



The Genome Database

Organism-centered listing of available genomic sequence records and projects

<https://www.ncbi.nlm.nih.gov/genome>

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

Scope

The Genome database at NCBI now uses a biologically-relevant schema to organize its records at the species-level. Records from the Genome database provide summary of the organisms, genomic information of available chromosomes and other replicons, and links to genomic sequences (in different stages of sequencing), their assembly, as well as annotation. Genome records also provide links to relevant entries in the BioProject and BioSample databases, which detail the research projects and biological specimens from which genomes were isolated and sequenced. Additionally, genome records connect to relevant records from the Assembly database, where individual records representing different versions of genomic assemblies are available along with download links. When relevant and available, genome records also provide links to data from sub-species, different strains/isolates, or from different sequencing attempts.



Links to Data and Help Documents

NCBI News on the release of the current schema:

<https://www.ncbi.nlm.nih.gov/About/news/17Nov2011.html>

Genome submission:

<https://www.ncbi.nlm.nih.gov/genbank/submit/>

See "Genomes" and "WGS" pull-down menu

Genome browser tool:

<https://www.ncbi.nlm.nih.gov/genome/browse/>

Genome FTP site:

<ftp://ftp.ncbi.nlm.nih.gov/genomes/>

Description of the updated genome FTP site:

<ftp://ftp.ncbi.nlm.nih.gov/genomes/README.txt>

BioProject submission site:

<https://submit.ncbi.nlm.nih.gov/subs/bioproject/>

Requires a My NCBI login

Accessing the Data

The Genome homepage (<https://www.ncbi.nlm.nih.gov/genome/>) provides access to records in the database through text searches (A) or browsing via the "Browse by Organism" link (B). A link to the genomes FTP site is available for data downloading purpose, with the FAQ page link below (<https://www.ncbi.nlm.nih.gov/genome/doc/ftpfaq/>) providing details on the organism-centric and assembly-based directory structure (C). An NCBI news entry summarizing this change is also available online at: <https://www.ncbi.nlm.nih.gov/news/08-26-2014-new-genomes-FTP-live/>

The screenshot shows the NCBI Genome Database homepage. At the top, there is a search bar with a dropdown menu set to "Genome" and a search button. Below the search bar, there is a navigation menu with "Limits" and "Advanced" options. The main content area features a large banner with the word "Genome" and a description: "This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations." Below the banner, there are several sections: "Using Genome" (with links for Help, Browse by Organism, Download / FTP, Download FAQ, and Submit a genome), "Genome Tools" (with links for BLAST the Human Genome, Microbial Nucleotide BLAST, and TaxPlot), "Custom resources" (with links for Human Genome, Microbes, Organelles, Viruses, and Prokaryotic reference genomes), "Genome Annotation and Analysis" (with links for Eukaryotic Genome Annotation, Prokaryotic Genome Annotation, and PASC), "Other Resources" (with links for Assembly, BioProject, BioSample, Map Viewer, and Protein Clusters), and "External Resources" (with links for GOLD, Ensembl Genome Browser, Bacteria Genomes at Sanger, and Large-Scale Genome Sequencing). Annotations A-F are placed on the page: A is on the search bar, B is on "Browse by Organism", C is on "Download FAQ", D is on "Genome Annotation and Analysis", E is on "Submit a genome", and F is on "External Resources".

The "Custom Resources" and "Other Resources" sections (D) provide links to other web pages and databases maintained by NCBI. The "Genome Tools" and "Genome Annotation and Analysis" sections (E) provide links to analysis tools available from NCBI. The "External Resources" section (F) lists links to representative resources outside NCBI.

Browsing or Searching for Genome Records

The “Browse by Organism” page provides an easy way to locate a specific genome record using the text input box. The example shows a display filtered by “pinus” (A), which retrieves three Overviews (B). The “Eukaryotes” tab provides details of projects with a summary on each assembly (C). Accession number in the Assembly column (D) links to the full report in the Assembly database with more details on that assembly. Filters button provides preset criteria to refine large list, and Columns button allows customization of the table display (E).

Genome > Genome Information by Organism <https://www.ncbi.nlm.nih.gov/genome/browse/>

pinus

Overview (3); Eukaryotes (3); Prokaryotes (0); Viruses (0); Plasmids (0); Organelles (21)

#	Organism Name	Organism Groups	Size(Mb)	Chromosomes	Organelles	Plasmids	Assemblies
1	Pinus lambertiana	Eukaryota;Plants;Land Plants	27.602.7	-	-	-	1
2	Pinus sylvestris	Eukaryota;Plants;Land Plants	0.965624	-	-	-	1

Overview (3); Eukaryotes (3); Prokaryotes (0); Viruses (0); Plasmids (0); Organelles (21)

#	Organism Name	Organism Groups	Strain	BioSample	BioProject	Assembly	Level	Size(Mb)	GC%	Replicons	WGS	Scaffolds	CDS	Release Date	FTP
1	Pinus taeda	Eukaryota;Plants;Land Plants		SAMN02981512	PRJNA17445C	GCA_000404065.3		22,103.6	34.80		APFE03	1,760,464		06-Aug-2015	G
2	Pinus sylvestris	Eukaryota;Plants;Land Plants		SAMEA18502918	PRJEB18435	GCA_900143225.1		0.965624	46.30		FRDG01	224		02-Jan-2017	G
3	Pinus lambertiana	Eukaryota;Plants;Land Plants		SAMN03354659	PRJNA17445C	GCA_001447015.2		27,602.7	3.10		LMP01	4,253,097		17-Nov-2015	G

The search box allows direct searching with query terms, and the example search with a generic term “rodent” retrieves almost 50 records (E). The “Limits” page (F) provides quick access to a list of preset attributes for use to restrict the search results.

The “Advanced” page (G) provides access to indexing fields through a pull-down menu (H) as well as terms indexed in each field and their record counts (in parentheses) through the “Show index list” link (I). The page automatically enters selected terms into the search box (J) along with selected Boolean operators (AND as the default). The “Edit” link (K) activates the search box to allow manual input. The example search combines search history #5 (representing “rodent”) with the fielded term “has annotation [properties]” to retrieve genome records with rodent text term as well as feature annotation. The “Add to history” link (L), when clicked, adds a new entry to the history section and shows how many genome records a given query retrieves.

Genome

Create alert Limits Advanced

See also 147 organelle- and plasmid-only records

Display Settings: Summary, 20 per page

Search results
Items: 1 to 20 of 49

1. [Mus musculus](#)
The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced
Kingdom: Eukaryota; Subgroup: Mammals
Sequence data: genome assemblies:22
Haploid chromosomes: 21; Organelles: 1
Date: 2002/05/15
ID: 52

2. [Peromyscus maniculatus](#)
The North American deer mouse is a widespread rodent with numerous subspecies
Kingdom: Eukaryota; Subgroup: Mammals
Sequence data: genome assemblies:1
Date: 2013/11/26
ID: 11397

Genome Advanced Search Builder

or

Builder

has annotation (35952)
has chromosome (12364)
has organelle (10515)
has plasmid (1422)
has wgs (23065)
organelle only (9571)
pathogen (1211)
plasmid only (65)

History www.ncbi.nlm.nih.gov/genome/advanced

Search	Add to builder	Query	Items found	Time
#6	<input type="button" value="Add"/>	Search #5 AND "has annotation"[Properties]	39	16:55:02
#5	<input type="button" value="Add"/>	Search rodent	49	16:39:41
#1	<input type="button" value="Add"/>	Search human[orgn]	1	14:35:45

Data Available from a Genome Record

The title of a genome record links to its Overview display as shown below. This display contains a list of commonly used FTP links at the top (A) to facilitate data download from the reorganized and assembly-based genomes ftp site.

The Overview itself provides:

- A brief description of the organism (B)
- A set of selected publications (manually closed, C)
- A summary table for the reference genome (D, shown in part)
- A diagram for the ideogram (E) with each chromosome linked to the NCBI MapViewer display
- The accession for a chromosome (F) in the "Reference genome" table links to the sequence of that chromosome in the Nucleotide database to allow more detailed examination of the annotation in supported formats.
- Links to graphical presentation of the genome through Genome Data Viewer (F) and BLAST search against this genome (G)
- Relevant entries from other NCBI databases such as Assembly and BioProject in the "Related information" section (H)

Mus musculus (house mouse)
 Reference genome: **Mus musculus (assembly GRCm38.p6)**
 Download sequences in FASTA format for genome, transcript, protein
 Download genome annotation in GFF, GenBank or tabular format
 BLAST against Mus musculus genome
 All 22 genomes for species:
 Browse the list
 Download sequence and annotation from RefSeq or GenBank

Display Settings: Overview Send to: ID: 52

Organism Overview: Genome Assembly and Annotation report [22]; Organelle Annotation Report [20]

Mus musculus (house mouse)
 The laboratory mouse is a major model organism for basic mammalian biology, human disease, and genome evolution, and its genome has been sequenced
 Lineage: Eukaryota[2606]; Metazoa[863]; Chordata[370]; Craniata[362]; Vertebrata[362]; Euteleostomi[356]; Mammalia[151]; Eutheria[146]; Euarchontoglires[67]; Glires[36]; Rodentia[34]; Myomorpha[21]; Murioidea[20]; Muridae[9]; Murinae[7]; Mus[4]; Mus[3]; Mus[1]; Mus[1]

The mouse is one of the major organisms for modeling human disease and comparative genome analysis. There are over 450 inbred strains of mice, providing a wealth of different genotypes and phenotypes for genetic and other studies. In addition, thousands of spontaneous, radiation- or chemically-induced, and transgenic mutants provide potential models [More...](#)

Summary
 Sequence data: genome assemblies: 22; sequence reads: 106 (see Genome Assembly and Annotation report)
 Statistics: median total length (Mb): 2671.82
 median protein count: 61940
 median GC%: 42.4891

NCBI Annotation Release: 105

Publications

Representative (genome information for reference and representative genomes)

Reference genome:
 Mus musculus GRCm38.p6
 Submitter: Genome Reference Consortium

Loc	Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr 1		NC_000067.6	CM000064.2		195.47	41.3	4,731	-	37	2,031	2,667	579
Chr 2		NC_000068.7	CM000065.2		182.11	42.2	6,282	-	8	2,783	3,491	809
Chr X		NC_000088.7	CM001013.2		171.03	39.7	3,010	-	17	933	2,291	912
Chr Y		NC_000087.7	CM001014.2		91.74	39.3	335	-	-	119	423	85
MT		NC_005089.1	AY172335.1		0.02	38.7	13	2	22	-	37	-
Un		-	-	-	93.44	43.6	2,471	2	10	-	-	-

Chromosomes

External Resources

NCBI Resources
 Genome Data Viewer
 Map Viewer
Tools
 BLAST Genome
Related information
 Assembly
 BioProject
 Gene
 Components
 Protein
 PubMed
 Taxonomy

Genome Data Viewer
 Mus musculus: GRCm38.p6 (GCF_000001635.26) Chr 1 (NC_000067.6): 1 - 195,471,971
 Ideogram View
 Exon Navigator
 Genes: NCBI Mus musculus Annotation Release 105
 RNA-seq exon coverage, aggregate (filtered), NCBI
 RNA-seq intron-spanning reads, aggregate (filtered)
 RNA-seq intron features, aggregate (filtered), NCBI

NCBI/BLAST/blastn suite
Mus musculus (house mouse) Nucleotide BLAST
 BLASTn programs search nucleotide databases using a nucleotide query. more...
 Enter Query Sequence
 Enter accession number(s), gi(s), or FASTA sequence(s)
 Or, upload file
 Choose Search Set
 Database: Genome (all assemblies top-level, Annotation Release 105) 12685 sequences
 Exclude Optional: Models (XM/XP)
 Entrez Query: Optional

Display formats (I) linked at the top of the genome record highlight other types of data available for this genome.

Prokaryotic Genomes

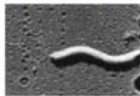
The display of prokaryotic organisms' genome records also contains a set of FTP links at the top (A) to allow quick download of relevant sequence data files. The Overview display (B) below is similar to that for eukaryotic organisms. The Dendrogram section (C) at the end contains a BLAST-based distance tree to address the relatedness of these strains or subspecies and highlights the complexity of this relationship, with the reference genome marked in yellow (D). Clicking a node highlights its branch (green, E). The context menu (activated upon hovering, F) provides functions for tree manipulation and access to specific genomes under the node, organized in a table similar to "Genome Assembly and Annotation report" (G).

Helicobacter pylori
Reference genome: Helicobacter pylori 26695

Download sequences in FASTA format for genome, protein
Download genome annotation in GFF, GenBank or tabular format
BLAST against Helicobacter pylori genome, protein
All 878 genomes for species:
Browse the list
Download sequence and annotation from RefSeq or GenBank

Display Settings: Overview

Organism Overview: Genome Assembly and Annotation report [878]; Genome Tree report [653]; Plasmid Annotation Report [52] ID:169



Helicobacter pylori
Causes peptic ulcers

Lineage: Bacteria[20550]; Proteobacteria[6298]; Epsilonproteobacteria[123]; Campylobacterales[103]; Helicobacteraceae[44]; Helicobacter[30]; Helicobacter pylori[1]

Helicobacter. This genus consists of organisms that colonize the mucosal layer of the gastrointestinal tract or are found enterohepatically (in the liver). *Helicobacter pylori*. It was only recently discovered (1983) by two Australians (Warren and Marshall) that this organism was associated with peptic ulcers. It is one of [More...](#)

Summary

Sequence data: genome assemblies: 878; sequence reads: 6 (See Genome Assembly and Annotation report)
Statistics: median total length (Mb): 1.6355
median protein count: 1453
median GC%: 38.9

Publications

Representative (genome information for reference and representative genomes)

Reference genome:

- Helicobacter pylori 26695

Submitter: TIGR
Human Pathogen
Morphology: Gram Negative, Shape:Spirilla, Motility:Yes
Environment: OxygenReq:Aerobic, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated
Phenotype: Disease:Gastric inflammation and peptic ulcer disease

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_000915.1	AE000511.1	1.67	38.9	1,445	7	36	1,555	65

Feedback

Please address comments and feedback to info@ncbi.nlm.nih.gov

Dendrogram (based on genomic BLAST)

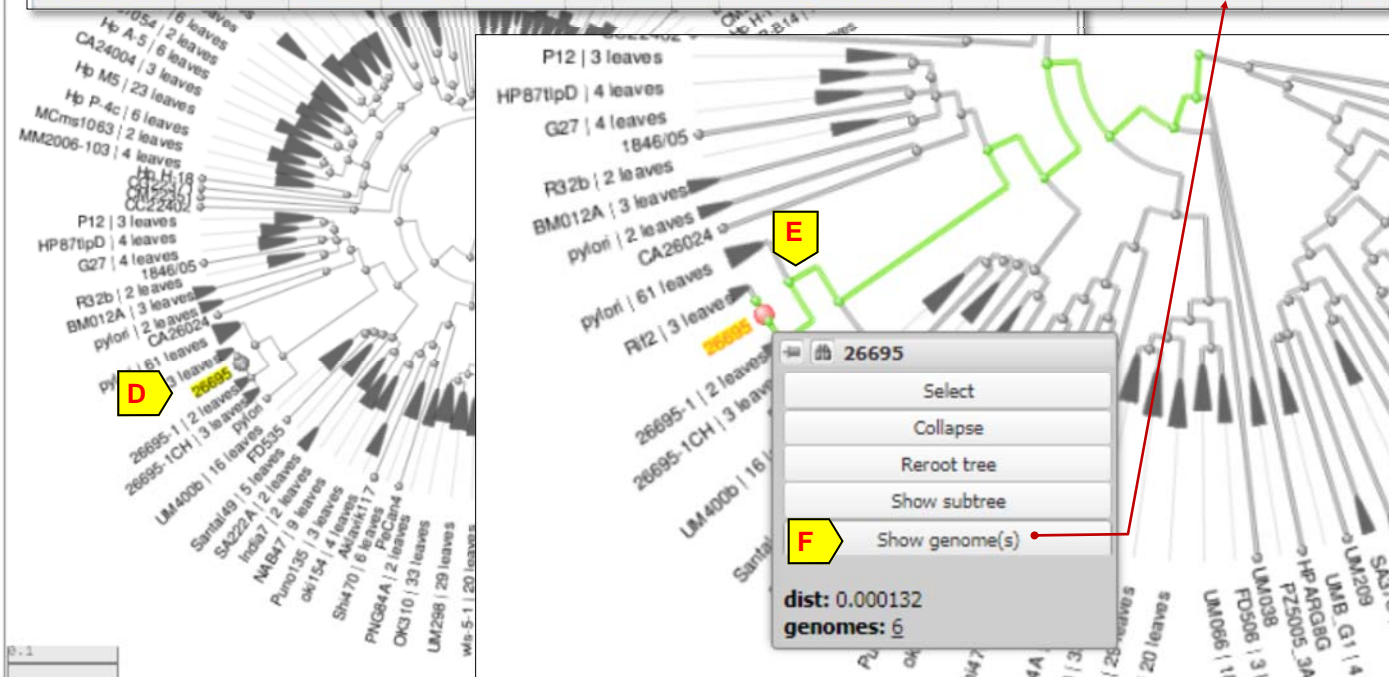
Find: all

Helicobacter pylori

Levels: All Complete [101] Chromosome [28] Scaffold [138] Contig [611]

Download table

Organism/Name	Strain	CladeID	BioSample	BioProject	Assembly	Level	Size (Mb)	GC%	Replicons	WGS	Scaffolds	Gene	Protein	Release Date	Modify Date	FTP
Helicobacter pylori 26695	26695	19252	SAMN02602995	PRJNA233	GCA_000008525.1	●	1.66787	38.90	chromosome_NC_000915.1 AE000511.1	-	-	1555	1445	1997/08/08	2016/08/02	◆◆
Helicobacter pylori J99	J99	19252	SAMN02602990	PRJNA234	GCA_000008785.1	●	1.64383	39.20	chromosome_NC_000921.1 AE001439.1	-	-	1607	1469	1999/01/12	2017/03/22	◆◆



dist: 0.000132
genomes: 6