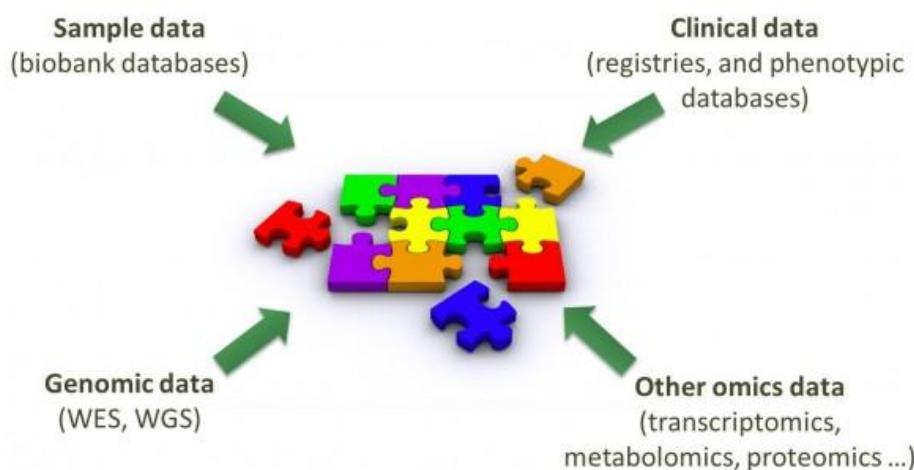


生物信息学： 组学时代的生物信息数据挖掘和理解

2021年春



有关信息

- 授课教师: 宁康, 刘欣, 陈卫华, 张礼斌,
陈鹏
 - Email: ningkang@hust.edu.cn
 - Office: 华中科技大学东十一楼504室
 - Phone: 87793041, 18627968927
- 课程网页
 - <http://www.microbioinformatics.org/Bioinformatics.html>
 - QQ群: “医学八年生物信息 2021”

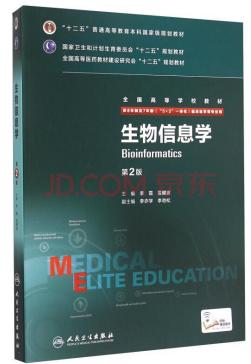


课程安排

周次	起止日期	教学内容	学时数	任课教师
9		生物信息学导论	8(讲课)	宁康
10		生物信息学基础问题和基本分析方法	8(讲课)	宁康
11		组学数据获取和蛋白组学分析(I)	8(讲课)	刘欣
12		组学数据获取和蛋白组学分析(II)	8(讲课)	刘欣
13		基因组和微生物组数据分析	8(讲课)	陈卫华
14		代谢组学数据获取和分析	8(讲课)	陈鹏
15		生物信息分析工具和数据库(I)	8(讲课)	张礼斌
16		生物信息分析工具和数据库(II)	8(讲课)	张礼斌

研究对象：
生物序列，
进化树，
生物网络，
基因表达
...

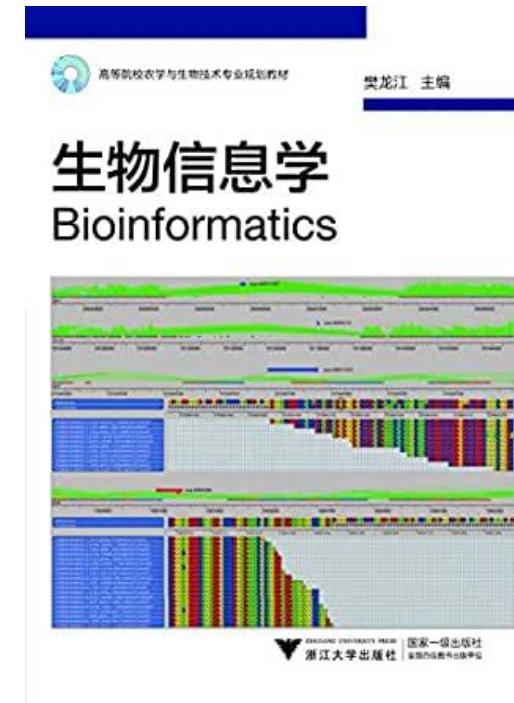
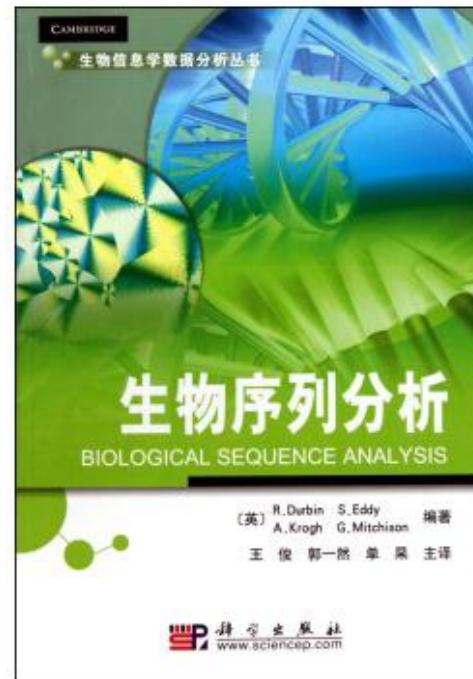
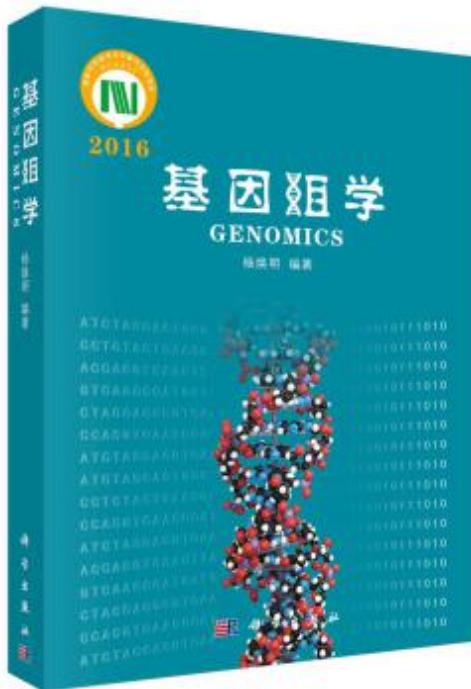
方法：
生物信息与大数据挖掘



教材及参考书目

- **教学参考书:**
 - 《生物信息学（八年制）》（第2版）. 人民卫生出版社. 2015年6月出版. 李霞等主编.
- **课外文献阅读:**
 - 《生物序列分析》（第1版）. 科学出版社. 2010年8月出版. R. Durbin等编著，王俊等主译.
 - 《生物信息学》（第1版）. 浙江大学出版社. 2017年3月出版. 樊龙江主编.
 - 《基因组学》（第1版）. 科学出版社. 2016年10月出版. 杨焕明主编.

References



Slides credits

- 生物信息学研究方法概述：北京大学生物信息中心
- 生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学
- 神经网络与深度学习：邱锡鹏@复旦大学
- Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Harvard University
- Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS
- Deep Learning in the Life Sciences: MIT

Omics databases

基因组可视化: Genome Browser, (<http://genome.ucsc.edu/>), (tracks, annotations, etc.)

序列保守性: WebLogo, (<http://weblogo.berkeley.edu/logo.cgi>),

基因预测: MEME, (<http://meme-suite.org/>).

进化树: iTOL, (<https://itol.embl.de/>),

基因调控网络: GeneNetwork, (<http://gn2.genenetwork.org/>), Cytoscape, (<https://cytoscape.org/>),

代谢通路: KEGG, (<https://www.kegg.jp/>); iPATH, (<https://pathways.embl.de/>),

蛋白结构与功能: PDB, (<http://www.rcsb.org>); pFAM, (<http://pfam.xfam.org/>),

微生物组: EBI Magnify. (<https://www.ebi.ac.uk/metagenomics/>),

蛋白和小分子互作数据: STITCH, (<http://stitch.embl.de/>); STRING, (<http://string-db.org>),

药物数据库: DrugBank, (<https://www.drugbank.ca/>),

生物数据分析平台: Galaxy, (<https://usegalaxy.org/>),

生物数据可视化: Echart, (<https://www.echartsjs.com/examples/zh/index.html>),

生物信息学常用数据库

- **一级数据库**
 - 数据库中的数据直接来源于实验获得的原始数据，只经过简单的归类整理和注释。
- **二级数据库**
 - 对原始生物分子数据进行整理、分类的结果，是在一级数据库、实验数据和理论分析的基础上针对特定的应用目标而建立的。

1.Nucleotide Sequence Databases

(1) 美国生物技术信息中心的GenBank

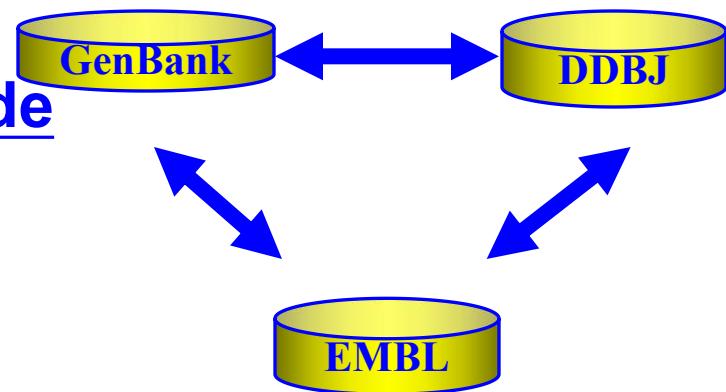
<http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>

(2) 欧洲分子生物学实验室的EMBL

<http://www.embl-heidelberg.de>

(3) 日本遗传研究所的DDBJ

<http://www.ddbj.nig.ac.jp/>



- 三个数据库中的数据基本一致，仅在数据格式上有所差别，对于特定的查询，三个数据库的响应结果一样。

文件



GenBank 1979年建设，1982年运行



← 输入关键字 直接搜索 → 转到 链接

Search

Launch

Foxit Messages

Foxit Online Services

Products

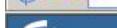
Images

Weather

News

Highlight

Pop-up Blocker



搜索



GenBank Overview

PubMed

Entrez

BLAST

OMIM

Books

Taxonomy

Structure

Search Entrez

for

Go

NCBI Home

NCBI Site Map

Submit to GenBank

Submit an update

Search GenBank

GenBank and RefSeq:
a comparison

BLAST

▶ What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2008 Jan;36\(Database issue\):D25-30](#)). There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in the WGS division as of August 2009.

The complete [release notes](#) for the current version of GenBank are available on the NCBI ftp site. A new release is made every two months. GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Molecular Biology Laboratory (EMBL), and GenBank at NCBI. These three organizations exchange data on a daily basis.

An example of a GenBank [record](#) may be viewed for a *Saccharomyces cerevisiae* gene.

In The News: 2009 H1N1 Flu Virus (Swine Flu)

The Centers for Disease Control and Prevention and other health officials are actively tracking the recent emergence of human cases of swine influenza A (H1N1) virus infection. Influenza A virus sequences from patients affected by this strain are being submitted to GenBank and can be accessed through the [NCBI Flu Resource](#)

▶ NLM/NCBI 2009 H1N1 Flu Resources:

- ④ Newest [2009 H1N1 influenza A sequences](#)
- ④ Citations [recently added](#) to PubMed
- ④ [MedlinePlus \(consumer health information\)](#)
- ④ [Enviro-Health Links](#)



▶ Submissions to GenBank

Many journals require [submission of sequence information](#) to a database prior to publication so



Searching GenBank

PubMed Entrez BLAST OMIM Books Taxonomy Structure

NCBI

SITE MAP

▶ Text and Similarity Searching

Entrez Browser

GenBank (nucleotides and proteins), PubMed (MEDLINE), 3D structures, genomes, and PopSet databases.

BLAST Sequence Similarity Searching

Nucleotide or protein query sequences against the specified database using the BLAST suite of algorithms.

dbEST Searching

dbEST (Database of Expressed Sequence Tags).

dbSTS Searching

dbSTS (Database of Sequence Tagged Sites).

dbGSS Searching

dbGSS (Database of Genome Survey Sequences).

▶ Information about Access to GenBank

Network Client/Server Applications

Network BLAST.

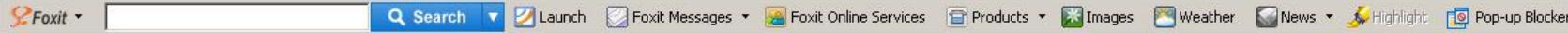
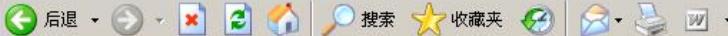
FTP

Full release and daily updates of GenBank.

- **Submissions to GenBank**
- Many journals require [submission of sequence information](#) to a database prior to publication so that an accession number may appear in the paper. NCBI has a WWW form, called [BankIt](#), for convenient and quick submission of sequence data. [Sequin](#), NCBI's stand-alone submission software for MAC, PC, and UNIX platforms, is also available by FTP. When using Sequin, the output files for direct submission should be sent to GenBank by electronic mail.
- There are specialized, streamlined procedures for batch submissions of sequences, such as [EST](#), [STS](#), and [HTG](#) sequences.
- **Updating or Revising a Sequence**
- Revisions or updates to GenBank entries can be made at any time and can be accepted as [BankIt](#) or [Sequin](#) files or as the text of an e-mail message. Click on the link for more information about [updating information on GenBank records](#).

- **Access to GenBank**
- GenBank is available for [searching](#) at NCBI via several methods.
- The GenBank database is designed to provide and encourage access within the scientific community to the most up to date and comprehensive DNA sequence information. Therefore, NCBI places no restrictions on the use or distribution of the GenBank data. However, some submitters may claim patent, copyright, or other intellectual property rights in all or a portion of the data they have submitted. NCBI is not in a position to assess the validity of such claims, and therefore cannot provide comment or unrestricted permission concerning the use, copying, or distribution of the information contained in GenBank.
- **New Developments**
- NCBI is continuously developing new tools and enhancing existing ones to improve both submission and access to GenBank. The easiest way to keep abreast of these and other developments is to check the "What's New" section of the NCBI Web page and to read the [NCBI News](#), which is also available by free subscription.

EMBL 1982年 运行



Search

EB-eye
Search

All Databases

Enter Text Here

Go Reset ?
Advanced SearchGive us
feedback

Databases

Tools

EBI Groups

Training

Industry

About Us

Help

Site Index



- EMBL-Bank Home
- Access
- Documentation
- News
- Submission
- Publications
- People
- Contact

EMBL Fetch

Fetch an EMBL record by id

Go

Hands-on Training**30th April - 1st May****2009:** Short ReadBioinformatics hands-on
EBI training course...[more](#)**Collaborations**

- ① **INSDC** - International Nucleotide Sequence Database Collaboration

- ② **NCBI** - The Nucleotide Sequence Database is produced in collaboration with GenBank (USA)

- ③ **DDBJ** - The Nucleotide Sequence Database is also produced in collaboration with the DNA Database of Japan (DDBJ)

EBI > Databases > EMBL-Bank

EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are [direct submissions](#) from individual researchers, genome sequencing projects and patent applications.



The database is produced in an international [collaboration](#) with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis. The [current database release](#) (Release 101, Sept 2009), with according [Release notes](#) and [user manual](#) are available from the EBI servers. A sample database entry is shown [here](#).

A publication in [Nucleic Acids Research 2009 37: D19-D25](#), provides further information and details.

The EMBL nucleotide sequence database forms part of the [European Nucleotide Archive](#), an EBI project led by [Guy Cochrane](#) as part of the [The Protein and Nucleotide Database Group \(PANDA\)](#) under [Ewan Birney](#).

Link	Explanation
Access	Database queries, Completed genomes webserver, FTP archives (EMBL release, alignments etc), EMBL sequence version archive (SVA), Browse by geography .
Submission	Primary sequence submissions, third party annotation, updates.
Documentation	Release notes user manual , Information for Submitters , FAQ , Release information , Forthcoming Changes , EMBL database statistics , Feature table , XML documentation , Sample entry , Accession Number Prefix Codes , Examples of annotation , EMBL Features & Qualifiers , DE line standards , Database Policies
Publications	Group publications
People	Group members
Contact	How to contact the EMBL Nucleotide Sequence Database
News	List of recent changes on this site

Contact

For information, comments and/or suggestions, please use the EBI Support Form page
<http://www.ebi.ac.uk/support/>

http://www.ebi.ac.uk/embl/index.html

DDBJ 1984年建立, 1987年启用

地址(D) http://www.ddbj.nig.ac.jp/ Norton AntiVirus

转到

DDBJ

Search

To Japanese Page

DDBJ Top Page

What's New

Welcome to DDBJ

DDBJing

SAKURA

Mass Data

Update/Correction

getentry

SRS

Homology Search

CIB Homepage

Site Map

DNA Data Bank of Japan

MEXT, Japan National Institute of Genetics

Center for Information Biology and DNA Data Bank of Japan

- Jump -

What's New

Hot Topics: [Homology Search] EST organism options change for DIVISION

- URL change of SQmatch and Data Retrieval by key words using SFgate-WAIS
- International Consortium Completes Human Genome Project

Welcome (to DDBJ)

- about DDBJ/CIB
- DDBJing (Japanese only)
- DDBJ Statistics
- DDBJ Publications Online
- Addresses Related to DDBJ Activities
- Conference and symposium related to DDBJ
 - H-invitational
 - MGED 5

Data Submission

- Introduction to data submission
- Submission via SAKURA
- Submission of mass data
- Guidance for Nucleotide Sequence
 - Data Submission to DDBJ
- From International Advisory Committee

Data Updates/Correction

Database Search

- getentry (data retrieval by accession numbers etc)
- SRS (data retrieval by key words)
- Homology Search (FASTA/BLAST/S&W SEARCH)
- TXSearch (retrieval of unified taxonomy database)
- XML Central of DDBJ (Web services by XML and SOAP)
- SQmatch (sequence search by regular expression)
- CAMUS Database (Compressed and Multiply Aligned DB)

Data Analysis

- ClustalW (multiple alignment and tree making)
- Use of Super Computer System (Japanese only)

Genome Analysis

- Human genome finished sequences
- GIB (integrated search of Bacteria, Archaea, Eukaryota)
- GTOP (Genome TO Protein structure and function)

Protein Database and Structure

- PMD (Protein Mutant Database)

2. Genome Databases

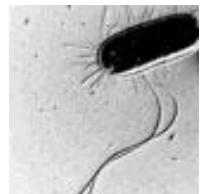
小鼠(Mouse)	http://www.informatics.jax.org/mgd.html
大鼠(Rat)	http://ratmap.gen.gu.se
狗(Dog)	http://mendel.berkeley.edu/dog.html
牛(Cow)	http://locus.jouy.inra.fr/cgi-bin/bovmap/intro2.pl
猪(Pig)	http://www.ri.bbsrc.ac.uk/pigmap/pigbase/pigbase.html
羊(Sheep)	http://dirk.invermay.cri.nz
鸡(Chicken)	http://www.ri.bbsrc.ac.uk/chickmap/chickbase/manager.html
斑马鱼(Zebra fish)	http://zfish.uoregon.edu
线虫(<i>C. elegans</i>)	http://www.ddbj.nig.ac.jp/htmls/celegans/html/CE_INDEX.html
果蝇(<i>Drosophila</i>)	http://morgan.harvard.edu
蚊子(Mosquito)	http://klab.agsci.colostate.edu
拟南芥(<i>Arabidopsis</i>)	http://genome-www.stanford.edu/Arabidopsis
棉花(Cotton)	http://algodon.tamu.edu
玉米(Maize)	http://www agron missouri edu
水稻(Rice)	http://www.staff.or.jp
大豆(Soya)	http://mendel agron iastate edu:8000/main html
杨树(Trees)	http://s27w007 pswfs gov

Model organism

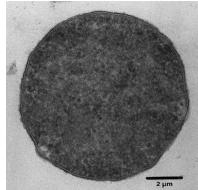
[Ureaplasma urealyticum](#)



Bacillus subtilis



[Buchnerasp.
APS](#)



[Thermoplasma acidophilum](#)

[Drosophila melanogaster](#)



[Escherichia coli](#)



[mouse](#)

[Plasmodium falciparum](#)



[Rickettsia prowazekii](#)



[Helicobacter pylori](#)



Arabidopsis



[Thermotoga maritima](#)



[Borrelia burgdorferi](#)



[Caenorhabditis elegans](#)



[Borrelia burgdorferi](#)

[Aquifex aeolicus](#)

[Neisseria meningitidis Z2491](#)

[Mycobacterium tuberculosis](#)

Model organism databases

- *Escherichia coli*
 - [E. coli Genome Center](#) (Wisconsin University, USA)
 - [The E. coli index](#) (University of Birmingham, UK)
- *S. cerevisiae* (Baker's yeast)
 - [SGD](#) (Yeast genome database at Stanford, USA)
 - [CYGD](#) (MIPS Comprehensive Yeast Genome Database, Neuherberg, Germany)
- *Arabidopsis thaliana*
 - [MATDB](#) (MIPS *A. thaliana* database, Munich, Germ.)
 - [TAIR](#) (The Arabidopsis Information Resource, previously AtDB, at Stanford, USA)
 - [KAOS](#) (Kazusa Arabidopsis data Opening Site at Kazusa DNA Research Institute, Jp)
 - [Arabidopsis Genome Analysis](#) (at Cold Spring Harbor laboratories, USA)
 - [TIGR Arabidopsis thaliana Database](#) (TIGR, Rockeville MD, USA)
- *Oryza sativa* (Rice)
 - [RGP](#) (Rice Genome Research Programme, Jp)
 - [Gramene](#) (Comparative mapping resource for graine)
 - [INE](#) (Integrated rice genome explorer: common database of the International Rice Genome Sequencing Project, IRGSP, Jp)

Model organism databases

- *Caenorhabditis elegans*
 - [WormBase](#) (*C. elegans* database at Cold Spring Harbor Laboratories, USA)
- *Drosophila melanogaster* (Fruit fly)
 - [FlyBase](#) (*Drosophila* genome database)
 - [BDGP](#) (Berkeley Drosophila genome project)
- *Danio rerio* (Zebrafish)
 - [ZFIN](#) (Zebrafish Information Network at University of Oregon, USA)
 - [WashU-Zebrafish Genome Resources](#) (Zebrafish EST database at Washington University, USA)
- *Mus musculus* (Mouse)
 - [MGI](#) (Mouse genome informatics)
- *Homo sapiens*
 - [GDB](#) (The human Genome Database, Toronto, Canada)
 - [HIB](#) (HumanInfoBase of annotated UniGene clusters - putative human gene transcripts - at MIPS, Germany)
 - [Human genome resources](#) (at NCBI, USA)
 - [Human genome browser](#) (at the University of California Santa Cruz, USA)
 - [HGP](#) (Human Genome Project at the Sanger Institute, Cambridge, UK)
 - [GeneLinks](#) (Portal to hyperlinks for each human gene at the Center for Genomics and Bioinformatics, Karolinska Institutet, Stockholm, Sweden)

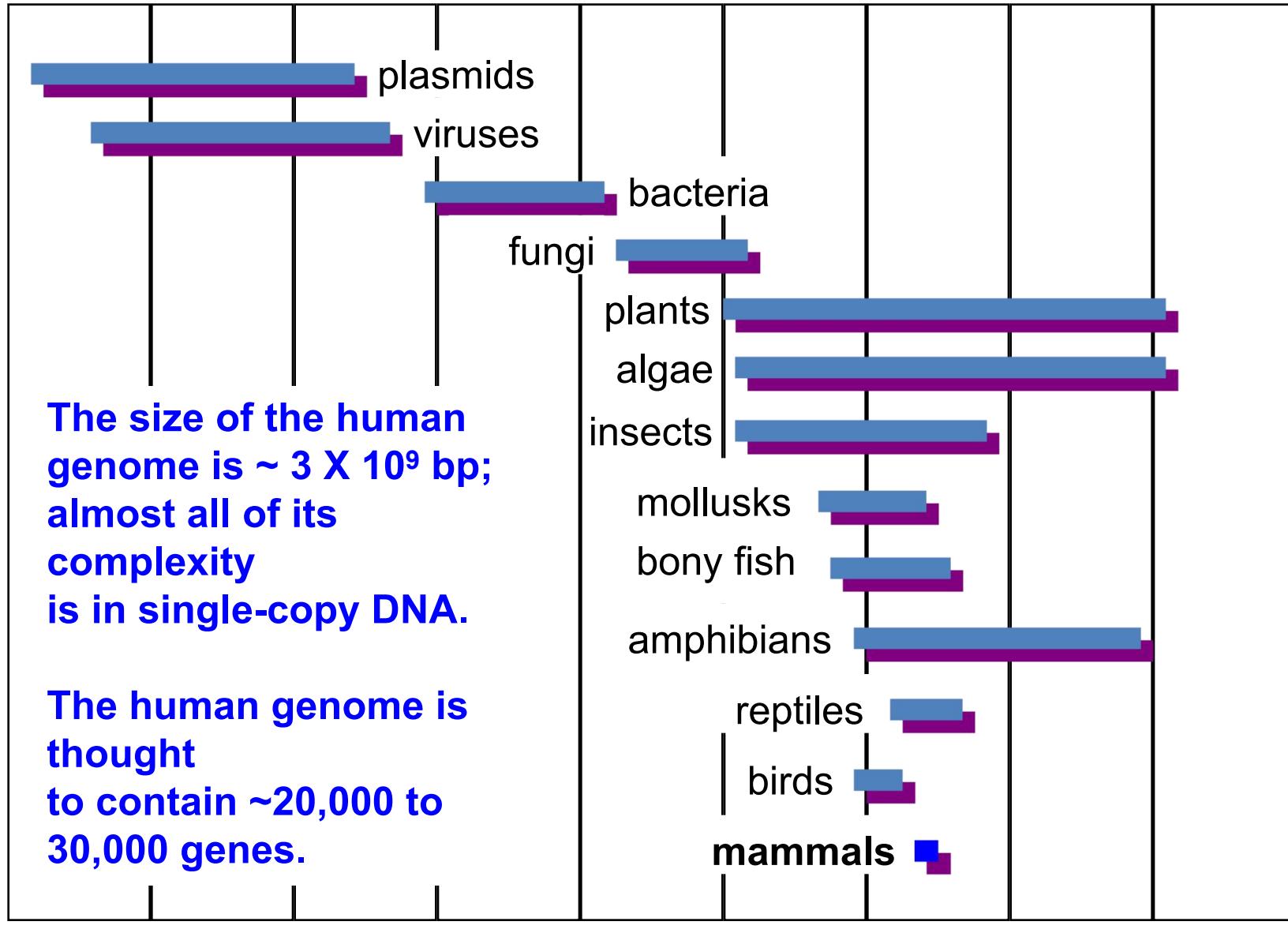
- Prokaryotes include:
 - *Escherichia coli* (*E. coli*) - This common, Gram-negative gut bacterium is the most widely-used organism in molecular genetics.
 - *Bacillus subtilis* - an endospore forming Gram-positive bacterium

Table of model genetic organisms

Organism	Genome Sequenced	Homologous Recombination	Biochemistry
Prokaryote			
<i>Escherichia coli</i>	Yes	Yes	Excellent
Eukaryote, unicellular			
<i>Dictyostelium discoideum</i>	Yes	Yes	Excellent
<i>Saccharomyces cerevisiae</i>	Yes	Yes	Good
<i>Schizosaccharomyces pombe</i>	Yes	Yes	Good
<i>Chlamydomonas reinhardtii</i>	Yes	No	Good
<i>Tetrahymena thermophila</i>	Yes	Yes	Good
Eukaryote, multicellular			
<i>Caenorhabditis elegans</i>	Yes	Difficult	Not so good
<i>Drosophila melanogaster</i>	Yes	Difficult	Good
<i>Arabidopsis thaliana</i>	Yes	No	Poor
Vertebrate			
<i>Danio rerio</i>	Yes	Difficult?	Good
<i>Mus musculus</i>	Yes	Yes	Good
<i>Homo sapiens</i>	Yes	Yes	Good 22

- The Genome database provides views for a variety of genomes, complete chromosomes, sequence maps with contigs, and integrated genetic and physical maps.
- The database is organized in six major organism groups: [Archaea](#), [Bacteria](#), [Eukaryotae](#), [Viruses](#), [Viroids](#), and [Plasmids](#) and includes complete chromosomes, organelles and plasmids as well as draft genome assemblies.

Genome sizes in nucleotide pairs (base-pairs)



The size of the human genome is $\sim 3 \times 10^9$ bp; almost all of its complexity is in single-copy DNA.

The human genome is thought to contain $\sim 20,000$ to $30,000$ genes.



地址(D) http://www.ncbi.nlm.nih.gov/genome/guide/

输入关键字 直接搜索 转到 链接



潘晶

NCBI Genome Resource Guides

Access to genome resource guides for selected organisms.

Mammals

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
human	Build 37.1	37.1	3 Aug 2009	
mouse	Build 37.1	37.1	5 Jul 2007	
rat	RGSC v3.4	4.1	6 Jul 2006	
cow	Btau_4.0	4.1	5 Aug 2008	
dog	Build 2.1	2.1	8 Sep 2005	

Show (+)

Birds

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
chicken	Build 2.1	2.1	29 Nov 2006	
zebra finch	Build 1.1	1.1	5 Mar 2009	

Amphibians

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
frog	na	na	na	

Echinoderms

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
sea urchin	Build 2.1	2.1	18 Oct 2006	

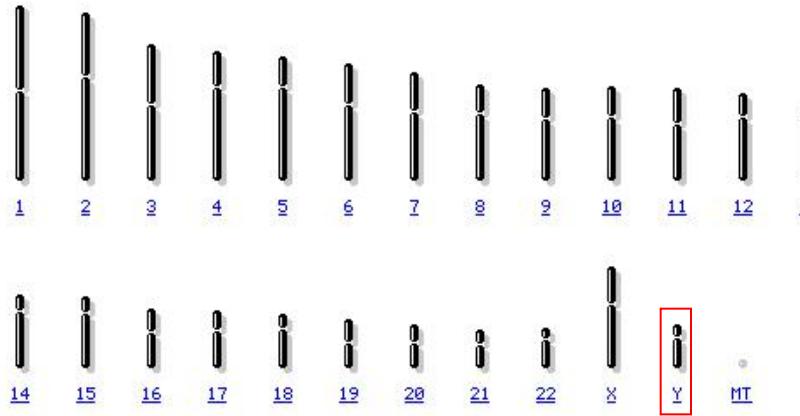
Fish

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
zebrafish	Zv7	3.1	12 Jul 2008	
fugu	Truy4.0	na	na	
pufferfish	Tniv7	na	na	
stickleback	Broad v1.0	na	na	

Insects

Organism	Reference Assembly	Current NCBI Build	Map Viewer Release date	Resource Links
fruit fly	Release 5.2	9.1	17 Apr 2008	
African malaria mosquito	AgamP3.3	3.1	10 Oct 2008	
red flour beetle	Build 2.1	2.1	9 Jun 2008	
Drosophila pseudoobscura	Release 2.0	1.1	17 Apr 2008	
honey bee	Amel_4.0	4.1	11 Aug 2006	

Show (+)



Lineage: Eukaryota; Metazoa; Chordata; Olfactory receptor genes; Homo sapiens

August 2009: NCBI released an updated version of the human genome assembly. This assembly update includes modifications to all chromosomes provided by the [Genome Reference Consortium](#) (GRC) and is used for display and for BLAST. For additional information

Human genome overview page (Build 37.1)
 Human genome overview page (Build 36.3)

[Map Viewer Home](#)
[Map Viewer Help](#)
[Human Maps Help](#)
[FTP](#)
[Data As Table View](#)
Maps & Options
 Compress Map
 Region Shown:

 You are here:
Ideogram
 Yp11.3
 Yp11.2
 YH1
 Yq11.2
 Yq12

Master Map: Genes On Sequence
 Region Displayed: 0-59M bp

Ideogram **Symbol** **E** **Cyto** **Description**

Contig	UniGene Cluster	Genes.seq	Symbol	Links	E	Cyto	Description
Yp11.32	NT_167199, NT_167200, NT_167201, NT_167202, NT_167203, NT_167204, NT_167205, NT_011896		ASMTL	+ OMIM HGNC sv pr dlev mm hm sts CCDS SNP best RefSeq	Xp22.3; Yp11.3	acetylserotonin O-methyltransferase-like protein	
Yp11.31			LOC359800	+ sv dlev mm sts	best RefSeq	Yp11.2	voltage-dependent anion channel 1
Yp11.2			LOC439957	+ svpr dlev mm hm	SNP	protein Y	similar to hCG174244
Yp11.1	NT_086998, NT_011878, NT_087001, NT_113819, NT_011875		LOC100287882	+ sv dlev mm	mRNA	Y	hypothetical protein LOC100287882
Yq11.1			LOC401629	+ sv dlev mm	best RefSeq	Y	hypothetical LOC401629
Yq11.21			HSFY1	+ OMIM HGNC svpr dlev mm hm sts CCDS	best RefSeq	Y	heat shock transcription factor Y1
Yq11.22			TTTY10	+ HGNC sv dlev mm sts	best RefSeq	Yq11.221	testis-specific transcript TTTY10
Yq11.23			CDY11P	+ HGNC sv dlev mm sts	best RefSeq	Y	testis-specific transcript CDY11P
Yq12			TTTY5	+ OMIM HGNC sv dlev mm	best RefSeq	Y	chromodomain protein TTTY5
			OFDYP9	+ HGNC sv dlev mm sts	best RefSeq	Yq11.223	OFD1 pseudogene Y
			RBMY2WP	+ HGNC sv dlev mm	best RefSeq	Yq11.223	RNA binding motif protein RBMY2WP
			CDY15P	+ HGNC sv dlev mm	best RefSeq	Yq11.223	chromodomain protein CDY15P
			LOC359999	+ sv dlev mm	best RefSeq	Yq11.2	CHRNA7 (cholinergic receptor, nicotinic, alpha 7)
			TRAPPC2P5	+ HGNC sv dlev mm sts	best RefSeq	Yq11.23	trafficking protein part TRAPPC2P5
			PRY4	+ HGNC sv dlev mm	best RefSeq	Yq11.223	PTPN13-like, Y-linked PRY4
			PPP1R12BP	+ HGNC sv dlev mm sts	best RefSeq	Yq11.23	protein phosphatase 1 regulatory subunit 12B
			PARP4P	+ HGNC sv dlev mm sts	best RefSeq	Y	poly (ADP-ribose) polymerase 4
			LOC100288402	+ sv dlev mm	mRNA	Y	hypothetical protein LOC100288402
			SPRY3	+ OMIM HGNC svpr dlev mm hm sts CCDS SNP best RefSeq	Xq28 and Yq12	sprouty homolog 3 (Drosophila)	
			LOC727856	+ sv dlev mm	best RefSeq	Xq28; Yq12	DEAD/H (Asp-Glu-A

Summary of Maps:
 Map 1: Ideogram
 Map 2: Contig
 Map 3: Human sapiens UniGene Clusters

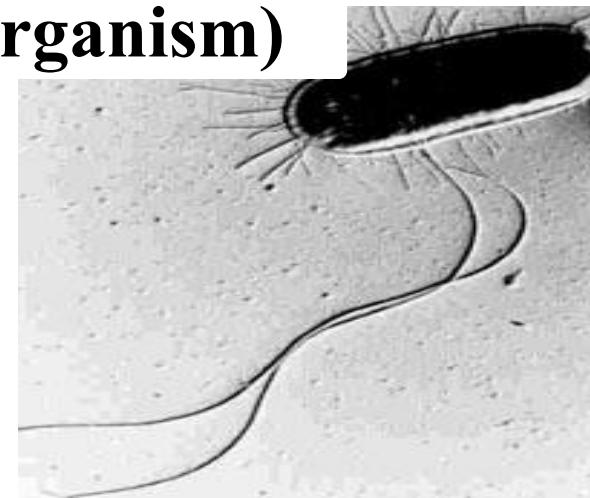
Region Displayed: Ypter-Yqter
 Region Displayed: 0-59M bp
 Total Contigs On Chromosome: 17
 Contigs Labeled: 17 Total Contigs in Region: 17
 Region Displayed: 0-59M bp

[Table View](#)
[Download/View Sequence/Evidence](#)
[Table View](#)
[Download/View Sequence/Evidence](#)

模式生物(Model Organism)



Escherichia coli 大肠杆菌



[Genome](#) > [Bacteria](#) > *Escherichia coli* O157:H7 str. FRIK2000, whole genome shotgun sequencing project

Lineage: [Bacteria](#); [Proteobacteria](#); [Gammaproteobacteria](#); [Enterobacteriales](#); [Enterobacteriaceae](#); [Escherichia](#); [Escherichia coli](#); [Escherichia coli](#) O157:H7; [Escherichia coli](#) O

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NZ_ACX000000000	Genes: 5383	COG	Genome Project	Publications: None
GenBank: ACX000000000	Protein coding: 5300	TaxMap	Refseq FTP	Refseq Status: WGS
Length: 5,408,690 nt	Structural RNAs: 83	TaxPlot	GenBank FTP	Seq. Status: Draft
GC Content: 50%	Pseudo genes: None	GenePlot	BLAST	Sequencing center: Medical Biofilm Research Institute
% Coding: 85%	Others: 56	gMap	TraceAssembly	Completed: 2009/08/20
Topology: other	Contigs: 247		CDD	Organism Group
Molecule: DNA			Other genomes for species: 170	

Gene Classification based on [COG functional categories](#)

Search gene, GenelD or locus_tag: Find Gene



Map Viewer - Microsoft Internet Explorer

文件(E) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退() 前进() 搜索() 收藏夹() Foxit Foxit Messages Foxit Online Services Products Images Weather News Highlight Pop-up Blocker

地址(D) http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?taxid=4932&chr=XVI

Search Launch Foxit Messages Foxit Online Services Products Images Weather News Highlight Pop-up Blocker

SOSO 搜索

PubMed Entrez BLAST OMIM Taxonomy Structure

Search Find Find in This View Advanced Search

Map Viewer Home Map Viewer Help Saccharomyces cerevisiae Maps Help FTP Data As Table View Maps & Options Compress Map Region Shown: Go

out zoom in

You are here: Genes_seq

Region Displayed: 0-950K bp

Master Map: Genes On Sequence

Contig Genes_seq Symbol Links E Description

YPL260W external Putative substrate of cAMP-dependent protein kinase (PKA); green fluorescent protein (GFP)-fusion protein
 ICY2 external Protein of unknown function; mobilized into polysomes upon a shift from a fermentable to nonfermentable carbon source
 NSL1 external Essential component of the MIND kinetochore complex (Mtw1p Including Nnf1p-Nsl1p-Dsn1p) which joins sister chromatids
 SAR1 external GTPase, GTP-binding protein of the ARF family, component of COPII coat of vesicles; required for transport of proteins from the ER to the Golgi
 SRP72 external Core component of the signal recognition particle (SRP) ribonucleoprotein (RNP) complex that functions in target protein import into the endoplasmic reticulum (ER)
 MF(ALPHA)1 external Mating pheromone alpha-factor, made by alpha cells; interacts with mating type a cells to induce cell cycle arrest
 PPO1 external Putative protein serine/threonine phosphatase; null mutation enhances efficiency of translational suppressors
 COX11 external Mitochondrial inner membrane protein required for delivery of copper to the Cox1p subunit of cytochrome c oxidase
 IDI1 external Isopentenyl diphosphate:dimethylallyl diphosphate isomerase (IPP isomerase), catalyzes an essential activation step in the biosynthesis of terpenoids
 YPL088W external Putative aryl alcohol dehydrogenase; transcription is activated by paralogous transcription factors Yrm1p and Yrm2p
 GPI2 external Protein involved in the synthesis of N-acetylglucosaminyl phosphatidylinositol (GlcNAc-PI), the first intermediate in the synthesis of glycosphingolipids
 PDR12 external Plasma membrane ATP-binding cassette (ABC) transporter, weak-acid-inducible multidrug transporter required for resistance to a variety of drugs
 NCR1 external Vacuolar membrane protein that transits through the biosynthetic vacuolar protein sorting pathway, involved in protein targeting and degradation
 MCM16 external Protein involved in kinetochore-microtubule mediated chromosome segregation, binds to centromere DNA
 ROX1 external Heme-dependent repressor of hypoxic genes; contains an HMG domain that is responsible for DNA bending
 TKL1 external Transketolase, similar to Tkl2p; catalyzes conversion of xylulose-5-phosphate and ribose-5-phosphate to sedoheptulose-7-phosphate
 YPR108W-A external Putative protein of unknown function; identified by fungal homology and RT-PCR
 CLB5 external B-type cyclin involved in DNA replication during S phase; activates Cdc28p to promote initiation of DNA synthesis
 VPS4 external AAA-type ATPase that is regulated by Vta1p; required for late endosome to vacuole transport, catalyzes the hydrolysis of inositol polyphosphates
 ATG13 external Phosphorylated protein that interacts with Vac8p, required for the cytoplasm-to-vacuole targeting (Cvt) pathway

Summary of Maps:

Map 1: Contig Table View

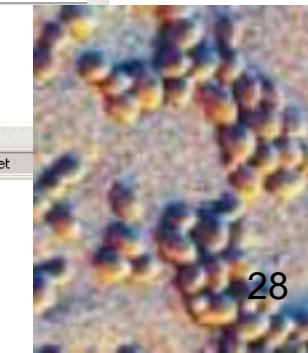
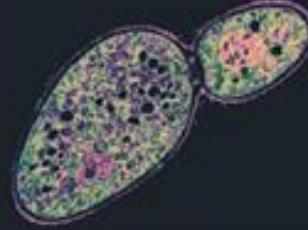
Region Displayed: 0-950K bp Download/View Sequence

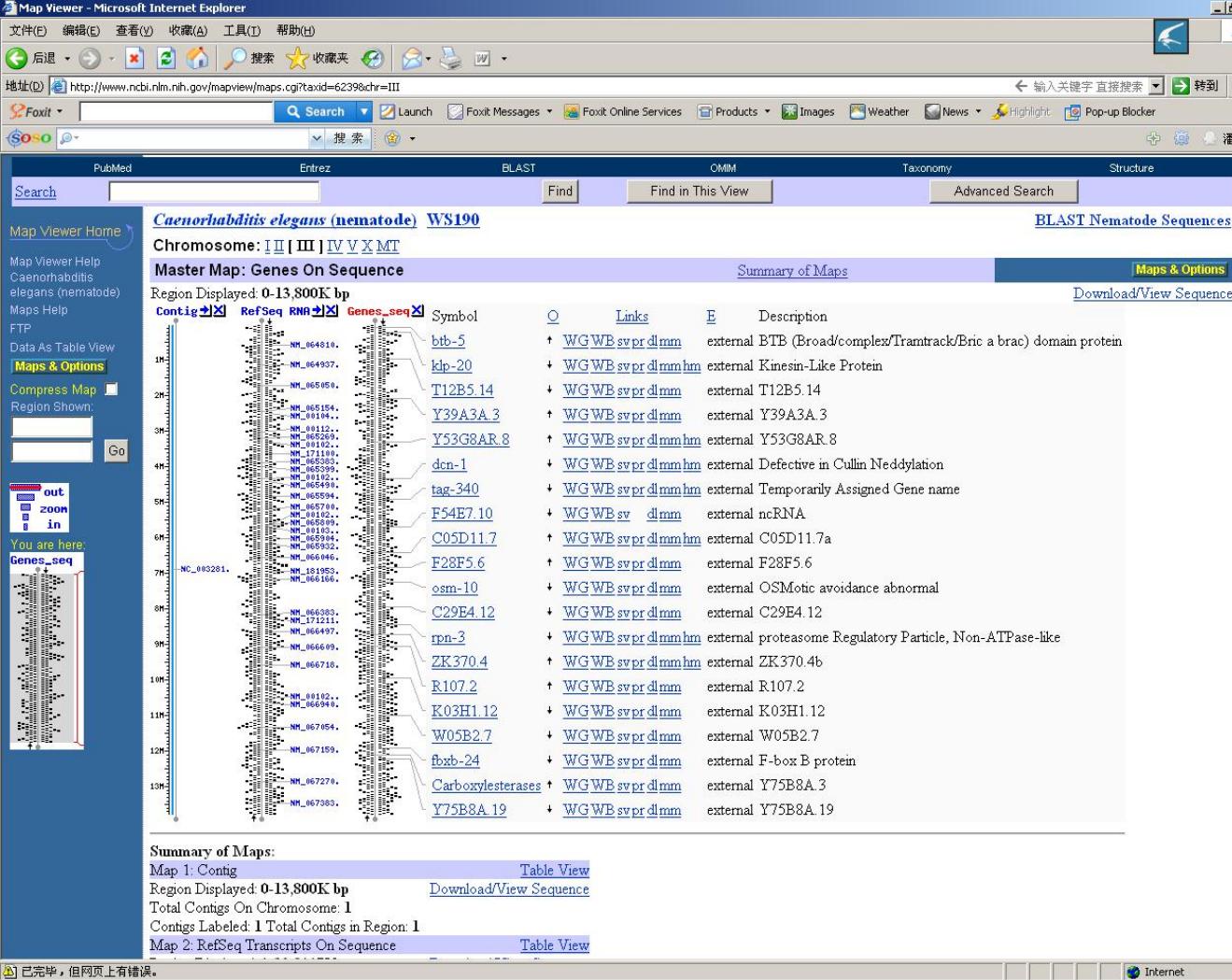
Total Contigs On Chromosome: 1

Contigs Labeled: 1 Total Contigs in Region: 1

网页上有错误。

酿酒酵母：16个染色体，全基因组1996年测定。





秀丽线虫：

雌雄同体成虫细胞数目只有959个，其中包括302个神经元；
 6条染色体，全基因组于1998年测定，长9.7Mb

果蝇: 繁殖很快, 基因组: 180Mb。

uid=6185 Genome Result - Microsoft Internet Explorer

文件(E) 编辑(E) 查看(V) 收藏(A) 工具(I) 帮助(H)

后退() 前进() 搜索() 收藏夹() Foxit Foxit Search Launch Foxit Messages Foxit Online Services Products Images Weather News Highlight Pop-up Blocker

地址() http://www.ncbi.nlm.nih.gov/sites/entrez?Db=genome&Cmd>ShowDetailView&TermToSearch=6185

输入关键字 直接搜索

All Databases PubMed Nucleotide Protein Genome Structure OMIM PMC Journals Books

Search Genome for Go Clear

Limits Preview/Index History Clipboard Details

Display Overview Show 20 Send to

All: 1

Genome > Eukarya > Drosophila melanogaster, whole genome shotgun sequencing project

Lineage: Eukarya; Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Protostomia; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea; Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha; Eremoneura; Cyclorrhapha; Schizophora; Acalyptratae; Ephyoidea; Drosophilidae; Drosophilinae; Drosophilini; Drosophilina; Drosophilidae; Drosophila; Sophophora; melanogaster group; melanogaster subgroup; Drosophila melanogaster

Chromosomes: master WGS, 2L, 2R, 3L, 3R, 4, X, Un

Organelles: mitochondrial-MT

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
RefSeq: NZ_AABU00000000	Genes: 837	COG	Genome Project	Publications: [8]
GenBank: AABU00000000	Protein coding: 702	TaxMap	RefSeq FTP	RefSeq Status: WGS
Length: 137,586,636 nt	Structural RNAs: 218	TaxPlot	GenBank FTP	Seq Status: Draft
GC Content: 39%	Pseudo genes: 43	GenePlot	BLAST	Sequencing center: The FlyBase Consortium/Berkeley Drosophila Genome Project/Celera Genomics
% Coding: N/A	Others: 132	gMap	TraceAssembly	Completed: 2009/01/27
Topology: other	Contigs: 2756		CDD	Organism Group
Molecule: DNA			Other genomes for species:	

Gene Classification based on COG functional categories

Search gene, GenelD or locus_tag: Find Gene

Zoom 1 nt 50,001 nt

Internet



Map Viewer - Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(I) 帮助(H)

地址(D) http://www.ncbi.nlm.nih.gov/mapview/maps.cgi?TAXID=3702&CHR=28MAPS=crtg-r,clone,tair_marker,genes[9750653..56%3A9947636..45]&ZOOM=100.0000

Search Launch Foxit Messages Foxit Online Services Products Images Weather News Highlight Pop-up Blocker

SOSO 搜索

NCBI NCBI Map Viewer

PubMed Entrez BLAST OMIM Taxonomy Structure

Search Find Find in This View Find in All Plants Find in All Eudicots Advanced Search

Arabidopsis thaliana (thale cress) Build 9.1

Chromosome: 1 [2] 3 4 5 MT Ptd

Master Map: Genes On Sequence

Region Displayed: 0-19,700K bp

Clone Marker Genes_seq

Symbol	Links	E	Description
AT2G02910	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	unknown protein
AT2G05435	TIGR MIPS TAIR SIGnAL sv dlevmm	external	
AT2G07570	TIGR MIPS TAIR SIGnAL sv dlevmm	external	
AT2G10560	TIGR MIPS TAIR SIGnAL svpr dlevmm	external	unknown protein
AT2G12832	TIGR MIPS TAIR SIGnAL sv dlevmm	external	
AT2G15070	TIGR MIPS TAIR SIGnAL sv dlevmm	external	
AT2G17870	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	cold-shock DNA-binding family protein
AT2G21010	TIGR MIPS TAIR SIGnAL svpr dlevmm	external	C2 domain-containing protein
AT2G24255	TIGR MIPS TAIR SIGnAL svpr dlevmm	external	unknown protein
AT2G27420	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	cysteine proteinase, putative
AT2G30984	TIGR MIPS TAIR SIGnAL sv dlevmm	external	misc_RNA
AtRLP23	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	AtRLP23 (Receptor Like Protein 23), kinase/ protein bin
AT2G34250	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	protein transport protein sec61, putative
FLA16	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	FLA16 (FASCICLIN-LIKE ARABINOGALACTAN)
AT2G37480	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	CPuORF52 (Conserved peptide upstream open reading
AT2G40800	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	unknown protein
MIR778A	TIGR MIPS TAIR SIGnAL sv dlevmm	external	misc_RNA
AT2G43900	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	inositol or phosphatidylinositol phosphatase/ inositol-poly
SAP18	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	SAP18 (SIN3 ASSOCIATED POLYPEPTIDE P18); p
AT2G47490	TIGR MIPS TAIR SIGnAL svpr dlevmm hm	external	mitochondrial substrate carrier family protein

Summary of Maps: Map 1: Contig

Table View

Internet

已完毕, 但网页上有错误。



The *Arabidopsis* genome shows extensive duplication of large chromosomal segments. The five chromosomes were analysed using a genome-wide sequence comparison tool. Segments with identical colours show apparent duplications, white regions represent unique portions, and hatched regions contain centromeric repeat elements. Telomeric repeats and nucleolar organizer regions are not shown.

拟南芥: 个体生命周期只有6周的十字花科小草, 是一种理想的模式植物。

非洲瓜蟾 (*Xenopus laevis*)

1个受精卵在24小时内分裂到各种器官初具雏形的程度；

uid=10422 Genome Result - Microsoft Internet Explorer

文件(F) 编辑(E) 查看(V) 收藏(A) 工具(T) 帮助(H)

后退() 前进() 搜索() 收藏夹()

地址(D): http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome&Cmd>ShowDetail&ViewTermToSearch=10422

Foxit Search Launch Foxit Messages Products Images Weather News Highlight Pop-up Blocker

Soso 搜索

Display Overview Show 20 Send to

All 1

Genome > Eukaryota > Xenopus laevis mitochondrion, complete genome

Lineage: Eukaryota, Fungi/Metazoa group; Metazoa; Eumetazoa; Bilateria; Coelomata; Deuterostoma; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Tetrapoda; Amphibia; Batrachia; Anura; Mesobatrachia; Pipoidea; Pipidae; Xenopodinae; Xenopus; Xenopus; Xenopus laevis

Genome Info:

Features:	BLAST homologs:	Links:	Review Info:
RefSeq: NC_001573 Genes: 13	COG	Genome Project	Publications: [1]
GenBank: M10217 Protein coding: 13	TaxMap	RefSeq FTP	RefSeq Status: Reviewed
Length: 17,553 nt Structural RNAs: 24	TaxPlot	GenBank FTP	Seq Status: Completed
GC Content: 36%	Pseudo genes: None	GenePlot	Sequencing center: None
% Coding: 64%	Others: 6	gMap	Completed: 1999-08-24
Topology: circular	Contigs: None	TraceAssembly	
Molecule: dsDNA		CDD	Organism Group
		Other genomes for species:	

Gene Classification based on COG functional categories

Search gene, GenID or locus_tag: Find Gene

Zoom

1 nt

16S ribosomal RNA

ND1

ND2

7,016 nt

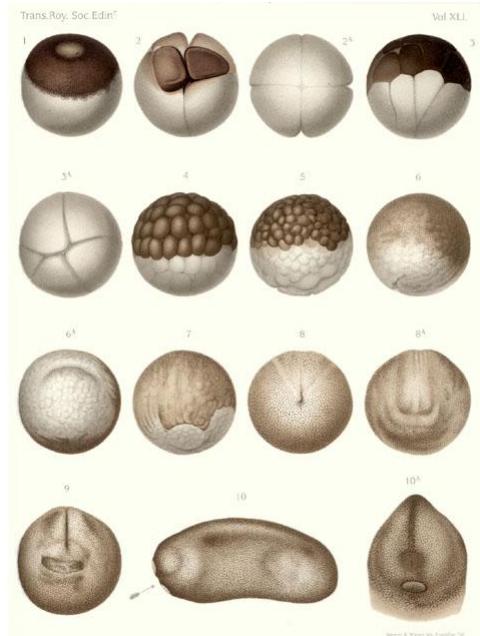
17,553 nt

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

Display Overview Show 20 Send to

Write to the Help Desk
NCBI | NLM | NIH
Department of Health & Human Services
Privacy Statement | Freedom of Information Act | Disclaimer

完毕 Internet



斑马鱼 (*Danio rerio*)

身体透明的小鱼，生活周期约3个月，是研究脊椎动物发育过程的良好对象。



NCBI NCBI Map Viewer

PubMed Entrez BLAST OMIM Taxonomy Structure

Search Find Find in This View Advanced Search

Danio rerio (zebrafish) Zv7

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 [25] MT

Master Map: Genes On Sequence Summary of Maps BLAST Zebrafish Sequences

Region Displayed: 0-33M bp Maps & Options Download/View Sequence/Evidence

Contig Dr UniG Genes seq Symbol Links E Description

Dr_00187..	zgc:153678	+ ZFIN ug svpr dl ev mm	hm	best RefSeq zgc:153678
Dr_00187..	LOC100148837	+ svpr dl ev mm	mRNA	similar to solute carrier family 35, member B4
Dr_00187..	cyp19a1b	+ ZFIN ug svpr dl ev mm sts hm	best RefSeq	cytochrome P450, family 19, subfamily A, polypeptide 1b
Dr_00187..	met	+ ZFIN ug svpr dl ev mm	hm	best RefSeq met proto-oncogene (hepatocyte growth factor receptor)
Dr_00187..	LOC558687	+ svpr dl ev mm	hm	mRNA similar to type II CAX cation/proton exchanger
Dr_00187..	LOC569055	+ ug svpr dl ev mm	hm	mRNA hypothetical LOC569055
Dr_00187..	LOC795008	+ svpr dl ev mm	hm	mRNA hypothetical protein LOC795008
Dr_00187..	ergic2	+ ZFIN ug svpr dl ev mm sts hm	best RefSeq	ERGIC and golgi 2
Dr_00187..	zgc:91794	+ ZFIN ug svpr dl ev mm sts hm	best RefSeq	zgc:91794
Dr_00187..	LOC797354	+ ug svpr dl ev mm	hm	mRNA similar to SMC1 beta protein
Dr_00187..	si_dkey-80c24.6	+ ZFIN ug svpr dl ev mm	best RefSeq	si_dkey-80c24.6
Dr_00187..	LOC569672	+ ug svpr dl ev mm	hm	mRNA similar to novel serine/threonine protein kinase
Dr_00187..	LOC100001260	+ svpr dl ev mm	mRNA	similar to type-1 protein phosphatase skeletal muscle glycogen targeting sub
Dr_00187..	wufd21f07	+ ZFIN ug svpr dl ev mm	hm	best RefSeq wufd21f07
Dr_00187..	zgc:110459	+ ZFIN ug svpr dl ev mm	hm	best RefSeq zgc:110459
Dr_00187..	LOC100150336	+ svpr dl ev mm	protein	similar to pol polyprotein
Dr_00187..	LOC564485	+ svpr dl ev mm	mRNA	hypothetical LOC564485
Dr_00187..	cox5aa	+ ZFIN ug svpr dl ev mm	best RefSeq	cytochrome c oxidase subunit Vaa
Dr_00187..	LOC792268	+ ug svpr dl ev mm	hm	best RefSeq hypothetical protein LOC792268
Dr_00187..	zgc:114037	+ ZFIN ug svpr dl ev mm	hm	best RefSeq zgc:114037

Summary of Maps: Map 1: Contig Table View

小鼠 (*Mus musculus*)
基因组大小与人类相近，有19条常染色体；



NCBI Map Viewer

PubMed Entrez BLAST OMM Taxonomy Structure

Search Find Find in This View Advanced Search

Mus musculus (laboratory mouse) Build 37.1 (Current)

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 [19] X Y MT

Master Map: Genes On Sequence

Region Displayed: 0-61M bp

Assembly Contig UniGene Genes_seq Symbol Links Description

129/Sv1a 129/SvEv... NT_082868. Fosl1 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq fos-like antigen 1

129/SvEv... NT_082868. Map3k11 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq mitogen activated protein kinase kinase kinase

NT_082868. Rcor2 MGI sv pr dl ev mm hm sts CCDS SNP GENSAT best RefSeq REST corepressor 2

BC014805 BC014805 MGI sv pr dl ev mm hm CCDS best RefSeq cDNA sequence BC014805

NT_082868. Slc3a2 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq solute carrier family 3 (activators of dibasic mRNA

NT_082868. Stzbp3b MGI sv pr dl ev mm mRNA syntaxin-binding protein 3B

NT_082868. LOC100040203 sv pr dl ev mm hm protein similar to prostatein C3 subunit

NT_082868. Fads2 MGI sv pr dl ev mm hm CCDS SNP best RefSeq fatty acid desaturase 2

NT_082868. 2810441K11Rik MGI sv pr dl ev mm hm sts CCDS best RefSeq RIKEN cDNA 2810441K11 gene

NT_082868. Plac11 MGI sv pr dl ev mm hm CCDS SNP GENSAT best RefSeq placenta-specific 1-like

NT_082868. Zfp91 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq zinc finger protein 91

NT_082868. Olf1454 MGI sv pr dl ev mm hm CCDS best RefSeq olfactory receptor 1454

NT_082868. Olf1470-ps1 MGI sv dl ev mm SNP best RefSeq olfactory receptor 1470, pseudogene 1

NT_082868. Ankrd15 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq ankyrin repeat domain 15

NT_082868. Rln1 MGI sv pr dl ev mm hm CCDS best RefSeq relaxin 1

NT_082868. Trpd5213 MGI sv pr dl ev mm hm CCDS best RefSeq tumor protein D52-like 3

NT_082868. Mms191 MGI sv pr dl ev mm hm sts CCDS SNP best RefSeq MMS19 (MET18 S. cerevisiae)-like mRNA

NT_082868. EG627889 MGI sv pr dl ev mm hm SNP predicted gene, EG627889

NT_082868. 1700011F14Rik MGI sv pr dl ev mm hm CCDS SNP best RefSeq RIKEN cDNA 1700011F14 gene

NT_082868. Emx2os MGI sv dl ev mm sts SNP best RefSeq empty spiracles homolog 2 (Drosophila) op

Summary of Maps

Map 1: Assembly

Table View

BLAST The Mouse Genome

Maps & Options

Download/View Sequence/Evidence

Region Shown:

Go

out zoom in

You are here: Ideogram

19A
19B
19C1
19C2
19C3
19D1
19D2
19D3

default master

34

BLAST

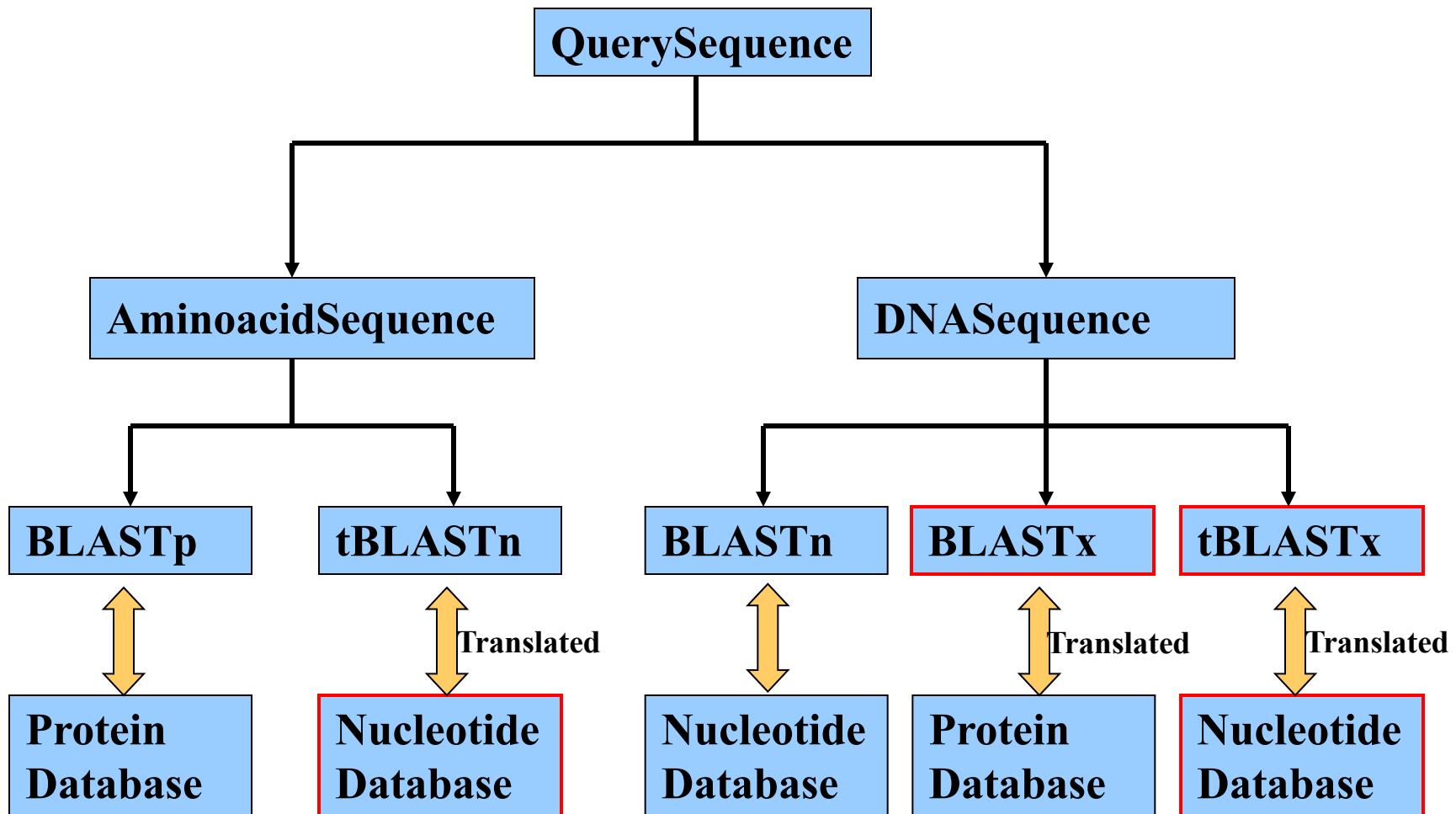
- 基本局部比对搜索工具
(Basic Local Alignment Search Tool)
- NCBI上BLAST服务的网址:
- <http://blast.ncbi.nlm.nih.gov/>
- NCBI上BLAST程序的下载:
[ftp://ftp.ncbi.nlm.nih.gov/blast/executables/
release/](ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/)
- NCBI的BLAST数据库下载网址:
<ftp://ftp.ncbi.nlm.nih.gov/blast/db/>



选择物种



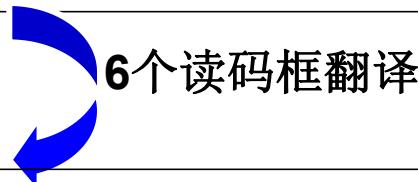
选择blast程序



程序名	搜索序列	数据库	内容	备注
blastp	Protein	Protein	比较氨基酸序列与蛋白质数据库	使用取代矩阵寻找较远的关系，进行 SEG 过滤
blastn	Nucleotide	Nucleotide	比较核酸序列与核酸数据库	寻找较高分值的匹配，对较远的关系不太适用
blastx	Nucleotide	Protein	比较核酸序列理论上的六个读码框的所有转换结果和蛋白质数据库	用于新的 DNA 序列和 ESTs 的分析，可转译搜索序列
tblastn	Protein	Nucleotide	比较蛋白质序列和核酸序列数据库，动态转换为六个读码框的结果	用于寻找数据库中没有标注的编码区，可转译数据库序列
tblastx	Nucleotide	Nucleotide	比较核酸序列和核酸序列数据库，经过两次动态转换为六个读码框的结果	转译搜索序列与数据库序列

以Blastx为例：

目标序列为**ATG AGT ACC GCT AAA TTA GTT AAA
TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC**



5'端到3'端

第一位起始:

ATG AGT ACC GCT AAA TTA GTT AAA TCA AAA GCG ACC AAT CTG CTT TAT ACC CGC

第二位起始:

TGA GTA CCG CTA AAT TAG TTA AAT CAA AAG CGA CCA ATC TGC TTT ATA CCC GC

第三位起始:

GAG TAC CGC TAA ATT AGT TAA ATC AAA AGC GAC CAA TCT GCT TTA TAC CCG C

3'端到5'端

第一位起始:

GCG GGT ATA AAG CAG ATT GGT CGC TTT TGA TTT AAC TAA TTT AGC GGT ACT CAT

第二位起始:

CGG GTA TAA AGC AGA TTG GTC GCT TTT GAT TTA ACT AAT TTA GCG GTA CTC AT

第三位起始:

GGG TAT AAA GCA GAT TGG TCG CTT TTG ATT TAA CTA ATT TAG CGG TAC TCA T

+3	E	Y	R	*	I	S	*	I	K	S	D	Q	S	A	L	Y	P		
+2	*	V	P	L	N	*	L	N	Q	K	R	P	I	C	F	I	P		
+1	M	S	T	A	K	L	V	K	S	K	A	T	N	L	L	Y	T	R	
5'-	ATG	AGT	ACC	GCT	AAA	TTA	GTT	AAA	TCA	AAA	GCG	ACC	AAT	CTG	CTT	TAT	ACC	CGC	-3'
	TAC	TCA	TGG	CGA	TTT	AAT	CAA	TTT	AGT	TTT	CGC	TGG	TTA	GAC	GAA	ATA	TGG	GCG	
	H	T	G	S	F	*	N	F	*	F	R	G	I	Q	K	I	G	A	-1
	L	V	A	L	N	T	L	D	F	A	V	T	R	S	*	V	R		-2
	S	Y	R	*	I	L	*	I	L	L	S	W	D	A	K	Y	G		-3



>[ref|YP_688332.1|](#) **G** DNA starvation/stationary phase protection protein Dps [Shigella flexneri 5 str. 8401]

[gb|ABF03027.1|](#) **G** global regulator, starvation conditions [Shigella flexneri 5 str. 8401]

Length=208

GENE ID: 4209430 **dps** | DNA starvation/stationary phase protection protein Dps [Shigella flexneri 5 str. 8401] (**10 or fewer PubMed links**)

Score = 37.4 bits (85), Expect = 0.36

Identities = 18/18 (100%), Positives = 18/18 (100%), Gaps = 0/18 (0%)

Frame = +1

Query	1	MSTAKLVKSATNLLYTR	54
		MSTAKLVKSATNLLYTR	
Sbjct	42	MSTAKLVKSATNLLYTR	59

blastn

BLAST® Basic Local Alignment Search Tool My 收集本页 [Sign in] [register]

Home Recent Results Saved Strategies Help

NCBI/ BLAST/ blastn suite Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
Query subrange From To
序列或目标序列的GI号

Or, upload file 浏览...
Job Title
Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set
Database Human genomic + transcript Mouse genomic + transcript Others (nr etc.)
Nucleotide collection (nr/nt)

Organism Optional
Enter organism name or id—completions will be suggested Exclude
Enter organism common name, taxon ID, or NCBI taxon ID—top taxa will be shown

Exclude Optional
 Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional
 Sequences from type material

Entrez Query Optional
Enter an Entrez query to limit search YouTube Create custom database

Program Selection
Optimize for
 Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)
Choose a BLAST algorithm

blastn算法选择

BLAST
 Show results in a new window

+ Algorithm parameters

42

▼ Algorithm parameters

General Parameters

Max target sequences: 100
Select the maximum number of aligned sequences to display ⓘ

Short queries: Automatically adjust parameters for short input sequences ⓘ

Expect threshold: 10

Word size: 28 ⓘ

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 1,-2 ⓘ 配对与错配

Gap Costs: Linear ⓘ 空位罚分

Filters and Masking

Filter: Human

Mask: Mask lower case letters ⓘ

BLAST Search database Human G+T using Megablast (Optimize for highly similar sequences)
 Show results in a new window

blastp

►NCBI/BLAST/blastp suite

Standard Protein BLAST

blastn blastp blastx tblastn tblastx

Enter Query Sequence BLASTP programs search protein databases using a protein query. [more...](#) [Reset page](#) [Bookmark](#)

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)

Query subrange [?](#)
From
To

Or, upload file 浏览... [?](#)

Job Title Enter a descriptive title for your BLAST search [?](#)

Align two or more sequences [?](#)

Choose Search Set

Database: Non-redundant protein sequences (nr) [?](#)

Organism Optional
Enter organism name or id-completions will be suggested Exclude [+](#)
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown. [?](#)

Exclude Optional
 Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional
Enter an Entrez query to limit search [?](#)

Program Selection

Algorithm: blastp (protein-protein BLAST) PSI-BLAST (Position-Specific Iterated BLAST) PHI-BLAST (Pattern Hit Initiated BLAST) DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm [?](#)

Blastp算法选择

BLAST

Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)

Show results in a new window

[Algorithm parameters](#)

Algorithm parameters

General Parameters

Max target sequences: 100 Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62 (selected)

Gap Costs: Extension: 1

Compositional adjustments: BLOSUM62, BLOSUM80, BLOSUM45

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only, Mask lower case letters

BLAST Search database Non-redundant protein sequences (nr) using Blastp (protein-protein BLAST)
 Show results in a new window

打分矩阵:

- PAM30
- PAM70
- BLOSUM80
- BLOSUM62
- BLOSUM45

选择打分矩阵 (scoring matrix)

	A	C	D	E	F	G	H →
A	4	0	-2	-1	-2	0	-2
C	0	9	-3	-4	-2	-3	-3
D	-2	-3	6	2	-3	-1	-1
E	-1	-4	2	5	-3	-2	0
F	-2	-2	-3	-3	6	-3	
G	0	-3	-1	-2	-3		
H	-2	-3	-1	0			

↓
BLOSUM 62

The PAM family

- Based on **global alignments**
- The PAM1 is the matrix calculated from comparisons of sequences with **no more than 1% divergence**.
- Other PAM matrices are extrapolated from PAM1.

The BLOSUM family

- Based on **local alignments**.
- BLOSUM62 is a matrix calculated from comparisons of sequences with **no less than 62% divergence**.
- All BLOSUM matrices are **based on observed alignments**; they are not extrapolated from comparisons of closely related proteins.

BLOSUM 80

PAM 1

Less divergent

BLOSUM 62

PAM 120



BLOSUM 45

PAM 250

More divergent

blastn结果

► NCBI/BLAST/blastn suite/Formatting Results - XD5X1JDU016

[Edit and Resubmit](#) [Save Search Strategies](#) [►Formatting options](#) [►Download](#)

lesson.seq.screen.Contig34

Query ID lcl|30513
Description lesson.seq.screen.Contig34
Molecule type nucleic acid
Query Length 1093

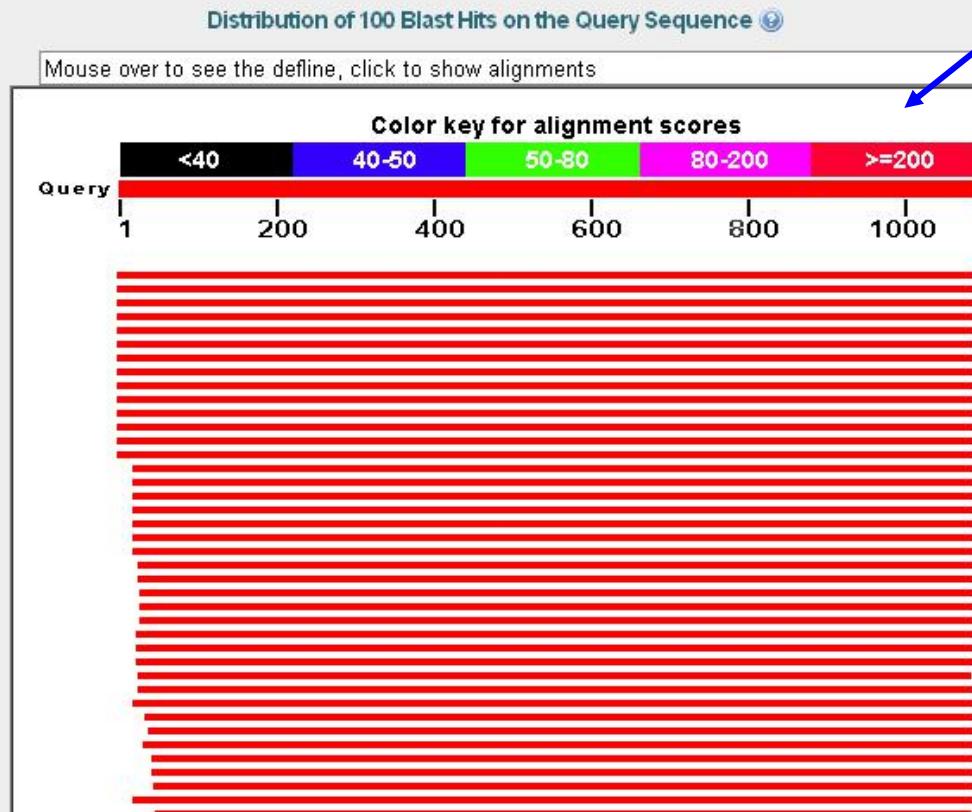
检索序列信息

Database Name nr
Description All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, environmental samples or phase 0, 1 or 2 HTGS sequences)
Program BLASTN 2.2.25+ [►Citation](#)

比对的数据库信息

Other reports: [►Search Summary](#) [\[Taxonomy reports\]](#) [\[Distance tree of results\]](#)

▼ Graphic Summary



blastn结果

Descriptions

Legend for links to other resources: UniGene GEO Gene Structure Map Viewer PubChem BioAssay

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E-value	Max ident	Links
BC000214.1	Homo sapiens guanine nucleotide binding protein (G protein)	1977	1977	99%	0.0	99%	
BC000366.2	Homo sapiens guanine nucleotide binding protein (G protein)	1973	1973	99%	0.0	99%	
BC010119.2	Homo sapiens guanine nucleotide binding protein (G protein)	1973	1973	99%	0.0	99%	
BC002620.1	Homo sapiens mRNA similar to guanine nucleotide binding pr	1971	1971	99%	0.0	99%	
BC014256.1	Homo sapiens guanine nucleotide binding protein (G protein)	1971	1971	99%	0.0	99%	
BC017287.2	Homo sapiens guanine nucleotide binding protein (G protein)	1969	1969	99%	0.0	99%	
BC000672.1	Homo sapiens cDNA clone IMAGE:3350045, **** WARNING	1960	1960	99%	0.0	99%	
BC073781.1	Homo sapiens guanine nucleotide binding protein (G protein)	1960	1960	99%	0.0	99%	
BC021034.1	Homo sapiens cDNA clone IMAGE:3834363, **** WARNING	1960	1960	99%	0.0	99%	
BC019093.2	Homo sapiens guanine nucleotide binding protein (G protein)	1960	1960	99%	0.0	99%	
BC019362.1	Homo sapiens guanine nucleotide binding protein (G protein)	1960	1960	99%	0.0	99%	
BC032006.1	Homo sapiens guanine nucleotide binding protein (G protein)	1960	1960	99%	0.0	99%	
NM_006098.4	Homo sapiens guanine nucleotide binding protein (G protein)	1954	1954	99%	0.0	99%	
AY159316.1	Homo sapiens lung cancer oncogene 7 mRNA, complete cds	1954	1954	99%	0.0	99%	
BC021993.1	Homo sapiens guanine nucleotide binding protein (G protein)	1947	1947	97%	0.0	99%	
CR607176.1	full-length cDNA clone CS0DA001YP03 of Neuroblastoma of	1947	1947	97%	0.0	99%	
CR603958.1	full-length cDNA clone CLOBB005ZG08 of Neuroblastoma of	1947	1947	97%	0.0	99%	
CR602411.1	full-length cDNA clone CLOBB012ZF09 of Neuroblastoma of	1947	1947	97%	0.0	99%	
CR601263.1	full-length cDNA clone CS0DI076YI19 of Placenta Cot 25-n	1947	1947	97%	0.0	99%	
CR600581.1	full-length cDNA clone CLOBB012ZC07 of Neuroblastoma of	1947	1947	97%	0.0	99%	
CR594655.1	full-length cDNA clone CS0DI053YK19 of Placenta Cot 25-n	1947	1947	97%	0.0	99%	
CR625098.1	full-length cDNA clone CS0DE012YB19 of Placenta of Homo	1940	1940	96%	0.0	99%	
CR604224.1	full-length cDNA clone CS0DH004YB11 of T cells (Jurkat cel	1940	1940	96%	0.0	99%	
CR616532.1	full-length cDNA clone CS0DE003YP09 of Placenta of Homo	1938	1938	96%	0.0	99%	
CR613365.1	full-length cDNA clone CS0DE012YI17 of Placenta of Homo	1938	1938	96%	0.0	99%	
CR594601.1	full-length cDNA clone CL0BB009ZE04 of Neuroblastoma of	1938	1938	96%	0.0	99%	

E值 (E-value) 表示仅仅因为随机性造成获得这一比对结果的可能性。这一数值越接近零，随机发生这一事件的可能性越小，结果可靠性越高。

blastn结果

▼ Alignments

Select All [Get selected sequences](#) [Distance tree of results](#)

>**gb|BC000214.11** **UEG** Homo sapiens guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1, mRNA (cDNA clone MGC:2416 IMAGE:2959178), complete cds
Length=1137

GENE ID: 10399 GNB2L1 | guanine nucleotide binding protein (G protein), beta polypeptide 2-like 1 [Homo sapiens] ([Over 100 PubMed links](#))

Score = 1977 bits (1070), Expect = 0.0
Identities = 1083/1088 (99%), Gaps = 5/1088 (0%)
Strand=Plus/Minus

Query	Subject	Sequence	Length
1	1137	tttttttttttttttagtgcagttttttttTTTGTAAAGCTCTGCCATAAACT	60
61	1079	TTTTTTTTTTTT--TGCCAGTTTTTTTTATTGTAAAGCTCTGCCATAAACT	1080
61	1079	TCTAGCGTGTGCCAATGGTCACCTGCCACACTCGCACCGAGGTTGCCAGCAA	120
61	1079	TCTAGCGTGTGCCAATGGTCACCTGCCACACTCGCACCGAGGTTGCCAGCAA	1020
121	1019	ACAGAGTCTGCCATCAGCAGACCAGGCCAGGGAGGTGCACTGGGGTGGTTCTGCCTTGC	180
121	1019	ACAGAGTCTGCCATCAGCAGACCAGGCCAGGGAGGTGCACTGGGGTGGTTCTGCCTTGC	960
181	959	TGCTGGTACTGATAACTCTTGTTCAGTTCATCTACAATGATCTTCCCTCTAAATCCC	240
181	959	TGCTGGTACTGATAACTCTTGTTCAGTTCATCTACAATGATCTTCCCTCTAAATCCC	900
241	899	AGATCTTGTGCTGGGGCCTGTGG-AGCACACAGCCAGTAGCGGTTAGGGCTGAAGCACA	299
241	899	AGATCTTGTGCTGGGGCCTGTGGCAGCACACAGCCAGTAGCGGTTAGGGCTGAAGCACA	840
300	839	GGGC GTT GAT GAT GTCCCC ACC AT CTAGCGTGTAAAGGTGTTGCCCTCGTTGAGATCCC	359
300	839	GGGC GTT GAT GAT GTCCCC ACC AT CTAGCGTGTAAAGGTGTTGCCCTCGTTGAGATCCC	780
360	779	ATAACATGGCCTGGCCATCTTGCCTCCAGAACAGCACAGAGGGATCCATCTGGAGAGACAG	419
360	779	ATAACATGGCCTGGCCATCTTGCCTCCAGAACAGCACAGAGGGATCCATCTGGAGAGACAG	720
420	719	TCACCGTGTTCAGATAGCCTGTGGCCAATGTGGTTGGTCTTCAGCTTGCAGTTAGCCA	479
420	719	TCACCGTGTTCAGATAGCCTGTGGCCAATGTGGTTGGTCTTCAGCTTGCAGTTAGCCA	660
480	659	GGTTCCATACCTTGACCAGCTTGTCCCAGCCACAGGAGACGATGATAGGGTTGCTGCTGT	539
480	659	GGTTCCATACCTTGACCAGCTTGTCCCAGCCACAGGAGACGATGATAGGGTTGCTGCTGT	600

练习1：网上运行blastx和blastn

(NCBIblast网址：<http://blast.ncbi.nlm.nih.gov/>)

>lesson. seq. screen. Contig34

```
TTTTTTTTTTTTTTAGTGCCAGTTTTTTATTGTAAAGCTCTGCCATAAACCTCTAGCGTGTGCCAATGGTCACCTGCCACA  
CTCGCACCAGGTTGTCCGTAGCCAGCAAACAGAGTCTGCCATCAGCAGACCAGGCCAGGGAGGTGCACTGGGTGGTTCTGCCTTGC  
TGCTGGTACTGATAACTTCTGCTTCAGTTCATCTACAATGATCTTCCCTCTAAATCCCAGATCTTGATGCTGGGCCTGTGGAGCACA  
CAGCCAGTAGCGGTTAGGGCTGAAGCACAGGGCGTTGATGATGTCCCCACCATCTAGCGTGTAAAGGTGTTGCCTCGTTGAGATCCA  
TAACATGGCCTGGCCATCCTGCCAGAACAGCACAGAGGGATCCATCTGGAGAGACAGTCACCGTGTTCAGATAGCCTGTGGCCAAT  
GTGGTTGGTCTTCAGCTGCAGTTAGCCAGGTTCCATACCTTGACCAGCTTGCCAGGCCACAGGAGACGATGATAGGTTGCTGCTGTT  
GGGCGAGAACGGACACAAGACACCCACTCTGAGTGGCTCTCATCCTGGACAGTGTATTGCACACACCCAGGGTATTCCATAGCTGAT  
GGTTTTATCTCGAGATCCAGAGACAATCTGCCGGTTGTCAGAGGAGAAGGCCACACTCAGCACATCCTGGTATGCCCAAAATCGCCT  
CGTGGTGGTGCCCGTTGAGATCCCAGAAGGCCAGGGTTCCATCCAGGAGCCTGAGAGGGCAAACCTGCCATCTGAGGAGATAACCA  
CATCACTAACAAAGTGGGAGTGACCCCGCAGAGCACGCTGTGAATTCCATAGTGGTCTCATCCCTGGTCAGTTCCACATGATGATGG  
TCTTATCTCGAGAGGCGGAGAGGATCATGTCGGGAAC TGCGGGTAGTAGCGATCTGGTTACCCAGCCGTGTGGCCCTGAGGGTGC  
CACGAAGGTCACTGCTCAGTCATGGCGCGGAGAGCGTGTTCGCTGCAGCGACGAGGATGGCACTGGATGGCTTAGAGAAACTAGC  
ACCACAGTCGACC
```

1. 对contig34进行网上blastn（演示），
2. blastx（自行操作）比对

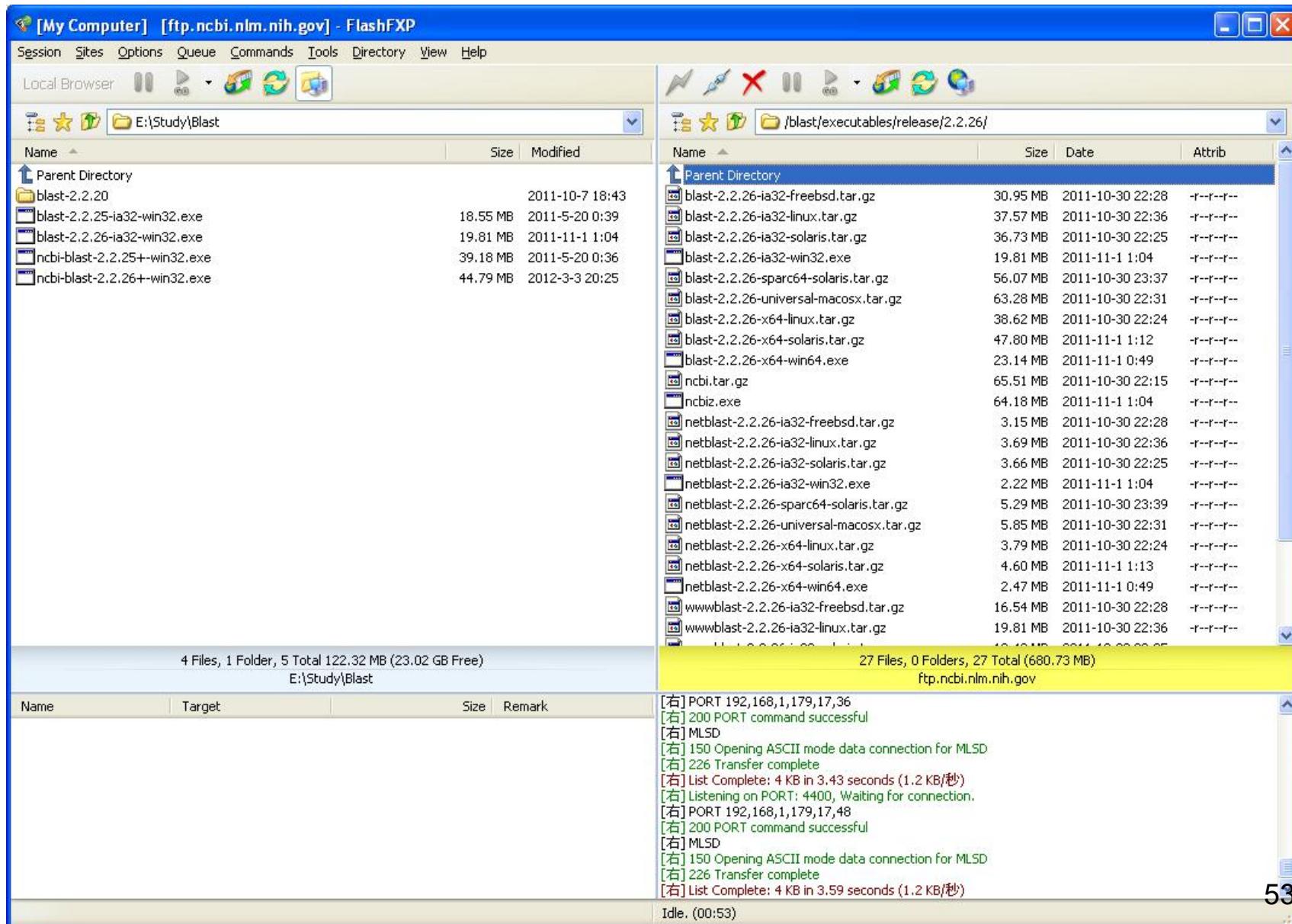
网页版BLAST的优缺点：

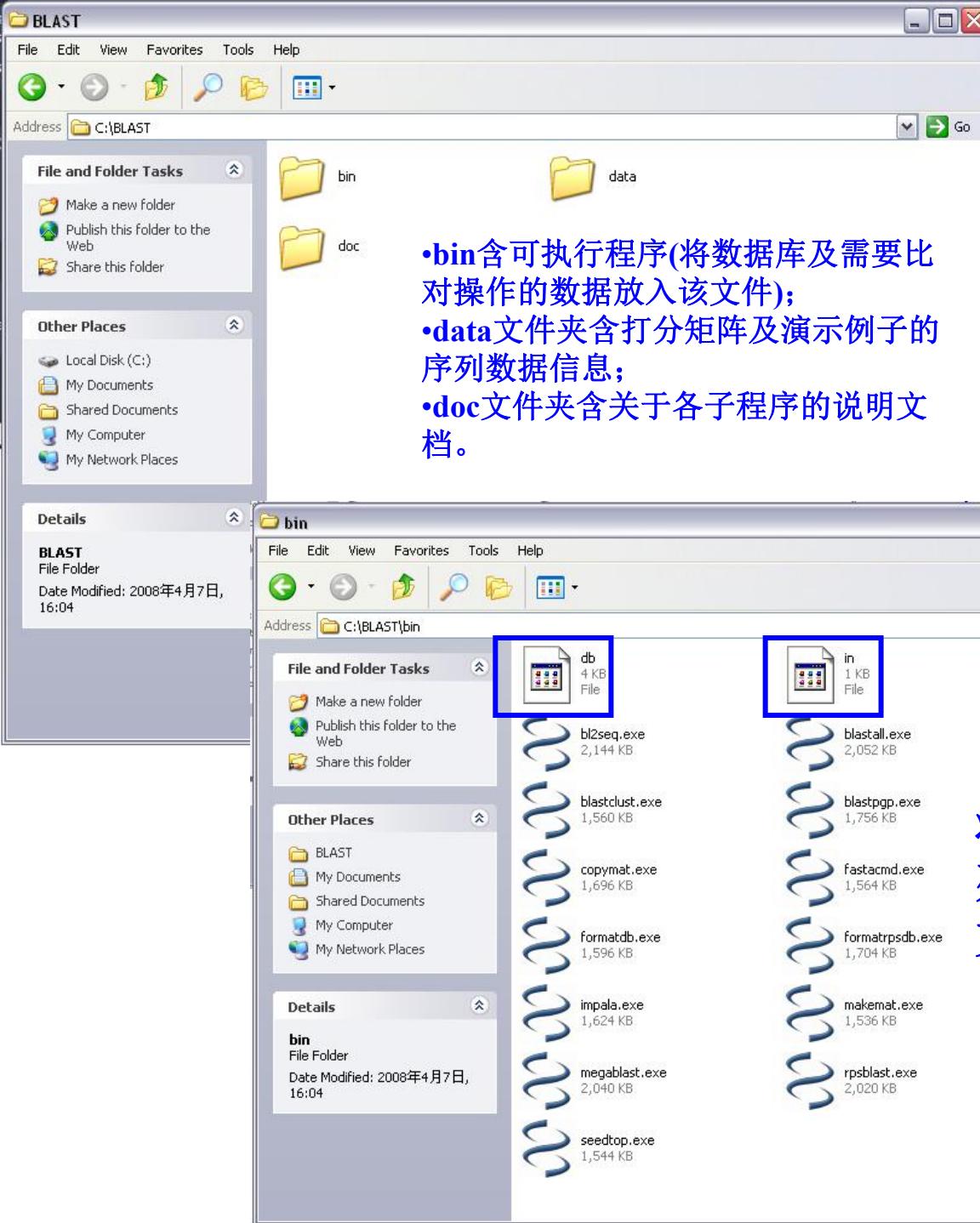
- 优点：直观方便，容易操作，数据库同步更新
- 缺点：不利于操作大批量的数据，同时也不能自己定义搜索的数据库，对网络依赖性太大。

本地运行BLAST

- 下载NCBI上blast程序:
- <ftp://ftp.ncbi.nlm.nih.gov/blast/executables/release/>
- 安装（安装到C: \blast）
- 数据库的格式化（formatdb）
- 程序运行（blastall）

登陆NCBI的FTP下载blast程序





双击安装到C盘
产生三个文件夹

- bin含可执行程序(将数据库及需要比对操作的数据放入该文件);
- data文件夹含打分矩阵及演示例子的序列数据信息;
- doc文件夹含关于各子程序的说明文档。

将数据库文件(db)及目标序列文件(in)保存在Blast/bin文件夹下

本地数据库的构建

由**fasta**格式的序列组成，以“>”开头，紧接着是序列描述信息，换行后即是核苷酸或蛋白质序列，直至下一个“>”前为止。

- 查看db文件

```
>gi|9506859|ref|NP_061932.1| 6.2 kd protein [Homo sapiens]
MVKLSKEAKQRLQQLFKGSQFAIRWGFPIPLVIYLGFKRGADPGMPEPTVLSLLWG
>gi|5441539|emb|CAB46824.1| Ribosomal protein [Canis familiaris]
EIANANSRQQILKLIKDGIIRKPVTVHSRARCRKNTLARRKGRHMGIGKRKGTAARMPEKVTVWMRRMR
ILRRLLTRYRESKKIDRHYHSLYLVKGKGVFKNKRILMEIH
>gi|13929184|ref|NP_114016.1| potassium large conductance calcium-activated channel, subfamily M, alpha member 1 [Rattus norvegicus]
MANGGGGGGGSSGSSGGGGGGGETALRMSSNIHANHLSLDASSSSSSSSSSSSSSSSSSSSVHEPKMDAL
IIPVTMEVPCDSRGQRMWVAFLASSMVTFGGLFILLWRTLKYLWTVCCCHCGGKTKEAQKINNGSSQAD
GTLKPVDEKEEVVAAEVGWMTSVKDAGVMISAQTLTGRVLVVLVFALSIGALVIYFIDSSNPIESCQNF
YKDFTLQIDMAFNVFFLLYFGLRFIAANDKLWFWLEVNSVVDFTVPPVFVSVYLNRSWLGLRFLRALRL
IQFSEILQFLNLKTSNSIKLVNLLSIFISTWLTAAGFIHLVENSGDPWENFQNNQALTYWECVYLLMVT
MSTVGYGDVYAKTTLGRFLMVFFILGGAMFASYVPEIEELIGNRKKYGGSYSAVSGRKHIVVCGHITLE
SVSNFLKDFLHKDRDDVNVEIVFLHNISPNELEALFKRHFTQVEFYQGSVLPNPHDLARVKIESADA CLI
LANKYCADCDAEDASNIMRVISIKNYHPKIRITQMLQYHNKAHLLNIPSWNWKEGDDAICLAEKLKGFI
AQSCLAQGLSTMLANLFSMRSFIKEEDTWQKYYLEGVSNEYTEYLSSAFVGLSPTVCELCFVKLKL
MIAIEYKSANRESRSRKRILINPGNHLKTQEGLGFFIASDAKEVKRAFFYCKACHDDVTDPKRIKKCGC
RRLIYSKMSIYKRAMSRACCFDCGRSERDCSCMSGRVRGNVDTLERNFPLSSVSVNDCTSFRAFEDEQPP
TLSPKKKQRNGGMRNSPNTSPKLMRHDPLIIPGNDQIDNMDSNVKKYDSTGMFHWCAPKEIEKVILTRSE
AAMTVLSGHVVVCIFGDVSSALIGRLNLMPLRASNPHYHELKHFVFGSIEYLREWETLHNFPKV SIL
PGTPLSRADLRNAVNILCDMCVILSANQNNIDDTSLQDKECILASNIKSMQFDDSIGVLQANSQGFTP P
GMDRSSPDNSPVHGMLRQPSITVGNIPIITELVNDTNQFLDQDDDDDPTELYLTQPFACGTAFAVSV
LDSLMSATYFNDNILTIRTLVTGGATPELEALIAEENALRGGYSTPQTLANRDRCRVAQLALLDGP FAD
LGDGGCYGDLFCALKTYNMLCFGYRLRDAHLSTPSQCTKRYVITNPPYEFELVPTDLIFCLMQFDHNA
GQSRASLSHSSHSSQSSKKSSVHSIPSTANRPNRPKSRESRDQKKEMVYR
>gi|627367|pir||A45259 desmoyokin - human (fragments)
MPGIKVGGSVNVNAKGQLDLGGRRGGVQVPAVDISSLGGGRPVEVQGPSLESGDHAKIKFPTMKVPKFGVS
TGREGQTPKAGLRSAPPEVSVGHKGKGPKLTIQAPQLEVSVP SANIEGLEGKLKGPOITGPSLEGDGLK
GAKPQGHIGVDASAPQIGGSITGPSVEVQAPDIDVQGPGSKLNVPKMKVPKFSVSGAKGEETGIDVTLPT
GEVTVPGVSGDVS LPEIATGGLEGKMKGKTKVTP EMIIQKPKISMQDV DLSLGSPKLKGDIKVSAPGVQG
DVKGPOVALKGSRV D IETPNLEG TLGPR LGSPSGK TGT CRISM SEV DLNVAAPKVKG GV D VTL PR VEGK
```

数据库的格式化

formatdb命令用于数据库的格式化：

formatdb [option1] [option2] [option3]…

formatdb常用参数

-i database_name 需要格式化的数据库名称

-p T\F 待格式化数据库的序列类型

(核苷酸选F； 蛋白质选T； 默认值为T)

例： **formatdb -i db -p T**

对蛋白质数据库 “**db**”进行格式化

程序运行

blastall命令用于运行五个blast子程序：

blastall [option1] [option2] [option3]

*可在dos下输入blastall查看各个参数的意义及使用

- blastall常用参数

- 四个必需参数

- p **program_name**, 程序名，根据数据库及搜索文件序列性质进行选择；

- d **database_name**, 数据库名称，比对完成格式化的数据库；

- i **input_file**, 搜索文件名称；

- o **output_file**,BLAST结果文件名称；

- 两个常用参数

- e **expectation**, 期待值, 默认值为10. 0, 可采用科学计数法来表示, 如 $2e-5$;

- m **alignment view options**: 比对显示选项, 其具体的说明可以用以下的比对实例说明

例：blastall -p blastx -d db -i in -o out -e 2e-5 -m 9 (表格显示比对结果)

采用blastx程序，将in中的序列到数据库db中进行比对，结果以表格形式输入到out文件

练习2:本地运行blastx

- 进入DOS命令行提示符状态 (“运行” → 输入cmd)
- 进入C盘， 输入： cd\
- 进入包含序列数据的bin目录下， 输入： cd blast\bin
- 查看目录下内容， 输入： dir
- 格式化数据库db： formatdb -i db -p T
 ↑
 输入 数据库类型： F/T
- 运行blastx
 - blastall -p blastx -i in -d db -o out -e 2e-5 -m 9
 ↑
 Blast程序 序列输入 数据库 结果输出
- 查看结果： 用写字板或者记事本打开out文件

out - Notepad

File Edit Format View Help

BLASTX 2.2.24 [Aug-08-2010]

Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

Query= lesson.seq.screen.Contig12
(819 letters)

Database: db
4 sequences; 3561 total letters

Searching.....done

Sequences producing significant alignments:

Score (bits)	E Value
-----------------	------------

gi|13929184|ref|NP_114016.1| potassium large conductance calcium... 297 e-151

>gi|13929184|ref|NP_114016.1| potassium large conductance
calcium-activated channel, subfamily M, alpha member 1
[Rattus norvegicus]
Length = 1243

Score = 297 bits (760), Expect(2) = e-151
Identities = 143/146 (97%), Positives = 144/146 (98%)
Frame = +1

Query: 382 G DW H I V V C G H I T L E S V S N F L K D F L H K D R D D V N V E I V F L H N I S P N L E A L F K R H F T Q V E F 561
G H I V V C G H I T L E S V S N F L K D F L H K D R D D V N V E I V F L H N I S P N L E A L F K R H F T Q V E F

Sbjct: 407 G R K H I V V C G H I T L E S V S N F L K D F L H K D R D D V N V E I V F L H N I S P N L E A L F K R H F T Q V E F 466

Query: 562 Y Q G S V L N P H D L L A R V K I E S A D A C L I L A N K Y C A D P D A E D A S N I M R V I S I K N Y H P K I R I I T Q M 741
Y Q G S V L N P H D L L A R V K I E S A D A C L I L A N K Y C A D P D A E D A S N I M R V I S I K N Y H P K I R I I T Q M
Sbjct: 467 Y Q G S V L N P H D L L A R V K I E S A D A C L I L A N K Y C A D P D A E D A S N I M R V I S I K N Y H P K I R I I T Q M 526

Query: 742 L Q Y H N K A H L L N M P S W N W K E G D D A I C L 819
L Q Y H N K A H L L N + P S W N W K E G D D A I C L
Sbjct: 527 L Q Y H N K A H L L N I P S W N W K E G D D A I C L 552

Score = 246 bits (629), Expect(2) = e-151
Identities = 123/127 (96%), Positives = 124/127 (97%)
Frame = +3

Query: 12 F L T V F P V F V S V Y S N R S W L G L R F L R A L R L I Q F S E I L Q F L N I L K T S N S I K L V N L L S I F I S T W 191
F T V P V F V S V Y M R S W L G L R F L R A L R L I Q F S E I L Q F L N I L K T S N S I K L V N L L S I F I S T W

3. 蛋白质序列数据库

- SWISS-PROT(欧洲)
- PIR(美国)

- **Protein Sequence Databases**
- [UniProt: United Protein Databases](#)

A single database that combines the information of the major international databases, European Bioinformatics Institute (EBI), Cambridge, UK; Protein Information Resource (PIR) - Georgetown University Medical Center (GUMC) & National Biomedical Research Foundation (NBRF), Washington, D.C.; and Swiss Institute of Bioinformatics (SIB) - Geneva, Switzerland. "The Universal Protein Resource (UniProt) provides the scientific community with a single, centralized, authoritative resource for protein sequences and functional information."

- [PIR Protein Sequence Database](#)

The database is described by its sponsor as "functionally annotated protein sequences, which grew out of the Atlas of Protein Sequence and Structure (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools." From the Protein Information Resource, the major U.S. source of protein informatics.

- [Swiss-Prot](#)

The major European protein sequence database, with accompanying annotations, from the Swiss Institute of Bioinformatics. "Swiss-Prot is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases." Also at this site is TrEMBL, which contains all translated nucleic acid protein coding sequences in EMBL that have not yet been annotated and incorporated into Swiss-Prot.

- SWISS-PROT
 - 只收录实际存在的蛋白质，有详细的注释（包括功能、结构域、翻译后的修饰等）及齐全的引文和到其它数据库的链接。
 - <http://www.expasy.org/sprot/>
 - <ftp://ftp.expasy.ch/databases/swiss-prot/>
- TrEMBL
 - 从EMBL库中的核酸序列翻译出来的氨基酸序列，已经完成自动注释。其中SP-TrEMBL条目已由专家人工分类并赋予SWISS-PROT索引号，但未通过人工审读被最终收入SWISS-PROT。
 - SWISS-PROT+TrEMBL非冗余库
 - <http://www.expasy.ch/sprot/>
 - ftp://ftp.expasy.ch/databases/sp_tr_nrdb/

• SWISS—PROT

1. 瑞士日内瓦大学医学生物化学系和欧洲生物信息学研究所(EBI)合作维护（1986年）；
2. 在EMBL和GenBank数据库上均建立了镜像站点；
3. 数据库包括了从EMBL翻译而来的蛋白质序列，这些序列经过检验和注释；
4. 数据记录包括两部分：
 序列
 注释(结构域、功能位点、跨膜区域、二硫键位置、翻译后的修饰、突变体等)
5. 数据存在滞后性 ➡ TrEMBL数据库的建立

SWISS-PROT的网址： <http://cn.expasy.org/sprot>

TrEMBL的网址： <http://www.ebi.ac.uk/trembl/index.html>

SWISS-PROT (<http://www.expasy.ch/sprot/sprot-top.html>)
是目前国际上比较权威的蛋白质序列数据库,其中的蛋白
质序列是经过注释的

SWISS-PROT中的数据来源于不同源地:

- (1) 从核酸数据库经过翻译推导而来;
- (2) 从蛋白质数据库PIR挑选出合适的数据;
- (3) 从科学文献中摘录;
- (4) 研究人员直接提交的蛋白质序列数据

SWISS-PROT有三个明显的特点 :

(1) 注释

在SWISS-PROT中，数据分为核心数据和注释两大类。

核心数据包括：

序列数据、参考文献、分类信息（蛋白质生物来源的描述）

注释包括：

- (A) 蛋白质的功能描述；
- (B) 翻译后修饰；
- (C) 域和功能位点，如钙结合区域、ATP结合位点等；
- (D) 蛋白质的二级结构；
- (E) 蛋白质的四级结构，如同构二聚体、异构三聚体等；
- (F) 与其它蛋白质的相似性；
- (G) 由于缺乏该蛋白质而引起的疾病；
- (H) 序列的矛盾、变化等。

(2) 最小冗余

- 尽量将相关的数据归并，降低数据库的冗余程度。
- 如果不同来源的原始数据有矛盾，则在相应序列特征表中加以注释。

(3) 与其它数据库的连接

- 对于每一个登录项，有许多指向其它数据库相关数据的指针，这便于用户迅速得到相关的信息。
- 现有的交叉索引有：
 - 到EMBL核酸序列数据库的索引，
 - 到PROSITE模式数据库的索引，
 - 到生物大分子结构数据库PDB的索引等。

TrEMBL (<http://www.ebi.ac.uk/trembl/index.html>) 是与SWISS-PROT相关的一个数据库。

包含从EMBL核酸数据库中根据编码序列(CDS)翻译而得到的蛋白质序列，并且这些序列尚未集成到SWISS-PROT数据库中。

TrEMBL有两个部分：

(1) SP-TrEMBL(SWISS-PROT TrEMBL)

包含最终将要集成到SWISS-PROT的数据，所有的SP-TrEMBL序列都已被赋予SWISS-PROT的登录号。

(2) REM-TrEMBL(REMaining TrEMBL)

包括所有不准备放入SWISS-PROT的数据，因此这部分数据都没有登录号。

TrEMBL - Microsoft Internet Explorer

文件(E) 编辑(E) 查看(V) 收藏(A) 工具(I) 帮助(H)

后退 前进 停止 刷新 复制 剪切 粘贴 主页 打印 打印预览 邮件 编辑 讨论

地址(D) <http://www.ebi.ac.uk/tremb/index.html>

Get Nucleotide sequences for Site search Go ? Norton AntiVirus

 Site Map SRS Start Session

EBI Home About EBI Research Services Toolbox Databases Downloads Submissions TrEMBL DATABASE

TrEMBL

The TrEMBL database contains the translations of all coding sequences (CDS) present in the [EMBL](#) Nucleotide Sequence Database, which are not yet integrated into [Swiss-Prot](#).

TrEMBL is split into two main sections .

- **SP-TrEMBL** (Swiss-Prot TrEMBL)
Contains the entries which should eventually be incorporated into Swiss-Prot and can be considered as a preliminary section of Swiss-Prot as all SP-TrEMBL entries have been assigned Swiss-Prot accession numbers.
- **REM-TrEMBL** (REMaining TrEMBL)
Contains the entries that we do not want to include in Swiss-Prot. REM-TrEMBL entries have no accession numbers.

The TrEMBL group is headed by: [Rolf Apweiler](#).

The current TrEMBL Release is version 24.12 as of 12-Sep-2003, and contains 938390 entries... [more stats](#)

TrEMBL Release 23.0 of 04-Mar-2003 contained 830525 entries... [more stats](#)

[Access the TrEMBL Database](#)

Swiss-Prot

 The Swiss-Prot protein knowledge-base is a curated protein sequence database that provides a high level of annotation, a minimal level of redundancy and high level of integration with other databases.

Nucleotide DB

 The EMBL Nucleotide Sequence Database constitutes Europe's primary nucleotide sequence resource.

69



Search Blast Align Retrieve ID Mapping *

Search in Query

Protein Knowledgebase (UniProtKB) hbsag Search Advanced Clear

1 - 25 of 2,396 results for hbsag in UniProtKB sorted by score descending

[Download](#)

Browse by taxonomy, keyword, gene ontology, enzyme class or pathway |

Reduce sequence redundancy to 100%, 90% or 50%

Page 1 of 96 | [Next](#)

Results [Customize](#)

- › Show only [reviewed](#) (86) ★(UniProtKB/Swiss-Prot) or [unreviewed](#) (2,310) ★(UniProtKB/TrEMBL) entries
- › Did you mean [bsag](#) (4,791)?
- › Restrict term "hbsag" to gene name (184), protein name (1,228), strain (5), taxonomy (5), tissue (5)

Entry	Entry name	Status	Protein names	Gene names	Organism	Length
P03138	HBSAG_HBVD3	★	Large envelope protein	S	Hepatitis B virus genotype D subtype ayw (isolate France/Tiollais/1979) (HBV-D)	389
P03141	HBSAG_HBVA3	★	Large envelope protein	S	Hepatitis B virus genotype A2 subtype adw2 (strain Rutter 1979) (HBV-A)	400
Q9E6S4	HBSAG_HBVC0	★	Large envelope protein	S	Hepatitis B virus genotype C (isolate Vietnam/3270/2000) (HBV-C)	400
P31869	HBSAG_HBVC2	★	Large envelope protein	S	Hepatitis B virus genotype C subtype ar (isolate Japan/S-207/1988) (HBV-C)	400
P03140	HBSAG_HBVC5	★	Large envelope protein	S	Hepatitis B virus genotype C subtype ad (isolate Japan/S-179/1988) (HBV-C)	400
P31873	HBSAG_HBVA1	★	Large envelope protein	S	Hepatitis B virus genotype A1 subtype adw2 (isolate Southern-Africa/Cai) (HBV-A)	400
Q91C35	HBSAG_HBVA6	★	Large envelope protein	S	Hepatitis B virus genotype A1 subtype adw2 (isolate South Africa/84/2001) (HBV-A)	389
Q91534	HBSAG_HBVA7	★	Large envelope protein	S	Hepatitis B virus genotype A2 (isolate Japan/11D11HCCW/1998) (HBV-A)	400
Q9PWW3	HBSAG_HBVB5	★	Large envelope protein	S	Hepatitis B virus genotype B2 (isolate Vietnam/16091/1992) (HBV-B)	400
P31868	HBSAG_HBVC1	★	Large envelope protein	S	Hepatitis B virus genotype C subtype adr (isolate Japan/Nishioka/1983) (HBV-C)	400
Q913A6	HBSAG_HBVC7	★	Large envelope protein	S	Hepatitis B virus genotype C subtype ayw (isolate China/Tibet127/2002) (HBV-C)	400
Q81162	HBSAG_HBVC8	★	Large envelope protein	S	Hepatitis B virus genotype C subtype adr (isolate Japan/A4/1994) (HBV-C)	389
Q9QMI0	HBSAG_HBVD4	★	Large envelope protein	S	Hepatitis B virus genotype D subtype ayw (isolate Japan/JYW796/1988) (HBV-D)	389
Q998M2	HBSAG_HBVD5	★	Large envelope protein	S	Hepatitis B virus genotype D subtype ayw (isolate Australia/AustKW/1991) (HBV-D)	389
Q92921	HBSAG_HBVD7	★	Large envelope protein	S	Hepatitis B virus genotype D (isolate Germany/1-91/1991) (HBV-D)	389
Q69603	HBSAG_HBVE1	★	Large envelope protein	S	Hepatitis B virus genotype E subtype ayw4 (isolate Kou) (HBV-E)	399
Q99HS3	HBSAG_HBVF3	★	Large envelope protein	S	Hepatitis B virus genotype F1 (isolate Argentina/sa11/2000) (HBV-F)	400

<http://www.uniprot.org/uniprot/?query=hbsag&sort=score>

- PIR (Protein Information Resource)
 - 国际蛋白质序列数据库，包含所有序列已知的自然界中野生型蛋白质信息。提供同源性和分类学组织的综合、非冗余的数据库。每周更新，每季度发行新版。
 - <http://pir.georgetown.edu/>
 - ftp://ftp.pir.georgetown.edu/pir_databases/
- UniProt
 - SWISS-PROT+TrEMBL+PIR
 - <http://www.ebi.uniprot.org/>
 - <ftp://ftp.ebi.ac.uk/pub/databases/uniprot/>

- **PIR(protein information resource)**
 1. 由美国NCBI翻译自GenBank的DNA序列(1984年);
 2. 在EMBL和GenBank数据库上均建立了镜像站点;
 3. 数据依据注释的质量分为4类。

网址: <http://www-nbrf.georgetown.edu/>

PIR数据库的分类情况(Release 51.03)

分类名称 (Name)	说明 (Comment)	记录数 (Number of entries)
PIR1	已分类、已注释 (Classified and annotated)	13572
PIR2	已注释(Annotated)	69368
PIR3	未核实(Unverified)	7508
PIR4	未翻译(Unencoded or untranslated)	196

PIR (Protein Information Resource)

- 目的：
帮助研究者鉴别和解释蛋白质序列信息，
研究分子进化、功能基因组。
- 它是一个全面的、经过注释的、非冗余的蛋白质序
列数据库。
- 所有序列数据都经过整理，超过99%的序列已按蛋
白质家族分类，一半以上还按蛋白质超家族进行
了分类。



PIR

Protein Information Resource

[About PIR](#)[Databases](#)[Search and Retrieval](#)[Download](#)[Support](#)

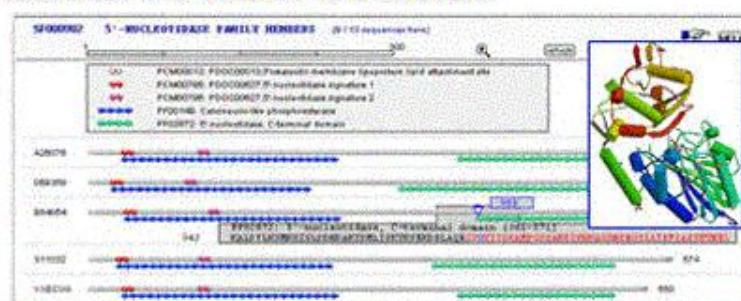
**AN INTEGRATED PUBLIC RESOURCE OF PROTEIN INFORMATICS TO SUPPORT
GENOMIC AND PROTEOMIC RESEARCH AND SCIENTIFIC DISCOVERY**

PIR produces the **Protein Sequence Database (PSD)** of functionally annotated protein sequences, which grew out of the *Atlas of Protein Sequence and Structure* (1965-1978) edited by Margaret Dayhoff and has been incorporated into an integrated knowledge base system of value-added databases and analytical tools.

ProClass, a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 50 biological databases. [Release 2.30, 8-Sep-2003, contains 1094,018 entries.](#)

PIR-NREF, a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. [Release 1.30, 8-Sep-2003, contains 1,332,284 entries.](#)

PIR has recently joined forces with [EBI](#) (European Bioinformatics Institute) and [SIB](#) (Swiss Institute of Bioinformatics) to establish the [UniProt](#) (United Protein Databases), the central resource of protein sequence and function.



PIR News Flash

PIR featured in Georgetown Blue & Gray newsletter

Text Search Protein Databases:

Find an Exact Peptide Match:

Type in a string of single letter amino acid code (at least 3 letters)



除了蛋白质序列数据之外，PIR还包含以下信息：

- (1) 蛋白质名称、蛋白质的分类、蛋白质的来源；
- (2) 关于原始数据的参考文献；
- (3) 蛋白质功能和蛋白质的一般特征，包括基因表达、翻译后处理、活化等；
- (4) 序列中相关的位点、功能区域。



ENTRY G00016 #type fragment
 TITLE FGF-receptor - common marmoset (fragment)
 ORGANISM #formal_name Callithrix jacchus #common_name common
 marmoset
 DATE 13-Mar-1997 #sequence_revision 13-Mar-1997 #text_change
 18-Jul-1997
 ACCESSIONS G00016
 REFERENCE H00018
 #authors Einspanier, R.
 #submission submitted to the EMBL Data Library, December 1995
 #accession G00016
 ##status preliminary; translated from GB/EMBL/DDBJ
 ##molecule_type mRNA
 ##residues 1-157 ##label EIN
 ##cross-references EMBL:268149; NID:gl279349
 CLASSIFICATION #superfamily basic fibroblast growth factor receptor 1;
 immunoglobulin homology; protein kinase homology
 FEATURE 3
 1-157 #domain protein kinase homology (fragment) #label
 KIN
 SUMMARY #length 157
 SEQUENCE
 5 10 15 20 25 30
 I/E M E U M K M I G K H K N I I N L L G A C T Q D G P L Y V I
 31 V E Y A S K G N L R E Y L R A R R P P G M E Y S Y D I N R V
 61 P E E Q M T F K D L V S C T Y Q L A R A M E Y L A S Q K C I
 91 H R D L A A R N V U L U T E N N U M K I A D F G L A R D I N N
 121 I D Y Y K K T T N G R L P V K W M A P E A L F D R U V Y T H Q
 151 S D V W S F G /

 Associated Alignments:
 DA0934 protein kinase homology
 DAL564 immunoglobulin homology - C2 type
 DAL565 immunoglobulin homology - V-type, Ig V regions
 FA1349 basic fibroblast growth factor receptor 1 - 555.0 1.0
 M06341 basic fibroblast growth factor receptor 1 - 524.0 1.0

 Related Links (Superfamily classification and Alignment):
 Protein Classification for Entry=G00016 at MIPS, Germany.
 ProClass for Entry=G00016 at Univ. of Texas, USA.

图 4.4 PIR 文件实例

PIR提供三种类型的检索服务：

一是基于文本的交互式查询，
用户通过关键字进行数据查询。

二是标准的序列相似性搜索，
包括BLAST、FastA等。

三是结合序列相似性、注释信息
和蛋白质家族信息的高级搜索，
包括按注释分类的相似性搜索、
结构域搜索等。

<http://pir.georgetown.edu/iproclass/>

<http://pir.georgetown.edu/pirwww/dbinfo/uniprot.shtml>

- PROSITE
 - 由专家审编的SWISS-PROT蛋白质序列中有生物意义的sites、patterns和profiles的数据库，可帮助确定新的蛋白质序列是否属于已知的家族。提供PrositeScan服务器搜索PROSITE库。
 - <http://www.expasy.ch/prosite/>
 - <ftp://au.expasy.org/databases/prosite/>
- ENZYME
 - 基于命名系统的酶数据库。可按酶的EC号、分类、学名和俗名、化合物、辅助因子等查询。每个条目下列出所催化的反应和酶的来源、功能等，并提供到其它数据库、MEDLINE和代谢途径图的链接。
 - <http://au.expasy.org/enzyme/>
 - <ftp://au.expasy.org/databases/enzyme/>

PROSITE

[Home](#) [ScanProsite](#) [ProRule](#) [Documents](#) [Downloads](#) [Links](#) [Funding](#)


Database of protein domains, families and functional sites

PROSITE consists of documentation entries describing protein domains, families and functional sites as well as associated **patterns** and **profiles** to identify them [[More details](#) / [References](#) / [Commercial users](#)].

PROSITE is complemented by **ProRule**, a collection of rules based on profiles and patterns, which increases the discriminatory power of profiles and patterns by providing additional information about functionally and/or structurally critical amino acids [[More details](#)].

Release 20.76, of 17-Oct-2011 (1627 documentation entries, 1308 patterns, 946 profiles and 944 ProRule)

PROSITE access

e.g: PDOC00022, PS50089, SH3, zinc finger

add wildcard (*)

Browse:

- by documentation entry
- by ProRule description
- by taxonomic scope
- by number of positive hit

PROSITE tools

- ScanProsite** - advanced scan
- PRATT** - allows to interactively generate conserved patterns from a series of unaligned proteins.

- MyDomains - Image Creator**  - allows to generate custom domain figures.



Scan a sequence against PROSITE patterns and profiles - quick scan

(Output includes graphical view and feature detection)



Enter your sequence or a [UniProtKB \(Swiss-Prot or TrEMBL\) ID or AC](#) [[help](#)]:

exclude patterns with a high probability of occurrence

<http://enzyme.expasy.org/>

4. 蛋白质结构数据库

显示分子结构 (**RasMol** , **ChemView**)

RasMol :Molecular Visualization Freeware for proteins, dna and macromolecules.

蛋白质结构数据库

- PDB
 - 蛋白质结构数据库，搜集由X射线衍射和核磁共振实验测定的生物大分子三维结构数据。
 - <http://www.rcsb.org/pdb/>
 - <ftp://ftp.rcsb.org/pub/pdb/>
- NRL-3D(*The NRL-3D database has been replaced with PDB.*)
 - 三维结构已经确定的蛋白质序列库，可以将新蛋白序列与库中序列比较，以判断是否与结构已知的蛋白质相似。
 - <http://pir.georgetown.edu/pirwww/dbinfo/nrl3d.html>
 - ftp://ftp.pir.georgetown.edu/pir_databases/other_databases/nrl_3d/

<http://www.pdb.org/pdb/home/home.do>

PDB (protein data bank)

- (1) 目前最主要的蛋白质分子结构数据库；
- (2) 1970年代建立，美国Brookhaven国家实验室维护管理；
- (3) 1988年，由美国RCSB(research collaborative for structural biology)管理；
- (4) 以文本格式存放数据，包括原子坐标、物种来源、测定方法、提交者信息、一级结构、二级结构等；
- (5) PDBsum数据库：PDB注释信息综合数据库，具有检索、分析、可视化的功能。

PDB的网址：<http://www.rcsb.org/pdb>(美国)

PDBsum的网址：<http://www.biochem.ucl.ac.uk/bsm/pdbsum>

PDB (Protein Data Bank)

PDB中含有通过实验（X射线晶体衍射，核磁共振NMR）测定的生物大分子的三维结构

- 蛋白质
- 核酸
- 糖类
- 其它复合物

- 隐式序列信息（*implicit sequence*）

PDB的隐式序列即为立体化学数据，包括每个原子的名称和原子的三维坐标。

- 显式序列信息（*explicit sequence*）

在PDB文件中，以关键字SEQRES作为显式序列标记，以该关键字打头的每一行都是关于序列的信息。

Contact Us | Print

PDB ID or Text

Search

[Advanced Search](#)

MyPDB Hide

[Login to your Account](#)
[Register a New Account](#)

Home Hide

[News & Publications](#)
[Usage/Reference Policies](#)
[Deposition Policies](#)
[Website FAQ](#)
[Deposition FAQ](#)
[Contact Us](#)
[About Us](#)
[Careers](#)
[External Links](#)
[Sitemap](#)
[New Website Features](#)

Deposition Hide

[All Deposit Services](#)
[Electron Microscopy](#)
[X-ray | NMR](#)
[Validation Server](#)
[BioSync Beamline](#)
[Related Tools](#)

Search Hide

[Advanced Search](#)
[Latest Release](#)
[New Structure Papers](#)
[Sequence Search](#)
[Chemical Components](#)
[Unreleased Entries](#)
[Browse Database](#)
[Histograms](#)

Tools Hide

[File Downloads](#)
[Compare Structures](#)
[FTP Services](#)
[File Formats](#)
[Services: RESTful | SOAP](#)
[Widacts](#)

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the [wwPDB](#), the RCSB PDB curates and annotates PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

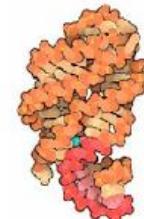
[Hide Welcome Message](#)

Featured Molecules (MotM Category View / Previous Features: MotM | PSI) Hide

Structural View of Biology



Protein
Synthesis



Molecule of the Month: Riboswitches

Why use two or more molecules when one will do? In our own cells, protein synthesis is controlled by thousands of regulatory proteins, which work together to decide when a particular protein will be made. Bacteria are masters of economy, however, and in some cases, they have figured out a way for messenger RNA to control itself, without the need for help by proteins.

[Full Article...](#)



Protein Structure Initiative Featured Molecule: Alpha and beta barrels

Analysis of the metabolic network of the bacterium *Thermatoga maritima* reveals the evolution of enzyme folding and function.

[Full Article](#) | [PSI Structural Biology Knowledgebase](#)

Latest Structures Hide

Customize This Page

New Features Hide

[PDBMobile: Save MyPDB Queries](#)

Latest features released:

[Website Release Archive](#)

RCSB PDB News Hide

[Weekly](#) | [Quarterly](#) | [Yearly](#)

2010-10-05

Latest Website Release

New and enhanced features are available, including improved navigation of the Molecule of the Month archive, PDBMobile for the iPhone, and new ways to explore search results using data distribution summaries.

[Poster Prize Awarded at ECM](#)

[Analyze small molecule interactions in the PDB with Ligand Explorer](#)

[Redesigned BioSync](#)

[Exploring Search Results](#)

wwPDB News Hide

[Statement on Retraction of PDB Entries](#)

2010-10-05

[Announcement: Chemical Shift Data Required for NMR Depositions Starting December 6, 2010](#)

[Upcoming Meeting: PDB Symposium at ACS](#)

- 转录因子数据库

RANSFAC <http://transfac.gbf.de>

ooTFD <http://www.ifti.org>

- 基因分类数据库

Gene Ontology (GO) <http://www.geneontology.org>

- 酶、代谢和调控路径数据库

KEGG <http://www.genome.ad.jp/kegg/>

Enzyme Nomenclature Database

<http://expasy.hcuge.ch/sprot/enzyme.html>

Protein Kinase Resource (PKR)

<http://www.sdsc.edu/kinases/>

- RNA数据库

miRBase <http://www.mirbase.org/>

mirna database <http://mirnadb.com/>

lncRNAdb <http://www.lncrna.org/Help.aspx>

Ensembl

Ensembl Genome Browser

http://www.ensembl.org/index.html

University at...o Libraries FlyBase Entrez Apple .Mac Amazon News eBay Yahoo! Google Scholar

Ensembl Genome Browser

e! Ensembl

Ensembl release 42 - Dec 2006

HOME · BLAST · BIOMART · SITEMAP · HELP

Your Ensembl

- Login or Register
- About User Accounts

Help & Documentation

- Table of Contents
- Helpdesk
- What's New
- About Ensembl
- Downloading data
- Displaying your own data
- Ensembl software

Select a species

- Mammals
- Other chordates
- Other eukaryotes

Ensembl Archive

e! View previous release of page in Archive!
e! Stable Archive! link for this page

Ensembl tools

- Start a sequence search → Search Ensembl for nucleotide and peptide sequences with BLAST and SSAHA.
- Mine Ensembl with BioMart → Cross-reference Ensembl datasets with BioMart, a powerful data-mining tool.
- Customise Your Ensembl → Register with Ensembl to bookmark your favourite pages, customise your home page and much more!
- Fetch data with the Ensembl API → Learn how to extract data from the public Ensembl database with this tutorial.

Search Ensembl

Search: All species for e.g. mouse chromosome 2 or X:10000..20000 or human gene BRCA2

Ensembl 42

Pre! species

Popular genomes

- Homo sapiens** NCBI 36 | Vega
- Mus musculus** NCBI m36 | Vega
- Danio rerio** Zv6 | Vega

More genomes

- Aedes aegypti** AaegL1
- Anopheles gambiae** AgamP3
- Bos taurus** Btau 2.0
- Caenorhabditis elegans** WS160
- Canis familiaris** CanFam 2.0 **UPDATED!**
- Clona intestinalis** JGI 2
- Clona savignyi** CSAV 2.0
- Dasyurus novemcinctus** ARMA

Ensembl headlines: Release 42 (December 2006)

- New - User accounts** (all species)
- New species - Duck-billed Platypus (*Ornithorhynchus anatinus*)**

Go to "http://www.ensembl.org/Drosophila_melanogaster/"

KEGG Pathway database

- <http://www.genome.jp/kegg/pathway.html>
- Breakdown into major categories:
 - metabolism (the most important one),
 - genetic information processing (including protein folding and sorting),
 - environmental information processing (including membrane transport and intracellular signaling),
 - cellular processes,
 - plus some others
- Broken down into subcategories, e.g. carbohydrate metabolism, and then into individual pathways, e.g. glycolysis/gluconeogenesis
 - (<http://www.genome.jp/kegg/pathway/map/map00010.html>)

Omics databases

