Mini Courses NCBI Resources one course at a time

Microbial Genomes Quick Start

An NCBI Mini-Course

Microbial Genomes Resources presents public data from prokaryotic genome sequencing projects. The sequence collection contains data from finished genomes as well as draft assemblies.

In this mini-course, we will learn:

1. To navigate through the Microbial Genomes resources and find:
   - listings of sequencing projects for microbial organisms and their sequencing status
   - an individual organism's sequencing projects information, summary of its biology, and quick links to access sequence and annotation.
2. To perform comparative analysis of protein sequences from different genomes, including:
   - functional assignments using the Clusters of Orthologous Groups (COGs) database
   - three-way homology-based comparison using the TaxPlot tool
   - pair-wise homology-based comparison using the GenePlot tool
   - taxonomic distribution of protein homologs using the TaxMap tool
   - proteins that may be specifically expressed in a genome or lost from a genome compared to a closer species

The course will address practical discovery questions such as 'Are there identifiable genes in microbial genomes that may be horizontally transferred?' and 'What are the differences between closely-related pathogenic and non-pathogenic bacteria?'

During the first hour, an overview will be given using the Escherichia coli O157:H7 Sakai genome, followed by an hour of hands-on session to practice using the Rickettsia prowazekii str. Madrid E genome. The following handout contains the screenshots of the overview.

URL: http://www.ncbi.nlm.nih.gov/Class/minicourses/microbialgenome.html

Problem 1

Access the Microbial Genomes Page (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). How many complete microbial genome sequences are available? How many are in progress and how many are in the assembly stage? From the "Organism Group" pull down menu, select Firmicutes. The complete genome sequences are available for how many Firmicutes?

Go back to the Complete genomes Page. Select "All Organisms" in the "Organism Group" pull down menu . From the resulting list, select the organism Escherichia coli O157:H7 str. Sakai. How many strains of Escherichia coli have been completely sequenced? Is the Sakai strain pathogenic in humans? Name a few differences between the pathogenic and non-pathogenic strains. For example, a type III secretion system, the locus of enterocyte effacement, numerous toxins and adhesins, as well as fimbrial gene clusters and iron uptake systems are found in pathogenecity islands. Click on the RefSeq accession number NC_002695.

From the resulting table, access the entry NC_002695 in the Nucleotide database. Access the Protein link. How many protein coding genes have been annotated on the genome? Go back to the Genomes Page. Access the Protein table that contains locations and detailed information such as the strand, product name, FASTA sequence etc. from the Protein coding link. Also, access and download the RNA table through the Structural RNA link. How many genes encoding arginine tRNA are annotated on the Sakai strain?

Return to Genome page. Click on the gene link for thrA in the genome map located at the bottom of the page to access the Gene database. What conserved domains are found in the protein? What is the pathway in which it is involved?

Go back to the Genome page. Click on the COG link. It provides an assignment of each protein to one or more COG functional categories. Note the variations in the %genome category compared to the %genus category. What would be its importance? Download the protein table for the proteins assigned to Intracellular trafficking and secretion.

Go back to the genome page. Click on the Tax Plot link. Tax Plot allows a three-way comparison of three different proteomes, including proteins encoded on all chromosomes and plasmids To study proteins specific to pathogenic O157 strains compared to K12, select the non pathogenic strain Escherichia coli K12 and other pathogenic strain, O157:H7. Assign a functional classification to these proteins by selecting a category from the pull down menu. Compare the assignments to the Translation and Cell envelope biogenesis or Cell motility functional category. What do you infer? Go back to the Genome page.

Access the GenePlot link. GenePlot uses precalculated BLAST results to determine the top hits in two organisms. The best hits of each protein from two genomes are calculated and plotted based on chromosomal location. Download all best hits between the K12 and Sakai strain proteins. Open the file in Excel and select the gi column for the Sakai strain. Use Batch Entrez http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Protein to download all the proteins in Entrez. Now obtain the proteins that are annotated on the Sakai strain that have no homolog in the K12 strain by using the Entrez history option and subtracting these proteins from all proteins annotated on the Sakai strain.

Go back to the Genome page. Access the TaxMap link. It reports the taxonomic distribution of the top BLAST hits for each of the protein in the entire proteome of the Sakai strain by comparing to the non-redundant database (excluding closely related species) by BLAST similarity. Compare the map to that of K12 http://www.ncbi.nlm.nih.gov/sutils/taxik.cgi?gi=115 . What do you conclude? Identify different phages inserted in the Sakai strain genome. Select one region that is more similar to Viruses. Select the link for integrase (gi 15800926). This link gives the top 200 protein BLAST similarity hits for the query protein. Select the Best hits option. Which organisms are represented in the similar proteins? Select the COG0582 link at the top to access the tabular format of the proteins assigned to that COG. Access the Initial Version of the COG database by clicking on the link on the left. Access the Phylogenetic Patten link. What is the major functional category that is ubiquitously present? Go back to COG home page. Select the Phylogenetic Patterns search. Access the pre-calculated comparison between Escherichia coli and Buchnera sp. APS. Obtain a list of COGs that contain proteins only from O157 strain and not from K12 or Buchnera. What is the major type of proteins in this list?

NCBI

National Center for Biotechnology Information

National Library of Medicine National Institutes of Health

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search All Databases for Go

SITE MAP
Alphabetical List
Resource Guide

About NCBI
An introduction to NCBI

GenBank
Sequence submission support and software

Literature databases
PubMed, OMIM, Books, and PubMed Central

Molecular databases
Sequences, structures, and taxonomy

Genomic biology
The human genome, whole genomes, and related resources

What does NCBI do?

Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

Whole Genome Association

The NCBI Whole Genome Association (WGA) resource provides researchers with access to genotype and associated phenotype information that will help elucidate the link between genes and disease. For more information, click here to see the the WGA resource page and click here to read the press release.

100 Gigabases

GenBank and its collaborating databases, the European Molecular Biology Laboratory and the DNA Data Bank of Japan, have reached a milestone of 100 billion bases from over 165,000 organisms. See the press release or find more information on GenBank.

Hot Spots

- Assembly Archive
- Clusters of orthologous groups
- Coffee Break, Genes & Disease, NCBI Handbook
- Electronic PCR
- Entrez Home
- Entrez Tools
- Gene expression omnibus (GEO)
- Human genome resources
- Influenza Virus Resource
- Map Viewer
- dbMHC
- Mouse genome

NCBI

Genomic Biology

PubMed | All Databases | BLAST | OMIM | Books | TaxBrowser | Structure

Search Entrez for Go

NCBI

Site Map
guide to NCBI resources

Cancer Chromosomes
chromosomal abnormalities

Clusters of Orthologous Groups
analysis of complete genomes

Gene
gene-related information

Genome
complete genome sequences

GEO
gene expression data

HomoloGene
orthologs between pairs of organisms

Map Viewer
map and genome displays

RefSeq
the reference sequence project

Software

Genomic Biology

NCBI provides several genomic biology tools and resources, including organism-specific pages that include links to many web sites and databases relevant to that species. We invite you to explore the links provided on this page.

Genome Resources Workshops

Meeting	Slides	Date
Plant and Animal Genome XIV	PDF	January 2006

Announcements and Updates

New Genome Resources:

NEW A genome resource guide page is now available for *Saccharomyces*

Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
Rattus norvegicus	RGSC v3.4	July 6, 2006
Macaca mulatta (rhesus macaque)	1.1	June 23, 2006
Caenorhabditis elegans	WS150	May 11, 2006
Mus musculus	36.1	May 8, 2006
Drosophila melanogaster	4.3	April 19, 2006
Tribolium castaneum (red flour beetle)	1.1	April 18, 2006
Homo sapiens	36.1	March 9, 2006
Dictyostelium discoideum	1.1	November 22, 2005
Arabidopsis thaliana	TAIR6.0	November 21,

Genome Resources

- Entrez Genome
- Fungal Genomes Central NEW
- Genome Projects Database
  - Eukaryotic
  - Fungi
  - Insects
  - Mammals
  - Microbial
  - Plants
- Map Viewer
- Organelles
- Plant Genomes Central
- Viral Resources
  - Influenza Virus Resource
  - Retroviruses
  - Viral Genomes

Organism-Specific

G Genome Resources
B BLAST
M Map Viewer
P Genome Project DB

- Arabidopsis B M P
- Aspergillus G B M P
- Bee G B M P
- Beetle NEW G B M P
- Cat G B M P
- Chicken G B M P

NCBI ENTREZ Genome Project

connection information discovery

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: – All –

Tools: D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMaps; F - FTP; R - Publications.

* size is estimated, otherwise genome size is calculated based on existing sequences

383 Complete Microbial Genomes selected: [A] - 28, [B] - 355 save

Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
Acidobacteria bacterium Ellin345	B	Acidobacteria	5.65	58.4	1		CP000360.1	NC_008009.1	05/04/06	05/10/06	DOE Joint Genome Institute	T P D L S G X F
Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	01/17/06	Genoscope	T P C D L S G X F R
Aeropyrum pernix K1	A	Crenarchaeota	1.67	56.3	1		BA000002.3	NC_000854.1	06/26/99	12/03/05	NITE	T P C D L S G X F R
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59.0	2	2	AE008688.1	NC_003304.1	12/18/01	06/27/06	University of Washington	T P C D L S G X M F R
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59.0	2	2	AE007869.1	NC_003062.1	12/18/01	12/02/05	Cereon	T P C D L S G X M F R
Alcanivorax borkumensis SK2	B	Gammaproteobacteria	3.12	54.7	1		AM286690.1	NC_008260.1	07/24/06	07/28/06	Bielefeld Univ	T P D L S G X F R
Alkalilimnicola ehrlichei MLHE-1	B	Gammaproteobacteria	3.28	67.5	1		CP000453.1	NC_008340.1	09/12/06	09/14/06	DOE Joint Genome Institute	T P D L S G X F
Anabaena variabilis ATCC 29413	B	Cyanobacteria	7.07	41.4	1	3	CP000117.1	NC_007413.1	09/15/05	06/16/06	DOE Joint Genome Institute	T P C D L S G X M F
Anaeromyxobacter dehalogenans 2CP-C	B	Deltaproteobacteria	5.01	74.9	1		CP000251.1	NC_007760.1	01/27/06	03/31/06	DOE Joint Genome Institute	T P D L S G X F
Anaplasma marginale str. St. Maries	B	Alphaproteobacteria	1.2	49.8	1		CP000030.1	NC_004842.2	12/08/04	12/02/05	Washington State University	T P C D L S G X M F R
Anaplasma phagocytophilum HZ	B	Alphaproteobacteria	1.47	41.6	1		CP000235.1	NC_007797.1	02/21/06	03/30/06	TIGR	T P D L S G X M F R
Aquifex aeolicus VF5	B	Aquificae	1.59	43.3	1	1	AE000657.1	NC_000918.1	04/16/98	12/02/05	DIVERSA	T P C D L S G X F R
Archaeoglobus fulgidus DSM 4304	A	Euryarchaeota	2.18	48.6	1		AE000782.1	NC_000917.1	12/06/97	01/17/06	TIGR	T P C D L S G X F R
Aster yellows witches'-broom phytoplasma AYWB	B	Firmicutes	0.72	26.8	1	4	CP000061.1	NC_007716.1	01/11/06	05/18/06	Ohio State University	T P C D L S G X M F R
Azoarcus sp. EbN1	B	Betaproteobacteria	4.73	64.7	1	2	CR555306.1	NC_006513.1	11/20/04	01/17/06	Max Planck Institute	T P C D L S G X F R
Bacillus anthracis str. 'Ames Ancestor'	B	Firmicutes	5.5	35.2	1	2	AE017334.2	NC_007530.2	05/20/04	04/03/06	TIGR	T P C D L S G X M F
Bacillus anthracis str. Ames	B	Firmicutes	5.23	35.4	1		AE016879.1	NC_003997.3	05/02/03	12/02/05	TIGR	T P C D L S G X M F R
Bacillus anthracis str. Sterne	B	Firmicutes	5.23	35.4	1		AE017225.1	NC_005945.1	06/24/04	04/03/06	DOE Joint Genome Institute	T P C D L S G X M F
Bacillus cereus ATCC 10987	B	Firmicutes	5.43	35.5	1	1	AE017194.1	NC_003909.8	02/13/04	08/17/06	TIGR	T P C D L S G X M F R

NCBI ENTREZ Genome Project

connection information discovery

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy | OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: – All –

Legend: ☐ - wgs assembly; ☐ - no sequence available.

* size is estimated, otherwise genome size is calculated based on existing sequences sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

655 Microbial Genomes In-Progress selected: [A] - 28, [B] - 627 save

Organism	King	Group	Contigs	* Size	GC	Accession	Blast	Center
Acaryochloris marina	B	Cyanobacteria	-	-	-	-	-	Arizona State University
Acaryochloris sp. CCMEE 5410	B	Cyanobacteria	-	-	-	-	-	J. Craig Venter Institute
Acidianus brierleyi	A	Crenarchaeota	-	* 1.8	-	-	-	University of Copenhagen
Acidiphilium cryptum JF-5	B	Alphaproteobacteria	76	-	67.1	NZ_AAOO00000000	BLAST	DOE Joint Genome Institute
Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	-	* 4.15	-	-	-	TIGR
Acidothermus cellulolyticus 11B	B	Actinobacteria	8	-	66.9	NZ_AAOL00000000	BLAST	DOE Joint Genome Institute
Acidovorax avenae subsp. citrulli AAC00-1	B	Betaproteobacteria	76	-	68.6	NZ_AASX00000000	BLAST	DOE Joint Genome Institute
Acidovorax sp. JS42	B	Betaproteobacteria	65	-	66.1	NZ_AASD00000000	BLAST	DOE Joint Genome Institute
Acinetobacter baumannii	B	Gammaproteobacteria	-	-	-	-	-	Genoscope
Acinetobacter baumannii	B	Gammaproteobacteria	-	-	-	-	-	Genoscope
Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	-	-	-	-	-	Yale University
Acinetobacter baumannii ICU	B	Gammaproteobacteria	-	-	-	-	-	CNR - National Research Council
Actinobacillus actinomycetemcomitans HK1651	B	Gammaproteobacteria	-	* 2.11	-	-	-	University of Oklahoma
Actinobacillus pleuropneumoniae serovar 1	B	Gammaproteobacteria	-	* 2.19	37	-	-	University of Oklahoma
Actinobacillus pleuropneumoniae serovar 1 str. 4074	B	Gammaproteobacteria	140	* 2.29	41.4	NZ_AACK00000000	BLAST	University of Oklahoma
Actinobacillus pleuropneumoniae serovar 5	B	Gammaproteobacteria	-	* 2.18	37	-	-	University of Oklahoma
Actinobacillus pleuropneumoniae serovar 7	B	Gammaproteobacteria	-	* 2.08	37	-	-	University of Oklahoma
Actinobacillus succinogenes 130Z	B	Gammaproteobacteria	117	-	44.9	NZ_AAKC00000000	BLAST	DOE Joint Genome Institute

NCBI ENTREZ Genome Project connection information discovery

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: – All –

Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes/Chlorobi
Chlamydiae/Verrucomicrobia
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Epsilonproteobacteria
Spirochaetes
Thermotogae
Other Bacteria

Legend: ☐ - wgs assembly; ☐ - no sequence available.

* size is estimated, o ed on existing sequences sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

655 Microbial ted: [A] - 28, [B] - 627 save

Org		Group	Contigs	* Size	GC	Accession	Blast	Center
Acaryochloris marin		ranobacteria	-	-	-	-	-	Arizona State University
Acaryochloris sp. C		ranobacteria	-	-	-	-	-	J. Craig Venter Institute
Acidianus brierleyi		enarchaeota	-	* 1.8	-	-	-	University of Copenhagen
Acidiphilium cryptu		aproteobacteria	76	-	67.1	NZ_AAOO00000000	BLAST	DOE Joint Genome Institute
Acidobacterium cap		cidobacteria	-	* 4.15	-	-	-	TIGR
Acidothermus cellul		ctinobacteria	8	-	66.9	NZ_AAOL00000000	BLAST	DOE Joint Genome Institute
Acidovorax avenae AAC00-1		proteobacteria	76	-	68.6	NZ_AASX00000000	BLAST	DOE Joint Genome Institute
Acidovorax sp. JS42		proteobacteria	65	-	66.1	NZ_AASD00000000	BLAST	DOE Joint Genome Institute
Acinetobacter baum		aproteobacteria	-	-	-	-	-	Genoscope
Acinetobacter baum		aproteobacteria	-	-	-	-	-	Genoscope
Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	-	-	-	-	-	Yale University
Acinetobacter baumannii ICU	B	Gammaproteobacteria	-	-	-	-	-	CNR - National Research Council

NCBI ENTREZ Genome Project connection information discovery

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: Firmicutes

Legend: ☐ - wgs assembly; ☐ - no sequence available.

* size is estimated, otherwise genome size is calculated based on existing sequences sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

155 Microbial Genomes In-Progress selected: [A] - 0, [B] - 155 clear filter save

Organism	Ring	Group	Contigs	* Size	GC	Accession	Blast	Center
Alkaliphilus metalliredigenes QYMF	B	Firmicutes	263	* 4	36.6	NZ_AAKU00000000	BLAST	DOE Joint Genome Institute
Ammonifex degensii	B	Firmicutes	-	* 3	-	-	-	University of California Los Angeles (UCLA)
Anaerobranca gottschalkii	B	Firmicutes	-	-	-	-	-	Technische Universitat Hamburg - Harburg
Bacillus alcalophilus	B	Firmicutes	-	-	-	-	-	Genencor
Bacillus amyloliquefaciens	B	Firmicutes	-	-	-	-	-	Competence Network
Bacillus anthracis Tsiankovskii-I	B	Firmicutes	-	-	-	-	-	TIGR
Bacillus anthracis str. A1055	B	Firmicutes	42	* 5.37	35.3	NZ_AAEO00000000	BLAST	TIGR

NCBI ENTREZ Genome Project

connection information discove

PubMed | Nucleotide | Protein | Genome | Structure | PopSet | Taxonomy

Search Genome Project | Go | Clear

Organism info | Complete genomes | Genomes in progress

organism group – All –

Tools legend: T - TaxMap; P - ProTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - P

* size is estimated, otherwise genome size is calculated based on existing sequences

383 Complete Microbial Genomes selected: [A] - 28, [B] - 355

Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
Acidobacteria bacterium Ellin345	B	Acidobacteria	5.65	58.4	1		CP000360.1	NC_008009.1	05/04/06	05/10/06	DOE Joint Genome Institute	T P D L S G
Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	01/17/06	Genoscope	T P C D L S G
Aeropyrum pernix K1	A	Crenarchaeota	1.67	56.3	1		BA000002.3	NC_000854.1	06/26/99	12/03/05	NITE	T P C D L S G
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59.0	2	2	AE008688.1	NC_003304.1	12/18/01	06/27/06	University of Washington	T P C D L S G
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59.0	2	2	AE007869.1	NC_003062.1	12/18/01	12/02/05	Cereon	T P C D L S G
Alcanivorax borkumensis SK2	B	Gammaproteobacteria	3.12	54.7	1		AM286690.1	NC_008260.1	07/24/06	07/28/06	Bielefeld Univ	T P D L S G
Alkalilimnicola ehrlichei MLHE-1	B	Gammaproteobacteria	3.28	67.5	1		CP000453.1	NC_008340.1	09/12/06	09/14/06	DOE Joint Genome Institute	T P D L S G
Anabaena variabilis ATCC 29413	B	Cyanobacteria	7.07	41.4	1	3	CP000117.1	NC_007413.1	09/15/05	06/16/06	DOE Joint Genome Institute	T P C D L S G
Ehrlichia chaffeensis str. Arkansas	B	Alphaproteobacteria	1.18	30.1	1		CP000236.1	NC_007799.1	02/21/06	03/30/06	TIGR	T P D L
Ehrlichia ruminantium str. Gardel	B	Alphaproteobacteria	1.5	27.5	1		CR925677.1	NC_006831.1	02/04/05	05/23/06	CIRAD, France	T P C D L
Ehrlichia ruminantium str. Welgevonden	B	Alphaproteobacteria	1.52	27.5	1		CR767821.1	NC_005295.2	01/07/05	12/03/05	University of Pretoria, South Africa	T P C D L
Ehrlichia ruminantium str. Welgevonden	B	Alphaproteobacteria	1.51	27.5	1		CR925678.1	NC_006832.1	02/04/05	05/22/06	CIRAD, France	T P C D L
Enterococcus faecalis V583	B	Firmicutes	3.36	37.4	1	3	AE016830.1	NC_004668.1	03/29/03	12/02/05	TIGR	T P C D L
Erwinia carotovora subsp. atroseptica SCRI1043	B	Gammaproteobacteria	5.06	51.0	1		BX950851.1	NC_004547.2	07/21/04	12/02/05	Welcome Trust Sanger Institute	T P C D L
Erythrobacter litoralis HTCC2594	B	Alphaproteobacteria	3.05	63.1	1		CP000157.1	NC_007722.1	01/11/06	03/30/06	J. Craig Venter Institute	T P C D L
Escherichia coli 536	B	Gammaproteobacteria	4.94	50.5	1		CP000247.1	NC_008253.1	07/13/06	07/25/06	University of Goettingen	T P D L
Escherichia coli CFT073	B	Gammaproteobacteria	5.23	50.5	1		AE014075.1	NC_004431.1	12/10/02	04/04/06	University of Wisconsin-Madison, USA	T P C D L
Escherichia coli K12	B	Gammaproteobacteria	4.64	50.8	1		U00096.2	NC_000913.2	09/05/97	09/22/06	University of Wisconsin-Madison, USA	T P C D L
Escherichia coli O157:H7 EDL933	B	Gammaproteobacteria	5.62	50.3	1	1	AE005174.2	NC_002655.2	02/24/01	07/19/06	University of Wisconsin-Madison, USA	T P C D L
Escherichia coli O157:H7 str. Sakai	B	Gammaproteobacteria	5.59	50.5	1	2	BA000007.2	NC_002695.1	03/29/00	05/23/06	Osaka Univ.	T P C D L
Escherichia coli UTI89	B	Gammaproteobacteria	5.18	50.6	1	1	CP000243.1	NC_007946.1	04/05/06	04/18/06	Washington University (WashU)	T P D L
Escherichia coli W3110	B	Gammaproteobacteria	* 4.65		1		AP009048.1	AC_000091.1	01/23/06	04/12/06	Nara Institute of Science and Technology	T P D L
Francisella tularensis subsp. holarctica	B	Gammaproteobacteria	1.9	32.2	1		AM233362.1	NC_007880.1	03/02/06	03/30/06	Lawrence Livermore National Laboratory	T P D L
Francisella tularensis subsp. holarctica OSU18	B	Gammaproteobacteria	1.9	32.2	1		CP000437.1	NC_008369.1	09/20/06	09/25/06	Baylor College of Medicine	P L

NCBI ENTREZ Genome Project

connection information discovery

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome Project for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Sort by Send to

All: 1 | Prokaryotes: 1

Links

Genome Project > *Escherichia coli* > ***Escherichia coli O157:H7 str. Sakai*** project at GIRC

Resource Links

NCBI Resources
- BLAST genome
- FTP
- TaxPlot

Organism data in GenBank
- Genomic
- mRNA
- Protein
- WGS

Sequencing Centers

Enterohemorrhagic Escherichia coli

Project data

Lineage: *Bacteria*; *Proteobacteria*; *Gammaproteobacteria*; *Enterobacteriales*; *Enterobacteriaceae*; *Escherichia*; *Escherichia coli O157:H7 str. Sakai*

Photo: CDC

- Osaka Univ.
- Sanger
- U. Wisc

Related Resources
- ASAP
- BIGS
- BRS
- CGSC
- Colibri
- CyberCell
- E. coli GenExp
- E. coli TF
- E. coli Wiki
- ECCE
- ECDC
- ECOR/DEC
- EPD
- ERIC
- Echo/coliBase
- EcoCyc
- EcoGene

Genome Projects

Genome sequencing:

- *Escherichia coli K12* at Univ. Wisconsin [Complete]
- ***Escherichia coli O157:H7 str. Sakai*** (Project ID: **226**) at GIRC [Complete]
- *Escherichia coli E2348/69* at Sanger Institute [In progress]
- *Escherichia coli B171* at TIGR [Assembly]
- *Escherichia coli 53638* at TIGR [Assembly]
- *Escherichia coli 101-1* at TIGR [Assembly]
- *Escherichia coli 536* at University of Goettingen [Complete]
- *Escherichia coli W3110* at Nara Institute of Science and Technology [Complete]
- *Escherichia coli CFT073* at Univ. Wisconsin [Complete]
- *Escherichia coli E110019* at TIGR [Assembly]
- *Escherichia coli UTI89* at Washington University (WashU) [Complete]
- *Escherichia coli O157:H7 EDL933* at Univ. Wisconsin [Complete]
- *Escherichia coli 042* at Sanger Institute [In progress]
- *Escherichia coli B7A* at TIGR [Assembly]
- *Escherichia coli F11* at TIGR [Assembly]
- *Escherichia coli E22* at TIGR [Assembly]
- *Escherichia coli HS* at TIGR [Assembly]
- *Escherichia coli E24377A* at TIGR [Assembly]

Genome information:

Name	RefSeq	GenBank	Publications	Length (Mbp)	GC content	Proteins	RNAs	3D Struct	TaxMap	CDD	COG
Chromosome	NC_002695	BA000007	4	5.49845	50.5%	5254	141	✔	✔	✔	✔
Plasmid pO157	NC_002128	AB011549	1	0.092721	47.6%	85	--				
Plasmid pOSAK1	NC_002127	AB011548	1	0.003306	43.4%	3	--				

Publications:

- Hayashi T *et al.*, "Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12.", *DNA Res*, 2001 Feb 28;8(1):11-22

Escherichia coli O157:H7 str. Sakai

Escherichia coli. This organism was named for its discoverer, Theodore Escherich, and is one of the premier model organisms used in the study of bacterial genetics, physiology, and biochemistry.

This enteric organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora. *E. coli*, is capable of causing various diseases in its host, especially when they acquire virulence traits. Strains of *E. coli* can cause urinary tract infections, neonatal meningitis, and many different intestinal diseases, usually by attaching to the host cell and introducing toxins that disrupt normal cellular processes. Virulence proteins may be encoded on extrachromosomal plasmids or within bacteriophages and distinct DNA segments termed pathogenicity islands (PAIs). PAIs are likely to have been transferred horizontally and may even have integrated into the chromosome through bacteriophage or plasmid integration or transposition.

***Escherichia coli* O157:H7**. This strain is associated with Hamburger disease, which is caused by the contamination of meat products by enterohemorrhagic *E. coli* (EHEC). The identifier O157:H7 refers to the serotype of EHEC, and reflects the specific antigenic markers found on the surface of the cell. EHEC attaches and effaces to cells in the large intestine.

There are numerous differences that distinguish O157:H7 from K-12, hundreds of them associated with genomic islands in either strain, including at least 9 large PAIs in O157:H7 that encode virulence factors. A type III secretion system, the locus of enterocyte effacement, numerous toxins and adhesins, as well as fimbrial gene clusters and iron uptake systems are found in these PAIs. Pathogenicity genes are also found on the plasmid pO157.

***Escherichia coli* O157:H7 strain Sakai**. This strain of O157:H7 was isolated in a 1997 outbreak in Sakai, Japan.

Cellular features					Environment			Temperature	
Gram stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Opt. temp.	Range
-	Rod	Singles, Pairs		Yes		Facultative	Host-associated	37C	Mesophilic

Pathogenic in: Human **Disease:** Hemorrhagic colitis

NCBI | Entrez Genome | My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

Genome > Bacteria > Escherichia coli O157:H7 str. Sakai, complete genome Links

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli O157:H7 ; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 bp**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

NCBI Entrez Nucleotide

My NCBI [Sign In] [Register]

PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | OMIM | Books

Search Nucleotide for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display GenBank Show 5 Send to Hide: Sequence Lesser features

Range: from begin to end Reverse complemented strand Features: + Refresh

This record may be shown in a reduced form. Use 'Hide' checkboxes for control.

1: NC_002695. Reports Escherichia coli ...[gi:15829254] Links

Comment Features

```
LOCUS       NC_002695            5498450 bp    DNA     circular BCT 23-MAY-2006
DEFINITION  Escherichia coli O157:H7 str. Sakai, complete genome.
ACCESSION   NC_002695
VERSION     NC_002695.1  GI:15829254
PROJECT     GenomeProject:226
KEYWORDS    .
SOURCE      Escherichia coli O157:H7 str. Sakai
  ORGANISM  Escherichia coli O157:H7 str. Sakai
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (sites)
  AUTHORS   Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K.,
            Yokoyama,K., Han,C.-G., Ohtsubo,E., Nakayama,K., Murata,T.,
            Tanaka,M., Tobe,T., Iida,T., Takami,H., Honda,T., Sasakawa,C.,
            Ogasawara,N., Yasunaga,T., Kuhara,S., Shiba,T., Hattori,M. and
            Shinagawa,H.
  TITLE     Complete genome sequence of enterohemorrhagic Escherichia coli
            O157:H7 and genomic comparison with a laboratory strain K-12
  JOURNAL   DNA Res. 8 (1), 11-22 (2001)
```

Links
- Gene
- Genome Project
- Full text in PMC
- GEO Profiles
- Protein
- PubMed
- Taxonomy
- LinkOut

NCBI Entrez Protein

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Protein for Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence project
Search for Genes
Clusters of orthologous groups
Protein reviews on the web

Display Summary Show 20 Sort by Relevance Send to

All: 5253 | Bacteria: 5253 | RefSeq: 5253 | Related Structures: 2886

Items 1 - 20 of 5253

Page 1 of 263 Next

1: NP_308028 Reports Links
thr operon leader peptide [Escherichia coli O157:H7 str. Sakai]
gi|15829255|ref|NP_308028.1|[15829255]

2: NP_308029 Reports BLink, Conserved Domains, Links
bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 str. Sakai]
gi|15829256|ref|NP_308029.1|[15829256]

3: NP_308030 Reports BLink, Conserved Domains, Links
homoserine kinase [Escherichia coli O157:H7 str. Sakai]
gi|15829257|ref|NP_308030.1|[15829257]

4: NP_308031 Reports BLink, Conserved Domains, Links
threonine synthase [Escherichia coli O157:H7 str. Sakai]
gi|15829258|ref|NP_308031.1|[15829258]

5: NP_308032 Reports BLink, Links
hypothetical protein ECs0005 [Escherichia coli O157:H7 str. Sakai]
gi|15829259|ref|NP_308032.1|[15829259]

NCBI Entrez Genome

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

Genome > Bacteria > *Escherichia coli O157:H7 EDL933, complete genome*

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli EDL933

Chromosomes: *chromosome*
Plasmids: pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002655	Genes: 5453	COG	Genome Project	Publications: [1]
GenBank: AE005174	Protein coding: 5324	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,528,445 nt**	Structural RNAs: 128	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: **Univ. Wisconsin**
% Coding: **87%**	Others: **328**	GenePlot	TraceAssembly	Completed: **2001/09/27**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

NCBI Entrez Genome

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Protein Table Show 20 Send to

All: 1

Position: from begin to end Length: from 0 to inf Refresh Reset

Escherichia coli O157:H7 str. Sakai, complete genome

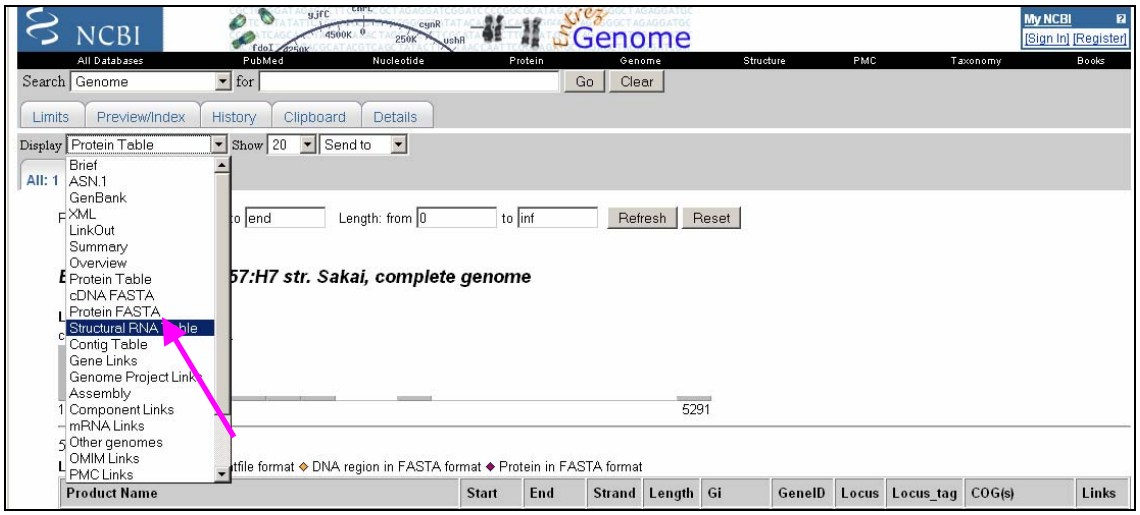
Length histogram.
Click on a bar to select length range.

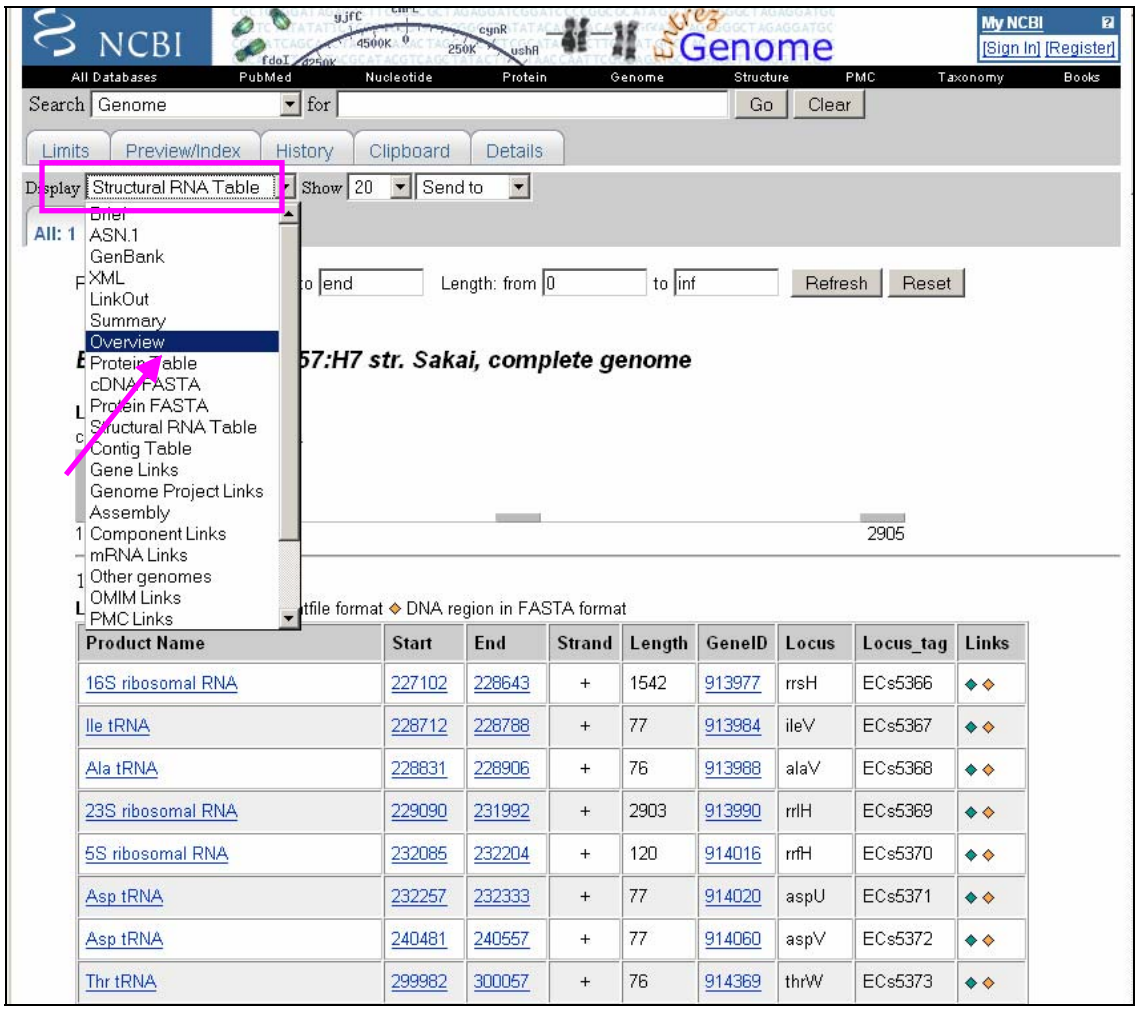
1 5291

5253 protein(s) shown

Legends: ◆ DNA region in flatfile format ◇ DNA region in FASTA format ◆ Protein in FASTA format

Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
thr operon leader peptide	190	273	+	27	15829255	913387	-	ECs0001	-	◆◇◆
bifunctional aspartokinase I/homeserine dehydrogenase I	354	2816	+	820	15829256	913388	thrA	ECs0002	COG0460E, COG0527E	◆◇◆
homoserine kinase	2818	3750	+	310	15829257	913390	-	ECs0003	COG0083E	◆◇◆
threonine synthase	3751	5037	+	428	15829258	913393	-	ECs0004	COG0498E	◆◇◆
hypothetical protein ECs0005	5251	5547	+	98	15829259	913394	-	ECs0005	-	◆◇◆
hypothetical protein ECs0006	5700	6476	-	258	15829260	913395	-	ECs0006	COG3022S	◆◇◆
putative inner membrane transport protein	6546	7976	-	476	15829261	913398	-	ECs0007	COG1115E	◆◇◆


NCBI
Genome
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search Genome for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display Protein Table
Show 20
Send to
Brief
ASN.1
GenBank
XML
LinkOut
Summary
Overview
Protein Table
cDNA FASTA
Protein FASTA
Structural RNA Table
Contig Table
Gene Links
Genome Project Links
Assembly
Component Links
mRNA Links
Other genomes
OMIM Links
PMC Links
Length: from 0 to inf
Refresh
Reset
57:H7 str. Sakai, complete genome
5291
Product Name
Start
End
Strand
Length
Gi
GeneID
Locus
Locus_tag
COG(s)
Links


NCBI
Genome
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search Genome for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display Structural RNA Table
Show 20
Send to
ASN.1
GenBank
XML
LinkOut
Summary
Overview
Protein Table
cDNA FASTA
Protein FASTA
Structural RNA Table
Contig Table
Gene Links
Genome Project Links
Assembly
Component Links
mRNA Links
Other genomes
OMIM Links
PMC Links
Length: from 0 to inf
Refresh
Reset
57:H7 str. Sakai, complete genome
2905
Product Name
Start
End
Strand
Length
GeneID
Locus
Locus_tag
Links
16S ribosomal RNA 227102 228643 + 1542 913977 rrsH ECs5366
Ile tRNA 228712 228788 + 77 913984 ileV ECs5367
Ala tRNA 228831 228906 + 76 913988 alaV ECs5368
23S ribosomal RNA 229090 231992 + 2903 913990 rrlH ECs5369
5S ribosomal RNA 232085 232204 + 120 914016 rrfH ECs5370
Asp tRNA 232257 232333 + 77 914020 aspU ECs5371
Asp tRNA 240481 240557 + 77 914060 aspV ECs5372
Thr tRNA 299982 300057 + 76 914369 thrW ECs5373

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

Genome > **Bacteria** > ***Escherichia coli O157:H7 str. Sakai, complete genome***

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Esche
; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt 9,910

ECs0003 ECs0004 ECs0008 mogA

thrA ECs0006 ECs0007

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt 9,910

ECs0003 ECs0004 ECs0008 mogA

thrA ECs0006 ECs0007

bifunctional aspartokinase I/homeserine dehydrogen...
Gene: thrA

Click here for S ... ino acids) of selected region

NCBI Entrez Gene

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books | OMIM

Search Gene for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Full Report Show 20 Send to

All: 1 | Current Only: 1 | Genes Genomes: 1 | SNP GeneView: 0

1: thrA bifunctional aspartokinase I/homeserine dehydrogenase I [*Escherichia coli O157:H7 str. Sakai* **]**

GeneID: 913388 updated 07-Jul-2006

Summary

Gene name	thrA
Locus tag	ECs0002
Gene type	protein coding
RefSeq status	Validated
Organism	*Escherichia coli O157:H7 str. Sakai (strain: Sakai, substrain: RIMD 0509952, serovar: O157:H7)*
Lineage	*Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia*

Genomic regions, transcripts, and products

Go to reference sequence details

NC_002695.1

[354 ▶ [2816 ▶

5' 3'

NP_308029 bifunctional aspartok...

■ - coding region

Entrez Gene Home

Table Of Contents

- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- General gene information
- General protein information
- Reference Sequences
- Related Sequences

Links Explain

- Conserved Domains
- Genome
- Nucleotide
- Full text in PMC
- Protein
- PubMed
- Taxonomy
- KEGG
- LinkOut

Entrez Gene Info

Feedback

Subscriptions

General protein information

Names

bifunctional aspartokinase I/homeserine dehydrogenase I

NP_308029.1

COG classification

multifunctional homotetrameric enzyme that catalyzes the phosphorylation of aspartate to form aspartyl-4-phosphate as well as conversion of aspartate semialdehyde to homoserine; functions in a number of amino acid biosynthetic pathways

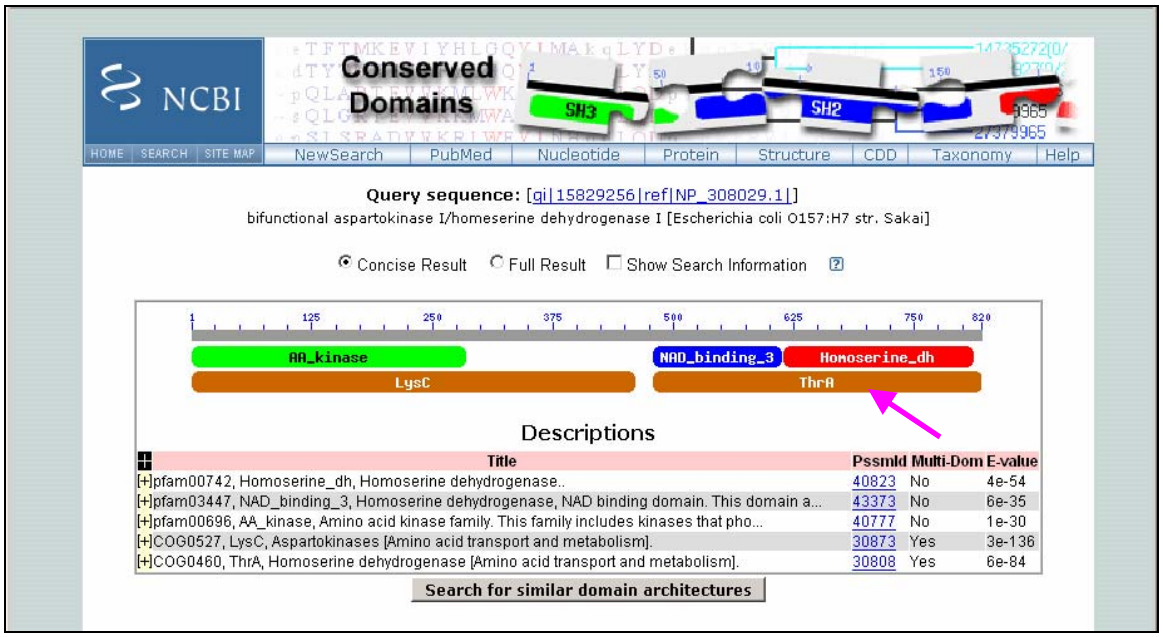
NCBI Reference Sequences (RefSeq)

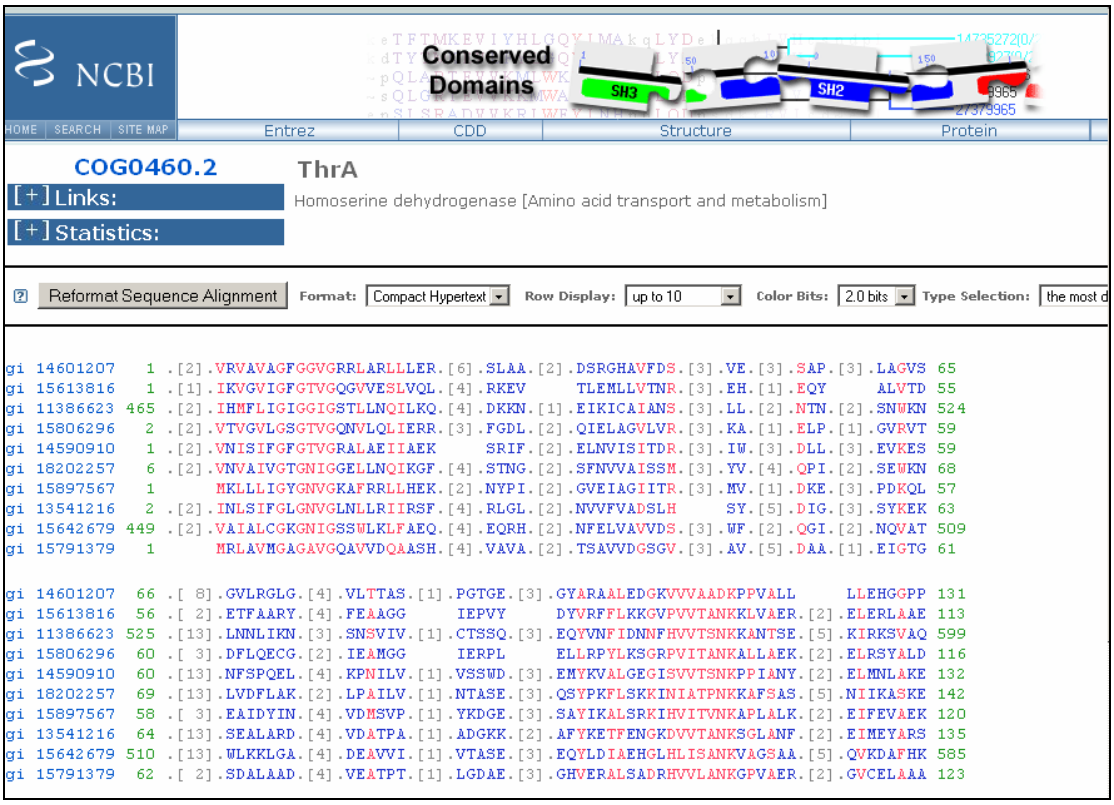
mRNA and Protein(s)

1. **NP_308029.1 bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 str. Sakai]**

Conserved Domains (5) summary

COG0460 Location:479-820 Blast Score:786	ThrA; Homoserine dehydrogenase [Amino acid transport and metabolism]
COG0527 Location:1-460 Blast Score:1237	LysC; Aspartokinases [Amino acid transport and metabolism]
pfam00696 Location:1-284 Blast Score:326	AA_kinase; Amino acid kinase family
pfam00742 Location:614-811 Blast Score:528	Homoserine_dh; Homoserine dehydrogenase
pfam03447 Location:479-612 Blast Score:363	NAD_binding_3; Homoserine dehydrogenase, NAD binding domain


NCBI
Conserved Domains
SH3
SH2
HOME SEARCH SITE MAP
NewSearch PubMed Nucleotide Protein Structure CDD Taxonomy Help
Query sequence: [gi|15829256|ref|NP_308029.1|]
bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 str. Sakai]
Concise Result
Full Result
Show Search Information
AA_kinase
NAD_binding_3
Homoserine_dh
LysC
ThrA
Descriptions
Title
PssmId Multi-Dom E-value
[+]pfam00742, Homoserine_dh, Homoserine dehydrogenase.. 40823 No 4e-54
[+]pfam03447, NAD_binding_3, Homoserine dehydrogenase, NAD binding domain. This domain a... 43373 No 6e-35
[+]pfam00696, AA_kinase, Amino acid kinase family. This family includes kinases that pho... 40777 No 1e-30
[+]COG0527, LysC, Aspartokinases [Amino acid transport and metabolism]. 30873 Yes 3e-136
[+]COG0460, ThrA, Homoserine dehydrogenase [Amino acid transport and metabolism]. 30808 Yes 6e-84
Search for similar domain architectures


NCBI
Conserved Domains
SH3
SH2
HOME SEARCH SITE MAP
Entrez CDD Structure Protein
COG0460.2
ThrA
[+] Links:
Homoserine dehydrogenase [Amino acid transport and metabolism]
[+] Statistics:
Reformat Sequence Alignment
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most d
gi 14601207 1 .[2].VRVAVAGFGGVGRRLARLLER.[6].SLAA.[2].DSRGHAVFDS.[3].VE.[3].SAP.[3].LAGVS 65
gi 15613816 1 .[1].IKVGVIGFGTVGQGVVESLVQL.[4].RKEV TLEMLLVTNR.[3].EH.[1].EQY ALVTD 55
gi 11386623 465 .[2].IHMFLIGIGGIGSTLLNQILKQ.[4].DKKN.[1].EIKICAIANS.[3].LL.[2].NTN.[2].SNWKN 524
gi 15806296 2 .[2].VTVGVLGSGTVGQNVLQLIERR.[3].FGDL.[2].QIELAGVLVR.[3].KA.[1].ELP.[1].GVRVT 59
gi 14590910 1 .[2].VNISIFGFGTVGRALAEIIAEK SRIF.[2].ELNVISITDR.[3].IW.[3].DLL.[3].EVKES 59
gi 18202257 6 .[2].VNVAIVGTGNIGGELLNQIKGF.[4].STNG.[2].SFNVVAISSM.[3].YV.[4].QPI.[2].SEWKN 68
gi 15897567 1 MKLLLIGYGNVGKAFRRLLHEK.[2].NYPI.[2].GVEIAGIITR.[3].MV.[1].DKE.[3].PDKQL 57
gi 13541216 2 .[2].INLSIFGLGNVGLNLLRIIRSF.[4].RLGL.[2].NVVFVADSLH SY.[5].DIG.[3].SYKEK 63
gi 15642679 449 .[2].VAIALCGKGNIGSSWLKLFAEQ.[4].EQRH.[2].NFELVAVVDS.[3].WF.[2].QGI.[2].NQVAT 509
gi 15791379 1 MRLAVMGAGAVGQAVVDQAASH.[4].VAVA.[2].TSAVVDGSGV.[3].AV.[5].DAA.[1].EIGTG 61
gi 14601207 66 .[8].GVLRGLG.[4].VLTTAS.[1].PGTGE.[3].GYARAALEDGKVVVAADKPPVALL LLEHGGPP 131
gi 15613816 56 .[2].ETFAARY.[4].FEAAGG IEPVY DYVRFFLKKGVPVVTANKKLVAER.[2].ELERLAAE 113
gi 11386623 525 .[13].LNNLIKN.[3].SNSVIV.[1].CTSSQ.[3].EQYVNFIDNNFHVVTSNKKANTSE.[5].KIRKSVAQ 599
gi 15806296 60 .[3].DFLQECG.[2].IEAMGG IERPL ELLRPYLKSGRPVITANKALLAEK.[2].ELRSYALD 116
gi 14590910 60 .[13].NFSPQEL.[4].KPNILV.[1].VSSWD.[3].EMYKVALGEGISVVTSNKPPIANY.[2].ELMNLAKE 132
gi 18202257 69 .[13].LVDFLAK.[2].LPAILV.[1].NTASE.[3].QSYPKFLSKKINIATPNKKAFSAS.[5].NIIKASKE 142
gi 15897567 58 .[3].EAIDYIN.[4].VDMSVP.[1].YKDGE.[3].SAYIKALSRKIHVITVNKAPLALK.[2].EIFEVAEK 120
gi 13541216 64 .[13].SEALARD.[4].VDATPA.[1].ADGKK.[2].AFYKETFENGKDVVTANKSGLANF.[2].EIMEYARS 135
gi 15642679 510 .[13].WLKKLGA.[4].DEAVVI.[1].VTASE.[3].EQYLDIAEHGLHLISANKVAGSAA.[5].QVKDAFHK 585
gi 15791379 62 .[2].SDALAAD.[4].VEATPT.[1].LGDAE.[3].GHVERALSADRHVVLANKGPVAER.[2].GVCELAAA 123

All: 1

Genome > **Bacteria** > ***Escherichia coli O157:H7 str. Sakai, complete genome***

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia co ; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

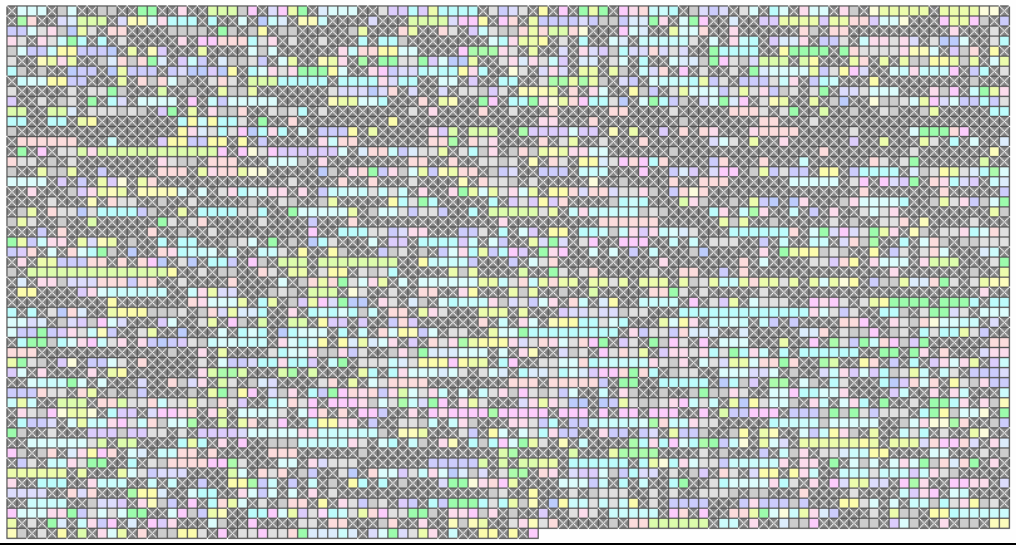
Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: thrC Find Gene

Escherichia coli O157:H7 str. Sakai **sequence** Microbial genomes
genus: ***Escherichia***
group: ***Gammaproteobacteria***
kingdom: *Bacteria*

5253 proteins: distribution by COGs functional categories
3306 proteins can be found in COGs data base



Code		COGs	Description	(% in sequence)	(% in genome)	(% in genus)	(% in Gammaproteobacteria)	(% in Bacteria)
☐	J	150	Translation	2.6046	2.5424	2.7363	3.3678	3.6565
☐	A	1	RNA processing and modification	0.0174	0.0169	0.0183	0.0242	0.0114
☐	K	281	Transcription	4.8793	4.8814	5.1309	4.8981	4.9477
☐	L	273	Replication, recombination and repair	4.7404	5.0508	4.5116	5.1961	4.4295
☐	B	0	Chromatin structure and dynamics	0.0000	0.0000	0.0000	0.0000	0.0044
☐	D	29	Cell cycle control, mitosis and meiosis	0.5036	0.5085	0.5369	0.6737	0.6910
☐	Y	-	Nuclear structure	-	-	-	-	-
☐	V	51	Defense mechanisms	0.8856	0.9068	0.9365	0.9882	1.1522
☐	T	168	Signal transduction mechanisms	2.9172	2.8729	2.9681	3.6227	3.4483
☐	M	242	Cell wall/membrane biogenesis	4.2021	4.2203	4.3072	4.3054	4.1094
☐	N	120	Cell motility	2.0837	2.1525	2.0835	2.0685	1.3776
☐	Z	0	Cytoskeleton	0.0000	0.0000	0.0000	0.0057	0.0052
☐	W	2	Extracellular structures	0.0347	0.0339	0.0214	0.0108	0.0071
☐	U	146	Intracellular trafficking and secretion	2.5352	2.7034	2.5166	2.2961	1.7173
☐	O	134	Posttranslational modification, protein turnover, chaperones	2.3268	2.3305	2.4251	2.8264	2.6152
☐	C	274	Energy production and conversion	4.7578	4.6695	5.0485	4.3095	4.3745
☐	G	261	Carbohydrate transport and metabolism	4.5320	4.4831	5.0577	3.5918	3.4880
☐	E	325	Amino acid transport and metabolism	5.6433	5.5169	6.0185	5.9370	5.9060
☐	F	77	Nucleotide transport and metabolism	1.3370	1.3136	1.4734	1.4513	1.6335
☐	H	135	Coenzyme transport and metabolism	2.3442	2.2881	2.4495	2.6556	2.6540
☐	I	106	Lipid transport and metabolism	1.8406	1.8051	1.9065	2.3512	2.6600
☐	P	220	Inorganic ion transport and metabolism	3.8201	3.7966	3.9351	3.6514	3.6268
☐	Q	75	Secondary metabolites biosynthesis, transport and catabolism	1.3023	1.2712	1.4002	1.6157	1.7937
☐	R	469	General function prediction only	8.1438	7.9322	7.9220	7.8357	8.1934
☐	S	273	Function unknown	4.7404	4.7034	4.4567	4.7426	4.8112
☒	-	1947	Not in COGs	33.8080	34.0000	32.1396	31.5744	32.6864

Escherichia coli O157:H7 str. Sakai

Intracellular trafficking and secretion genes:

Location	strand	Length	PID	Locus	LocusTag	COG	Product
19298..21748	-	816	15829276	-	ECs0022	COG3188NU	putative outer membrane usher protein precursor
21761..22444	-	227	15829277	-	ECs0023	COG3121NU	putative fimbrial chaperone
22494..23027	-	177	15829278	-	ECs0024	COG3539NU	putative fimbrial protein
29618..30112	+	164	15829284	lspA	ECs0030	COG0597MU	signal peptidase II
112884..115589	+	901	15829356	-	ECs0102	COG0653U	translocase
119011..120213	-	400	15829364	-	ECs0110	COG1459NU	HofC
120203..121588	-	461	15829365	-	ECs0111	COG2804NU	HofB
121598..122038	-	146	15829366	-	ECs0112	COG4969NU	prelipin peptidase dependent protein
156679..157239	-	186	15829396	-	ECs0142	COG3539NU	putative fimbrial protein
157241..159841	-	866	15829397	-	ECs0143	COG3188NU	putative outer membrane porin protein
159883..160608	-	241	15829398	-	ECs0144	COG3121NU	putative pilin chaperone
160695..161291	-	198	15829399	-	ECs0145	COG3539NU	putative fimbrial-like protein

All: 1

Genome > **Bacteria** > ***Escherichia coli O157:H7 str. Sakai, complete genome***

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia ; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

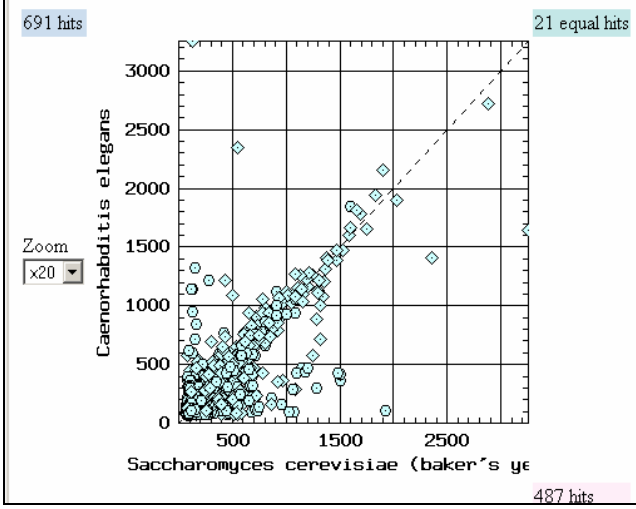
386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

6239 Caenorhabditis elegans

4932 Saccharomyces cerevisiae (baker's yeast)

Distribution of *E. coli O157:H7 str. Sakai* homologs



BLAST PubMed Nucleotide Protein Genome Structure Taxonomy

Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

155864 E. coli O157:H7 EDL933

Choose two species for comparison

6239 Caenorhabditis elegans

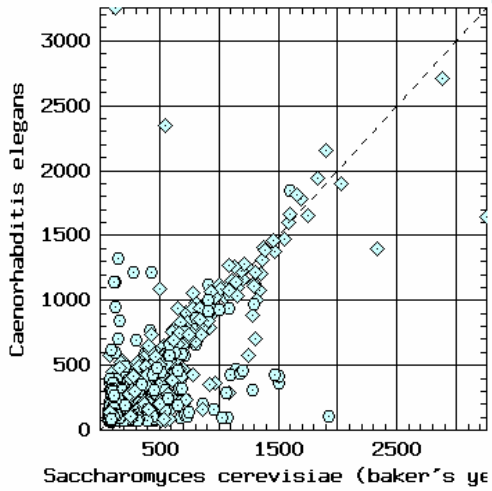
4932 Saccharomyces cerevisiae (baker's yeast)

Distribution of *E. coli O157:H7 EDL933* homologs

691 hits

22 equal hits

Zoom x20



501 hits

Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 E. coli K12

4932 Saccharomyces cerevisiae (baker's yeast)

—Erythrobacter—
E. litoralis HTCC2594
E. sp. NAP1
—Escherichia—
E. coli 101-1
E. coli 536
E. coli 53638
E. coli B171
E. coli B7A
E. coli CFT073
E. coli E110019
E. coli E22
E. coli E24377A
E. coli F11
E. coli HS
E. coli K12
E. coli O157:H7
E. coli O157:H7 EDL933
E. coli O157:H7 str. Sakai
E. coli UTI89

Distributi

691 hits

Zoom x20

Protein homologs in Complete Microbial / Eukaryotic genomes

To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu

Select your query genome

386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 E. coli K12

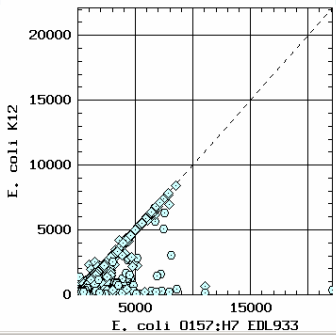
155864 E. coli O157:H7 EDL933

Distribution of *E. coli O157:H7 str. Sakai* homologs

329 hits

992 equal hits

Zoom x20



Done

Select your query genome

386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 E. coli K12

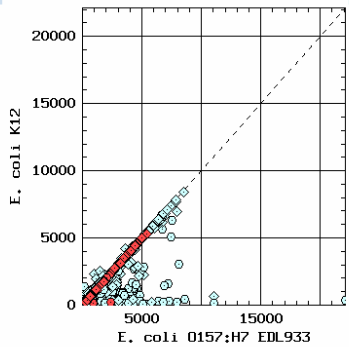
155864 E. coli O157:H7 EDL933

Distribution of *E. coli O157:H7 str. Sakai* homologs

329 hits

992 equal hits

Zoom x20



2996 hits

5340 query proteins produced **4317 hits**, from which 161 are selected (first 100 shown).
Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Symmetrical hits are shown as diamonds. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Cutoff: 10 Linear scale compare

Query: compare clear

Translation, ribosomal structure and biogenesis compare

386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 E. coli K12

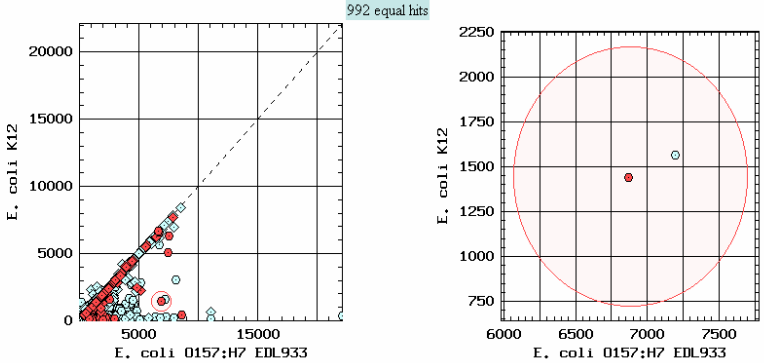
155864 E. coli O157:H7 EDL933

Distribution of *E. coli O157:H7 str. Sakai* homologs

329 hits

992 equal hits

Zoom x20



2996 hits

5340 query proteins produced **4317 hits**, from which 1 is selected .

Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Symmetrical hits are shown as diamonds. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Cutoff: 10 Linear scale compare

Query: compare clear

Cell envelope biogenesis, outer membrane compare

Blast2Seq	Accession	Blink	Definition
	NP_308389	=>	**AidA-I adhesin-like protein [Escherichia coli O157:H7 str. Sakai]**
1440	NP_416026	=>	conserved protein, predicted pseudogene [Escherichia coli K12]
6869	NP_286049	=>	putative beta-barrel outer membrane protein [Escherichia coli O157:H7 EDL933]

All: 1

Genome > ***Bacteria*** > ***Escherichia coli O157:H7 str. Sakai, complete genome***

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli O157:H ; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: **GIRC**
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

NCBI Entrez Genome

PubMed | Nucleotide | Protein | Genome | Structure | Gene | Taxonomy | OMIM

Search Genome Go Clear

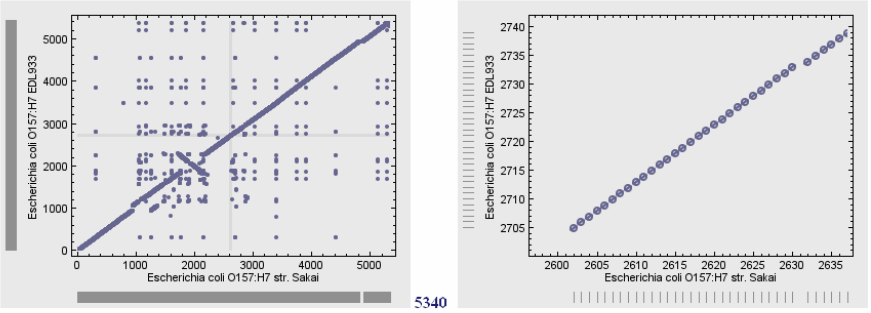
Pairwise genome comparison of protein homologs (symmetrical best hits)

Query organism: Escherichia coli O157:H7 str. Sakai

Select two organisms to compare

E. coli O157:H7 str. Sakai versus E. coli O157:H7 EDL933 compare selected pair

5423 proteins total



Zoom 2x

click on the left picture to recenter
click on the right picture to zoom and recenter

shaded rectangle from the left picture is shown on the right picture

switch axis

Total number of bets 5341. Save all bets in order on genome.

bl2seq	Locus tags	Protein name
●	Z3024 - ECs2675	hypothetical protein ECs2675 [Escherichia coli O157:H7 str. Sakai]
●	Z3027 - ECs2676	flagellar basal body protein [Escherichia coli O157:H7 str. Sakai]

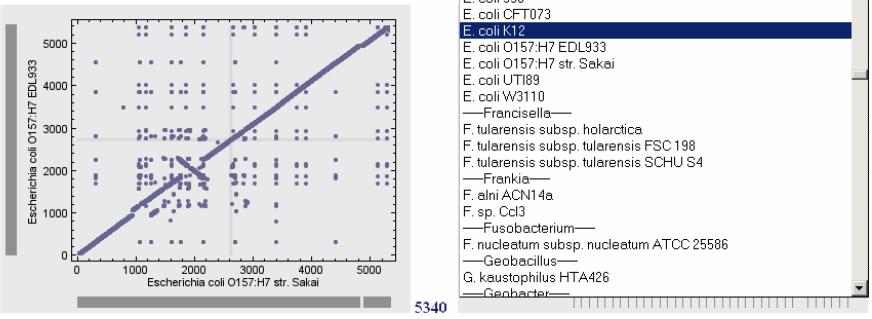
Pairwise genome comparison of protein homologs (symmetrical best hits)

Query organism: Escherichia coli O157:H7 str. Sakai

Select two organisms to compare

E. coli O157:H7 str. Sakai versus E. coli O157:H7 EDL933 compare selected pair

5423 proteins total



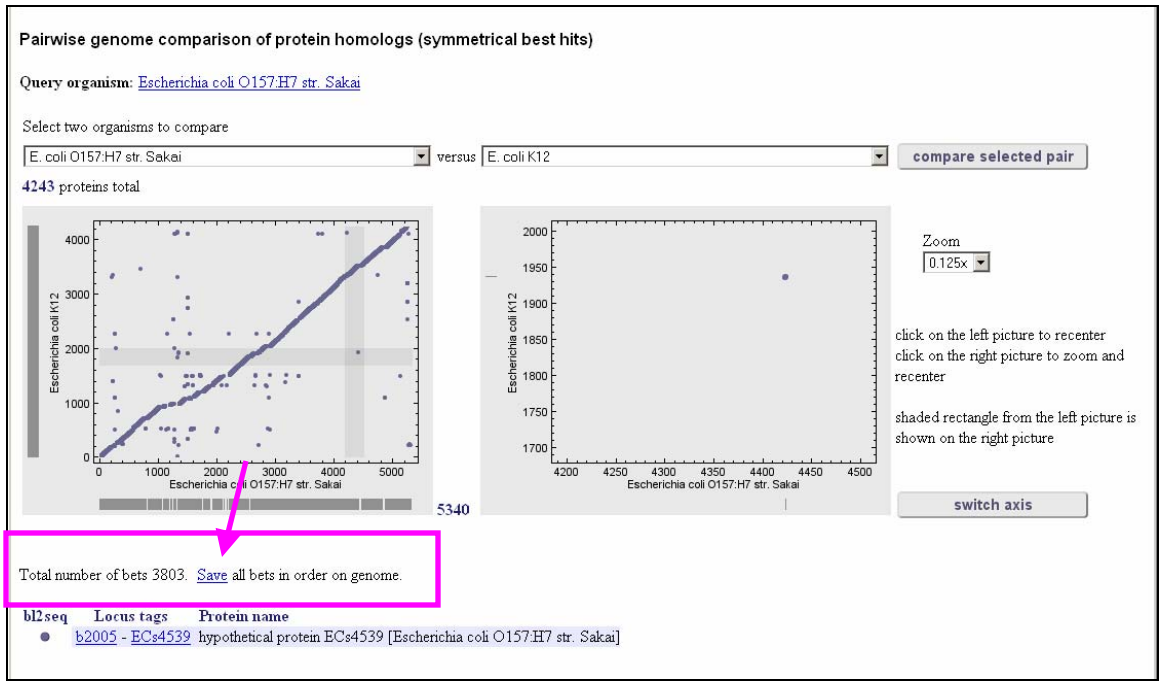
—Escherichia—
E. coli 536
E. coli CFT073
E. coli K12
E. coli O157:H7 EDL933
E. coli O157:H7 str. Sakai
E. coli UTI89
E. coli W3110
—Francisella—
F. tularensis subsp. holarctica
F. tularensis subsp. tularensis FSC 198
F. tularensis subsp. tularensis SCHU S4
—Frankia—
F. alni ACN14a
F. sp. Ccl3
—Fusobacterium—
F. nucleatum subsp. nucleatum ATCC 25586
—Geobacillus—
G. kaustophilus HTA426
—Geobacter—

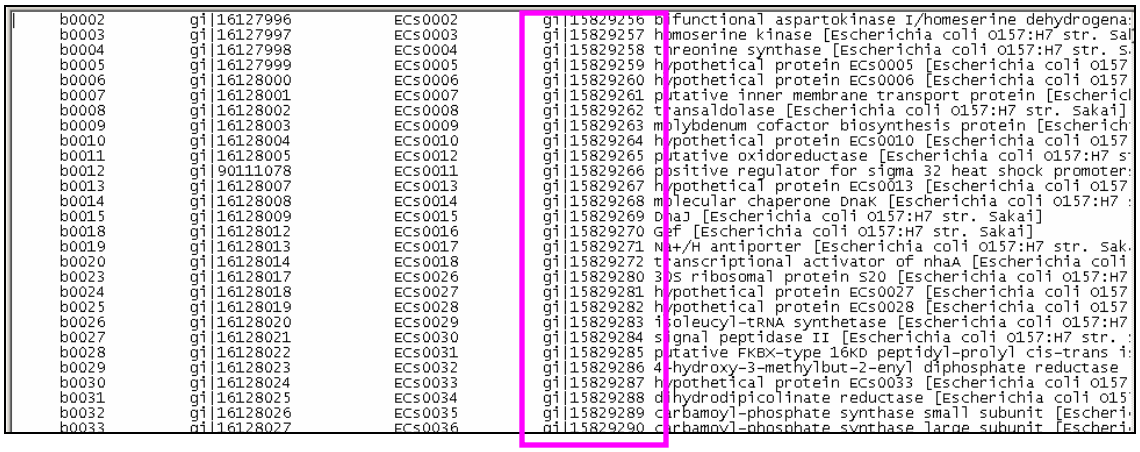
Zoom 2x

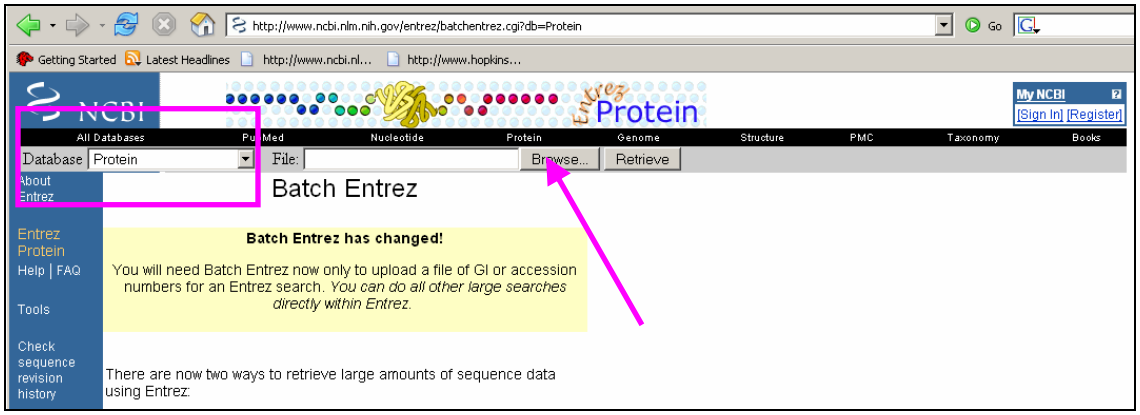
click on the left picture to recenter
click on the right picture to zoom and recenter

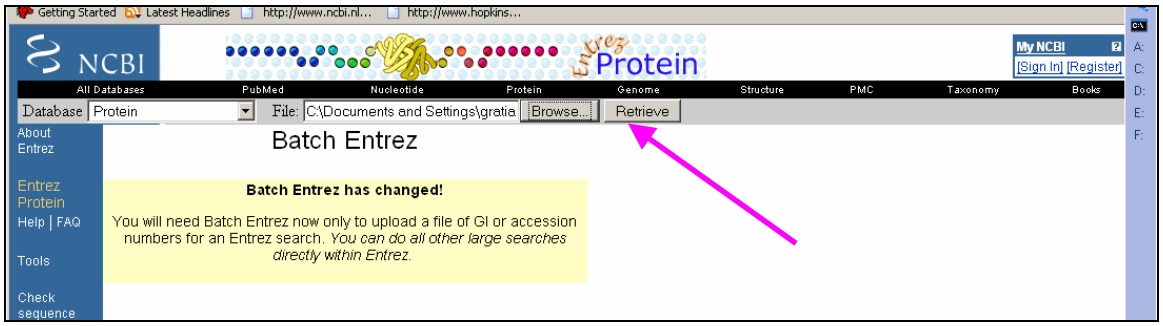
shaded rectangle from the left picture is shown on the right picture

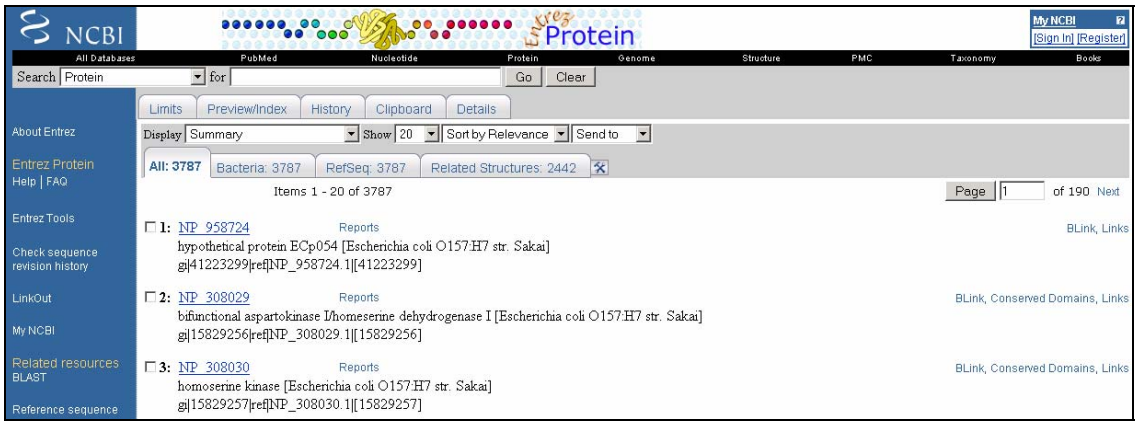
switch axis

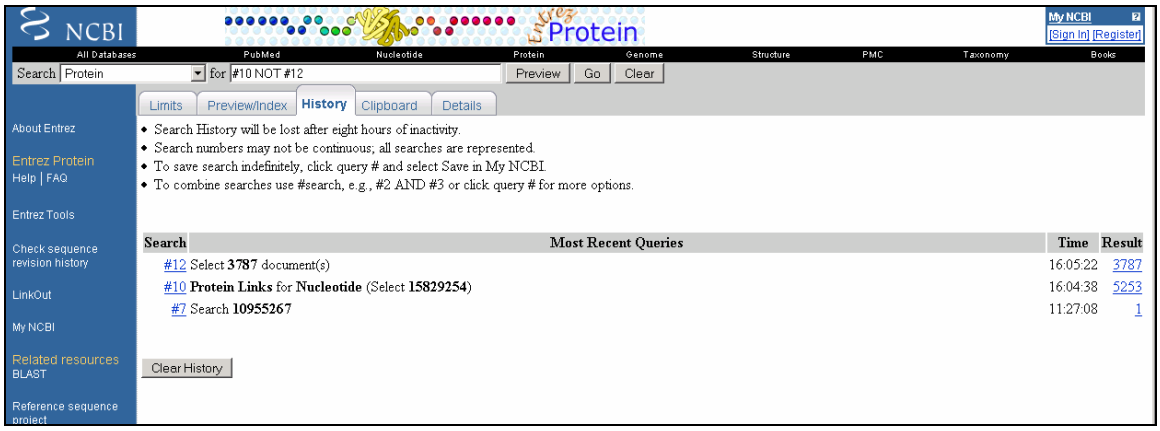

Pairwise genome comparison of protein homologs (symmetrical best hits)
Query organism: Escherichia coli O157:H7 str. Sakai
Select two organisms to compare
E. coli O157:H7 str. Sakai
versus
E. coli K12
compare selected pair
4243 proteins total
Escherichia coli K12
Escherichia coli O157:H7 str. Sakai
5340
Zoom
0.125x
click on the left picture to recenter
click on the right picture to zoom and recenter
shaded rectangle from the left picture is shown on the right picture
switch axis
Total number of bets 3803. Save all bets in order on genome.
bl2seq
Locus tags
Protein name
b2005 - ECs4539 hypothetical protein ECs4539 [Escherichia coli O157:H7 str. Sakai]

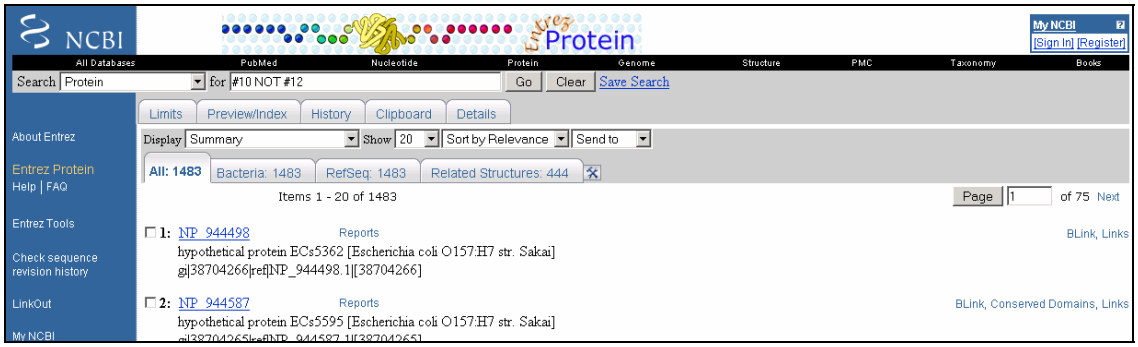

b0002 gi|16127996 ECs0002 gi|15829256 bifunctional aspartokinase I/homeserine dehydrogena
b0003 gi|16127997 ECs0003 gi|15829257 homoserine kinase [Escherichia coli O157:H7 str. Sal
b0004 gi|16127998 ECs0004 gi|15829258 threonine synthase [Escherichia coli O157:H7 str. S
b0005 gi|16127999 ECs0005 gi|15829259 hypothetical protein ECs0005 [Escherichia coli O157
b0006 gi|16128000 ECs0006 gi|15829260 hypothetical protein ECs0006 [Escherichia coli O157
b0007 gi|16128001 ECs0007 gi|15829261 putative inner membrane transport protein [Escherich
b0008 gi|16128002 ECs0008 gi|15829262 transaldolase [Escherichia coli O157:H7 str. Sakai]
b0009 gi|16128003 ECs0009 gi|15829263 molybdenum cofactor biosynthesis protein [Escherich
b0010 gi|16128004 ECs0010 gi|15829264 hypothetical protein ECs0010 [Escherichia coli O157
b0011 gi|16128005 ECs0012 gi|15829265 putative oxidoreductase [Escherichia coli O157:H7 s
b0012 gi|90111078 ECs0011 gi|15829266 positive regulator for sigma 32 heat shock promoter
b0013 gi|16128007 ECs0013 gi|15829267 hypothetical protein ECs0013 [Escherichia coli O157
b0014 gi|16128008 ECs0014 gi|15829268 molecular chaperone DnaK [Escherichia coli O157:H7
b0015 gi|16128009 ECs0015 gi|15829269 DnaJ [Escherichia coli O157:H7 str. Sakai]
b0018 gi|16128012 ECs0016 gi|15829270 Gef [Escherichia coli O157:H7 str. Sakai]
b0019 gi|16128013 ECs0017 gi|15829271 Na+/H antiporter [Escherichia coli O157:H7 str. Sak
b0020 gi|16128014 ECs0018 gi|15829272 transcriptional activator of nhaA [Escherichia coli
b0023 gi|16128017 ECs0026 gi|15829280 30S ribosomal protein S20 [Escherichia coli O157:H7
b0024 gi|16128018 ECs0027 gi|15829281 hypothetical protein ECs0027 [Escherichia coli O157
b0025 gi|16128019 ECs0028 gi|15829282 hypothetical protein ECs0028 [Escherichia coli O157
b0026 gi|16128020 ECs0029 gi|15829283 isoleucyl-tRNA synthetase [Escherichia coli O157:H7
b0027 gi|16128021 ECs0030 gi|15829284 signal peptidase II [Escherichia coli O157:H7 str.
b0028 gi|16128022 ECs0031 gi|15829285 putative FKBX-type 16KD peptidyl-prolyl cis-trans i
b0029 gi|16128023 ECs0032 gi|15829286 4-hydroxy-3-methylbut-2-enyl diphosphate reductase
b0030 gi|16128024 ECs0033 gi|15829287 hypothetical protein ECs0033 [Escherichia coli O157
b0031 gi|16128025 ECs0034 gi|15829288 dihydrodipicolinate reductase [Escherichia coli O15
b0032 gi|16128026 ECs0035 gi|15829289 carbamoyl-phosphate synthase small subunit [Escheri
b0033 gi|16128027 ECs0036 gi|15829290 carbamoyl-phosphate synthase large subunit [Escheri


http://www.ncbi.nlm.nih.gov/entrez/batchentrez.cgi?db=Protein
Getting Started
Latest Headlines
NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Database
Protein
File:
Browse...
Retrieve
About Entrez
Batch Entrez
Entrez Protein
Help | FAQ
Tools
Check sequence revision history
Batch Entrez has changed!
You will need Batch Entrez now only to upload a file of GI or accession numbers for an Entrez search. You can do all other large searches directly within Entrez.
There are now two ways to retrieve large amounts of sequence data using Entrez:


Getting Started
Latest Headlines
http://www.ncbi.nl...
http://www.hopkins...
NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Database Protein
File: C:\Documents and Settings\gratia
Browse...
Retrieve
About Entrez
Entrez Protein
Help | FAQ
Tools
Check sequence
Batch Entrez
Batch Entrez has changed!
You will need Batch Entrez now only to upload a file of GI or accession numbers for an Entrez search. You can do all other large searches directly within Entrez.


NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search Protein
for
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Display Summary
Show 20
Sort by Relevance
Send to
All: 3787
Bacteria: 3787
RefSeq: 3787
Related Structures: 2442
Items 1 - 20 of 3787
Page 1 of 190 Next
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence
1: NP_958724 Reports BLink, Links
hypothetical protein ECp054 [Escherichia coli O157:H7 str. Sakai]
gi|41223299|ref|NP_958724.1|[41223299]
2: NP_308029 Reports BLink, Conserved Domains, Links
bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 str. Sakai]
gi|15829256|ref|NP_308029.1|[15829256]
3: NP_308030 Reports BLink, Conserved Domains, Links
homoserine kinase [Escherichia coli O157:H7 str. Sakai]
gi|15829257|ref|NP_308030.1|[15829257]


NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search Protein
for #10 NOT #12
Preview
Go
Clear
Limits
Preview/Index
History
Clipboard
Details
Search History will be lost after eight hours of inactivity.
Search numbers may not be continuous; all searches are represented.
To save search indefinitely, click query # and select Save in My NCBI
To combine searches use #search, e.g., #2 AND #3 or click query # for more options.
Search
Most Recent Queries
Time
Result
#12 Select 3787 document(s) 16:05:22 3787
#10 Protein Links for Nucleotide (Select 15829254) 16:04:38 5253
#7 Search 10955267 11:27:08 1
Clear History
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence project


NCBI
Entrez Protein
My NCBI
[Sign In] [Register]
All Databases
PubMed
Nucleotide
Protein
Genome
Structure
PMC
Taxonomy
Books
Search Protein
for #10 NOT #12
Go
Clear
Save Search
Limits
Preview/Index
History
Clipboard
Details
Display Summary
Show 20
Sort by Relevance
Send to
All: 1483
Bacteria: 1483
RefSeq: 1483
Related Structures: 444
Items 1 - 20 of 1483
Page 1 of 75 Next
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
1: NP_944498 Reports BLink, Links
hypothetical protein ECs5362 [Escherichia coli O157:H7 str. Sakai]
gi|38704266|ref|NP_944498.1|[38704266]
2: NP_944587 Reports BLink, Conserved Domains, Links
hypothetical protein ECs5595 [Escherichia coli O157:H7 str. Sakai]

NCBI Entrez Genome

My NCBI [Sign In] [Register]

All Databases | PubMed | Nucleotide | Protein | Genome | Structure | PMC | Taxonomy | Books

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

Genome > Bacteria > Escherichia coli O157:H7 str. Sakai, complete genome Links

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli O157:H7 ; Escherichia coli O157:H7 str. Sakai

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002695	Genes: 5395	COG	Genome Project	Publications: [4]
GenBank: BA000007	Protein coding: 5253	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,498,450 nt**	Structural RNAs: 141	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **50%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: GIRC
% Coding: **85%**	Others: **170**	GenePlot	TraceAssembly	Completed: **2001/10/02**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

BLAST | PubMed | Nucleotide | Protein | Genome | Structure | Ta

- [Escherichia (5081)] - [Enterobacteriaceae (4729)] - [Enterobacteriales (4729)] - [Gammaproteobacteria (4611)] - [Proteobacteria (4436)] - [Bacteria [cellular organisms (1934)]

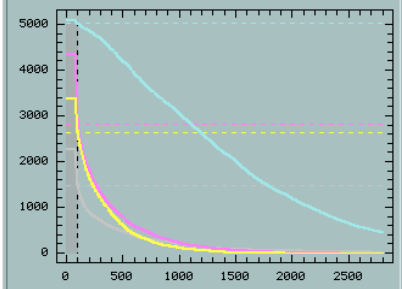
5253 *Escherichia coli O157:H7 str. Sakai* proteins: taxonomic distribution of the homologs

Click on a region in the map, to see blast results for that region

COGTable Cut-Off: 95 Cut-Off+: 10 Redraw HELP

Click on the graph at left, to change the blast threshold (cut-off) used to find related sequences

Distribution of database hits

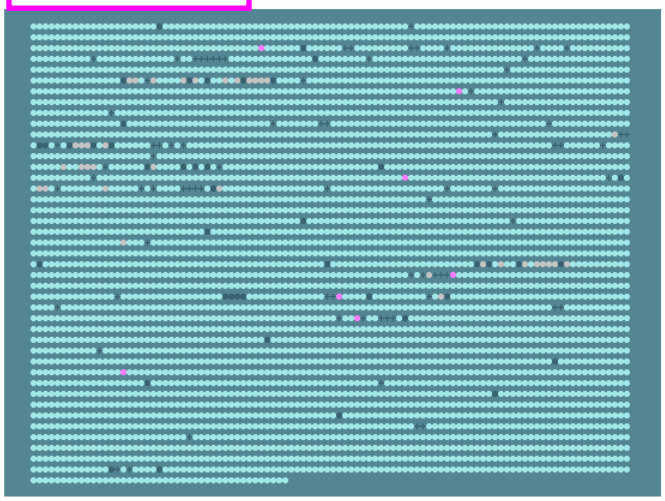


Name		Above cut-off	Best
Virus	V	1462	383
Eukaryota	E	2800	8
Eubacteria	B	5031	4541
Archaea	A	2602	0
Eubacteria + Archaea	BA	5031	4541
Archaea + Eukaryota	AE	3273	8
Eukaryota + Eubacteria	EB	5034	4549
Eukaryota + Eubacteria + Archaea	EBA	5034	4549

BLAST PubMed Nucleotide Protein Genome

- [*Escherichia (4194)*] - [*Enterobacteriaceae (3964)*] - [*Enterobacteriales (3964)*] - [*Gammaproteobacteria (3855)*] - [*Proteobacteria (1250)*]

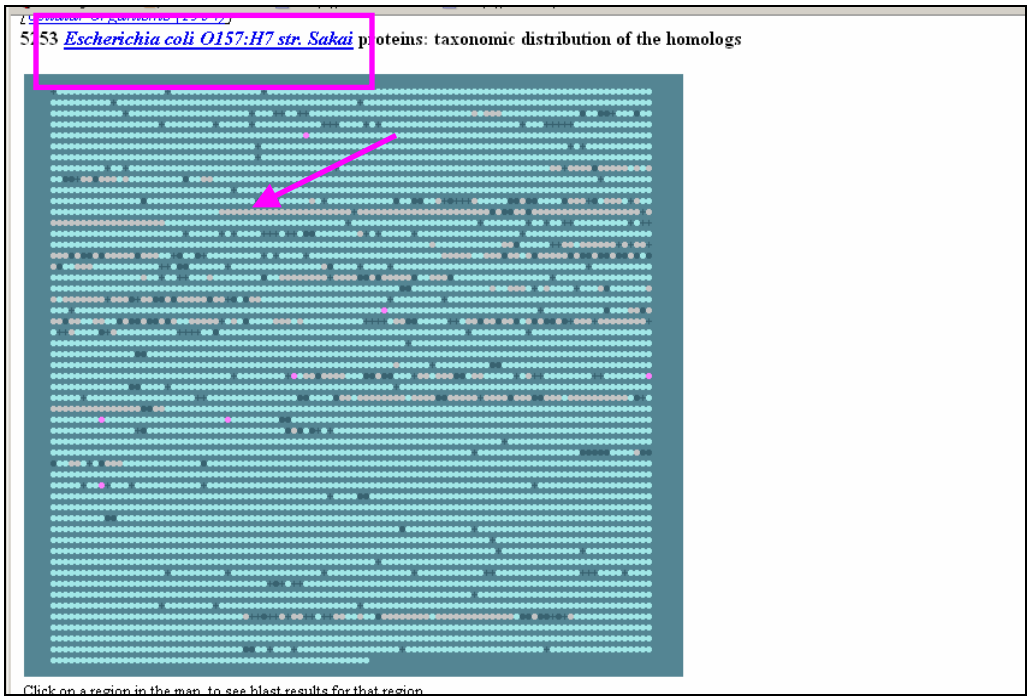
4243 ***Escherichia coli K12*** **proteins: taxonomic distribution of the homologs**



Click on a region in the map, to see blast results for that region

COGTable Cut-Off: 95 Cut-Off+: 10 Redraw HELP

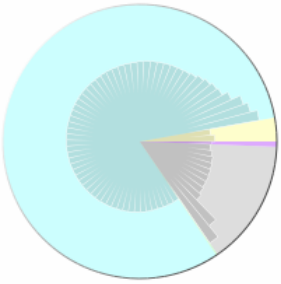
Click on the graph at left, to change the blast threshold (cut-off) used to find related sequences


5253 Escherichia coli O157:H7 str. Sakai proteins: taxonomic distribution of the homologs
Click on a region in the map, to see blast results for that region

5253 Escherichia coli O157:H7 str. Sakai proteins: taxonomic distribution of the homologs
Click on a region in the map, to see blast results for that region
COGTable
Cut-Off: 95
Cut-Off+: 10
Redraw
HELP
Gene - BLAST alignments against 'nr'; gi - link to GenBank record; 3-D - BLAST alignments against 3-D structure database

	Virus		Eukaryota		Eubacteria		Archaea			
100	91(85) V		21(0) E		78(9) B		13(0) A			
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
15830408 B	-		-		30062536	(517)	-		-	hypothetical protein ECs1154
15830409 B	55819138	(149)	50915630	(327)	82544678	(1603)	110604763	(426)	5	curved DNA-binding protein
15830412 B	-		108880530	(121)	74311559	(2138)	-		3	periplasmic glucose-1-phosphatase
15830413 B	-		-		24112416	(393)	-		-	hypothetical protein ECs1159
15830414 V	9632467	(2271)	-		16764254	(1475)	110622801	(117)	-	putative integrase
15830415 V	9632468	(504)	-		16764255	(261)	-		-	putative excisionase
15830416 V	9632469	(549)	-		-		-		-	hypothetical protein ECs1162
15830417 V	9632470	(606)	-		56414547	(429)	-		-	hypothetical protein ECs1163
15830418 V	9632471	(1101)	-		24111740	(173)	-		-	hypothetical protein ECs1164
15830419 V	9632472	(504)	-		-		-		-	hypothetical protein ECs1165
15830420 V	9632473	(375)	-		74312280	(341)	-		-	hypothetical protein ECs1166

Query: gi|15830414 putative integrase [Escherichia coli O157:H7 str. Sakai]



- 18 Archaea
- 525 Bacteria
- 0 Metazoa
- 0 Fungi
- 1 Viridiplantae
- 95 Viruses
- 4 Other Eukaryota

	Virus		Eukaryota		Eubacteria		Archaea			
100	91(85) V		21(0) E		78(9) B		13(0) A			
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
15830408 B	-		-		30062536	(517)	-		-	hypothetical protein ECs1154
15830409 B	55819138	(149)	50915630	(327)	82544678	(1603)	110604763	(426)	5	curved DNA-binding protein
15830412 B	-		108880530	(121)	74311559	(2138)	-		3	periplasmic glucose-1-phosphatase
15830413 B	-		-		24112416	(393)	-		-	hypothetical protein ECs1159
15830414 V	9632467	(2271)	-		16764254	(1475)	110622801	(117)	-	putative integrase
15830415 V	9632468	(504)	-		16764255	(261)	-		-	putative excisionase
15830416 V	9632469	(549)	-		-		-		-	hypothetical protein ECs1162
15830417 V	9632470	(606)	-		56414547	(429)	-		-	hypothetical protein ECs1163

Query: gi|15830414 putative integrase [Escherichia coli O157:H7 str. Sakai]
Matching gi: 13360620

COG0582 assigned by Cognitor (7 best hits)

Show identical | Best hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list | Run BLAST

200 BLAST hits to 148 unique species Sort by taxonomy proximity

0 Archaea | 181 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 19 Viruses | 0 Other Eukaryotae

Keep only | Cut-Off 95 | Select | Reset

New search by GI: 15830414 | Go

436 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
Conserved Domain Database hits				
2271	1	NP 859328	32171084	hypothetical protein Stx2IIp083 [Stx2 conve
1588	9	ZP 007...	75513013	COG0582: Integrase [Escherichia coli 53638]
1582	9	ZP 007...	75209455	COG0582: Integrase [Escherichia coli B171]
1582	1	NP 543053	18249865	putative integrase [Bacteriophage P27]
1541	1	ABB70738	81301988	Int protein [Phage 258-320]
1536	1	CAD24081	28803669	putative integrase [Escherichia coli phage
1497	11	NP 311040	15832267	putative integrase [Escherichia coli O157:H
1493	1	CAD88804	49523586	putative integrase [Phage BP-4795]
1475	7	NP 459869	16764254	putative integrase [Salmonella typhimurium
1389	7	NP 459980	16764365	integrase [Salmonella typhimurium LT2]
1385	7	NP 805696	29142354	integrase [Salmonella enterica subsp. enter
1373	7	YP 215950	62179533	Gifsy-2 prophage integrase [Salmonella ente
1120	9	AP 002200	89108420	predicted defective integrase [Escherichia
1116	9	NP 416096	16129537	Qin prophage; predicted defective integrase
1113	9	P76168	9910723	Putative lambdoid prophage Qin defective in
1073	7	YP 647490	108811723	hypothetical protein YPN 1560 [Yersinia pes

Query: gi|15830414 putative integrase [Escherichia coli O157:H7 str. Sakai]
Matching gi: 13360620

COG0582 assigned by Cognitor (7 best hits)

Show identical | All hits | Common Tree | Taxonomy Report | 3D structures | CDD-Search | GI list | Run BLAST

200 BLAST hits to 148 unique species Sort by taxonomy proximity

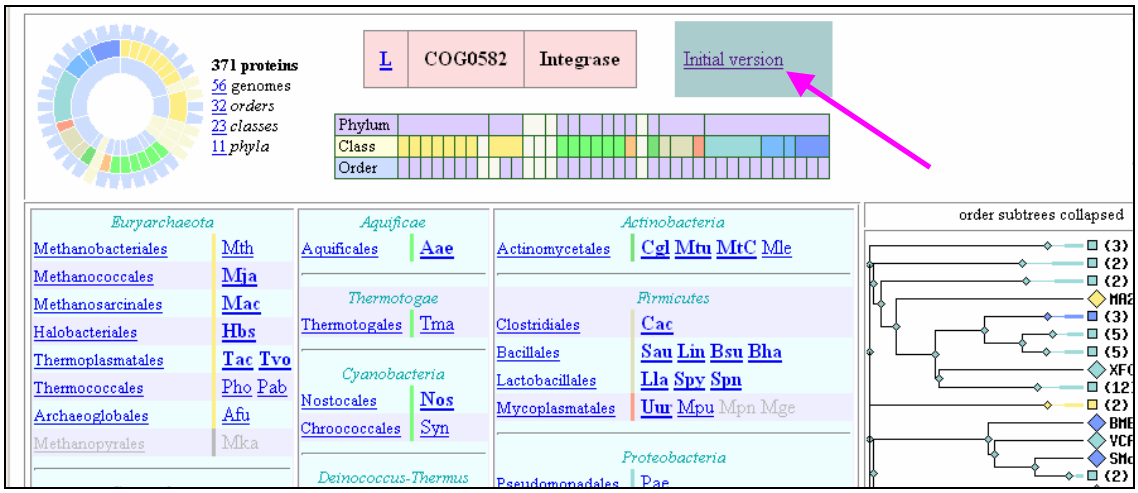
0 Archaea | 181 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 19 Viruses | 0 Other Eukaryotae

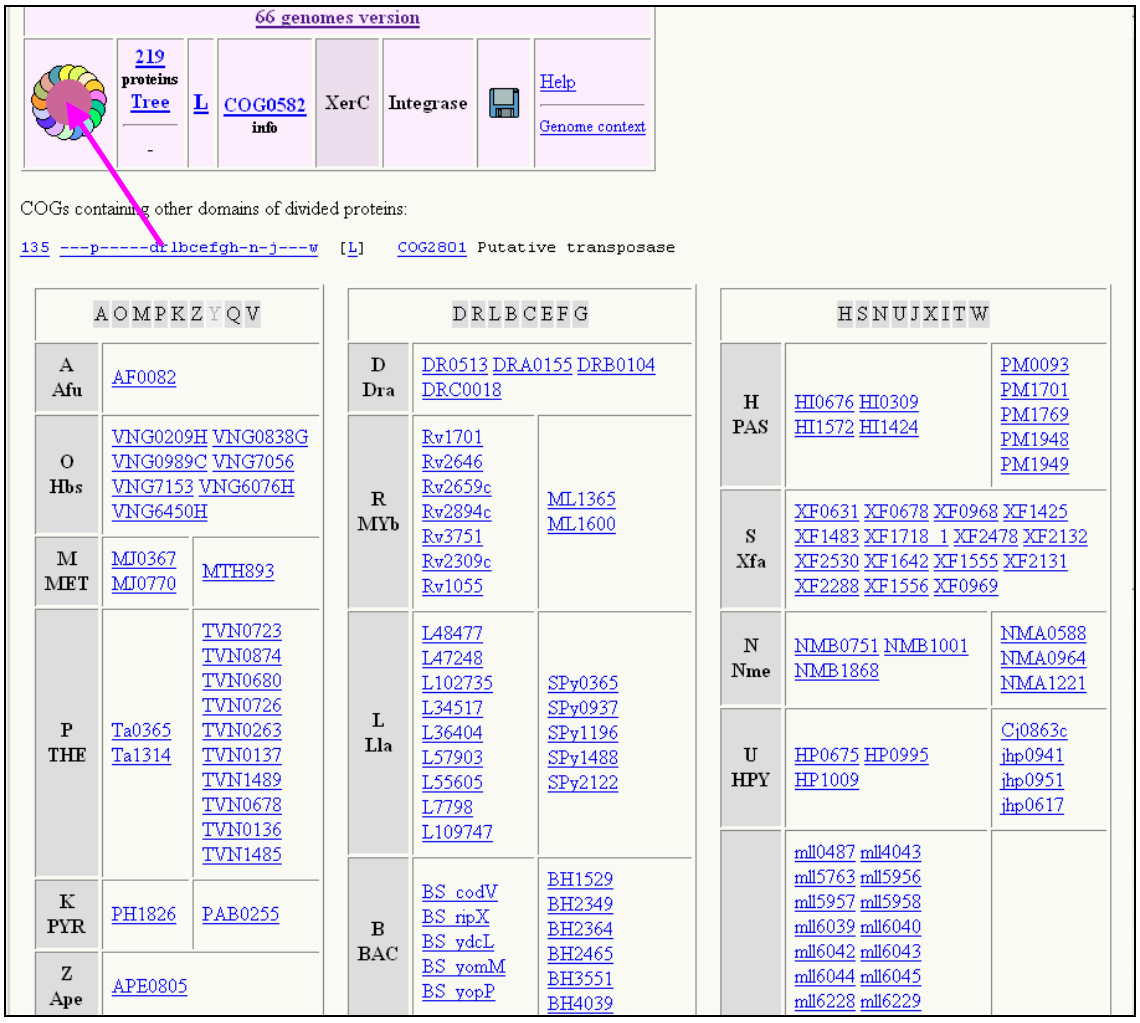
Keep only | Cut-Off 95 | Select | Reset

New search by GI: 15830414 | Go

436 aa

SCORE	P	ACCESSION	GI	N	ORGANISM
Conserved Domain Database hits					
2271	1	NP 859328	32171084	1	*Stx2 converting bacteriophage II*
1588	9	ZP 007...	75513013	1	*Escherichia coli 53638*
1582	9	ZP 007...	75209455	1	*Escherichia coli B171*
1582	1	NP 543053	18249865	1	*Bacteriophage P27*
1541	1	ABB70738	81301988	1	*Phage 258-320*
1536	1	CAD24081	28803669	1	*Escherichia coli phage EH297*
1497	11	NP 311040	15832267	3	*Escherichia coli O157:H7 str. Sakai*
1493	1	CAD88804	49523586	1	*Phage BP-4795*
1475	7	NP 459869	16764254	2	*Salmonella typhimurium LT2*
1385	7	NP 805696	29142354	1	*Salmonella enterica subsp. enterica ser*
1373	7	YP 215950	62179533	1	*Salmonella enterica subsp. enterica ser*
1120	9	AP 002200	89108420	2	*Escherichia coli W3110*
1116	9	NP 416096	16129537	1	*Escherichia coli K12*
1113	9	P76168	9910723	1	*Escherichia coli*
1073	7	YP 647490	108811723	1	*Yersinia pestis Nepal516*
1063	7	YP 311068	74312649	1	*Shigella sonnei Ss046*
997	7	ZP 011...	89103031	1	*Yersinia pestis biovar Orientalis str.*
906	5	YP 129516	54308496	1	*Photobacterium profundum SS9*
887	7	YP 689623	110806103	2	*Shigella flexneri 5 str. 8401*
852	5	ZP 012...	90414999	2	*Photobacterium profundum 3TCK*


371 proteins
56 genomes
32 orders
23 classes
11 phyla
L
COG0582
Integrase
Initial version
Phylum
Class
Order
Euryarchaeota
Methanobacteriales Mth
Methanococcales Mja
Methanosarcinales Mac
Halobacteriales Hbs
Thermoplasmatales Tac Tvo
Thermococcales Pho Pab
Archaeoglobales Afu
Methanopyrales Mka
Aquificae
Aquificales Aae
Thermotogae
Thermotogales Tma
Cyanobacteria
Nostocales Nos
Chroococcales Syn
Deinococcus-Thermus
Actinobacteria
Actinomycetales Cgl Mtu MtC Mle
Firmicutes
Clostridiales Cac
Bacillales Sau Lin Bsu Bha
Lactobacillales Lla Spy Spn
Mycoplasmatales Uur Mpu Mpn Mge
Proteobacteria
Pseudomonadales Pae
order subtrees collapsed


66 genomes version
219 proteins
Tree
L
COG0582
info
XerC
Integrase
Help
Genome context
COGs containing other domains of divided proteins:
135 ---p-----drlbcefgh-n-j---w [L] COG2801 Putative transposase
A O M P K Z Y Q V
A Afu AF0082
O Hbs VNG0209H VNG0838G VNG0989C VNG7056 VNG7153 VNG6076H VNG6450H
M MET MJ0367 MJ0770 MTH893
P THE Ta0365 Ta1314 TVN0723 TVN0874 TVN0680 TVN0726 TVN0263 TVN0137 TVN1489 TVN0678 TVN0136 TVN1485
K PYR PH1826 PAB0255
Z Ape APE0805
D R L B C E F G
D Dra DR0513 DRA0155 DRB0104 DRC0018
R MYb Rv1701 Rv2646 Rv2659c Rv2894c Rv3751 Rv2309c Rv1055 ML1365 ML1600
L Lla L48477 L47248 L102735 L34517 L36404 L57903 L55605 L7798 L109747 SPy0365 SPy0937 SPy1196 SPy1488 SPy2122
B BAC BS_codV BS_ripX BS_ydcL BS_yomM BS_yopP BH1529 BH2349 BH2364 BH2465 BH3551 BH4039
H S N U J X I T W
H PAS HI0676 HI0309 HI1572 HI1424 PM0093 PM1701 PM1769 PM1948 PM1949
S Xfa XF0631 XF0678 XF0968 XF1425 XF1483 XF1718_1 XF2478 XF2132 XF2530 XF1642 XF1555 XF2131 XF2288 XF1556 XF0969
N Nme NMB0751 NMB1001 NMB1868 NMA0588 NMA0964 NMA1221
U HPY HP0675 HP0995 HP1009 Cj0863c jhp0941 jhp0951 jhp0617
mll0487 mll4043 mll5763 mll5956 mll5957 mll5958 mll6039 mll6040 mll6042 mll6043 mll6044 mll6045 mll6228 mll6229

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

Science 1997 Oct 24;278(5338):631-7,
Nucleic Acids Res 2001 Jan 1; 29(1):22-28.

Help

COGnitor

Protein/Gene name: Text search:

Code	Name	Proteins	in COGs
A	*Archaeoglobus fulgidus*	2420	1872
O	*Halobacterium sp. NRC-1*	2605	1701
M	*Methanococcus jannaschii*	1786	1330
	Methanobacterium thermoautotrophicum	1873	1388
P	*Thermoplasma acidophilum*	1482	1230
	Thermoplasma volcanium	1499	1243
K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
Z	*Aeropyrum pernix*	1841	1178
Y	*Saccharomyces cerevisiae*	5955	2290
Q	*Aquifex aeolicus*	1560	1329
V	*Thermotoga maritima*	1858	1527
D	*Deinococcus radiodurans*	3187	2226
R	*Mycobacterium tuberculosis*	3927	2585
	Mycobacterium leprae	1605	1134
L	*Lactococcus lactis*	2267	1618
	Streptococcus pyogenes	1697	1211
B	*Bacillus subtilis*	4118	2870
	Bacillus halodurans	4066	2878
C	*Synechocystis*	3167	2159
E	*Escherichia coli K12*	4275	3414
	Escherichia coli O157	5315	3662
	Buchnera sp. APS	575	568
F	*Pseudomonas aeruginosa*	5567	4392

Principal component analysis of genomes

List of COGs

Distribution

Co-occurrences

Phylogenetic patterns

Phylogenetic patterns search

Functional categories

J K L

D O M N P T

G C E F H I Q

R S

Pathways and functional systems

FTP

2223 Phylogenetic patterns in COGs Group species: Help

AEB		AB		B		A	
452	622	785	891	780	1284	34	196
aompkzyqvdrlbcefghsnujxitw	84	aomp------r--c-f-----j----	6	----------e-gh--------	41	a-m----------	39
aompkzyqvdrlbcefghsnuj----	11	aompkz-q-dr-bcefgh--uj----	6	-------qvdrlbcefghsnujxitw	30	aompkz----------	28
aompk-yqvdrlbcefghsnuj----	8	a-m-----v----------	4	----------efg----j----	22	a-m-k----------	16
aom---yqvdrlbcefghsnuj----	7	a-m-----v----c----------	4	----------efg--------	18	--m-k----------	12
aompkzyqvdrlbcefghsnujxit-	7	aom-kz-qvdrlbcefghs-uj--tw	4	----------efgh--------	18	aom-kz----------	12
aompkzyqvdrlbcefghsnuj-i--	6	aompkz-qvdrlbcefghsnujxitw	4	----------f-----j----	15	a-m-kz----------	11
aomp--yqvdrlbcefghsnuj----	5	--m--------lb----------	3	-------qvdrlbcefghsnujxit-	15	aompk----------	10
aomp-zyq-dr-bcefghsnujxi--	5	--m----qvdrlbcefghsnujxit-	3	----------fg----j----	13	aom----------	9
-omp--yqvdrlbcefghsnujxitw	4	--m-k-------b----------	3	----------ef-----j----	13	aom-k----------	7
a-m-kzyqvdrlbcefghsnuj----	4	--mpk---v-r----------	3	----------efghsn------	11	a---kz----------	6
aompkzy--d-l----------it-	4	-o----------e-gh--------	3	-------qv---b-efg---uj--t-	10	a-mpkz----------	6
aompkzyqvdrlbcefghsnuj-itw	4	-o----------b-efg----j----	3	-------qv---b-efg---uj-it-	9	aomp----------	5
-o-pkzyqvdr-bcef--sn-j----	3	-o--------r-b----------	3	----------ef--s-------	8	-ompk----------	4
a-m---yqvdrlbcefg-s-uj----	3	-om----qvdrlbcefghsnujxitw	3	----------efghs-------	8	ao-p----------	4
aom--zyqvdrlbcefghsnuj-i--	3	-om-k---vd--b--fg-----x---	3	----------r----f-----j----	8	-om-k----------	3
aompkzyq----------	3	a-m----------c----------	3	----------h-nu-----	7	ao-pk----------	3
aompkzyqvdrlbcefghsnuj-it-	3	a-m-k---v----------	3	----------efgh-n------	7	--m-kz----------	2
aompkzyqvdrlbcefghsnujx-t-	3	a-m-k--q-----ce-----u-----	3	----------r----------j----	7	-om----------	2
----k-y--dr-b-efgh-nuj----	2	a-m-kz-q----------	3	----------jx---	6	a-m--z----------	2
-o----y------cefghsn-jx---	2	a-m-kz-qv----------	3	----------c-f-----j----	6		
-o---zyq-dr-bcef--s--jx---	2	ao--kz--vd-lb--f-----j--tw	3	----------lb----------w	6	15 single patterns	
-o-p-zy--drlbc-f-----jxi-w	2	ao-p-z-q-dr--cef--snujx---	3	----------s-ujx---	5		
-o-p-zyq-dr-bc-fg-snujx---	2	ao-pkz---dr-b-e------j----	3	----------fg--n------	5		
-om---y--drlbcefghsnuj--tw	2	ao-pkz-qvdr--cef--snujx---	3	----------e--h-n------	5		
-ompkzyqvdrlbcefghsnuj-it-	2	ao-pkz-qvdrlbcefgh--uj-itw	3	----------e-gh---j----	5		
a-m-k-yqvdrlbcefghsnuj----	2	aompkz-qvdrlbcefghsnuj----	3	----------efg-s-------	5		
a-mpkzyq-d--bcefgh-nujxi--	2	-----z---dr-b-ef-----j--t-	2	----------efghsn-j----	5		
ao--k-yqvdrlbcefghsnujxi--	2	----k----d-lb-efgh-n------	2	----------lb-e----------	5		
ao-p--yqvdrlbcefghsnujxitw	2	----k--qv----------	2	-------qvdrlbcefghsnujx-t-	5		
ao-p-zy-vdr-b-ef--sn-j----	2	---p--------b-e-----uj----	2	----------hs--j----	4		
aom---yqvdrlbcefghsnujxit-	2	--m----------c-------j----	2	----------f--s--j----	4		
aom--zyqvdrlbc-f---nujxitw	2	--m---------b----------	2	----------e-gh-n------	4		

AE	
11	102
aompkzy----------	72
aom-kzy----------	8
a-m-kzy----------	7
a-mpkzy----------	4
aompk-y----------	3
----kzy----------	2

84 COGs Protein/Gene name: Select Help

		A O M P K Z Y Q V D R L B C E F G H S N U J X I T W				
phy:		a o m p k z y q v d r l b c e f g h s n u j x i t w				
-	43	aompkzyqvdrlbcefghsnujxitw	AlaS	[J]	COG0013	Alanyl-tRNA synthetase
-	45	aompkzyqvdrlbcefghsnujxitw	ArgS	[J]	COG0018	Arginyl-tRNA synthetase
-	49	aompkzyqvdrlbcefghsnujxitw	KsgA	[J]	COG0030	Dimethyladenosine transferase (rRNA methylation)
-	78	aompkzyqvdrlbcefghsnujxitw	TufB	[JE]	COG0050	GTPases - translation elongation factors
-	71	aompkzyqvdrlbcefghsnujxitw	GlnS	[J]	COG0008	Glutamyl- and glutaminyl-tRNA synthetases
-	49	aompkzyqvdrlbcefghsnujxitw	HisS	[J]	COG0124	Histidyl-tRNA synthetase
-	44	aompkzyqvdrlbcefghsnujxitw	IleS	[J]	COG0060	Isoleucyl-tRNA synthetase
-	45	aompkzyqvdrlbcefghsnujxitw	LeuS	[J]	COG0495	Leucyl-tRNA synthetase
-	51	aompkzyqvdrlbcefghsnujxitw	Map	[J]	COG0024	Methionine aminopeptidase
-	45	aompkzyqvdrlbcefghsnujxitw	MetG	[J]	COG0143	Methionyl-tRNA synthetase
-	44	aompkzyqvdrlbcefghsnujxitw	PheS	[J]	COG0016	Phenylalanyl-tRNA synthetase alpha subunit
-	44	aompkzyqvdrlbcefghsnujxitw	PheT	[J]	COG0072	Phenylalanyl-tRNA synthetase beta subunit
-	44	aompkzyqvdrlbcefghsnujxitw	ProS	[J]	COG0442	Prolyl-tRNA synthetase
-	42	aompkzyqvdrlbcefghsnujxitw	TruB	[J]	COG0130	Pseudouridine synthase
-	45	aompkzyqvdrlbcefghsnujxitw	RplA	[J]	COG0081	Ribosomal protein L1
-	45	aompkzyqvdrlbcefghsnujxitw	RplJ	[J]	COG0244	Ribosomal protein L10
-	46	aompkzyqvdrlbcefghsnujxitw	RplK	[J]	COG0080	Ribosomal protein L11
-	45	aompkzyqvdrlbcefghsnujxitw	RplM	[J]	COG0102	Ribosomal protein L13
-	45	aompkzyqvdrlbcefghsnujxitw	RplN	[J]	COG0093	Ribosomal protein L14
-	44	aompkzyqvdrlbcefghsnujxitw	Rpl0	[J]	COG0200	Ribosomal protein L15
-	44	aompkzyqvdrlbcefghsnujxitw	RplP	[J]	COG0197	Ribosomal protein L16/L10E
-	43	aompkzyqvdrlbcefghsnujxitw	RplR	[J]	COG0256	Ribosomal protein L18
-	45	aompkzyqvdrlbcefghsnujxitw	RplB	[J]	COG0090	Ribosomal protein L2
-	45	aompkzyqvdrlbcefghsnujxitw	RplV	[J]	COG0091	Ribosomal protein L22
-	43	aompkzyqvdrlbcefghsnujxitw	RplW	[J]	COG0089	Ribosomal protein L23
-	44	aompkzyqvdrlbcefghsnujxitw	RplX	[J]	COG0198	Ribosomal protein L24
-	44	aompkzyqvdrlbcefghsnujxitw	RpmC	[J]	COG0255	Ribosomal protein L29
-	44	aompkzyqvdrlbcefghsnujxitw	RplC	[J]	COG0087	Ribosomal protein L3
-	45	aompkzyqvdrlbcefghsnujxitw	RplD	[J]	COG0088	Ribosomal protein L4
-	45	aompkzyqvdrlbcefghsnujxitw	RplE	[J]	COG0094	Ribosomal protein L5
-	45	aompkzyqvdrlbcefghsnujxitw	RplF	[J]	COG0097	Ribosomal protein L6
-	45	aompkzyqvdrlbcefghsnujxitw	RpsJ	[J]	COG0051	Ribosomal protein S10
-	45	aompkzyqvdrlbcefghsnujxitw	RpsK	[J]	COG0100	Ribosomal protein S11
-	46	aompkzyqvdrlbcefghsnujxitw	RpsL	[J]	COG0048	Ribosomal protein S12

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in 43 complete genomes, representing 30 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

Science 1997 Oct 24;278(5338):631-7,
Nucleic Acids Res 2001 Jan 1; 29(1):22-28.

Help

COGnitor

Protein/Gene name: Text search:

Code	Name	Proteins	in COGs
A	*Archaeoglobus fulgidus*	2420	1872
O	*Halobacterium sp. NRC-1*	2605	1701
M	*Methanococcus jannaschii*	1786	1330
	Methanobacterium thermoautotrophicum	1873	1388
P	*Thermoplasma acidophilum*	1482	1230
	Thermoplasma volcanium	1499	1243
K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
Z	*Aeropyrum pernix*	1841	1178
Y	*Saccharomyces cerevisiae*	5955	2290

Principal component analysis of genomes
List of COGs
Distribution
Co-occurrences
Phylogenetic patterns
Phylogenetic patterns search

Functional categories

J K L

Phylogenetic patterns search

NCBI

Search dc - don't care Help

Org	dc	Yes	No	Org	dc	Yes	No	Org	dc	Yes	No
A Afu	●	○	○	D Dra	●	○	○	H Hin	●	○	○
O Hbs	●	○	○	R Mtu	●	○	○	S Xfa	●	○	○
M Met	●	○	○	B BAC	●	○	○	N Nme	●	○	○
P THE	●	○	○	L STR	●	○	○	U HPY	●	○	○
K PYR	●	○	○	C Ssp	●	○	○	J Cje	●	○	○
Z Ape	●	○	○	E ENT	●	○	○	X Rpr	●	○	○
Y Sce	●	○	○	F Pae	●	○	○	I CLA	●	○	○
Q Aae	●	○	○	G Vch	●	○	○	T SPI	●	○	○
V Tma	●	○	○					W Myc	●	○	○

Differences in closely related genomes

J ALPHA	*Mesorhizobium loti*	vs.	*Caulobacter crescentus*
L STR	*Lactococcus lactis*	vs.	*Streptococcus pyogenes*
P THE	*Thermoplasma acidophilum*	vs.	*Thermoplasma volcanium*
H PAS	*Haemophilus influenzae*	vs.	*Pasteurella multocida*
M MET	*Methanococcus jannaschii*	vs.	*Methanobacterium thermoautotrophicum*
R MYb	*Mycobacterium tuberculosis*	vs.	*Mycobacterium leprae*
N NEI	*Neisseria meningitidis MC58*	vs.	*Neisseria meningitidis Z2491*
K PYR	*Pyrococcus horikoshii*	vs.	*Pyrococcus abyssi*
B BAC	*Bacillus subtilis*	vs.	*Bacillus halodurans*
E ENT	*Escherichia coli*	vs.	*Buchnera sp. APS*
U HPY	*Helicobacter pylori 26695*	vs.	*Helicobacter pylori J99*
I CLA	*Chlamydia trachomatis*	vs.	*Chlamydia pneumoniae*
W MYC	*Mycoplasmas*	vs.	*Ureaplasma urealyticum*
T SPI	*Treponema pallidum*	vs.	*Borrelia burgdorferi*

Escherichia coli k12	-	x	-	x	-	x	-	x	**E** - Enterobacteriaceae
Escherichia coli O157	-	-	x	x	-	-	x	x	
Buchnera sp. APS	-	-	-	-	x	x	x	x	
Function	1290	55	71	1339	2	1	-	549	
J K L	191	8	10	153	-	-	-	171	**Information storage and processing**
J	81	-	-	20	-	-	-	116	Translation, ribosomal structure and biogenesis
K	50	2	4	61	-	-	-	15	Transcription
L	60	6	6	72	-	-	-	40	DNA replication, recombination and repair
D O M N P T	185	9	14	346	-	-	-	133	**Cellular processes**
D	7	-	-	15	-	-	-	10	Cell division and chromosome partitioning
O	29	-	3	46	-	-	-	32	Posttranslational modification, protein turnover, chaperones
M	36	3	1	88	-	-	-	27	Cell envelope biogenesis, outer membrane
N	43	4	2	43	-	-	-	41	Cell motility and secretion
P	31	1	6	104	-	-	-	18	Inorganic ion transport and metabolism
T	39	1	2	50	-	-	-	5	Signal transduction mechanisms
C G E F H I Q	249	14	10	496	1	1	-	233	**Metabolism**
C	74	-	1	104	-	-	-	45	Energy production and conversion
G	33	5	2	98	-	1	-	32	Carbohydrate transport and metabolism
E	47	2	4	111	1	-	-	68	Amino acid transport and metabolism
F	14	1	-	40	-	-	-	30	Nucleotide transport and metabolism
H	33	1	-	81	-	-	-	39	Coenzyme metabolism
I	20	2	1	38	-	-	-	14	Lipid metabolism

71 COGs Protein/Gene name: Select Help

Only *Escherichia coli 0157*

-	5	----------l--e--h-n------	-	[K]	COG3561	Phage anti-repressor protein
-	24	----------l--ef-hsn-j----	-	[K]	COG2932	Predicted transcriptional regulator
-	22	aompkz----r-b-e---s--j----	-	[K]	COG3609	Predicted transcriptional regulators containing the CopG/
-	14	----------l--ef-hs-------	-	[K]	COG3617	Prophage antirepressor
-	8	----------lb-e------j----	XtmA	[L]	COG3728	Phage terminase, small subunit
-	5	---p-----d----e----n------	-	[L]	COG3440	Predicted restriction endonuclease
-	33	---p------rlb-e------j---w	-	[L]	COG3328	Predicted transposase
-	60	a--p-----d--bce----n-j----	-	[L]	COG3335	Predicted transposase
-	47	-o-p-----d-l--e------j----	-	[L]	COG3436	Transposase
-	29	-omp-----d--b-e----nuj--t-	-	[L]	COG1002	Type II restriction enzyme, methylase subunits
-	11	---------d--bcef-h--uj---w	UreE	[O]	COG2371	Urease accessory protein UreE
-	12	---------dr-bcef-h--uj---w	UreF	[O]	COG0830	Urease accessory protein UreF
-	11	---------d--bcef-h--uj---w	UreH	[O]	COG0829	Urease accessory protein UreH
-	22	----------ef-hsn------	FhaB	[M]	COG3210	Putative hemagglutinin/hemolysin
-	19	----------cef-hsnuj----	FhaC	[N]	COG2831	Hemolysin activation/secretion protein
-	25	----------r---efg-s--j----	ParE	[N]	COG3668	Plasmid stabilization system protein
-	51	aom-kz--vdrlbce-gh------tw	CbiO	[P]	COG1122	ABC-type cobalt transport system, ATPase component
-	31	aom-kz--vdrlbce--h-------w	CbiQ	[P]	COG0619	ABC-type cobalt transport system, permease component CbiC
-	8	----------b-ef----------	-	[P]	COG3546	Mn-containing catalase
-	3	----------ef-----j----	HemS	[P]	COG3720	Putative heme degradation protein
-	3	----------e-gh--------	HugX	[P]	COG3721	Putative heme iron utilization protein
-	12	---------d----e------j----	TerB	[P]	COG3793	Tellurite resistance protein
-	6	----------efg----j----	-	[T]	COG3456	FHA-domain-containing protein
-	21	---------d--bce----------	TerZ	[T]	COG2310	Uncharacterized proteins involved in stress response, hom
-	10	------y-v--lb-e-gh--------	SacC	[G]	COG1621	Beta-fructosidases (levanase/invertase)
-	26	--m---y--dr-bcefg----j----	-	[GC]	COG1819	Glycosyl transferases, related to UDP-glucuronosyltransfe
-	4	-o----------e------j----	Mal	[E]	COG3799	Methylaspartate ammonia-lyase
-	13	---------dr-bcef-h--uj---w	UreC	[E]	COG0804	Urea amidohydrolase (urease) alpha subunit
-	13	---------dr-bcef-h--uj---w	UreB	[E]	COG0832	Urea amidohydrolase (urease) beta subunit
-	13	---------dr-bcef-h--uj---w	UreA	[E]	COG0831	Urea amidohydrolase (urease) gamma subunit
-	49	----------rlb-ef-hsn-j----	-	[I]	COG1835	Predicted acyltransferases
-	19	----------lbcefg-sn-j---w	SunT	[Q]	COG2274	ABC-type bacteriocin/lantibiotic exporters, contain an N-
-	4	----------b-ef----------	-	[Q]	COG3435	Gentisate 1,2-dioxygenase
-	14	a-m-k-y--dr-b-e----------	-	[R]	COG1205	Distinct helicase family with a unique C-terminal domain
-	7	----------rl--e------j----	-	[R]	COG3740	Phage head maturation protease
-	4	----------ef--s-------	-	[R]	COG3500	Phage protein D

Problem 2:

Access the Microbial Genomes Page (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi). How many complete microbial genome sequences are available? How many are in progress and how many are in the assembly stage? From the "Organism Group" pull down menu, select Spirochaetes. The complete genome sequences are available for how many Spirochaetes?

Go back to the Complete genomes Page by selecting "All Organisms" in the "Organism Group" pull down menu. How many Rickettsia genomes are completely sequenced? Note their sizes. Select the organism Rickettsia prowazekii str. Madrid E. Is the organism pathogenic in humans? How are the bacteria transmitted to humans? Click on the RefSeq accession number NC_000963.

From the resulting table, access the entry NC_000963 in the nucleotide database. Access the Proteins link. How many protein coding genes have been annotated on the genome? What do you think about the number of proteins compared to some other bacteria such as Escherichia coli? Go back to the Genomes Page. Access the Protein table that contains locations and detailed information such as the strand, product name, FASTA sequence etc. from the Protein coding link. Also, access and download the RNA table through the Structural RNA link. How many genes encoding arginine tRNA are annotated? Go back to the Genome page.

Click on the gene lpxD in the genome map at the bottom of the page. What conserved domain is found in the protein? What is the pathway in which it is involved? Go back to the Genome page.

Click on the COG link. It provides an assignment of each protein to one or more COG functional categories. Note the variations in the % in genome category compared to the % in Bacteria category. What would be its importance? Download the protein table for the proteins assigned to Cell wall/membrane biogenesis.

Go back to the genome page. Click on the Tax Plot link. Tax Plot allows a three-way comparison of three different proteomes, including proteins encoded on all chromosomes and plasmids. Choose Rickettsia typhi str. Wilmington and Rickettsia conorii str. Malish 7 from the pull down menu. Which Rickettsia species is more similar to Rickettsia prowazekii str. Madrid E? Now select the organisms Oryza sativa (taxonomy id 4530) and Escherichia coli K12. What does the result indicate?

Access the GenePlot link. GenePlot uses precalculated BLAST results to determine the top hits in two organisms. The best hits of each protein from two

genomes are calculated and plotted based on chromosomal location. Download all best hits between the Rickettsia prowazekii str. Madrid E and Rickettsia typhi str. Wilmington proteins. Zoom in the plot .125X. Note the inversion in the two genomes. Which region of the genome is inverted? Go back to Genome page.

Access the TaxMap link. It reports the taxonomic distribution of the top BLAST hits for each of the protein in the entire proteome by comparing to the non-redundant database (excluding closely related species) by BLAST similarity. Exclude all hits that are most similar to Bacteria. What does the graph now indicate? Click on the first line. Search for the tlc1 gene in the result. This protein is most similar to the protein from which lineage? Click on the gi number 15603932 for the protein. This link gives the top 200 protein BLAST similarity hits for the query protein. Select the Best hits option. Which organisms are represented in the similar proteins? The protein is most homologous to protein from which organism after bacteria? Select the COG3202 link at the top to access the tabular format of the proteins assigned to that COG. What are the source organisms for the homologs to this protein? Access the Initial Version of the COG database by clicking on the link on the left. Select the Phylogenetic Patterns search. Obtain a list of COGs that are absent in Rickettsia. What are the major functional categories?