Mini Courses NCBI Resources one course at a time

Microbial Genomes Quick Start

An NCBI Mini-Course

In this mini-course, we will learn how to access the microbial genome sequences and annotations, how to navigate through and download the gene and protein datasets, and will be introduced to the available genomic and comparative genomic analysis tools. 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 EDL933 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 by selecting "All Organisms" in the "Organism Group" pull down menu . Select the organism Escherichia coli O157:H7 EDL933. How many strains of Escherichia coli have been completely sequenced? Is the 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_002655.

From the resulting table, access the entry NC_002655 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 EDL933 strain?

Return to Genome page. Click on the gene link for thrC 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 %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 EDL933 strain proteins. Open the file in Excel and select the gi column for the EDL933 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 EDL933 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 EDL933 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 EDL933 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 EDL933 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 –

T - Tools; M - TaxMap; P - PartTable; C - COGTable; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; 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	#ch	#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.

644 Microbial Genomes In-Progress selected: [A] - 28, [B] - 616 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
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
Actinomyces naeslundii MG1	B	Actinobacteria	-	*3.06	-	-	-	TIGR
Aeromonas salmonicida subsp. salmonicida A449	B	Gammaproteobacteria	-	*5	54.6	-	-	National Research Council Canada - Institute of Marine Biosciences
Agrobacterium radiobacter K84	B	Alphaproteobacteria	-	-	-	-	-	University of Washington
Agrobacterium vitis S4	B	Alphaproteobacteria	-	-	-	-	-	University of Washington
Akkermansia muciniphila	B	Other Bacteria	-	-	47.6	-	-	DOE Joint Genome Institute
Alkaliphilus metalliredigenes QYMF	B	Firmicutes	263	*4	36.6	NZ_AAKU00000000	BLAST	DOE Joint Genome Institute
Alteromonas macleodii 'Deep ecotype'	B	Gammaproteobacteria	111	*3.5	44.9	NZ_AAOD00000000	BLAST	J. Craig Venter 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 –

Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria

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

* size is estimated, based on existing sequences

sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

644 Microbial Genomes In-Progress selected: [A] - 28, [B] - 616 save

	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 hospitalis	A	Crenarchaeota	-	*1.8	- -		-	University of Copenhagen
Acidiphilium cryptum	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
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
Actinomyces naeslundii MG1	B	Actinobacteria	-	*3.06	- -		-	TIGR
Aeromonas salmonicida subsp. salmonicida A449	B	Gammaproteobacteria	-	*5	54.6 -		-	National Research Council Canada - Institute of Marine Biosciences
Agrobacterium radiobacter K84	B	Alphaproteobacteria	-	-	- -		-	University of Washington
Agrobacterium vitis S4	B	Alphaproteobacteria	-	-	- -		-	University of Washington
Akkermansia muciniphila	B	Other Bacteria	-	-	47.6 -		-	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: Firmicutes

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

sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

147 Microbial Genomes In-Progress selected: [A] - 0, [B] - 147 clear filter save

Organism	King	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_AAEC00000000	BLAST	TIGR
Bacillus anthracis str. A2012	B	Firmicutes	1	*5.37	34.8	NZ_AAAC00000000	BLAST	TIGR
Bacillus anthracis str. Australia 94	B	Firmicutes	49	*5.5	35.2	NZ_AAES00000000	BLAST	TIGR
Bacillus anthracis str. CNEVA-9066	B	Firmicutes	30	*5.49	35.2	NZ_AAEN00000000	BLAST	TIGR
Bacillus anthracis str. Kruger B	B	Firmicutes	64	*5.47	35.1	NZ_AAEQ00000000	BLAST	TIGR
Bacillus anthracis str. Vollum	B	Firmicutes	52	*5.39	35.2	NZ_AAEP00000000	BLAST	TIGR
Bacillus anthracis str. Western North America USA6153	B	Firmicutes	44	*5.51	35.2	NZ_AAER00000000	BLAST	TIGR
Bacillus cereus AH1134	B	Firmicutes	-	-	- -		-	TIGR
Bacillus cereus AH820	B	Firmicutes	-	-	- -		-	TIGR
Bacillus cereus B4264	B	Firmicutes	-	-	- -		-	TIGR
Bacillus cereus F0837/76	B	Firmicutes	-	-	35.4 -		-	Genoscope
Bacillus cereus G9241	B	Firmicutes	207	*5.93	35.2	NZ_AAEK00000000	BLAST	TIGR
Bacillus cereus G9842	B	Firmicutes	-	-	- -		-	TIGR
Bacillus cereus Q1	B	Firmicutes	-	-	38 -		-	Microbial Genome Center of ChMPH
Bacillus cereus W	B	Firmicutes	-	-	35.5 -		-	TIGR
Bacillus cereus subsp. cytotoxis NVH 391-98	B	Firmicutes	114	-	35.8	NZ_AALL00000000	BLAST	DOE Joint Genome Institute
Bacillus coagulans 36D1	B	Firmicutes	-	-	46 -		-	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: Firmicutes

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

sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

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

147 Microbial Genomes In-Progress selected: [A] - 0, [B] - 147 clear filter save

Organism	King	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_AAEC00000000	BLAST	TIGR
Bacillus anthracis str. A2012	B	Firmicutes	1	*5.37	34.8	NZ_AAAC00000000	BLAST	TIGR
Bacillus anthracis str. Australia 94	B	Firmicutes	49	*5.5	35.2	NZ_AAES00000000	BLAST	TIGR
Bacillus anthracis str. CNEVA-9066	B	Firmicutes	30	*5.49	35.2	NZ_AAEN00000000	BLAST	TIGR
Bacillus anthracis str. Kruger B	B	Firmicutes	64	*5.47	35.1	NZ_AAEQ00000000	BLAST	TIGR
Bacillus anthracis str. Vollum	B	Firmicutes	52	*5.39	35.2	NZ_AAEP00000000	BLAST	TIGR
Bacillus anthracis str. Western North America USA6153	B	Firmicutes	44	*5.51	35.2	NZ_AAER00000000	BLAST	TIGR
Bacillus cereus AH1134	B	Firmicutes	-	-	- -	-	-	TIGR
Bacillus cereus AH820	B	Firmicutes	-	-	- -	-	-	TIGR
Bacillus cereus B4264	B	Firmicutes	-	-	- -	-	-	TIGR
Bacillus cereus F0837/76	B	Firmicutes	-	-	35.4	-	-	Genoscope
Bacillus cereus G9241	B	Firmicutes	207	*5.93	35.2	NZ_AAEK00000000	BLAST	TIGR
Bacillus cereus G9842	B	Firmicutes	-	-	- -	-	-	TIGR

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 legend: T - TaxMap; P - ProfTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; 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	5.05	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

Erwinia carotovora subsp. atroseptica SCRI1043	B	Gammaproteobacteria	5.06	51	1		BX950851	NC_004547	07/21/2004	Welcome Trust Sanger Institute	T	P	C	D
Erythrobacter litoralis HTCC2594	B	Alphaproteobacteria	3.05	63.1	1		CP000157	NC_007722	03/11/2005	J. Craig Venter Institute	T	P	C	D
Escherichia coli CFT073	B	Gammaproteobacteria	5.23	50	1		AE014075	NC_004431	12/10/2002	University of Wisconsin-Madison, USA	T	P	C	D
Escherichia coli K12	B	Gammaproteobacteria	4.64	50	1		U00096	NC_000913	09/05/1997	University of Wisconsin-Madison, USA	T	P	C	D
Escherichia coli O157:H7 EDL933	B	Gammaproteobacteria	5.62	50	1	1	AE005174	NC_002655	02/24/2001	University of Wisconsin-Madison, USA	T	P	C	D
Escherichia coli O157:H7 str. Sakai	B	Gammaproteobacteria	5.59	50	1	2	BA000007	NC_002695	03/29/2000	Osaka Univ.	T	P	C	D
Escherichia coli UTI89	B	Gammaproteobacteria	5.18	51.0	1	1	CP000243	NC_007946	04/05/2006	Washington University (WashU)	T	P		D
Escherichia coli W3110	B	Gammaproteobacteria	*4.65		1		AP009048	AC_000091	01/23/2006	Nara Institute of Science and Technology	T	P		D
Francisella tularensis subsp. holarctica	B	Gammaproteobacteria	1.9	32.2	1		AM233362	NC_007880	03/02/2006	Lawrence Livermore National Laboratory	T	P		D

NCBI ENTREZ Genome Project connection information discovery My NCBI [Sign In] [R

All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy

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 EDL933*** **project at Univ. Wisconsin**

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 EDL933*

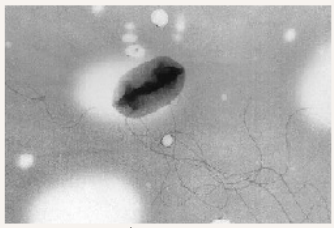


Photo: CDC

- Osaka Univ.
- Sanger
- U. Wisc

Related Resources

- ASAP
- BIGS
- BRS
- CGSC
- Colibri
- CyberCell
- E. coli Community
- E. coli Forum
- E. coli TF
- E. coli Wiki
- ECCE
- ECDC
- ECOR/DEC
- EGED
- EPD
- ERIC

Genome Projects

Genome sequencing:

- *Escherichia coli K12* at Univ. Wisconsin [Complete]
- *Escherichia coli O157:H7 str. Sakai* 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*** (Project ID: **259**) 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
Plasmid pO157	NC_007414	AF074613	1	0.092077	47%	99	1				
Chromosome	NC_002655	AE005174	1	5.52844	50%	5324	128	✔	✔	✔	✔

Publications:

- Perna NT *et al.*, "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7.", *Nature*, 2001 Jan 25;409(6819):529-33

Escherichia coli O157:H7 EDL933

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 EDL933**. This strain (substrain) is considered a reference strain of O157:H7, which was first isolated during an outbreak in 1982.

Cellular features					Environment			Temperature	
Gram stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Opt. temp.	Range
-	Rod	Singles, Pairs		Yes		Facultative	Host-associated	25-35C	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 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 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

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

1: NC_002655. Reports Escherichia coli ...[gi:16445223]

Comment | Features | Sequence

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

```
LOCUS       NC_002655            5528445 bp    DNA     circular BCT 19-JUL-2006
DEFINITION  Escherichia coli O157:H7 EDL933, complete genome.
ACCESSION   NC_002655
VERSION     NC_002655.2  GI:16445223
PROJECT     GenomeProject:259
KEYWORDS    .
SOURCE      Escherichia coli O157:H7 EDL933
  ORGANISM  Escherichia coli O157:H7 EDL933
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 5528445)
  AUTHORS   Perna,N.T., Plunkett,G. III, Burland,V., Mau,B., Glasner,J.D.,
            Rose,D.J., Mayhew,G.F., Evans,P.S., Gregor,J., Kirkpatrick,H.A.,
            Posfai,G., Hackett,J., Klink,S., Boutin,A., Shao,Y., Miller,L.,
```

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
Tools
Check sequence revision history
LinkOut
My NCBI (Cubby)
Related resources
BLAST
Reference sequence project
Entrez Genes
Clusters of orthologous groups

Display Summary Show 20 Sort by Relevance Send to

All: 5324 | Bacteria: 5324 | RefSeq: 5324 | Related Structures: 2927

Items 1 - 20 of 5324

Page 1 of 267 Next

1: NP_285693 Reports Links
thr operon leader peptide [Escherichia coli O157:H7 EDL933]
gi|15799681|ref|NP_285693.1|[15799681]

2: NP_285694 Reports BLink, Conserved Domains, Links
bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 EDL933]
gi|15799682|ref|NP_285694.1|[15799682]

3: NP_285695 Reports BLink, Conserved Domains, Links
homoserine kinase [Escherichia coli O157:H7 EDL933]
gi|15799683|ref|NP_285695.1|[15799683]

4: NP_285696 Reports BLink, Conserved Domains, Links
threonine synthase [Escherichia coli O157:H7 EDL933]
gi|15799684|ref|NP_285696.1|[15799684]

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

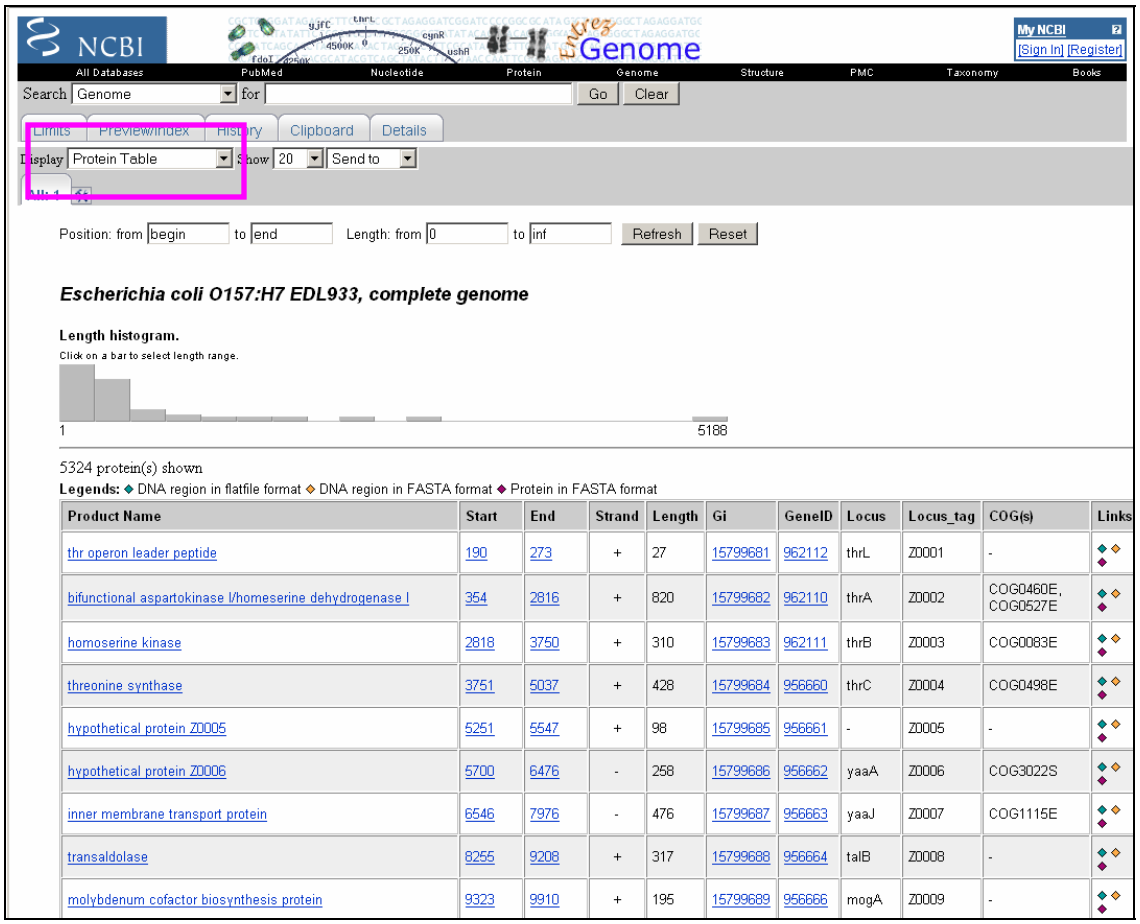
All: 1

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

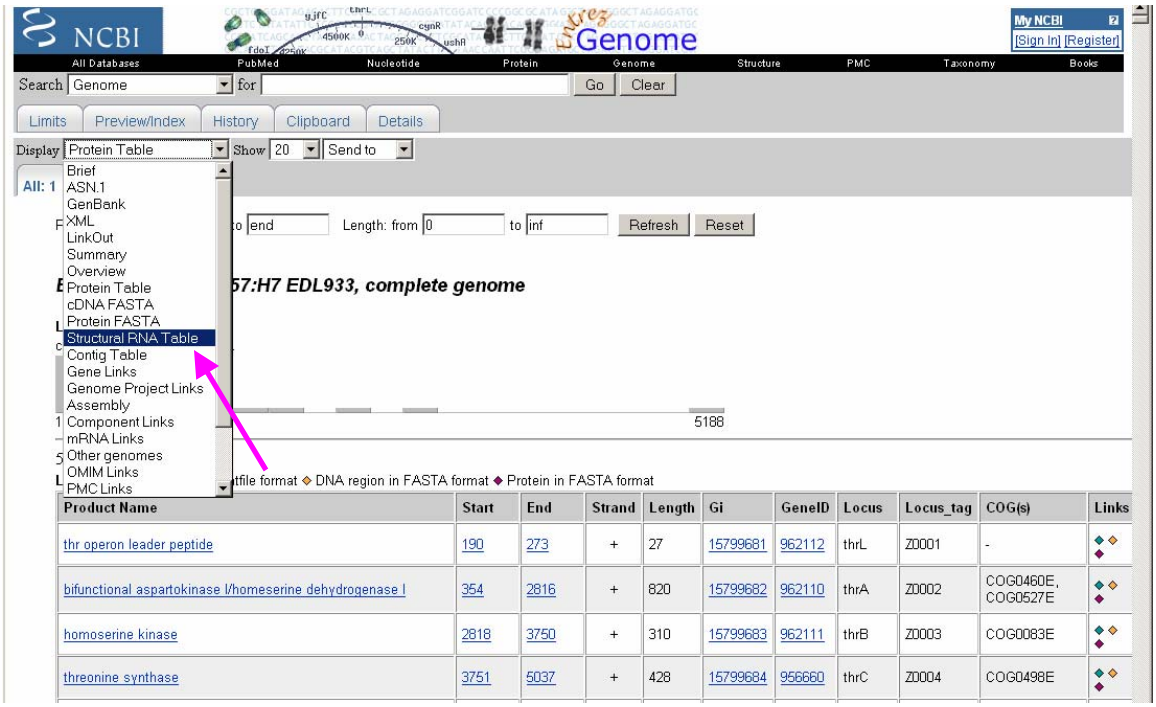
Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli EDL933

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	



Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
thr operon leader peptide	190	273	+	27	15799681	962112	thrL	Z0001	-	
bifunctional aspartokinase I/homeserine dehydrogenase I	354	2816	+	820	15799682	962110	thrA	Z0002	COG0460E, COG0527E	
homoserine kinase	2818	3750	+	310	15799683	962111	thrB	Z0003	COG0083E	
threonine synthase	3751	5037	+	428	15799684	956660	thrC	Z0004	COG0498E	
hypothetical protein Z0005	5251	5547	+	98	15799685	956661	-	Z0005	-	
hypothetical protein Z0006	5700	6476	-	258	15799686	956662	yaaA	Z0006	COG3022S	
inner membrane transport protein	6546	7976	-	476	15799687	956663	yaaJ	Z0007	COG1115E	
transaldolase	8255	9208	+	317	15799688	956664	talB	Z0008	-	
molybdenum cofactor biosynthesis protein	9323	9910	+	195	15799689	956666	mogA	Z0009	-	



Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
thr operon leader peptide	190	273	+	27	15799681	962112	thrL	Z0001	-	
bifunctional aspartokinase I/homeserine dehydrogenase I	354	2816	+	820	15799682	962110	thrA	Z0002	COG0460E, COG0527E	
homoserine kinase	2818	3750	+	310	15799683	962111	thrB	Z0003	COG0083E	
threonine synthase	3751	5037	+	428	15799684	956660	thrC	Z0004	COG0498E	

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 Structural RNA Table Show 20 Send to

All: 1

- 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

to end Length: from 0 to inf Refresh Reset

57:H7 EDL933, complete genome

5188

tfile 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	15799681	962112	thrL	Z0001	-	
bifunctional aspartokinase I/homeserine dehydrogenase I	354	2816	+	820	15799682	962110	thrA	Z0002	COG0460E, COG0527E	
homoserine kinase	2818	3750	+	310	15799683	962111	thrB	Z0003	COG0083E	
threonine synthase	3751	5037	+	428	15799684	956660	thrC	Z0004	COG0498E	
hypothetical protein Z0005	5251	5547	+	98	15799685	956661	-	Z0005	-	
hypothetical protein Z0006	5700	6476	-	258	15799686	956662	yaaA	Z0006	COG3022S	
inner membrane transport protein	6546	7976	-	476	15799687	956663	yaaJ	Z0007	COG1115E	
transaldolase	8255	9208	+	317	15799688	956664	talB	Z0008	-	

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

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

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

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt 9,910 nt

thrB thrC talB mogA thrA yaaA yaaJ

5528445 nt

Click here for Sequence Viewer presentation (base sequence and aligned amino acids) of selected region

Comment and suggestions to: [info@ncbi.nlm.nih.gov]

Display Overview Show 20 Send to

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

Links

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

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt — 9,910 nt

thrB — thrC — threonine synthase [Escherichia coli O157:H7 EDL93... — Gene: thrC — talB — mogA — thrA — yaaA — yaaJ

5528445 nt

Click here for Sequence Viewer presentation (base sequence and aligned amino 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: thrC threonine synthase [*Escherichia coli O157:H7 EDL933* **]**

GeneID: 956660 — updated 07-Jul-2006

Summary

Gene name	thrC
Locus tag	Z0004
Gene type	protein coding
RefSeq status	Validated
Organism	*Escherichia coli O157:H7 EDL933 (strain: EDL933, serotype: O157:H7)*
Lineage	*Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia*

Genomic regions, transcripts, and products

Go to reference sequence details

NC_002655.2

[3751 ▶ — [5037 ▶ — 5' — 3' — NP_285696 threonine synthase — = coding region

Genomic context

[354 ▶ — [6476 ▶ — thrA — thrB — thrC — Z0005 — yaaR

Entrez Gene Home

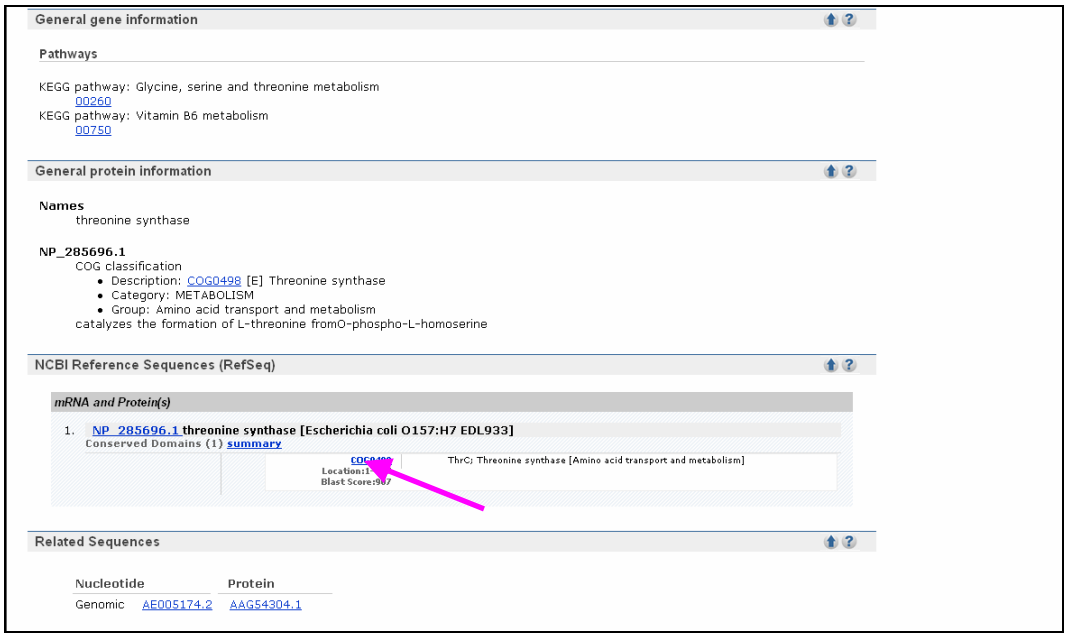
Table Of Contents
- Summary
- Genomic regions, transcripts...
- Genomic context
- Bibliography
- General gene information
- General protein information
- Reference Sequences
- Related Sequences

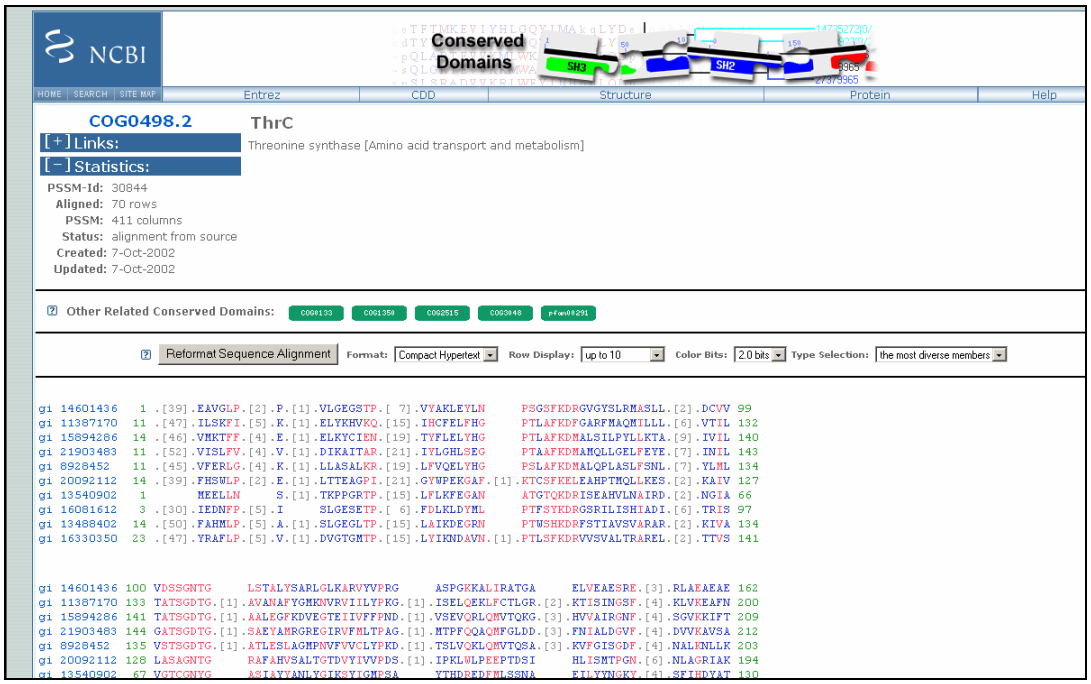
Links — Explain
- Conserved Domains
- Genome
- Nucleotide
- Protein
- PubMed
- Taxonomy
- KEGG
- LinkOut

Entrez Gene Info

Feedback

Subscriptions


General gene information
Pathways
KEGG pathway: Glycine, serine and threonine metabolism
00260
KEGG pathway: Vitamin B6 metabolism
00750
General protein information
Names
threonine synthase
NP_285696.1
COG classification
Description: COG0498 [E] Threonine synthase
Category: METABOLISM
Group: Amino acid transport and metabolism
catalyzes the formation of L-threonine fromO-phospho-L-homoserine
NCBI Reference Sequences (RefSeq)
mRNA and Protein(s)
1. NP_285696.1 threonine synthase [Escherichia coli O157:H7 EDL933]
Conserved Domains (1) summary
ThrC; Threonine synthase [Amino acid transport and metabolism]
Related Sequences
Nucleotide
Protein
Genomic
AE005174.2
AAG54304.1


NCBI
Conserved Domains
HOME
SEARCH
SITE MAP
Entrez
CDD
Structure
Protein
Help
COG0498.2
ThrC
[+] Links:
[−] Statistics:
Threonine synthase [Amino acid transport and metabolism]
PSSM-Id: 30844
Aligned: 70 rows
PSSM: 411 columns
Status: alignment from source
Created: 7-Oct-2002
Updated: 7-Oct-2002
Other Related Conserved Domains:
Reformat Sequence Alignment
Format: Compact Hypertext
Row Display: up to 10
Color Bits: 2.0 bits
Type Selection: the most diverse members

All: 1

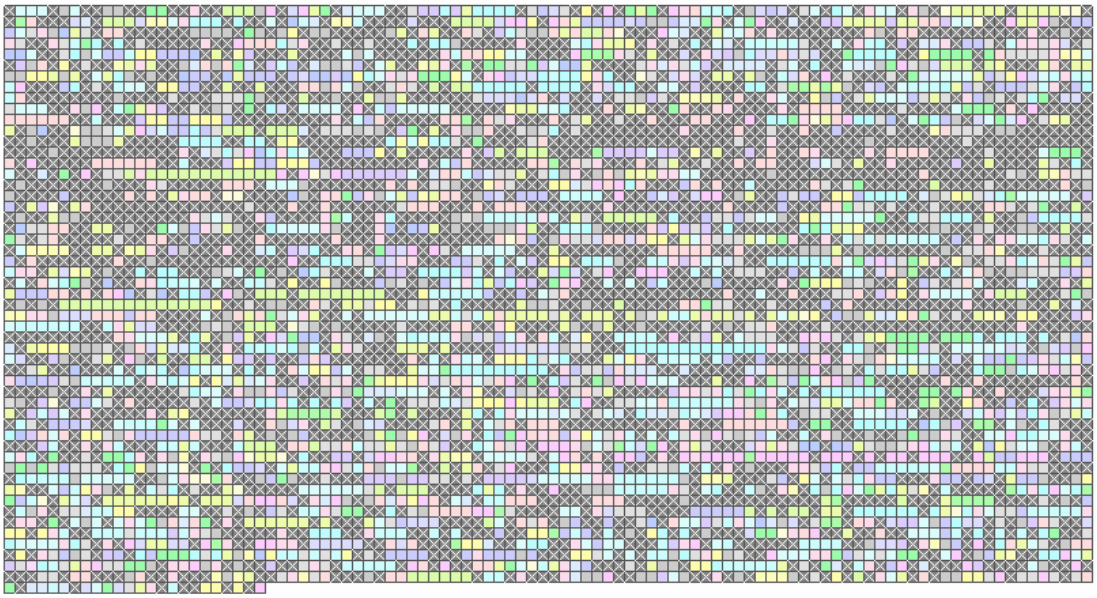
Genome > **Bacteria** > ***Escherichia coli O157:H7 EDL933, complete genome***

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

5324 proteins: distribution by COGs functional categories
3362 proteins can be found in COGs data base



Code		COGs	Description	(% in sequence)	(% in genome)	(% in genus)	(% in Gammaproteobacteria)	(% in Bacteria)
☐	J	150	Translation	2.5681	2.5424	2.7363	3.3678	3.6565
☐	A	1	RNA processing and modification	0.0171	0.0169	0.0183	0.0242	0.0114
☐	K	287	Transcription	4.9135	4.8814	5.1309	4.8981	4.9477
☐	L	293	Replication, recombination and repair	5.0163	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	0.4965	0.5085	0.5369	0.6737	0.6910
☐	Y	-	Nuclear structure	-	-	-	-	-
☐	V	54	Defense mechanisms	0.9245	0.9068	0.9365	0.9882	1.1522
☐	T	171	Signal transduction mechanisms	2.9276	2.8729	2.9681	3.6227	3.4483
☐	M	246	Cell wall/membrane biogenesis	4.2116	4.2203	4.3072	4.3054	4.1094
☐	N	122	Cell motility	2.0887	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	0.0342	0.0339	0.0214	0.0108	0.0071
☐	U	153	Intracellular trafficking and secretion	2.6194	2.7034	2.5166	2.2961	1.7173
☐	O	137	Posttranslational modification, protein turnover, chaperones	2.3455	2.3305	2.4251	2.8264	2.6152
☐	C	277	Energy production and conversion	4.7423	4.6695	5.0485	4.3095	4.3745
☐	G	268	Carbohydrate transport and metabolism	4.5883	4.4831	5.0577	3.5918	3.4880
☐	E	326	Amino acid transport and metabolism	5.5812	5.5169	6.0185	5.9370	5.9060
☐	F	78	Nucleotide transport and metabolism	1.3354	1.3136	1.4734	1.4513	1.6335
☐	H	135	Coenzyme transport and metabolism	2.3112	2.2881	2.4495	2.6556	2.6540
☐	I	107	Lipid transport and metabolism	1.8319	1.8051	1.9065	2.3512	2.6600

Intracellular trafficking and secretion genes:

Location	strand	Length	PID	Locus	LocusTag	COG	Product
19299..21749	-	816	15799701	-	Z0022	COG3188NU	putative usher protein
21762..22445	-	227	15799702	-	Z0023	COG3121NU	putative chaperone protein
22495..23028	-	177	15799703	-	Z0024	COG3539NU	putative type-1 fimbrial protein
29619..30113	+	164	15799709	lspA	Z0031	COG0597MU	signal peptidase II
112885..115590	+	901	15799782	secA	Z0108	COG0653U	translocase
119012..120214	-	400	15799790	hofC	Z0116	COG1459NU	putative integral membrane protein involved in biogenesis of fimbriae, protein transport, DNA uptake
120204..121589	-	461	15799791	hofB	Z0117	COG2804NU	putative integral membrane protein involved in biogenesis of fimbriae, protein transport, DNA uptake
121599..122039	-	146	15799792	ppdD	Z0118	COG4969NU	prelipin peptidase dependent protein
156680..157240	-	186	15799822	yadM	Z0149	COG3539NU	putative fimbrial protein

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	ODD	Organism Group
Molecule: **DNA**			Other genomes for species	

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

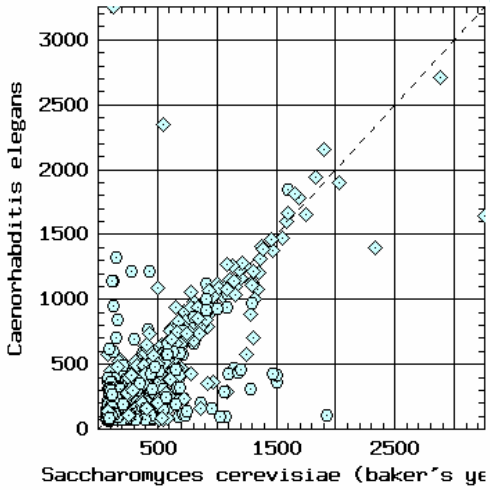
4932 Saccharomyces cerevisiae (baker's yeast)

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

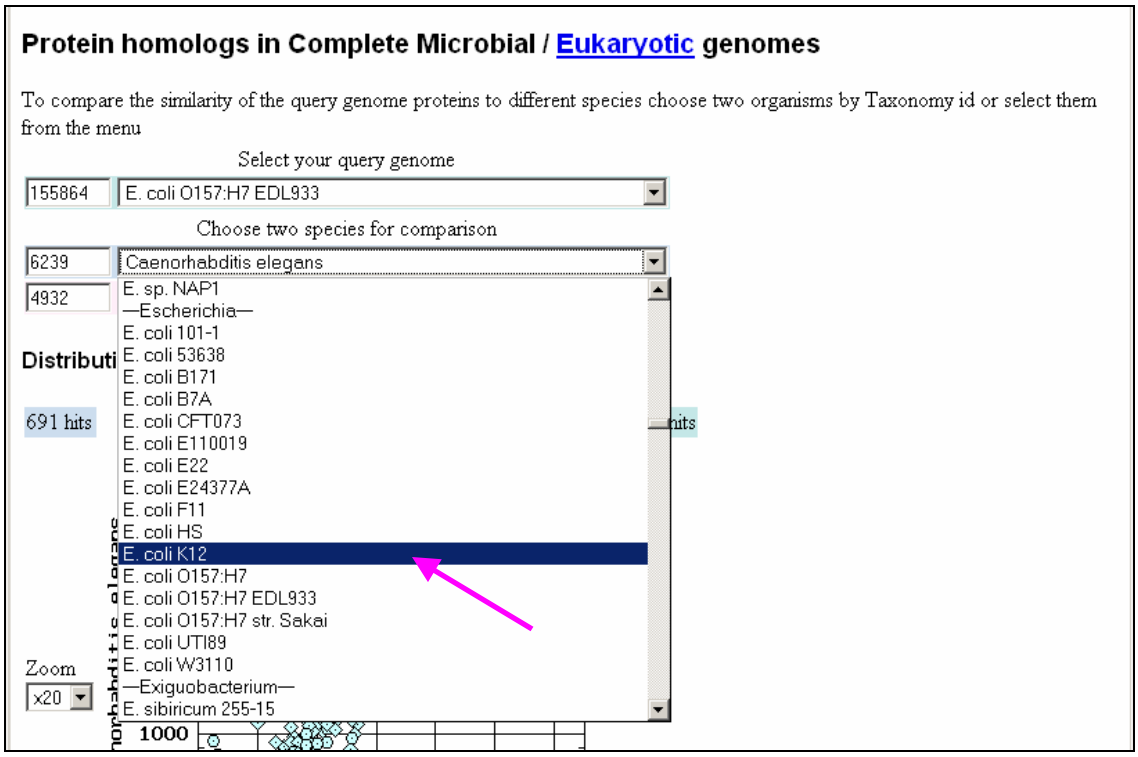
691 hits

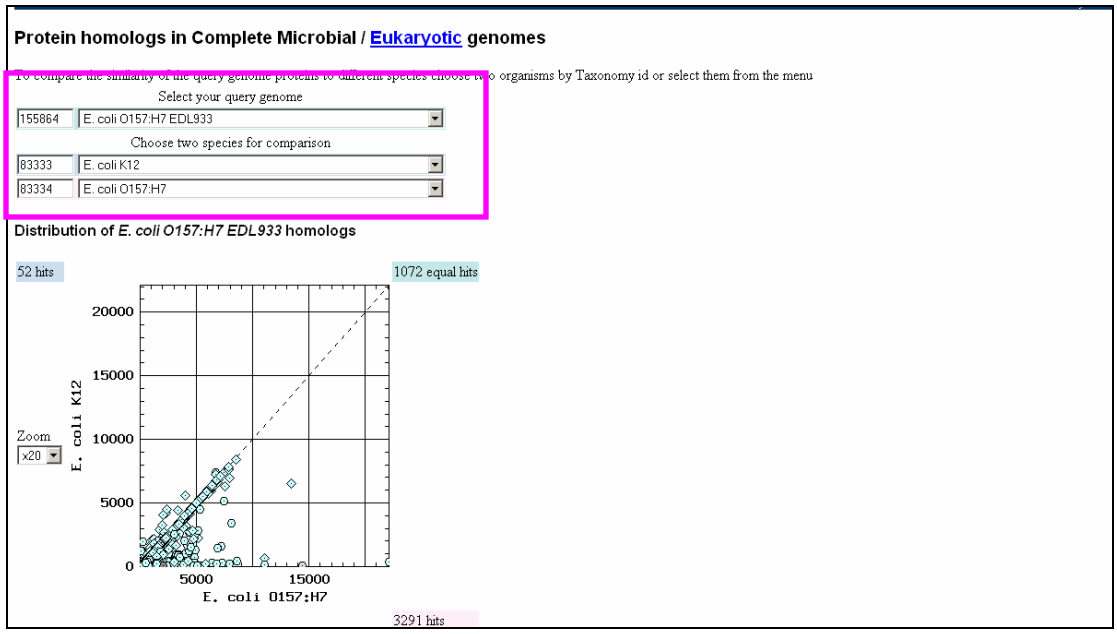
22 equal hits

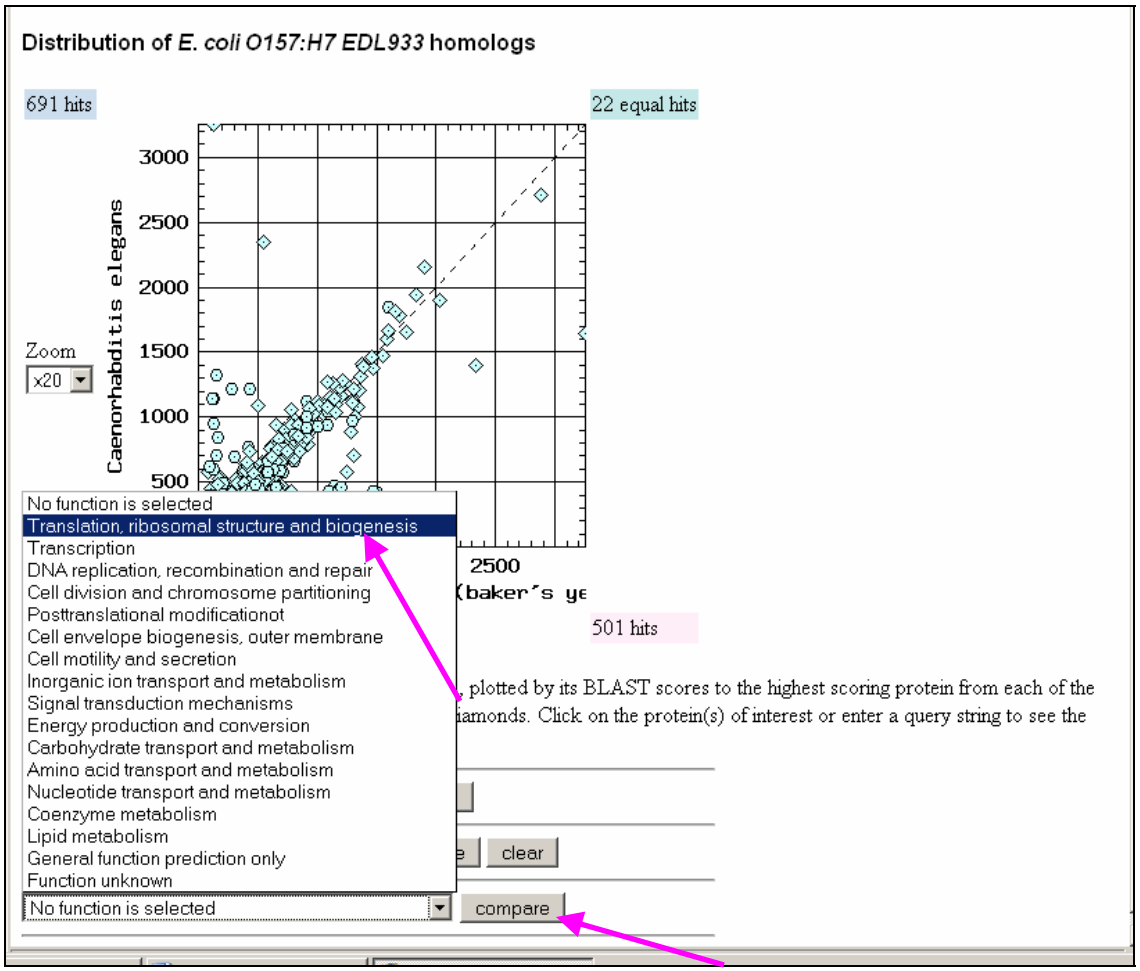
Zoom x20



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
155864
E. coli O157:H7 EDL933
Choose two species for comparison
6239
Caenorhabditis elegans
4932
E. sp. NAP1
—Escherichia—
E. coli 101-1
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
E. coli W3110
—Exiguobacterium—
E. sibiricum 255-15
Distributi
691 hits
Zoom
x20
1000


Protein homologs in Complete Microbial / Eukaryotic genomes
Select your query genome
155864
E. coli O157:H7 EDL933
Choose two species for comparison
83333
E. coli K12
83334
E. coli O157:H7
Distribution of E. coli O157:H7 EDL933 homologs
52 hits
1072 equal hits
Zoom
x20
E. coli K12
20000
15000
10000
5000
0
5000
15000
E. coli O157:H7
3291 hits


Distribution of E. coli O157:H7 EDL933 homologs
691 hits
22 equal hits
Caenorhabditis elegans
3000
2500
2000
1500
1000
500
Zoom
x20
2500
(baker's ye
501 hits
No function is selected
Translation, ribosomal structure and biogenesis
Transcription
DNA replication, recombination and repair
Cell division and chromosome partitioning
Posttranslational modificationot
Cell envelope biogenesis, outer membrane
Cell motility and secretion
Inorganic ion transport and metabolism
Signal transduction mechanisms
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme metabolism
Lipid metabolism
General function prediction only
Function unknown
No function is selected
plotted by its BLAST scores to the highest scoring protein from each of the
iamonds. Click on the protein(s) of interest or enter a query string to see the
clear
compare

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

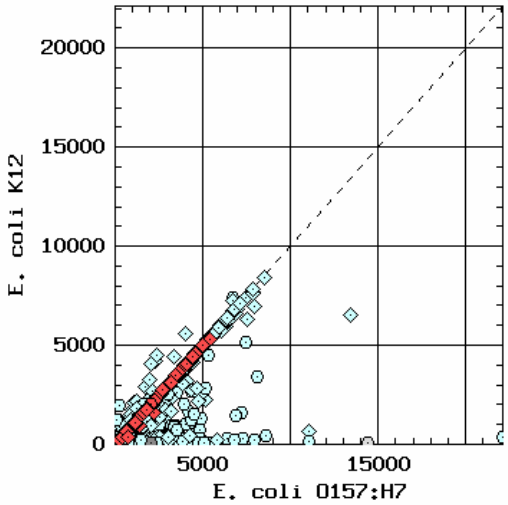
83333	E. coli K12
83334	E. coli O157:H7

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

52 hits

1072 equal hits

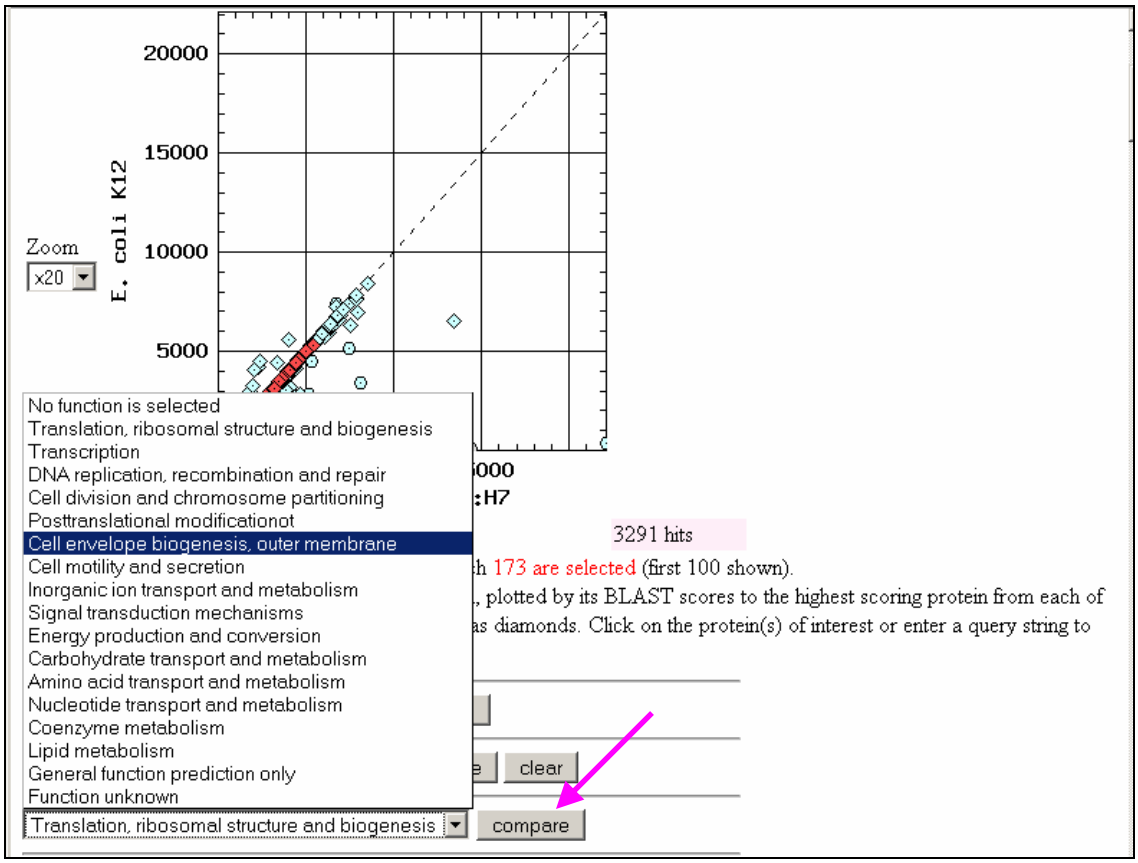
Zoom x20



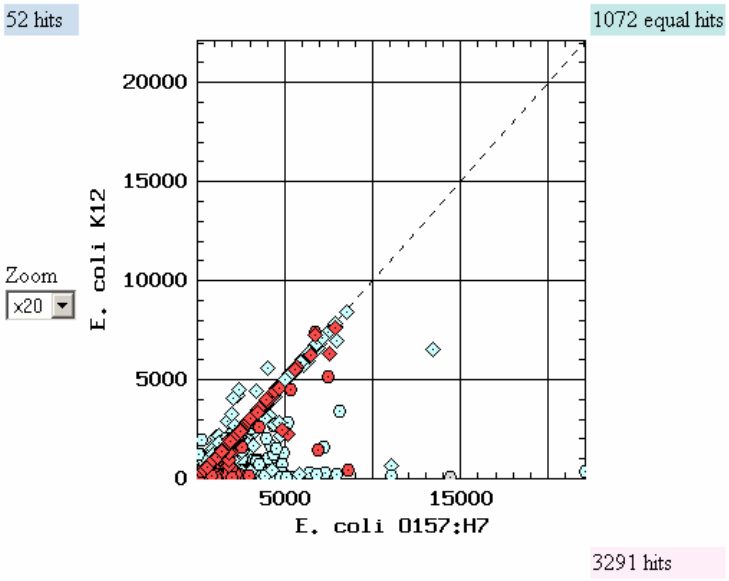
3291 hits

5423 query proteins produced **4415 hits**, from which 173 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.

Blast2Seq	Accession	Blink	Definition
	NP_285717	=>	**30S ribosomal protein S20 [Escherichia coli O157:H7 EDL933]**
425	NP_414564	=>	30S ribosomal protein S20 [Escherichia coli K12]
425	NP_308053	=>	30S ribosomal protein S20 [Escherichia coli O157:H7 str. Sakai]
	NP_285720	=>	**isoleucyl-tRNA synthetase [Escherichia coli O157:H7 EDL933]**
4948	NP_414567	=>	isoleucyl-tRNA synthetase [Escherichia coli K12]
4973	NP_308056	=>	isoleucyl-tRNA synthetase [Escherichia coli O157:H7 str. Sakai]
	NP_285748	=>	**dimethyladenosine transferase [Escherichia coli O157:H7 EDL933]**
1401	NP_414593	=>	dimethyladenosine transferase [Escherichia coli K12]
1407	NP_308083	=>	dimethyladenosine transferase [Escherichia coli O157:H7 str. Sakai]


20000
15000
10000
5000
E. coli K12
Zoom
x20
No function is selected
Translation, ribosomal structure and biogenesis
Transcription
DNA replication, recombination and repair
Cell division and chromosome partitioning
Posttranslational modificationot
Cell envelope biogenesis, outer membrane
Cell motility and secretion
Inorganic ion transport and metabolism
Signal transduction mechanisms
Energy production and conversion
Carbohydrate transport and metabolism
Amino acid transport and metabolism
Nucleotide transport and metabolism
Coenzyme metabolism
Lipid metabolism
General function prediction only
Function unknown
Translation, ribosomal structure and biogenesis
compare
clear
3291 hits
173 are selected (first 100 shown).
, plotted by its BLAST scores to the highest scoring protein from each of
as diamonds. Click on the protein(s) of interest or enter a query string to

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



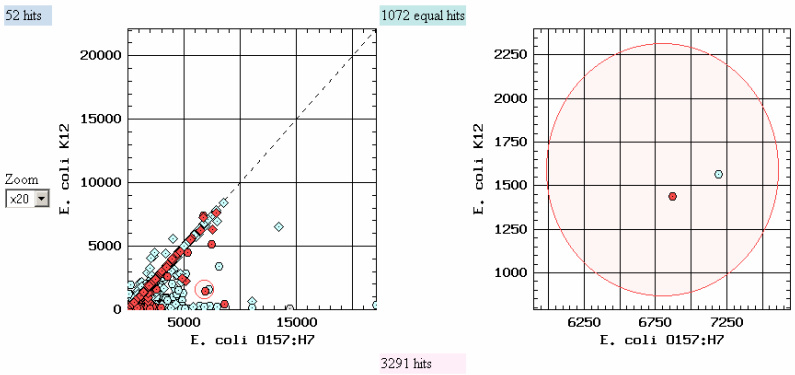
5423 query proteins produced **4415 hits**, from which 261 are selected (first 100 shown).
Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Symmetrical hits are shown as diamonds. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Cutoff: 10 Linear scale compare

Query: compare clear

Cell envelope biogenesis, outer membrane compare

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



5423 query proteins produced **4415 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

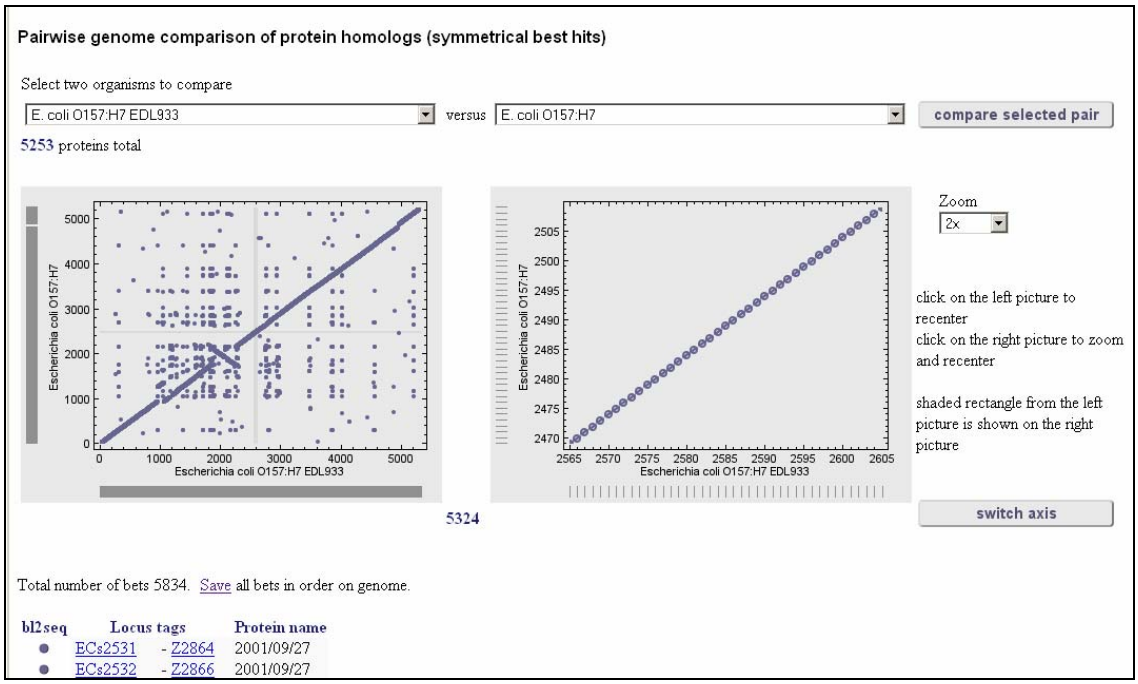
Blast2Seq	Accession	Blink	Definition
	NP_286049	=>	**putative beta-barrel outer membrane protein [Escherichia coli O157:H7 EDL933]**
1440	NP_416026	=>	conserved protein, predicted pseudogene [Escherichia coli K12]
6869	NP_308389	=>	AidA-I adhesin-like protein [Escherichia coli O157:H7 str. Sakai]

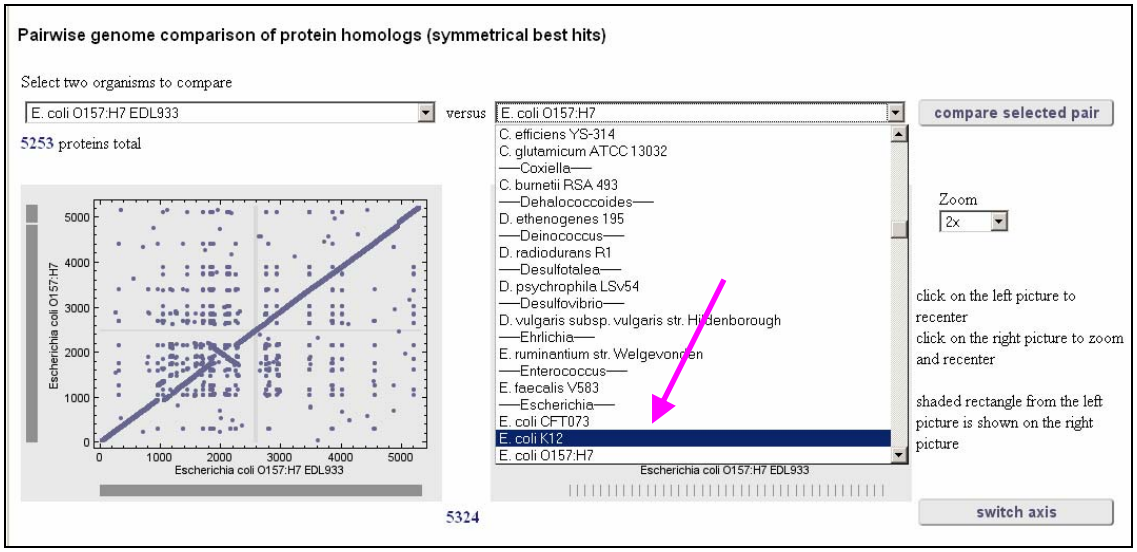
Genome > Bacteria > *Escherichia coli O157:H7 EDL933, complete genome*

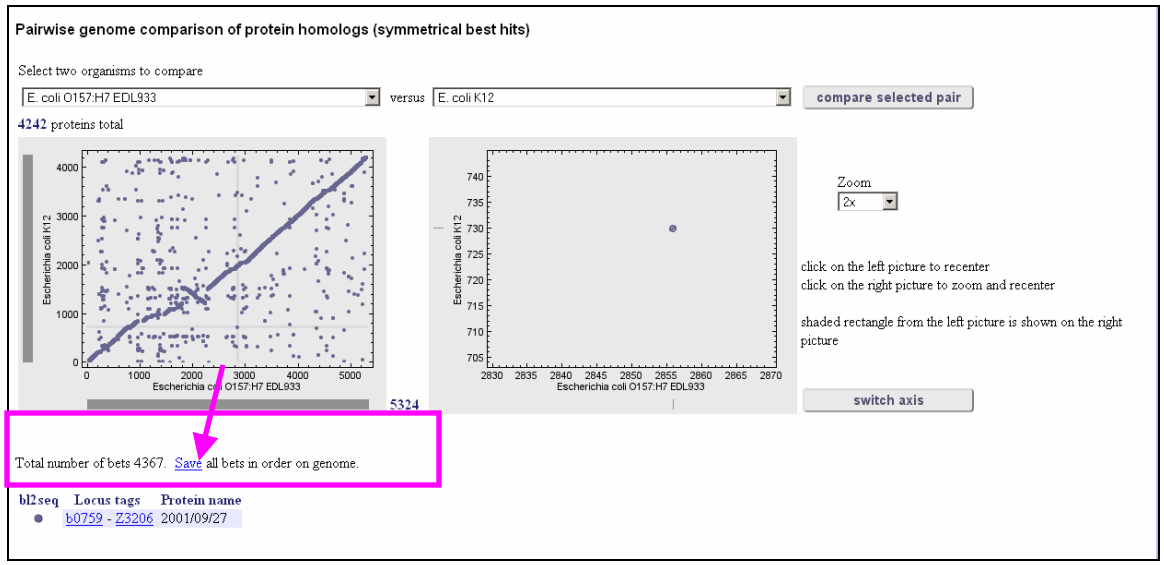
Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia EDL933

Chromosomes: *chromosome*
Plasmids: pO157

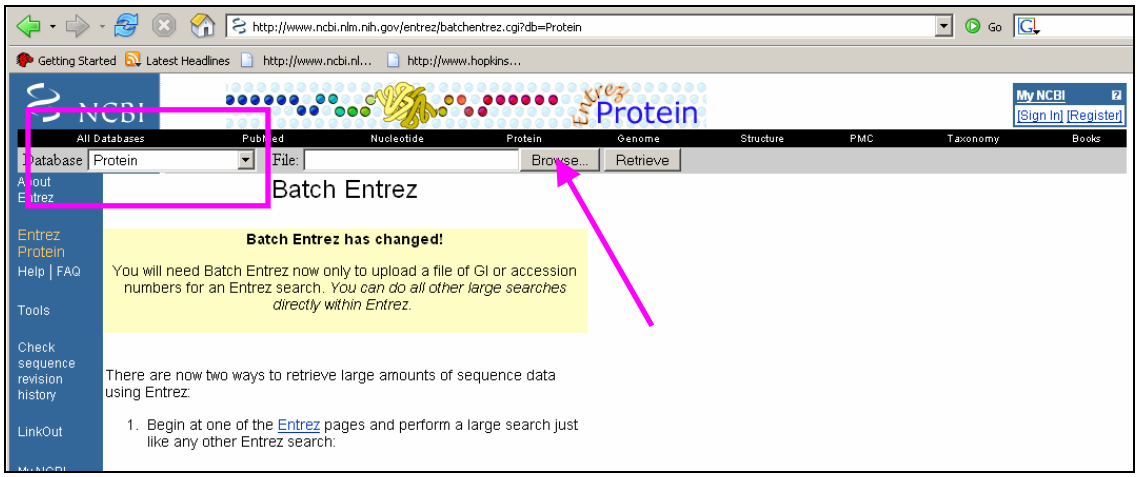
Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_002655	Genes: 5453	COG	Genome Project	Publications: [1]
GenBank: AE005174	Protein coding: 5324	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,528,445 nt**	Structural RNAs: 128	TaxMap	GenBank FTP	Seq.Status: **Completed**
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	

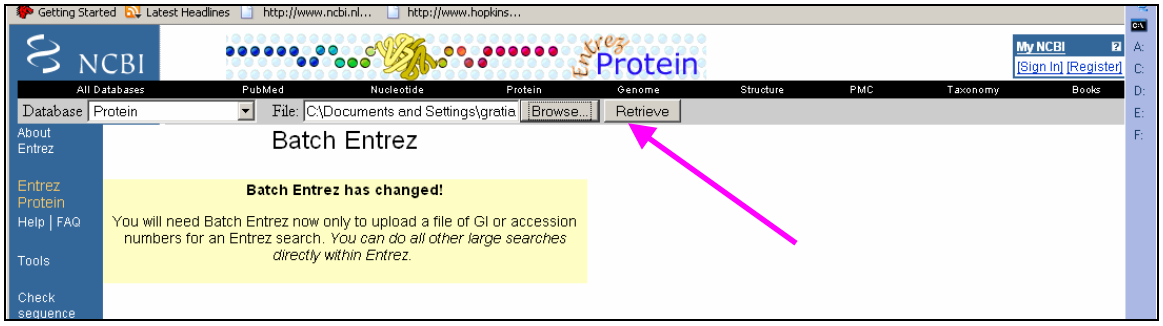

Pairwise genome comparison of protein homologs (symmetrical best hits)
Select two organisms to compare
E. coli O157:H7 EDL933
versus
E. coli O157:H7
compare selected pair
5253 proteins total
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
Zoom
2x
click on the left picture to recenter
click on the right picture to zoom and recenter
shaded rectangle from the left picture is shown on the right picture
5324
switch axis
Total number of bets 5834. Save all bets in order on genome.
bl2seq
Locus tags
Protein name
ECs2531 - Z2864 2001/09/27
ECs2532 - Z2866 2001/09/27

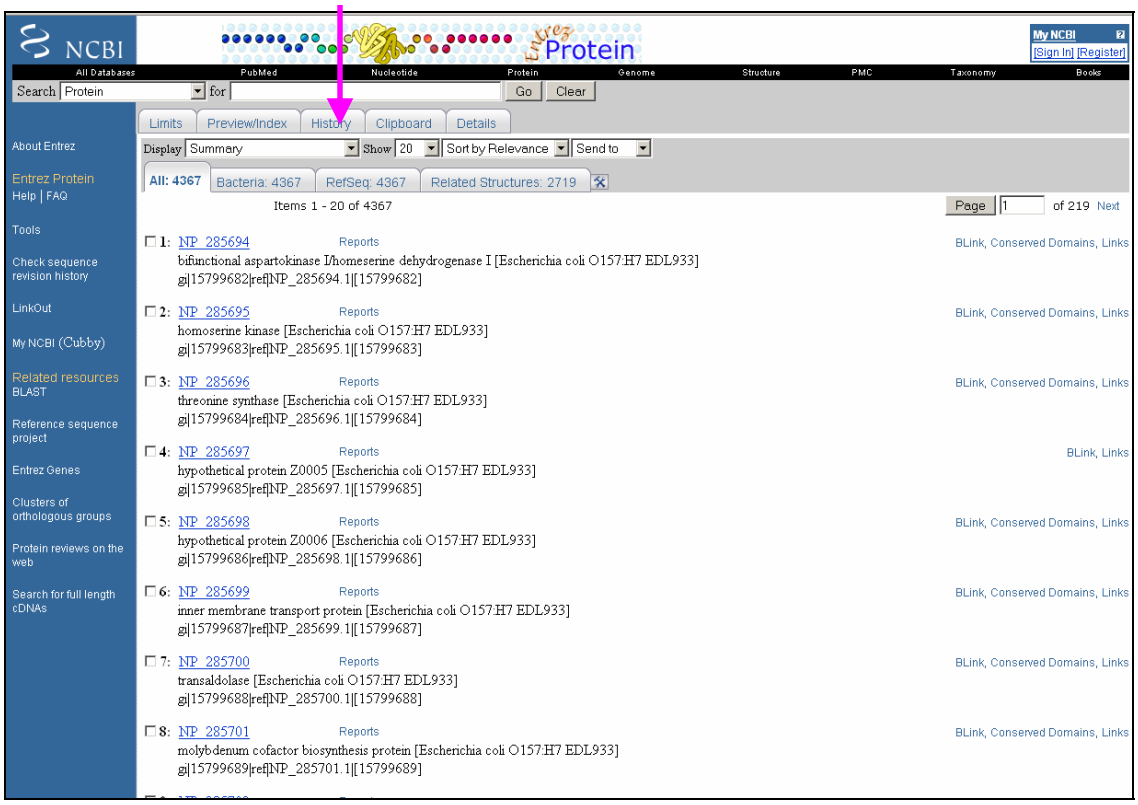

Pairwise genome comparison of protein homologs (symmetrical best hits)
Select two organisms to compare
E. coli O157:H7 EDL933
versus
E. coli O157:H7
compare selected pair
5253 proteins total
C. efficiens YS-314
C. glutamicum ATCC 13032
—Coxiella—
C. burnetii RSA 493
—Dehalococcoides—
D. ethenogenes 195
—Deinococcus—
D. radiodurans R1
—Desulfotalea—
D. psychrophila LSv54
—Desulfovibrio—
D. vulgaris subsp. vulgaris str. Hildenborough
—Ehrlichia—
E. ruminantium str. Welgevonden
—Enterococcus—
E. faecalis V583
—Escherichia—
E. coli CFT073
E. coli K12
E. coli O157:H7
Escherichia coli O157:H7
Escherichia coli O157:H7 EDL933
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
5324
switch axis



b0002	gi|16127996	Z0002	gi|15799682		9/27/2001
b0003	gi|16127997	Z0003	gi|15799683		9/27/2001
b0004	gi|16127998	Z0004	gi|15799684		9/27/2001
b0005	gi|16127999	Z0005	gi|15799685		9/27/2001
b0006	gi|16128000	Z0006	gi|15799686		9/27/2001
b0007	gi|16128001	Z0007	gi|15799687		9/27/2001
b0008	gi|16128002	Z0008	gi|15799688		9/27/2001
b0009	gi|16128003	Z0009	gi|15799689		9/27/2001
b0010	gi|16128004	Z0010	gi|15799690		9/27/2001
b0011	gi|16128005	Z0011	gi|15799691		9/27/2001
b0012	gi|16128006	Z0012	gi|15799692		9/27/2001
b0013	gi|16128007	Z0013	gi|15799693		9/27/2001
b0014	gi|16128008	Z0014	gi|15799694		9/27/2001
b0015	gi|16128009	Z0015	gi|15799695		9/27/2001
b0018	gi|16128012	Z0016	gi|15799696		9/27/2001
b0019	gi|16128013	Z0018	gi|15799697		9/27/2001
b0020	gi|16128014	Z0019	gi|15799698		9/27/2001
b2109	gi|16130047	Z0022	gi|15799701		9/27/2001
b2110	gi|16130048	Z0023	gi|15799702		9/27/2001
b2111	gi|16130049	Z0024	gi|15799703		9/27/2001
b0023	gi|16128017	Z0027	gi|15799705		9/27/2001
b0024	gi|16128018	Z0028	gi|15799706		9/27/2001
b0025	gi|16128019	Z0029	gi|15799707		9/27/2001
b0026	gi|16128020	Z0030	gi|15799708		9/27/2001
b0027	gi|16128021	Z0031	gi|15799709		9/27/2001
b0028	gi|16128022	Z0033	gi|15799710		9/27/2001
b0029	gi|16128023	Z0034	gi|15799711		9/27/2001
b0030	gi|16128024	Z0035	gi|15799712		9/27/2001
b0031	gi|16128025	Z0036	gi|15799713		9/27/2001
b0032	gi|16128026	Z0037	gi|15799714		9/27/2001
b0033	gi|16128027	Z0038	gi|15799715		9/27/2001
b0034	gi|49175992	Z0040	gi|15799717		9/27/2001
b0035	gi|16128029	Z0041	gi|15799718		9/27/2001
b0036	gi|16128030	Z0042	gi|15799719		9/27/2001
b0037	gi|49175993	Z0043	gi|15799720		9/27/2001
b0038	gi|16128032	Z0044	gi|15799721		9/27/2001
b0039	gi|16128033	Z0045	gi|15799722		9/27/2001
b0040	gi|16128034	Z0046	gi|15799723		9/27/2001
b0041	gi|16128035	Z0047	gi|15799724		9/27/2001
b0042	gi|16128036	Z0048	gi|15799725		9/27/2001
b0043	gi|16128037	Z0049	gi|15799726		9/27/2001
b0044	gi|16128038	Z0050	gi|15799727		9/27/2001
b0045	gi|16128039	Z0051	gi|15799728		9/27/2001
b0046	gi|16128040	Z0052	gi|15799729		9/27/2001
b0047	gi|16128041	Z0053	gi|15799730		9/27/2001
b0048	gi|16128042	Z0055	gi|15799731		9/27/2001
b0049	gi|16128043	Z0058	gi|15799734		9/27/2001
b0050	gi|16128044	Z0059	gi|15799735		9/27/2001
b0051	gi|16128045	Z0060	gi|15799736		9/27/2001
b0052	gi|16128046	Z0061	gi|15799737		9/27/2001
b0053	gi|16128047	Z0062	gi|15799738		9/27/2001
b0054	gi|16128048	Z0063	gi|15799739		9/27/2001


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
Batch Entrez
Entrez Protein
Help | FAQ
Tools
Check sequence revision history
LinkOut
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:
1. Begin at one of the Entrez pages and perform a large search just like any other Entrez search:


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: C:\Documents and Settings\gratia
Browse...
Retrieve
Batch Entrez
Entrez Protein
Help | FAQ
Tools
Check sequence
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: 4367
Bacteria: 4367
RefSeq: 4367
Related Structures: 2719
Items 1 - 20 of 4367
Page 1 of 219 Next
About Entrez
Entrez Protein
Help | FAQ
Tools
Check sequence revision history
LinkOut
My NCBI (Cubby)
Related resources
BLAST
Reference sequence project
Entrez Genes
Clusters of orthologous groups
Protein reviews on the web
Search for full length cDNAs
1: NP_285694 Reports BLink, Conserved Domains, Links
bifunctional aspartokinase I/homeserine dehydrogenase I [Escherichia coli O157:H7 EDL933]
gi|15799682|ref|NP_285694.1|[15799682]
2: NP_285695 Reports BLink, Conserved Domains, Links
homoserine kinase [Escherichia coli O157:H7 EDL933]
gi|15799683|ref|NP_285695.1|[15799683]
3: NP_285696 Reports BLink, Conserved Domains, Links
threonine synthase [Escherichia coli O157:H7 EDL933]
gi|15799684|ref|NP_285696.1|[15799684]
4: NP_285697 Reports BLink, Links
hypothetical protein Z0005 [Escherichia coli O157:H7 EDL933]
gi|15799685|ref|NP_285697.1|[15799685]
5: NP_285698 Reports BLink, Conserved Domains, Links
hypothetical protein Z0006 [Escherichia coli O157:H7 EDL933]
gi|15799686|ref|NP_285698.1|[15799686]
6: NP_285699 Reports BLink, Conserved Domains, Links
inner membrane transport protein [Escherichia coli O157:H7 EDL933]
gi|15799687|ref|NP_285699.1|[15799687]
7: NP_285700 Reports BLink, Conserved Domains, Links
transaldolase [Escherichia coli O157:H7 EDL933]
gi|15799688|ref|NP_285700.1|[15799688]
8: NP_285701 Reports BLink, Conserved Domains, Links
molybdenum cofactor biosynthesis protein [Escherichia coli O157:H7 EDL933]
gi|15799689|ref|NP_285701.1|[15799689]

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

Lineage: Bacteria ; Proteobacteria ; Gammaproteobacteria ; Enterobacteriales ; Enterobacteriaceae ; Escherichia ; Escherichia coli ; Escherichia coli O157:H7 EDL933

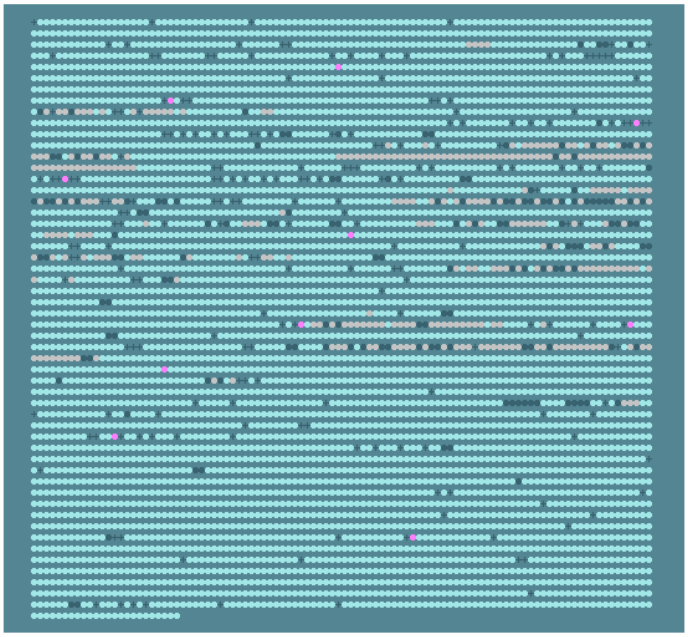
Chromosomes: *chromosome*
Plasmids: pO157

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	

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy

- *[Escherichia (5135)]* - *[Enterobacteriaceae (4763)]* - *[Enterobacteriales (4763)]* - *[Gammaproteobacteria (4643)]* - *[Proteobacteria (4465)]* - *[Bacteria (3915)]* - *[cellular organisms (1858)]*

5324 *Escherichia coli O157:H7 EDL933* 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

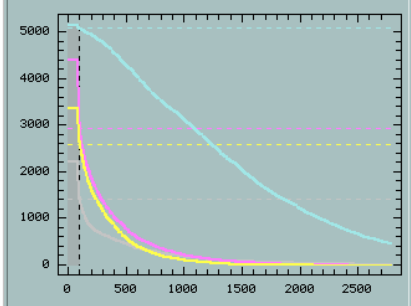
Click on the graph at left, to change the blast threshold (cut-off) used to find related sequences

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

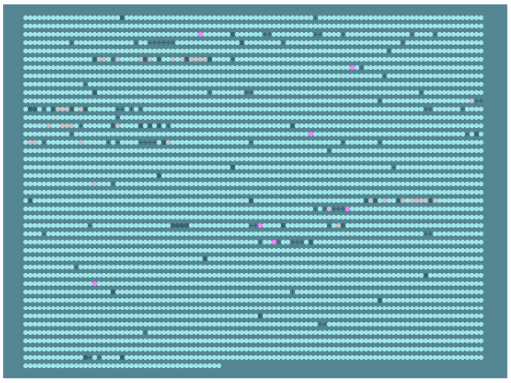


Name		Above cut-off	Best
Virus	V	1390	296
Eukaryota	E	2892	11
Eubacteria	B	5088	4664
Archaea	A	2574	0
Eubacteria + Archaea	BA	5088	4664
Archaea + Eukaryota	AE	3324	11
Eukaryota + Eubacteria	EB	5094	4675
Eukaryota + Eubacteria + Archaea	EBA	5094	4675

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy

[Enterobacteriaceae (3964)] - [Enterobacteriales (3964)] - [Gammaproteobacteria (3855)] - [Proteobacteria (3707)] - [Bacteria (3237)] - [cellular organisms (1250)]

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

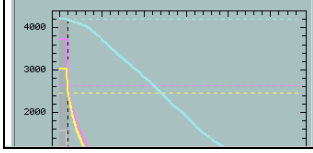


Click on a region in the map, to see blast results for that region

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

Distribution of database hits

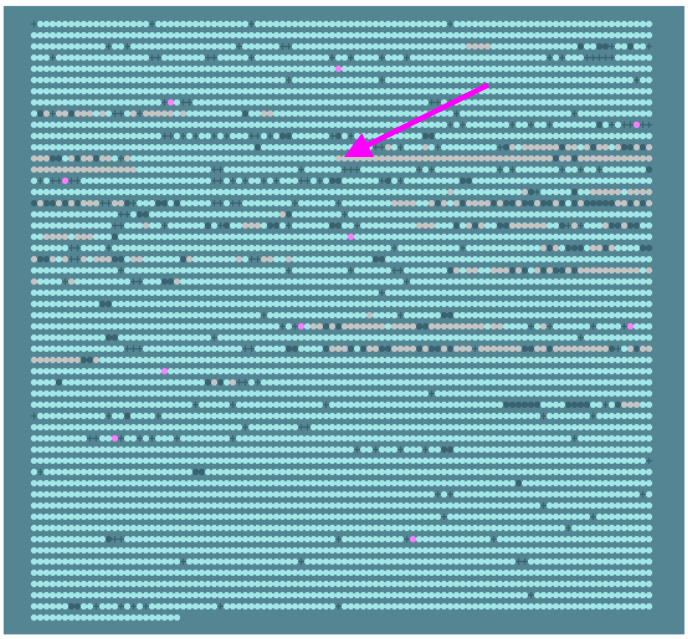


Name		Above cut-off	Best
Virus	V	837	36
Eukaryota	E	2614	8
Eubacteria	B	4189	4093
Archaea	A	2444	0
Eubacteria + Archaea	BA	4189	4093
Archaea +	AE	3030	8

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy

- [Escherichia (5135)] - [Enterobacteriaceae (4763)] - [Enterobacteriales (4763)] - [Gammaproteobacteria (4643)] - [Proteobacteria (4465)] - [Bacteria (3915)] - [cellular organisms (1858)]

5324 *Escherichia coli O157:H7 EDL933* proteins: taxonomic distribution of the homologs

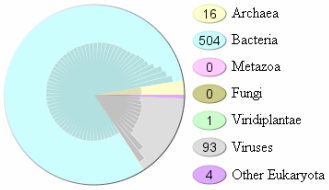


Click on a region in the map, to see blast results for that region

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

	Virus		Eukaryota		Eubacteria		Archaea			
100	73 (66) V		32 (0) E		83 (29) B		21 (0) A			
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
15800923 B	-		108880530	(121)	74311559	(2138)	-		3	periplasmic glucose-1-phosphatase
15800924 B	-		-		24051292	(393)	-		-	hypothetical protein Z1422
15800925 B	-		53749369	(469)	82776291	(916)	72396581	(441)	-	TrpR binding protein WrbA
15800926 V	32128262	(2311)	-		16764254	(1504)	14324479	(111)	-	integrase for bacteriophage BP-933W
15800927 V	32128263	(528)	-		16764255	(261)	-		-	putative excisionase for bacteriophage BP-933W
15800928 V	32128264	(549)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800929 V	32128266	(638)	-		56128804	(429)	-		-	unknown protein encoded by bacteriophage BP-933W
15800930 V	32128267	(1101)	-		24050523	(173)	-		-	unknown protein encoded by bacteriophage BP-933W
15800931 V	32128269	(504)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800932 V	32128271	(500)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800933 V	32128274	(1327)	108705972	(111)	16760392	(311)	-		-	unknown protein encoded by bacteriophage BP-933W
15800934 V	32128278	(489)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800935 V	20065894	(313)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800936 V	20065897	(1203)	54637203	(120)	82776669	(1174)	-		1	putative exonuclease of bacteriophage BP-933W
15800937 V	20065900	(1370)	-		16759879	(1307)	-		-	putative Bet recombination protein of bacteriophage BP-933W
15800938 V	31044238	(514)	-		82776671	(447)	-		-	putative host-nuclease inhibitor protein Gam of bacteriphage BP-933W
15800939 V	31044239	(447)	-		-		-		-	putative Kil protein of bacteriphage BP-933W
15800940 V	31044240	(646)	-		67158861	(144)	-		-	putative single-stranded DNA binding protein
15800941 V	31044241	(849)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800942 V	20065909	(645)	-		-		-		-	putative antitermination protein N of bacteriophage BP-933W
15800943 V	20065910	(271)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800944 V	20065911	(1861)	37575481	(257)	91779963	(649)	48429855	(215)	68	putative serine/threonine kinase encoded by bacteriophage BP-933W
15800945 V	20065912	(759)	-		91779964	(281)	-		-	unknown protein encoded by bacteriophage BP-933W
15800946 V	20065914	(615)	-		37198778	(267)	-		-	unknown protein encoded by bacteriophage BP-933W
15800947 V	20065916	(1222)	-		36786244	(505)	-		-	putative repressor protein CI of bacteriophage BP-933W
15800948 V	20065917	(384)	-		32030721	(143)	-		-	regulatory protein Cro of bacteriophage BP-933W
15800949 V	20065918	(491)	-		-		-		-	putative regulatory protein CII of bacteriophage BP-933W
15800950 V	20065919	(1662)	-		1143412	(649)	-		-	putative replication protein O of bacteriophage BP-933W
15800951 V	20065922	(1207)	-		50846084	(597)	-		-	putative replication protein P of bacteriophage BP-933W
15800952 V	31044253	(505)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800953 V	9632499	(783)	-		56128782	(730)	-		-	unknown protein encoded by bacteriophage BP-933W

Query: gi|15800926 integrase for bacteriophage BP-933W [Escherichia coli O157:H7 EDL933]



- 16 Archaea
- 504 Bacteria
- 0 Metazoa
- 0 Fungi
- 1 Viridiplantae
- 93 Viruses
- 4 Other Eukaryota

	Virus		Eukaryota		Eubacteria		Archaea			
100		73 (66) V		32 (0) E		83 (29) B		21 (0) A		
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
15800923 B	-		108880530	(121)	74311559	(2138)	-		3	periplasmic glucose-1-phosphatase
15800924 B	-		-		24051292	(393)	-		-	hypothetical protein Z1422
15800925 B	-		53749369	(469)	82776291	(916)	72396581	(441)	-	TrpR binding protein WrbA
15800926 V	32128262	(2311)	-		16764254	(1504)	14324479	(111)	-	integrase for bacteriophage BP-933W
15800927 V	32128263	(528)	-		16764255	(261)	-		-	putative excisionase for bacteriophage BP-933W
15800928 V	32128264	(549)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800929 V	32128266	(638)	-		56128804	(429)	-		-	unknown protein encoded by bacteriophage BP-933W
15800930 V	32128267	(1101)	-		24050523	(173)	-		-	unknown protein encoded by bacteriophage BP-933W
15800931 V	32128269	(504)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800932 V	32128271	(500)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800933 V	32128274	(1327)	108705972	(111)	16760392	(311)	-		-	unknown protein encoded by bacteriophage BP-933W
15800934 V	32128278	(489)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800935 V	20065894	(313)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800936 V	20065897	(1203)	54637203	(120)	82776669	(1174)	-		1	putative exonuclease of bacteriophage BP-933W
15800937 V	20065900	(1370)	-		16759879	(1307)	-		-	putative Bet recombination protein of bacteriophage BP-933W
15800938 V	31044238	(514)	-		82776671	(447)	-		-	putative host-nuclease inhibitor protein Gam of bacteriphage BP-933W
15800939 V	31044239	(447)	-		-		-		-	putative Kil protein of bacteriphage BP-933W
15800940 V	31044240	(646)	-		67158861	(144)	-		-	putative single-stranded DNA binding protein
15800941 V	31044241	(849)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800942 V	20065909	(645)	-		-		-		-	putative antitermination protein N of bacteriophage BP-933W
15800943 V	20065910	(271)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800944 V	20065911	(1861)	37575481	(257)	91779963	(649)	48429855	(215)	68	putative serine/threonine kinase encoded by bacteriophage BP-933W
15800945 V	20065912	(759)	-		91779964	(281)	-		-	unknown protein encoded by bacteriophage BP-933W
15800946 V	20065914	(615)	-		37198778	(267)	-		-	unknown protein encoded by bacteriophage BP-933W
15800947 V	20065916	(1222)	-		36786244	(505)	-		-	putative repressor protein CI of bacteriophage BP-933W
15800948 V	20065917	(384)	-		32030721	(143)	-		-	regulatory protein Cro of bacteriophage BP-933W
15800949 V	20065918	(491)	-		-		-		-	putative regulatory protein CII of bacteriophage BP-933W
15800950 V	20065919	(1662)	-		1143412	(649)	-		-	putative replication protein O of bacteriophage BP-933W
15800951 V	20065922	(1207)	-		50846084	(597)	-		-	putative replication protein P of bacteriophage BP-933W
15800952 V	31044253	(505)	-		-		-		-	unknown protein encoded by bacteriophage BP-933W
15800953 V	9632499	(783)	-		56128782	(730)	-		-	unknown protein encoded by bacteriophage BP-933W

NCBI | BLAST | Protein | Structure | PubMed | Taxonomy | Genome | Nucleotide | 3D-Domains | Books | Help

Query: gi|15800926 integrase for bacteriophage BP-933W [Escherichia coli O157:H7 EDL933]
Matching gi: 32171084, 32170917, 32128262, 32128095, 20065877, 19911669, 12514276, 9633397, 9632467, 7649821, 5881593, 4585378

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 150 unique species Sort by taxonomy proximity

0 Archaea | 182 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 18 Viruses | 0 Other Eukaryotae

Keep only | Cut-Off 95 | Select | Reset | New search by GI: 15800926 | Go

444 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
Conserved Domain Database hits				
2271	9	NP_309187	15830414	putative integrase [Escherichia coli O157:H7 str. Sakai]
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 K12 phage EH297]
1504	7	NP_459869	16764254	putative integrase [Salmonella typhimurium LT2]
1497	9	NP_311040	15832267	putative integrase [Escherichia coli O157:H7 str. Sakai]
1493	1	CAD88804	49523586	putative integrase [Phage BP-4795]
1390	7	NP_805696	29142354	integrase [Salmonella enterica subsp. enterica serovar Typhi T
1389	7	NP_459980	16764365	integrase [Salmonella typhimurium LT2]
1373	7	YP_215950	62179533	Gifsy-2 prophage integrase [Salmonella enterica subsp. enteric
1120	9	AP_002200	89108420	predicted defective integrase [Escherichia coli W3110]
1116	9	NP_416096	16129537	Qin prophage; predicted defective integrase [Escherichia coli

NCBI

BLAST | Protein | Structure | PubMed | Taxonomy
Genome | Nucleotide | 3D-Domains | Books | Help

Query: gi|15800926 integrase for bacteriophage BP-933W [Escherichia coli O157:H7 EDL933]
Matching gi: 32171084, 32170917, 32128262, 32128095, 20065877, 19911669, 12514276, 9633397, 9632467, 7649821, 5881593, 4585378

COG582 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 150 unique species Sort by taxonomy proximity

0 Archaea | 182 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 18 Viruses | 0 Other Eukaryotae

Keep only | Cut-Off 95 | Select | Reset | New search by GI: 15800926 | Go

444 aa

SCORE	P	ACCESSION	GI	N	ORGANISM
Conserved Domain Database hits					
2271	9	NP_309187	15830414	- 4	*Escherichia coli O157:H7 str. Sakai*
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*
1504	7	NP_459869	16764254	- 2	*Salmonella typhimurium LT2*
1493	1	CAD88804	49523586	- 1	*Phage BP-4795*
1390	7	NP_805696	29142354	- 1	*Salmonella enterica subsp. enterica serovar Typhi Ty2*
1373	7	YP_215950	62179533	- 2	*Salmonella enterica subsp. enterica serovar Choleraesuis str*
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. IP275*
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*
822	7	YP_408146	82544199	- 1	*Shigella boydii Sb227*
544	9	ZP_009...	83587486	- 1	*Escherichia coli 101-1*
543	9	ZP_007...	75259458	- 3	*Escherichia coli E22*
542	7	ZP_009...	83569428	- 1	*Shigella dysenteriae 1012*
538	7	NP_837464	30063293	- 1	*Shigella flexneri 2a str. 2457T*
510	5	ZP_012...	91224291	- 1	*Vibrio alginolyticus 12G01*
453	10	NP_287542	15801525	- 2	*Escherichia coli O157:H7 EDL933*
445	5	YP_125555	54293140	- 1	*Legionella pneumophila str. Lens*
445	5	YP_122540	54296171	- 2	*Legionella pneumophila str. Paris*
416	1	NP_490644	17313264	- 2	*Pseudomonas phage phi CTX*
376	5	ZP_008...	78687994	- 2	*Shewanella sp. ANA-3*

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
J	mll0487 mll4043 mll5763 mll5956 mll5957 mll5958 mll6039 mll6040 mll6042 mll6043 mll6044 mll6045 mll6228 mll6229 mll6230 mll6374 mll6431 mll6432 mll7973 mll8495 mlr0475	CC0344 CC1766

66 genomes version

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		
	452 [illegible]		785 891		780 1284	
aompkzyqvdrlbcefghsnujxitw	84	aomp------r--c-f-----j----	6	----------e-gh--------	41	a-m----------
aompkzyqvdrlbcefghsnuj----	11	aompkz-q-dr-bcefgh--uj----	6	------qvdrlbcefghsnujxitw	30	aompkz-------
aompk-yqvdrlbcefghsnuj----	8	a-m-----v----------	4	----------efg----j----	22	a-m-k--------
aom---yqvdrlbcefghsnuj----	7	a-m-----v----c----------	4	----------efg--------	18	--m-k--------
aompkzyqvdrlbcefghsnujxit-	7	aom-kz-qvdrlbcefghs-uj--tw	4	----------efgh-------	18	aom-kz-------
aompkzyqvdrlbcefghsnuj-i--	6	aompkz-qvdrlbcefghsnujxitw	4	----------f-----j----	15	a-m-kz-------
aomp--yqvdrlbcefghsnuj----	5	--m--------lb----------	3	------qvdrlbcefghsnujxit-	15	aompk--------
aomp-zyq-dr-bcefghsnujxi--	5	--m----qvdrlbcefghsnujxit-	3	----------fg----j----	13	aom----------
-omp--yqvdrlbcefghsnujxitw	4	--m-k-------b----------	3	----------ef-----j----	13	aom-k--------
a-m-kzyqvdrlbcefghsnuj----	4	--mpk---v-r----------	3	----------efghsn------	11	a---kz-------
aompkzy--d-l----------it-	4	-o----------e-gh-------	3	------qv---b-efg---uj--t-	10	a-mpkz-------
aompkzyqvdrlbcefghsnuj-itw	4	-o----------b-efg----j----	3	------qv---b-efg---uj-it-	9	aomp---------
-o-pkzyqvdr-bcef--sn-j----	3	-o--------r-b----------	3	----------ef--s-------	8	-ompk--------
a-m---yqvdrlbcefg-s-uj----	3	-om----qvdrlbcefghsnujxitw	3	----------efghs-------	8	ao-p---------
aom--zyqvdrlbcefghsnuj-i--	3	-om-k---vd--b--fg-----x---	3	----------r----f-----j----	8	-om-k--------
aompkzyq----------	3	a-m----------c----------	3	----------h-nu-----	7	ao-pk--------
aompkzyqvdrlbcefghsnuj-it-	3	a-m-k---v----------	3	----------efgh-n------	7	--m-kz-------
aompkzyqvdrlbcefghsnujx-t-	3	a-m-k--q-----ce-----u-----	3	----------r----------j----	7	-om----------
----k-y--dr-b-efgh-nuj----	2	a-m-kz-q----------	3	----------jx---	6	a-m--z-------
-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 patte
-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	aompkzy-------
ao--k-yqvdrlbcefghsnujxi--	2	----k----d-lb-efgh-n------	2	----------lb-e----------	5	aom-kzy-------
ao-p--yqvdrlbcefghsnujxitw	2	----k--qv----------	2	------qvdrlbcefghsnujx-t-	5	a-m-kzy-------
ao-p-zy-vdr-b-ef--sn-j----	2	---p--------b-e-----uj----	2	----------hs--j----	4	a-mpkzy-------
aom---yqvdrlbcefghsnujxit-	2	--m----------c-------j----	2	----------f--s--j----	4	aompk-y-------
aom--zyqvdrlbc-f---nujxitw	2	--m---------b----------	2	----------e-gh-n------	4	----kzy-------
aompk-yqvdrlbcefghsnujxitw	2	--m-----v-----efgh--------	2	----------efg-s--j----	4	--m-kzy-------
aompkzy-vdr-b----h--------	2	--m-----v--lb----------	2	----------b--f-----j----	4	
aompkzy-vdrlbcefghs-uj----	2	--m----qvdr--cef--sn-jx---	2	----------b-e------j----	4	4 single patter
aompkzyqv-rlbce-g---uj--t-	2	--m-k--q----------	2	----------b-e--h---j----	4	
aompkzyqvdrlbcefg-snuj----	2	--mpk-----r---e----------	2	----------l--e-gh--------	4	
aompkzyqvdrlbcefghsnuj--tw	2	-o--------dr-b--f-----j----	2	---------d-----f-----j----	4	
aompkzyqvdrlbcefghsnujxi--	2	-o-p-----d-lb--fg----j----	2	---------d----e------j----	4	
		-o-p-----dr--cef-----j----	2	---------dr-bcef-h--uj---w	4	
413 single patterns		-o-p---q-dr--cef--snujx---	2	--------v---b--------j----	4	------yqvdrlbc
		-o-pkz--vdrlbcefghs--j--tw	2	--------vdrlbcefghsnujxitw	4	------yqv-rlbc
		-om-k---vd--bc-fg-----x---	2	-------q------efghsnujxi--	4	------yqvdrlbc
		-om-k--qvdr-bcefghsnujxit-	2	-------q-----cefghsnujxi--	4	------yqvd-lbc

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  RplO  [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
-  45 aompkzyqvdrlbcefghsnujxitw  RpsM  [J]  COG0099 Ribosomal protein S13
-  49 aompkzyqvdrlbcefghsnujxitw  RpsN  [J]  COG0199 Ribosomal protein S14
-  45 aompkzyqvdrlbcefghsnujxitw  RpsO  [J]  COG0184 Ribosomal protein S15P/S13E
-  45 aompkzyqvdrlbcefghsnujxitw  RpsQ  [J]  COG0186 Ribosomal protein S17
-  44 aompkzyqvdrlbcefghsnujxitw  RpsS  [J]  COG0185 Ribosomal protein S19
-  45 aompkzyqvdrlbcefghsnujxitw  RpsB  [J]  COG0052 Ribosomal protein S2
-  43 aompkzyqvdrlbcefghsnujxitw  RpsC  [J]  COG0092 Ribosomal protein S3
-  46 aompkzyqvdrlbcefghsnujxitw  RpsD  [J]  COG0522 Ribosomal protein S4 and related proteins
-  44 aompkzyqvdrlbcefghsnujxitw  RpsE  [J]  COG0098 Ribosomal protein S5
```

66 genomes version

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

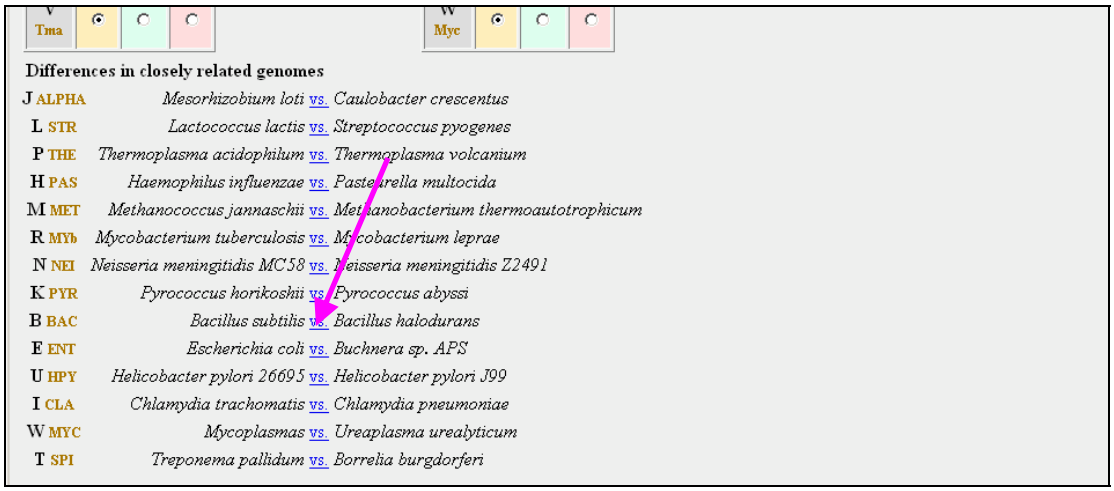
Principal component analysis of genomes
List of COGs
Distribution
Co-occurrences
Phylogenetic patterns
Phylogenetic patterns search

Functional categories

J K L

D O M N P T

Phylogenetic patterns search
NCBI
Search
dc - don't care
Help
Org dc Yes No
A Afu
O Hbs
M Met
P THE
K PYR
Z Ape
Y Sce
Q Aae
V Tma
D Dra
R Mtu
B BAC
L STR
C Ssp
E ENT
F Pae
G Vch
H Hin
S Xfa
N Nme
U HPY
J Cje
X Rpr
I CLA
T SPI
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
Q	28	3	2	24	-	-	-	5	Secondary metabolites biosynthesis, transport and catabolism
R S	681	25	38	404	1	-	-	50	**Poorly characterized**
R	214	7	12	182	-	-	-	34	General function prediction only
S	467	18	26	222	1	-	-	16	Function unknown

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/Arc/MetJ DNA-binding domain
-	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 CbiQ and related transporters
-	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, homologs of TerZ and putative cAMP-bindin
-	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-glucuronosyltransferase
-	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-terminal double-glycine peptidase doma
-	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 including a metal-binding cysteine clu
-	7	----------rl--e------j----	-	[R]	COG3740	Phage head maturation protease
-	4	----------ef--s-------	-	[R]	COG3500	Phage protein D
-	4	----------ef--s-------	-	[R]	COG3499	Phage protein U
-	5	----------ef--s--j----	-	[R]	COG3497	Phage tail sheath protein FI
-	4	----------ef--s-------	-	[R]	COG3498	Phage tail tube protein FII
-	7	-o--------rl--e---s-------	-	[R]	COG3378	Predicted ATPase
-	11	aompkz--------e-g----j----	-	[R]	COG1661	Predicted DNA-binding proteins with PD1-like DNA-binding motif
-	8	---------d--b-ef---nuj----	-	[R]	COG2819	Predicted hydrolase of the alpha/beta superfamily
-	48	-om-kz-qvdrlbcefg----j-it-	-	[R]	COG1266	Predicted metal-dependent membrane protease
-	3	----------efg---------	-	[R]	COG3529	Predicted nucleic-acid-binding protein containing a Zn-ribbon domain
-	9	----------lb-efg-s-uj----	-	[R]	COG2910	Putative NADH-flavin reductase

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?