D32–D36 *Nucleic Acids Research, 2009, Vol. 37, Database issue* *Published online 16 October 2008*
doi:10.1093/nar/gkn721

NCBI Reference Sequences: current status, policy and new initiatives

Kim D. Pruitt*, Tatiana Tatusova, William Klimke and Donna R. Maglott

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health,
Rm 4As.47B, 45 Center Drive, Bethesda, MD, USA

Received September 15, 2008; Revised September 26, 2008; Accepted September 30, 2008

ABSTRACT

NCBI's Reference Sequence (RefSeq) database (http://www.ncbi.nlm.nih.gov/RefSeq/) is a curated non-redundant collection of sequences representing genomes, transcripts and proteins. RefSeq records integrate information from multiple sources and represent a current description of the sequence, the gene and sequence features. The database includes over 5300 organisms spanning prokaryotes, eukaryotes and viruses, with records for more than 5.5 × 10^6 proteins (RefSeq release 30). Feature annotation is applied by a combination of curation, collaboration, propagation from other sources and computation. We report here on the recent growth of the database, recent changes to feature annotations and record types for eukaryotic (primarily vertebrate) species and policies regarding species inclusion and genome annotation. In addition, we introduce RefSeqGene, a new initiative to support reporting variation data on a stable genomic coordinate system.

INTRODUCTION

NCBI's Reference Sequence (RefSeq) is a public database of nucleotide and protein sequences with feature and bibliographic annotation. The RefSeq database is built and distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine located at the US National Institutes of Health. RefSeq records are made publicly available, at no cost, by multiple methods: (i) interactive query over the internet using text via Entrez (1); (ii) Basic Local Alignment Search Tool (BLAST) (2,3) programs; (iii) scripted query using E-Utilities (4); and (iv) download by FTP. There is a formal bi-monthly release cycle (odd-numbered months) for RefSeq, with each release, incremental daily updates and special reports for some species provided from the FTP site. RefSeqs are an integral part of many resources at NCBI, including the Gene database (5), the Map Viewer and HomoloGene.

NCBI builds RefSeqs from the sequence data available in the archival database GenBank (6), which is a comprehensive public repository of sequences submitted to, and exchanged among the International Nucleotide Sequence Database Collaboration (INSDC). RefSeq records are not part of GenBank although they can be retrieved from NCBI using the same interfaces, such as Entrez Nucleotide and Entrez Protein. For a comparison of the two databases, please see the GenBank chapter, appendix, in the online NCBI Handbook (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid = handbook.section.GenBank_ASM).

The RefSeq collection is unique in providing a heavily curated, non-redundant, explicitly linked nucleotide and protein database representing significant taxonomic diversity. The RefSeq database provides a critical foundation for integrating sequence, genetic and functional information and is used internationally as a standard for genome annotation. RefSeq records can be identified by a distinct accession format, which includes an underscore ('_') at the third position. A full definition of the RefSeq accession space and availability is available on the RefSeq Web site. This web site provides many details about the RefSeq project including links to additional documentation about the curation process in the RefSeq chapter of the NCBI Handbook (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid = handbook.chapter.ch18).

GROWTH

The size of the comprehensive bi-monthly RefSeq release continues to grow in pace with the large-scale genome and cDNA sequencing projects. As of July 2008, the release included records from 5395 species and represented 5 590 364 protein records with the majority from bacterial genomes (4 120 701 proteins) and the next largest number provided for fungal, then mammalian species (346 470 and 313 549, respectively; see Table 1). From July 2007

*To whom correspondence should be addressed. Tel: +1 301 435 5898; Fax: +1 301 480 2918; Email: pruitt@ncbi.nlm.nih.gov

Table 1. Annual growth of the RefSeq ftp release[a]

Release date	June 2003	July 2004	July 2005	July 2006	July 2007	July 2008
Release number	1	6	12	18	24	30
Number of species	2005	2467	2969	3695	4511	5395
Annual percent growth		23	20	24	22	20
Number of proteins	785 143	1 050 975	1 695 929	2 762 164	3 866 210	5 590 364
Annual percent growth		34	61	63	40	45

[a] RefSeq statistics are reported in the release notes provided for each release (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/) and archives are also available (ftp://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/archive/).

Table 2. Annual growth of curated records

	Total	Microbial	Mammalian
Release 24 (July 2007)	210 503	113 640	27 069
Release 30 (July 2008)	265 002	156 834	34 027
Annual growth (%)	26	38	26

Table 3. Percent curation of release 30 per taxonomic group

Release node	Curated species[a] (%)	Curated proteins (%)
Complete	41	5
Fungi	20	2
Invertebrate[b]	89	20
Microbial	36	4
Plant	32	2
Protozoa	16	0
Vertebrate_mammalian	86	11
Vertebrate_other	92	10
Viral	17	10

[a] The total number of species per RefSeq release node is calculated by counting distinct NCBI tax_ids annotated for all RefSeq records available in that node.
[b] Records for drosophila species are tracked as curated by FlyBase by default.

(RefSeq release 24) to July 2008 (RefSeq release 30) the number of species included in the RefSeq release increased by 20% and the total number of records increased by 53% with a 45% growth in the protein collection. The number of curated records also continues to increase (Table 2); this growth represents curation by NCBI staff in addition to curation by collaborators, and an expansion in the number of genomes that are supported by collaborations. As shown in Table 3, there is at least one curated record for 41% of the species represented in the RefSeq release 30 and 5% of the 5.5×10^6 proteins have been curated. These percentages vary with taxonomic groups, because there may be more significant support for some by collaborators as well as NCBI curation staff. For example, there is at least one curated record for 86% of the species in the mammalian group but when calculated in terms of the total number of mammalian proteins, only11% are curated. This discrepancy reflects both the large amount of curation at the species level (88%) for 233 mammalian mitochondrial genomes, which encode a relatively small number of proteins (3055, 88% of which are curated), and a smaller amount of curation of the large number of proteins from the nuclear genomes of some of the mammals included in RefSeq, because the focuses are on the human and mouse reference genomes. For the human genome, 53% of all proteins annotated on all human genome assemblies have been curated. This includes the set of proteins that are generated as a product of NCBI's genome annotation pipeline (with a XP_ accession prefix) in addition to the proteins that are based on transcript data and publications, which are the primary focus of curation (with a NP_ accession prefix). When considering curation of proteins annotated on the human reference genome assembly or proteins with a NP_ accession prefix, then 79% of the human RefSeq proteins have been curated. The focus on human and mouse is supported by the Consensus CDS (CCDS) collaboration (see http://www.ncbi.nlm.nih.gov/projects/CCDS).

NEW FEATURES FOR EUKARYOTIC REFSEQ RECORDS

Support for eukaryotic, primarily vertebrate, records was modified to represent a larger number of pseudogene and non-coding transcripts, and to add feature annotation, as described below.

Non-transcribed pseudogenes

For human and mouse, the sequence defining each non-transcribed pseudogene is derived from the reference genome assembly when possible. Previously, pseudogenes were defined based on any genomic record available in GenBank. A subset may still be defined on records other than those used for the reference assembly if the reference assembly is incomplete at that location, or if the pseudogene is known not to occur in the reference assembly (e.g. a known haplotype or strain difference).

Non-coding transcripts

More non-coding RNAs are being included in the RefSeq collection; this subset includes transcribed pseudogenes, antisense transcripts, known functional RNAs and transcribed loci of unknown function that do not appear to be protein coding. In addition, alternatively spliced transcripts of protein-coding loci that severely truncate or otherwise render the transcript unlikely to be capable of supporting translation are also represented as a non-coding sequence, including transcripts from protein-coding loci that are candidates for nonsense-mediated decay [NMD; (7)]. However, proteins are still represented for a subset of NMD candidate transcripts if there is publication support or if all transcripts available consistently exhibit the same extended UTR pattern for a known gene. For example, see NR_024147.1 and NM_001005845.1. Non-coding RNA records use the accession prefix NR_ or XR_. Representative records can be retrieved with the Entrez nucleotide query 'srcdb_refseq[prop] AND biomol_RNA[prop] NOT biomol_mRNA[prop]'. Coverage of this molecule type is known to be incomplete.

Exons

Exon feature annotation is now calculated for transcripts and some non-transcribed pseudogenes for human and mouse records. Exon annotation on transcript records is computed by aligning the transcript record to the reference genome assembly, using the program Splign (8) and interpreting the alignment result. Exon names are incremented according to 5′ to 3′ order of all exons identified based on available RefSeq transcripts for the gene. This annotation highlights transcript variant differences; for example, it is more apparent when a variant omits an exon as there is a gap in the exon names. Exon annotation (pseudoexons) for non-transcribed pseudogenes is calculated by aligning the RefSeq transcript from the corresponding functional gene (using Splign) to the pseudogene genomic region and interpreting the alignment result. Exon information is displayed in the flat file and included in files provided for FTP. This annotation provides a more complete description of RefSeq transcript variants by providing information on the locations on a spliced transcript that correspond to gene exons and exon names. Exon features are calculated on a weekly basis after a record becomes publicly available in NCBI databases.

Primary block

Multiple submissions to the INSDC are often used to construct the RefSeq record in order to represent a more complete transcript; to assemble a genomic region that is manually annotated (with NG_ accession prefix); or to select a nucleotide polymorphic variant that is thought to be the better representative. The PRIMARY block displayed on a flat file record indicates the specific coordinates in the RefSeq record (REFSEQ_SPAN) and the corresponding coordinates from each GenBank source that was used to assemble the RefSeq (PRIMARY_IDENTIFIER and PRIMARY_SPAN). The PRIMARY block follows the COMMENT block and is available in the ASN.1 as a seq hist assembly block. This information is provided for vertebrates and a small number of other species. For example, see accessions: (i) NG_008407.1 which represents a RefSeqGene genomic record (see below); (ii) NM_000539.3 which is a human transcript record that was assembled from genomic sequence based on transcript alignments; or, (iii) NM_000207.2 which is a human transcript record that was assembled from two GenBank transcripts to represent a 3′-UTR that is both more complete and more consistent with the reference genome assembly. Although a given RefSeq transcript may be assembled from more than one GenBank record, please note that the set of GenBank transcripts that derive from the gene in question must generally support, and not contradict, the final exon combination represented in RefSeq transcripts.

Comments

Comments about curation decisions made by the CCDS collaboration are now included in human and mouse RefSeq transcript and protein records. These comments are provided when the CDS structure, as annotated on the reference genome, is modified resulting in a change to the protein to include or exclude alternate coding exons, use alternate splice donor or acceptor sites or modify the N-terminal length by using a different in-frame start codon. For example, see NM_000031.5.

REFSEQ POLICIES

Species included

Species are considered for inclusion in the RefSeq collection based on a combination of factors including the availability and quality of a whole genome assembly, the number of available cDNA submissions to INSDC, medical relevance and the availability of additional support data from the research community. The RefSeq project is supported by several different process flows and a species may be selected to proceed via a particular process flow depending on the type and abundance of available data as well as the availability of community-maintained annotation. In general, the RefSeq project includes representation of genomes for species that have any of these features:

(1) Model organism.
(2) Causative agent of or otherwise associated with disease.
(3) Eukaryotic genomes that are considered to be a reasonably complete representation of the whole genome with a sequence project goal of >6× sequence depth.
(4) Prokaryotic genomes that are considered to be either a complete genome sequence or a whole genome shotgun sequence (WGS) assembly. WGS assemblies that represent strain-specific differences are included in the RefSeq collection.
(5) Complete organelles and plastid genomes.
(6) Targeted sequence regions that support specific reporting or identification needs; for example, the RefSeqGene project (below) or other gene-specific benchmarks that are used for identification purposes.

Genome annotation decisions

It is important to note that RefSeq records are maintained by NCBI whereas the primary sequence records available in the international nucleotide sequence databases (INSD) are maintained and updated by the submitters of those records. Thus, annotation presented on a RefSeq genomic record may differ from the original submission for some species or loci. Annotation is provided on genome assemblies represented in RefSeq by several different process flows including:

(1) Collaboration.
(2) Annotation by NCBI.
(3) Propagation of annotation from GenBank with targeted curation.

We collaborate with model organism database groups that have the resources and interest in maintaining and submitting genome annotation updates at some frequency and represent their curated annotation in RefSeq. The model organism group is considered the primary authority for the annotated RefSeq genome. For some species, including *Saccharomyces cerevisiae* and *Caenhorhabditis elegans*, annotation updates are submitted directly to RefSeq. For other species, including *Drosophila melanogaster* and *Arabidopsis thaliana*, annotation updates are submitted directly to the INSDC and propagated to RefSeq.

NCBI's computational genome annotation pipelines are used to annotate some prokaryotic and eukaryotic genomes. Genome annotation is calculated by NCBI upon request from a genome sequencing project, for genomes that are submitted without comprehensive whole genome annotation (and the submitting group does not plan to provide annotation) and for genomes that are considered to be of critical importance to continuing medical and basic research. For more information, see the Genome Annotation chapter of the NCBI Handbook (http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.chapter.ch14). A prokaryotic annotation pipeline has been developed to support annotation collaborations with interested sequencing centres wherein the NCBI annotation pipeline is used to calculate the genome annotation submitted to the INSDC and propagated to RefSeq. Annotation of the larger eukaryotic genomes is more complex and additional considerations are given to genomes that are submitted to the INSDC unannotated but with intent to submit annotation at a future date. If annotation is not submitted in a reasonable period of time (~12 months), then NCBI may proceed to calculate annotation for the RefSeq genome records as this provides valuable information to the research community and puts the RefSeq genome into the queue for periodic updates and ongoing annotation maintenance. Note that the eukaryotic genomes annotated by NCBI are updated periodically and RefSeqs for RNAs and proteins may be updated independent of the annotated genome as new transcript data are submitted to the archival INSDC.

Most genome submissions do include annotation information, and this is frequently propagated to the RefSeq representation for the genome. The RefSeq record may differ from the original INSD record in small details to conform to RefSeq feature annotation standards. In these cases, NCBI staff may curate individual loci based on requests from the scientific community or based on on-going curation efforts that are oriented on protein homology approaches (see Targeted annotation for microbial and organelle genomes section).

RELATED INITIATIVES

CCDS

Collaborators for the Consensus CDS (CCDS) project have made a significant focus on coordinating curation of the human proteome in the past 2 years. The last annotation update for the human genome, NCBI build 36.3, resulted in 20 159 tracked CCDS IDs (for over 17 000 genes). A significant outcome of the CCDS collaboration is convergence towards common curation criteria for coding sequence (CDS) annotation by all curators. Comments (displayed as 'Public Note') and an expanded history report are now provided; public notes explain the curation logic applied when modifying, or removing, a CCDS ID. For example, see CCDS7726.2 (http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi).

RefSeqGene

RefSeqGene is a subset of NCBI's RefSeq project. RefSeqGene records provide a stable standard genomic sequence for reporting sequence variants, especially those of clinical relevance. As reported by Gulley *et al.* (9), reference sequence standards are needed to support clear communication about the position of variation that is biologically significant, without uncertainty that may result from identification of the position of the initiation codon or splice sites. RefSeq mRNA sequences are being used as a reporting standard, but they have the obvious limitation of not providing explicit coordinates for flanking or intronic sequence. RefSeq chromosome sequences are also used, but the coordinate system is unappealingly large and, perhaps more importantly, not as stable because the human reference sequence continues to be updated. When requested by an authoritative group for a gene, RefSeqGene sequences are constructed to differ from that of the reference genome to represent a specific allele. The default range of a RefSeqGene sequence begins 5 kb upstream of the first exon and extends 2 kb downstream of the last exon, but curators can modify this upon request. RefSeqGene records can be retrieved from Entrez nucleotide using the query 'refseqgene[keyword]' and are provided for ftp as weekly updates (see ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/RefSeqGene/). The RefSeqGene web site (http://www.ncbi.nlm.nih.gov/projects/RefSeq/RSG/) provides a report of the >650 genes currently included, and the RefSeqGene accession version, links to view the record in graphical or traditional flat file format, and links to access more information in Entrez Gene or OMIM. Records are considered curated and stable; provision of RefSeqGene records is done in consultation with locus-specific databases or other experts as needed.

Targeted annotation for microbial and organelle genomes

During the last 2 years, NCBI has been working on automatic methods to improve the quality of microbial and organellar genome annotation. Curated clusters from the Protein Clusters database (10) are used to provide updated and consistent functional annotation to RefSeq genomic records. Structural RNAs are added by using tRNAscanSE for tRNAs, Infernal and covariance models for 5S rRNAs and an internal database for 16S and 23S rRNAs (11–13). Existing annotations are checked by comparison to this tool and the results have been used to correct a number of annotations on numerous RefSeq records including additions (751 rRNAs; 3679 tRNAs), deletions (4 rRNAs, 47 tRNAs) and strand corrections (103 rRNAs, 115 tRNAs). NCBI has also developed a genome validation tool for assessing the quality of functional annotation (incorrectly annotated RNAs as above, overlapping features and potential frameshifts) with both a web component (http://www.ncbi.nlm.nih.gov/genomes/frameshifts/frameshifts.cgi) as well as standalone package for local installation (ftp://ftp.ncbi.nih.gov/genomes/TOOLS/subcheck/). This tool is used in-house for curation of RefSeq records as well as by various submitters of prokaryotic genomic records to improve annotations prior to submission.

FUTURE DIRECTIONS

The RefSeq collection is expected to grow as more genomes are sequenced. For human, collaborations among CCDS, HUGO Gene Nomenclature Committee (HGNC), Swiss-Prot, the ENCODE project, locus-specific databases and others will continue to improve the coverage and accuracy of RefSeq sequence and annotation. For microbial genomes, a major initiative to curate bacterial protein annotation across the group of related proteins (10) is expected to result in a significant increase in the number of curated bacterial proteins. In addition, improvements to the RefSeq release processing pipeline and quality assurance testing provided are in progress.

FUNDING

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

Conflict of interest statement. None declared.

REFERENCES

1. Schuler,G.D., Epstein,J.A., Ohkawa,H. and Kans,J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol.*, **266**, 141–162.
2. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
3. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., Dicuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, in press.
5. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2007) Entrez gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **35**, D26–D31.
6. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell, J. and Sayers,E.W. (2009) GenBank. *Nucleic Acids Res.*, in press.
7. Maquat,L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.*, **5**, 89–99
8. Kapustin,Y., Souvorov,A., Tatusova,T. and Lipman,D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct.*, **21**, 20.
9. Gulley,M.L., Braziel,R.M., Halling,K.C., Hsi,E.D., Kant,J.A., Nikiforova,M.N., Nowak,J.A., Ogino,S., Oliveira,A., Polesky,H.F. *et al.* (2007) Clinical laboratory reports in molecular pathology. *Arch. Pathol. Lab Med.*, **131**, 852–863.
10. 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.*, in press.
11. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
12. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
13. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.