



# 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/](http://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. 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. Other attributes that distinguish RefSeq from INSDC records are listed below.



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 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.

Taxonomic Node	Release 97 Count	Attribute	INSDC*	RefSeq
Archaea	1323	Accession includes underscore ('_')	No	Yes
Bacteria	60021	INSDC member	Yes	No
Fungi	12561	Source of sequence	Submitter	INSDC
Invertebrate	3817	Source of annotation	Submitter	Multiple
Mitochondrion	10048	Owner of sequence record	Submitter	NCBI
Plant	4343	Redundancy	High	Low
Plasmid	4476	Archive of experimental data	Yes	No
Plastid	4506	Review of experimental data	No	Yes
Protozoa	565	Curation by NCBI	No	Yes
Vertebrate_mammalian	1170	Regular update schedule	No	Yes
Vertebrate_other	4278			
Virus	9305			

\* The INSDC includes NCBI (USA), ENA (Europe), and DDBJ (Japan)

RefSeq also started to annotated experimentally verified functional elements onto genomic records, with NG\_ initialed accession. More information on this effort is at: [www.ncbi.nlm.nih.gov/refseq/functionalelements/](http://www.ncbi.nlm.nih.gov/refseq/functionalelements/)

## Access

RefSeq records are available by

- Searching in the nucleotide or protein databases ([www.ncbi.nlm.nih.gov/nucleotide](http://www.ncbi.nlm.nih.gov/nucleotide) and [www.ncbi.nlm.nih.gov/protein](http://www.ncbi.nlm.nih.gov/protein))
- BLASTing against selected databases ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)), and
- Downloading through FTP ([ftp.ncbi.nlm.nih.gov/refseq/](ftp://ftp.ncbi.nlm.nih.gov/refseq/) and [ftp.ncbi.nlm.nih.gov/genomes/refseq/](ftp://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, Protein Clusters, and more. RefSeq filters in the Entrez Nucleotide or 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/](http://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:

- Reference RNA sequences (refseq\_rna)
- Reference proteins (refseq\_protein)
- Reference genomic sequences (refseq\_genomic)
- NCBI genomes (chromosome) includes only the chromosome level records
- 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 by FTP ([ftp.ncbi.nlm.nih.gov/refseq/release/](ftp://ftp.ncbi.nlm.nih.gov/refseq/release/) and [ftp.ncbi.nlm.nih.gov/genomes/refseq/](ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/)).

## Record Display

The Nucleotide database displays Reference Sequences 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 this 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](http://www.pubmed.gov/26553804).

### Homo sapiens potassium voltage-gated channel subfamily C member 1 (KCNC1), transcript variant 2, mRNA

NCBI Reference Sequence: NM\_004976.4

[FASTA](#) [Graphics](#)

Go to:

[www.ncbi.nlm.nih.gov/nuccore/NM\\_004976.4](http://www.ncbi.nlm.nih.gov/nuccore/NM_004976.4)  
Reference section removed for brevity.

**LOCUS** NM\_004976  
**DEFINITION** Homo sapiens potassium voltage-gated channel subfamily C member 1 (KCNC1), transcript variant 2, mRNA.  
**ACCESSION** NM\_004976  
**VERSION** NM\_004976.4 GI:163792199  
**KEYWORDS** RefSeq.  
**SOURCE** Homo sapiens (human)  
**TAXONOMY** [Homo sapiens](#)  
 Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

**COMMENT** REVIEWED REFSEQ: This record has been curated by NCBI staff. The reference sequence was derived from [DN993192.1](#), [BC107129.2](#), [AC124056.8](#), [AK125480.1](#) and [AW452561.1](#). On Dec 20, 2007 this sequence version replaced gi:115527118.

**Summary:** This gene encodes a member of a family of integral membrane proteins that mediate the voltage-dependent potassium ion permeability of excitable membranes. Alternative splicing is thought to result in two transcript variants encoding isoforms that differ at their C-termini. These isoforms have had conflicting names in the literature: the longer isoform has been called both 'b' and 'alpha', while the shorter isoform has been called both 'a' and 'beta' (PMIDs 1432046, 12091563). [provided by RefSeq, Oct 2014].

**Transcript Variant:** This variant (2) lacks two exons and its 3' exon extends past a splice site used in variant 1, resulting in a distinct 3' coding region and 3' UTR, compared to variant 1. The encoded isoform (2) has a shorter and distinct C-terminus compared to isoform 1.

**Sequence Note:** This RefSeq record was created from genomic sequence data because no single transcript was supported by transcript alignments.

**Publication Note:** This RefSeq record includes a list of publications that are available for this gene.

**Biological Attributes**

- Bicistronic transcript CDS
- Uses downstream AUG
- Inferred exon combination
- Gene product(s) localized to mitochondrion
- Non-AUG initiation codon
- Nonsense-mediated mRNA decay
- PolyA required for stop codon
- Protein contains selenocysteine
- Readthrough transcript
- Ribosomal slippage
- Undergoes RNA editing
- Unitary pseudogene
- Regulatory uORF
- Multifunctional protein

**Reference sequence information**

RefSeq alternative splicing  
See 3 reference mRNA sequence splice variants for the KCNC1 gene.

**FEATURES** Location/Qualifiers

```

gene             1..6807
                 /gene="KCNC1"
                 /gene_synonym="KV3.1; KV4; NGK2"
                 /note="potassium voltage-gated channel, Shaw-related subfamily, member 1"
                 /db_xref="GeneID:3746"
                 /db_xref="HGNC:6233"
                 /db_xref="HPRD:15936"
                 /db_xref="MIM:176258"
exon             1..625
                 /gene="KCNC1"
                 /gene_synonym="KV3.1; KV4; NGK2"
                 /inference="alignment:Splice:1.39.8"
misc feature     41..43
                 /gene="KCNC1"
                 /gene_synonym="KV3.1; KV4; NGK2"
                 /note="upstream in-frame stop codon"
CDS              56..1591
                 /gene="KCNC1"
                 /gene_synonym="KV3.1; KV4; NGK2"
                 /note="isoform B is encoded by transcript variant B; voltage-gated potassium channel protein KV3.1; potassium voltage-gated channel subfamily C member 1; voltage-gated potassium channel subunit Kv4; voltage-gated potassium channel subunit Kv3.1"
                 /codon_start=1
  
```

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

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 (left, H) sums up link names and resources they represent.

## 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) display (E), with

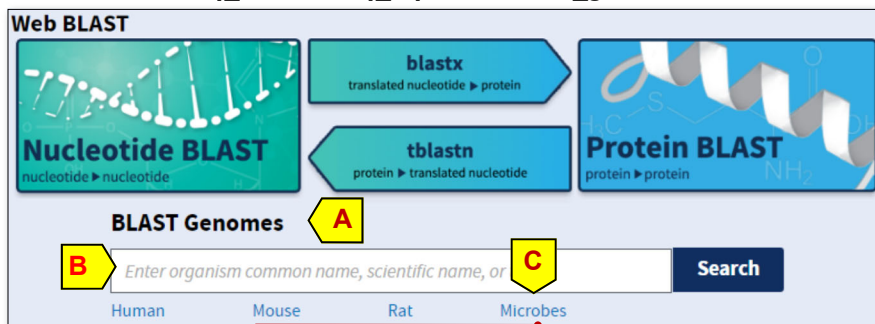
the alternatively spliced transcripts shown at the top (F). Tracks below the transcripts provide additional supporting evidence from RNA-seq data (G).

Complete RefSeq proteomes from a selected subset of organisms representing key taxonomic groups, are also input into HomoloGene, which identifies homologs through a comprehensive set of calculations. In addition, the NCBI eukaryotic genome annotation pipeline automatically analyzes the RefSeq proteins it processes to identify homologous genes. Both sets of homologs are available under the “Homology” heading (H) of the “General gene information” section of the Gene full report page.

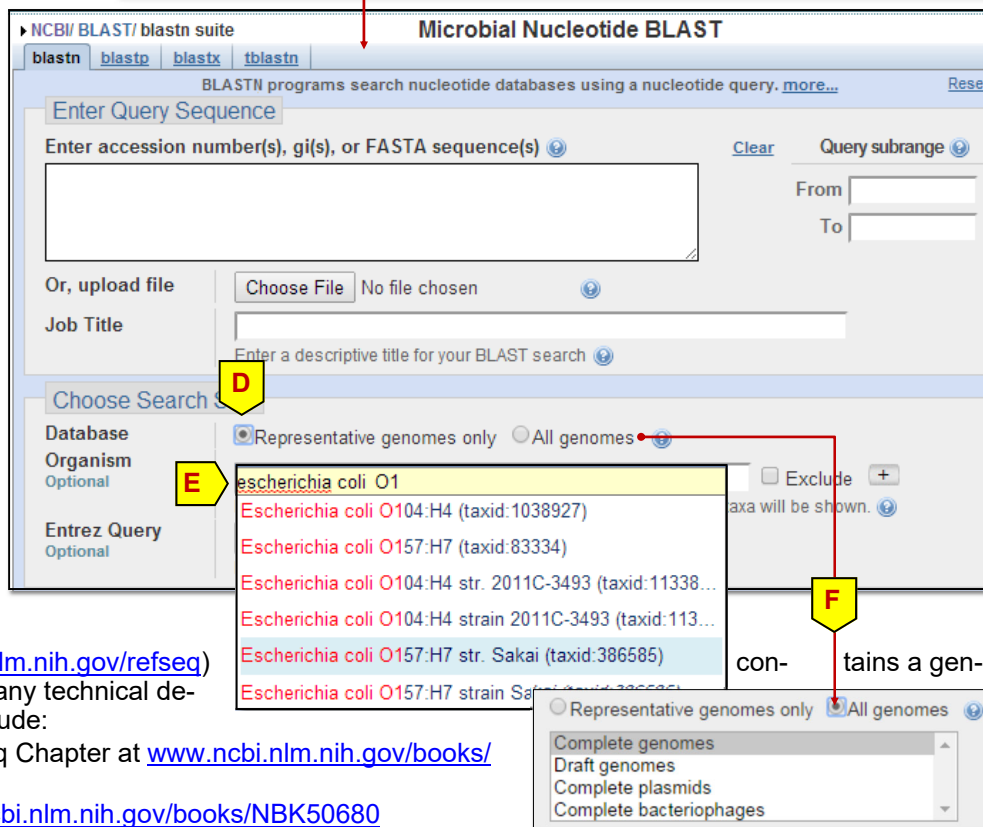


## Searching RefSeq Using BLAST

The NCBI BLAST web service ([blast.ncbi.nlm.nih.gov](http://blast.ncbi.nlm.nih.gov)) provides a way to search the public databases maintained at NCBI using sequences as the 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\_genomic*, or *refseq\_protein*. 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. To locate the best genomic data available for organisms not listed, type the name in the search box (B), select from the suggested list, and click “Search”. This will retrieve a customized search page with that dataset as the target database.



To search against RefSeq genomes for bacteria and archaea, use the Microbes link (C) from the BLAST homepage, which points to the Microbial Genome BLAST page (shown in part). In this page, 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 by typing and selecting from the suggested list. Check the “All Genomes” radio button (F) to see other microbial genomes available for selection. Hold ctrl key and click to select multiple datasets.



## Additional Information

### Documentation

The RefSeq homepage ([www.ncbi.nlm.nih.gov/refseq](http://www.ncbi.nlm.nih.gov/refseq)) provides 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](http://www.ncbi.nlm.nih.gov/books/NBK21091)
- The RefSeq FAQ at [www.ncbi.nlm.nih.gov/books/NBK50680](http://www.ncbi.nlm.nih.gov/books/NBK50680)
- Release notes at [ftp.ncbi.nlm.nih.gov/refseq/release/release-notes](http://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.ncbi.nlm.nih.gov/pubmed/24259432](http://www.ncbi.nlm.nih.gov/pubmed/24259432)).
- RefSeq: an update on prokaryotic genome annotation and curation. Haft DH, et al. *Nucleic Acids Res.* 2018 Jan 4;46(D1):D851-D860 ([www.ncbi.nlm.nih.gov/pubmed/29112715](http://www.ncbi.nlm.nih.gov/pubmed/29112715))

### Alternative ways for data access

The RefSeq sequence data are released on a regular schedule. Data files from the release are available from the RefSeq FTP site ([ftp.ncbi.nlm.nih.gov/refseq](http://ftp.ncbi.nlm.nih.gov/refseq)) and the genomes ftp site ([ftp.ncbi.nlm.nih.gov/genomes/refseq](http://ftp.ncbi.nlm.nih.gov/genomes/refseq)). The Aspera download site ([www.ncbi.nlm.nih.gov/public](http://www.ncbi.nlm.nih.gov/public)) provides a faster alternative, which requires Aspera browser plugin ([downloads.asperasoft.com/connect2/](http://downloads.asperasoft.com/connect2/)). In addition, EUtils can be used to access the RefSeq dataset from the Nucleotide and Protein databases programmatically. The EDirect package provides command line access to RefSeq data and allows flexible workflow and data parsing ([www.ncbi.nlm.nih.gov/books/NBK179288/](http://www.ncbi.nlm.nih.gov/books/NBK179288/)).

### Feedback and technical assistance

- For RefSeq record update requests, contact [www.ncbi.nlm.nih.gov/RefSeq/update.cgi](http://www.ncbi.nlm.nih.gov/RefSeq/update.cgi)
- For questions and feedback on NCBI resources in general, contact [info@ncbi.nlm.nih.gov](mailto:info@ncbi.nlm.nih.gov)
- For questions and problems encountered during BLAST searches, contact [blast-help@ncbi.nlm.nih.gov](mailto:blast-help@ncbi.nlm.nih.gov)