxa or data types. For example, all viral RefSeqs undergo taxonomic review by NCBI staff before public release. RefSeq accessions are widely cited in scientific publications and genetic databases because they provide a stable and consistent coordinate system that can be used as a baseline for reporting gene specific data, clinical variation, and cross-species comparisons. These reference sequence standards are increasingly important because accurate reporting and reproducibility are vital components for best practices in biomedical research (1).

In recent years advanced sequencing techniques have facilitated a substantial increase in whole genome assembly submissions to the public databases. As a result, the RefSeq project has concordantly expanded the depth and breadth of taxa included in the dataset primarily through improvements to several in-house annotation pipelines. All taxa are in scope for RefSeq inclusion; however, annotation is often limited to those organisms for which a high quality primary genome assembly is available with uncontested organism information. Thus, we may exclude some categories of data that don't meet our quality standards. Excluded datasets include: metagenomes, assemblies with low contig N50 values or especially high number of unplaced scaffolds/contigs (i.e. high fragmentation), or genomes that have significant mismatch or indel variation compared to other closely related genomes for the species (e.g. some prokaryotes).

A unique aspect of the RefSeq dataset is the combined approach of leveraging computation, collaboration and curation by NCBI scientific staff. As a large bioinformatics facility, NCBI has invested in developing robust process flows to generate annotation and perform quality assurance tests for eukaryotic and prokaryotic genomes, transcripts, and proteins. Improvements to the viral genomes process flow are in progress. The RefSeq group collaborates with numerous expert groups including official nomenclature authorities (e.g. HUGO Gene Nomenclature Committee (HGNC) and Zebrafish Information Network (ZFIN) for human and zebrafish gene names respectively), UniProtKB (protein names) and miRBase (microRNAs) (2–5). These, and other, collaborations help maintain and improve on the quality of the RefSeq data set through QA reports, exchanges of gene and sequence information, and exchanges of functional information. NCBI staff also provide curation support for viruses, prokaryotes, eukaryotes, organelles, plasmids, and targeted projects including curating genes and sequences for *Homo sapiens*, *Mus musculus* and other organisms. RefSeq curators improve the quality of the database through review of QA test results, involvement in the selection of certain inputs for genome annotation processing, sequence analysis, taxonomic analysis, and functional review. Curation also supports improvements to genome annotation pipelines as content experts help define programmatic approaches to model both typical and atypical biology. For eukaryotes, particularly mammals, transcript-based curation defines 'best' sequence representatives (as 'known' RefSeqs; Table 1 footnote) which are used as a primary input reagent to the eukaryotic genome annotation pipeline (http://www.ncbi.nlm.nih.gov/books/NBK169439/). Improvements in input reagent quality in turn add significant quality and reproducibility to the resultant genome annotation. This type of manual curation has historically been focused on human and mouse because of their unique biomedical importance (6). More recently these curation efforts have given greater attention to *Rattus norvegicus*, *Danio rerio*, *Bos taurus*, and *Gallus gallus*. These species are relevant to human health as well as agricultural sustainability.

In this paper, we report on our progress in expanding the RefSeq dataset to include more diverse organisms, describe improvements in data access, and provide examples illustrating an increased focus on providing phylogenetically useful datasets as well as functional feature annotation on RefSeq transcript and protein records. We anticipate these efforts and improvements in the RefSeq dataset will continue to contribute to the advancement of medical translational research, agricultural improvements, phylogenetic identification, and evolutionary studies.

GENERATING THE REFSEQ DATASET

RefSeq sequence records are generated by different methods depending on the sequence class and organism. Archaeal and bacterial genomes (see Prokaryotes section) are annotated using NCBI's prokaryotic genome annotation pipeline (http://www.ncbi.nlm.nih.gov/books/NBK174280/), while a small number of reference bacterial genomes are supported by collaboration and manual curation. RefSeq eukaryotic genomes are provided using two process flows. The majority of plant, animal, insect and arthropod genomes are annotated by the eukaryotic genome annotation pipeline. This pipeline generates annotation results based on available transcript data (including RNA-Seq and transcriptome shotgun assembly (TSA) data), as well as protein homology, *ab initio* prediction (largely when transcriptome data are unavailable), and available known (curated) RefSeq transcripts and proteins (see Table 1). Pipeline-generated annotation (model RefSeqs) may or may not have support for the complete exon combination from a single evidence alignment but may have RNA-Seq support for exon pairs. The eukaryotic genomes which have been annotated by this pipeline are reported publicly with links to download the data by FTP, to view or perform a BLAST query against the annotated genome, or to access a detailed annotation report summary (http://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/). The pipeline for a subset of eukaryotes including fungi, protozoa, and nematodes involves propagating annotation that has been submitted to the International Nucleotide Sequence Database Collaboration (INSDC), with format standardization, to a RefSeq copy of the submitted genome assembly (see Algae, Fungi, Nematodes and Protozoa).

NCBI staff provide the bulk of RefSeq organelle genome annotation through propagation from the INSDC submission. Mammalian mitochondria annotation is often supplemented with manual curation. The RefSeq project also maintains reference sequences for targeted loci projects

Table 1. RefSeq accession prefixes

Prefix	Molecule type	Use context
NC_[1]	DNA	Chromosomes Linkage Groups
AC_[1]	DNA	Chromosomes Linkage Groups
NZ_[2]	DNA	Chromosomes Scaffolds Used predominantly for prokaryotic genomes.
NT_[3]	DNA	Scaffolds
NW_[3]	DNA	Scaffolds
NG_[1]	DNA	Genomic regions. A genomic region record may represent a single or multiple genetic loci (e.g. rRNA targeted locus, RefSeqGene, non-transcribed pseudogene)
NM_[3,4]	mRNA	protein-coding transcripts
XM_[3,5]	mRNA	protein-coding transcripts
NR_[3,4]	RNA	non-protein-coding transcripts including lncRNAs, structural RNAs, transcribed pseudogenes, and transcripts with unlikely protein-coding potential from protein-coding genes
XR_[3,5]	RNA	non-protein-coding transcripts, as above
NP_[3,4]	protein	Proteins annotated on NM_ transcript accessions or annotated on genomic molecules without an instantiated transcript (e.g. some mitochondrial genomes, viral genomes, and reference bacterial genomes
AP_[3]	protein	Proteins annotated on AC_ genomic accessions or annotated on genomic molecules without an instantiated transcript record
XP_[3,5]	protein	Proteins annotated on XM_ transcript accessions or annotated on genomic molecules without an instantiated transcript record
YP_[3]	protein	Proteins annotated on genomic molecules without an instantiated transcript record
WP_[6]	protein	Proteins that are non-redundant across multiple strains and species. A single protein of this type may be annotated on more than one prokaryotic genome

[1]The complete accession number format consists of the prefix, including the underscore, followed by 6 numbers followed by the sequence version number.
[2]The complete accession format consists of the prefix followed by the INSDC accession number that the RefSeq record is based on followed by the RefSeq sequence version number.
[3]The complete accession number format consists of the prefix, including the underscore, followed by 6 or 9 numbers followed by the sequence version number.
[4]Records with this accession prefix have been curated by NCBI staff or a model organism database, or are in the pool of accessions that curators work with. These records are referred to as the 'known' RefSeq dataset.
[5]Records with this accession prefix are generated through either the eukaryotic genome annotation pipeline, or the small eukaryotic genome annotation pipeline. Records generated via the first method are referred to as the 'model' RefSeq dataset.
[6]The complete accession number format consists of the prefix, including the underscore, followed by 9 numbers followed by the version number. The version number is always '.1' as these records are not subject to update. See online documentation for additional information: www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/.

such as RefSeqGene, which is a member of the Locus Reference Genomic (LRG) collaboration (7), for bacterial and fungal ribosomal rRNA loci, and for fungal internal transcribed spacer sequences (ITS) (8). In addition, a significant number of human, mouse, and other transcripts and proteins are provided through collaboration and manual curation which includes sequence analysis and literature review.

NCBI's prokaryotic (see below) and eukaryotic annotation pipelines have kept pace with the increasing number of genome assemblies submitted to INSDC by providing consistent annotation onto RefSeq copies of selected high quality submitted genome assemblies. To date, 245 eukaryotic genomes, including 170 vertebrate genomes, have been annotated by this pipeline, of which more than 120 species were annotated in the past 20 years. Among this group are 52 bird species that include representative species of most avian orders (9,10). There has also been a significant expansion in the number of RefSeq-annotated assemblies for non-human primates, other mammals, fish, plants, and arthropods.

ACCESSING THE REFSEQ DATASET

The RefSeq homepage http://www.ncbi.nlm.nih.gov/refseq/ is a central hub for all aspects of the RefSeq dataset. This site provides links that guide users through a general description of the project as well as factsheets, growth statistics and information on more focused RefSeq projects such as the Prokaryotic genome re-annotation initiative, the Consensus Coding Sequence (CCDS) project (11) the RefSeqGene project, and Targeted Loci (http://www.ncbi.nlm.nih.gov/refseq/targetedloci/) projects. Links to the most current comprehensive FTP release and detailed documentation on the format and content of the release can be found in the 'Announcements' section of the RefSeq homepage. Previous RefSeq announcements are also available from this page. We strongly encourage downloading RefSeq data directly from NCBI, as downloads from other bioinformatics and genome browser resources may not include all of the available data, or may merely reflect alignments of RefSeq transcripts to a genome rather than the genome annotation results that are generated by NCBI.

RefSeq sequence data can be accessed interactively using NCBIs Nucleotide and Protein databases, in BLAST

databases, through NCBI's programmatic interface (E-utilities), or through file transfer protocol (FTP). E-utilities support scripted access to download RefSeq data in a variety of formats based on either search terms or accession lists; extensive documentation is available in the NCBI Handbook (www.ncbi.nlm.nih.gov/books/NBK25501/) and training videos are available from NCBI's YouTube channel (https://www.youtube.com/user/NCBINLM). Both the Nucleotide and Protein databases allow for query results to be restricted to only RefSeq records by selecting 'RefSeq' under the 'Source database' in the filters sidebar. RefSeq data may also be accessed from other NCBI databases including Assembly, BioProject, Gene, and Genome by following the links provided to Nucleotide, Protein, or FTP resources Information on curation changes within the RefSeq group or NCBI updates that impact the RefSeq database are reported through several sources including RefSeq FTP release notes, periodic published reports, the NCBI Announcements News feed http://www.ncbi.nlm.nih.gov/news/ and through the NCBI Insights Blog http://ncbiinsights.ncbi.nlm.nih.gov/. Users may also subscribe to the refseq-announce mail list to receive periodic updates about the project and a summary of the content of each RefSeq FTP release (http://www.ncbi.nlm.nih.gov/mailman/listinfo/refseq-announce/).

RefSeq data are distributed via FTP through two sites, *refseq* (ftp://ftp.ncbi.nlm.nih.gov/refseq/) and *genomes* (ftp://ftp.ncbi.nlm.nih.gov/genomes/). The *refseq* FTP site provides daily updates of all new and updated RefSeq records, weekly updates of some data types, and a bi-monthly comprehensive RefSeq release (/refseq/release/). In addition, select organism-specific transcript and protein datasets, including human and mouse, are updated weekly. The *RefSeqGene* subdirectory is updated daily, with alignments to the genome released with each annotation run. The comprehensive bi-monthly RefSeq release is organized by taxonomic (e.g. vertebrate mammals) or other groupings (e.g. mitochondria). Data may also be downloaded for the entire RefSeq collection from the /refseq/release/complete/ directory. The RefSeq release offers an advantage for those who want to maintain periodic updates of either the complete collection or a single group. It also includes records that are not available from the companion *genomes* FTP site, such as transcripts in the collection that are maintained independently from, and may not be currently annotated on, a genome assembly. The release is provided with significant documentation of the files installed (/refseq/release/release-catalog/) including MD5 checksums, a list of all installed files, as well as release notes and announcements (/refseq/release/release-notes/).

RefSeq data can also be downloaded from the *genomes* FTP site. In August 2014 NCBI announced a major reorganization of this FTP site which now provides assembly and organism-based access to both GenBank and RefSeq genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/). This directory is further divided into subdirectories based on the same groups that are used in the RefSeq release, each of which provides additional sub-divisions by species. The *genomes* FTP site provides files representing all RefSeq genome assemblies reported in NCBI's Assembly resource (www.ncbi.nlm.nih.gov/assembly/). The advantage of the *genomes* site is that the data can be accessed in an assembly- or organism-specific manner. Data provided includes genome and product (transcript/protein) sequence, annotation, assembly reports and statistics, and MD5 checksums; these data are updated when the genome assembly and/or annotation are updated. This area does not include RefSeq sequences that are outside the scope of a genome assembly or products that are not annotated on a genome.

GROWTH AND STATISTICS

RefSeq FTP release 71 (July 2015) includes more than 77 million sequence records for more than 55 000 organisms. Table 2 summarizes the growth of the RefSeq dataset in the last year in terms of the organisms and number of sequence records represented per each RefSeq release FTP directory area. Bacterial genomes and proteins comprise the bulk of the RefSeq dataset (56% of the total accessions and 76% of the >52 million protein accessions). Significant increases in the number of organisms, proteins, and total records are seen for invertebrate, plant, and eukaryotic organisms which is consistent with the increased number and throughput of genome sequencing projects. A significant factor for the continued high rate of growth of RefSeq data are improvements in genome pipelines that generate annotated RefSeq genomes. Most notably, this includes increased capacity in NCBI's prokaryotic genome annotation pipeline, re-development of the process flow that propagates annotation from eukaryotic GenBank genomes onto RefSeq genomes, and the incorporation of RNA-Seq evidence in NCBI's eukaryotic genome annotation pipeline and its impact on generating model RefSeqs (XM_, XR_ and XP_ accessions, Table 1).

The dramatic decrease in the number of plasmid protein records, and thus in the number of total accessions, reflects the completion of a RefSeq bacterial genome re-annotation project (http://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/) and the adoption of the new data model for prokaryotes, including their plasmids. In this new data model a single RefSeq non-redundant protein accession may be annotated on more than one genomic sequence record when translation of those genomic protein-coding regions results in an identical protein (see http://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/). Redundancy in all bacterial proteins also significantly decreased; however, it is not apparent here due to continued significant increases in the number of bacterial genomes included in the dataset. These changes also resulted in an overall drop in the number of archaeal protein records.

VERTEBRATES

A select group of vertebrates including *Homo sapiens, Mus musculus*, *Rattus norvegicus*, *Gallus gallus*, *Bos taurus* and *Danio rerio* are the major focus of our transcript- and literature-based manual curation efforts. Curators generally work from lists of genes with data conflicts identified by quality assurance (QA) tests, some of which were previously described (12). They follow a detailed set of guide-

Table 2. Annual growth in the number organisms, proteins, and transcripts represented in the comprehensive RefSeq release, per FTP release directory

Release Directory	Organisms	% Change	Transcripts	% Change	Proteins	% Change
Archaea	952	12	1109	318	1037407	-5
Bacteria	39660	40	19650	488	40194748	14
Fungi	3367	18	1438749	17	1440956	17
Invertebrate	1786	29	1435978	76	1367317	74
Mitochondrion	5732	24	112	-15	83208	24
Plant	847	59	2181963	86	2067971	75
Plasmid	2139	31	12	9	126725	-62
Plastid	843	54	120	0	72579	50
Protozoa	273	27	849678	46	865048	45
Vertebrate_mammalian	776	14	3778288	44	3266845	39
Vertebrate_other	2755	26	2097939	85	2023378	84
Viral	4850	17	0	0	230360	15
Complete	55267	34	11803354	56	52494032	20

[a]Counts are based on statistics reports that are available from the RefSeq FTP site at ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-statistics/ (e.g. archaea.acc_taxid_growth.txt and related files). The percent annual change is based on comparing data counts for RefSeq release 71 (July 2015) and RefSeq release 66 (July 2014).

lines when analyzing each gene in order to ensure cross-person consistency in the curated dataset. This analysis involves in-depth sequence assessment and literature review to create reference transcripts, proteins, pseudogenes and RefSeqGene records. RefSeq curators generate transcript variants, resolve sequence errors, remove inaccurate information, update records to correctly represent the biology of the locus, and add valuable functional information to some RefSeq records such as improved protein names, a summary of the function of the gene product, functional features of the gene, and/or relevant publications. Manual curation and literature review by the RefSeq group can result in the representation of unique variants and isoforms that would not be predicted when based solely on computational analysis. For instance, literature review of the human tumor suppressor gene, *PTEN* (phosphatase and tensin homolog, GeneID: 5728) revealed the existence of a longer protein isoform resulting from use of an alternative in-frame upstream CUG initiation codon found at the center of a palindromic sequence upstream of the canonical mRNA translation start codon (13). Strong experimental data indicated that this mitochondrial-specific isoform initiates with a leucine, rather than a methionine (14). The RefSeq data model for eukaryotes provides one transcript explicitly linked to one protein. Therefore, two identical transcript records were provided to reflect translation from the alternate initiation codons; NP_000305.3 represents the 403 amino acid protein that uses the canonical methionine start codon, while NP_001291646.2 represents the mitochondrial-localized 576 amino acid protein that initiates with a leucine. Thus, the curation process serves a dual purpose of providing accurate reference sequences that facilitate precise and reproducible genome annotation and providing records that include relevant biological information. In this section we discuss recent updates, improvements we have made to our manual curation process, and examples of focused curation projects.

RefSeqGene project

The RefSeqGene sub-project defines human genomic sequences to be used as reference standards for well-characterized genes, particularly for use by the clinical genetics community. These sequences serve as a stable foundation for reporting pathogenic variants, for establishing conventions for numbering exons and introns, and for defining the coordinates of other variants. Each RefSeqGene record focuses on a gene-specific genomic region and typically is annotated with a subset of RefSeq transcripts and proteins selected by domain experts. Those selections determine exon features. Alignments of older versions of the canonical RefSeq transcript/protein, as well as other known RefSeqs, are included. These records typically include 5 kilobases (kb) of sequence upstream of the focus gene, and 2 kb of sequence downstream, to support representation of potential regulatory sites or deletions extending beyond the gene feature. A RefSeqGene record may include annotation information for other genes that are located within its boundaries. RefSeqGene records are reviewed initially by locus-specific databases and NCBI staff. RefSeqGene is a member of the LRG collaboration (7) which provides additional review of the sequence data before adding an LRG accession. A recent work focus expanded the number of RefSeqGene records to represent all genes for which at least two clinical tests have been submitted to the NIH Genetic Testing Registry (GTR). At this time there are 5596 RefSeqGene records, of which 633 have an LRG accession. RefSeqGene records can be retrieved by searching the Nucleotide database with 'refseqgene[keyword]', by their LRG accessions, by browsing the RefSeqGene web site (www.ncbi.nlm.nih.gov/refseq/rsg/), or by FTP (ftp://ftp.ncbi.nlm.nih.gov/refseq/H_sapiens/RefSeqGene/).

Incorporation of RNA-Seq and other data types in transcript-based curation

A major goal of the RefSeq curation project is to represent high quality and full-length transcript and protein reference sequences. As such, our curation criteria are primarily based on conventional transcript (mRNA and ESTs) and protein alignments and published evidence. However, vertebrate transcriptome projects have become ever more complex with the majority of new transcript data currently generated by short read sequencing technology. Genome-wide studies looking at global patterns of promoter-associated epigenetic marks also provide evidence of active promot-

ers and/or active transcription. The RefSeq group has adjusted curation practices to incorporate these new data types to enhance our manual annotation, particularly in cases where a gene or variant lacks abundant conventional transcript support. These RNA-Seq and epigenomic studies have generated enormous datasets that present a challenge for gene annotation groups for example through potential false positives and the lack of support for long range exon combinations (15). RefSeq curators mitigate against false positives by selectively incorporating only high quality datasets for consideration into our genome annotation pipeline and into the manual annotation process. RefSeq curators visualize transcript alignments, variation data, and filtered RNA-Seq data in customized displays within an in-house alignment tool incorporated into the NCBI Genome Workbench platform (http://www.ncbi.nlm.nih.gov/tools/gbench/). Curation of human genes utilizes analyzed RNA-Seq reads from the Illumina BodyMap 2 (BioProject: PRJEB2445) and Human Protein Atlas projects (BioProject: PRJEB4337) (16). Additionally curators use promoter-associated histone modification marks such as H3K4me3 from the NIH Roadmap Epigenomic Mapping Consortium (REMC; (17) and the ENCODE (Encyclopedia of DNA Elements) project (18) to verify the presence of an active promoter. RefSeq curators also evaluate polyA-seq data to affirm 3′ completeness of transcripts lacking a polyA tail (19). Additional data types, including PhyloCSF (20), CpGIslands, RepeatMasker (21) and Cap analysis of gene expression (CAGE) data (22), are sometimes used as additional support.

Long non-coding RNAs (lncRNAs)

The RefSeq group continues to significantly expand on the representation of non-coding structural- and micro-RNAs, transcribed pseudogenes, and the largely uncharacterized lncRNAs. This class of genes is generally defined as being transcripts >200 nt in length that lack strong protein-coding potential (23). lncRNA RefSeq records are generated by curation and through the eukaryotic genome annotation pipeline for lncRNA genes. NCBI currently maintains over 540 000 eukaryotic lncRNA RefSeq records, of which over 6700 have been curated and only a few hundred have been functionally characterized. Of these, many have been implicated in human disease, such as BACE1-AS which may play a role in the pathophysiology of Alzheimer's disease, and HOTAIR which has been associated with multiple cancers (24,25). The vast majority of lncRNAs have unknown functions and the absence of long open reading frames presents a challenge in terms of confirming the completeness of the transcript. Furthermore, lncRNA submissions to INSDC are largely based on TSAs from short read datasets that may include artifactual exon combinations. RefSeq curators take a conservative approach to representing lncRNA genes, only manually creating RefSeqs (with a NR_ accession prefix) for high quality transcripts for which we have some certainty of the exon structure. Ideally, the transcript support should be spliced with at least three exons but two-exon and intronless transcripts can be represented if they are supported by promoter-associated epigenomics, poly(A) evidence, additional cDNAs, and/or RNA-Seq data. RefSeq lncRNA records for non-coding genes can be retrieved from NCBI's Nucleotide database using the search string 'biomol ncrna lncrna[Properties]' and selecting the RefSeq filter from the left column.

Functional annotation

The unique contribution of curated eukaryotic RefSeq transcript records is that they integrate functional information with a reference sequence. RefSeq curation staff adds gene summaries, nomenclature, transcript variant text, gene and sequence attributes, and functional features that are available on the RefSeq record and/or through the Gene resource (http://www.ncbi.nlm.nih.gov/gene). In the past year, RefSeq staff have pursued several in-depth annotation projects, some of which are briefly described in the following paragraphs, to add functional data to specific sets of genes where computational tools are unable to accurately represent biological knowledge. These projects include annotation of antimicrobial peptides, endogenous retroviruses, replication-dependent histones, regulatory uORFs, and antizymes.

Antimicrobial peptides (AMPs). AMPs were a recent curation focus (http://ncbiinsights.ncbi.nlm.nih.gov/2015/05/21/) (26). AMPs are naturally occurring peptides that are found in a diverse array of species and have been implicated in many immune roles including bactericidal, antiviral, antifungal and even antitumor activities. A list of over 130 human genes encoding one or more experimentally proven AMPs was gathered from several publicly available AMP datasets and also mined from publications. Most of these AMPs had not been previously identified in the RefSeq database, and none of the AMP databases connected the peptides to their encoding gene. RefSeq curators manually annotated the RefSeq records for each AMP-encoding human gene to ensure that the functional peptide was annotated, to include a publication describing the antimicrobial activity of the peptide, to add a brief summary describing the antimicrobial activity of the encoded AMP, and to store a new RefSeq attribute 'Protein has antimicrobial activity' which is included in the RefSeq attribute structured comment (e.g. NM_001124.2 for ADM; GeneID: 133). To access all of the curated human transcript or protein AMP records, search the nucleotide or protein database using 'Protein has antimicrobial activity[properties]'. Currently, this search will find 191 RefSeq records, including splice variants and protein isoforms.

Endogenous retroviruses (ERVs). Endogenous retroviruses (ERVs) are genomic loci that are derived from the ancestral insertion of an exogenous retrovirus into the host genome. ERV loci are generally out of scope for RefSeq; however, we annotate full-length ERV protein-coding loci that map to a single genomic location if they have evolved to serve a host function, are associated with a known disease, and/or if they have been assigned nomenclature by an official nomenclature committee. About 8% of the human genome is of retroviral origin (27); however due to their ancient origins most human ERV loci have accumulated nonsense mutations and can no longer encode

a protein. The syncytin proteins, which are involved in placental development (28), are a well-known exception to this. Human syncytin-1 and syncytin-2 proteins are encoded by the ERVW-1 (NM_001130925.1, NM_014590.3) and ERVFRD-1 (NM_207582.2) genes. To date we have created 67 RefSeqs for ERV loci, which includes records representing ERV genes from a diverse set of mammals. A new RefSeq attribute category entitled 'endogenous retrovirus' was created for these records and appears in a structured comment on the RefSeq record. These records can be retrieved from the Nucleotide database by searching for 'endogenous retrovirus [properties]'.

Replication-dependent histones. A rapid synthesis of histone mRNAs is required during cell division in order to produce large amounts of histone proteins. Critical to this process are the replication-dependent histone genes that are upregulated during the G1/S phase of cell cycle (29). A specific RefSeq project was undertaken with the aim of curating the full set of replication-dependent histone protein coding genes in human and mouse. These genes have a canonical 3′ histone downstream element (HDE) sequence in the genomic sequence and the resultant mature mRNAs characteristically lack poly(A) tails and instead terminate shortly after an RNA stem-loop structure (30). The HDE element is found on the precursor transcript but is not included on the processed transcript represented by RefSeq. The location of the conserved 16 nucleotide stem-loop structure sequence is indicated on the RefSeq record as a feature annotation entitled 'stem-loop'. An example can be seen on the RefSeq entry NM_003539.3 for HIST1H4D (GeneID: 8360). To date, 127 human and mouse replication-dependent histone RefSeq records have been curated and a RefSeq attribute added which can be used to retrieve these records from the Nucleotide database using the search string 'replication-dependent histone[properties]'.

Regulatory upstream open reading frames (uORFs). Translation of an upstream open reading frame (uORF) can negatively affect translation of the primary protein-coding open reading frame (pORF) (31). This effect does not always completely silence pORF translation and may be dependent on cell type, developmental state or cellular condition. Therefore, although uORFs may be predicted from the six-frame translation of a transcript, the regulatory effect of this element must be determined through experimental validation. RefSeq curators reviewed the literature to find transcripts with experimental evidence of regulatory uORFs and updated the corresponding RefSeq transcript records to add a misc_feature denoting the location of these uORFs. An example is the RefSeq entry NM_000392.4 for ABCC2 (GeneID: 1244). A new RefSeq attribute category entitled 'regulatory uORF' was created and appears in a structured comment on these RefSeq records. Both the annotated feature and the attribute cite the supporting publication by PubMed ID. To date, 260 records have been annotated with this attribute and these records can be retrieved from the Nucleotide database by searching for 'regulatory uORF [properties]'

Antizyme genes. One of the goals of the RefSeq project is to represent genes with exceptional biology that do not follow standard decoding rules of protein synthesis. The ornithine decarboxylase antizyme gene is such an example, where a programmed +1 ribosomal frameshifting mechanism occurs and cannot be predicted by conventional computational tools. A set of vertebrate antizyme transcript and protein records were recently the subject of a manual annotation effort to create standards to improve annotation of these gene products by the eukaryotic genome annotation pipeline (32). The RefSeq records were manually annotated with the split CDS feature to reflect ribosomal slippage, and include a 'ribosomal slippage' attribute with published evidence, various miscellaneous feature annotations (such as the location of the frameshift site) and a brief summary describing the function and novel properties of the gene (e.g. NM_139081.2). These records can be retrieved from either the Nucleotide or Protein database with the search query: vertebrates[orgn] refseq[filter] ribosomal slippage[prop] antizyme[title]. This search currently finds 242 RefSeq records (NM or NP), which includes transcript variants and protein isoforms.

INVERTEBRATES

Invertebrate species represent the vast majority of extant metazoans (33); however, only a relatively small number are represented by sequenced genomes. This despite the fact that many species have critical biomedical importance such as *Anopheles gambiae*, a vector for malaria and *Biomphalaria glabrata*, a vector for schistosomiasis (34,35). Other invertebrates including *Apis mellifera*, *Bombyx mori* and *Crassostrea gigas* have significant commercial value (36–38). The RefSeq group has made efforts to increase the number and scope of invertebrate genomes represented in the dataset by providing annotation via the eukaryotic genome annotation pipeline or by propagating annotation from INSDC submissions onto the RefSeq copy of those genomes. For both process flows we are dependent on the public availability of high quality genomes in INSDC databases and NCBI's Assembly database (www.ncbi.nlm.nih.gov/assembly/). To date 46 invertebrate genomes have been annotated by NCBI including representative species of insects, arachnids, mollusks and basal chordates. We anticipate a significant expansion in the number of insect and other invertebrate genomes annotated as a result of genome initiatives such as the i5k (39), 1KITE (1K Insect Transcriptome Evolution, http://www.1kite.org/) and the Global Invertebrate Genome Alliance (http://giga.nova.edu/) (40).

PLANTS

RefSeq continues to expand the diversity of plant species represented in the dataset. To date, 61 plant species have been included in the RefSeq genomes dataset (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plant/) of which 33 species were annotated through the eukaryotic genome annotation pipeline; the remainder are RefSeq copies of annotated genomes submitted to INSDC. In the future, more plant genomes selected for RefSeq inclusion will be processed by the eukaryote annotation pipeline, rather than propagating

annotation from the INSDC submission. This is a change of policy for the RefSeq plant genomes and will result in greater overall consistency of plant annotation data within the RefSeq dataset. The majority of the RefSeq transcripts and proteins available for plant species are 'model' records (XM_, XP_ and XR_ accessions; Table 1), with a smaller subset of 'known' records (NM_, NR_, NP_) that are maintained independently of the annotation process by a combination of automated processing and manual review. Manual curation of plant transcript and protein data are currently provided for *Zea mays* and *Solanum lycopersicum*. The current curation focus entails extensive sequence review and is targeted toward resolving QA concerns in the current set of transcripts. Error resolution is focused on identifying and removing chimeric transcripts, redundant transcripts and genes, and improving the quality of the represented sequence by assessing indels and mismatches among the RefSeq transcript, the genomic sequence, and orthologous data. For plants, we strive to provide a curated transcript and protein dataset that is consistent with the cultivar selected for genome sequencing and assembly. The curation protocol used for vertebrate data is also used for plants. Thus, RefSeq transcript records may be updated to be based on a different INSDC source sequence, or may be assembled from more than one INSDC sequence record in order to provide a transcript from the preferred cultivar. If INSDC transcript data are not available for the genomic cultivar then a RefSeq transcript may be generated from the assembled genomic sequence based on a combination of transcript or protein alignments, RNA-Seq, and/or published data. A second area of focus is to increase the number of supported known protein-coding transcripts and proteins as this provides a curated reagent that can be used when annotating other plant genomes. Lastly, we are making more RefSeqs representing splice variants when there is sufficient supporting evidence. These efforts will significantly improve the quality of the plant RefSeq dataset and will contribute to improvements in future genome annotations. The current set of plant genomes annotated by the pipeline can be accessed at NCBI's eukaryotic genome annotation pipeline website http://www.ncbi.nlm.nih.gov/genome/annotation_euk/all/ with links to the detailed annotation report and other resources such as species BLAST and FTP.

ALGAE, FUNGI, NEMATODES AND PROTOZOA

The NCBI small eukaryotic genome pipeline is a new automated pipeline designed for the generation of RefSeq records as a result of direct propagation of annotated INSDC records. The RefSeq records thus generated are copies of the GenBank data with some format changes to adhere to RefSeq requirements. The most notable difference between the original INSDC record and the RefSeq record is the addition of the RefSeq transcript product. Although not designed to generate *de novo* genome annotation, the small eukaryotic genome pipeline draws from several of the NCBI eukaryotic genome annotation pipeline modules and their code (http://www.ncbi.nlm.nih.gov/books/NBK169439/).

The 'Small Eukaryotes' designation refers to the pipeline's primary use to generate RefSeq genomes for relatively smaller eukaryotic genomes (compared to those of plants and vertebrates) such as those of algae, protozoa, fungi, nematodes, and some arthropods. However, some large plant genomes are also processed using this pipeline. This pipeline processes high-quality assemblies consisting of chromosomes and/or scaffolds and their components. Those assemblies with high contig and scaffold N50, high quality sequence, and reasonably good INSDC-submitted annotation are prioritized. This pipeline, which replaces a historical process flow that required more manual support, has only recently reached a public production phase and is already yielding an increased number of 'small' eukaryotic genomes represented in RefSeq. Work is ongoing to optimize the pipeline throughput and to add more automation and further minimize curator processing tasks. Longer-term plans include implementing a protein-name management system in order to provide, correct, or improve on the INSDC submitted names over time. Many of the genomes that are in scope for the small eukaryotes pipeline cannot currently be processed by the (large) eukaryotic genome annotation pipeline due to taxonomic diversity and limited availability of transcript data needed to train the de novo annotation pipeline.

Fungal targeted loci

Fungal morphology is highly diverse, ranging from complex multicellular structures to very simple single cells. A variety of morphological structures and spore types can be produced by a single species. Conversely, many species produce similar morphologies (morphs) but are in fact genetically very distant. Until recently, a single species could be validly described with more than one binomial name based on sexual or asexual morphs. In many cases, only a single morph has been described and recorded for a given species, although species closely related to it could have several morphs described and recorded. Consequently, sequence comparisons have been applied in the fungal community to differentiate between species, to track species as they proceed through complex life cycles and to identify cryptic species. As part of the dynamic process of taxonomic re-evaluation, many fungal species corrections are not always up-to-date in GenBank sequence data.

To be a more reliable resource for DNA-based identification, reference sequences derived from type specimens (which act as references for species) need to be labelled with the correct and most up-to-date species name. The Fungi RefSeq targeted loci databases provide this valuable resource. For example, PRJNA177353 is a BioProject that specifically focuses on the internal transcribed spacer (ITS) regions in the nuclear ribosomal cistron which has been used for many years as a phylogenetic marker and recently approved as the formal barcode sequence of Fungi (41). The ITS RefSeq database started out as a collaboration with Index Fungorum, MycoBank and UNITE, as well as a large group of taxonomic specialists. Sequences were selected, mostly from type specimens of valid descriptions, and then current correct species names were associated with the sequences with the aim to represent most of the accepted fun-

gal orders (8). Results from this curation effort have been used and cited by various publications (42–46) and have aided additional efforts at validating subsets of reference sequences, e.g. medically significant species (47).

The aim, with continued curation, is to add sequences from newly described orders and to extend representation to include most of the accepted families with a focus on medically important Fungi. The process also includes making corrections, replacing sequence from verified material with sequence from type material as it becomes available and editing definition lines or removing RefSeq records as taxonomic classifications changes. This ensures that BLAST search results correctly display the current name. The RefSeq ITS records have been extended to represent 3,060 sequences representing 270 families from 39 classes. During the initial collaborative ITS RefSeq effort, a smaller set of sequence accessions from the 28S nuclear large subunit ribosomal gene (LSU) were also collected but not verified. A workflow similar to the ITS record curation process was followed and during continued curation these LSU records have been verified for sequence quality, correct identification, and accurate source data. Close to 500 records (from 800 potential records) representing >100 families from 21 classes were verified and recently released. The 28S dataset can be retrieved from BioProject PRJNA51803 (48).

PROKARYOTES

The NCBI RefSeq prokaryotic genome collection represents assembled prokaryotic genomes with different levels of quality and sampling density. For prokaryotes, based on past community feedback our current policy is to provide genome annotation for all prokaryotic genomes that meet our quality criteria. In recent years, we have faced two major challenges: (i) keeping up with the rapid escalation of submitted prokaryotic genomes; and, (ii) addressing a growing inconsistency in genome annotation due to the use of both an INSDC propagation-based pipeline and different versions of a NCBI *de novo* genome annotation pipeline as developed over time.

With the increasing interest in human pathogens and advancement of DNA sequencing technology, the number of sequenced prokaryotic genomes has rapidly increased in the last decade. Some bacterial strains are often indistinguishable using current genotyping approaches, but minor genetic differences can be detected on the basis of whole-genome sequencing, which is useful for characterizing transmission pathways, identifying antibiotic resistance, and surveying outbreaks. To investigate food-borne pathogens or infection outbreaks, large numbers of nearly identical bacterial genomes have been sequenced and annotated in recent years, resulting in numerous identical proteins, each having a distinct accession number. In 2013 NCBI introduced a new protein data model and accession prefix (WP_) for the RefSeq collection. This change reduced the redundancy in RefSeq prokaryotic proteins and facilitated identification of proteins that were identically found on more than one genome. It also allowed for an improved strategy for managing prokaryotic protein names. These non-redundant records represent unique prokaryotic protein sequences that are independent of any particular bacterial genome and may be annotated on multiple strains or species (www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/).

Historically, RefSeq bacterial genomes annotation was propagated from INSDC submissions, when available, or generated using different versions of NCBI's Prokaryotic Genome Annotation Pipeline (which is also offered as a service for GenBank submissions). This resulted in accumulated inconsistencies in both structural and functional annotation in the RefSeq prokaryotic dataset. Over the past two years NCBI improved several aspects of the Prokaryotic Genome Annotation Pipeline to increase capacity and further standardize annotation rules. Our pipeline combines a gene calling algorithm, GeneMarkS+ (49,50), with an alignment-based gene detection approach and is capable of annotating both complete and draft WGS genomes. The pipeline currently predicts protein-coding genes, structural RNAs (5S, 16S and 23S), tRNAs and small non-coding RNAs.

In 2015, we released a comprehensive annotation update for RefSeq prokaryotic genomes in order to harmonize genome annotation and complete the transition to the new protein data model. A new prokaryotic protein name database, name specifications, and an evidence-based strategy were developed and are currently in the process of being deployed. Thus far, over 3 million protein records have updated names in an initial demonstration of the approach. The new prokaryotic data model offers a significant advantage to name management as the protein name is carried with the protein sequence record; updating the name on that protein record results in automatically propagating the update to all genomes that are annotated with that accession number.

RefSeq prokaryotic genomes are organized in several new categories such as reference genomes and representative genomes based on curated attributes and assembly and annotation quality measures (www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/) (51). Reference genomes are manually selected 'gold standard' complete genomes with high quality annotation and the highest level of experimental support for structural and functional annotation. Currently, a small dataset of 122 reference genomes are manually annotated by collaborating groups and NCBI staff. The reference genomes are available at: http://www.ncbi.nlm.nih.gov/genome/browse/reference/. Representative genomes are computationally calculated and selected to represent diverse species. The representative genomes are available at: www.ncbi.nlm.nih.gov/genome/browse/representative/.

RefSeq prokaryotic genome data can be accessed in BLAST databases, web resources (Assembly, BioProject, Genome, Nucleotide and Protein), through NCBI's programming utilities, or can be downloaded from the genomes or refseq FTP sites. A custom 'Microbes' BLAST page, accessed from the BLAST home page, provides options to search against all RefSeq prokaryotic genomes, the Reference and Representative genomes subset, or to restrict the search to a specific taxa. A subset of prokaryotic genomes are annotated with a NCBI Gene ID and can be retrieved in NCBI's Gene resource or from the Gene FTP site. For archaea, this is provided for most complete genomes. For bacteria, this is provided for reference genomes and the rep-

resentative genomes for species that have at least 10 genome submissions.

Prokaryotic targeted loci

In prokaryotes, the 16S ribosomal RNA sequence has become a standard molecular marker for the description of a new species. While these marker sequences have become widely used, the quality of the sequence data and the associated meta-data being submitted to INSDC databases varies considerably. Recognizing the importance of access to high quality data for these markers, NCBI has expanded its targeted loci project to provide an up-to-date source of curated data. The targeted loci project currently maintains nearly 18 000 16S ribosomal RNA reference sequences of which over 95% are from type strains. The type strains are considered the exemplar of the species and it is essential that type strain data be annotated with correct metadata and be free from contamination.

This work involved an exhaustive review and update to the underlying taxonomy database which was used in conjunction with NCBI's type strain Entrez filter to retrieve candidate sequences. The sequence data and their associated taxonomy/meta-data have been reviewed and corrected to include the most up-to-date information. If a sequence failed validation or could not be accurately validated, it was excluded. These reference sequences can now be used as 'gold standards' for the analysis of existing and new rRNA sequences.

Bacterial and Archaeal 16S rRNA datasets are available from BioProject (PRJNA33175 and PRJNA33317, respectively). A custom BLAST database is also available ('16S ribosomal RNA sequences (Bacteria and Archaea)').

VIRUSES

The RefSeq data model for viruses differs from that of other organisms. In general, only one complete RefSeq genome is created for each viral species. Occasionally multiple RefSeq records are created within a given viral species to reflect well defined genotypes or important laboratory and/or wild strains. Additional genomes for a given species are validated for taxonomy and completeness and then indexed as sequence 'neighbors' (52). Both RefSeq and neighbor genomes are retrievable through the specialized Viral Genome Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/) and from Entrez Nucleotide and Genome pages using 'RefSeq Genome for Species' and 'Other INSDC Genome Sequences' links (52).

Taxonomy is a major concern to viral genomics as there are 3186 viral species officially recognized by the International Committee on Taxonomy of Viruses (ICTV) (53) and 4834 complete genomes from both official and provisional viral species available from INSDC databases. The NCBI Pairwise Sequence Comparison (PASC) tool was developed to assist in the classification of viral genomes based on global and/or local alignments between genomes (http://www.ncbi.nlm.nih.gov/sutils/pasc/). The scope of this tool has been expanded to include a number of virus families and other taxonomic groups, and it has been used to help support the demarcation of new taxonomic criteria (54–57).

Another emerging problem in viral genomics is inconsistent and/or inaccurate annotation among related viral genome sequences. This issue often reflects differing annotation processes and ongoing experimental work and can lead to confusion among data consumers and make comparative analysis between genomes difficult. This problem is addressed within the NCBI Virus Variation Resource (http://www.ncbi.nlm.nih.gov/genome/viruses/variation/) where computational pipelines are employed to provide up-to-date, standardized annotation for several viruses (58). Currently, these pipelines calculate standardized gene and protein boundaries for all Influenza virus, Dengue virus, and West Nile virus sequences and standardized gene and protein names and metadata terms for these and two other viruses, Middle East respiratory coronavirus and Ebolavirus. This standardized data is then leveraged within a specialized, metadata-centric search interface that facilitates the easy retrieval of sequences based on specific biological criteria.

Maintaining up-to-date, widely accepted annotation standards requires continuous collaboration with the greater scientific community. The NCBI Viral Genome Annotation Working Group was established to leverage consortia of public databases, sequencing centers, and research groups to develop standardized sequence annotation as well as isolate naming schemes for different groups of viruses (59–63). This approach not only establishes standards for viral annotation but also represents these standards within the current RefSeq record, ensuring accessibility for all database users and submitters. Similar collaborations are also necessary to support value added, interpretive resources such as the HIV-1, human interaction database (http://www.ncbi.nlm.nih.gov/genome/viruses/retroviruses/hiv-1/interactions/) (64). Collaborators from the Southern Research Institute provide documented HIV-1, human molecular interactions curated from the literature and NCBI maintains a user friendly resource where users can query for specific types of interactions and find more information about the genes involved.

FUTURE DIRECTIONS

The RefSeq project is unique in offering a reference sequence dataset of transcripts, proteins and genomes that encompasses all kingdoms of life and has been actively maintained and updated over time to incorporate improved computational strategies, new data types, and new knowledge. We have demonstrated the capability and capacity to respond to recent rapid increases in the number of sequenced genomes submitted to INSDC databases. We have defined a diverse set of policies and strategies for the curation and annotation of eukaryotic, prokaryotic, and viral species to meet the different needs of organism-specific communities. The RefSeq dataset is widely used as a reference standard for many different analyses including human and pathogen clinical applications, comparative genomics, expression assays, sequence variation interpretation, and both array and probe construction. At NCBI, the RefSeq dataset is integrated into multiple resources including Assembly, BLAST, Epigenomics, Gene (where RefSeq annotation is the pri-

mary basis for most Gene entries), Genome, dbSNP, dbVar, Variation Viewer, and more.

We will continue to target manual curation to improve structural and functional information for human and other vertebrate genomes. Our conservative manual curation approach ensures the continued high quality and reliability of the human, mouse, and other 'known' RefSeq records which serves the needs of those who need a well-supported definition of alternate exons (fewer false positives). The addition of RNA-Seq data to our annotation pipeline significantly increased our annotation of alternate splice variants as model RefSeqs to serve the needs of those who want a more comprehensive, but still well-supported, definition of the exome (fewer false negatives). While both known and model RefSeqs report the support evidence on the sequence record, they use distinct approaches to do so. Future efforts will be directed toward harmonizing evidence reporting for both 'known' and 'model' RefSeqs so that users can more easily identify this information. We will also be adding a new data type to the human and mouse RefSeq collection in the near future to represent experimentally reported regulatory and functional elements with known (or reasonably inferred) functional consequences.

For prokaryotic genomes, we continue to work on refining aspects of the structural annotation that is generated by the Prokaryotic Genome Annotation Pipeline. Our work toward a new approach to manage functional information is still being refined and will be described elsewhere. We anticipate re-annotating the entire RefSeq prokaryotic genomes dataset when new versions of our prokaryotic annotation pipeline become available (to improve structural annotation). The decision to annotate all RefSeq prokaryotes using a single method, together with the sheer volume of this dataset, necessitates a different approach that leverages multiple sources of evidence to provide functional information. Protein names will be updated on an ongoing basis as organized by protein families or categories of evidence type. Our goals for the coming year include greater integration of Rfam (65) in our annotation pipeline, expanded collaboration, improved protein names, and reporting support evidence on the protein sequence record.

ACKNOWLEDGEMENTS

We would like to thank the scientific community for constructive feedback, suggestions, error reports, and collaborations over the last 15 years that have contributed toward the quality and accuracy of the represented sequence, structural annotation, and functional annotation.

FUNDING

Intramural Research Program of the NIH, National Library of Medicine. Funding for open access charge: The Intramural Research Program of the National Institutes of Health, National Library of Medicine.
Conflict of interest statement. None declared.

REFERENCES

1. Nosek,B.A., Alter,G., Banks,G.C., Borsboom,D., Bowman,S.D., Breckler,S.J., Buck,S., Chambers,C.D., Chin,G., Christensen,G. *et al.* (2015) SCIENTIFIC STANDARDS. Promoting an open research culture. *Science*, **348**, 1422–1425.
2. Gray,K.A., Yates,B., Seal,R.L., Wright,M.W. and Bruford,E.A. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res.*, **43**, D1079–D1085.
3. Ruzicka,L., Bradford,Y.M., Frazer,K., Howe,D.G., Paddock,H., Ramachandran,S., Singer,A., Toro,S., Van Slyke,C.E., Eagle,A.E. *et al.* (2015) ZFIN, The zebrafish model organism database: Updates and new directions. *Genesis*, **53**, 498–509.
4. UniProt,C. (2015) UniProt: a hub for protein information. *Nucleic acids Res.*, **43**, D204–212.
5. Kozomara,A. and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, D68–73.
6. McGarvey,K.M., Goldfarb,T., Cox,E., Farrell,C.M., Gupta,T., Joardar,V.S., Kodali,V.K., Murphy,M.R., O'Leary,N.A., Pujar,S. *et al.* (2015) Mouse genome annotation by the RefSeq project. *Mamm. Genome.*, **26**, 379–390.
7. Dalgleish,R., Flicek,P., Cunningham,F., Astashyn,A., Tully,R.E., Proctor,G., Chen,Y., McLaren,W.M., Larsson,P., Vaughan,B.W. *et al.* (2010) Locus Reference Genomic sequences: an improved basis for describing human DNA variants. *Genome Med.*, **2**, 24.
8. Schoch,C.L., Robbertse,B., Robert,V., Vu,D., Cardinali,G., Irinyi,L., Meyer,W., Nilsson,R.H., Hughes,K., Miller,A.N. *et al.* (2014) Finding needles in haystacks: linking scientific names, reference specimens and molecular data for Fungi. *Database*, 1–21.
9. Zhang,G., Li,C., Li,Q., Li,B., Larkin,D.M., Lee,C., Storz,J.F., Antunes,A., Greenwold,M.J., Meredith,R.W. *et al.* (2014) Comparative genomics reveals insights into avian genome evolution and adaptation. *Science*, **346**, 1311–1320.
10. Jarvis,E.D., Mirarab,S., Aberer,A.J., Li,B., Houde,P., Li,C., Ho,S.Y., Faircloth,B.C., Nabholz,B., Howard,J.T. *et al.* (2014) Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science*, **346**, 1320–1331.
11. Farrell,C.M., O'Leary,N.A., Harte,R.A., Loveland,J.E., Wilming,L.G., Wallin,C., Diekhans,M., Barrell,D., Searle,S.M., Aken,B. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, **42**, D865–D872.
12. Pruitt,K.D., Tatusova,T. and Maglott,D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.
13. Hopkins,B.D., Fine,B., Steinbach,N., Dendy,M., Rapp,Z., Shaw,J., Pappas,K., Yu,J.S., Hodakoski,C., Mense,S. *et al.* (2013) A secreted PTEN phosphatase that enters cells to alter signaling and survival. *Science*, **341**, 399–402.
14. Liang,H., He,S., Yang,J., Jia,X., Wang,P., Chen,X., Zhang,Z., Zou,X., McNutt,M.A., Shen,W.H. *et al.* (2014) PTENalpha, a PTEN isoform translated through alternative initiation, regulates mitochondrial function and energy metabolism. *Cell Metab.*, **19**, 836–848.
15. Bolouri,H. (2014) Modeling genomic regulatory networks with big data. *Trends Genet.: TIG*, **30**, 182–191.
16. Fagerberg,L., Hallstrom,B.M., Oksvold,P., Kampf,C., Djureinovic,D., Odeberg,J., Habuka,M., Tahmasebpoor,S., Danielsson,A., Edlund,K. *et al.* (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol. Cell. Proteomics : MCP*, **13**, 397–406.
17. Bernstein,B.E., Stamatoyannopoulos,J.A., Costello,J.F., Ren,B., Milosavljevic,A., Meissner,A., Kellis,M., Marra,M.A., Beaudet,A.L., Ecker,J.R. *et al.* (2010) The NIH roadmap epigenomics mapping consortium. *Nat. Biotechnol.*, **28**, 1045–1048.
18. Hoffman,M.M., Ernst,J., Wilder,S.P., Kundaje,A., Harris,R.S., Libbrecht,M., Giardine,B., Ellenbogen,P.M., Bilmes,J.A., Birney,E. *et al.* (2013) Integrative annotation of chromatin elements from ENCODE data. *Nucleic Acids Res.*, **41**, 827–841.
19. Derti,A., Garrett-Engele,P., Macisaac,K.D., Stevens,R.C., Sriram,S., Chen,R., Rohl,C.A., Johnson,J.M. and Babak,T. (2012) A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, **22**, 1173–1183.
20. Lin,M.F., Jungreis,I. and Kellis,M. (2011) PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics*, **27**, i275–282.

21. Price,A.L., Jones,N.C. and Pevzner,P.A. (2005) De novo identification of repeat families in large genomes. *Bioinformatics*, **21 Suppl 1**, i351–358.
22. Kodzius,R., Kojima,M., Nishiyori,H., Nakamura,M., Fukuda,S., Tagami,M., Sasaki,D., Imamura,K., Kai,C., Harbers,M. *et al.* (2006) CAGE: cap analysis of gene expression. *Nat. Methods*, **3**, 211–222.
23. Morris,K.V. and Mattick,J.S. (2014) The rise of regulatory RNA. *Nat. Rev. Genet.*, **15**, 423–437.
24. Evin,G. and Hince,C. (2013) BACE1 as a therapeutic target in Alzheimer's disease: rationale and current status. *Drugs Aging*, **30**, 755–764.
25. Yu,X. and Li,Z. (2015) Long non-coding RNA HOTAIR:a novel oncogene (review). *Mol. Med. Rep.*, **12**, 5611–5618.
26. Zasloff,M. (2002) Antimicrobial peptides in health and disease. *N. Engl. J. Med.*, **347**, 1199–1200.
27. Lander,E.S., Linton,L.M., Birren,B., Nusbaum,C., Zody,M.C., Baldwin,J., Devon,K., Dewar,K., Doyle,M., FitzHugh,W. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
28. Mi,S., Lee,X., Li,X., Veldman,G.M., Finnerty,H., Racie,L., LaVallie,E., Tang,X.Y., Edouard,P., Howes,S. *et al.* (2000) Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*, **403**, 785–789.
29. Marzluff,W.F., Gongidi,P., Woods,K.R., Jin,J. and Maltais,L.J. (2002) The human and mouse replication-dependent histone genes. *Genomics*, **80**, 487–498.
30. Kari,V., Karpiuk,O., Tieg,B., Kriegs,M., Dikomey,E., Krebber,H., Begus-Nahrmann,Y. and Johnsen,S.A. (2013) A subset of histone H2B genes produces polyadenylated mRNAs under a variety of cellular conditions. *PLoS One*, **8**, e63745.
31. Barbosa,C., Peixeiro,I. and Romao,L. (2013) Gene expression regulation by upstream open reading frames and human disease. *PLoS Genet.*, **9**, e1003529.
32. Rajput,B., Murphy,T.D. and Pruitt,K.D. (2015) RefSeq curation and annotation of antizyme and antizyme inhibitor genes in vertebrates. *Nucleic Acids Res.*, **43**, 7270–7279.
33. Zhang,Z.Q. (2013) Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (Addenda 2013). *Zootaxa*, **3703**, 1–82.
34. Holt,R.A., Subramanian,G.M., Halpern,A., Sutton,G.G., Charlab,R., Nusskern,D.R., Wincker,P., Clark,A.G., Ribeiro,J.M., Wides,R. *et al.* (2002) The genome sequence of the malaria mosquito Anopheles gambiae. *Science*, **298**, 129–149.
35. Knight,M., Arican-Goktas,H.D., Ittiprasert,W., Odoemelam,E.C., Miller,A.N. and Bridger,J.M. (2014) Schistosomes and snails: a molecular encounter. *Front. Genet.*, **5**, 230.
36. Honeybee Genome Sequencing, C. (2006) Insights into social insects from the genome of the honeybee Apis mellifera. *Nature*, **443**, 931–949.
37. Xia,Q., Zhou,Z., Lu,C., Cheng,D., Dai,F., Li,B., Zhao,P., Zha,X., Cheng,T., Chai,C. *et al.* (2004) A draft sequence for the genome of the domesticated silkworm (Bombyx mori). *Science*, **306**, 1937–1940.
38. Zhang,G., Fang,X., Guo,X., Li,L., Luo,R., Xu,F., Yang,P., Zhang,L., Wang,X., Qi,H. *et al.* (2012) The oyster genome reveals stress adaptation and complexity of shell formation. *Nature*, **490**, 49–54.
39. i5K,Consortium (2013) The i5K Initiative: advancing arthropod genomics for knowledge, human health, agriculture, and the environment. *J. Heredity*, **104**, 595–600.
40. Scientists,G.C.o., Bracken-Grissom,H., Collins,A.G., Collins,T., Crandall,K., Distel,D., Dunn,C., Giribet,G., Haddock,S., Knowlton,N. *et al.* (2014) The Global Invertebrate Genomics Alliance (GIGA): developing community resources to study diverse invertebrate genomes. *J. Heredity*, **105**, 1–18.
41. Schoch,C.L., Seifert,K.A., Huhndorf,S., Robert,V., Spouge,J.L., Levesque,C.A., Chen,W., Bolchacova,E., Voigt,K., Crous,P.W. *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 6241–6246.
42. Visagie,C.M., Houbraken,J., Frisvad,J.C., Hong,S.B., Klaassen,C.H., Perrone,G., Seifert,K.A., Varga,J., Yaguchi,T. and Samson,R.A. (2014) Identification and nomenclature of the genus Penicillium. *Stud. Mycol.*, **78**, 343–371.
43. Corte,L., di Cagno,R., Groenewald,M., Roscini,L., Colabella,C., Gobbetti,M. and Cardinali,G. (2015) Phenotypic and molecular diversity of Meyerozyma guilliermondii strains isolated from food and other environmental niches, hints for an incipient speciation. *Food Microbiol.*, **48**, 206–215.
44. Federhen,S. (2015) Type material in the NCBI Taxonomy Database. *Nucleic Acids Res.*, **43**, D1086–D1098.
45. Nilsson,R.H., Tedersoo,L., Ryberg,M., Kristiansson,E., Hartmann,M., Unterseher,M., Porter,T.M., Bengtsson-Palme,J., Walker,D.M., de Sousa,F. *et al.* (2015) A comprehensive, automatically updated fungal ITS sequence dataset for reference-based chimera control in environmental sequencing efforts. *Microb. Environ/JSME*, **30**, 145–150.
46. Mittelbach,M., Yurkov,A.M., Nocentini,D., Nepi,M., Weigend,M. and Begerow,D. (2015) Nectar sugars and bird visitation define a floral niche for basidiomycetous yeast on the Canary Islands. *BMC Ecol.*, **15**, 2.
47. Irinyi,L., Serena,C., Garcia-Hermoso,D., Arabatzis,M., Desnos-Ollivier,M., Vu,D., Cardinali,G., Arthur,I., Normand,A.C., Giraldo,A. *et al.* (2015) International Society of Human and Animal Mycology (ISHAM)-ITS reference DNA barcoding database–the quality controlled standard tool for routine identification of human and animal pathogenic fungi. *Med. Mycol.*, **53**, 313–337.
48. Schoch,C.L., Seifert,K.A., Huhndorf,S., Robert,V., Spouge,J.L., Levesque,C.A., Chen,W., Fungal Barcoding,C. and Fungal Barcoding Consortium Author, L. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 6241–6246.
49. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
50. Borodovsky,M. and Lomsadze,A. (2014) Gene identification in prokaryotic genomes, phages, metagenomes, and EST sequences with GeneMarkS suite. *Curr. Protoc. Microbiol.*, **32**, Unit 1E 7.
51. Tatusova,T., Ciufo,S., Federhen,S., Fedorov,B., McVeigh,R., O'Neill,K., Tolstoy,I. and Zaslavsky,L. (2015) Update on RefSeq microbial genomes resources. *Nucleic Acids Res.*, **43**, D599–D605.
52. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic Acids Res.*, **43**, D571–D577.
53. Adams,M.J., Lefkowitz,E.J., King,A.M., Bamford,D.H., Breitbart,M., Davison,A.J., Ghabrial,S.A., Gorbalenya,A.E., Knowles,N.J., Krell,P. *et al.* (2015) Ratification vote on taxonomic proposals to the International Committee on Taxonomy of Viruses (2015). *Arch. Virol.*, **160**, 1837–1850.
54. Bao,Y., Chetvernin,V. and Tatusova,T. (2012) PAirwise Sequence Comparison (PASC) and its application in the classification of filoviruses. *Viruses*, **4**, 1318–1327.
55. Bao,Y., Chetvernin,V. and Tatusova,T. (2014) Improvements to pairwise sequence comparison (PASC): a genome-based web tool for virus classification. *Arch. Virol.*, **159**, 3293–3304.
56. Kuhn,J.H., Durrwald,R., Bao,Y., Briese,T., Carbone,K., Clawson,A.N., deRisi,J.L., Garten,W., Jahrling,P.B., Kolodziejek,J. *et al.* (2015) Taxonomic reorganization of the family Bornaviridae. *Arch. Virol.*, **160**, 621–632.
57. Radoshitzky,S.R., Bao,Y., Buchmeier,M.J., Charrel,R.N., Clawson,A.N., Clegg,C.S., DeRisi,J.L., Emonet,S., Gonzalez,J.P., Kuhn,J.H. *et al.* (2015) Past, present, and future of arenavirus taxonomy. *Arch. Virol.*, **160**, 1851–1874.
58. Brister,J.R., Bao,Y., Zhdanov,S.A., Ostapchuck,Y., Chetvernin,V., Kiryutin,B., Zaslavsky,L., Kimelman,M. and Tatusova,T.A. (2014) Virus Variation Resource–recent updates and future directions. *Nucleic Acids Res.*, **42**, D660–665.
59. Seto,D., Chodosh,J., Brister,J.R., Jones,M.S. and Members of the Adenovirus Research, C. (2011) Using the whole-genome sequence to characterize and name human adenoviruses. *J. Virol.*, **85**, 5701–5702.
60. Matthijnssens,J., Ciarlet,M., McDonald,S.M., Attoui,H., Banyai,K., Brister,J.R., Buesa,J., Esona,M.D., Estes,M.K., Gentsch,J.R. *et al.* (2011) Uniformity of rotavirus strain nomenclature proposed by the Rotavirus Classification Working Group (RCWG). *Arch. Virol.*, **156**, 1397–1413.
61. Brister,J.R., Bao,Y., Kuiken,C., Lefkowitz,E.J., Le Mercier,P., Leplae,R., Madupu,R., Scheuermann,R.H., Schobel,S., Seto,D. *et al.* (2010) Towards viral genome annotation standards, report from the 2010 NCBI Annotation Workshop. *Viruses*, **2**, 2258–2268.

62. Brister,J.R., Le Mercier,P. and Hu,J.C. (2012) Microbial virus genome annotation-mustering the troops to fight the sequence onslaught. *Virology*, **434**, 175–180.
63. Kuhn,J.H., Andersen,K.G., Bao,Y., Bavari,S., Becker,S., Bennett,R.S., Bergman,N.H., Blinkova,O., Bradfute,S., Brister,J.R. *et al.* (2014) Filovirus RefSeq entries: evaluation and selection of filovirus type variants, type sequences, and names. *Viruses*, **6**, 3663–3682.
64. Ako-Adjei,D., Fu,W., Wallin,C., Katz,K.S., Song,G., Darji,D., Brister,J.R., Ptak,R.G. and Pruitt,K.D. (2015) HIV-1, human interaction database: current status and new features. *Nucleic Acids Res.*, **43**, D566–570.
65. 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.