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 (<u>www.ncbi.nlm.nih.gov/refseq/</u>). 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 se-

guence 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 RefSeg 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	Bedundancy	High	Low		
Plasmid	4476		Vaa	Ne		
Plastid	4506	Archive of experimental data	No No	Yes Yes Yes		
Protozoa	565	Review of experimental data				
Vertebrate_mammalian	1170	Curation by NCBI				
Vertebrate_other	4278	Regular update schedule	No			
Virus	9305	* The INSDC includes NCBI (USA), ENA (Europe), and DDBJ (Japan)				

RefSeq also started to anno-

tated 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/

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)
- BLASTing against selected databases (<u>blast.ncbi.nlm.nih.gov</u>), and
- Downloading through FTP (<u>ftp.ncbi.nlm.nih.gov/refseq/</u> and <u>ftp.ncbi.nlm.nih.gov/genomes/refseq/</u>)

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, <u>www.ncbi.nlm.nih.gov/books/NBK25501/</u>) 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 (<u>ftp.ncbi.nlm.nih.gov/refseq/release/</u> and <u>ftp.ncbi.nlm.nih.gov/genomes/refseq/</u>).

Page 2

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.

Homo sapiens potassium voltage-gated channel subfamily C member 1 (KCNC1), script variant 2, mRNA					Change region shown			
FASTA Gra	uphics B			Ana	lyze this sequence			
<u>Go to:</u> ♥		www.ncbi.nlm.nih	.gov/nuccore/NM_004976.4	Run	BLA Biological Attributes			
LOCUS	NM 004976	Reference sec	tion removed for brevity.	Pick	Pri Bicistronic transcript CDS			
DEFINITION	Homo sapiens potassium v	oltage-gated channe	el subfamily C member 1	High	Uses downstream AUG			
ACCESSION	(KCNC1), transcript var: NM 004976	iant 2, mRNA.		Find	In t Inferred exon combination			
VERSION	NM_004976.4 GI:16379219	99		0 mtia	Gene product(s) localized to mitochondrion			
KEYWORDS	RetSeq. Homo sapiens (human)			Ky3	Non-AUG initiation codon			
E JANISM	Homo sapiens			secu	Rea Nonsense-mediated mRNA decay			
	Eukaryota; Metazoa; Cho Mammalia; Eutheria; Eua	rdata; Craniata; Ver rchontoglires; Prima	rtebrata; Euteleostomi; ates; Haplorrhini;	A rec	Curr PolyA required for stop codon			
	Catarrhini; Hominidae; H	lomo.		prog	Protein contains selenocysteine			
COMMENT	REVIEWED <u>REFSEQ</u> : This re	cord has been curat	ed by NCBI staff. The	Geno	nde Readthrough transcript			
	AC124056.8, AK125480.1 a	erived from <u>DN99319</u> and AW452561.1.	<u>92.1, BC107129.2</u> ,		Ribosomal slippage			
	On Dec 20, 2007 this sec	uence version repla	aced gi: <u>115527118</u> .		Undergoes RNA editing			
	Summary: This gene encod	les a member of a fa	mily of integral D	Path	unitary pseudogene			
	membrane proteins that m permeability of excitabl	ediate the voltage- le membranes. Altern	dependent potassium ion Native splicing is	Volta	ge Regulatory uORF			
	thought to result in two	transcript variant	s encoding isoforms that	Pota	ssiu Multifuncational protein			
	names in the literature:	1. These isoforms h the longer isoform	nave had conflicting 1 has been called both	Neur	onal System			
	'b' and 'alpha', while t	the shorter isoform	has been called both 'a'					
	and beta (PMIDS 143204	6, 12091563). [prov	laed by RetSeq, Oct 2014].	Reference sequence information				
	Transcript Variant: This evon extends past a spli	; variant (2) lacks ice site used in var	two exons and its 3'	RefS	Seq alternative splicing			
	distinct 3' coding regio	ared to variant 1. The	See for th	be KCNC1 gene.				
	encoded isoform (2) has to isoform 1.	a shorter and disti	gene 16807	on/Qualif	iers			
	Sequence Note: This Refs	eq record was creat	ed fr /gene="//gene a	KCNC1"	KV3.1; KV4; NGK2"			
	for the full length of t	he gene. The extent	note="/note="	/note="potassium voltage-gated channel, Shaw-relat subfamily, member 1"				
	supported by transcript	alignments.	/db_xref="GeneID: <u>3746</u> " /db_xref="HGNC:6233"					
	Publication Note: This	RefSeq record inclu	udes a /db_xre /db_xre	f="HPRD: f="MIM:1	<u>15936</u> " 76258"			
I internet	publications that are av	ailable for this ge	ene. P exon 1625	KCNC1"				
	NCBI Cone database		/gene_s /infere	ynonym="	KV3.1; KV4; NGK2" gnment:Splign:1.39.8"			
HGNC	HUGO Gene Nomenclatu	re Committee	-3G707 misc feature 4143 /gene="	KCNC1"				
MIM	OMIM database		sup /gene_s SAM /note="	upstream	KV3.1; KV4; NGK2" in-frame stop codon"			
protein id	NCBI protein database		CDS 56159 /gene=	KCNC1"	W19 1. W14. NOVOI			
CCDS	Consensus CDS		/note="	isoform	B is encoded by transcript variant B; otassium channel protein KV3.1: potassium			
variation	NCBI SNP database		ARY_S voltage	-gated c	hannel subfamily C member 1; voltage-gated el subunit Kv4; voltage-gated potassium			
IMGT/Gene-D	B ImMunoGeneTics		5 channel /codon	subunit start=1	Kv3.1"			
FlyBase	Database of Drosophila 0	Gene & Genomics						
RGD	Rat Genome Database		The FEATURES section	ι (<mark>G</mark>) co	ntains detailed annotation of exons,			
EcoGene	EcoGene Database		<pre>polyadenylation signals (misc_feature) of the ge</pre>	or sites ne, cod	, miscellaneous teatures ing sequences (CDS) and protein			
TAIR	The Arabidopsis Informat	ion Resource	include	e cross-references to related records				
SGD	Saccharomyces Genome	Database	CBI or e	elsewhere. The table (left, H) sums				
BeeBase	hymenoptera genome da	tabase		แต่สุรุเท				

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.

	Full Report - A	Send to: +	Hide sidebar >>		
	KCNC1 potassi	um voltage-gated channel subfamily C member 1 [Homo sapiens (human)]			
Format	www.ncbi.nlm	.nih.gov/gene/3746	Table of contents		
Full Report	Summary	8 7	Summary		
Eall Report (text)	Official Symbol	KONO1	Genomic context		
Gene Table D	Official Full Name	notassium voltage-nated channel subfamily C member 1 provided by HGNC	Genomic regions, transcripts, and products		
Gene Table (text)	Primary source	HGNC:HGNC:6233	Expression		
GeneRIF	See related	Ensembl:ENSG00000129159 MIM:176258; Vega:OTTHUMG00000166359	Bibliography		
Summary	Gene type	protein coding	Phenotynes		
Summary (text)	RefSeq status		Variation		
Tabular	Urganism	Homo saplens Fukarvota: Metazoa: Chordata: Craniata: Vertebrata: Futeleostomi: Mammalia: Futberia: Fuarchontonlires:	variation		
	Lineage	Primates: Haplorrhini: Catarrhini: Hominidae: Homo	Pathways from BioSystems		
© ASIN. I	Also known as	KV4; EPM7; NGK2; KV3.1	Interactions		
UI List	Summary	This gene encodes a member of a family of integral membrane proteins that mediate the voltage-dependent	General gene information		
		potassium ion permeability of excitable membranes. Alternative splicing is thought to result in two transcript	Markers, Clone Names, Homology, Gene Ontology		
		variants encoding isotorms that differ at their C-termini. These isotorms have had conflicting names in the	General protein information		
		called both "a" and "beta" (PMIDs 1432046, 12091563), [provided by RefSeq. Oct 2014]	NCBI Reference Sequences (RefSeq)		
	Expression	Biased expression in brain (RPKM 5.1) and testis (RPKM 0.7) See more	Related sequences		
	Orthologs	mouse all	Additional links		

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**).

2	5 NC_0000	11.10 - Find	1:		Ě E P			କ୍ ୩୮୦ ដ	1	X Tools	 Trac 	(S • 🥰
	17,730 K	17,735 K	17,740 K	17,745 K	17,750	17,755 K	17,760 K	17,765 K	17,770 K	17,775 K	17,780 K	17,785 K
lenes	, NCBI HO	omo sapiens	s Annotati	on Release	109, 2018-	-03	<u>}</u>					_
	VP 938966 2	_				KCN	401		-			
	NM_0011127	41.1	<u></u>	<u> </u>	<u> </u>			- <u>></u>		>		00110521
	NM_00497	76.4	*		>					NP_0	04967.1	
Senes	, Ensembl	l release 9	ЭĞ			14						J
03.3												
ISG000	00129159 [>	>						-	>	> _	>
Clini	cal, dbSN	NP b152 v2										
it ad	Vaniatio	II dbeND	b152 ++2								() ()	
JILEO	Variatio	JIIS, UDSNP	DIJ2 V2									
Live	RefSNPs,	dbSNP b152	2 v2									
RNA-S	eq exon o	coverage, a	aggregate	(filtered),	NCBI Homo	sapiens	Annotation	E749	109 - log	base 2 sca	led	~
		A Antonia Acada	and the second second		- Adamin Jan		-	128	And a fail and	Statistics of the local division of the loca	A state of the	
A shife	the Los and Dates									0 les be	0	<u> </u>
(NA-S	eq intro	n-spanning	reads, ag	gregate (11	lterea), 1	NCBI Homo	sapiens An	notation	Release 10	19 - log ba	se 2 scale	ed 🔹
		_						128)
	17 700 K	17 705 K	17 740 K	12 245 K	17 750 K	12 255 K	17 700 K	17 705 K	17 770 K	12 225 K	17 700 K	17 705 K
	17,730 K	17,735 K	17,740 K	17,745 K	17,750 K	17,755 K	17,760 K	17,765 K	17,770 K	17,775 K	17,780 K	17,785 K

Complete RefSeq proteomes from a

selected subset of organisms representing key taxonomic groups, are also input into HomoloGene, which identifies homologd 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.

General gene information	Homologs from comprehensive analy	sis of complete RefSeq	* ?				
Markers	proteomes from selected organisms F	HomolGene.					
🗆 Homology 🖁 🛔							
Homologs of the KCNC1 gene: The KCNC1 gene is conserved in chimpanzee, Rhesus monkey, dog, cow, mouse, rat, chicken, zebrafish,							
mosquito, and C.elegans.							
Orthologs from Annotation Pipeline: 246 organisms have orthologs with human gene KCNC1							
Orthologs							
E Clone Names		Homoloas identified from	n eukarvotic genome	e `			
I Cione Names		annotation nineline	,	-			
Gene Ontology Provided by GO							

Searching RefSeg Using BLAST

The NCBI BLAST web service (blast.ncbi.nlm.nih.gov) provides a way to search the public databases maintained at NCBI using sequences as the guery. You can search against RefSeg entries using search pages listed under the "Basic BLAST" section by selecting RefSeq-only databases, i.e., refseq rna, refseq representative genomes, ref-

seq 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 RefSeg 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 RefSeg homepage (www.ncbi.nlm.nih.gov/refseg) eral description of the project and many technical details. Other documents available include:

- The NCBI Handbook RefSeg 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. RefSeg: an update on mammalian reference sequences. Nucleic Acids Res. 42 (Database issue):D756-63 (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)

Alternative ways for data access

The RefSeg 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) and the genomes ftp site (ftp.ncbi.nlm.nih.gov/genomes/refseq). The Aspera download site (www.ncbi.nlm.nih.gov/public) provides a faster alternative, which requires Aspera browser plugin (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/).

Feedback and technical assistance

- For RefSeq record update requests, contact 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 <u>blast-help@ncbi.nlm.nih.gov</u>



Draft genomes

Complete plasmids Complete bacteriophages

Page 4

Web BLAST blastx Protein BLAS Nucleotide BLAST tblastn ▶ translated nucleotide rotein 🕨 proteir **BLAST Genomes** С Search Enter organism common name, scientific name, o Human Mouse Rat Microbes NCBI/ BLAST/ blastn suite Microbial Nucleotide BLAST blastn blastp blastx tblastn BLASTN programs search nucleotide databases using a nucleotide query. more... Reset Enter Query Sequence Enter accession number(s), gi(s), or FASTA sequence(s) 😡 Query subrange 😡 Clear From То Or, upload file Choose File No file chosen \bigcirc Job Title Enter a descriptive title for your BLAST search 🙆 D Choose Search Database Representative genomes only OAII genomes Organism Exclude + escherichia coli O1 Optional axa will be sh<mark>own. 😡</mark> Escherichia coli O104:H4 (taxid:1038927) Entrez Query Escherichia coli O157:H7 (taxid:83334) Optional Escherichia coli O104:H4 str. 2011C-3493 (taxid:11338. F Escherichia coli O104:H4 strain 2011C-3493 (taxid:113) Escherichia coli O157:H7 str. Sakai (taxid:386585) tains a gencon-Escherichia coli O157:H7 strain Sa Representative genomes only All genomes 0 Complete genomes