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 project information, summary of its biology, and links to 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 close 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 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, which 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 "Links" → "Protein". How many protein coding genes have been annotated on the genome? Go back to the Genome Overview table.
Access the "Protein Table", which contains locations and detailed information such as the strand, product name, FASTA sequence etc., by clicking on "5253" next to "Protein coding". Also, access the "Structural RNA Table" through the "141" 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 the other pathogenic strain, O157:H7 EDL933, and click "compare". You can filter the comparison by functional classification by typing general terms such as

“translation” or more specific terms such as “adhesin” in the “Query” text box. 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. Select “Sort by Accession” and read through some of the entries. Do these make sense? 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 protein 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 hits for the integrase protein. Select the “Best hits” option. Which organisms are represented in the similar proteins?

Return to the NCBI homepage and select “Clusters of Orthologous Groups (COGs)” from the Hot Spots list on the right. Follow the link to the “Initial Version” of the COG database, then select the Phylogenetic patterns search. Access the pre-calculated comparison between Escherichia coli and Buchnera sp. APS near the bottom of the page. Note the COGs that contain proteins only from O157 strain and not from K12 or Buchnera. What are the major types of proteins in this list?

http://www.ncbi.nlm.nih.gov/

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 **Protein Clusters**
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similiarity and protein function. Click here to find out more about the Protein Clusters database.

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

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

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

Genomic Biology

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

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

July, 2007

New Genome Resource Page: A genome resource guide is now available for *Ornithorhynchus anatinus* (platypus)

Map Viewer - genome annotation updates:

Species	Build	Map Viewer Release
Caenorhabditis elegtans (nematode)	WS170	July 18, 2007
Equus caballus (domestic horse)	1.1	July 11, 2007
Ornithorhynchus anatinus (platypus)	1.1	July 11, 2007
Mus musculus (mouse)	37.1	July 5, 2007
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
Dictyostelium discoideum	2.1	November 2, 2006
Strongylocentrotus purpuratus (sea urchin)	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

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 - Publicat

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

? 635 Complete Microbial Genomes selected: [A] - 49, [B] - 586 save

GPID	Organism	King	Group	Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36		1	9	CP000828.1	NC_009925.1	10/16/07	12/09/07	Washington University (WashU) [more]	T P D L S G X
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	12/06/07	DOE Joint Genome Institute	T P D L S G X
15771	Acidobacteria bacterium Ellin345	B	Acidobacteria	5.7	58.4	1		CP000360.1	NC_008009.1	05/04/06	12/02/07	DOE Joint Genome Institute	T P D L S G X
16097	Acidothermus cellulolyticus 11B	B	Actinobacteria	2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	12/06/07	DOE Joint Genome Institute	T P D L S G X
15708	Acidovorax avenae subsp. citrulli AAC00-1	B	Betaproteobacteria	5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	12/10/07	DOE Joint Genome Institute	T P D L S G X M
15685	Acidovorax sp. JS42	B	Betaproteobacteria	4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	12/03/07	DOE Joint Genome Institute [more]	T P D L S G X M
17477	Acinetobacter baumannii ATCC 17978	B	Gammaproteobacteria	4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	12/10/07	Yale Univ.	T P D L S G X M
12352	Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	11/29/07	Genoscope	T P C D L S G X M
18221	Actinobacillus pleuropneumoniae L20	B	Gammaproteobacteria	2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	12/12/07	National Research Council of Canada [more]	T P D L S G X M

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

964 Microbial Genomes In-Progress selected: [A] - 34, [B] - 930

save

GPID	Organism	King	Group	Conts	* Size	GC	Accession	Blast	Released	Center
16707	Acaryochloris sp. CCMEE 5410	B	Cyanobacteria	-	-	- -		-	-	J. Craig Venter Institute
19259	Acholeplasma laidlawii PG-8A	B	Firmicutes	-	-	- -		-	-	Research Institute for Physico-Chemical Medicine, Federal Agency of Public Health and Social Development of the Russian Federation
199	Acidianus brierleyi	A	Crenarchaeota	-	*1.8	- -		-	-	University of Copenhagen
1	Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	-	*4.15	- -		-	-	TIGR
28085	Acidobacterium capsulatum ATCC 51196	B	Acidobacteria	-	-	- -		-	-	J. Craig Venter Institute
19277	Aciduliprofundum boonei T469	A	Euryarchaeota	-	-	- -		-	-	J. Craig Venter Institute
13000	Acinetobacter baumannii	B	Gammaproteobacteria	-	-	- -		-	-	Genoscope
13001	Acinetobacter baumannii	B	Gammaproteobacteria	-	-	- -		-	-	Genoscope
21111	Acinetobacter baumannii AB0057	B	Gammaproteobacteria	1	-	38.9	NZ_ABJM00000000	BLAST	01/25/08	Case Western Reserve University
17827	Acinetobacter baumannii ACICU	B	Gammaproteobacteria	-	-	- -		-	-	CNR - National Research Council

NCBI ENTREZ Genome Project

connection information discovery

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

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: -- 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

Map; P - ProtTable; C - COG Table; D - 3-D neighbors; L - BLAST; S - CDD search; G - GenePlot; X - TaxPlot; M - gMap; F - FTP; R - Publicat

* size is estimated, o based on existing sequences

635 Compl ed: [A] - 49, [B] - 586

save

GPID		Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
12997	Acaryochlori	Cyanobacteria	8.36		1	9	CP000828.1	NC_009925.1	10/16/07	12/09/07	Washington University (WashU) [more]	T P D L S G X
15753	Acidiphilium	haproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	12/06/07	DOE Joint Genome Institute	T P D L S G X
15771	Acidobacteri	Acidobacteria	5.7	58.4	1		CP000360.1	NC_008009.1	05/04/06	12/02/07	DOE Joint Genome Institute	T P D L S G X
16097	Acidothermu	Actinobacteria	2.4	66.9	1		CP000481.1	NC_008578.1	11/09/06	12/06/07	DOE Joint Genome Institute	T P D L S G X
15708	Acidovorax a AAC00-1	taproteobacteria	5.4	68.5	1		CP000512.1	NC_008752.1	01/04/07	12/10/07	DOE Joint Genome Institute	T P D L S G X M
15685	Acidovorax s	taproteobacteria	4.54	66.1	1	2	CP000539.1	NC_008782.1	01/04/07	12/03/07	DOE Joint Genome Institute [more]	T P D L S G X M
17477	Acinetobacte 17978	mmaproteobacteria	4.02	38.9	1	2	CP000521.1	NC_009085.1	03/01/07	12/10/07	Yale Univ.	T P D L S G X M
12352	Acinetobacte	mmaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	11/29/07	Genoscope	T P C D L S G X M
18221	Actinobacillu L20	mmaproteobacteria	2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	12/12/07	National Research Council of Canada [more]	T P D L S G X M

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

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 - Pub

132 Complete Microbial Genomes selected: [A] - 0, [B] - 132 clear filter

GPID	Organism	King	Group	Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
13006	Alkaliphilus metalliredigens QYMF	B	Firmicutes	4.9	36.8	1		CP000724.1	NC_009633.1	06/28/07	12/06/07	DOE Joint Genome Institute	T P D L S G
16083	Alkaliphilus oremlandii OhILAs	B	Firmicutes	3.1	36.2	1		CP000853.1	NC_009922.1	10/11/07	12/06/07	DOE Joint Genome Institute	T P D L S G
13478	Aster yellows witches'-broom phytoplasma AYWB	B	Firmicutes	0.73	26.8	1	4	CP000061.1	NC_007716.1	01/11/06	12/08/07	Ohio State University [more]	T P C D L S G
13403	Bacillus amyloliquefaciens FZB42	B	Firmicutes	3.9		1		CP000560.1	NC_009725.1	08/01/07	12/09/07	Competence Network [more]	T P D L S G
10784	Bacillus anthracis str. 'Ames Ancestor'	B	Firmicutes	5.5	35.2	1	2	AE017334.2	NC_007530.2	05/20/04	12/03/07	TIGR	T P C D L S G
309	Bacillus anthracis str. Ames	B	Firmicutes	5.23	35.4	1		AE016879.1	NC_003997.3	05/02/03	12/04/07	TIGR	T P C D L S G
10878	Bacillus anthracis str. Sterne	B	Firmicutes	5.23	35.4	1		AE017225.1	NC_005945.1	06/24/04	12/03/07	DOE Joint Genome Institute	T P C D L S G
74	Bacillus cereus ATCC 10987	B	Firmicutes	5.43	35.5	1	1	AE017194.1	NC_003909.8	02/13/04	12/03/07	TIGR	T P C D L S G
384	Bacillus cereus ATCC 14579	B	Firmicutes	5.42	35.3	1	1	AE016877.1	NC_004722.1	05/02/03	12/03/07	Institut National de la Recherche Agronomique (INRA)	T P C D L S G
12468	Bacillus cereus E33L	B	Firmicutes	5.85	35.1	1	5	CP000001.1	NC_006274.1	09/16/04	12/05/07	DOE Joint Genome Institute	T P C D L S G

NCBI ENTREZ Genome Project

connection information discover

PubMed Nucleotide Protein Genome Structure PopSet Taxonomy OMIM

Search Genome Project Go Clear

Organism info | Complete genomes | Genomes in progress

organism group: Firmicutes

-- 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

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 - Publicatio

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

? 635 Complete Microbial Genomes selected: [A] - 49, [B] - 586 save

GPID	Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Modified	Center	Tools
12997	Acaryochloris marina MBIC11017	B	Cyanobacteria	8.36		1	9	CP000828.1	NC_009925.1	10/16/07	12/09/07	Washington University (WashU) [more]	T P D L S G X F
15753	Acidiphilium cryptum JF-5	B	Alphaproteobacteria	3.97	67.1	1	8	CP000697.1	NC_009484.1	05/11/07	12/06/07	DOE Joint Genome Institute	T P D L S G X F
15771	Acidobacteria bacterium Ellin345	B	Acidobacteria	5.7	58.4	1		CP000360.1	NC_008009.1	05/04/06	12/02/07	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	12/06/07	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	12/10/07	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	12/03/07	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	12/10/07	Yale Univ.	T P D L S G X M F
12352	Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861.1	NC_005966.1	10/30/04	11/29/07	Genoscope	T P C D L S G X M F
18221	Actinobacillus pleuropneumoniae L20	B	Gammaproteobacteria	2.3	41.3	1		CP000569.1	NC_009053.1	02/23/07	12/12/07	National Research Council of Canada [more]	T P D L S G X M F

13355	Ehrlichia ruminantium str. Welgevonden	B	Alphaproteobacteria	1.51	27.5	1		CR925678.1	NC_006832.1	02/04/05	05/22/06	CIRAD, France	T P C D L S G X M F R
12720	Enterobacter sakazakii ATCC BAA-894	B	Gammaproteobacteria	4.56		1	2	CP000783.1	NC_009778.1	08/28/07	09/07/07	Washington University (WashU)	P L G X F
17461	Enterobacter sp. 638	B	Gammaproteobacteria	4.66	52.9	1	1	CP000653.1	NC_009436.1	04/18/07	08/14/07	DOE Joint Genome Institute [more]	P L S G X F
70	Enterococcus faecalis V583	B	Firmicutes	3.36	37.4	1	3	AE016830.1	NC_004668.1	03/29/03	12/02/05	TIGR	T P C D L S G X M F R
350	Erwinia carotovora subsp. atroseptica CRI1043	B	Gammaproteobacteria	5.06	51.0	1		BX950851.1	NC_004547.2	07/21/04	12/02/05	Welcome Trust Sanger Institute	T P C D L S G X M F R
8480	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
6235	Escherichia coli 536	B	Gammaproteobacteria	4.9	50.5	1		CP000247.1	NC_008253.1	07/13/06	08/16/07	University of Goettingen	T P D L S G X M F R
6718	Escherichia coli APEC O1	B	Gammaproteobacteria	5.51	50.5	1	2	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
8960	Escherichia coli E24377A	B	Gammaproteobacteria	5.27	50.6	1	6	CP000800.1	NC_009801.1	09/11/07	09/14/07	TIGR	P L G X F
8959	Escherichia coli HS	B	Gammaproteobacteria	4.6	50.8	1		CP000802.1	NC_009800.1	09/10/07	09/14/07	TIGR	P L G X F
225	Escherichia coli K12	B	Gammaproteobacteria	4.6	50.8	1		U00096.2	NC_000913.2	09/05/97	05/01/07	University of Wisconsin-Madison, USA	T P C D L S G X M F R
259	Escherichia coli O157:H7 EDL933	B	Gammaproteobacteria	5.59	50.3	1	1	AE005174.2	NC_002655.2	02/24/01	09/14/07	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
6259	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
6351	Escherichia coli W3110	B	Gammaproteobacteria	*4.65	50.8	1		AP009048.1	AC_000091.1	01/23/06	08/16/07	Nara Institute of Science and Technology	T P D L S G X M F R
6719	Fervidobacterium nodosum Rt17-B1	B	Thermotogae	1.9	34.9	1		CP000771.1	NC_009718.1	07/30/07	09/05/07	DOE Joint Genome Institute	P L S G X F
16082	Flavobacterium johnsoniae UW101	B	Bacteroidetes/Chlorobi	6.1	34.1	1		CP000685.1	NC_009441.1	05/01/07	08/14/07	DOE Joint Genome Institute	P L S G X F
19979	Flavobacterium psychrophilum JIP02/86	B	Bacteroidetes/Chlorobi	2.9	32.5	1		AM398681.1	NC_009613.1	06/25/07	06/29/07	Institut National de la Recherche Agronomique (INRA)	P L S G X F
16421	Francisella tularensis subsp. holarctica	B	Gammaproteobacteria	1.9	32.2	1		AM233362.1	NC_007880.1	03/02/06	06/28/07	Lawrence Livermore National Laboratory [more]	T P D 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*

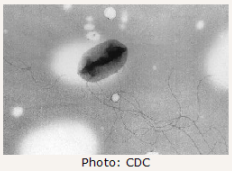


Photo: CDC

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	5.5	50.5%	5253	141	✔	✔	✔	✔
Plasmid pO157	NC_0021[illegible]	AB011549	1	0.092721	47.6%	85	--				
Plasmid pOSAK1	NC_0021[illegible]	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 O157:H7

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 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

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

```
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.
```

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

NCBI Entrez Protein

All Databases | PubMed | Nucleotide | Protein | Genome

Search Protein for Go Clear

Limits Preview/Index History Clipboard Details

Display Summary Show 20 Sort by Relevance Send to

All: 5253 | Bacteria: 5253 | RefSeq: 5253 | Related Structures: 3168

Items 1 - 20 of 5253

1: NP_308064 Reports
DNA-binding transcriptional activator CaiF [Escherichia coli O157:H7 str. Sakai]
gi|162139821|ref|NP_308064.2|[162139821]

2: NP_308071 Reports
putative electron transfer flavoprotein FixA [Escherichia coli O157:H7 str. Sakai]
gi|162139820|ref|NP_308071.2|[162139820]

3: NP_308107 Reports
leucine transcriptional activator [Escherichia coli O157:H7 str. Sakai]
gi|162139819|ref|NP_308107.2|[162139819]

4: NP_308128 Reports
SecA regulator SecM [Escherichia coli O157:H7 str. Sakai]
gi|162139818|ref|NP_308128.2|[162139818]

5: NP_308150 Reports
hypothetical protein ECs0123 [Escherichia coli O157:H7 str. Sakai]
gi|162139817|ref|NP_308150.2|[162139817]

6: NP_308174 Reports
poly(A) polymerase I [Escherichia coli O157:H7 str. Sakai]
gi|162139816|ref|NP_308174.2|[162139816]

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	◆◆◆

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

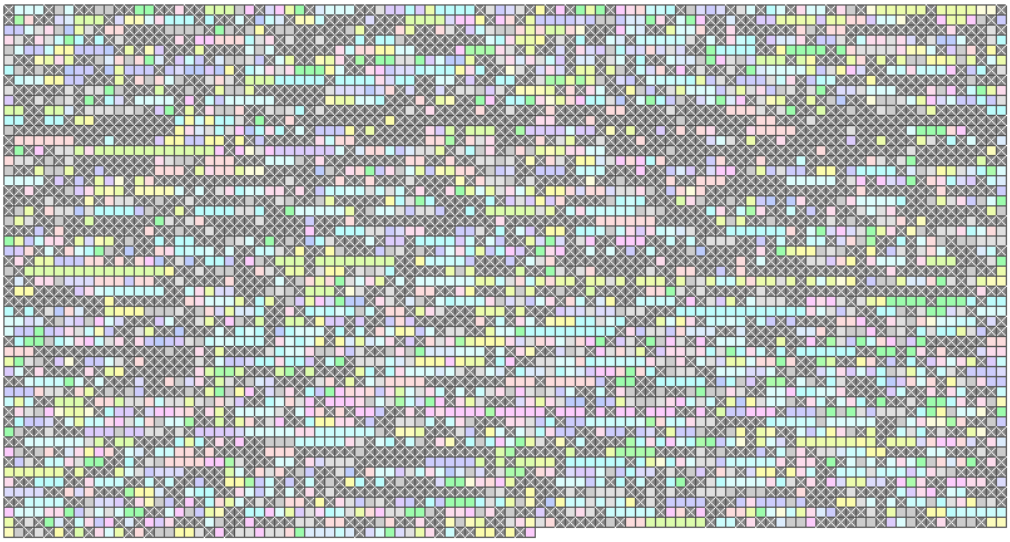
Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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

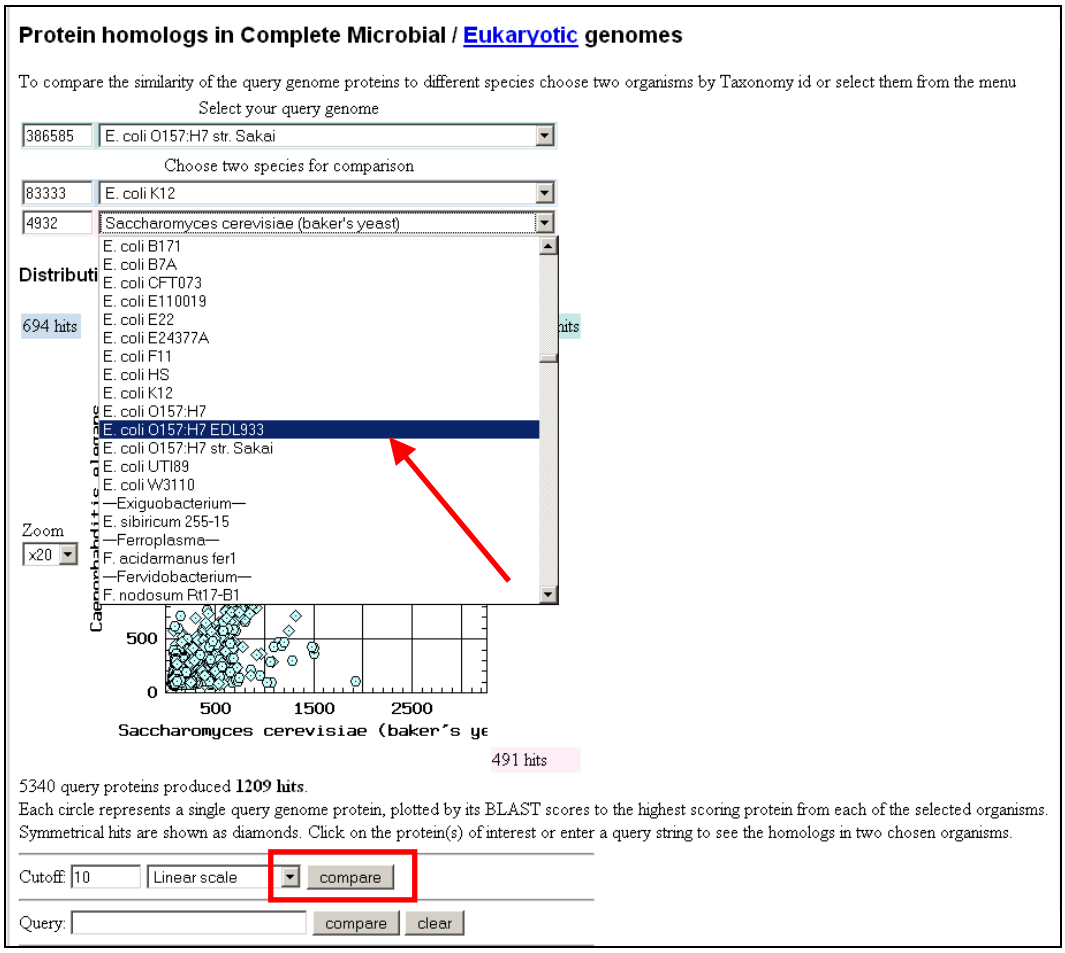
All: 1

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
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

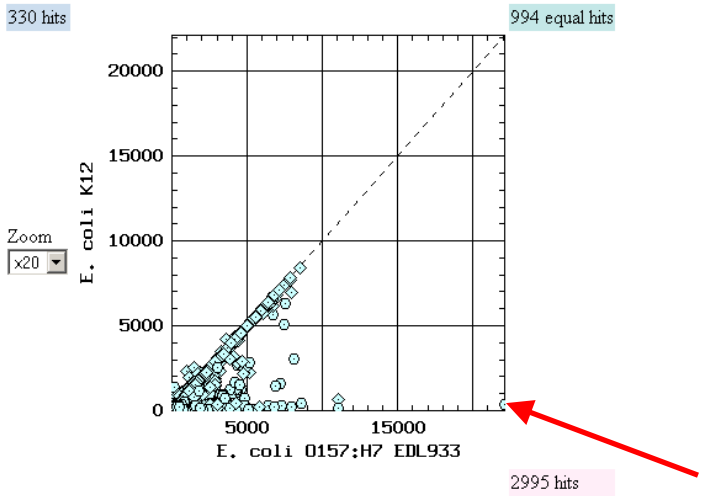
386585 | E. coli O157:H7 str. Sakai

Choose two species for comparison

83333 | E. coli K12

155864 | E. coli O157:H7 EDL933

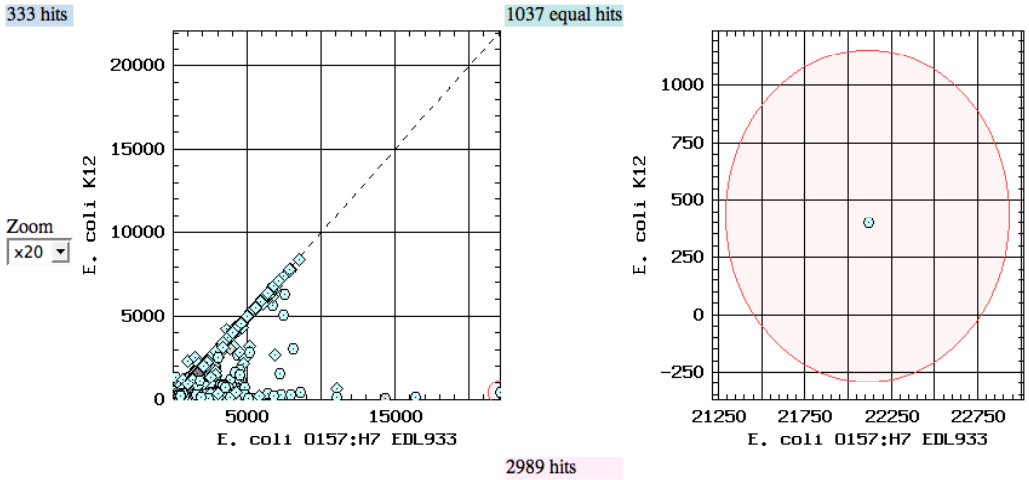
Distribution of *E. coli O157:H7 str. Sakai* homologs



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.

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



5340 query proteins produced **4359 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

Blast2Seq	Accession	Blink	Definition
	NP_308569	=>	**hypothetical protein ECs0542 [Escherichia coli O157:H7 str. Sakai]**
401	NP_416485	=>	adhesin [Escherichia coli K12]
22114	NP_286230	=>	putative RTX family exoprotein [Escherichia coli O157:H7 EDL933]

5340 query proteins produced **4359 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: translation compare 2

1

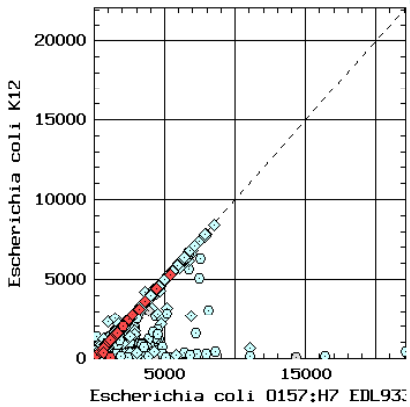
Blast2Seq	Accession	Blink	Definition
	NP_308569	=>	**hypothetical protein ECs0542 [Escherichia coli O157:H7 str. Sakai]**
401	NP_416485	=>	adhesin [Escherichia coli K12]
22114	NP_286230	=>	putative RTX family exoprotein [Escherichia coli O157:H7 EDL933]

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

333 hits

1037 equal hits

Zoom x20



2989 hits

5340 query proteins produced **4359 hits**, from which 51 are 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: translation compare clear

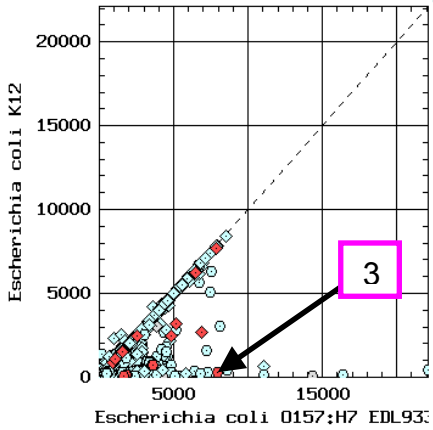
Blast2Seq	Accession	Blink	Definition
	NP_308083	=>	**dimethyladenosine transferase [Escherichia coli O157:H7 str. Sakai]**
1401	NP_414593	=>	S-adenosylmethionine-6-N',N'-adenosyl (rRNA) dimethyltransferase [Escherichia coli K12]
1407	NP_285748	=>	dimethyladenosine transferase [Escherichia coli O157:H7 EDL933]

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

332 hits

1037 equal hits

Zoom x20



2986 hits

5340 query proteins produced **4355 hits**, from which 14 are selected .
Each circle represents a single query genome protein, plotted by its BLAST scores to the highest scoring protein from eac
protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Cutoff: 10 Linear scale compare

Query: adhesin compare 2

1

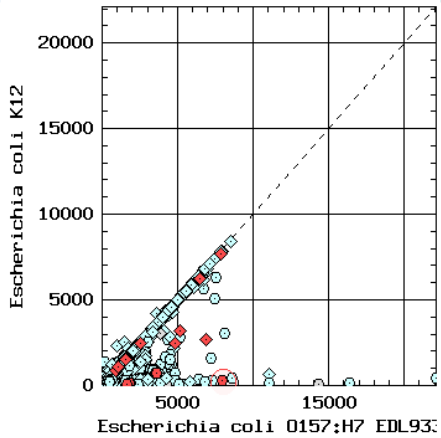
Blast2Seq	Accession	Blink	Definition
	NP_308377	=>	**putative adhesin [Escherichia coli O157:H7 str. Sakai]**
136	NP_415720	=>	predicted adhesin [Escherichia coli K12]
1799	NP_286038	=>	hypothetical protein Z0390 [Escherichia coli O157:H7 EDL933]
	NP_308389	=>	**AidA-I adhesin-like protein [Escherichia coli O157:H7 str. Sakai]**
2660	Q9JMS3	=>	Uncharacterized protein yuaQ
6869	NP_286049	=>	putative beta-barrel outer membrane protein [Escherichia coli O157:H7 EDL933]

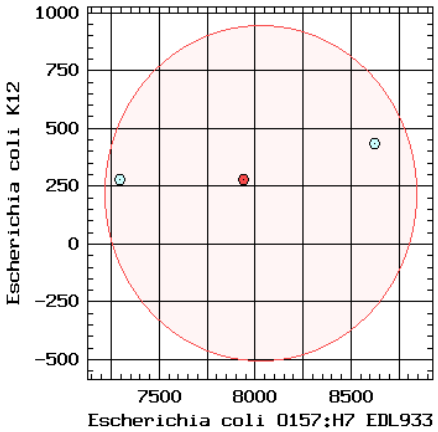
Distribution of *Escherichia coli O157:H7 str. Sakai* homologs

332 hits

1037 equal hits

Zoom x20





2986 hits

5340 query proteins produced **4355 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. S
protein(s) of interest or enter a query string to see the homologs in two chosen organisms.

Cutoff: 10 Linear scale compare

Query: adhesin compare clear

Blast2Seq	Accession	Blink	Definition
	NP_312507	=>	**hypothetical protein ECs4480 [Escherichia coli O157:H7 str. Sakai]**
274	Q9JMS5	=>	Uncharacterized protein yuaO precursor
7942	NP_290185	=>	putative adhesin [Escherichia coli O157:H7 EDL933]

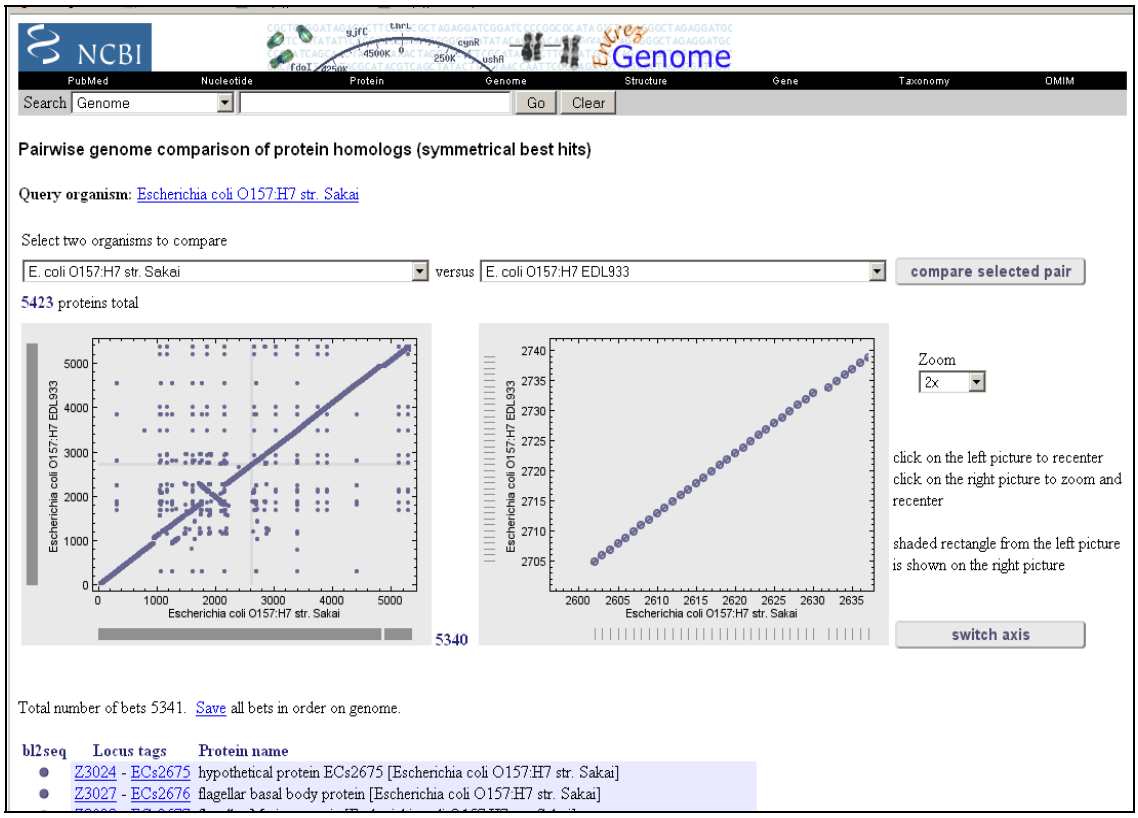
All: 1

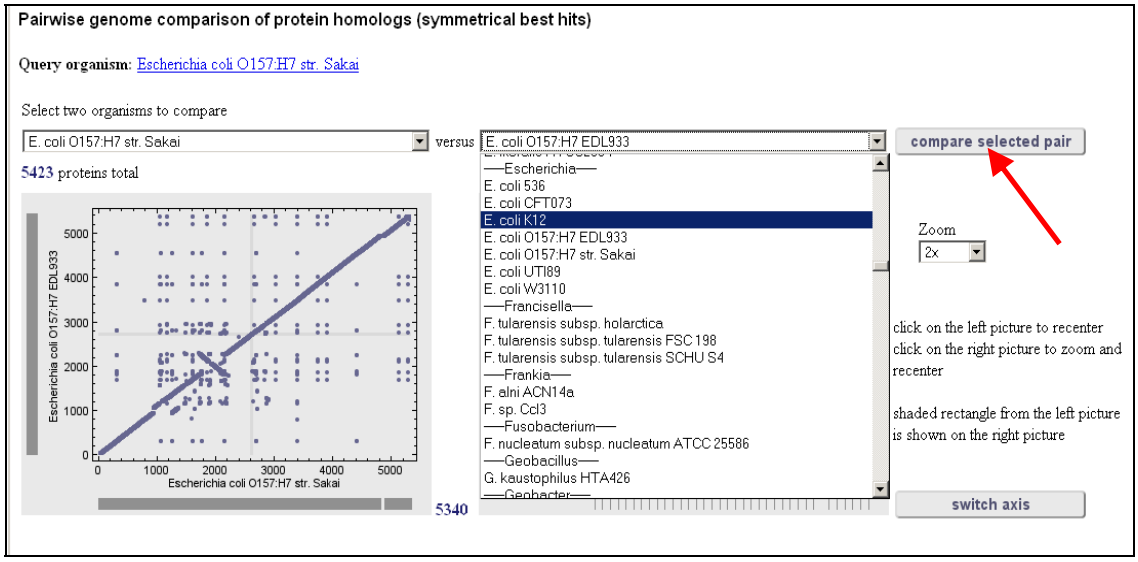
Genome* > *Bacteria* > *Escherichia coli O157:H7 str. Sakai, complete genome

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
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7 str. Sakai
5340
Zoom
2x
click on the left picture to recenter
click on the right picture to zoom and recenter
shaded rectangle from the left picture is shown on the right picture
switch axis
Total number of bets 5341. Save all bets in order on genome.
bl2seq
Locus tags
Protein name
Z3024 - ECs2675 hypothetical protein ECs2675 [Escherichia coli O157:H7 str. Sakai]
Z3027 - ECs2676 flagellar basal body protein [Escherichia coli O157:H7 str. Sakai]


Pairwise genome comparison of protein homologs (symmetrical best hits)
Query organism: Escherichia coli O157:H7 str. Sakai
Select two organisms to compare
E. coli O157:H7 str. Sakai
versus
E. coli O157:H7 EDL933
compare selected pair
5423 proteins total
—Escherichia—
E. coli 536
E. coli CFT073
E. coli K12
E. coli O157:H7 EDL933
E. coli O157:H7 str. Sakai
E. coli UTI89
E. coli W3110
—Francisella—
F. tularensis subsp. holarctica
F. tularensis subsp. tularensis FSC 198
F. tularensis subsp. tularensis SCHU S4
—Frankia—
F. alni ACN14a
F. sp. Ccl3
—Fusobacterium—
F. nucleatum subsp. nucleatum ATCC 25586
—Geobacillus—
G. kaustophilus HTA426
Escherichia coli O157:H7 EDL933
Escherichia coli O157:H7 str. Sakai
5340
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

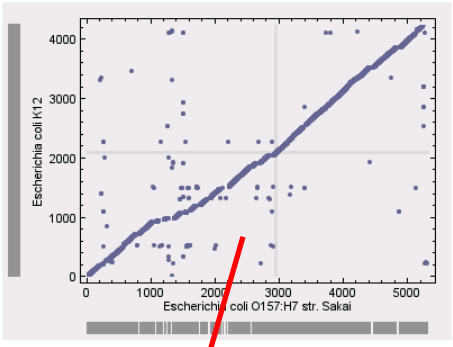
Pairwise genome comparison of protein homologs (symmetrical best hits)

Query organism: Escherichia coli O157:H7 str. Sakai

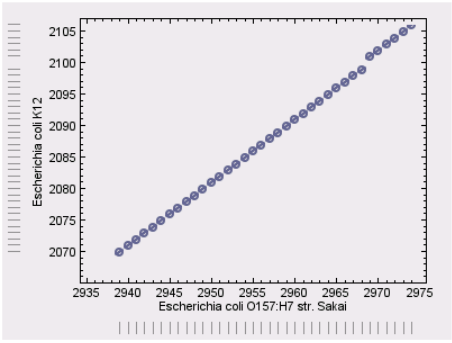
Select two organisms to compare

E. coli O157:H7 str. Sakai versus E. coli K12 compare selected

4243 proteins total



5340



Zoom 2x

click on the left picture to recenter
click on the right picture t
and 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.

bl2seq	Locus tags	Protein name
●	b2141 - ECs3033	hypothetical protein ECs3033 [Escherichia coli O157:H7 str. Sakai]
●	b2142 - ECs3034	putative seritonin 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. Sal
b0004	gi|16127998	ECs0004	gi|15829258 threonine synthase [Escherichia coli O157:H7 str. S
b0005	gi|16127999	ECs0005	gi|15829259 hypothetical protein ECs0005 [Escherichia coli O157
b0006	gi|16128000	ECs0006	gi|15829260 hypothetical protein ECs0006 [Escherichia coli O157
b0007	gi|16128001	ECs0007	gi|15829261 putative inner membrane transport protein [Escherich
b0008	gi|16128002	ECs0008	gi|15829262 transaldolase [Escherichia coli O157:H7 str. Sakai]
b0009	gi|16128003	ECs0009	gi|15829263 molybdenum cofactor biosynthesis protein [Escherich
b0010	gi|16128004	ECs0010	gi|15829264 hypothetical protein ECs0010 [Escherichia coli O157
b0011	gi|16128005	ECs0012	gi|15829265 putative oxidoreductase [Escherichia coli O157:H7 s
b0012	gi|90111078	ECs0011	gi|15829266 positive regulator for sigma 32 heat shock promoter
b0013	gi|16128007	ECs0013	gi|15829267 hypothetical protein ECs0013 [Escherichia coli O157
b0014	gi|16128008	ECs0014	gi|15829268 molecular chaperone DnaK [Escherichia coli O157:H7
b0015	gi|16128009	ECs0015	gi|15829269 DnaJ [Escherichia coli O157:H7 str. Sakai]
b0018	gi|16128012	ECs0016	gi|15829270 Gef [Escherichia coli O157:H7 str. Sakai]
b0019	gi|16128013	ECs0017	gi|15829271 Na+/H antiporter [Escherichia coli O157:H7 str. Sak
b0020	gi|16128014	ECs0018	gi|15829272 transcriptional activator of nhaA [Escherichia coli
b0023	gi|16128017	ECs0026	gi|15829280 30S ribosomal protein S20 [Escherichia coli O157:H7
b0024	gi|16128018	ECs0027	gi|15829281 hypothetical protein ECs0027 [Escherichia coli O157
b0025	gi|16128019	ECs0028	gi|15829282 hypothetical protein ECs0028 [Escherichia coli O157
b0026	gi|16128020	ECs0029	gi|15829283 isoleucyl-tRNA synthetase [Escherichia coli O157:H7
b0027	gi|16128021	ECs0030	gi|15829284 signal peptidase II [Escherichia coli O157:H7 str.
b0028	gi|16128022	ECs0031	gi|15829285 putative FKBX-type 16KD peptidyl-prolyl cis-trans i
b0029	gi|16128023	ECs0032	gi|15829286 4-hydroxy-3-methylbut-2-enyl diphosphate reductase
b0030	gi|16128024	ECs0033	gi|15829287 hypothetical protein ECs0033 [Escherichia coli O157
b0031	gi|16128025	ECs0034	gi|15829288 dihydrodipicolinate reductase [Escherichia coli O15
b0032	gi|16128026	ECs0035	gi|15829289 carbamoyl-phosphate synthase small subunit [Escheri
b0033	gi|16128027	ECs0036	gi|15829290 carbamoyl-phosphate synthase large subunit [Escheri

15829257
15829258
15829259
15829260
15829261
15829262
15829263
15829264
15829265
15829266
15829267
15829268
15829269
15829270
15829271
15829272
15829280
15829281
15829282
15829283
15829284
15829285
15829286
15829287
15829288
15829289
15829290
15829291
15829292
15829293
15829294
15829295
15829296
15829297

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

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...

GenBank vs. RefSeq

Confused about the distinctions between GenBank, RefSeq, TPA and UniProt? Click here for a brief description of the databases and their differences.

New **Protein Clusters**
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similiarity and protein function. Click here to find out more about the Protein Clusters database.

PubMed Central
An archive of biomedical and life sciences journals

- Free fulltext
- Over 1,100,000 articles from over 340 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.

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

NCBI

Entrez | PubMed | Nucleotide | Protein | Genome | Structure | OMIM | PMC | Journals | Books

Advanced Entrez Tools

Web Tools:

Batch Entrez - Upload a file of GI or accession numbers to retrieve sequences.

PubMed Batch Citation Matcher - Send citation information to Entrez and retrieve PubMed IDs for linking, citation display, or other applications.

Advanced Entrez Searching - Advanced searching techniques for Web Entrez.

My NCBI - includes automatic e-mailing of search updates and filters for search results.

Programming Tools:

E-Utilities - Run Entrez queries and download data from your own scripts over the Web.

Linking to Entrez - Link to specific Entrez pages from your own web pages or applications.

Entrez Client/Server - C language library for embedding Entrez calls into your programs.

NCBI

Batch Entrez

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

Database: Protein File: Browse... Retrieve

PubMed
Protein
CoreNucleotide
GSS
EST
Structure
Genome
Books
CancerChromosomes
Conserved Domains
dbGaP
3D Domains
Gene
Genome Project
GENSAT
GEO Profiles
GEO DataSets
HomoloGene
MeSH
NCBI Web Site

Batch E

Use Batch GIs or accession numbers from the Nucleotide or Protein database d identifiers from other Entrez databases.

Tips : Sor s can be tens of thousands of lines long and Batch Entrez may not r e list. Split the list of identifiers into smaller files using a file splitting s hand at the command prompt in UNIX or LINUX systems. Put several th per file, left-formatted, and one per line. This may be necessar ge numbers of genome records. You can check the NCBI FTP site l e genome records. Also, use GIs rather than 'accessio making lists for batch entrez to fetch.

Please no heck for duplicate identifiers when reporting results from a list that yo

When reti accessions, you must select the specific component database s or GIs were saved. For Nucleotide, choose either the CoreNucl S selection from the database menu. If you have a mixed list of nucleotide accessions or UIDs, you will need to run the Batch Entrez search three times. Select the database from the pull-down menu, CoreNucleotide, EST, and GSS separately.

In all cases, be certain to select the correct database for uploaded identifiers when using Batch Entrez, to ensure the expected records. For example, if you have saved a list of protein GIs, be sure to select the Protein database.

- Create a file with a list of GI or accession numbers and save it locally
- Select the database from which the list of accessions or UIDs originated
- Use the 'Browse' button to select the filename containing the list of UIDs from your system directory
- Press the Retrieve button and you will see a list of document summaries
- Select a format in which to display the data for viewing, and/or saving
- Select 'Send to file' to save the file.

The following records can't be retrieved:

Id=gi|109552: protein: This record was removed because it was deleted from the PIR sequence database. Please contact pirmail@nbrf.georgetown.edu for further details.

The following records may be obsolete:

Id=gi|1095526: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095526: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095526: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095527: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095528: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095528: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.
Id=gi|1095530: This record was removed because it was deleted from the PRF sequence database. Please contact isoyama@prf.or.jp for further details.

Received lines: 3803
Rejected lines: 1
Passed to Entrez: 3802
Retrieve records for 3802 UID(s)

NCBI Entrez Protein

My NCBI [Sign In] [Register]

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

Search Protein for Go Clear

Limits | Preview/Index | History | Clipboard | Details

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

Display Summary Show 20 Sort by Relevance Send to

All: 3775 | Bacteria: 3771 | RefSeq: 3771 | Related Structures: 2619

Items 1 - 20 of 3775 Page 1 of 189 Next

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

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

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

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

5: NP_308033 Reports BLink, Conserved Domains, Links
hypothetical protein ECs0006 [Escherichia coli O157:H7 str. Sakai]
gi|15829260|ref|NP_308033.1|[15829260]

6: NP_308034 Reports BLink, Conserved Domains, Links
putative inner membrane transport protein [Escherichia coli O157:H7 str. Sakai]
gi|15829261|ref|NP_308034.1|[15829261]

7: NP_308035 Reports BLink, Conserved Domains, Links
transaldolase [Escherichia coli O157:H7 str. Sakai]
gi|15829262|ref|NP_308035.1|[15829262]

NCBI Entrez Protein

My NCBI [Sign In] [Register]

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

Search Protein for #3 NOT #5 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
BLAST

- 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
#5	Select **3775** document(s)	16:48:01	3775
#4	Select **3802** document(s)	16:47:57	3802
#3	**Protein Links** for **CoreNucleotide** (Select **15829254**)	16:47:31	5253

Clear History

NCBI Entrez Protein

All Databases | PubMed | Nucleotide | Protein | Genome

Search Protein for (#3) NOT (#5) Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Relevance Send to

All: 1599 | Bacteria: 1599 | RefSeq: 1599 | Related Structures: 603

Items 1 - 20 of 1599

1: NP_308064 Reports
DNA-binding transcriptional activator CaiF [Escherichia coli O157:H7 str. Sakai]
gi|162139821|ref|NP_308064.2|[162139821]

2: NP_308071 Reports
putative electron transfer flavoprotein FixA [Escherichia coli O157:H7 str. Sakai]
gi|162139820|ref|NP_308071.2|[162139820]

3: NP_308107 Reports
leucine transcriptional activator [Escherichia coli O157:H7 str. Sakai]
gi|162139819|ref|NP_308107.2|[162139819]

4: NP_308128 Reports
SecA regulator SecM [Escherichia coli O157:H7 str. Sakai]
gi|162139818|ref|NP_308128.2|[162139818]

NCBI Entrez Protein

All Databases | PubMed | Nucleotide | Protein | Genome

Search Protein for (#3) NOT (#5) Go Clear Save Search

Limits | Preview/Index | History | Clipboard | Details

Display Summary Show 20 Sort by Accession Send to

All: 1599 | Bacteria: 1599 | RefSeq: 1599 | Related Structures: 603

Items 1 - 20 of 1599

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

2: NP_308046 Reports
hypothetical protein ECs0019 [Escherichia coli O157:H7 str. Sakai]
gi|15829273|ref|NP_308046.1|[15829273]

3: NP_308048 Reports
hypothetical protein ECs0021 [Escherichia coli O157:H7 str. Sakai]
gi|15829275|ref|NP_308048.1|[15829275]

4: NP_308049 Reports
putative outer membrane usher protein precursor [Escherichia coli O157:H7 str. Sakai]
gi|15829276|ref|NP_308049.1|[15829276]

5: NP_308050 Reports
putative fimbrial chaperone [Escherichia coli O157:H7 str. Sakai]
gi|15829277|ref|NP_308050.1|[15829277]

6: NP_308051 Reports
putative fimbrial protein [Escherichia coli O157:H7 str. Sakai]
gi|15829278|ref|NP_308051.1|[15829278]

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

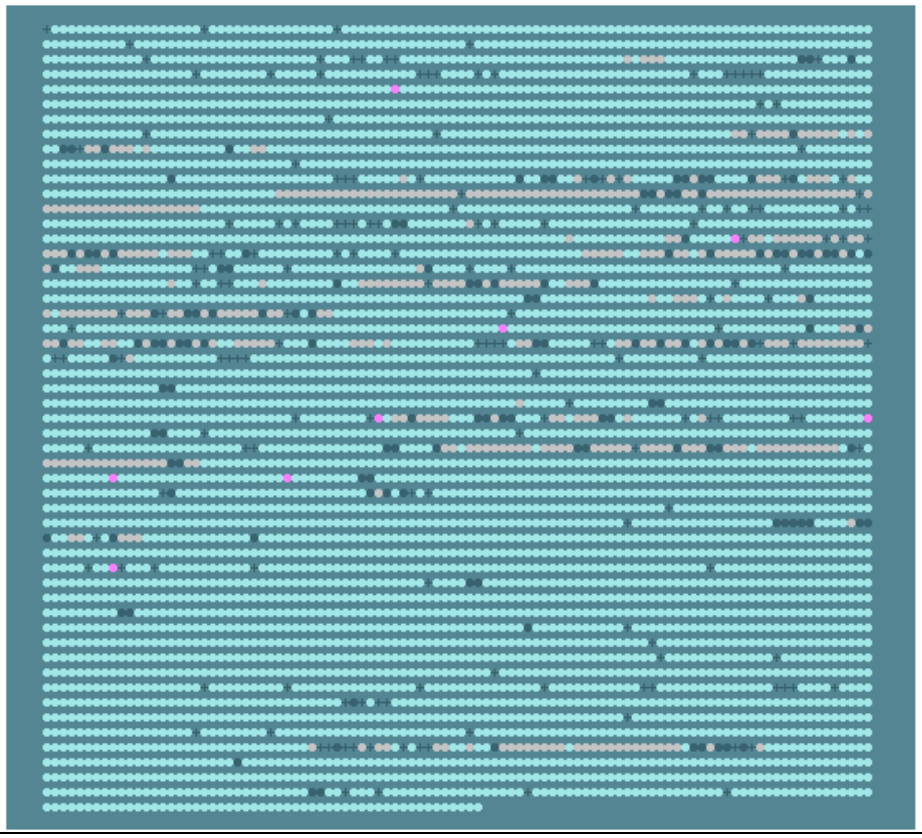
Chromosomes: *chromosome*
Plasmids: pOSAK1, pO157

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	

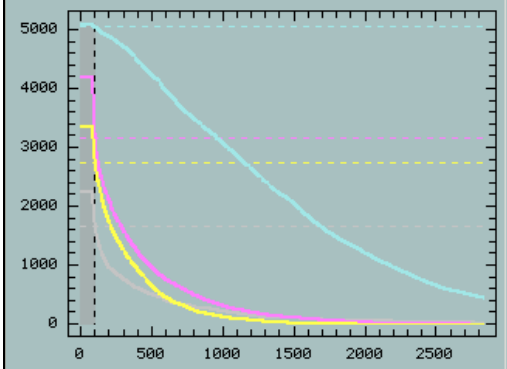
BLAST PubMed Nucleotide Protein

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

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



Distribution of database hits

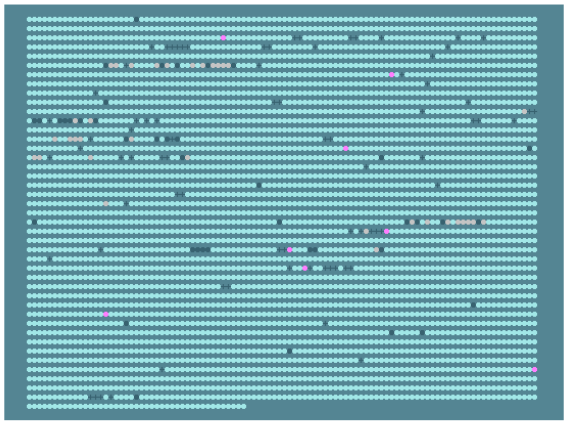


Name		Above cut-off	Best
Virus	V	1651	402
Eukaryota	E	3128	9
Eubacteria	B	5057	4535
Archaea	A	2724	0
Eubacteria + Archaea	BA	5057	4535
Archaea + Eukaryota	AE	3465	9
Eukaryota + Eubacteria	EB	5059	4544
Eukaryota + Eubacteria + Archaea	EBA	5059	4544

BLAST PubMed Nucleotide Protein Genome Structure Taxonomy

- [Escherichia (4203)] - [Enterobacteriaceae (3977)] - [Enterobacteriales (3977)] - [Gammaproteobacteria (3892)] - [Proteobacteria (3706)] - [Bacteria (3243)] - [cellular organisms (1306)]

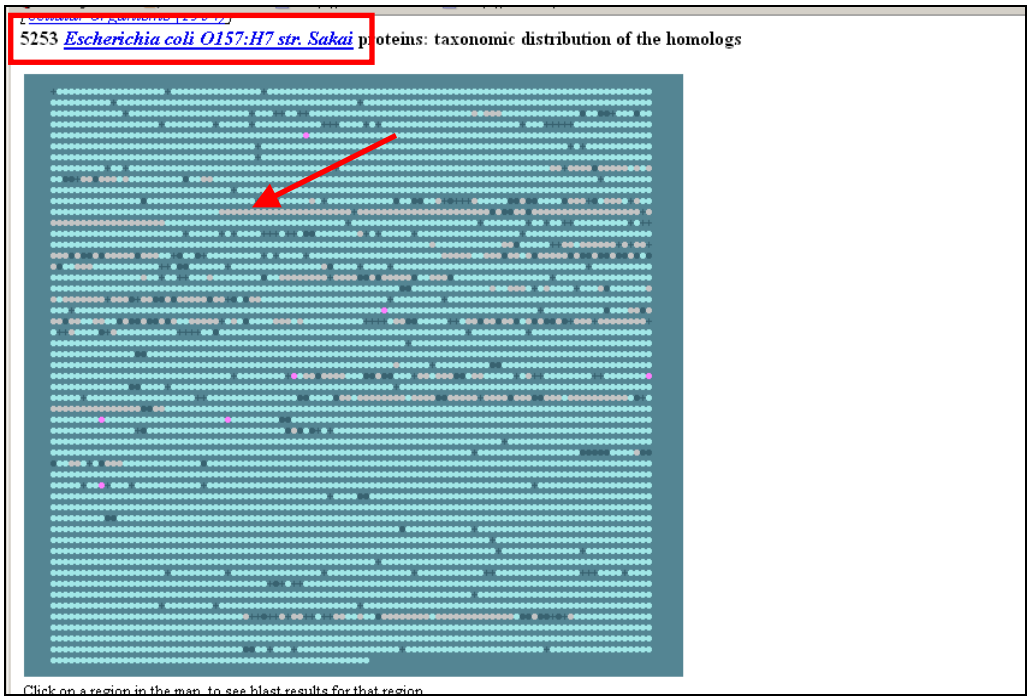
4243 *Escherichia coli K12* proteins: taxonomic distribution of the homologs

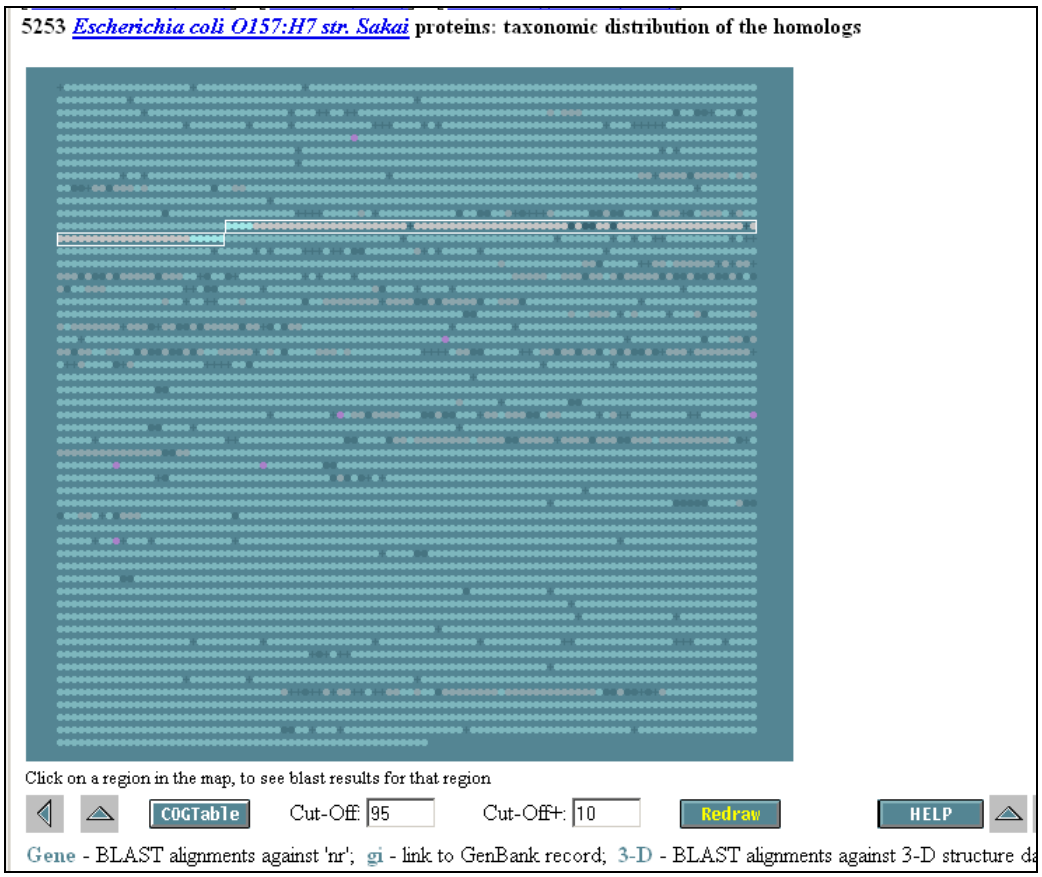


Click on a region in the map, to see blast results for that region

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

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

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

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

	Virus		Eukaryota		Eubacteria		Archaea			
100	91(84) V		24(0) E		80(9) B		16(0) A			
Gene	gi	score	gi	score	gi	score	gi	score	3-D	protein name
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
15830424 V	20065892	(381)	-		42631014	(136)	-		-	C4-type zinc finger protein (TraR family)
15830425 V	9632476	(361)	-		118740494	(120)	-		-	hypothetical protein ECs1171
15830426 V	9633407	(312)	-		-		-		-	hypothetical protein ECs1172
15830427 V	9633408	(318)	-		37526827	(115)	-		-	hypothetical protein ECs1173
15830428 V	32170937	(1206)	54637203	(129)	82776669	(1183)	-		-	exonuclease
15830429 V	9633410	(1369)	-		6759957	(1305)	-		-	recombination protein Bet
15830430 V	32170944	(727)	-		82776671	(436)	-		-	Gam
15830431 V	9633412	(446)	-		-		-		-	Kil protein
15830432 V	9626284	(297)	-		56414538	(200)	-		-	regulatory protein cIII
15830433 V	9633414	(649)	-		67158861	(141)	-		-	Ea10
15830434 V	9633416	(442)	-		-		-		-	hypothetical protein ECs1180
15830435 V	9633417	(449)	-		56414533	(110)	-		-	N
15830436 .	-		-		-		-		-	hypothetical protein ECs1182
15830437 V	19262907	(277)	-		-		-		-	hypothetical protein ECs1183
15830438 V	9633418	(856)	-		103485769	(166)	-		-	hypothetical protein ECs1184
15830439 V	23343453	(1149)	-		85059205	(577)	-		-	putative cI repressor protein
15830440 V	9633420	(373)	-		113461548	(185)	15922518	(97)	-	CRO
15830441 V	9633421	(493)	-		-		-		-	CII

NCBI

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

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

COG0582 assigned by Cognitor (4 best hits)

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

200 BLAST hits to 145 unique species Sort by taxonomy proximity

0 Archaea 170 Bacteria 0 Metazoa 0 Fungi 0 Plants 8 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_049461	9632467	integrase [Bacteriophage 933W]
1588	9	ZP_007...	75513013	COG0582: Integrase [Escherichia coli 53638]
1582	1	CAC83519	18152332	putative integrase [Bacteriophage P27]
1582	9	ZP_007...	75209455	COG0582: Integrase [Escherichia coli B171]
1541	1	ABB70737	81301986	Int protein [Phage 258-320]
1536	1	CAD24081	28803669	putative integrase [Escherichia coli phage EH297]
1497	11	BAB36436	13362482	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	AAO69545	29137984	integrase [Salmonella enterica subsp. enterica serovar Typhi Ty2]
1373	7	AAX64869	62127166	Gifsy-2 prophage integrase [Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67]
1281	9	ABV21122	157081414	site-specific recombinase, phage integrase family [Escherichia coli E24377A]
1272	9	ZP_017...	124529267	phage integrase family protein [Escherichia coli B]
1272	7	YP_001...	156933782	hypothetical protein ESA_01608 [Enterobacter sakazakii ATCC BAA-894]
1249	7	YP_001...	123442607	putative phage integrase [Yersinia enterocolitica subsp. enterocolitica 8081]
1120	9	AP_002200	89108420	predicted defective integrase [Escherichia coli W3110]
1113	9	P76168	9910723	Putative lambdoid prophage Qin defective integrase
1106	7	YP_001...	157370504	integrase family protein [Serratia proteamaculans 568]
1073	7	ABG17890	108775371	hypothetical protein YPN_1560 [Yersinia pestis Nepal516]
1063	7	YP_311068	74312649	putative integrase [Shigella sonnei Ss046]
997	7	ZP_011...	89103031	COG0582: Integrase [Yersinia pestis biovar Orientalis str. IP275]
906	5	CAG19714	46912925	hypothetical integrase for prophage CP-933V [Photobacterium profundum SS9]
887	7	ABF04318	110615651	putative integrase [Shigella flexneri 5 str. 8401]

NCBI

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

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

COG0582 **assigned by Cognitor (4 best hits)**

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

200 BLAST hits to 145 unique species Sort by taxonomy proximity

0 Archaea | 170 Bacteria | 0 Metazoa | 0 Fungi | 0 Plants | 8 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_049461	9632467	1	*Enterobacteria phage 933W*
1588	9	ZP_007...	75513013	1	*Escherichia coli 53638*
1582	1	CAC83519	18152332	1	*Enterobacteria phage phiP27*
1582	9	ZP_007...	75209455	1	*Escherichia coli B171*
1541	1	ABB70737	81301986	1	*Phage 258-320*
1536	1	CAD24081	28803669	1	*Escherichia coli phage EH297*
1497	11	BAB36436	13362482	2	*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	AAO69545	29137984	1	*Salmonella enterica subsp. enterica serovar Typhi Ty2*
1373	7	AAX64869	62127166	1	*Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67*
1281	9	ABV21122	157081414	1	*Escherichia coli E24377A*
1272	9	ZP_017...	124529267	1	*Escherichia coli B*
1272	7	YP_001...	156933782	1	*Enterobacter sakazakii ATCC BAA-894*
1249	7	YP_001...	123442607	1	*Yersinia enterocolitica subsp. enterocolitica 8081*
1120	9	AP_002200	89108420	2	*Escherichia coli W3110*
1113	9	P76168	9910723	1	*Escherichia coli K12*
1106	7	YP_001...	157370504	1	*Serratia proteamaculans 568*
1073	7	ABG17890	108775371	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	CAG19714	46912925	1	*Photobacterium profundum SS9*
887	7	ABF04318	110615651	2	*Shigella flexneri 5 str. 8401*

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,

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...

GenBank® Celebrating 25 Years

NCBI will hold a scientifc meeting to celebrate the 25th anniversary of GenBank.
April 7-8, 2008
Natcher Auditorium, NIH Campus, Bethesda MD

click here for more information

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

COGs

Phylogenetic classification of proteins encoded in complete genomes

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

66 genomes
38 *orders*
28 *classes*
14 *phyla*

Unicellular clusters FTP

Science 1997 Oct 24;278(5338):631-7,
BMC Bioinformatics 2003 Sep 11;4(1):41.

Initial version

Euryarchaeota	
Methanobacteriales	Mth
Methanococcales	Mja
Halobacteriales	Hbs
Thermoplasmatales	Tac Tvo
Thermococcales	Pho Pab
Archaeoglobales	Afu
Methanopyrales	Mka
Methanosarcinales	Mac
Crenarchaeota	
Thermoproteales	Pya
Sulfolobales	Sso
Desulfurococcales	Ape
Ascomycota	
Saccharomycetales	Sce
Schizosaccharomycetales	Spo
Microsporidia	
Apansporoblastina	Ecu
Aquificae	
Aquificales	Aae
Thermotogae	
Thermotogales	Tma
Cyanobacteria	
Nostocales	Nos
Chroococcales	Syn
Deinococcus-Thermus	
Deinococcales	Dra
Fusobacteria	
Fusobacterales	Fnu
Spirochaetes	
Spirochaetales	Tpa Bbu
Chlamydiae	
Chlamydiales	Ctr Cpn
Actinobacteria	
Actinomycetales	Cgl Mtu MtC Mle
Firmicutes	
Clostridiales	Cac
Bacillales	Sau Lin Bsu Bha
Lactobacillales	Lla Spy Spn
Mycoplasmatales	Uur Mpu Mpn Mge
Proteobacteria	
Pseudomonadales	Pae
Enterobacteriales	Eco EcZ Ecs Ype Sty Buc
Xanthomonadales	Xfa
Vibrionales	Vch
Pasteurellales	Hin Pmu
Burkholderiales	Rso
Neisseriales	Nme NmA
Campylobacterales	Hpy jHp Cje
Caulobacterales	Ccr
Rhizobiales	Atu Sme Bme Mlo
Rickettsiales	Rpr Rco

Eukaryotic Clusters FTP

Code	*Name*	Abbre
A	*Arabidopsis thaliana* (thale cress)	at
C	*Caenorhabditis elegans* (worm)	ce
D	*Drosophila melanogaster* (fruit fly)	dm
H	*Homo sapiens* (human)	hs
Y	*Saccharomyces cerevisiae* (baker yeast)	sc
P	*Schizosaccharomyces pombe* (fission yeast)	sp
E	*Encephalitozoon cuniculi* (Microsporidia)	ec
Upcoming eukaryotic genomes		
O	*Oryza sativa* (rice)	os
Q	*Anopheles gambiae* (mosquito)	ag

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	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

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 ----------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
- 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? Return to the NCBI homepage and select “Clusters of Orthologous Groups (COGs)” from the Hot Spots list on the right. Follow the link to the “Initial Version” of the COG database, then select the Phylogenetic patterns search. Obtain a list of COGs that are absent in Rickettsia. What are the major functional categories?