Mini Courses NCBI Resources one course at a time

Microbial Genomes Quick Start

An NCBI Mini-Course

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

In this mini-course, we will learn:

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

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

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

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

Problem 1

From the NCBI homepage, access the "Genomic biology" site and from there select the "Microbial" Genome Resource link. The link leads to the microbial projects in the Entrez Genome Project database:
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?

On the Complete genomes page change the selection from "Firmicutes" back to "All Organisms". Scroll down the resulting list to the Escherichia coli strains. How many strains of E. coli have been completely sequenced? Select the "Escherichia coli O157:H7 str. Sakai" link that leads to the Genome Project Overview for this organism.

Is the Sakai strain pathogenic in humans? Name a few differences between the pathogenic and non-pathogenic strains. For example, a type III secretion system, the locus of enterocyte effacement, numerous toxins and adhesins, as well as fimbrial gene clusters and iron uptake systems are found in pathogenecity islands. Click on the RefSeq accession number NC_002695 to access the organism's information in the Entrez Genome database.

From the resulting Genome Overview table, access the entry NC_002695 in the Entrez Nucleotide database. Access the Entrez Protein database through the "Protein" link. How many protein coding genes have been annotated on the genome? Go back to the Genome Overview table.
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 Structural RNA table through the "Structural RNAs" link. How many genes encoding arginine tRNA are annotated on the Sakai strain? Go back to the Genome Overview table.

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 Overview table.

Click on the "TaxPlot" link. TaxPlot 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 Overview table.

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

Select the "TaxMap" link. TaxMap reports the taxonomic distribution of the top BLAST hits for each of the proteins in the entire proteome of the Sakai strain by comparing to the non-redundant database (excluding closely related species) by BLAST similarity. Compare the map to that of K12 http://www.ncbi.nlm.nih.gov/sutils/taxik.cgi?gi=115 . What do you conclude? Identify different phages inserted in the Sakai strain genome. On the Sakai map, select one region that is more similar to Viruses. The page will reload to display the homologs for the selected region. Select the link for integrase (gi 15830414). 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. 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

Tools
Data mining

Research at NCBI
People, projects, and seminars

Software engineering
Tools, R&D, and databases

Education
Teaching resources and on-line tutorials

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...

New dbGaP

NCBI's dbGaP Genome Wide Association Database

NCBI's dbGaP (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies. The resource is intended to help elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and pre-computed associations between subjects' phenotypes and genotypes. Click here to read the press release. To read more about GWA projects, see NCBI's GWA resource page

NCBI Citations

Two papers authored by researchers at NCBI are among the top 40 most cited articles for 2005 according to ScienceWatch's "Hottest Research of 2004-05" list. The papers describe NCBI's Conserved Domain Database and Reference Sequence database. Click on the links above to view the papers.

PubMed Central

An archive of biomedical and life sciences journals

- Free fulltext
- Over 900,000 articles from over 300 journals
- Linked to PubMed and fully searchable

Use of PubMed Central requires no registration or fee. Access it from any computer with an Internet connection.

NCBI News

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 resources
- My NCBI
- ORF finder
- Rat genome resources
- Reference sequence project
- SAGEmap
- SKY/CGH database

NCBI

Genomic Biology

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

Search All Databases 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
for genetic analysis

TaxPlot
3-way genome

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.

Assembly and Annotation Information

- AGP Resources
- Annotation Information
- Assembly Information
- Genome Glossary
- NCBI Handbook, Chapter 14: Genome Assembly and Annotation Process

Announcements

February, 2006

Human Consensus CDS (CCDS): An update for the human CCDS dataset is now available from the CCDS web site and FTP site. [more..]

New Genome Resource Page: A genome resource guide and BLAST page are now available for *Equus caballus* (horse)

Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
Monodelphis domesticus (opossum)	MonDom5	March 8, 2007
Danio rerio (zebrafish)	Zv6	February 27, 2007
Populus trichocarpa (black cottonwood)	1.1	January 12, 2007
Bos taurus (cow)	3.1	January 3, 2007
Gallus gallus (chicken)	2.1	November 30, 2006
Drosophila melanogaster (fruit fly)	5.1	November 21, 2006
Caenorhabditis elegans (nematode)	WS160	November 21, 2006
Dictyostelium discoideum	2.1	November 2, 2006
Strongylocentrotus purpuratus (sea	2.1	October 18, 2006

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 G B M P
- Cat G B M P
- Chicken G B M P
- Chimpanzee G B M P
- Cow G B M P
- Dictyostelium G B M P
- Dog G B M P

NCBI ENTREZ Genome Project

connection information discovery

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

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: -- All --

Tools legend: T - TaxMap; P - ProtTable; 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

487 Complete Microbial Genomes selected: [A] - 38, [B] - 449

save

PID	Organism	King	Group	*Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
15771	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
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	11/22/06	DOE Joint Genome Institute	T P D L S G X F
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	Betaproteobacteria	5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	01/05/07	DOE Joint Genome Institute	T P D L S G X M F
15685	Acidovorax sp. JS42	B	Betaproteobacteria	4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	01/11/07	DOE Joint Genome Institute [more]	T P D L S G X M F
17477	Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	03/13/07	Yale Univ.	P L S G X M F R
12352	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 M F R
18221	Actinobacillus pleuropneumoniae L20	B	Gammaproteobacteria	2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	02/27/07	National Research Council of Canada [more]	P L S G X M F
16697	Aeromonas hydrophila subsp. hydrophila ATCC 7966	B	Gammaproteobacteria	4.7	61.5	1		CP000462.1	NC_008570.1	11/06/06	11/22/06	TIGR	T P D L S G X F R
16723	Aeromonas salmonicida subsp. salmonicida A449	B	Gammaproteobacteria	5.03	58.2	1	2	CP000644.1	NC_009348.1	04/06/07	04/19/07	National Research Council Canada - Institute of Marine Biosciences	P L S G X F
211	Aeropyrum pernix K1	A	Crenarchaeota	1.7	56.3	1		BA000002.3	NC_000854.2	06/26/99	11/28/06	NITE	T P C D L S G X F R
282	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
283	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
13005	Alcanivorax borkumensis SK2	B	Gammaproteobacteria	3.1	54.7	1		AM286690.1	NC_008260.1	07/24/06	01/26/07	Bielefeld Univ	T P D L S G X F R
15763	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
10642	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

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.

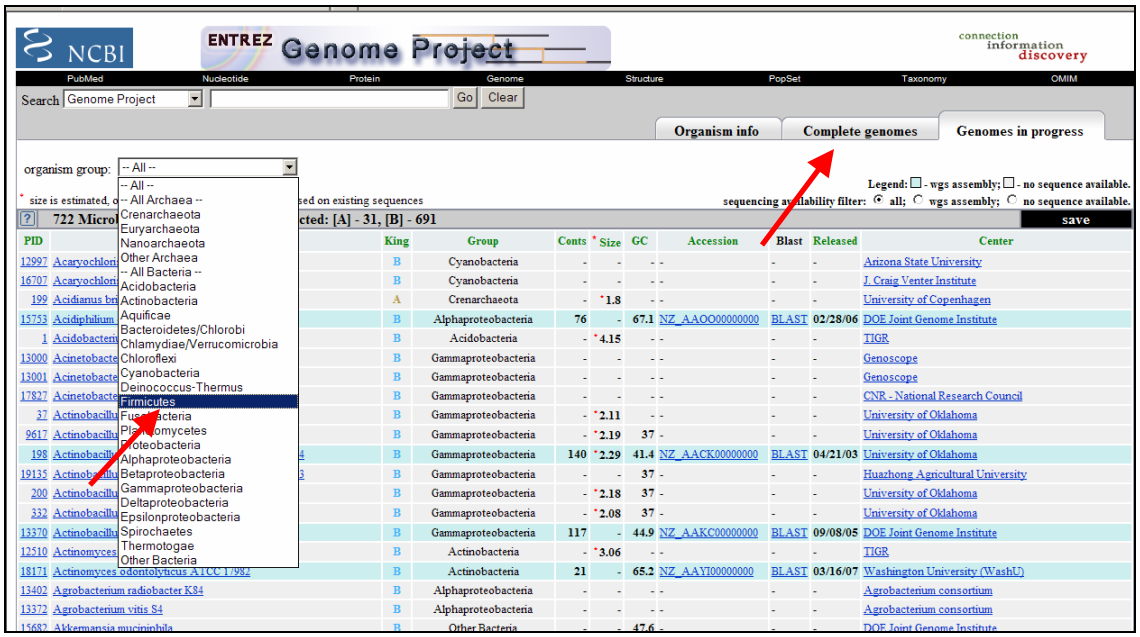
* size is estimated, otherwise genome size is calculated based on existing sequences

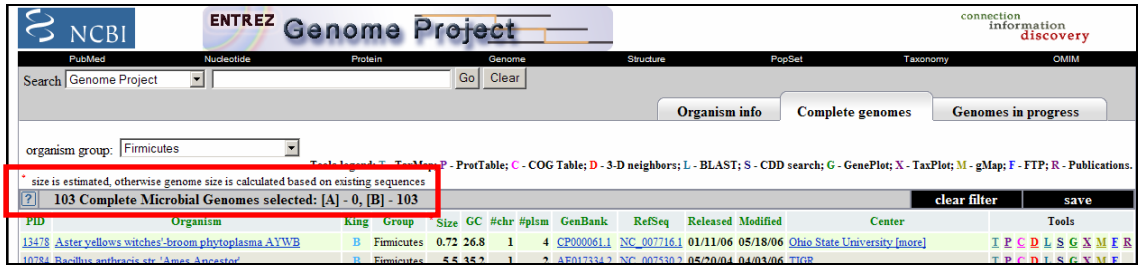
sequencing availability filter: ◉ all; ○ wgs assembly; ○ no sequence available.

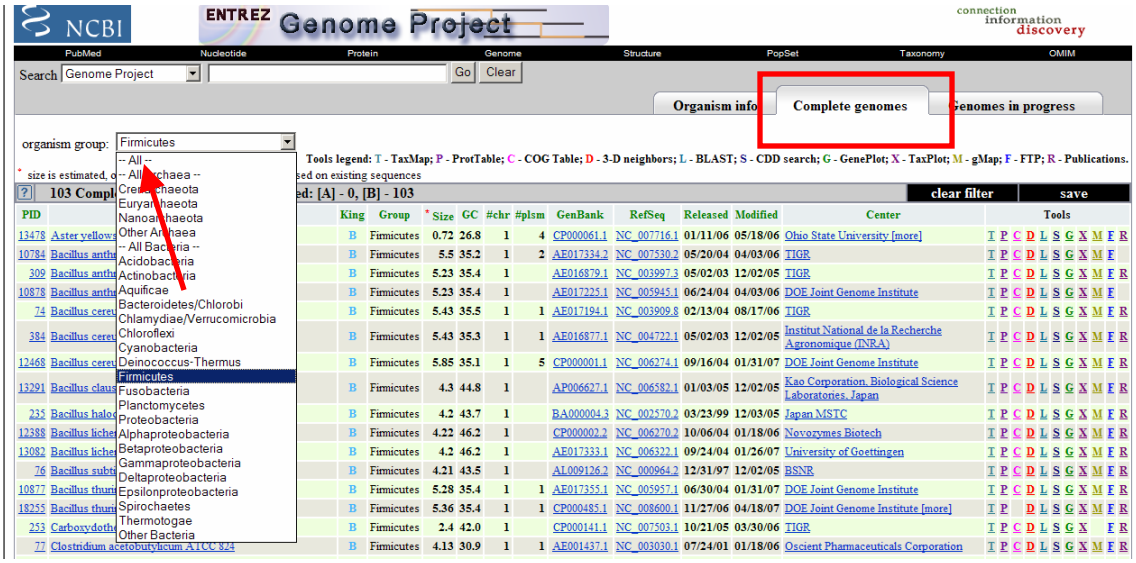
722 Microbial Genomes In-Progress selected: [A] - 31, [B] - 691

save

PID	Organism	King	Group	Conts	*Size	GC	Accession	Blast	Released	Center
12997	Acaryochloris marina	B	Cyanobacteria	-	-	- -		-	-	Arizona State University
16707	Acaryochloris sp. CCMEE 5410	B	Cyanobacteria	-	-	- -		-	-	J. Craig Venter Institute
199	Acidianus brierleyi	A	Crenarchaeota	-	*1.8	- -		-	-	University of Copenhagen
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	76	-	67.1	NZ_AAOO00000000	BLAST	02/28/06	DOE Joint Genome Institute
1	Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	-	*4.15	- -		-	-	TIGR
13000	Acinetobacter baumannii	B	Gammaproteobacteria	-	-	- -		-	-	Genoscope
13001	Acinetobacter baumannii	B	Gammaproteobacteria	-	-	- -		-	-	Genoscope
17827	Acinetobacter baumannii ICU	B	Gammaproteobacteria	-	-	- -		-	-	CNR - National Research Council
37	Actinobacillus actinomycetemcomitans HK1651	B	Gammaproteobacteria	-	*2.11	- -		-	-	University of Oklahoma
9617	Actinobacillus pleuropneumoniae serovar 1	B	Gammaproteobacteria	-	*2.19	37 -		-	-	University of Oklahoma
198	Actinobacillus pleuropneumoniae serovar 1 str. 4074	B	Gammaproteobacteria	140	*2.29	41.4	NZ_AACK00000000	BLAST	04/21/03	University of Oklahoma
19135	Actinobacillus pleuropneumoniae serovar 3 str. JL03	B	Gammaproteobacteria	-	-	37 -		-	-	Huazhong Agricultural University
200	Actinobacillus pleuropneumoniae serovar 5	B	Gammaproteobacteria	-	*2.18	37 -		-	-	University of Oklahoma
332	Actinobacillus pleuropneumoniae serovar 7	B	Gammaproteobacteria	-	*2.08	37 -		-	-	University of Oklahoma
13370	Actinobacillus succinogenes 130Z	B	Gammaproteobacteria	117	-	44.9	NZ_AAKC00000000	BLAST	09/08/05	DOE Joint Genome Institute
12510	Actinomyces naeslundii MG1	B	Actinobacteria	-	*3.06	- -		-	-	TIGR
18171	Actinomyces odontolyticus ATCC 17982	B	Actinobacteria	21	-	65.2	NZ_AAYI00000000	BLAST	03/16/07	Washington University (WashU)
13402	Agrobacterium radiobacter K84	B	Alphaproteobacteria	-	-	- -		-	-	Agrobacterium consortium
13372	Agrobacterium vitis S4	B	Alphaproteobacteria	-	-	- -		-	-	Agrobacterium consortium
15682	Akkermansia muciniphila	B	Other Bacteria	-	-	47.6 -		-	-	DOE Joint Genome Institute
18947	Alcanivorax sp. PR1	B	Bacteroidetes/Chlorobi	12	-	38.7	NZ_AAYU00000000	BLAST	03/06/07	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 --
-- All --
-- All Archaea --
Crenarchaeota
Euryarchaeota
Nanoarchaeota
Other Archaea
-- All Bacteria --
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes/Chlorobi
Chlamydiae/Verrucomicrobia
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Alphaproteobacteria
Betaproteobacteria
Gammaproteobacteria
Deltaproteobacteria
Epsilonproteobacteria
Spirochaetes
Thermotogae
Other Bacteria
Legend: - wgs assembly; - no sequence available.
sequencing availability filter: all; wgs assembly; no sequence available.
save


NCBI
ENTREZ Genome Project
connection information discovery
Search Genome Project Go Clear
Organism info
Complete genomes
Genomes in progress
organism group: Firmicutes
Tools legend: T - TaxMap; P - ProtTable; 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
103 Complete Microbial Genomes selected: [A] - 0, [B] - 103
clear filter
save


NCBI
ENTREZ Genome Project
connection information discovery
Search Genome Project Go Clear
Organism info
Complete genomes
Genomes in progress
organism group: Firmicutes
Tools legend: T - TaxMap; P - ProtTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - Publications.
clear filter
save

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 - ProtTable; 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

487 Complete Microbial Genomes selected: [A] - 38, [B] - 449 save

PID	Organism	King	Group	*Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
15771	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
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	11/22/06	DOE Joint Genome Institute	T P D L S G X F
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	Betaproteobacteria	5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	01/05/07	DOE Joint Genome Institute	T P D L S G X M F
15685	Acidovorax sp. JS42	B	Betaproteobacteria	4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	01/11/07	DOE Joint Genome Institute [more]	T P D L S G X M F
17477	Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	03/13/07	Yale Univ.	P L S G X M F R
12352	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 M F R
18221	Actinobacillus pleuropneumoniae L20	B	Gammaproteobacteria	2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	02/27/07	National Research Council of Canada [more]	P L S G X M F
16697	Aeromonas hydrophila subsp. hydrophila ATCC 7966	B	Gammaproteobacteria	4.7	61.5	1		CP000462.1	NC_008570.1	11/06/06	11/22/06	TIGR	T P D L S G X F R
16723	Aeromonas salmonicida subsp. salmonicida A449	B	Gammaproteobacteria	5.03	58.2	1	2	CP000644.1	NC_009348.1	04/06/07	04/19/07	National Research Council Canada - Institute of Marine Biosciences	P L S G X F
211	Aeropyrum pernix K1	A	Crenarchaeota	1.7	56.3	1		BA000002.3	NC_000854.2	06/26/99	11/28/06	NITE	T P C D L S G X F R
282	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
283	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
13005	Alcanivorax borkumensis SK2	B	Gammaproteobacteria	3.1	54.7	1		AM286690.1	NC_008260.1	07/24/06	01/26/07	Bielefeld Univ	T P D L S G X F R
15763	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

13480	Erythrobacter litoralis HTCC2594	B	Alphaproteobacteria	3.05	63.1	1		CP000157.1	NC_007722.1	01/11/06	03/30/06	J. Craig Venter Institute [more]	T P C D L S G X M F
16235	Escherichia coli 536	B	Gammaproteobacteria	4.94	50.5	1		CP000247.1	NC_008253.1	07/13/06	07/25/06	University of Goettingen	T P D L S G X M F
16718	Escherichia coli APEC O1	B	Gammaproteobacteria	5.1	50.5	1		CP000468.1	NC_008563.1	10/04/06	04/09/07	Iowa State University	T P D L S G X M F R
313	Escherichia coli CFT073	B	Gammaproteobacteria	5.23	50.5	1		AE014075.1	NC_004431.1	12/10/02	04/04/06	University of Wisconsin-Madison, USA	T P C D L S G X M F R
225	Escherichia coli K12	B	Gammaproteobacteria	4.6	50.8	1		U00096.2	NC_000913.2	09/05/97	04/24/07	University of Wisconsin-Madison, USA	T P C D L S G X M F R
259	Escherichia coli O157:H7 EDL933	B	Gammaproteobacteria	5.62	50.3	1	1	AE005174.2	NC_002655.2	02/24/01	07/19/06	University of Wisconsin-Madison, USA	T P C D L S G X M F R
226	Escherichia coli O157:H7 str. Sakai	B	Gammaproteobacteria	5.6	50.5	1	2	BA000007.2	NC_002695.1	03/29/00	03/30/07	Osaka Univ.	T P C D L S G X M F R
16259	Escherichia coli UTI89	B	Gammaproteobacteria	5.21	50.6	1	1	CP000243.1	NC_007946.1	04/05/06	03/12/07	Washington University (WashU)	T P D L S G X M F R
16351	Escherichia coli W3110	B	Gammaproteobacteria	*4.65	50.8	1		AP009048.1	AC_000091.1	01/23/06	04/12/06	Nara Institute of Science and Technology	T P D L S G X M F R
16421	Francisella tularensis subsp. holarctica	B	Gammaproteobacteria	1.9	32.2	1		AM233362.1	NC_007880.1	03/02/06	03/30/06	Lawrence Livermore National Laboratory [more]	T P D L S G X M F
17265	Francisella tularensis subsp. holarctica OSU18	B	Gammaproteobacteria	1.9	32.2	1		CP000437.1	NC_008369.1	09/20/06	09/25/06	Baylor College of Medicine	T P D L S G X M F R
16088	Francisella tularensis subsp. novicida U112	B	Gammaproteobacteria	1.9	32.5	1		CP000439.1	NC_008601.1	11/28/06	11/30/06	University of Washington	T P D L S G X M F
17375	Francisella tularensis subsp. tularensis FSC198	B	Gammaproteobacteria	1.9	32.3	1		AM286280.1	NC_008245.1	07/12/06	07/25/06	School of Biological Sciences, University of Birmingham, UK	T P D L S G X M F
9	Francisella tularensis subsp. tularensis SCHU S4	B	Gammaproteobacteria	1.9	32.3	1		AJ749949.1	NC_006570.1	12/17/04	01/26/07	Swedish Defence Research Agency [more]	T P C D L S G X M F R
18459	Francisella tularensis subsp. tularensis WY96-3418	B	Gammaproteobacteria	1.9	32.3	1		CP000608.1	NC_009257.1	03/24/07	03/31/07	BioHealthBase [more]	P L S G X M F

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

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

Search Genome Project for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1 | Environmental: 0 | Eukaryotes: 0 | Prokaryotes: 1

Links

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

Resource Links

NCBI Resources
- BLAST genome
- FTP
- TaxPlot

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

Sequencing Centers

Enterohemorrhagic Escherichia coli

Project data

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

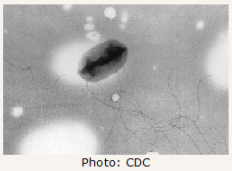


Photo: CDC

Photo: CDC

Genome Projects

Plasmid genome:

- *Escherichia coli* at University of Bristol, Department of Pathology and Microbiology, UK
- *Escherichia coli* at Molecular Biology, Institute of Biochemistry and Physiology of Microorganisms, IBPM RAS, RUSSIA
  - Institute of Biochemistry and Physiology of Microorganisms, Molecular Biology, Russia, Pushchino

Genome sequencing:

- *Escherichia coli 042* at Sanger Institute [In progress]
- *Escherichia coli 101-1* at TIGR [Assembly]
- *Escherichia coli 536* at University of Goettingen [Complete]
- *Escherichia coli 53638* at TIGR [Assembly]
- *Escherichia coli APEC O1* at Iowa State University [Complete]
- *Escherichia coli B* at US DOE Joint Genome Institute (JGI-PGF) [Assembly]
  - ⊞ Expand to show the list of Centers
- *Escherichia coli B str. REL606* at International E. coli B Consortium [In progress]
  - ⊞ Expand to show the list of Centers
- *Escherichia coli B171* at TIGR [Assembly]
- *Escherichia coli B7A* at TIGR [Assembly]
- *Escherichia coli CFT073* at Univ. Wisconsin [Complete]
- *Escherichia coli E110019* at TIGR [Assembly]
- *Escherichia coli E22* at TIGR [Assembly]
- *Escherichia coli E2348/69* at Sanger Institute [In progress]
- *Escherichia coli E24377A* at TIGR [Assembly]
- *Escherichia coli F11* at TIGR [Assembly]
- *Escherichia coli HS* at TIGR [Assembly]
- *Escherichia coli K12* at Univ. Wisconsin [Complete]
- *Escherichia coli O157:H7 EDL933* at Univ. Wisconsin [Complete]
- *Escherichia coli O157:H7 str. GZ-021210/cattle* at South China Agricultural University [In progress]
  - INVITROGEN shanghai.
- ***Escherichia coli O157:H7 str. Sakai O157:H7*** (Project ID: **226**) at GIRC [Complete]
- *Escherichia coli RS218* at University of Wisconsin-Madison [In progress]
  - ⊞ Expand to show the list of Centers
- *Escherichia coli UTI89* at Washington University (WashU) [Complete]

Genome information:

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

Publications:

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

Escherichia coli O157:H7 str. Sakai

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

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

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

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

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

Cellular features					Environment			Temperature	
Gram stain	Shape	Arrangement	Endospores	Motility	Salinity	Oxygen Req.	Habitat	Opt. temp.	Range
-	Rod	Singles, Pairs		Yes		Facultative	Host-associated	37C	Mesophilic
Pathogenic in: Human **Disease:** Hemorrhagic colitis									

NCBI Entrez Genome

My NCBI [Sign In] [Register]

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

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

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

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

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

NCBI Entrez Nucleotide

[Sign In] [Reg

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

Search Nucleotide for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display GenBank Show 5 Send to Hide: sequence all but gene, CDS and mRNA features

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

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

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

Comment Features

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

```
LOCUS       NC_002695            5498450 bp    DNA     circular BCT 30-MAR-2007
DEFINITION  Escherichia coli O157:H7 str. Sakai, complete genome.
ACCESSION   NC_002695
VERSION     NC_002695.1  GI:15829254
PROJECT     GenomeProject:226
KEYWORDS    .
SOURCE      Escherichia coli O157:H7 str. Sakai
  ORGANISM  Escherichia coli O157:H7 str. Sakai
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (sites)
  AUTHORS   Hayashi,T., Makino,K., Ohnishi,M., Kurokawa,K., Ishii,K.,
            Yokoyama,K., Han,C.G., Ohtsubo,E., Nakayama,K., Murata,T.,
            Tanaka,M., Tobe,T., Iida,T., Takami,H., Honda,T., Sasakawa,C.,
            Ogasawara,N., Yasunaga,T., Kuhara,S., Shiba,T., Hattori,M. and
            Shinagawa,H.
  TITLE     Complete genome sequence of enterohemorrhagic Escherichia coli
            O157:H7 and genomic comparison with a laboratory strain K-12
  JOURNAL   DNA Res. 8 (1), 11-22 (2001)
   PUBMED   11258796
  REMARK    Erratum:[DNA Res 2001 Apr 27;8(2):96]
REFERENCE   2  (sites)
  AUTHORS   Yokoyama,K., Makino,K., Kubota,Y., Watanabe,M., Kimura,S.,
            Yutsudo,C.H., Kurokawa,K., Ishii,K., Hattori,M., Abe,H., Iida,T.,
            Yamamoto,K., Hayashi,T., Yasunaga,T., Honda,T., Sasakawa,C. and
            Shinagawa,H.
  TITLE     Complete nucleotide sequence of the prophage VT1-Sakai carrying the
            Shiga toxin 1 genes of the enterohemorrhagic Escherichia coli
            O157:H7 strain derived from the Sakai outbreak
  JOURNAL   Gene 258 (1-2), 127-139 (2000)
   PUBMED   11111050
REFERENCE   3  (sites)
  AUTHORS   Ohnishi,M., Murata,T., Nakayama,K., Kuhara,S., Hattori,M.
```

NCBI

Entrez Protein

All Databases | PubMed | Nucleotide | Protein | Genome | Structure

Search Protein for Go Clear

Limits | Preview/Index | History | Clipboard | Details

About Entrez

Entrez Protein
Help | FAQ

Entrez Tools

Check sequence revision history

LinkOut

My NCBI

Related resources
BLAST

Reference sequence project

Search for Genes

Clusters of orthologous groups

Protein reviews on the web

Search for full length cDNAs

Display Summary Show 20 Sort by Relevance Send to

All: 5253 | Bacteria: 5253 | RefSeq: 5253 | Related Structures: 2959

Items 1 - 20 of 5253

1: NP_944498 Reports
hypothetical protein ECs5362 [Escherichia coli O157:H7 str. Sakai]
gi|38704266|ref|NP_944498.1|[38704266]

2: NP_944587 Reports
hypothetical protein ECs5595 [Escherichia coli O157:H7 str. Sakai]
gi|38704265|ref|NP_944587.1|[38704265]

3: NP_313355 Reports
hypothetical protein ECs5328 [Escherichia coli O157:H7 str. Sakai]
gi|38704264|ref|NP_313355.2|[38704264]

4: NP_313351 Reports
putative structural protein [Escherichia coli O157:H7 str. Sakai]
gi|38704263|ref|NP_313351.2|[38704263]

5: NP_313338 Reports
hypothetical protein ECs5311 [Escherichia coli O157:H7 str. Sakai]
gi|38704262|ref|NP_313338.2|[38704262]

6: NP_944586 Reports
hypothetical protein ECs5593 [Escherichia coli O157:H7 str. Sakai]
gi|38704261|ref|NP_944586.1|[38704261]

7: NP_313325 Reports
hypothetical protein ECs5298 [Escherichia coli O157:H7 str. Sakai]
gi|38704260|ref|NP_313325.2|[38704260]

NCBI Entrez Genome

My NCBI [Sign In] [Register]

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

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All: 1

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

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

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

Gene Classification based on COG functional categories Search gene, GeneID or locus_tag: Find Gene

Zoom

1 nt — 9,910 nt

ECs0003 ECs0004 ECs0008 mogA thrA ECs0006 ECs0007

5498450 nt

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

NCBI Entrez Genome

My NCBI [Sign In] [Register]

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

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Protein Table Show 20 Send to

All: 1

Escherichia coli O157:H7 str. Sakai, complete genome

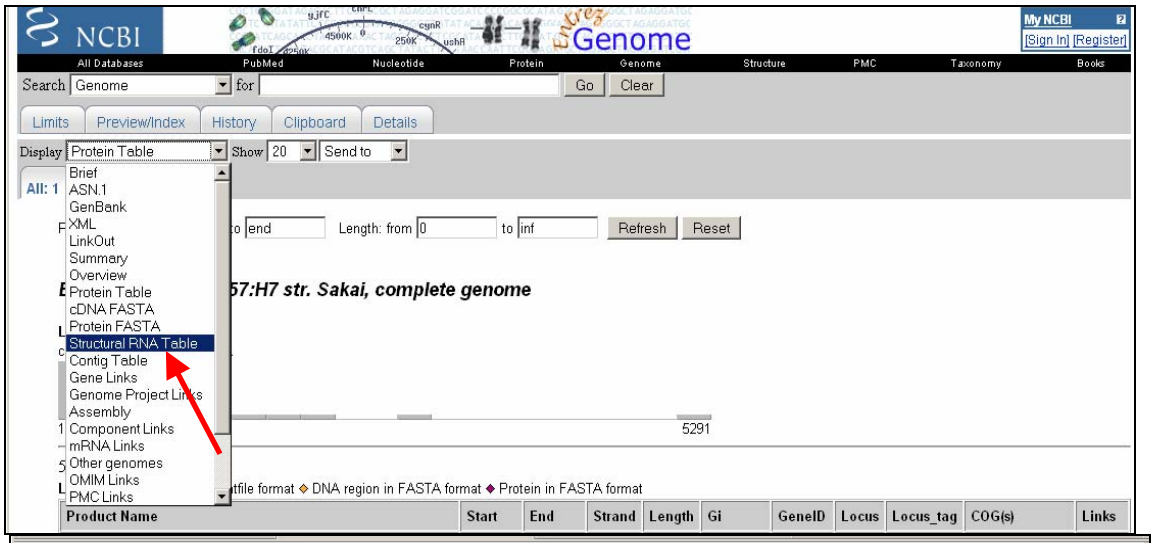
Position: from begin to end Length: from 0 to inf Refresh Reset

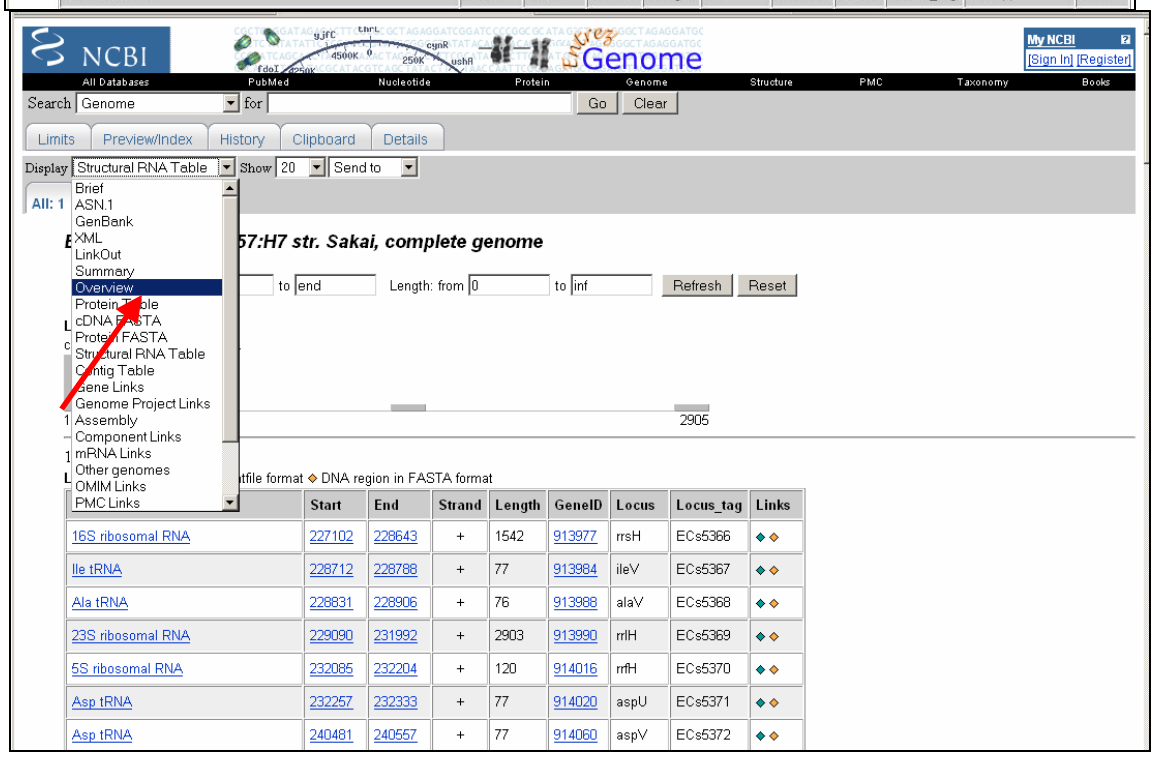
Length histogram.
Click on a bar to select length range.

1 — 5291

5253 protein(s) shown
Legends: ◆ DNA region in flatfile format ◆ DNA region in FASTA format ◆ Protein in FASTA format

Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
thr operon leader peptide	190	273	+	27	15829255	913387	-	ECs0001	-	◆◆◆
bifunctional aspartokinase I/homeserine dehydrogenase I	354	2816	+	820	15829256	913388	thrA	ECs0002	COG0460E, COG0527E	◆◆◆
homoserine kinase	2818	3750	+	310	15829257	913390	-	ECs0003	COG0083E	◆◆◆
threonine synthase	3751	5037	+	428	15829258	913393	-	ECs0004	COG0498E	◆◆◆
hypothetical protein ECs0005	5251	5547	+	98	15829259	913394	-	ECs0005	-	◆◆◆
hypothetical protein ECs0006	5700	6476	-	258	15829260	913395	-	ECs0006	COG3022S	◆◆◆
putative inner membrane transport protein	6546	7976	-	476	15829261	913398	-	ECs0007	COG1115E	◆◆◆
transaldolase	8255	9208	+	317	15829262	913400	-	ECs0008	-	◆◆◆
molybdenum cofactor biosynthesis protein	9323	9910	+	195	15829263	913401	mogA	ECs0009	-	◆◆◆
hypothetical protein ECs0010	9945	10511	-	188	15829264	913402	-	ECs0010	COG1584S	◆◆◆
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]


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


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
Brief
ASN.1
GenBank
XML
LinkOut
Summary
Overview
Protein Table
cDNA FASTA
Protein FASTA
Structural RNA Table
Contig Table
Gene Links
Genome Project Links
Assembly
Component Links
mRNA Links
Other genomes
OMIM Links
PMC Links
57:H7 str. Sakai, complete genome
to end Length: from 0 to inf Refresh Reset
2905
tfile format ◆ DNA region in FASTA format
Start End Strand Length GeneID Locus Locus_tag Links
16S ribosomal RNA 227102 228643 + 1542 913977 rrsH ECs5366
Ile tRNA 228712 228788 + 77 913984 ileV ECs5367
Ala tRNA 228831 228906 + 76 913988 alaV ECs5368
23S ribosomal RNA 229090 231992 + 2903 913990 rrlH ECs5369
5S ribosomal RNA 232085 232204 + 120 914016 rrfH ECs5370
Asp tRNA 232257 232333 + 77 914020 aspU ECs5371
Asp tRNA 240481 240557 + 77 914060 aspV ECs5372

NCBI Entrez Genome

My NCBI [Sign In] [Register]

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

Search Genome for Go Clear

Limits | Preview/Index | History | Clipboard | Details

Display Overview Show 20 Send to

All 1

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

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

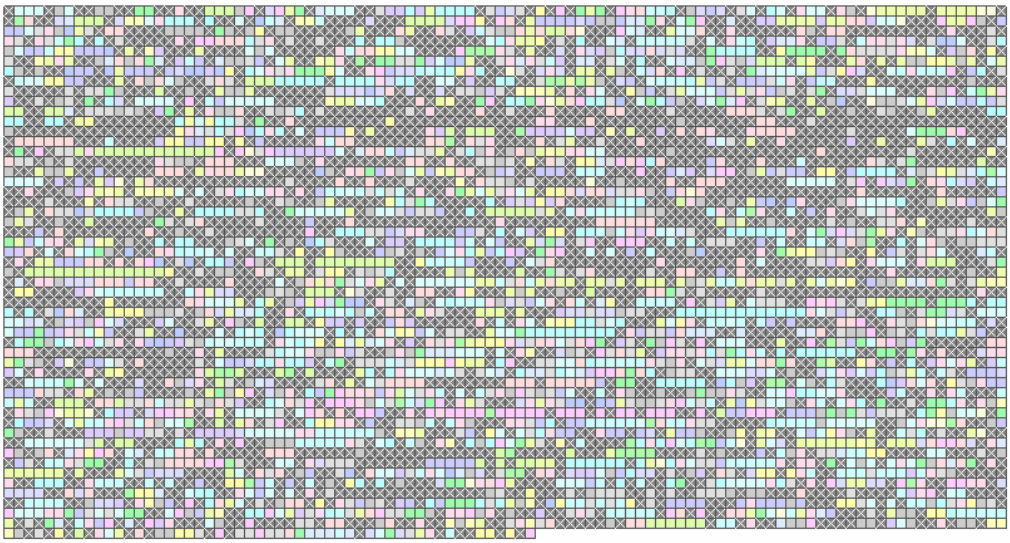
Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

Escherichia coli O157:H7 str. Sakai **sequence** Microbial genomes

genus: ***Escherichia***
group: ***Gammaproteobacteria***
kingdom: *Bacteria*

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



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

Escherichia coli O157:H7 str. Sakai

Intracellular trafficking and secretion genes:

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

All: 1

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

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

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

Protein homologs in Complete Microbial / Eukaryotic genomes

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

Select your query genome

386585 E. coli O157:H7 str. Sakai

Choose two species for comparison

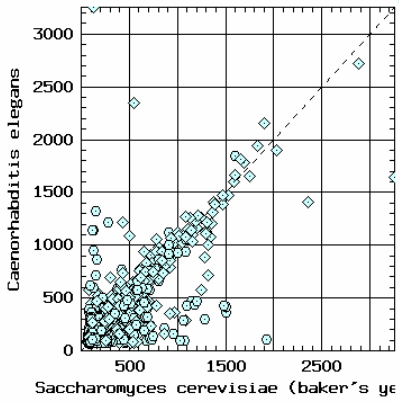
6239 Caenorhabditis elegans

4932 Saccharomyces cerevisiae (baker's yeast)

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

694 hits

24 equal hits

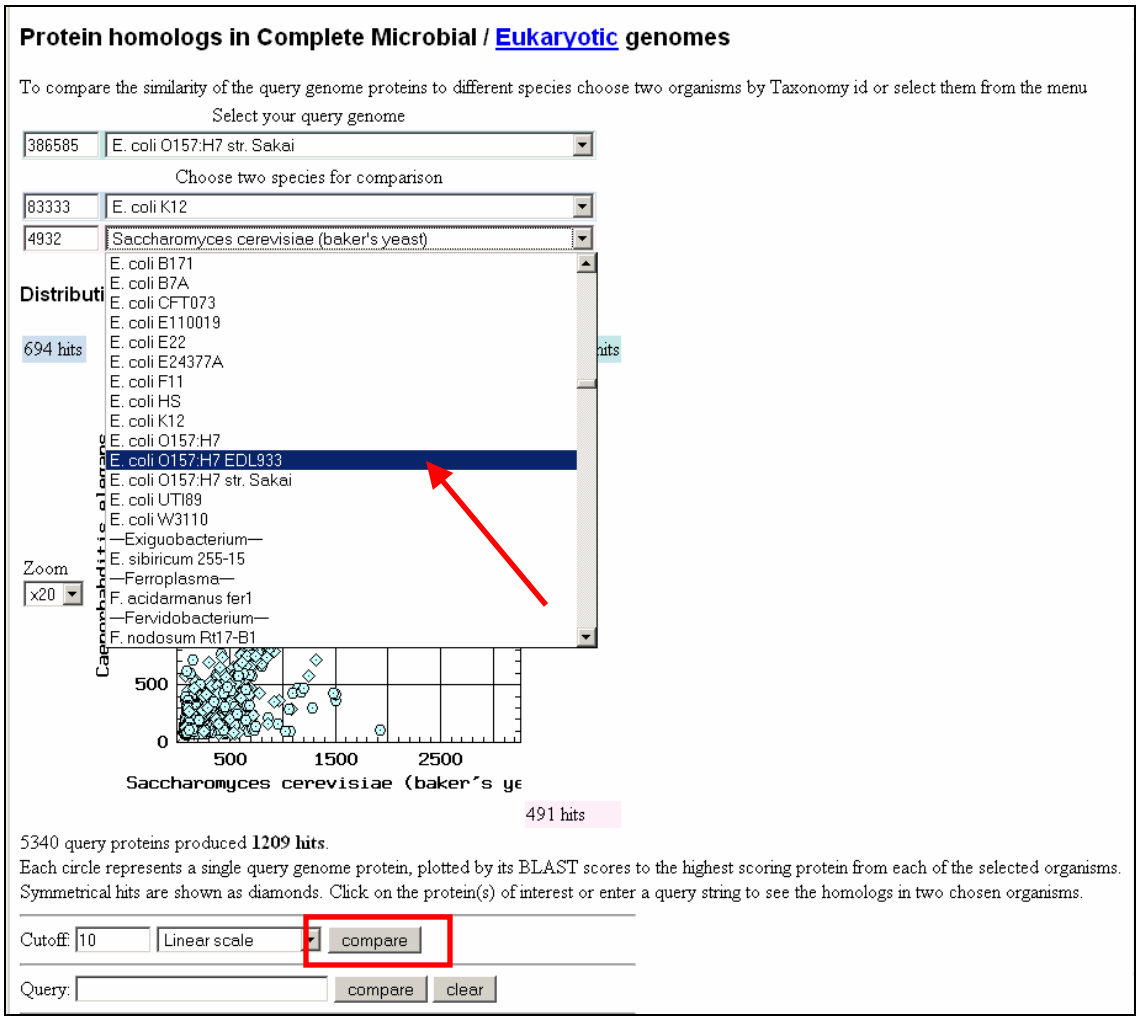


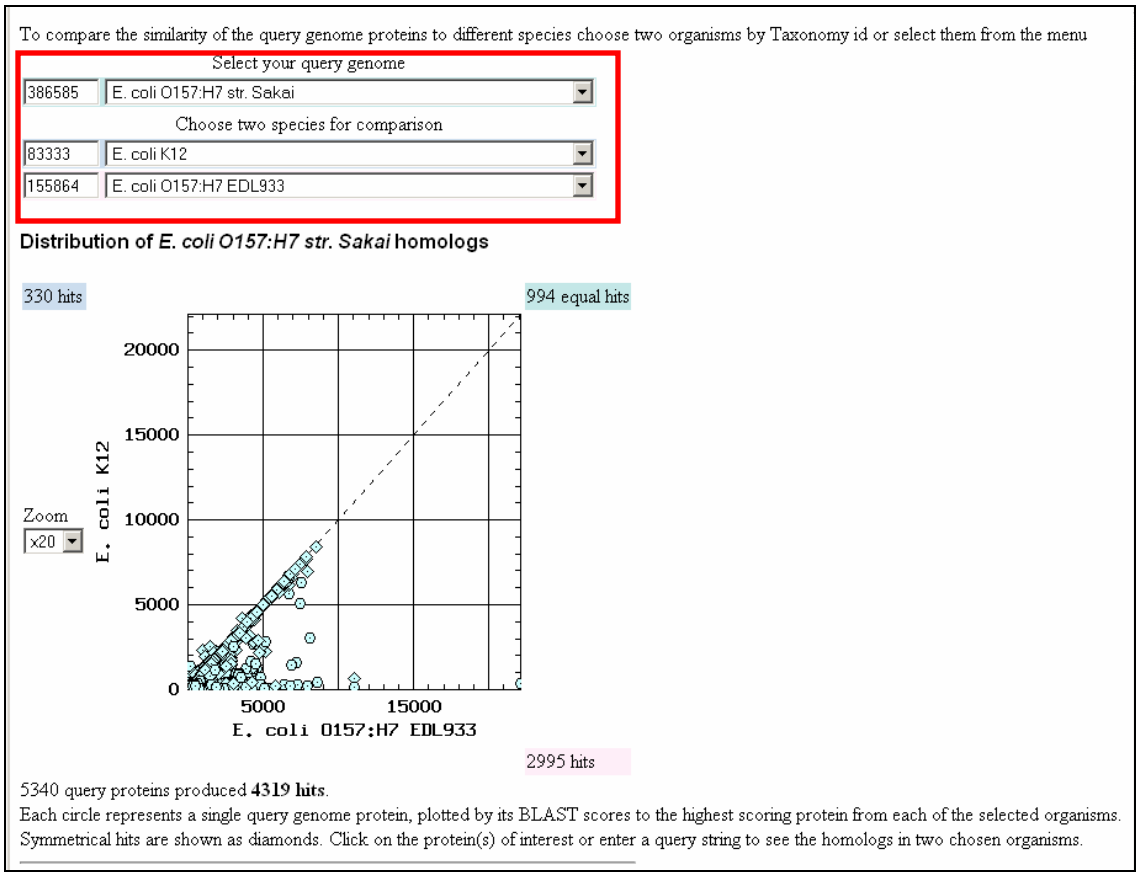
Zoom x20

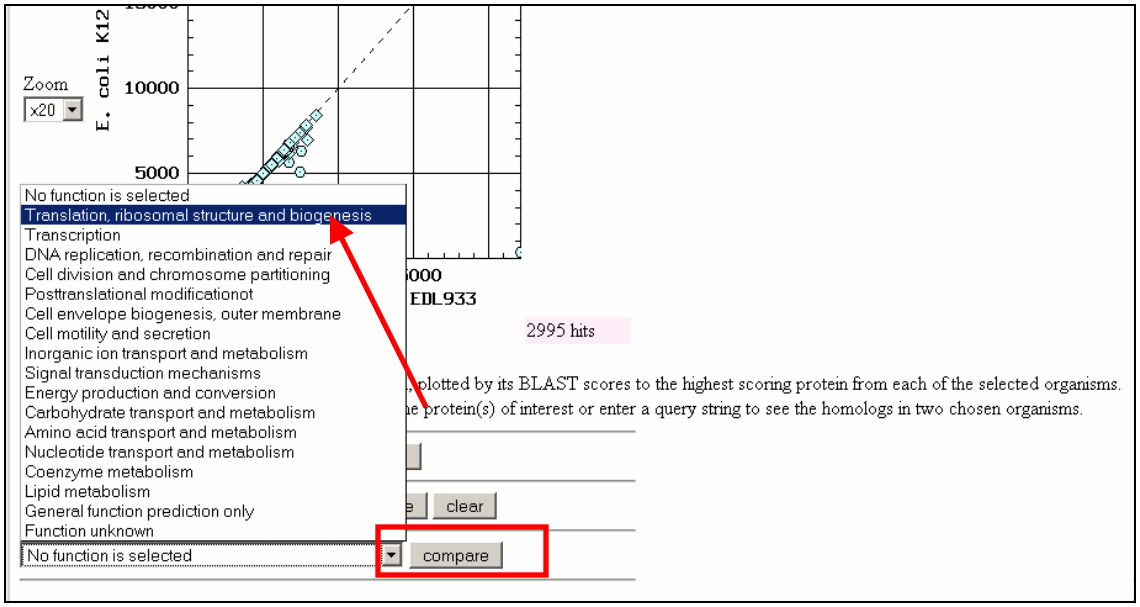
491 hits

5340 query proteins produced **1209 hits**.

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


Protein homologs in Complete Microbial / Eukaryotic genomes
To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu
Select your query genome
386585
E. coli O157:H7 str. Sakai
Choose two species for comparison
83333
E. coli K12
4932
Saccharomyces cerevisiae (baker's yeast)
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
—Ferroplasma—
F. acidarmanus fer1
—Fervidobacterium—
F. nodosum Rt17-B1
Distributi
694 hits
Zoom
x20
500
0
500
1500
2500
Saccharomyces cerevisiae (baker's ye
491 hits
5340 query proteins produced 1209 hits.
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


To compare the similarity of the query genome proteins to different species choose two organisms by Taxonomy id or select them from the menu
Select your query genome
386585
E. coli O157:H7 str. Sakai
Choose two species for comparison
83333
E. coli K12
155864
E. coli O157:H7 EDL933
Distribution of E. coli O157:H7 str. Sakai homologs
330 hits
994 equal hits
Zoom
x20
E. coli K12
20000
15000
10000
5000
0
5000
15000
E. coli O157:H7 EDL933
2995 hits
5340 query proteins produced 4319 hits.
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.


Zoom
x20
E. coli K12
10000
5000
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
EDL933
2995 hits
plotted by its BLAST scores to the highest scoring protein from each of the selected organisms.
protein(s) of interest or enter a query string to see the homologs in two chosen organisms.
clear
No function is selected
compare

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

Select your query genome

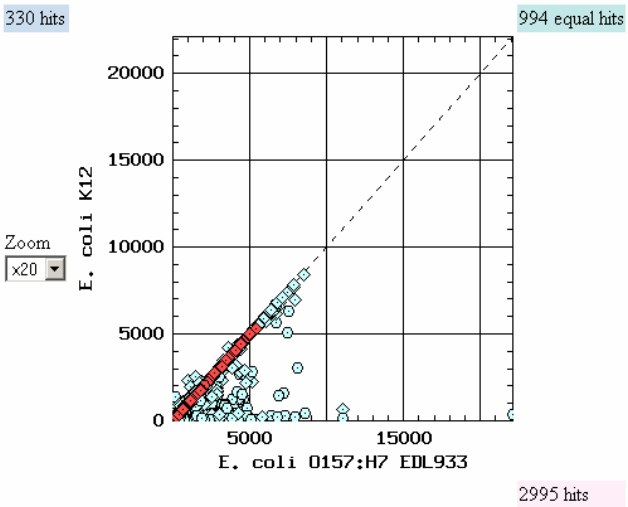
386585 | E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 | E. coli K12

155864 | E. coli O157:H7 EDL933

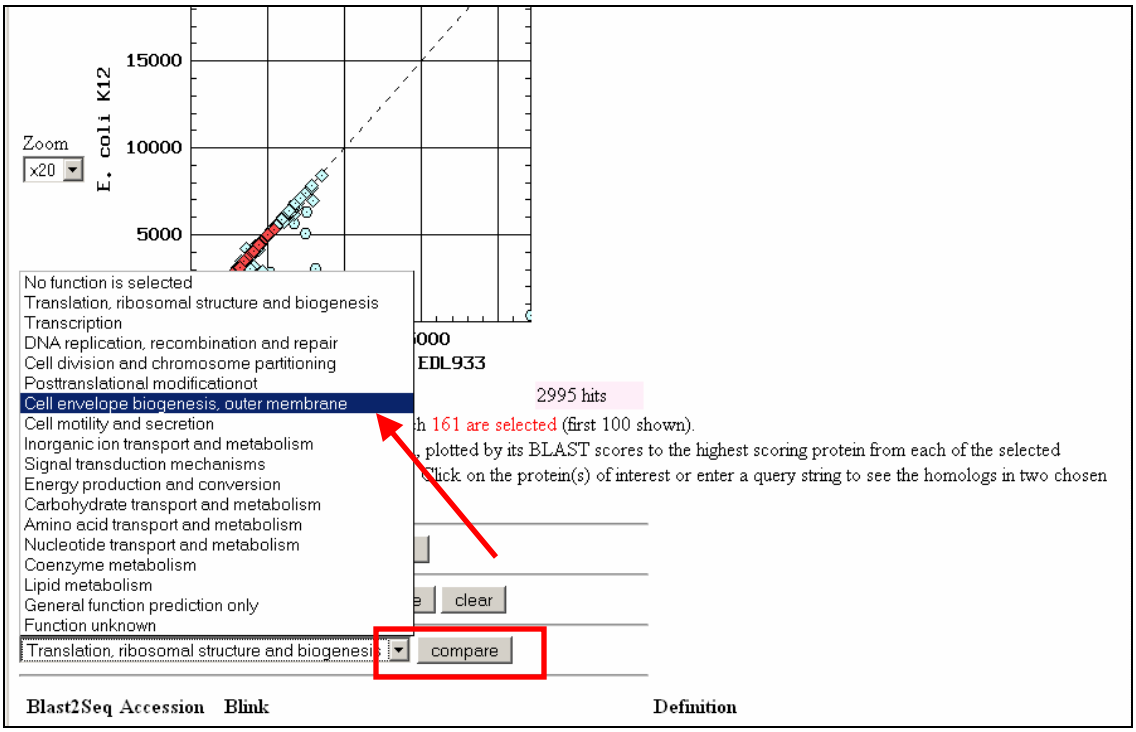
Distribution of *E. coli O157:H7 str. Sakai* homologs

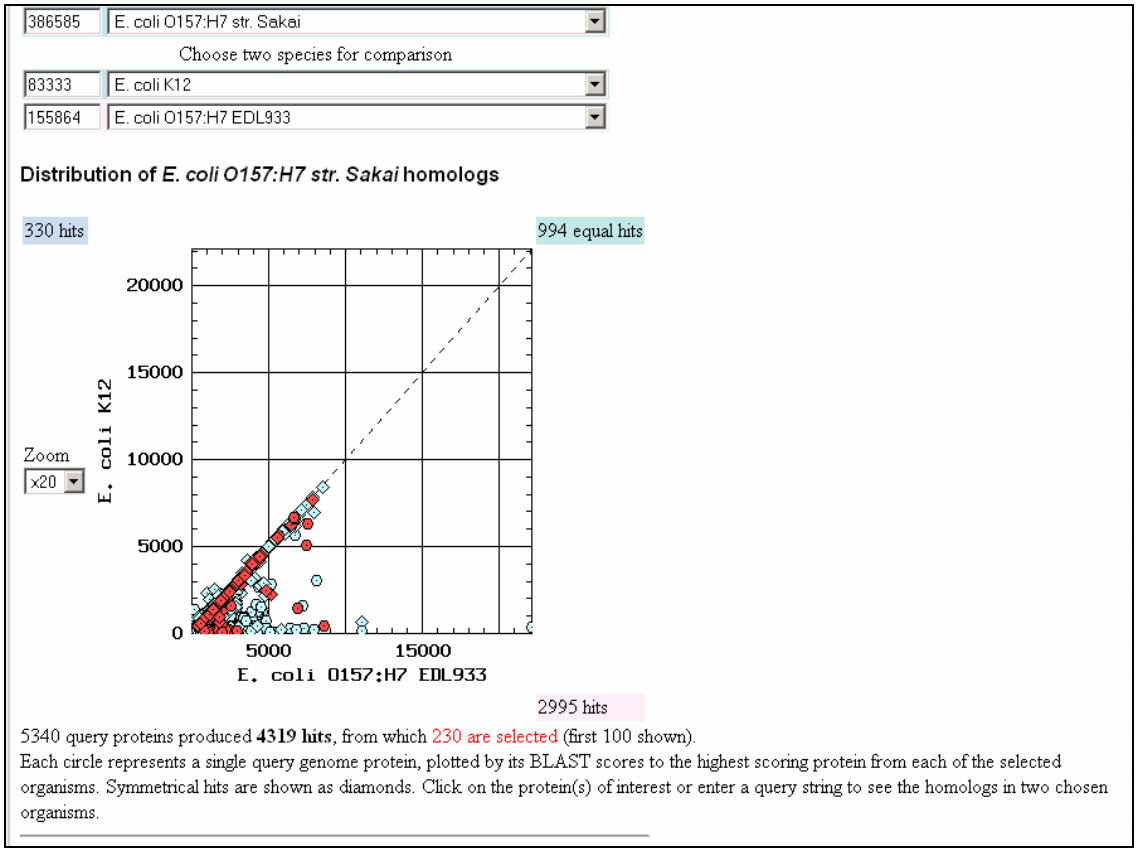


5340 query proteins produced **4319 hits**, from which 161 are selected (first 100 shown).
Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from each of the selected organisms. Symmetrical hits are shown as diamonds. Click on the protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Translation, ribosomal structure and biogenesis | compare

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


Zoom
x20
E. coli K12
15000
10000
5000
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
EDL933
2995 hits
h 161 are selected (first 100 shown).
, plotted by its BLAST scores to the highest scoring protein from each of the selected
Click on the protein(s) of interest or enter a query string to see the homologs in two chosen
clear
Translation, ribosomal structure and biogenesis
compare
Blast2Seq Accession Blink
Definition


386585
E. coli O157:H7 str. Sakai
Choose two species for comparison
83333
E. coli K12
155864
E. coli O157:H7 EDL933
Distribution of E. coli O157:H7 str. Sakai homologs
330 hits
994 equal hits
Zoom
x20
E. coli K12
20000
15000
10000
5000
0
5000
15000
E. coli O157:H7 EDL933
2995 hits
5340 query proteins produced 4319 hits, from which 230 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.

Protein homologs in Complete Microbial / Eukaryotic genomes

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

Select your query genome

386585 | E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 | E. coli K12

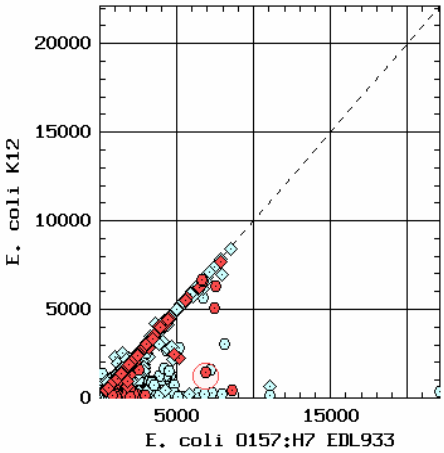
155864 | E. coli O157:H7 EDL933

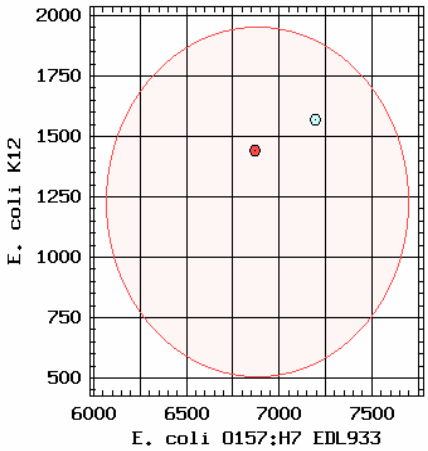
Distribution of *E. coli O157:H7 str. Sakai* homologs

330 hits

994 equal hits

Zoom x20





2995 hits

5340 query proteins produced **4319 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

Cell envelope biogenesis, outer membrane | compare

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

All: 1

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

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

Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

NCBI Entrez Genome

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

Search Genome Go Clear

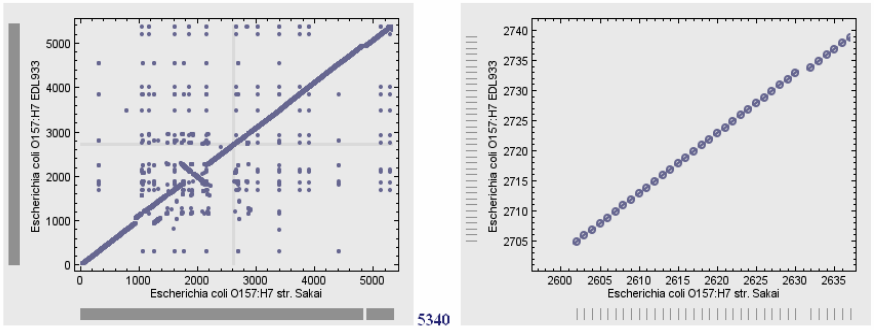
Pairwise genome comparison of protein homologs (symmetrical best hits)

Query organism: Escherichia coli O157:H7 str. Sakai

Select two organisms to compare

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

5423 proteins total



Zoom 2x

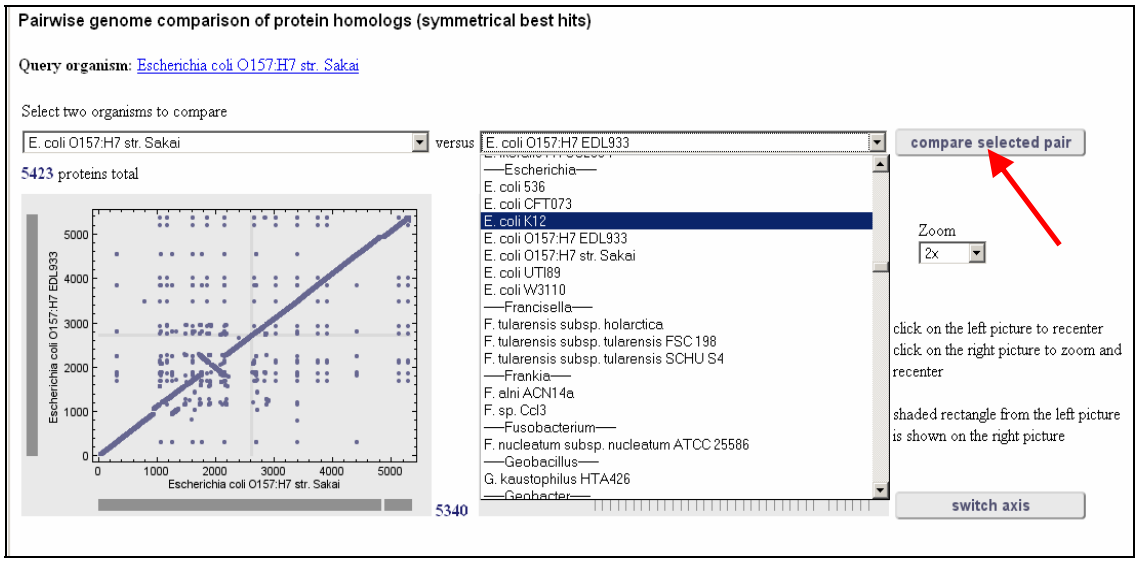
click on the left picture to recenter
click on the right picture to zoom and recenter

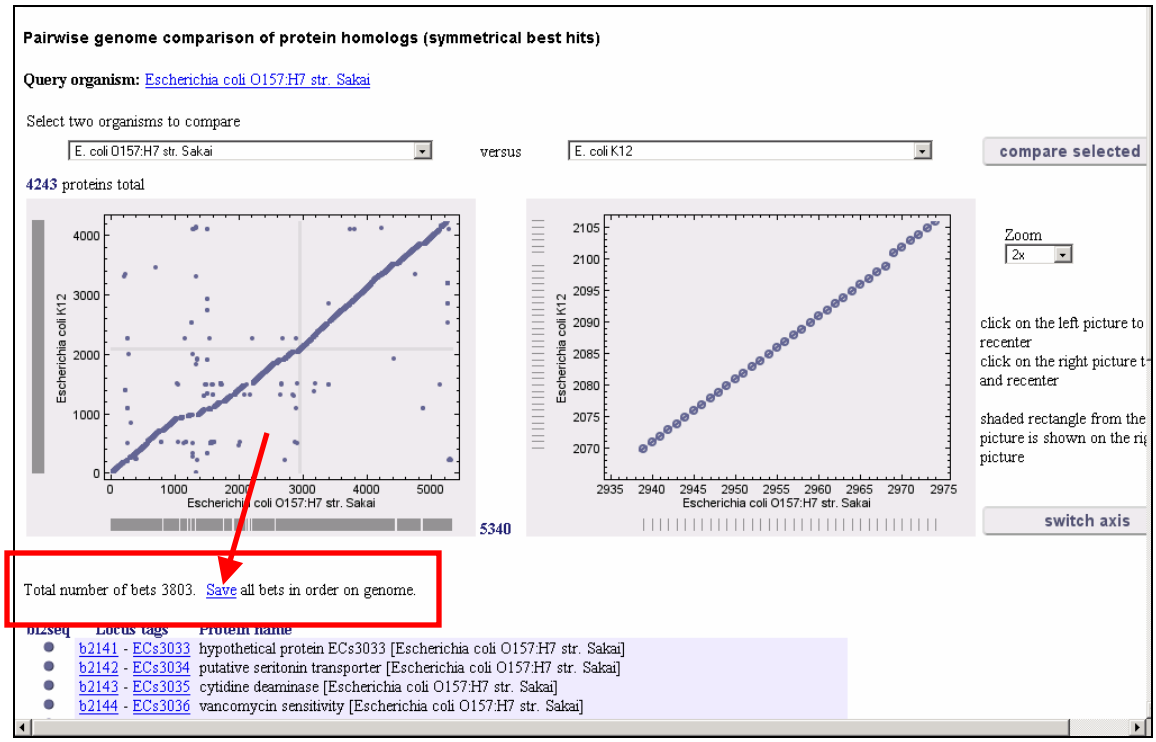
shaded rectangle from the left picture is shown on the right picture

switch axis

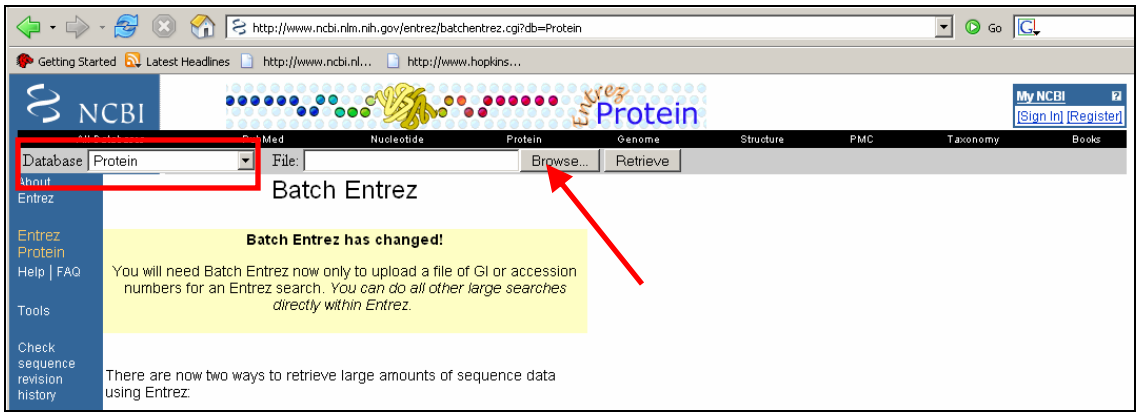
Total number of bets 5341. Save all bets in order on genome.

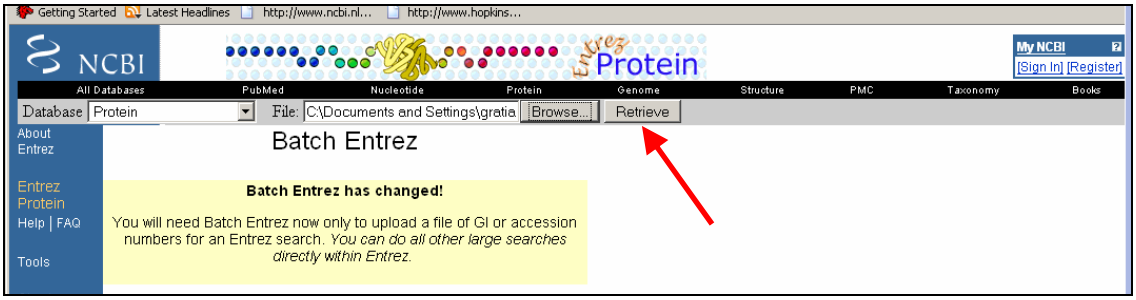
bl2seq	Locus tags	Protein name
•	Z3024 - ECs2675	hypothetical protein ECs2675 [Escherichia coli O157:H7 str. Sakai]
•	Z3027 - ECs2676	flagellar basal body protein [Escherichia coli O157:H7 str. Sakai]

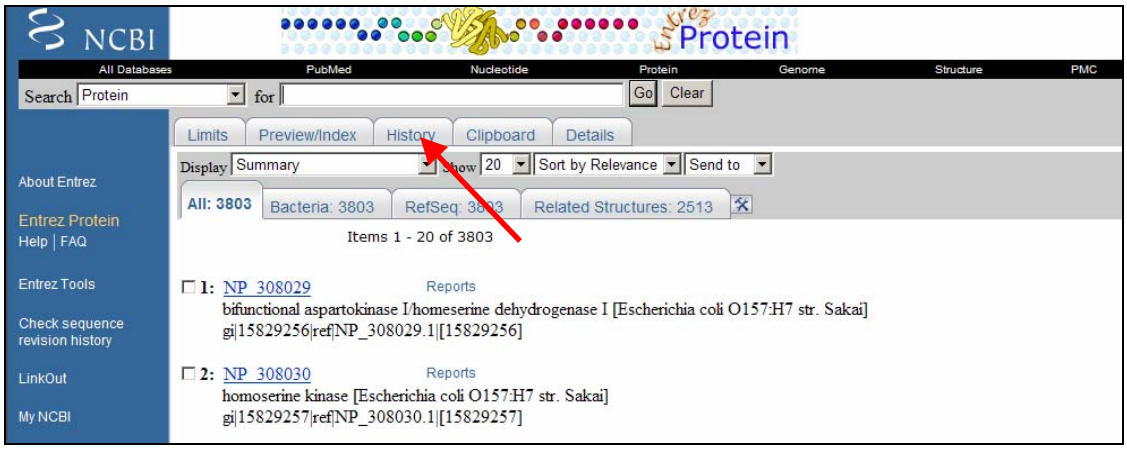

Pairwise genome comparison of protein homologs (symmetrical best hits)
Query organism: Escherichia coli O157:H7 str. Sakai
Select two organisms to compare
E. coli O157:H7 str. Sakai
versus
E. coli O157:H7 EDL933
compare selected pair
5423 proteins total
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7 str. Sakai
5340
—Escherichia—
E. coli 536
E. coli CFT073
E. coli K12
E. coli O157:H7 EDL933
E. coli O157:H7 str. Sakai
E. coli UTI89
E. coli W3110
—Francisella—
F. tularensis subsp. holarctica
F. tularensis subsp. tularensis FSC 198
F. tularensis subsp. tularensis SCHU S4
—Frankia—
F. alni ACN14a
F. sp. Ccl3
—Fusobacterium—
F. nucleatum subsp. nucleatum ATCC 25586
—Geobacillus—
G. kaustophilus HTA426
Zoom
2x
click on the left picture to recenter
click on the right picture to zoom and recenter
shaded rectangle from the left picture is shown on the right picture
switch axis

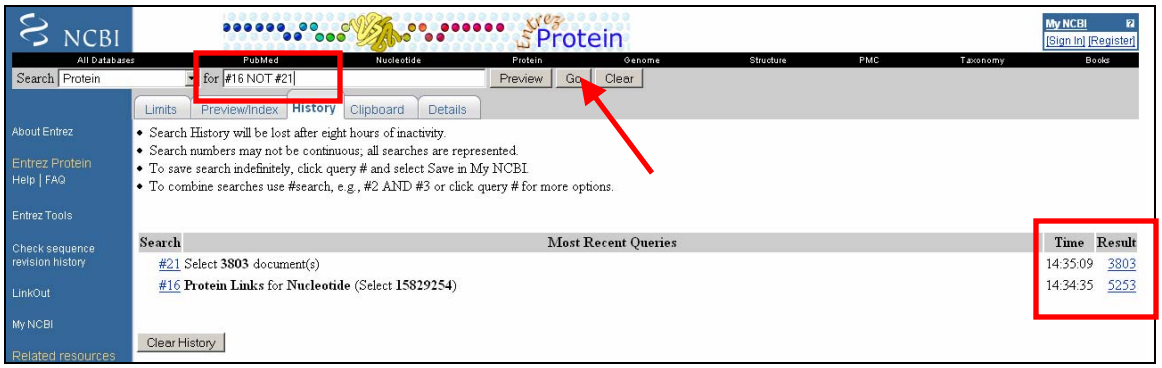

Pairwise genome comparison of protein homologs (symmetrical best hits)
Query organism: Escherichia coli O157:H7 str. Sakai
Select two organisms to compare
E. coli O157:H7 str. Sakai
versus
E. coli K12
compare selected
4243 proteins total
Escherichia coli K12
Escherichia coli O157:H7 str. Sakai
5340
Zoom
2x
click on the left picture to recenter
shaded rectangle from the
picture is shown on the ri
picture
switch axis
Total number of bets 3803. Save all bets in order on genome.
b2141 - ECs3033 hypothetical protein ECs3033 [Escherichia coli O157:H7 str. Sakai]
b2142 - ECs3034 putative serotonin transporter [Escherichia coli O157:H7 str. Sakai]
b2143 - ECs3035 cytidine deaminase [Escherichia coli O157:H7 str. Sakai]
b2144 - ECs3036 vancomycin sensitivity [Escherichia coli O157:H7 str. Sakai]

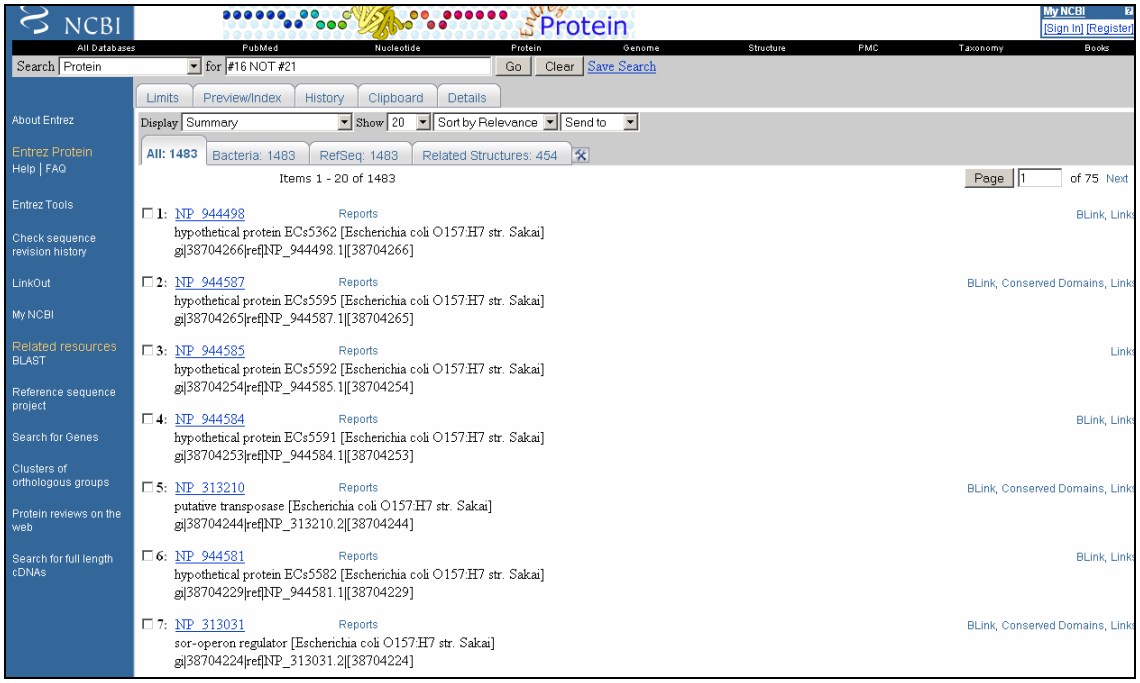
b0002	gi|16127996	ECs0002	gi|15829256 bifunctional aspartokinase I/homeserine dehydrogena
b0003	gi|16127997	ECs0003	gi|15829257 homoserine kinase [Escherichia coli O157:H7 str. Sa
b0004	gi|16127998	ECs0004	gi|15829258 threonine synthase [Escherichia coli O157:H7 str. S
b0005	gi|16127999	ECs0005	gi|15829259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

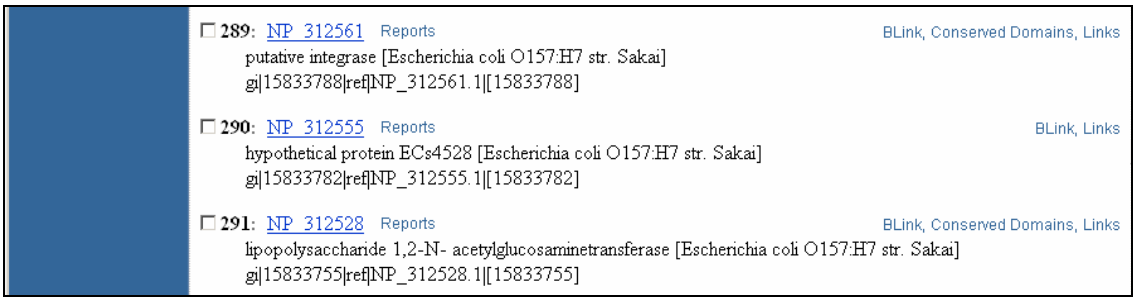







NCBI
Protein
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Protein for #16 NOT #21 Preview Go Clear
Limits Preview/Index History Clipboard Details
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
Search History will be lost after eight hours of inactivity.
Search numbers may not be continuous; all searches are represented.
To save search indefinitely, click query # and select Save in My NCBI.
To combine searches use #search, e.g., #2 AND #3 or click query # for more options.
Search
Most Recent Queries
Time
Result
#21 Select 3803 document(s)
14:35:09
3803
#16 Protein Links for Nucleotide (Select 15829254)
14:34:35
5253
Clear History


NCBI
Protein
My NCBI
[Sign In] [Register]
All Databases PubMed Nucleotide Protein Genome Structure PMC Taxonomy Books
Search Protein for #16 NOT #21 Go Clear Save Search
Limits Preview/Index History Clipboard Details
Display Summary Show 20 Sort by Relevance Send to
All: 1483 Bacteria: 1483 RefSeq: 1483 Related Structures: 454
Items 1 - 20 of 1483
Page 1 of 75 Next
About Entrez
Entrez Protein
Help | FAQ
Entrez Tools
Check sequence revision history
LinkOut
My NCBI
Related resources
BLAST
Reference sequence project
Search for Genes
Clusters of orthologous groups
Protein reviews on the web
Search for full length cDNAs
1: NP_944498 Reports BLink, Links
hypothetical protein ECs5362 [Escherichia coli O157:H7 str. Sakai]
gi|38704266|ref|NP_944498.1|[38704266]
2: NP_944587 Reports BLink, Conserved Domains, Links
hypothetical protein ECs5595 [Escherichia coli O157:H7 str. Sakai]
gi|38704265|ref|NP_944587.1|[38704265]
3: NP_944585 Reports Links
hypothetical protein ECs5592 [Escherichia coli O157:H7 str. Sakai]
gi|38704254|ref|NP_944585.1|[38704254]
4: NP_944584 Reports BLink, Links
hypothetical protein ECs5591 [Escherichia coli O157:H7 str. Sakai]
gi|38704253|ref|NP_944584.1|[38704253]
5: NP_313210 Reports BLink, Conserved Domains, Links
putative transposase [Escherichia coli O157:H7 str. Sakai]
gi|38704244|ref|NP_313210.2|[38704244]
6: NP_944581 Reports BLink, Links
hypothetical protein ECs5582 [Escherichia coli O157:H7 str. Sakai]
gi|38704229|ref|NP_944581.1|[38704229]
7: NP_313031 Reports BLink, Conserved Domains, Links
sor-operon regulator [Escherichia coli O157:H7 str. Sakai]
gi|38704224|ref|NP_313031.2|[38704224]


289: NP_312561 Reports BLink, Conserved Domains, Links
putative integrase [Escherichia coli O157:H7 str. Sakai]
gi|15833788|ref|NP_312561.1|[15833788]
290: NP_312555 Reports BLink, Links
hypothetical protein ECs4528 [Escherichia coli O157:H7 str. Sakai]
gi|15833782|ref|NP_312555.1|[15833782]
291: NP_312528 Reports BLink, Conserved Domains, Links
lipopolysaccharide 1,2-N- acetylglucosaminetransferase [Escherichia coli O157:H7 str. Sakai]
gi|15833755|ref|NP_312528.1|[15833755]

Display Overview Show 20 Send to

All: 1

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

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

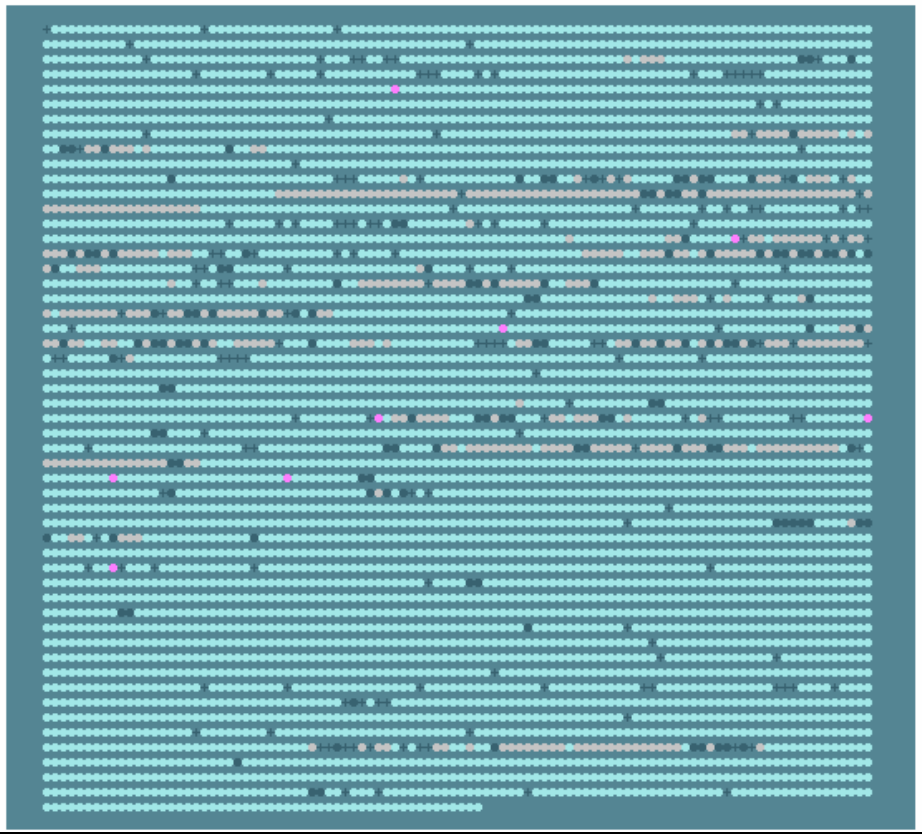
Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

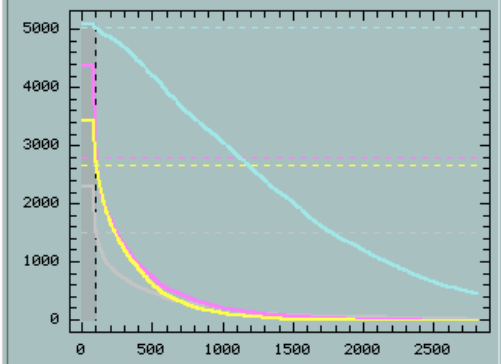
BLAST PubMed Nucleotide Protein

- [Escherichia (5087)] - [Enterobacteriaceae (4750)] - [Enterobacteriales (4750)] - [Gammaproteobact

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

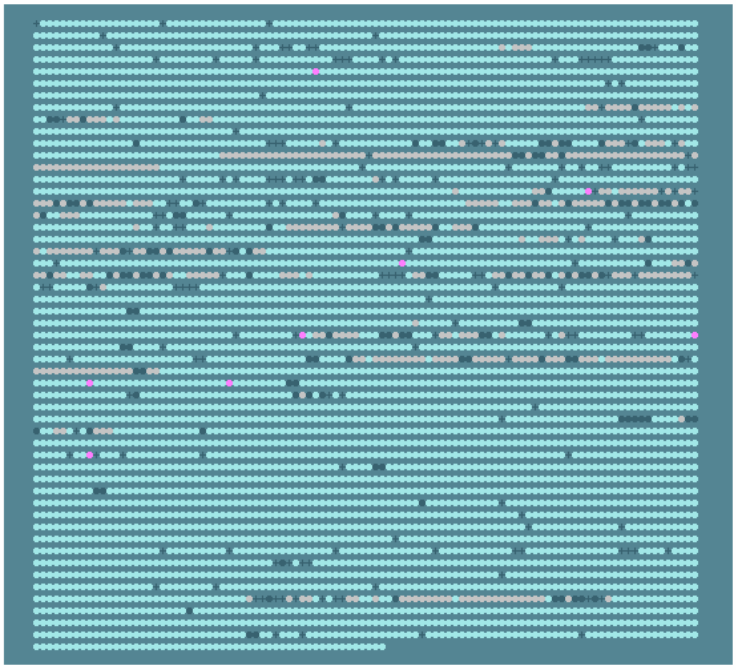


Distribution of database hits



Name		Above cut-off	Best
Virus	V	1487	379
Eukaryota	E	2763	9
Eubacteria	B	5041	4552
Archaea	A	2652	0
Eubacteria + Archaea	BA	5041	4552
Archaea + Eukaryota	AE	3271	9
Eukaryota + Eubacteria	EB	5045	4561
Eukaryota + Eubacteria + Archaea	EBA	5045	4561

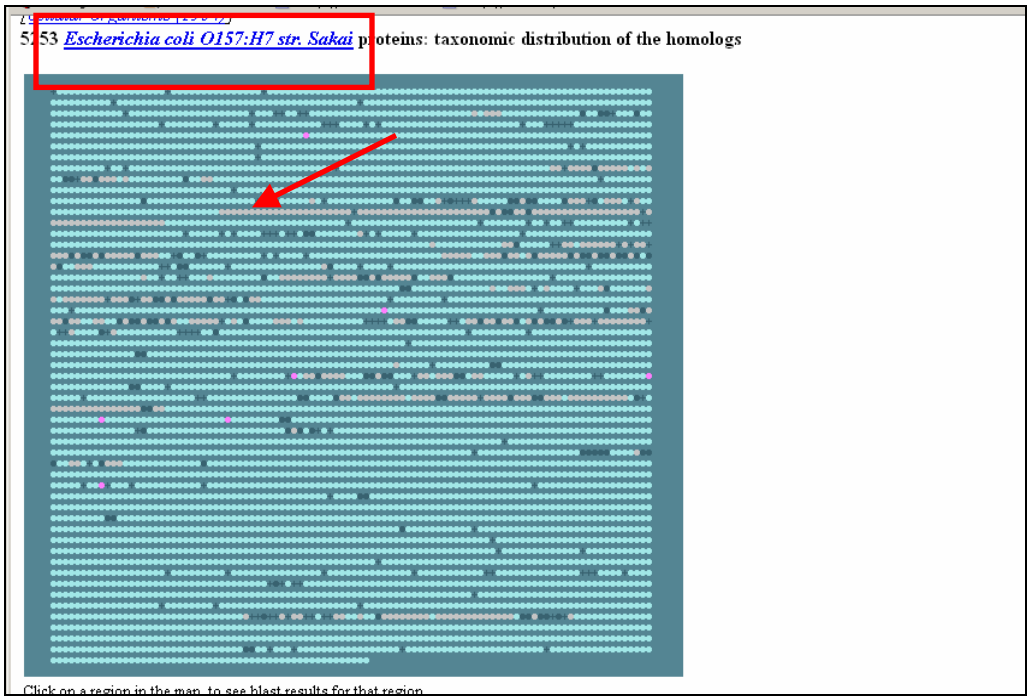
- [*Escherichia (5087)*] - [*Enterobacteriaceae (4750)*] - [*Enterobacteriales (4750)*] - [*Gammaproteobacteria (4648)*] - 5253 *Escherichia coli O157:H7 str. Sakai* **proteins: taxonomic distribution of the homologs**

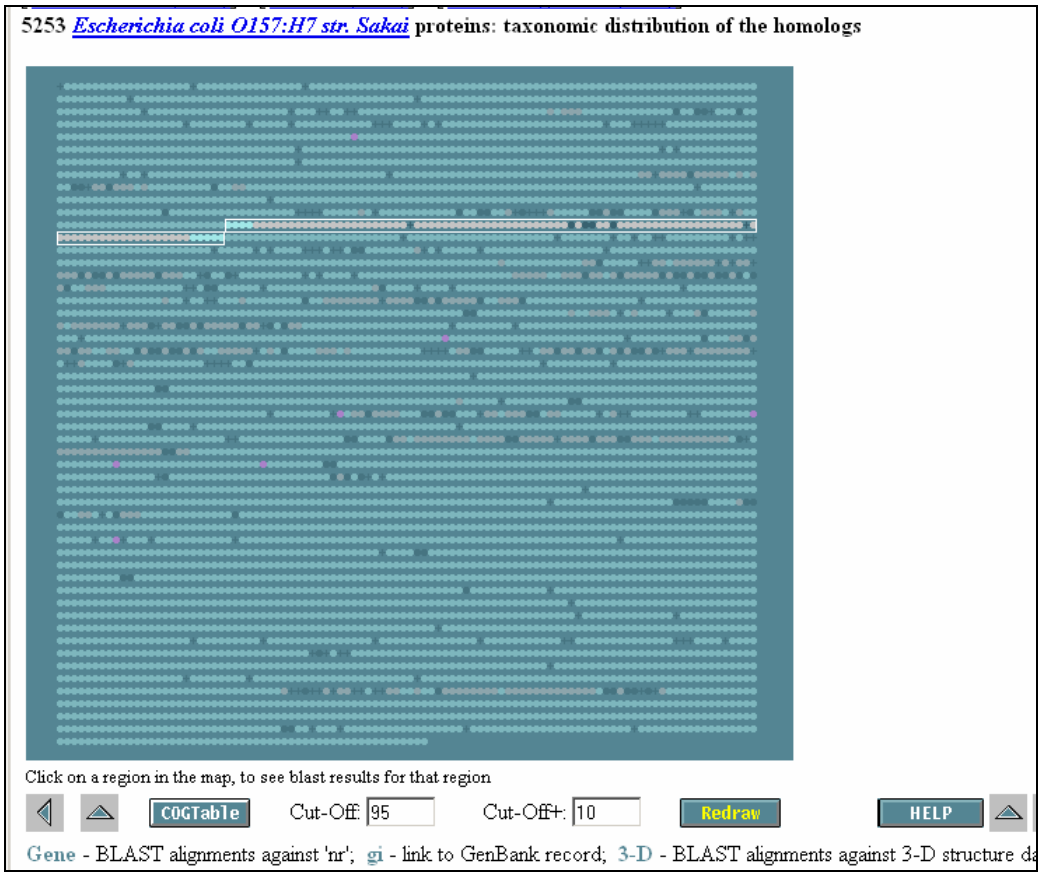


Click on a region in the map, to see blast results for that region

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

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


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


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

Gene - BLAST alignments against nr , gi - link to GenBank record, 3-D - BLAST alignments against 3-D structure database

	Virus		Eukaryota		Eubacteria		Archaea			
100	91(84) V		24(0) E		80(9) B		13(0) A			
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
15830404 B	-		11465394	(343)	74311552	(1163)	91772718	(201)	-	response transcriptional regulator for tor
15830405 B	-		86279341	(95)	82544681	(2061)	52549486	(106)	-	trimethylamine N-oxide reductase cytochrom
15830406 B	-		70998767	(165)	110805012	(4477)	11498802	(333)	-	trimethylamine N-oxide reductase subunit
15830407 B	-		-		82776283	(989)	15789983	(98)	-	chaperone protein TorD
15830408 B	-		-		30062536	(517)	-		-	hypothetical protein ECs1154
15830409 B	55819138	(149)	124418217	(330)	82544678	(1603)	110604763	(426)	-	curved DNA-binding protein
15830412 B	-		108880530	(121)	74311559	(2138)	-		-	periplasmic glucose-1-phosphatase
15830413 B	-		-		24112416	(393)	-		-	hypothetical protein ECs1159
15830414 V	9632467	(2271)	-		16764254	(1475)	110622801	(117)	-	putative integrase
15830415 V	9632468	(504)	-		16764255	(261)	-		-	putative excisionase
15830416 V	9632469	(549)	-		-		-		-	hypothetical protein ECs1162
15830417 V	9632470	(606)	-		56414547	(429)	-		-	hypothetical protein ECs1163
15830418 V	9632471	(1101)	-		24111740	(173)	-		-	hypothetical protein ECs1164
15830419 V	9632472	(504)	-		-		-		-	hypothetical protein ECs1165
15830420 V	9632473	(375)	-		74312280	(341)	-		-	hypothetical protein ECs1166
15830421 V	9632474	(516)	-		-		-		-	hypothetical protein ECs1167
15830422 V	9633404	(845)	-		-		-		-	hypothetical protein ECs1168
15830423 V	9632475	(1327)	20330752	(111)	16760392	(311)	-		-	hypothetical protein ECs1169

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

COG0582 assigned by Cognitor (5 best hits)

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

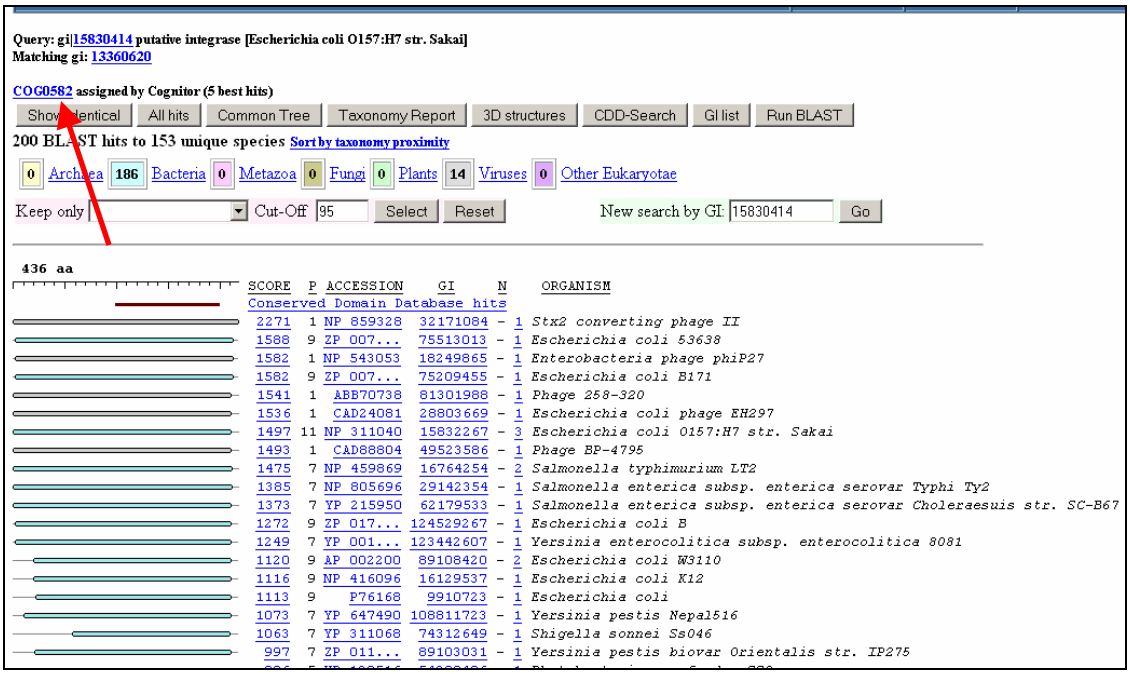
200 BLAST hits to 15[illegible] unique species **Sort by taxonomy proximity**

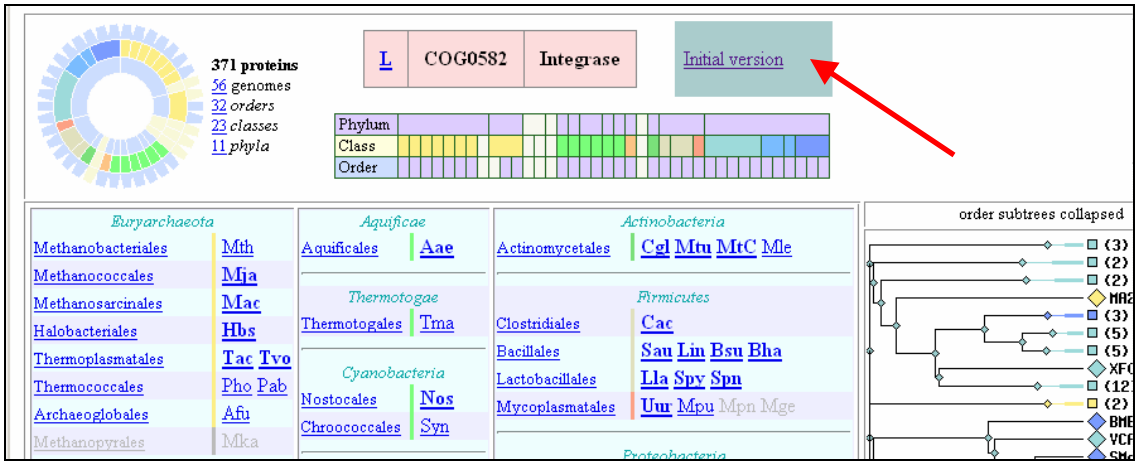
0 Archaea | 186 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 14 Viruses | 0 Other Eukaryotae

Keep only | Cut-Off 95 | Select | Reset | New search by GI: 15830414 | Go

436 aa

SCORE	P	ACCESSION	GI	PROTEIN DESCRIPTION
Conserved Domain Database hits				
2271	1	NP 859328	32171084	hypothetical protein Stx2IIp083 [Stx2 converting bacteriophage II]
1588	9	ZP 007...	75513013	COG0582: Integrase [Escherichia coli 53638]
1582	1	NP 543053	18249865	putative integrase [Bacteriophage P27]
1582	9	ZP 007...	75209455	COG0582: Integrase [Escherichia coli B171]
1541	1	ABB70738	81301988	Int protein [Phage 258-320]
1536	1	CAD24081	28803669	putative integrase [Escherichia coli phage EH297]
1497	11	NP 311040	15832267	putative integrase [Escherichia coli O157:H7 str. Sakai]
1493	1	CAD88804	49523586	putative integrase [Phage BP-4795]
1475	7	NP 459869	16764254	putative integrase [Salmonella typhimurium LT2]
1389	7	NP 459980	16764365	integrase [Salmonella typhimurium LT2]
1385	7	NP 805696	29142354	integrase [Salmonella enterica subsp. enterica serovar Typhi Ty2]
1373	7	YP 215950	62179533	Gifsy-2 prophage integrase [Salmonella enterica subsp. enterica serov
1272	9	ZP 017...	124529267	phage integrase family protein [Escherichia coli B]
1249	7	YP 001...	123442607	putative phage integrase [Yersinia enterocolitica subsp. enterocolitic
1120	9	AP 002200	89108420	predicted defective integrase [Escherichia coli W3110]
1116	9	NP 416096	16129537	Qin prophage; predicted defective integrase [Escherichia coli K12]
1113	9	P76168	9910723	Putative lambdoid prophage Qin defective integrase
1073	7	YP 647490	108811723	hypothetical protein YPN_1560 [Yersinia pestis Nepal516]
1063	7	YP 311068	74312649	putative integrase [Shigella sonnei Ss046]


Query: gi|15830414 putative integrase [Escherichia coli O157:H7 str. Sakai]
Matching gi: 13360620
COG0582 assigned by Cognitor (5 best hits)
Show identical
All hits
Common Tree
Taxonomy Report
3D structures
CDD-Search
GI list
Run BLAST
200 BLAST hits to 153 unique species
Sort by taxonomy proximity
0 Archaea
186 Bacteria
0 Metazoa
0 Fungi
0 Plants
14 Viruses
0 Other Eukaryotae
Keep only
Cut-Off 95
Select
Reset
New search by GI: 15830414
Go
436 aa
SCORE P ACCESSION GI N ORGANISM
Conserved Domain Database hits
2271 1 NP_859328 32171084 - 1 Stx2 converting phage II
1588 9 ZP_007... 75513013 - 1 Escherichia coli 53638
1582 1 NP_543053 18249865 - 1 Enterobacteria phage phiP27
1582 9 ZP_007... 75209455 - 1 Escherichia coli B171
1541 1 ABB70738 81301988 - 1 Phage 258-320
1536 1 CAD24081 28803669 - 1 Escherichia coli phage EH297
1497 11 NP_311040 15832267 - 3 Escherichia coli O157:H7 str. Sakai
1493 1 CAD88804 49523586 - 1 Phage BP-4795
1475 7 NP_459869 16764254 - 2 Salmonella typhimurium LT2
1385 7 NP_805696 29142354 - 1 Salmonella enterica subsp. enterica serovar Typhi Ty2
1373 7 YP_215950 62179533 - 1 Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67
1272 9 ZP_017... 124529267 - 1 Escherichia coli B
1249 7 YP_001... 123442607 - 1 Yersinia enterocolitica subsp. enterocolitica 8081
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


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

66 genomes version

	219 proteins Tree -	L	COG0582 info	XerC	Integrase		Help Genome context

COGs containing other domains of divided proteins:

135 ---p-----drlbcefgh-n-j---w [L] COG2801 Putative transposase

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

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

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	51	-	-	-	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

71 COGs Protein/Gene name: Select Help

```
Only Escherichia coli O157
-   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 doma
-  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 putativ
-  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 p
-   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
-   7 ----------r1--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--------r1--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
-  37 -om-k---v-----e-gh-nu----w   -    [S]   COG1479 Uncharacterized ACR
```

Problem 2

From the NCBI homepage, access the "Genomic biology" site and from there select the "Microbial" Genome Resource link. The link leads to the microbial projects in the Entrez Genome Project database:
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?

On the Complete genomes page change the selection from "Spirochaetes" back to "All Organisms". Scroll down the resulting list to the Rickettsia strains. How many Rickettsia genomes are completely sequenced? Note their sizes. Select the "Rickettsia prowazekii str. Madrid E" link that leads to the Genome Project Overview for this organism.

Is the organism pathogenic in humans? How are the bacteria transmitted to humans? Click on the RefSeq accession number NC_000963 to access the organism's information in the Entrez Genome database.

From the resulting Overview table, access the entry NC_000963 in the Entrez Nucleotide database. Access the Entrez Protein database through the "Protein" 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 Genome Overview table. 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 Structural RNA table through the "Structural RNA" link. How many genes encoding arginine tRNA are annotated? Go back to the Genome Overview table.

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 Overview table.

Click on the "TaxPlot" link. TaxPlot 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? Go back to the Genome Overview table.

Select 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 the Genome Overview table.

Select the “TaxMap” link. TaxMap 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. The page will reload to display the homologs for the selected region. 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?