WIKIPEDIA
The Free Encyclopedia

RefSeq

The Reference Sequence (**RefSeq**) database[1] is an open access, annotated and curated collection of publicly available nucleotide sequences (DNA, RNA) and their protein products. RefSeq was introduced in 2000.[2][3] This database is built by National Center for Biotechnology Information (NCBI), and, unlike GenBank, provides only a single record for each natural biological molecule (i.e. DNA, RNA or protein) for major organisms ranging from viruses to bacteria to eukaryotes.

For each model organism, *RefSeq* aims to provide separate and linked records for the genomic DNA, the gene transcripts, and the proteins arising from those transcripts. *RefSeq* is limited to major organisms for which sufficient data are available (121,461 distinct "named" organisms as of July 2022),[4] while GenBank includes sequences for any organism submitted (approximately 504,000 formally described species).[5]

Refseq

Content	
Description	curated non-redundant sequence database of genomes.
Contact	
Research center	National Center for Biotechnology Information
Primary citation	Pruitt KD & al. (2005)[1]
Access	
Website	https://www.ncbi.nlm.nih.gov/RefSeq

RefSeq categories

RefSeq collection comprises different data types, with different origins, so it is necessary to establish standard categories and identifiers to store each data type. The most important categories are:

RefSeq accession categories and molecule types

Category	Description
NC	Complete genomic molecules
NG	Incomplete genomic region
NM	mRNA
NR	ncRNA
NP	Protein
XM	predicted mRNA model
XR	predicted ncRNA model
XP	predicted Protein model (eukaryotic sequences)
WP	predicted Protein model (prokaryotic sequences)

For more details and more categories, see Table 1 (https://www.ncbi.nlm.nih.gov/books/NBK21091/table/ch18.T.entrez_queries_to_retrieve_sets_o) in Chapter 18 of the book *The Reference Sequence (RefSeq) Database* (https://www.ncbi.nlm.nih.gov/books/NBK21091).

RefSeq Projects

Several projects to improve *RefSeq* services are currently in development by the NCBI, often in collaboration with research centers such as EMBL-EBI:

- **Consensus CDS (CCDS):** This project aims to identify a core set of human and mouse protein-coding regions and standardize sets of genes with high and consistent levels of genomic annotation quality. This project was announced in 2009 and is still in development.[6][7]

- **RefSeq Functional Elements (RefSeqFE):** It is focused on describing non-genic functional elements which are gene regulatory regions such as: enhancers, silencers, DNase I hypersensitive regions, DNA replication origins etc.). The current scope of this project is restricted to the human and mouse genomes.[8]
- **RefSeqGene:** Its main goal is to define genomic sequences to be used as reference standards for well-characterized genes. Previously described mRNA, protein and chromosome sequences have the weaknesses of not providing explicit genomic coordinates of gene flanking and intronic regions as well as showing awkwardly large coordinates that change with every new genome assembly. The RefSeqGene project is designed to eliminate these errors.[9]
- **Targeted Loci:** This project records molecular markers, specially protein-coding and ribosomal RNA loci that are used for phylogenetic and barcoding analysis. The scope of this project includes sequences for Archaea, Bacteria and Fungi organisms, accessible via Entrez and BLAST queries. It also includes GenBank sequences for Animals, Plants and Protists, accessible via BLAST queries.[10]
- **Virus Variation (ViV):** It is a specific resource of sequence data processing pipelines and analysis tools for display and retrieval of sequences from several viral groups such as influenza virus, ebolavirus, MERS coronavirus or Zika virus. New viruses, processing pipelines, tools and other features are included regularly.[11]
- **RefSeq Select:** This project aims to select datasets of **RefSeq Select** transcripts, as the most representative for every protein-coding gene, based on multiple criteria: prior use in clinical databases, transcript expression, evolutionary conservation of the coding region etc. Since many genes are represented by multiple *RefSeq* transcripts/proteins due to the biological process of alternative splicing, this complexity is problematic for studies such as comparative genomics or exchange of clinical variant data.[12]
- **MANE** (**M**atched **A**nnotation from the **N**CBI and **E**MBL-EBI): It is a collaborative project between NCBI and EMBL-EBI whose main goal is to define a set of transcripts and their proteins for all the protein-coding genes in the human genome. By doing that, the differences in transcripts annotation between *RefSeq* and Ensembl/GENCODE annotation systems are reduced. A **MANE Select** transcripts set are created as a useful universal standard for clinical reporting and comparative or evolutionary genomics. A second **MANE Plus Clinical** set are also created with additional transcripts to report all *Pathogenic* (P) or *Likely Pathogenic* (LP) clinical variants available in public resources.[13] This project was announced in 2018 and is expected to finish in 2022.

Statistics

According to the RefSeq release 213 (July 2022), the number of species represented in the database by counting distinct taxonomic IDs are as follows:[4]

Taxonomic ID	Species
Archaea	1443
Bacteria	69122
Complete	121461
Fungi	16869
Invertebrate	5715
Mitochondrion	13648
Plant	9177
Plasmid	6073
Plastid	9430
Protozoa	746
Vertebrate (mammalian)	1509
Viral	11620
Vertebrate (other)	5237
Other	4

The counts of accession and basepairs per molecule type are:[4]

Molecule type	Accessions	Basepairs/residues
Genomics	40,758,769	$2.923212393984 \times 10^{12}$
RNA	45,781,716	$1.22253022047 \times 10^{11}$
Protein	234,520,053	$9.129062394 \times 10^{10}$

See also

- GenBank
- Sequence analysis
- Sequence profiling tool
- Sequence motif
- UniProt
- List of sequenced eukaryotic genomes
- List of sequenced archaeal genomes

References

1. Pruitt KD, Tatusova T, Maglott DR (January 2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539979). *Nucleic Acids Research*. **33** (Database issue): D501–D504. doi:10.1093/nar/gki025 (https://doi.org/10.1093%2Fnar%2Fgki025).

PMC 539979 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC539979). PMID 15608248 (https://pubmed.ncbi.nlm.nih.gov/15608248).

2. Maglott DR, Katz KS, Sicotte H, Pruitt KD (January 2000). "NCBI's LocusLink and RefSeq" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102393). *Nucleic Acids Research*. **28** (1): 126–128. doi:10.1093/nar/28.1.126 (https://doi.org/10.1093%2Fnar%2F28.1.126). PMC 102393 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC102393). PMID 10592200 (https://pubmed.ncbi.nlm.nih.gov/10592200).
3. Pruitt KD, Katz KS, Sicotte H, Maglott DR (January 2000). "Introducing RefSeq and LocusLink: curated human genome resources at the NCBI". *Trends in Genetics*. **16** (1): 44–47. doi:10.1016/s0168-9525(99)01882-x (https://doi.org/10.1016%2Fs0168-9525%2899%2901882-x). PMID 10637631 (https://pubmed.ncbi.nlm.nih.gov/10637631).
4. RefSeq Release 213 Statistics (http://ftp.ncbi.nlm.nih.gov/refseq/release/release-notes/) (Report). National Library of Medicine. 11 July 2022. Retrieved 20 July 2022.
5. Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (January 2022). "GenBank" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8690257). *Nucleic Acids Research*. **50** (D1): D161–D164. doi:10.1093/nar/gkab1135 (https://doi.org/10.1093%2Fnar%2Fgkab1135). PMC 8690257 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8690257). PMID 34850943 (https://pubmed.ncbi.nlm.nih.gov/34850943).
6. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. (July 2009). "The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2704439). *Genome Research*. **19** (7): 1316–1323. doi:10.1101/gr.080531.108 (https://doi.org/10.1101%2Fgr.080531.108). PMC 2704439 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2704439). PMID 19498102 (https://pubmed.ncbi.nlm.nih.gov/19498102).
7. Pujar S, O'Leary NA, Farrell CM, Loveland JE, Mudge JM, Wallin C, et al. (January 2018). "Consensus coding sequence (CCDS) database: a standardized set of human and mouse protein-coding regions supported by expert curation" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753299). *Nucleic Acids Research*. **46** (D1): D221–D228. doi:10.1093/nar/gkx1031 (https://doi.org/10.1093%2Fnar%2Fgkx1031). PMC 5753299 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5753299). PMID 29126148 (https://pubmed.ncbi.nlm.nih.gov/29126148).
8. Farrell CM, Goldfarb T, Rangwala SH, Astashyn A, Ermolaeva OD, Hem V, et al. (January 2022). "RefSeq Functional Elements as experimentally assayed nongenic reference standards and functional interactions in human and mouse" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8744684). *Genome Research*. **32** (1): 175–188. doi:10.1101/gr.275819.121 (https://doi.org/10.1101%2Fgr.275819.121). PMC 8744684 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8744684). PMID 34876495 (https://pubmed.ncbi.nlm.nih.gov/34876495).
9. Gulley ML, Braziel RM, Halling KC, Hsi ED, Kant JA, Nikiforova MN, et al. (June 2007). "Clinical laboratory reports in molecular pathology". *Archives of Pathology & Laboratory Medicine*. **131** (6): 852–863. doi:10.5858/2007-131-852-CLRIMP (https://doi.org/10.5858%2F2007-131-852-CLRIMP). PMID 17550311 (https://pubmed.ncbi.nlm.nih.gov/17550311).
10. "NCBI RefSeq Targeted Loci Project" (https://www.ncbi.nlm.nih.gov/refseq/targetedloci/). *www.ncbi.nlm.nih.gov*. Retrieved 2022-07-27.
11. Hatcher EL, Zhdanov SA, Bao Y, Blinkova O, Nawrocki EP, Ostapchuck Y, et al. (January 2017). "Virus Variation Resource - improved response to emergent viral outbreaks" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210549). *Nucleic Acids Research*. **45** (D1): D482–D490. doi:10.1093/nar/gkw1065 (https://doi.org/10.1093%2Fnar%2Fgkw1065). PMC 5210549 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5210549). PMID 27899678 (https://pubmed.ncbi.nlm.nih.gov/27899678).
12. "NCBI RefSeq Select" (https://www.ncbi.nlm.nih.gov/refseq/refseq_select/). *www.ncbi.nlm.nih.gov*. Retrieved 2022-07-27.

13. Morales J, Pujar S, Loveland JE, Astashyn A, Bennett R, Berry A, et al. (April 2022). "A joint NCBI and EMBL-EBI transcript set for clinical genomics and research" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9007741). *Nature*. **604** (7905): 310–315. doi:10.1038/s41586-022-04558-8 (https://doi.org/10.1038%2Fs41586-022-04558-8). PMC 9007741 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9007741). PMID 35388217 (https://pubmed.ncbi.nlm.nih.gov/35388217).

Sources

- This article incorporates public domain material from *NCBI Handbook* (https://www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2). National Center for Biotechnology Information.

External links

- RefSeq (https://www.ncbi.nlm.nih.gov/RefSeq)
- GenBank, RefSeq, TPA and UniProt: What's in a Name? (https://www.ncbi.nlm.nih.gov/books/NBK21105/#ch1.Appendix_GenBank_RefSeq_TPA_and_UniP)

Retrieved from "https://en.wikipedia.org/w/index.php?title=RefSeq&oldid=1174898842"