

18. The Reference Sequence (RefSeq) Project

Kim Pruitt, Tatiana Tatusova, and James Ostell

Created: October 9, 2002
Updated: August 13, 2003

Summary

The Reference Sequence (RefSeq) database provides a biologically non-redundant collection of DNA, RNA, and protein sequences. Each RefSeq represents a single, naturally occurring molecule from a particular organism. RefSeqs are frequently based on GenBank records but differ in that each RefSeq is a synthesis of information, not a piece of a primary research data in itself. Similar to a review article in the literature, a RefSeq is an interpretation by a particular group at a particular time. RefSeqs can be retrieved in several different ways: by searching the Entrez Nucleotide or Protein database, by BLAST searching, by FTP, or through links from other NCBI resources.

Introduction

The goal of the NCBI RefSeq [<http://www.ncbi.nih.gov/RefSeq/>] project is to provide the best non-redundant and comprehensive collection of naturally occurring DNA, RNA, and protein molecules for major organisms. The collection explicitly links the nucleotide and protein sequences. Ideally, all molecule types will be available for each well-studied organism, but the current database collection pragmatically includes those molecules and organisms that are most readily identified, either by collaborators or NCBI processing of public sequence data. Thus, different amounts of information are available for different organisms at any given time. Intermediate records are provided for some organisms when the genome sequence is not yet finished.

As a non-redundant collection of sequences, RefSeq offers a significant advantage during database searches or when identifying sequences (whether by BLAST, text, or accession queries, or inclusion in a local custom database). RefSeq represents an objective and experimentally verifiable definition of non-redundancy by providing one example of each natural biological molecule per organism. For some organisms, the RefSeq collection includes alternatively spliced transcripts that share some identical exons, or identical proteins expressed from these alternatively spliced transcripts, or close paralogs or homologs. RefSeq provides the substrate for a variety of objective conclusions about non-redundancy based on clustering identical sequences or families of related sequences.

RefSeq is unique in providing a large, multi-species, curated sequence database that explicitly links chromosome, transcript, and protein information; it establishes a baseline for integrating sequence, genetic, expression, and functional information into a single, consistent framework. RefSeq is substantially based on GenBank sequence records (Chapter 1). Hereafter, the shorter term “GenBank” is used to indicate the full set of archival sequence data that is submitted to, and redistributed by, the three collaborating databases DDBJ, EMBL, and GenBank. RefSeq records

include attribution to the original sequence data; however, RefSeq differs from GenBank in the same way that a review article differs from the relevant collection of primary research articles on the same subject. RefSeq represents a synthesis and summary of information by a person or group based on the primary information that was gathered by others. Other organizing principles or standards of judgment are possible, which is why such a work should be attributed to the synthesizing "editors". RefSeq does not exclude other syntheses based on the same primary information. But, similar to a review article, it allows the comparison of many different observations taken over time. RefSeq also has the advantage of organizing a large body of diverse data into a single consistent framework with a uniform set of conventions and standards.

Note that although based upon GenBank, RefSeq is distinct from GenBank. GenBank represents the sequence and annotations that are supplied by the original authors and is never changed by others. GenBank remains the primary sequence repository. RefSeq is one of many possible "review articles" based on that essential archive.

The RefSeq collection establishes a consistent baseline and clear model of the central dogma. RefSeq standards support genome annotation, gene characterization, mutation analysis, expression studies, and polymorphism discovery. The RefSeq collection offers advantages in:

- Facile identification of sequence standards for genomes, transcripts, or proteins
- Genome annotation
- Comparative genomics
- Reduction of redundancy in clustering approaches
- Providing a foundation for unambiguous association of functional information (supports navigation)

Database Content: Background

The current RefSeq collection includes sequences from over 2,000 distinct taxonomic identifiers, ranging from viruses to bacteria to eukaryotes. It represents chromosomes, organelles, plasmids, transcripts, and over 700,000 proteins. Every sequence is assigned a stable Accession number, version number, and gi number; older versions remain available if a sequence is updated over time. Table 1 indicates the types of sequence molecules and the corresponding RefSeq Accession number formats. Also see the RefSeq [<http://www.ncbi.nih.gov/RefSeq/>] Web site.

Table 1. The RefSeq Accession number format and molecule types.

Accession prefix	Molecule type
NC_	Complete genomic molecule
NG_	Genomic region
NM_	mRNA
NP_	Protein

Accession prefix	Molecule type
NR_	RNA
NT_ ^a	Genomic contig
NW_ ^a	Genomic contig (WGS ^b)
XM_ ^a	mRNA
XP_ ^a	Protein
XR_ ^a	RNA
NZ_ ^c	Genomic (WGS)
ZP_ ^a	Protein, on NZ_

^a Computed.

^b Assembly of Whole Genome Shotgun (WGS) sequence data.

^c An ordered collection of WGS for a genome.

Updates

RefSeq updates are provided on a daily basis, as needed. New records may be added to the collection, or existing records may be updated to reflect sequence or annotation changes or as part of a bulk update from a collaborator. New and updated records are made available in Entrez as soon as possible. The FTP site also provides daily update information (see below).

Flat File Format and Annotated Features

RefSeq records appear similar in format to the GenBank records from which they are derived; distinguishing features include a unique accession prefix that includes an underscore, which is never present in a GenBank accession (Table 1), and a COMMENT field that indicates the RefSeq status and the source of the sequence information (Figure 1). Some RefSeq records include feature annotation that is not present on the underlying GenBank record; new annotation is provided by computation as well as manual curation. For example, nucleotide variation features are computed using the data available in the dbSNP database (Chapter 5), and protein domains are computed using NCBI's Conserved Domain Database (Chapter 3). Further nucleotide and protein features, publications, and comments may be added by collaborating groups or NCBI staff (see Box 2).

LOCUS	GCNT2	4691 bp	mRNA	linear	PRI 23-APR-2003
DEFINITION	Homo sapiens glucosaminyl (N-acetyl) transferase 2, I-branching enzyme (GCNT2), transcript variant 2, mRNA				
ACCESSION	NM_001491				
VERSION	NM_001491.2	GI:30061504			
KEYWORDS	.				
SOURCE	Homo sapiens (human)				
ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.				
REFERENCE	1 (bases 1 to 4691)				
AUTHORS	Inaba,N., Hiruma,T., Togayachi,A., Iwasaki,H., Wang,X.H., Furukawa,Y., Sumi,R., Kudo,T., Fujimura,K., Iwai,T., Gotoh,M., Nakamura,M. and Narimatsu,H.				
TITLE	A novel I-branching beta-1,6-N-acetylglucosaminyltransferase involved in human blood group I antigen expression				
JOURNAL	Blood 101 (7), 2870-2876 (2003)				
MEDLINE	22528541				
PUBMED	12468428				
REFERENCE	2 (bases 1 to 4691)				
AUTHORS	Potter,K.N., Hobby,P., Klijn,S., Stevenson,F.K. and Sutton,B.J.				
TITLE	Evidence for involvement of a hydrophobic patch in framework region 1 of human V4-34-encoded Igs in recognition of the red blood cell I antigen				
JOURNAL	J. Immunol. 169 (7), 3777-3782 (2002)				
MEDLINE	22229534				
PUBMED	12244172				
RE MARK	GeneRIF: The I carbohydrate antigen interacts simultaneously with the entire hydrophobic patch in framework region 1 and with the outside surface of Ig heavy chain complementarity-determining region 3, leaving most of the site available for binding other antigens.				
REFERENCE	3 (bases 1 to 4691)				
AUTHORS	Yu,L.C., Twu,Y.C., Chang,C.Y. and Lin,M.				
TITLE	Molecular basis of the adult i phenotype and the gene responsible for the expression of the human blood group I antigen				
JOURNAL	Blood 98 (13), 3840-3845 (2001)				
MEDLINE	21599759				
PUBMED	11739194				
REFERENCE	4 (bases 1 to 4691)				
AUTHORS	Yeh,J.C., Ong,E. and Fukuda,M.				
TITLE	Molecular cloning and expression of a novel beta-1,6-N-acetylglucosaminyltransferase that forms core 2, core 4, and 1 branches				
JOURNAL	J. Biol. Chem. 274 (5), 3215-3221 (1999)				
MEDLINE	99115671				
PUBMED	9915862				
REFERENCE	5 (bases 1 to 4691)				
AUTHORS	Bierhuizen,M.F., Maemura,K., Kudo,S. and Fukuda,M.				
TITLE	Genomic organization of core 2 and I branching beta-1,6-N-acetylglucosaminyltransferases. Implication for evolution of the beta-1,6-N-acetylglucosaminyltransferase gene family				
JOURNAL	Glycobiology 5 (4), 417-425 (1995)				
MEDLINE	96078409				
PUBMED	7579796				
REFERENCE	6 (bases 1 to 4691)				
AUTHORS	Bierhuizen,M.F., Mattei,M.G. and Fukuda,M.				
TITLE	Expression of the developmental I antigen by a cloned human cDNA encoding a member of a beta-1,6-N-acetylglucosaminyltransferase				

Figure 1, continued.

COMMENT	gene family Genes Dev. 7 (3), 468-478 (1993)	Status and sequence data source
JOURNAL	Genes Dev. 7 (3), 468-478 (1993)	
MEDLINE	93194065	
PUBMED	8449405	
COMMENT	REVIEWED REFSEQ : This record has been curated by NCBI staff. The reference sequence was derived from L41607.1 and AL832719.1 . On Apr 23, 2003 this sequence version replaced gi: 4503962 .	History
COMMENT	Summary: The enzyme encoded by this gene is responsible for the formation of the blood group I antigen. The i and I antigens are determined by linear and branched poly-N-acetyllactosaminoglycans, respectively. During embryonic development in human erythrocytes, the fetal i antigen is replaced by the adult I antigen as the result of the appearance of a beta-1,6-N-acetylglucosaminyltransferase, the I-branching enzyme. This gene encodes the I-branching enzyme that converts the linear form into the branched form. Defects in this gene have been associated with adult i blood group phenotype. Several transcript variants encoding different isoforms have been described.	Gene Summary
COMMENT	Transcript Variant: This variant (2) contains a different 5' end exon compared to variants 1 and 3, resulting in an isoform (B) with a different N-terminus compared to isoforms A and C. COMPLETENESS: full length.	Transcript variant description
FEATURES	Location/Qualifiers	
source	1..4691 /organism="Homo sapiens" /mol_type="mRNA" /db_xref="taxon:9606" /chromosome="6" /map="6p24"	
gnc	1..4691 /gene="GCNT2" /note="synonyms: IGNT, AIGnT, BIGnT, CIGnT, NACGT1, NAGCT1" /db_xref="LocusID:2651" /db_xref="MIM:600429"	Links to more information
CDS	709..1911 /gene="GCNT2" /EC_number="2.4.1.150" /note="isoform B is encoded by I-branching enzyme; blood group Ii; N-acetyllactosaminide beta-1,6-N-acetylglucosaminyltransferase" /codon_start=1 /product="I beta-1,6-N-acetylglucosaminyltransferase isoform B" /protein_id="NP_001482.1" /db_xref="GI:4503963" /db_xref="LocusID:2651" /db_xref="MIM:600429"	Links to more information Link to protein

Figure 1: Features of a RefSeq record. *Dialog balloons* indicate distinguishing features including the Accession number format, the COMMENT text indicating the status and source of the sequence information, the (optional) gene description (Summary text), the (optional) description of transcript variants, and links to additional information. Links may be provided to external sources of information (e.g., OMIM and the Expasy Enzyme Commission Web site) as well as to other NCBI pages (e.g., the protein record and LocusLink), as seen in this example.

Assembling and Maintaining the RefSeq Collection

Summary

The RefSeq database is compiled through several processes including collaboration, extraction from GenBank, and computation. Each molecule is annotated as accurately as possible with the correct organism name, correct gene symbol for that organism, and informative protein names whenever possible. Collaborations with authoritative groups outside of NCBI provide a variety of information ranging from curated sequence data, nomenclature, feature annotations, and links to external organism-specific resources. If a collaboration has not been established, then NCBI staff assembles the data from GenBank. Each record has a tag indicating the level of curation it has received (Table 2), and the collaborating group is attributed. Thus, the RefSeq record may be an essentially unchanged validated copy of the original GenBank record, or it may include corrected or additional information that has been added by collaborators or experts at NCBI.

Table 2. RefSeq status codes.

Code	Description
GENOME ANNOTATION	The RefSeq record is provided via automated processing and is not subject to individual review or revision between builds.
INFERRED	The RefSeq record has been predicted by genome sequence analysis, but it is not yet supported by experimental evidence. The record may be partially supported by homology data.
PREDICTED	The RefSeq record has not yet been subject to individual review, and some aspect of the RefSeq record is predicted.
PROVISIONAL	The RefSeq record has not yet been subject to individual review. The initial sequence-to-gene name associations have been established by outside collaborators or NCBI staff.
REVIEWED	The RefSeq record has been reviewed by NCBI staff or by a collaborator. The NCBI review process includes assessing available sequence data and the literature. Some RefSeq records may incorporate expanded sequence and annotation information.
VALIDATED	The RefSeq record has undergone an initial review to provide the preferred sequence standard. The record has not yet been subject to final review, at which time additional functional information may be provided.
WGS	The RefSeq record is provided to represent a collection of whole genome shotgun sequences. These records are not subject to individual review or revisions between genome updates.

In cases when a molecule is represented by multiple sequences for an organism in GenBank, an effort is made by NCBI staff to select the “best” sequence to instantiate as a RefSeq. The goal is to avoid known mutations, sequencing errors, cloning artifacts, and erroneous annotation; should an existing RefSeq be identified with a problem of this type, it is corrected. Sequences are validated to confirm that the genomic sequence corresponding to an annotated mRNA feature matches the mRNA sequence record, and that coding region features really can be translated into the corresponding protein sequence.

RefSeq records may be added to the collection, or existing records may be updated, on a daily basis. Separate working groups that use distinct process pipelines compile the RefSeq collection for different organisms. RefSeq records are provided by collaboration and the following three pipelines:

- Genome Annotation pipeline
- LocusLink pipeline
- Entrez Genomes pipeline

Collaboration

We welcome collaborations whenever authoritative groups outside NCBI are willing to provide sequences, nomenclature, annotations, or links to phenotypic or organism-specific resources. For some species, the RefSeq collection is curated entirely by a collaborating authoritative group that provides both the sequences and annotations. Other species may be provided via varying levels of collaborative efforts. For example, a Viral Genome Advisory group [<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viradvisors.html>] has been established to support curation of the viral RefSeq collection. For some species, some information is provided by the external authoritative source, and some information is provided by NCBI. This process may be automated in that data are periodically downloaded and subject to validation to detect errors and apply annotation in a more uniform way; NCBI does not otherwise carry out additional curation to add annotation or make sequence changes to records supplied by collaborators. As the collaborating group supplies updates, changes are reflected in the RefSeq collection for that organism. Collaborator-supplied records have a REVIEWED status, and the collaborating group is identified. Table 3 lists examples of curated and annotated sequence records provided wholly or in part through collaboration.

Table 3. Selected examples of collaborator-contributed RefSeq records.

Organism	Collaborator
<i>Saccharomyces cerevisiae</i>	Saccharomyces Genome Database (SGD)
<i>Arabidopsis thaliana</i>	The Institute for Genomic Research (TIGR)
<i>Pseudomonas aeruginosa</i>	<i>Pseudomonas aeruginosa</i> Community Annotation Project (PseudoCAP)
<i>Drosophila melanogaster</i>	Drosophila Sequencing Consortium, FlyBase, and NCBI staff

Genome Annotation Pipeline

NCBI is providing annotation of genomic sequence data for some genomes including some microbial species, human, and mouse. These pipelines are automated and yield genomic, transcript, and protein RefSeq records (records that are provided vary by organism). Data are refreshed periodically, and for the eukaryotic annotation pipeline, records are not subject to individual incremental updates or manual curation (see Table 1; see Chapter 14 for more information on the eukaryotic genome annotation pipeline).

LocusLink Pipeline

LocusLink supports the generation of RefSeq records for human, mouse, rat, fly, zebrafish, and cow; the RefSeq process for these organisms takes advantage of the descriptive information available in LocusLink. Multiple collaborations support the collection of this descriptive information (Box 1; see also Chapter 19). The *Drosophila* RefSeq collection is provided in collaboration with FlyBase, and numerous additional collaborations have occurred for individual genes and gene families, primarily for the human RefSeq collection.

This data set consists of genomic regions, transcripts, and protein. Records representing genomic regions are provided primarily to support more comprehensive genome-level annotation and may represent gene clusters or single genes or pseudogenes. Records are annotated with the level of curation it has undergone; records may have an INFERRED, PREDICTED, PROVISIONAL, REVIEWED, or VALIDATED status (Table 2).

Sequences in LocusLink records enter RefSeq by a mixture of computational analysis process, collaboration, and in-house curation. As illustrated in Figure 2, generation of the initial RefSeq record is dependent on first identifying a representative sequence for a gene in the LocusLink database. New genes are identified and added to the collection by collaborators or in-house processing that mines information available in UniGene, the Genome Annotation pipeline, and new GenBank submissions (Box 1). Once sequence data are associated with a LocusID, it is used as a query sequence for automated BLAST analysis. The longest mRNA that meets our stringent matching criteria is taken from the BLAST results; this is either a cDNA sequence or a mRNA feature annotated on a DNA sequence record. The identified sequence is subsequently used to provide the initial RefSeq record. This stage of the process includes detection of conflicts and problems including:

- Sequence-to-locus association conflicts (e.g., close paralogs)
- Vector contamination
- The GenBank record used is a genomic record with “not experimental” annotation
- The protein sequence is annotated as partial
- The RefSeq transcript would be suspiciously long

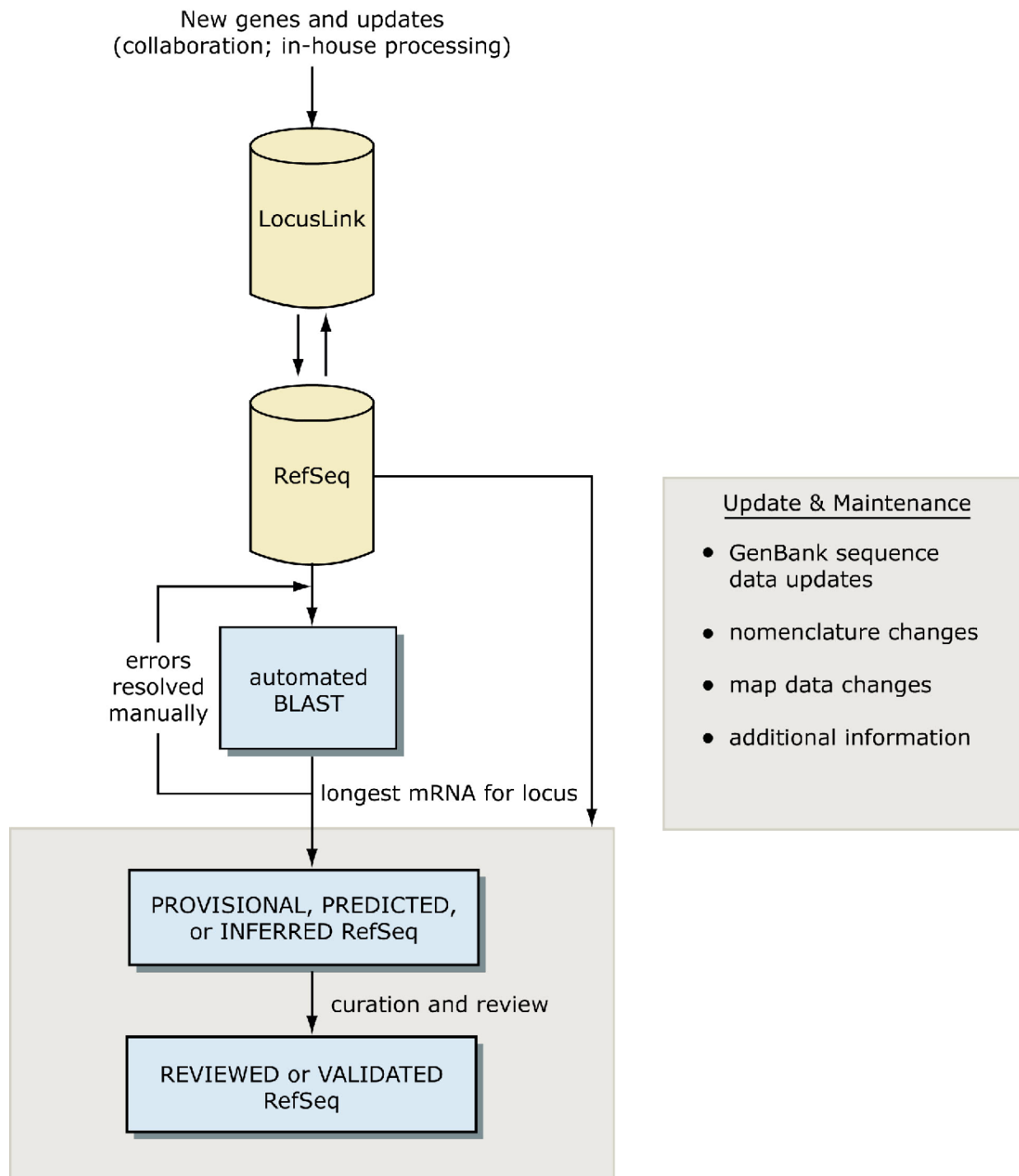



Figure 2: LocusLink-supported RefSeq pipeline. Once a gene is defined and associated with sufficient sequence information in LocusLink, it can be pushed into the RefSeq pipeline. New genes are added to LocusLink by collaborators and in-house review. The RefSeq process is initiated by an automated BLAST step, which uses the sequence information in LocusLink as a query against GenBank to identify the longest mRNA for each locus. This sequence is represented as a provisional, predicted, or inferred RefSeq record. Subsequent review and curation may result in a sequence or annotation update (as described in Box 2). Records are refreshed if the underlying GenBank Accession number is updated or if an official nomenclature or cytogenetic map location is updated in LocusLink.

Conflicts and problems of these types must be resolved before the RefSeq can become public. Records are subject to validation to correct annotation errors and to provide annotation in a more consistent format. LocusLink descriptive information including official nomenclature, alias symbols, alternate descriptive names, map location, and additional citations are applied to the records. Records at this stage have a PROVISIONAL, PREDICTED, or INFERRED status.

These initial records then enter into the in-house review pipeline, where additional manual curation may be applied. The review process prioritizes known genes, gene families, and problem cases that are identified through user input or analysis. The curation process includes analysis of a suite of precomputed BLAST results, literature review, review of available database and Web resources (both in-house and external), and collaboration to curate the nucleotide and protein sequence data and to apply additional annotation and descriptive information to both the RefSeq sequence record and to the LocusLink database. Box 2 lists more detailed information concerning the type of errors corrected and information added by the manual curation process. Sequence review is carried out primarily by NCBI staff, but some sequences and annotations are provided by collaboration including a portion of the *Drosophila melanogaster* RefSeq collection. The curation process also provides additional sequence records to represent splice variants when sufficient information about their full-length composition is available. Records that have undergone manual curation, either by NCBI staff or a collaborating group, have a VALIDATED or REVIEWED status. Note that for many genes, intermediate levels of manual curation occur to correct sequence mis-associations, to use a more optimal GenBank record, and to provide additional data to LocusLink before full review of a RefSeq sequence record.

Additional ongoing review is applied to identified problem sets; for example, periodic analysis may be carried out to identify sequences that include repeats, that have poor-quality splice sites, are very short or very long, or that are extremely similar to sequences associated with a different LocusID. Review of problem sets may result in discontinuing a RefSeq, LocusID, or both. A RefSeq is suppressed if it is found to represent a transcribed repeat element, to be derived from the wrong organism (i.e., the GenBank sequence it was based on does not have accurate organism annotation), or to not represent a “gene”. An Entrez query will still retrieve a suppressed record, with a disclaimer appearing on the query result document summary (Figure 3a), but the record is not included in the BLAST databases nor in the calculation of related sequences or the BLink display (precomputed protein BLAST results). If a RefSeq is found to be redundant with another public RefSeq, then one is retained and the other becomes a secondary Accession number (Figure 3b). If the sequences were associated with two different LocusIDs, then the LocusIDs are merged so that in LocusLink, a query with either of the original LocusIDs will retrieve the remaining single record.

(a)
 1: NM_032931

ref[NM_032931.1][[14249727]
 This record was removed at the submitters request. 

(b)

```


LOCUS          NM_002729                1821 bp      mRNA   linear   PRI 10-FEB-2002
DEFINITION    Homo sapiens hematopoietically expressed homeobox (HHEX), mRNA.
ACCESSION    NM_002729 NM_001529 
VERSION      NM_002729.2  GI:17978473
KEYWORDS     .
SOURCE       human.
ORGANISM     Homo sapiens
  
```

Figure 3: How to recognize suppressed and redundant RefSeq records. (a) A standard text statement is included on the Entrez document summary for suppressed RefSeq records (*red arrow*). (b) If redundant RefSeq records are merged, then both Accession numbers appear on the flat file ACCESSION line (*green arrow*). The first Accession number listed is the primary identifier, and all other listed accessions are “secondary” Accession numbers.

Input from the research community is welcome to further improve the quality of this data set. Interested parties are invited to contact us by sending an email to the NCBI Help Desk (info@ncbi.nlm.nih.gov). See also the RefSeq Web site at <http://www.ncbi.nlm.nih.gov/RefSeq/>.

Entrez Genomes Pipeline

The Entrez Genomes database represents a collection of complete, or nearly complete, genomes and chromosomes. It is divided into six large taxonomic groups: Archaea, Eubacteria, Eukaryotae, Viroids, Viruses, and Plasmids. See the Prominent Organisms [<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.html>] page for a complete list of organisms included in this database; this list is automatically updated daily. Entrez Genomes RefSeq records include genomic, transcript, and protein records and are provided by collaboration, in-house automatic processing, and in-house curation. The Entrez Genomes Web site includes custom displays, analysis, and tools for some genomes (see Table 4).

Table 4. Selected Entrez Genomes Resources.

Web Page	Web Site
Entrez Genomes Home	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome
Prominent Organisms	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/org.html
Microbes	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/micr.html
Viral Genomes	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viruses.html
Organelle Genome Resources	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/organelles.html
Plant Genomes Central	http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/PlantList.html

In general, these RefSeq records undergo an initial automated validation step before being released. The resulting record is a copy of a GenBank entry, but validation may make some corrections and provides more consistent feature annotation. Records provided via collaboration have a status of REVIEWED and are attributed to the collaborating group. Records provided by in-house processing have a PROVISIONAL, VALIDATED, or REVIEWED status.

Entrez Genomes record processing falls into four primary categories: chromosomes, microbial genomes, small complete genomes, and viruses.

Chromosomes

RefSeq records in this category are usually submitted directly to Entrez Genomes as a complete chromosome sequence representing an assembly of individual clones that are themselves available in GenBank. Examples of this type of RefSeq include *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, and *Caenorhabditis elegans*. In addition, RefSeq records may be available for some genomes that are not yet fully complete but for which complete sequence is available for individual chromosomes; for example, a RefSeq chromosome 1 record is provided for *Leishmania major*.

These records are curated by the organism-specific collaborating group and undergo NCBI validation before being released. The validation process checks for logical conflicts in the annotation (which are reported back to the submitting group) and makes small modifications to format the submission as a RefSeq record. This category of RefSeq record is displayed using the NCBI Map Viewer (Chapter 20) in addition to being available in the main Entrez sequence databases. An organism-specific genome BLAST service is provided for these genomic, transcript, and protein records.

Microbial Genomes

Microbial complete genomes are submitted to GenBank, but because of the GenBank/EMBL/DDBJ collaboration agreement, which limits the size to 350 kb, they are made available in GenBank as a series of Accession numbers. RefSeq does not need to adhere to this upper limit, and therefore complete genome sequences are available as RefSeq records in Entrez Genomes.

These records are subject to additional automated validation and computational analysis; the computational analysis results are then manually reviewed, resulting in more complete annotation of the RefSeq record (see Table 5).

Table 5. Selected Examples of REVIEWED Microbial Genomes.

RefSeq	Organism
NC_003450	<i>Corynebacterium glutamicum</i>
NC_001802	<i>Human immunodeficiency virus 1</i>
NC_004193	<i>Oceanobacillus iheyensis</i>
NC_002689	<i>Thermoplasma volcanium</i>
NC_004718	<i>SARS coronavirus</i>

Small Complete Genomes

Smaller complete genomic sequences, including organelles, plasmids, and viruses, are based on single GenBank records. Automatic processing scans GenBank daily for complete genome updates and new submissions; identified records are candidates for a complete genome RefSeq. These records are manually evaluated to make the final decision; if more than one genomic sequence is available for the genome, then only one is selected to become the RefSeq standard. This selection takes into account various factors including the level of annotation, strain information, and community input.

Some of these RefSeq records undergo manual curation and have a REVIEWED status. For example, viral genomes are re-annotated using GeneMarkS in collaboration with Mark Borodovsky's Bioinformatics Group at the Georgia Institute of Technology. Following GeneMarkS [<http://opal.biology.gatech.edu/GeneMark/>] analysis, viral genomes are subject to manual review by NCBI staff. Viral annotation also relies on an established Viral Genome Advisors [<http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/viradvisors.html>] group. For example, the HIV-1 RefSeq was curated by NCBI Staff in collaboration with the authors of the book *Retroviruses* [<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=rv.TOC>], and NCBI curators expanded the mature protein annotation for several viruses, including the poliovirus and hepatitis C, based on observations reported in the literature that have not been included in a GenBank submission.

Access and Retrieval

RefSeq records can be accessed by direct query, BLAST and FTP download, or indirectly through links provided from several NCBI resources. In addition, RefSeqs are included in some computed resources, and therefore links may be found from those pages to individual RefSeq records. For example, the RefSeq collection is included in Clusters of Orthologous Groups of proteins (COG; Chapter 22) analysis and in Conserved Domain Database (CDD; Chapter 3) analysis to identify proteins with similar domain architecture. Some links to RefSeq records are based on LocusLink associations (e.g., links from OMIM; Chapter 7), whereas others are based on sequence similarity. Links to RefSeq records may be found in the following resources:

- BLAST results (Chapter 16)
- BLink (precomputed protein blast results)
- CDD
- dbSNP (Chapter 5)
- Entrez (Chapter 15)
- LocusLink (Chapter 19)
- Map Viewer (Chapter 20)

Table 6. Entrez queries to retrieve sets of RefSeq records.

Query	Accession prefix	RefSeq category retrieved
srcdb_refseq[prop]	NC_, NG_, NT_, NW_, NZ_, NM_, NR_, XM, XR_, NP_, XP_, ZP_	All
srcdb_refseq_known[prop]	NC_, NG_, NM_, NR_, NP_	REVIEWED, PROVISIONAL, PREDICTED, INFERRED, and VALIDATED
srcdb_refseq_reviewed[prop]	NC_, NG_, NM_, NR_, NP_	REVIEWED records
srcdb_refseq_validated[prop]	NC_, NM_, NR_, NP_	VALIDATED records
srcdb_refseq_provisional[prop]	NC_, NG_, NM_, NR_, NP_	PROVISIONAL records
srcdb_refseq_predicted[prop]	NM_, NR_, NP_	PREDICTED records
srcdb_refseq_inferred[prop]	NM_, NR_, NP_	INFERRED records
srcdb_refseq_model[prop] ^a	NT_, NW_, XM_, XR_, XP_, ZP_	Genome annotation model records

^a Retrieves those NT_ and NW_ records that have gene annotation.

LocusLink Query Access

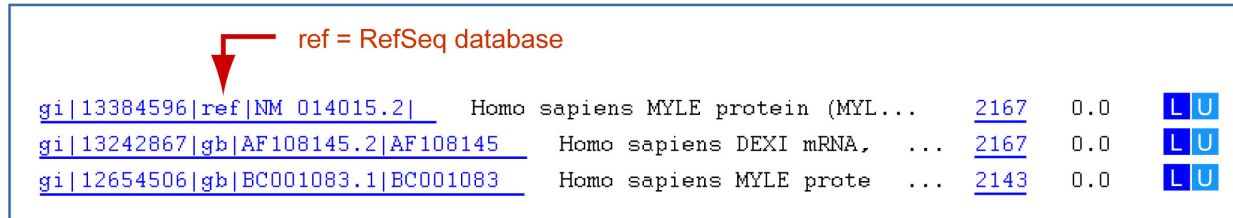
A subset of the RefSeq collection is represented in LocusLink, a gene-centered database, including human, mouse, rat, Drosophila, zebrafish, cow, and HIV-1 (Chapter 19); additional organisms continue to be added to LocusLink and RefSeq over time. LocusLink-supported RefSeq records include a link to the LocusLink gene report page (note the LocusID link on the gene and CDS features in Figure 1). Loci with associated RefSeq records can be retrieved using the controlled query term “has_refseq”. LocusLink query results and gene reports indicate when a RefSeq is available, with links provided to the nucleotide and protein sequences and to related resources, including the Map Viewer and BLink (precomputed protein alignments) and Conserved Domain Database (CDD). The process of RefSeq curation also expands the data available in LocusLink by providing a range of information including:

- Alternate names
- Enzyme Committee numbers
- Gene summaries
- Publications
- Related GenBank accessions
- Transcript variant descriptions

BLAST

RefSeq transcript and protein records are included in the non-redundant nucleotide and protein BLAST databases, and genomic sequences are included in the “chromosome” database; therefore, when a query sequence matches a RefSeq record, the hit is included in the BLAST results

(see Figure 5). Additional organism-specific BLAST pages provide access to specific custom databases, which vary by organism as needed. For example, the human genome BLAST page provides access to query the RefSeq contigs, transcripts, or proteins in addition to (human) sequences in the GenBank High-Throughput Genomic (HTG) division.



gi 13384596 ref NM_014015.2 	Homo sapiens MYLE protein (MYL...	2167	0.0	L U
gi 13242867 gb AF108145.2 AF108145	Homo sapiens DEXI mRNA, ...	2167	0.0	L U
gi 12654506 gb BC001083.1 BC001083	Homo sapiens MYLE prote ...	2143	0.0	L U

Figure 5: RefSeq records are included in NCBI BLAST databases. In a BLAST summary list of results, the abbreviation *ref* identifies records that are provided by the RefSeq collection.

FTP

Currently, RefSeq records generated by different pipelines are available in different FTP areas [RefSeq LocusLink pipeline [<ftp://ftp.ncbi.nih.gov/refseq/>]; Entrez Genomes RefSeqs [<ftp://ftp.ncbi.nih.gov/genomes/>]; Models [<ftp://ftp.ncbi.nih.gov/genomes/>] (Genome Annotation pipeline)]. In the future, the entire collection will be made available in a regular release cycle similar to GenBank releases. The RefSeq release will be available in the “refseq” FTP directory. Separate files are provided for nucleotide and protein records in a variety of formats. See the available README files for additional information.

Related Reading

- Besemer J, Lomsadze A, Borodovski M. 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; 2001.
- Blake JA, Eppig JT, Richardson JE, Davisson MT. The Mouse Genome Database (MGD): expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res* 28:108-111; 2000.
- Boguski MS, Schuler GD. ESTablishing a human transcript map. *Nat Genet* 10:369-371; 1995.
- Coffin JM, Hughes SH, Varmus E. *Retroviruses*. Plainview (NY): Cold Spring Harbor Laboratory Press; 1997.
- FlyBase Consortium. The FlyBase database of the Drosophila Genome Projects and community literature. The FlyBase Consortium. *Nucleic Acids Res* 27:85-88; 1999.
- Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* 15:57-61; 2000.
- Lukashin A, Borodovski M. GeneMark.hmm new solutions for gene finding. *Nucleic Acids Res* 26:1107-1115; 1998.

- Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiesse PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res* 30:281-283; 2002.
- Pruitt KD, Katz KS, Sicotte H, Maglott DR. Introducing RefSeq and LocusLink: curated human genome resources at the NCBI. *Trends Genet* 16:44-47; 2000.
- Tatusova TA, Karsch-Mizrachi I, Ostell JA. Complete genomes in WWW Entrez: data representation and analysis. *Bioinformatics* 15:536-543; 1999.
- Twigger S, Lu J, Shimoyama M, Chen D, Pasko D, Long H, Ginster J, Chen CF, Nigam R, Kwitek A, Eppig J, Maltais L, Maglott D, Schuler G, Jacob H, Tonellato PJ. Rat Genome Database (RGD): mapping disease into the genome. *Nucleic Acids Res* 30:125-128; 2002.
- Westerfield M, Doerry E, Kirkpatrick AE, Douglas SA. Zebrafish informatics and the ZFIN database. *Methods Cell Biol* 60:339-355; 1999.
- White JA, McAlpine PJ, Antonarakis S, Cann H, Eppig JT, Frazer K, Frezal J, Lancet D, Nahmias J, Pearson P, Peters J, Scott A, Scott H, Spurr N, Talbot C Jr, Povey S. Guidelines for human gene nomenclature (1997). HUGO Nomenclature Committee. *Genomics* 45:468-471; 1997.

Box 1: Approaches to associate sequence data with LocusLink loci.

Collaboration with authoritative groups including:

- FlyBase
- Human Gene Nomenclature Committee (HGNC)
- Mouse Genome Informatics (MGI)
- OMIM
- RATMAP
- Rat Genome Database (RGD)
- WormBase
- Zebrafish Information Network (ZFIN)
- Gene Family Authorities

In-House Processing:

- Extraction from GenBank
- Genome Annotation pipeline
- Homology analysis
- In-house curation
- UniGene analysis

Box 2: RefSeq curation: error correction and data addition.

The RefSeq curation process may result in correction of errors as well as provision of additional sequence information and feature annotation, as indicated below.

Curation of Sequences

Error Corrections

1. Remove chimeric sequences.
2. Remove vector and linker sequence.
3. Remove sequences annotated with incorrect organism information.
4. Resolve sequence-to-locus mis-associations.
5. Correct apparent sequence errors as identified through sequence analysis and personal communication; an attempt is made to reconcile sequence differences to finished genomic sequence.
6. Modify the extent of the original GenBank CDS annotation, as determined through sequence analysis (including homology considerations), literature review, and personal communication.

Provide Additional Information

1. Extend UTRs based on publicly available sequence data or literature review.
2. Provide RefSeq with full-length CDS from a series of overlapping partial GenBank sequences.
3. Provide splice variants and corresponding protein variants.

Curation of Annotations (information added)

Nucleotide and Protein

1. Alias symbols and names
2. Brief description of transcript and protein variants
3. Publications
4. Summary of gene, protein function

Protein Only

1. Alternate translation start sites
2. Enzyme Commission number
3. Mature peptide products
4. Non-AUG translation start sites
5. Protein domains
6. Selenocysteine proteins

Nucleotide Only

1. Indication of transcript completeness, as known

2. Poly(A) signal, site
3. RNA editing
4. Variation features