



The RefSeqGene Project

A collection of sequences as foundation for gene-based coordinates

<https://www.ncbi.nlm.nih.gov/refseq/rsg/>

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

Scope of the RefSeqGene project

RefSeqGene, a subset of NCBI's Reference Sequence (RefSeq) project, defines genomic sequences to be used as reference standards for well-characterized genes. These sequences provide a stable foundation for reporting mutations, establishing conventions for numbering exons and introns, and defining the coordinates of variations such as single nucleotide (SNV), multiple nucleotide (MNV), insertions, and deletions. Sequences from this project solve the drawbacks and limitations present in mRNA- or whole chromosome-based systems by providing gene-specific genomic sequences for each gene with its upstream, intronic, and downstream flanking regions. Modifications made to RefSeqGene sequences are versioned and a tool is provided to facilitate conversion of coordinates among versions: www.ncbi.nlm.nih.gov/genome/tools/remap/#tab=rsg



The RefSeqGene project is an active member of the Locus Reference Genomic (LRG) collaboration. Input and leadership from Dr. M. L. Gulley and the Molecular Pathology Resource Committee of the College of American Pathologists has greatly facilitated its implementation [1].

Sequence selection

Sequences in the RefSeqGene set represent well-supported, naturally occurring haplotypes, and prevalent alleles. The RefSeqGene group collaborates with multiple Locus Specific Databases (LSDBs) and the LRG project of GEN2PHEN [2] to establish and maintain these standard sequences. The RefSeqGene-annotated genes with identified LSDB counterparts can be retrieved from NCBI Gene (www.ncbi.nlm.nih.gov/gene) using the term *refseqgene*. Sequences of RefSeqGene entries can be retrieved using fielded term *refseqgene[keyword]* from the NCBI Nucleotide database (www.ncbi.nlm.nih.gov/nucleotide).

Data access

Over 5,300 RefSeqGene records are available to serve as the foundation for gene-based coordinates. The RefSeqGene homepage (A) provides access to browse and search for RefSeqGene entries (B). It lists the available entries in a summary table and provides official symbols, gene name, GeneID, and other information. The list can be filtered by terms entered in the text box (C) or browsed by paging (D). RefSeqGene sequences are also accessible through the BLAST homepage.

The screenshot shows the RefSeqGene homepage. Callout A points to the 'Gene' dropdown menu. Callout B points to the 'Browse Genes with RefSeqGene Sequences' link. Callout C points to the search text box. Callout D points to the pagination controls.

1. Gulley et al. *Clinical laboratory reports in molecular pathology*. 2007. *Arch Pathol Lab Med*. 131(6): 852-63.
2. Dalgleish et al. *Locus Reference Genomic sequences: an improved basis for describing human DNA variants*. 2010.

The screenshot shows the 'RefSeqGene Records' table. Callout C points to the search filter text box. Callout D points to the pagination controls.

Symbol	Name	GeneID	LRG	RSGID	Views	GTR	Associated Diseases
A1CE	APOBEC1 complementation factor	29974		NG_029916.1	graphic, sequence		
A2M	alpha-2-macroglobulin	2		NG_011717.1	graphic, sequence	GTR	Alzheimer's disease (OMIM 104300) Alpha-2-macroglobulin deficiency (OMIM 614036)
A2ML1	alpha-2-macroglobulin-like 1	144568		NG_042857.1	graphic, sequence		

Contents of a gene record

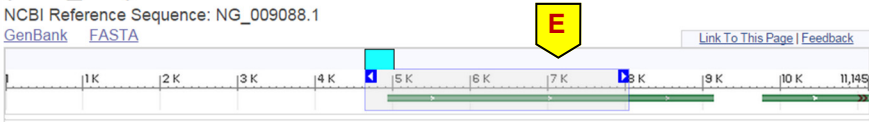
In the RefSeqGene Browse list (shown below), summary information and links to records in other databases are provided for each entry. Here, official gene symbols (**A**) in the “Symbol” column link to Gene records. The “graphic” (**B**) in the “Views” column displays sequences in the NCBI Sequence Viewer. The “GTR” column (**C**) provides links to the NIH Genetic Testing Registry (GTR). Additionally, IDs (**D**) in the “Associated Diseases” columns link to OMIM records with disease-specific information.

RefSeqGene Records

File table: **B** **C** **D**

Symbol	Name	GeneID	LRG	RSID	Views	GTR	Associated Diseases
IL2RA	interleukin 2 receptor, alpha	3559	LRG_73	NG_007403.1	graphic, sequence	GTR	Interleukin 2 receptor, alpha, deficiency of (OMIM 606367) Diabetes mellitus, insulin-dependent, 10 (OMIM 601942)
IL2RG	interleukin 2 receptor, gamma	3561	LRG_150	NG_009088.1	graphic, sequence	GTR	X-linked severe combined immunodeficiency (OMIM 300400) Combined immunodeficiency, X-linked (OMIM 312863)

Homo sapiens interleukin 2 receptor, gamma (IL2RG), RefSeqGene (LRG_150) on chromosome X



Sequence Viewer display

In the Sequence Viewer display (left), the complete gene locus, represented by the RefSeqGene, is depicted at the top (**E**) and details for the Gene are shown in the

The screenshot shows the NCBI Sequence Viewer interface. At the top, the gene model for IL2RG is displayed (E). Below this, several tracks are visible: SNPs, ClinVar Short Variations based on dbSNP 141 (G), Cited Variations based on dbSNP 141, and Genes. A 'Tools' menu is open (H), showing options such as 'Go To', 'Search', 'Flip Strands', 'Markers', 'Set Origin', 'Sequence Text View' (I), 'Add new Panel', 'Add new Panel on Range', 'BLAST and Primer Search', 'Download', and 'Printer-Friendly PDF'. The bottom panel shows the 'Sequence View (positive strand)' with the DNA sequence and its corresponding protein translation.

panel below (F). The set of variation tracks (G) provides an comprehensive overview of existing data in the context of gene features.

Clicking the “Tools” icon (H) opens a popup menu allowing the selection of “Sequence Text View” option to display the annotated sequences in a floating text window (I) within the Sequence viewer display.