Secondary Nucleotide Databases

Shatakshi Kulkarni

Introduction

- Secondary nucleotide databases are databases that curate and organize nucleotide sequences from primary databases, such as GenBank, EMBL, and DDBJ.
- They provide added value through annotation, analysis, and the integration of additional information.
- These databases often offer user-friendly interfaces and tools for sequence alignment, similarity searches, and functional predictions.
- These databases often focus on specific types of sequences, specific organisms, or specific types of analysis.
- E.g., Annotations and functional information about the identified proteins is given.

Secondary Nucleotide Databases

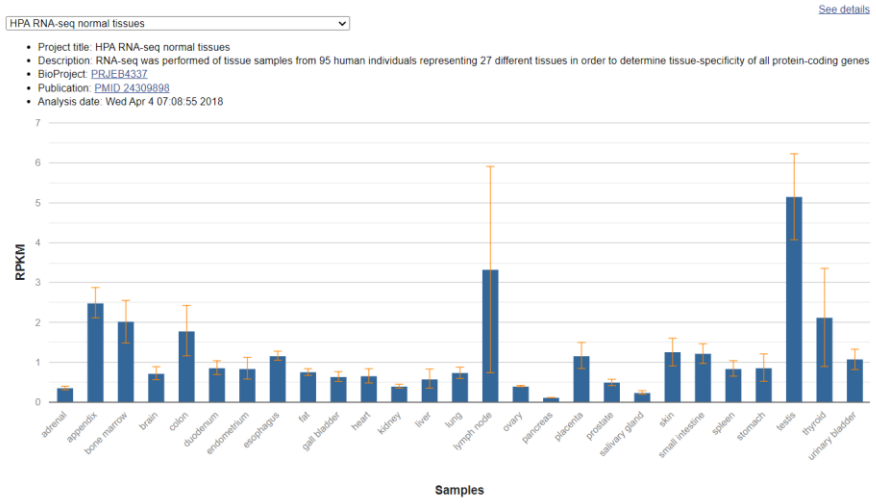
- **Gene:** concentrate on individual genes and their related data, including sequences, structure, function, and expression.
- **Genome:** Genome secondary databases focus on entire genomes, providing comprehensive resources for the analysis and comparison of whole-genome sequences.
- **EST:** Expressed sequence tags. Maintains expressed sequence tags (ESTs) and short, single-pass reads from mRNA (cDNA)
- **STS:** Sequenced Tagged Site. Is a relatively short, easily PCR-amplified sequence (200 to 500 bp) which can be specifically amplified by PCR and detected in the presence of all other genomic sequences and whose location in the genome is mapped.
- **GSS:** Genome Survey Sequence. The GSS division of GenBank is similar to the EST division, with the exception that most of the sequences are genomic in origin, rather than cDNA (mRNA).

Gene

- https://www.ncbi.nlm.nih.gov/gene/?term=
- The Gene database is a resource of the National Center for Biotechnology Information (NCBI) that centralizes gene-related information into individual records.
- The NCBI Gene database has information on gene sequences, gene alleles and mutations, genomes, amino acid sequences for proteins, and much more genetic data on humans, as well as many other animal species.
- RefSeq and other primary databases provide data to Gene database. They integrate, annotate, and often reanalyze data from primary databases like GenBank, EMBL, and DDBJ to provide additional layers of information, context, and utility for researchers.
- RefSeq provides a comprehensive, integrated, non-redundant set of sequences, including genomic DNA, transcripts, and proteins. It also provides sequences curated giving annotated functional elements, gene expression studies.

Gene Database Results

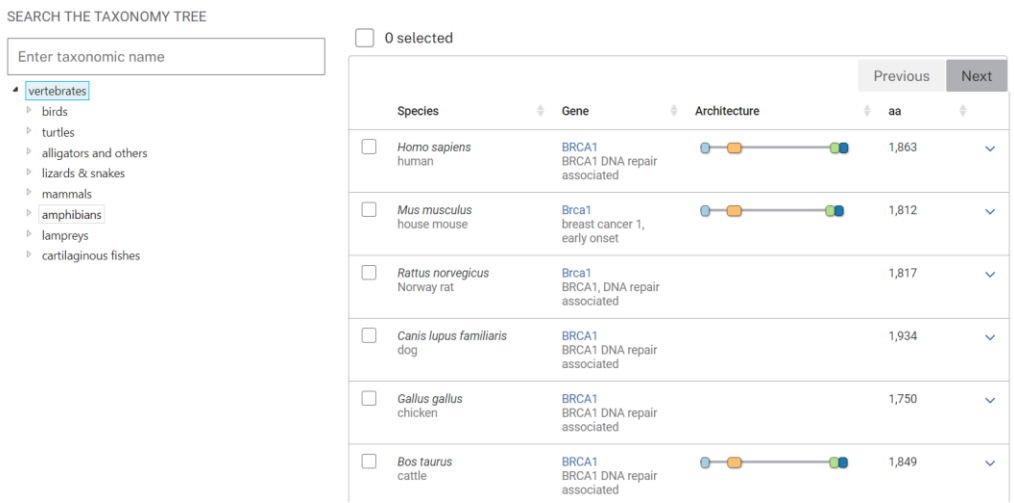
- ❑ Summary: Pathophysiological role of gene, disease mechanism and molecular pathways
- ❑ Expression: gene expression of gene in different tissues

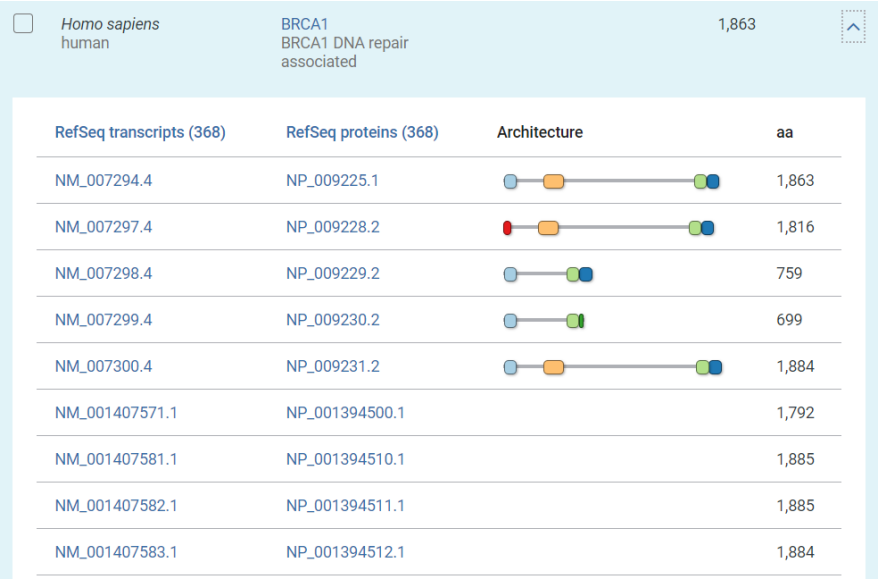


Sample	BioSample	RPKM	Count	Links
adrenal	3 samples	0.355 ± 0.042	29579	SRA, BioSample
appendix	3 samples	2.495 ± 0.381	181185	SRA, BioSample
bone marrow	4 samples	2.018 ± 0.532	432170	SRA, BioSample
brain	3 samples	0.73 ± 0.163	72250	SRA, BioSample
colon	5 samples	1.794 ± 0.632	396724	SRA, BioSample
duodenum	2 samples	0.867 ± 0.173	44311	SRA, BioSample
endometrium	3 samples	0.852 ± 0.271	89581	SRA, BioSample
esophagus	3 samples	1.166 ± 0.113	160671	SRA, BioSample

Gene Database Results

- <u>Orthologs:</u> comparison of searched gene in different species with possible common ancestry. For each species, all transcript variants and the transcripts coding for a protein are provided. No. of Amino acids, Architecture (1 transcript containing different exons part (amino acids) coding for different proteins).





Gene Database Results

- ❑ Gene Table: Provides access to ortholog sets and transcripts table
- ❑ Transcript Table: For the particular species that is searched, it provides information about how many different mRNA molecules can be transcribed from that particular gene, each transcript variant eventually leading to distinct protein isoforms. The variants arise due to alternative splicing, alternative promoter usage, post-transcriptional modification, etc.

Download | Edit genes | Select columns | 1 Gene | Rows per page 20 | 1-1 of 1

Gene ID	Symbol	Gene name	Scientific name	Common Name	Gene type	Transcripts	Ortholog set	Input
672	BRCA1	BRCA1 DNA repair associated	Homo sapiens (human)	human	protein-coding	368	672	672

Transcripts and Proteins

BRCA1 – BRCA1 DNA repair associated

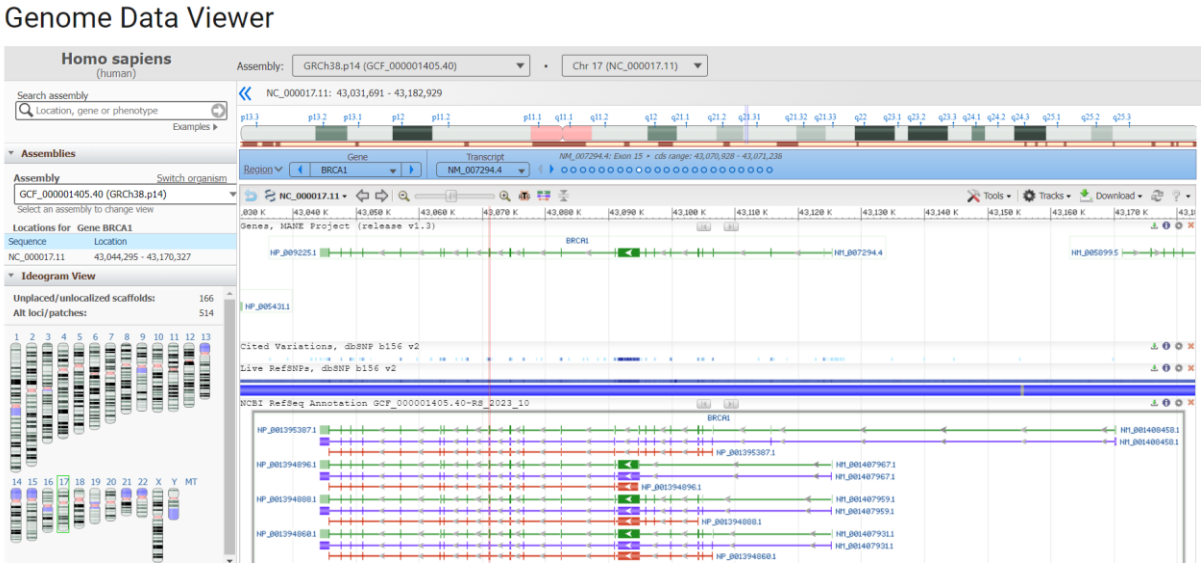
Homo sapiens (human)

Download | Select columns | 368 Transcripts 368 selected

Gene ID	Gene symbol	Transcript	Length (nt)	Protein	Length (aa)	Protein name	Isoform
672	BRCA1	NM_001408458.1	3785	NP_001395387.1	712	breast cancer type 1 susceptibil...	116
672	BRCA1	NM_001407967.1	6390	NP_001394896.1	1566	breast cancer type 1 susceptibil...	85
672	BRCA1	NM_001407959.1	6851	NP_001394888.1	1736	breast cancer type 1 susceptibil...	78
672	BRCA1	NM_001407931.1	6803	NP_001394860.1	1774	breast cancer type 1 susceptibil...	59
672	BRCA1	NM_001407747.1	6926	NP_001394676.1	1815	breast cancer type 1 susceptibil...	37
672	BRCA1	NM_001407962.1	6902	NP_001394891.1	1735	breast cancer type 1 susceptibil...	80
672	BRCA1	NM_001408512.1	3473	NP_001395441.1	592	breast cancer type 1 susceptibil...	145

Gene Database Results

- ❑ <u>Genomic Context:</u> Chromosome location, reference assembly, Genomic regions, transcripts, and products



- ❑ <u>Bibliographic Data:</u> Related research articles
- ❑ <u>Phenotypes:</u> Associated disease conditions provides information on diseases in which the gene is related

Gene Database Results

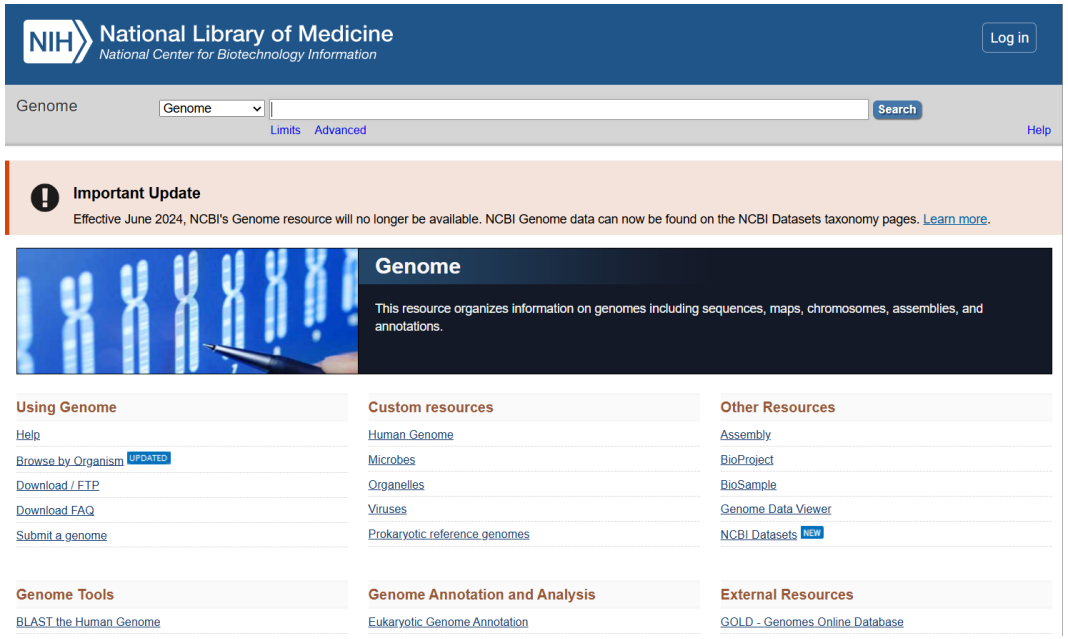
- ❑ Variations: ClinVar, dbVar
- ❑ Pathways: Cellular and molecular pathways in which they are involved. E.g., cell cycle, cell proliferation, reproduction
- ❑ Interactions: Interaction of searched gene with other genes
- ❑ General Gene Information:

→ **Gene Ontology:** info about pathways (molecular, cellular) pathway in which it is involved. E.g., DNA repair, checkpoint in cell cycle

- ❑ General Protein Information: info related to proteins, enzymes, activities of that particular gene
- ❑ NCBI RefSeq: information of each protein coded by transcript variants of that particular gene

Genome

- Home - Genome - NCBI (nih.gov)
- Genome is a Secondary nucleotide databases of NCBI
- Whole genome sequence of an organism can be obtained



Genome List

Genome

Download a genome data package including genome, transcript and protein sequence, annotation and a data report

Selected taxa

Escherichia coli (E. coli) Enter one or more taxonomic names

Filters

Download Select columns 278,189 Genomes Rows per page 20 1-20 of 278,189

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action	Level	Release ...	WGS accession	Scaffolds count
ASM584v2	GCA_000005845.2	GCF_000005845.2	Escherichia coli str. K-12 substr...	K-12 substr. MG165...	NCBI RefSeq Submitter		Complete	Sep, 2013		1
ASM886v2	GCA_000008865.2	GCF_000008865.2	Escherichia coli O157:H7 str. S...	Sakai substr. RIMD ...	NCBI RefSeq Submitter		Complete	Jun, 2018		3
ASM285371v1	GCA_002853715.1	GCF_002853715.1	Escherichia coli (E. coli)	14EC020 (strain)	NCBI RefSeq Submitter		Complete	Jan, 2018		3
ASM1326v1	GCA_000013265.1	GCF_000013265.1	Escherichia coli UTI89	UTI89 (strain)	NCBI RefSeq Submitter		Complete	Apr, 2006		2

Level of assembly
(complete, scaffold or contig)

Genome Assembly

Genome assembly ASM584v2 reference

Download datasets URL FTP

Actions

NCBI RefSeq assembly	GCF_000005845.2
Submitted GenBank assembly	GCA_000005845.2
Taxon	Escherichia coli str. K-12 substr. MG1655
Strain	K-12 substr. MG1655
Submitter	Univ. Wisconsin
Date	Sep 26, 2013

View the legacy Assembly page

Assembly statistics

	RefSeq	GenBank
Genome size	4.6 Mb	4.6 Mb
Total ungapped length	4.6 Mb	4.6 Mb
Number of chromosomes	1	1
Number of scaffolds	1	1
Scaffold N50	4.6 Mb	4.6 Mb
Scaffold L50	1	1
Number of contigs	1	1
Contig N50	4.6 Mb	4.6 Mb
Contig L50	1	1
GC percent	51	51
Assembly level	Complete Genome	Complete Genome
View sequences	view RefSeq sequences	view GenBank sequences

Annotation details

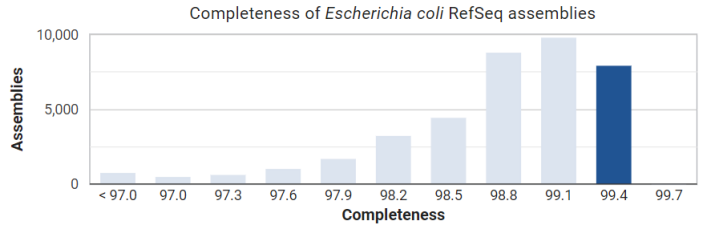
	RefSeq	GenBank
Provider	NCBI RefSeq	Univ. Wisconsin
Name	Annotation submitted by NCBI RefSeq	Annotation submitted by Univ. Wisconsin
Date	Mar 9, 2022	Nov 8, 2022
Genes	4,639	4,639
Protein-coding	4,288	4,288
	View RefSeq annotation	View GenBank annotation

Quality analysis

CheckM analysis (v1.2.2)

Completeness: 99.48% (89th Percentile, dark blue bar)

Contamination: 0.15%



Genome Annotation

Genome Annotation

Genes annotated on Escherichia coli str. K-12 substr. MG1655 ASM584v2 (GCF_000005845.2)

Annotation Name: Annotation submitted by NCBI RefSeq (March 9, 2022)

Filters

Download | Select columns | 4,639 Genes | Rows per page 20 | 1-20 of 4,639

n	Name	Symbol	Locus Tag	Gene ID	Gene type	Proteins	Length (aa)	Action
	thr operon lea...	thrL	b0001	944742	protein-coding	1 NP_414542.1	21	
	fused aspartat...	thrA	b0002	945803	protein-coding	1 NP_414543.1	820	
	homoserine ki...	thrB	b0003	947498	protein-coding	1 NP_414544.1	310	
	threonine synt...	thrC	b0004	945198	protein-coding	1 NP_414545.1	428	
	DUF2502 dom...	yaaX	b0005	944747	protein-coding	1 NP_414546.1	98	

Gene | Gene | Advanced

Full Report | Send to:

thrL thr operon leader peptide [*Escherichia coli str. K-12 substr. MG1655*]

Download Dataset

Gene ID: 944742, updated on 2-May-2024

Summary

Gene symbol	thrL
Gene description	thr operon leader peptide
Primary source	ASAP:ABE-0000006
Locus tag	b0001
See related	ECOCYC:EG11277
Gene type	protein coding
RefSeq status	PROVISIONAL
Organism	Escherichia coli str. K-12 substr. MG1655 (strain: K-12, substrain: MG1655)
Lineage	Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia
Also known as	ECK0001
Summary	The ThrL leader peptide controls by attenuation the expression of the thrLABC operon, which encodes four out of the five enzymes of threonine biosynthesis pathway, in response to the threonine and isoleucine levels . [More information is available at EcoCyc: EG11277].

NEW **Try the new Gene table**
Try the new Transcript table

Protein | Protein | Advanced

GenPept

thr operon leader peptide [Escherichia coli str. K-12 substr. MG1655]

NCBI Reference Sequence: NP_414542.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       NP_414542                 21 aa            linear   CON 09-MAR-2022
DEFINITION  thr operon leader peptide [Escherichia coli str. K-12 substr.
            MG1655].
ACCESSION   NP_414542
VERSION     NP_414542.1
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
DBSOURCE    REFSEQ: accession NC_000913.3
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Pseudomonadota; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (residues 1 to 21)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
  TITLE     Escherichia coli K-12: a cooperatively developed annotation
            snapshot--2005
  JOURNAL   Nucleic Acids Res. 34 (1), 1-9 (2006)
```

Genome Search Results

- **Genomes List:**

  (Obtained from different strains, samples of same species and derived from different experimental methodologies)

  Assembly, GenBank, RefSeq ID, Scientific name, Modifier (strain), Annotations (submitter), level(Indicates the completeness of the genome assembly [e.g., complete, scaffold, contig]), scaffold count (No. of scaffolds, gives an idea about gaps in the sequence and hence completeness)

→ **Assembly:**

- **Reference genomes:** (green tick)

  High quality, well annotated genome assemblies that serve as a standard for the species. They are derived from multiple sources to create a consensus sequence. This reference genome provides a baseline for comparison which facilitates genetic variation such as mutations, insertions, deletions, SNPs/single nucleotide polymorphisms. Reference genomes are used in a wide range of applications, including gene discovery, comparative genomics, evolutionary studies, and medical research.

- **Other genome assemblies**: (remaining)

  Help in identifying genetic variation, evolutionary relationships and functional genomics when compared with reference genome.

Genome Search Results

- **Exploring Assembly (Reference or others):**
  - → **Assembly statistics:** Genome size, GC%, Assembly level, View Sequences (RefSeq or GenBank)
  - → **Sample details**
  - → **Quality Analysis:** % of contaminants like adapters, primers given which gives an idea about completeness of sequencing
  - → **Annotations details:** View annotations of genes present in the genome (RefSeq, GenBank)
    - Genome location and name of gene. Each gene can be explored in Gene secondary nucleotide database
    - Protein: Protein coded by that particular gene. Each protein can be explored in Genpept database