Reference Sequence (RefSeq) Database

Non-redundant genomic, RNA and protein sequence records, from microbes to human

https://www.ncbi.nlm.nih.gov/refseq

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Scope

The Reference Sequence (RefSeq) database is a collection of standard sequences intended to represent genomic sequences, transcripts, and translated products that occur naturally in an organism (www.ncbi.nlm.nih.gov/refseq/). The taxonomic range of the collection spans much of the planet's diversity and includes eukaryotes, bacteria, archaea, and viruses. The table to the right sums up entries available for key taxonomic nodes for release 207. More up-to-date information is available from the release note at ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/.

Taxonomic Node	Release 212 Count
Archaea	1410
Bacteria	68260
Fungi	16581
Invertebrate	5480
Mitochondrion	13144
Plant	8743
Plasmid	5958
Plastid	8997
Protozoa	701
Vertebrate_mammalian	1443
Vertebrate_other	5112
Virus	11620

Automatic processing of public sequence records, collaboration with authoritative scientists or groups outside NCBI, and curation by biological experts at NCBI are at the center of the RefSeq processing pipelines. The product is a collection that is richly annotated with structural and functional features, actively maintained, and flexible enough to provide sequence standards to meet the needs of different research communities. For example, RefSeqGene records support clinical laboratory testing for human disease, curated vertebrate transcript records support genome annotation pipelines and genome browsers, and ribosomal RNA records are provided as type standards for bacterial and archaeal genomes. The RefSeq collection is also a key element unifying several of the molecular resources provided by NCBI.

RefSeq also started to annotated experimentally verified functional elements onto genomic records, with NG_ initialed accession (www.ncbi.nlm.nih.gov/refseq/functionalelements/).

Each sequence record is based on public sequence data submitted to the International Nucleotide Sequence Database Collaboration (INSDC), but in contrast to the public sequence archives, the RefSeq collection is largely non-redundant. The table to the right lists other attributes that distinguish RefSeq from INSDC records.

Attribute	INSDC*	RefSeq
Accession includes underscore ('_')	No	Yes
INSDC member	Yes	No
Source of sequence	Submitter	INSDC
Source of annotation	Submitter	Multiple
Owner of sequence record	Submitter	NCBI
Redundancy	High	Low
Archive of experimental data	Yes	No
Review of experimental data	No	Yes
Curation by NCBI	No	Yes
Regular update schedule	No	Yes
* The INSDC includes NCBI (USA), ENA (Europe), and DDBJ (Japan)		

Access

RefSeq records are available by

- Searching in the nucleotide or protein databases (www.ncbi.nlm.nih.gov/nuccore and www.ncbi.nlm.nih.gov/protein)
- Searching with query sequences against selected databases through BLAST (blast.ncbi.nlm.nih.gov), and
- Downloading the release through FTP (ftp.ncbi.nlm.nih.gov/refseq/ and ftp.ncbi.nlm.nih.gov/genomes/refseq/)

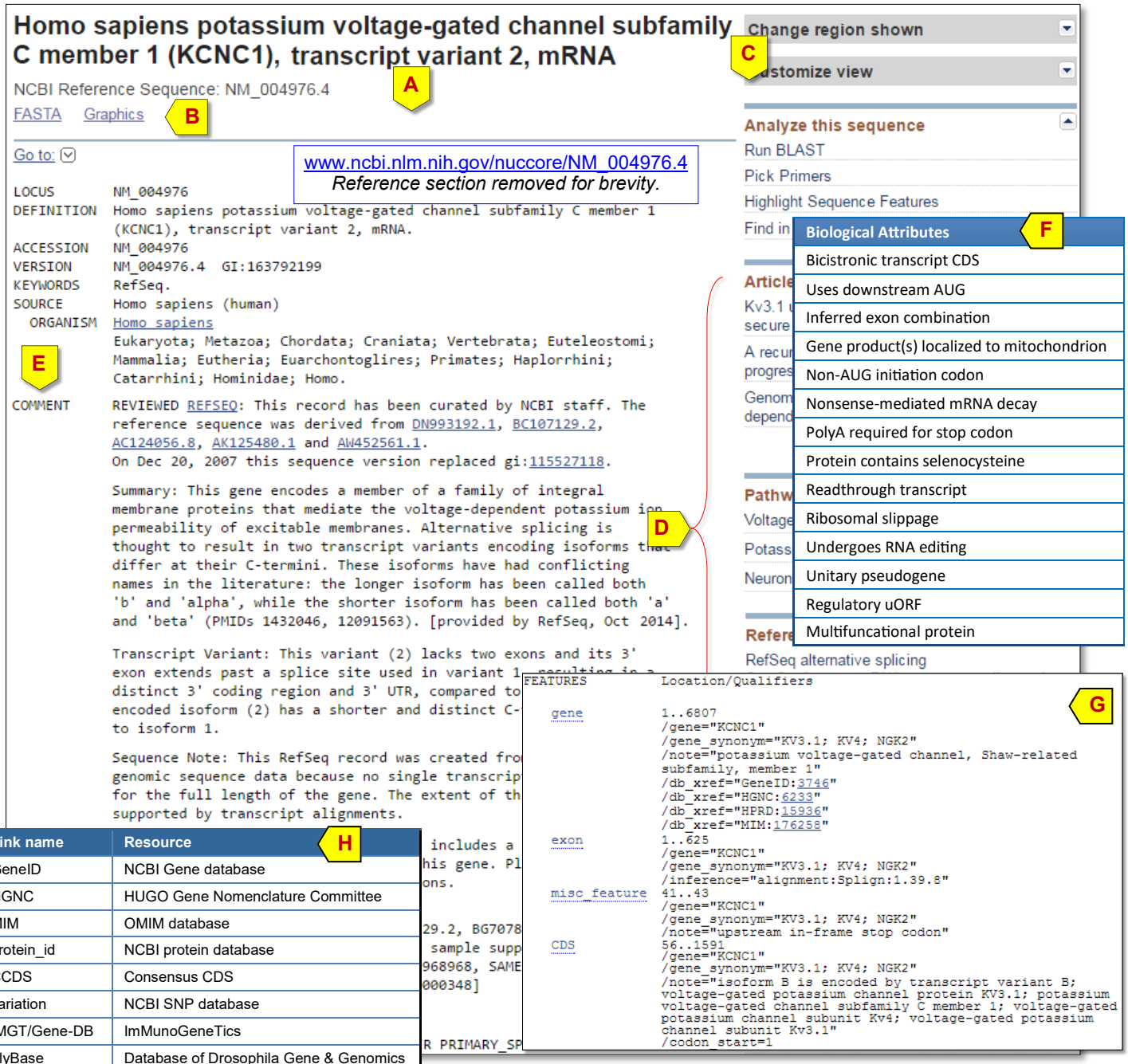
RefSeq records are also available through hyperlinks displayed from many NCBI resources, including Gene, Genome, BioProject, dbSNP, ClinVar, Genome Data Viewer tool, and more. RefSeq filters in the Entrez Nucleotide and Protein databases allows the quick retrieval of relevant RefSeq records from the query result. This filter (refseq[filter]) is also available to the Entrez Programming Utilities (EUtils, www.ncbi.nlm.nih.gov/books/NBK25501/) for programmatic access. RefSeq records are found in the general nucleotide (nt) and non-redundant protein (nr) BLAST databases. RefSeq-only databases are also available from the Database pull-down list, such as:

- Reference RNA sequences (refseq_rna)
- Reference protein sequences (refseq_protein)
- Reference genome sequences (refseq_genomes)
- Reference representative genomes (refseq_representative_genomes, less redundant)
- 16S ribosomal RNA sequences (Bacteria and Archaea)

In addition, RefSeq sub-project and organism-oriented BLAST databases are available from organism-specific BLAST pages, e.g., Human genome, Microbes, as well as the RefSeqGene project page. The complete RefSeq collection, subsets defined by taxonomic node (e.g., plants) or type of molecule, or key model organisms can be downloaded through FTP (ftp.ncbi.nlm.nih.gov/refseq/release/ and ftp.ncbi.nlm.nih.gov/genomes/refseq/).

Record Display

The Nucleotide database displays a retrieved Reference Sequence record in GenBank format by default (**A**). You can access the FASTA and graphical displays using links (**B**) at the top of the page. Use links in the collapsible sections to the right (**C**) to access relevant records from other sources, such as PubMed and Gene (**D**). Refer to the COMMENT section (**E**) for information about the INSDC source data used to derive the record. For some records, this section may also include a review status, a summary of the gene function in the summary paragraph, as well as structured reports of transcript evidence, gene and sequence attributes of biological interest, and more. Refer to the table (**F**) for summary of biological attributes that may be included in the COMMENT section. For more details, see www.pubmed.gov/26553804.



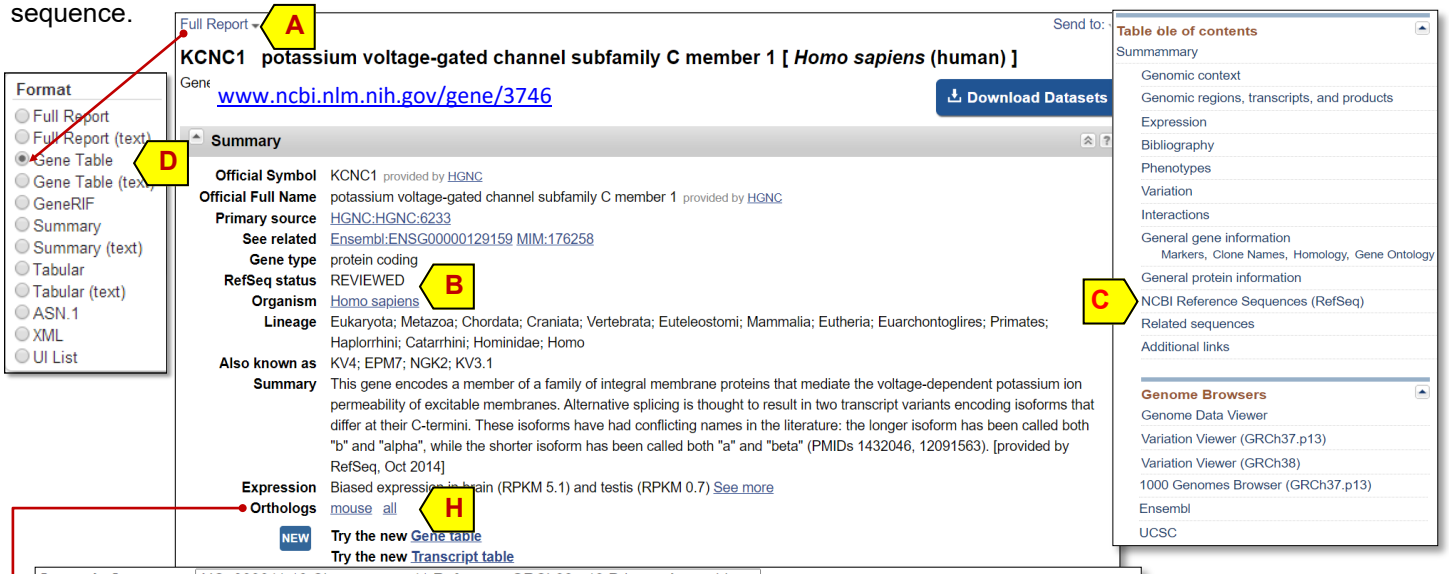
Link name	Resource
GeneID	NCBI Gene database
HGNC	HUGO Gene Nomenclature Committee
MIM	OMIM database
protein_id	NCBI protein database
CCDS	Consensus CDS
variation	NCBI SNP database
IMGT/Gene-DB	ImMunoGeneTics
FlyBase	Database of Drosophila Gene & Genomics
RGD	Rat Genome Database
EcoGene	EcoGene Database
TAIR	The Arabidopsis Information Resource
SGD	Saccharomyces Genome Database
BeeBase	hymenoptera genome database

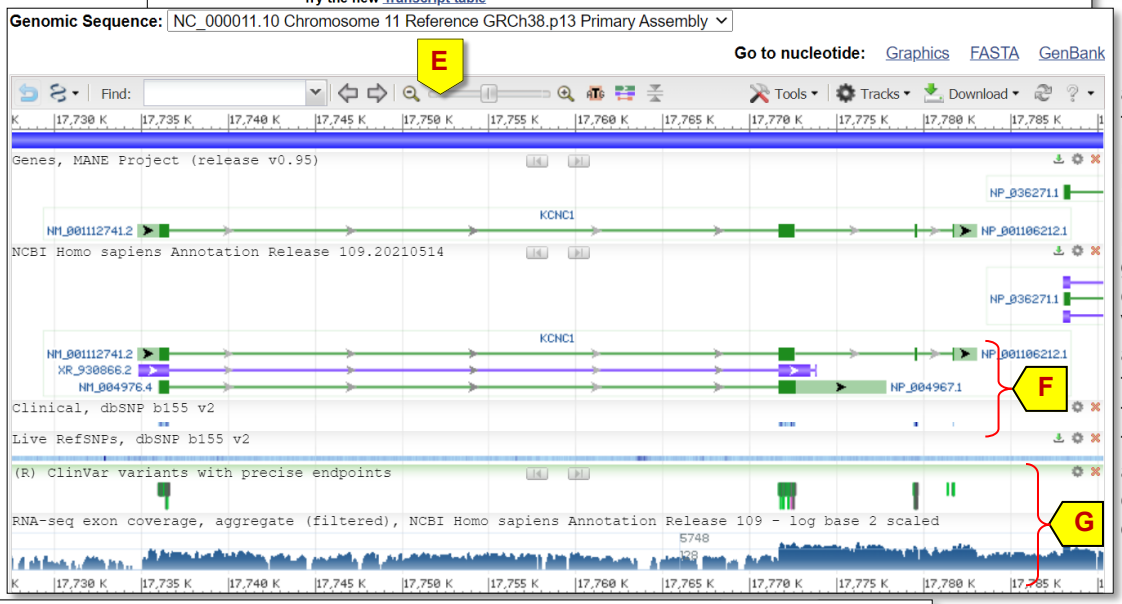
H

The FEATURES section (**G**) contains detailed annotation of exons, polyadenylation signals or sites, miscellaneous features (misc_feature) of the gene, coding sequences (CDS) and protein products. Many features include cross-references to related records in other databases at NCBI or elsewhere. The table (**H**, left) sums up link names and resources they represent. Feature annotated on RefSeq genomic records will be more complex and comprehensive. The functional element records focus on the annotation of regulatory elements based on literature report. More information is at: www.ncbi.nlm.nih.gov/refseq/functionalelements/.

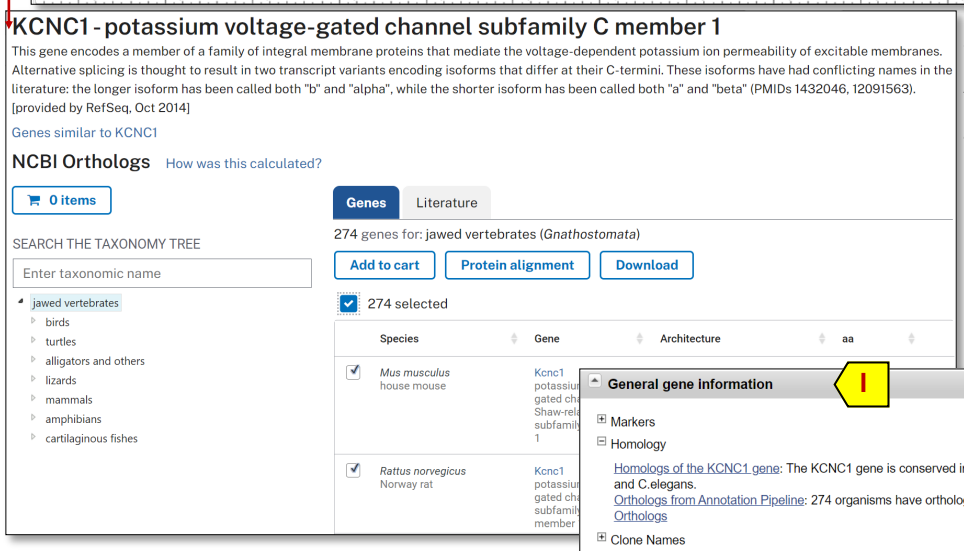
Links to RefSeq Entries from Other Resources

RefSeq records are integral parts of many other NCBI resources. For example, the Gene full report page (**A**) reports the RefSeq curation status (**B**). Click the in-page navigation link “NCBI Reference Sequences (RefSeq)” (**C**) to bring the section in focus and see a detailed summary of individual RefSeq entries for the Gene record. Display the record in “Gene Table” format (**D**) to get details on the exon/intron structure, their genomic coordinates, and links to FASTA sequence.

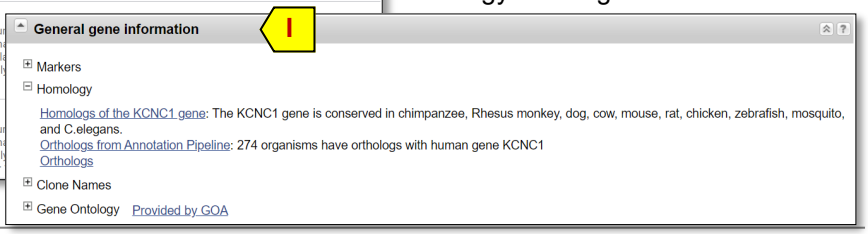




RefSeq mRNA records are an integral part of the genome annotation process. The “Genomic regions, transcripts, and products” section of the report presents this graphically using the embedded Sequence Viewer (SV, **E**), with the alternatively spliced transcripts shown at the top (**F**). Tracks below the transcripts provide additional supporting evidence from RNA-seq data (**G**).

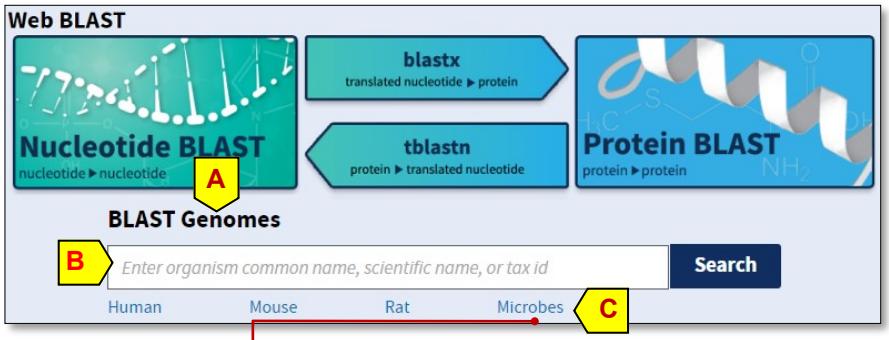


The NCBI Eukaryotic Genome Annotation pipeline analyzes available complete RefSeq proteomes to derive the orthologs gene list, which is available from gene as a link in the Summary section (**H**), which points to a new Datasets powered display. This and a smaller set from HomoloGene are also available in the “General gene information” section (**I**), under the Homology heading.

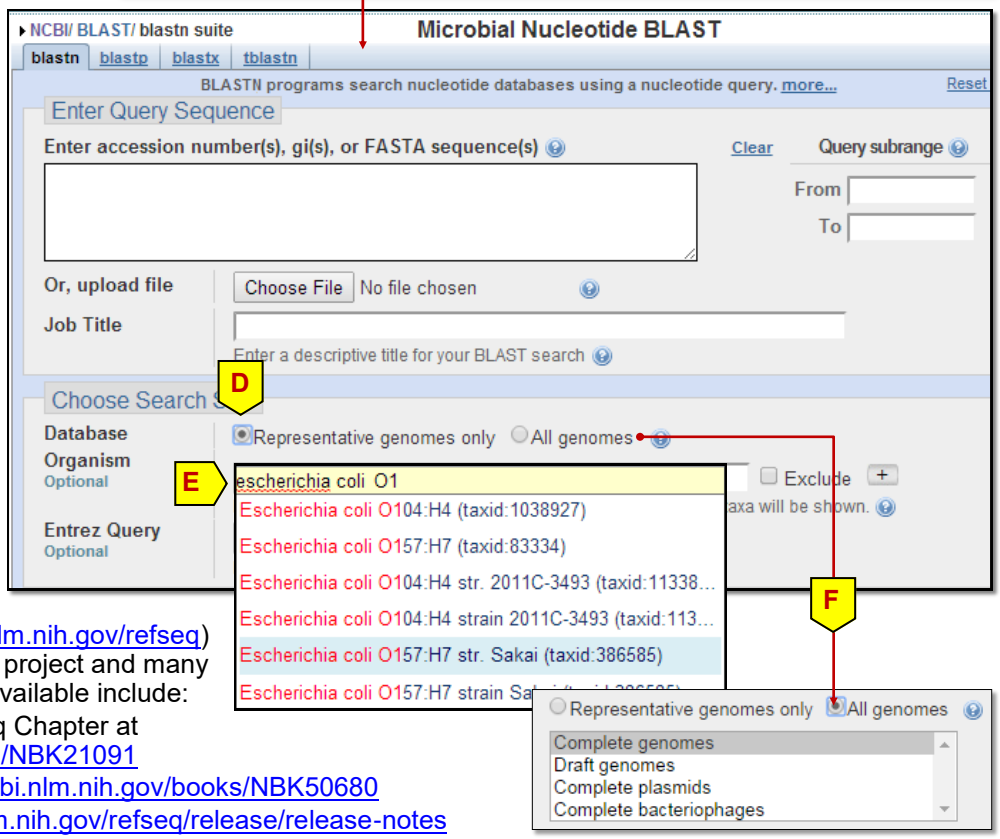


Searching RefSeq Using BLAST

The NCBI BLAST web service (blast.ncbi.nlm.nih.gov) provides a way to search the available NCBI RefSeq datasets using sequences as query. You can search against RefSeq entries using search pages listed under the "Basic BLAST" section by selecting RefSeq-only databases, i.e., ***refseq_rna***, ***refseq_representative_genomes, refseq_genomes***, ***refseq_protein***, and the newly introduced ***refseq_select*** for RNA and protein entries. To search against reference genome assemblies of a specific organism, use the organism-specific BLAST pages listed in the "BLAST Genomes" section (**A**) of the BLAST homepage. For organisms not listed, search with the organism name and select from the suggested list (**B**). This will retrieve a customized search page with the best dataset set as the target database.



RefSeq genomes for bacteria and archaea are also available through the Microbes link (**C**) from the BLAST homepage. This link points to the Microbial Genome BLAST page (partially shown to the right). The default database is set to a representative subset (**D**) selected by the research community and/or by NCBI computation. Use the Organism input box (**E**) to limit the search to specific taxa of interest. Check the "All Genomes" radio button (**F**) to see other microbial genomes available for selection. Hold ctrl key and click to select multiple databases.



Additional Information

Documentation

The RefSeq homepage (www.ncbi.nlm.nih.gov/refseq) contains a general description of the project and many technical details. Other documents available include:

- The NCBI Handbook RefSeq Chapter at www.ncbi.nlm.nih.gov/books/NBK21091
- The RefSeq FAQ at www.ncbi.nlm.nih.gov/books/NBK50680
- Release notes at ftp.ncbi.nlm.nih.gov/refseq/release/release-notes
- Pruitt KD, et. al. 2014. RefSeq: an update on mammalian reference sequences. Nucleic Acids Res. 42 (Database issue):D756-63 (www.pubmed.gov/24259432).
- RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation.
- RefSeq: an update on prokaryotic genome annotation and curation. Li W, et al. Nucleic Acids Res. 2021 Jan 8;49(D1):D1020-D1028 (www.pubmed.gov/33270901)

Alternative ways for data access

Data files from regularly updated RefSeq releases are available from the RefSeq FTP site (ftp.ncbi.nlm.nih.gov/refseq) and the genomes ftp site (ftp.ncbi.nlm.nih.gov/genomes/refseq). The Entrez Programming Utilities can be used to access the RefSeq dataset from the Nucleotide and Protein databases programmatically (www.ncbi.nlm.nih.gov/books/NBK25501/). The EntrezDirect package provides Unix command line access to RefSeq data and allows flexible workflow and data parsing (www.ncbi.nlm.nih.gov/books/NBK179288). The new datasets service (www.ncbi.nlm.nih.gov/datasets) provides a new way to access the data, through the web, by command line tool, or programmatically through API calls.

Feedback and technical assistance

- For RefSeq record update requests, use this web form www.ncbi.nlm.nih.gov/RefSeq/update.cgi
- For questions and feedback on NCBI resources in general, contact info@ncbi.nlm.nih.gov
- For questions and problems encountered during BLAST searches, contact blast-help@ncbi.nlm.nih.gov