

Chapter-02. 重要生物信息资源

本章内容

2.1 引言

2.2 NCBI系列数据库及数据资源

2.3 UCSC基因组浏览器与数据资源

2.4 EMBL-EBI数据库与数据资源

2.5 其他重要生物信息学资源

2.6 数据批量获取方法



第1节：引言----数据是源头、软件是手段

重要知识点

- ✓ Database and Biological database
- ✓ Database language and tool
- ✓ Primary and secondary databases
- ✓ NCBI, EMBL, DDBJ
- ✓ Software, tools

❖ 1.1 关于生物信息学数据库

1). 数据库 (Database)

用于收集、整理、储存、加工、发布和检索数据的系统。



A. 不同层次的数据库



B. 通过互联网访问数据库

2). 数据库工具

- ✓ SQL（**结构化查询语言**） -世界上最流行的标准化的数据库语言，能快速存储记录文件和图像等。
- ✓ Access, Oracle等
- ✓ 生物类的数据库种类很多（**序列、结构、分子间的互作等**）。
- ✓ 文章投稿之前，需要先将各种核酸和蛋白质序列提交至特定的公开数据库中。
- ✓ 数据库的记录，通常包括两部分：原始数据 + 注释
- ✓ 一个数据库，通常会链接到多个相关的数据库。

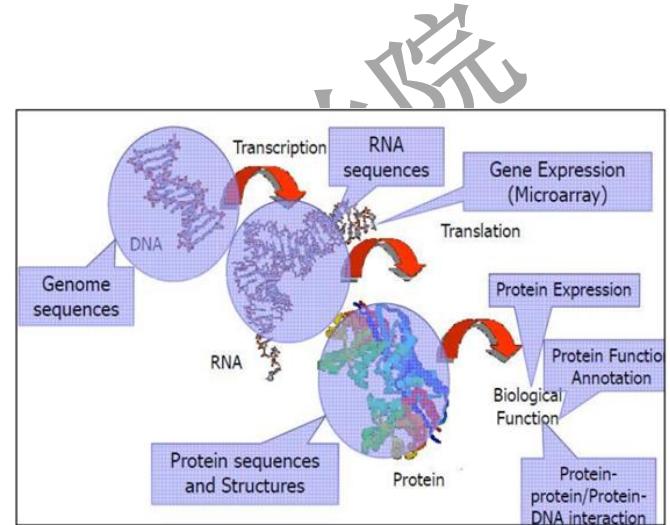
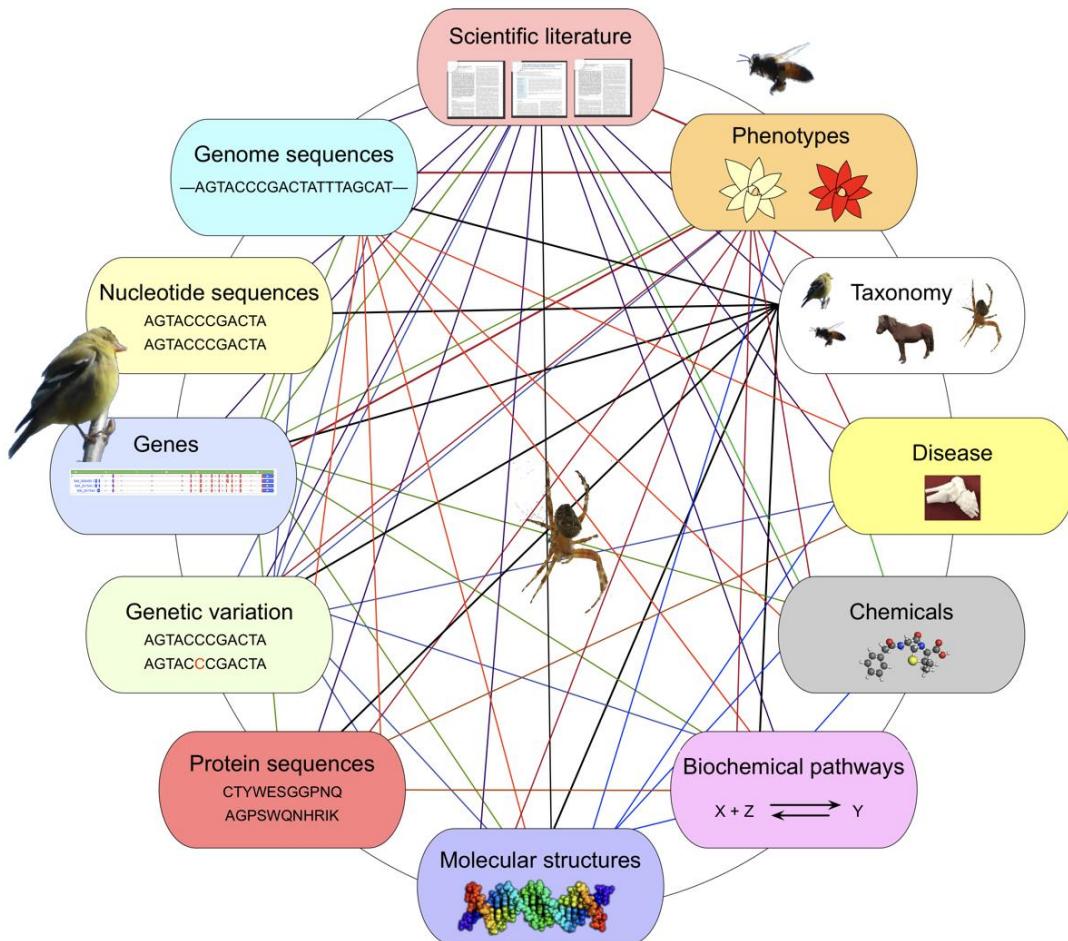
核苷酸序列数据库----水稻抗病相关基因**OSDR8**

DQ176424





Many kinds of bioinformatics databases



These are the biological data types

- nucleotide sequences
- gene level or expression data
- protein sequences
- proteins sequence patterns or motifs
- macromolecular 3D structure
- metabolic pathways

常见的生物信息学数据库

- ✓ 基因组数据库，如

Human (<https://www.ncbi.nlm.nih.gov/genome/guide/human/>)

MGD (<http://www.informatics.jax.org>)

Wormbase (<http://www.wormbase.org>)

Flybase (<http://flybase.net>)

- ✓ 核酸序列数据库，如

GenBank (<https://www.ncbi.nlm.nih.gov/genbank/>)

- ✓ 蛋白质序列数据库，如PIR-PSD、Swiss-Prot、TrEMBL、UniProt等

Uniprot (<http://www.uniprot.org>)

- ✓ 生物大分子三维结构数据库

- ✓ 其他二次专业数据库，如药物靶点数据库TTD等

3). Biomolecular Data Resources:

Bioinformatics Infrastructure for Biomedical Data Science

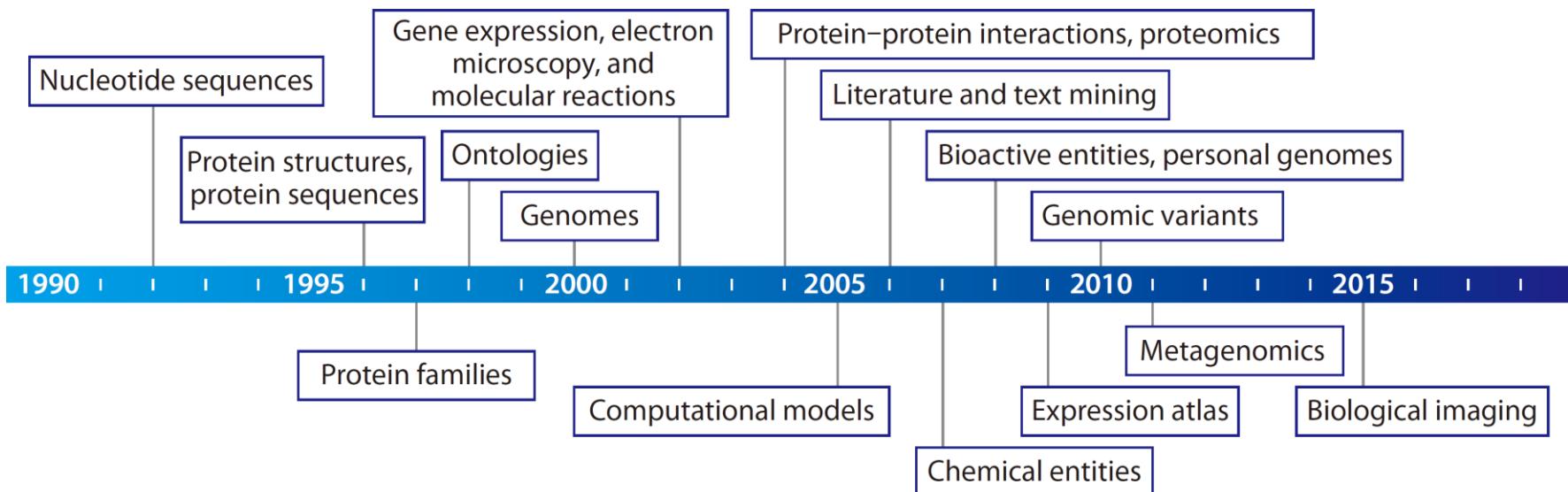


Figure 1

Time line of biomolecular data types and the establishment of data resources. From the emergence of the first shotgun nucleotide sequencing reads, advances in various types of technologies have resulted in the subsequent development of several types of data resources at the European Bioinformatics Institute of the European Molecular Biology Laboratory (EMBL-EBI). These include both archival and added-value data resources, as well as cross-supporting resources such as ontologies and literature resources.

❖ 1.3 国际三大核苷酸序列数据库

- ✓ International Nucleotide Sequence Database Collaboration (INSDC,
国际核苷酸序列数据库联盟)

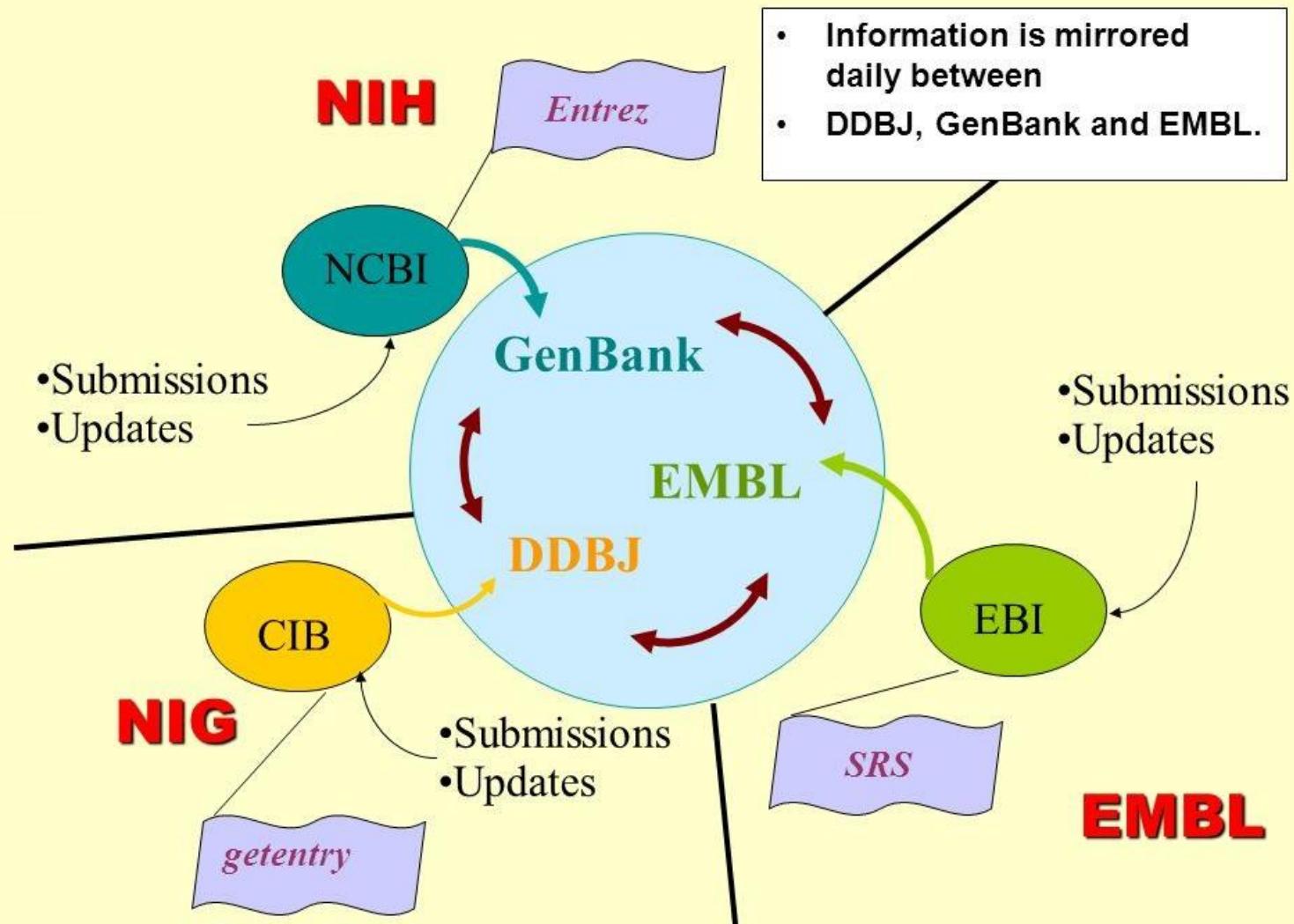
GenBank [美] <http://www.ncbi.nlm.nih.gov/Genbank>

EMBL [欧] <http://www.ebi.ac.uk/embl/>

DDBJ [日] <http://www.ddbj.nig.ac.jp/>

- ✓ 每天这三个数据库作数据同步操作
- ✓ 在任何一个数据库操作(查找、投递数据等)是等效的。

4). 三大核酸数据库



5). NIG

- 日本国立遗传学研究所(National Institute of Genetics)
- 创立于1949年7月1日,属文部省管辖
- 信息服务始于1984年
- 维护管理着DDBJ
 - 1987年1月发行DDBJ第一版
- <http://www.nig.ac.jp>



Research Highlights

2019/09/11 **New**

A new homologous recombination factor, HROB, controls the MCM8-MCM9 pathway

2019/09/05 **New**

Neural signatures of sleep in zebrafish

2019/08/30 **New** **Press release**

Glia-neuron interactions underlie state transitions to generalized seizures

2019/08/28 **New** **Press release**

The making of ‘Fancy Mouse’: Study reveals true cause of colorful hair on popular East Asian pet mice

2019/08/21 **New**

DFAST: Prokaryotic genome annotation pipeline for data submission to DDBJ

More

Seminar Meetings

More



Seminar•Meetings

2019/09/20 13:30–14:30

Patterning chromatin with histone variants controls transcription

2019/09/24 11:00–12:00

Genetic basis of nervous-system response to stress and injury

2019/09/27 10:00–11:00

Cell-type-specific patterned activities specify gene expression patterns for olfactory circuit formation

2019/09/27 15:30–16:30

細胞内部の分子レベルイメージングを目指して
(Towards Molecular-level Imaging in Whole Cell)



Information

2019/09/09 **New**

Message From NIGINTERN 2019

2019/07/10

Summer Holiday (Aug. 15,16)

2019/07/01

Faculty member SHIMAMOTO at the Center for Frontier Research has been awarded tenure

2019/06/27 **Recruitment**

NIG-GS (NIG Global Scholar) selection

2019/04/26 **Past**

Assistant Professor (Cell Dynamics and Signaling Laboratory)

More

Bioinformation and DDBJ Center

[BI-DDBJ Web Sites](#)[Google Custom Search](#)

Bioinformation and DDBJ Center provides sharing and analysis services for data from life science researches and advances science.

Search & Analysis



Submissions



Downloads



SuperComputer



Statistics



Activities



Training



About Us



News from Bioinformation and DDBJ Center

Search

[getentry](#)[ARSA](#)[DRA Search](#)[TXSearch](#)[BLAST](#)

Analysis

[Vector Screening System](#)[ClustalW](#)[WABI \(Web API for Biology\)](#)[DDBJ FTP Site](#)

Databases

[Annotated/Assembled Sequences \(DDBJ\)](#)[Sequence Read Archive \(DRA\)](#)[Genomic Expression Archive \(GEA\)](#)[BioProject](#)[BioSample](#)[Japanese Genotype-phenotype Archive \(JGA\)](#)[Submission portal D-way](#)

NIG SuperComputer

[NIG SuperComputer](#)

DDBJ Information

Training

[DDBJ RSS](#)[DDBJ on Twitter](#)[DDBJ on facebook](#)[DDBJ on Youtube](#)[DDBJ on GitHub](#)[DDBJ on Google Drive](#)

Partners

 大学院共同情報基盤・情報セキュリティ研究センター
国立遺伝学研究所 大学共同利用機関法人
情報・システム研究機構
Research Organization of Information and Systems JBI portal
Japan Service for Biocience Information

NBDC

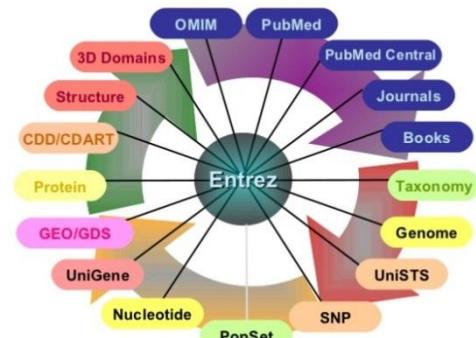
 DBCLS
Database Center
for Life Science PDBj
Protein Data Bank Japan

第2节：NCBI数据库与数据资源

- NCBI是由美国国立卫生研究院、国家医学图书馆（NLM）发起的，旨在推动分子生物学、生化、遗传学知识储存和文献整理。
- 1988年NCBI建立,1992年GenBank成立，随后有PubMed（生物医学文献公共检索和分析平台）、OMIM（人类孟德尔遗传在线）、MMDB（3D蛋白结构分子模型数据库）、UniGene（特有人类基因序列集）、CGAP（癌症基因组剖析计划）等二级数据库。



The (ever expanding) Entrez System

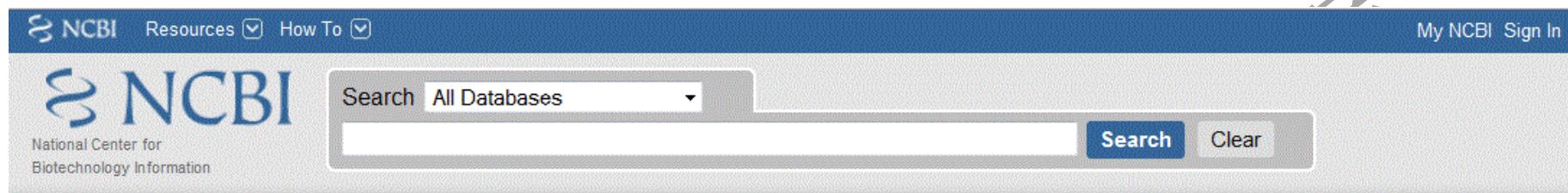


❖ 2.1 美于NCBI

✿ 美国国立生物技术信息中心



- 美国国家生物技术信息中心(National Center for Biotechnology Information)
- 前身是NIH所属的一个计算生物学研究室，1988年独立为NCBI，形式上属于国家医学图书馆(National Library of Medicine/NLM)
- 管理着许多著名数据库，如*GenBank*、PubMed、Medline、dbSNP、COG、OMIM等
- 提供Entrez、BLAST等服务
- 地址：<http://www.ncbi.nlm.nih.gov>



The screenshot shows the NCBI homepage. At the top, there's a blue header bar with the NCBI logo, a search bar labeled "Search All Databases", and links for "Resources" and "How To". On the right, there are "My NCBI" and "Sign In" buttons. Below the header is the main content area.

Resources

NCBI Home

All Resources (A-Z)

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Small Molecules

Taxonomy

Training & Tutorials

Variation

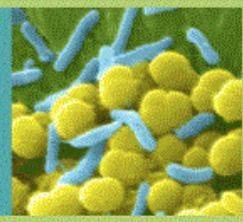
Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[More about the NCBI](#) | [Mission](#) | [Organization](#) | [Research](#) | [RSS](#)

Genome

1000 prokaryotic genomes are now completed and available in the Genome database.



1 2 3 4

How To...

- Determine conserved synteny between the genomes of two organisms
- Find a homolog for a gene in another organism
- Obtain the full text of an article
- Design PCR primers and check them for specificity

Popular Resources

- BLAST
- Bookshelf
- Gene
- Genome
- Nucleotide
- OMIM
- Protein
- PubChem
- PubMed
- PubMed Central
- SNP

NCBI News

Education resource information in the May NCBI News

07 Jun 2010

May NCBI News is available.

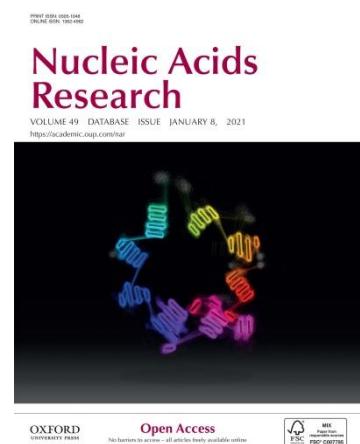
OMIM's new look, Epigenomics in April NCBI News

❖ 2.2 NCBI旗下的生物信息学相关数据库

数据库名称	生物学功能分类
EST (Expressed Sequence Tag)	cDNA和cDNA特征序列
Nucleotide	核酸序列数据库
GSS (Genome Survey Sequence)	测序起始阶段的各种短reads（用于序列示踪、重复序列和基因数量预判等）
GEO	基因表达谱数据库
GenBank	公开可获得的已注释DNA序列（包括Nucleotide、GSS和EST三个子库）
RefSeq	在GenBank基础上给予每个基因一个可靠的注释条目，构成该数据库
Gene	收录全部已测序物种的基因注释信息，是目前最权威的基因注解数据库，标识码为Entrez ID
Genome	完整的基因组数据库
遗传多态数据库（dbSNP、dbVar、dbGap和ClinVar）	分别收录所有物种中发现的短序列多态和突变信息、较大规模的基因组变异、以遗传多态为分子标记物的基因型和表型关联性研究数据，以及临幊上发现或报道的有证据支持的与人类健康有关的变异位点
蛋白质数据库（Protein Cluster、Structure等）	蛋白质分类、三维结构等信息
Unigene	针对每个基因建立一个独立的数据体系，分别将不同来源的基因序列、蛋白相似性、表达、定位等信息罗列和比较，提供一个整合的数据资源

Table 1. The Entrez Databases (as of 9 September 2020)

Database	Records	Description
Literature		
PubMed	31 471 600	scientific and medical abstracts/citations
PubMed Central	6 447 271	full-text journal articles
NLM catalog	1 619 856	index of NLM collections
Books	825 385	books and reports
MeSH	300 500	ontology used for PubMed indexing
Genomes		
Nucleotide	429 731 711	DNA and RNA sequences
BioSample	14 628 076	descriptions of biological source materials
SRA	11 807 161	high-throughput DNA and RNA sequence read archive
Taxonomy	2 401 136	taxonomic classification and nomenclature catalog
Assembly	837 406	genome assembly information
BioProject	458 893	biological projects providing data to NCBI
Genome	55 580	genome sequencing projects by organism
BioCollections	8 138	museum, herbaria and other biorepository collections
Genes		
GEO Profiles	128 414 055	gene expression and molecular abundance profiles
Gene	28 377 759	collected information about gene loci
GEO datasets	4 002 373	functional genomics studies
PopSet	350 627	sequence sets from phylogenetic and population studies
HomoloGene	141 268	homologous gene sets for selected organisms
Genetics		
SNP	720 643 623	short genetic variations
dbVar	6 030 887	genome structural variation studies
ClinVar	845 008	human variations of clinical significance
MedGen	335 277	medical genetics literature and links
GTR	76 814	genetic testing registry
dbGaP	1 397	genotype/phenotype interaction studies
Proteins		
Protein	874 272 642	protein sequences
Identical protein groups	329 946 078	protein sequences grouped by identity
Protein clusters	1 137 329	sequence similarity-based protein clusters
Structure	167 650	experimentally-determined biomolecular structures
Sparcle	149 462	conserved domain architectures
Conserved domains	59 951	conserved protein domains
Chemicals		
PubChem substance	285 048 146	deposited substance and chemical information
PubChem compound	111 325 418	chemical information with structures, information and links
PubChem BioAssay	1 229 071	bioactivity screening studies
BioSystems	983 968	molecular pathways with links to genes, proteins and chemicals



OXFORD
UNIVERSITY PRESS
Open Access
No barrier to access - all articles freely available online
ISSN 0309-1364
ISSN 1362-1418

Fig. Landing page for the new NCBI Datasets product (<https://www.ncbi.nlm.nih.gov/datasets/>) that provides packaged downloads of genomic datasets using either a web interface, an API, or a LINUX command-line tool.

Welcome to NCBI Datasets

BETA

NCBI Datasets is an experimental resource for finding and building datasets - and we're just getting started! Our web interface allows you to download genome sequence and annotation for eukaryotic organisms and our recently added SARS-CoV-2 genome and protein datasets. ... [more](#)

Programmatic access

Bacterial and viral data are not yet supported for online browsing. For access to data for all organisms, including bacteria and viruses, use our command line tool and RESTful APIs.

Command-line

Our Datasets command-line tool, is available for Windows, Mac, and Linux.

GitHub

Explore Datasets with our Python library and Jupyter notebooks.

Datasets API

Use our RESTful APIs to add functionality to your applications.

SELECT COLUMNS				
EDIT	DOWNLOAD	VIEW	SELECT COLUMNS	
	Gene ID	Symbol	Gene name	Chromosome
<input type="checkbox"/>	6794	STK11	serine/threonine kinase 11	19
<input checked="" type="checkbox"/>	1499	CTNNB1	catenin beta 1	3
<input checked="" type="checkbox"/>	4089	SMAD4	SMAD family member 4	18
<input type="checkbox"/>	4436	MSH2	mutS homolog 2	2
<input checked="" type="checkbox"/>	207	AKT1	AKT serine/threonine kinase 1	14
<input type="checkbox"/>	11200	CHEK2	checkpoint kinase 2	22
			...kinase alpha 1	5

Data tables

Build a table of genes or transcripts and choose from a variety of custom columns.

[GET STARTED](#)

Browsing genome datasets



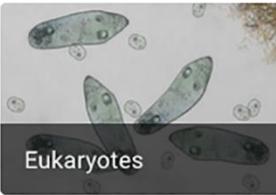
Animals



Plants



Fungi



Eukaryotes

Homo sapiens
human

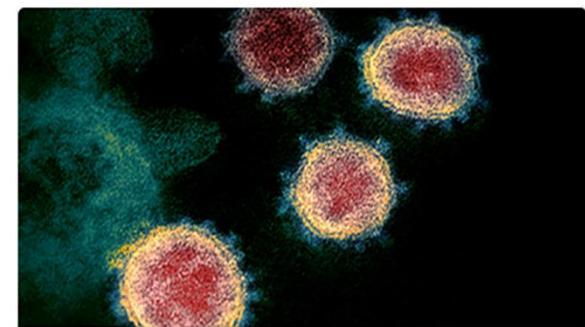
129 assemblies

Mus musculus
house mouse

22 assemblies

Arabidopsis thaliana
thale cress

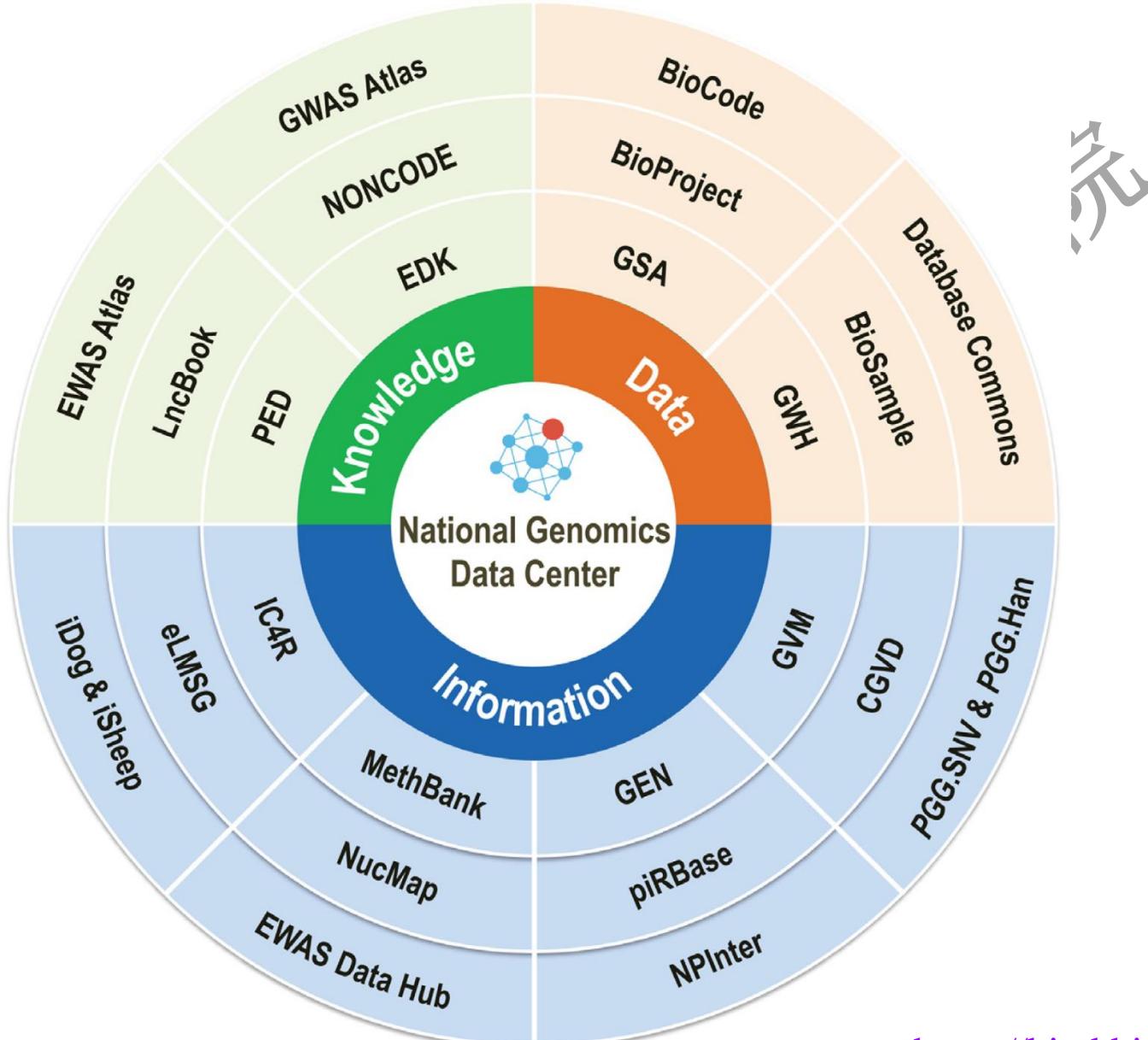
30 assemblies



Coronavirus datasets

Download SARS-CoV-2 genome and protein sequences, annotation and a data report for all complete genomes.

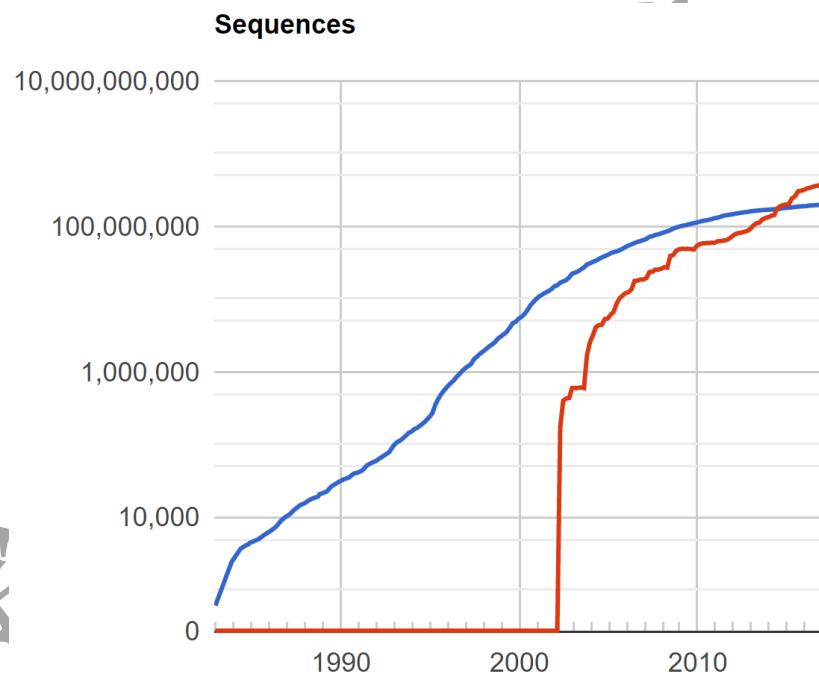
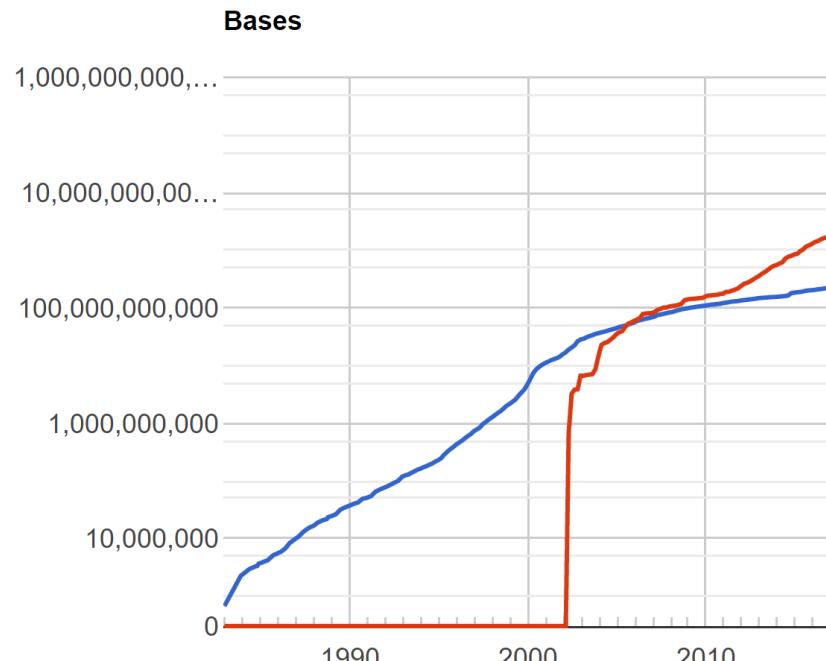
[GET DATA](#)



<https://bigd.big.ac.cn/>

Fig. The National Genomics Data Center's core data resources.

GenBank statistics



Above figures showed that the number of bases and the number of sequence records in each release of GenBank, beginning with Release 3 in 1982. CON-division records are not represented in these statistics: because they are constructed from the non-CON records in the database, their inclusion here would be a form of double-counting. From 1982 to the present, **the number of bases in GenBank has doubled approximately every 18 months.**

❖ 2.3 生物信息学数据库的使用实例

【实例1】

查询引起2019冠状病毒病（COVID-19）的病原体的全基因组序列，并了解其基因组的基本结构。

人类新型冠状病毒：**SARS-CoV-2**

Nucleotide

Search



COVID-19 Information



① 选择数据库

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)Ending
Structural
Racism

UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

[LEARN MORE](#)

NCBI Home

[Resource List \(A-Z\)](#)[All Resources](#)[Chemicals & Bioassays](#)[Data & Software](#)[DNA & RNA](#)[Domains & Structures](#)[Genes & Expression](#)[Genetics & Medicine](#)[Genomes & Maps](#)[Homology](#)[Literature](#)[Proteins](#)[Sequence Analysis](#)[Taxonomy](#)[Training & Tutorials](#)[Variation](#)

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts
into NCBI databases

Download

Transfer NCBI data to your
computer

Learn

Find help documents, attend a
class or watch a tutorial

Develop

Use NCBI APIs and code
libraries to build applications

Analyze

Identify an NCBI tool for your
data analysis task

Research

Explore NCBI research and
collaborative projects

Popular Resources

[PubMed](#)[Bookshelf](#)[PubMed Central](#)[BLAST](#)[Nucleotide](#)[Genome](#)[SNP](#)[Gene](#)[Protein](#)[PubChem](#)

NCBI News & Blog

NCBI on YouTube: ClinVar API, check
data with GaPTools, get genetic context
with Sequence Viewer

27 Aug 2021

Every so often, we gather our most[GenBank release 245.0](#)

26 Aug 2021

GenBank release 245.0 (8/18/2021) is
now available on the NCBI FTP site. This

Nucleotide

Nucleotide

COVID-19

Create alert Advanced

Search

Help



COVID-19 Information

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

② 输入检索对象

③ 执行

Species

Animals (1,519)

Fungi (127)

Protists (2)

Bacteria (6,669)

Viruses (1,215,467)

Customize ...

Molecule types

genomic DNA/RNA (1,222,391)

mRNA (1,223)

Customize ...

Source databases

INSDC (GenBank) (1,216,287)

RefSeq (7,407)

Customize ...

Sequence Type

Nucleotide (1,223,794)

Genetic compartments

Mitochondrion (1)

Sequence length

Custom range...

Release date

Custom range...

Revision date

Custom range...

[Clear all](#)[Show additional filters](#)

Summary ▾ 20 per page ▾ Sort by Default order ▾

Send to: ▾

Filters: [Manage Filters](#)

Results by taxon

Top Organisms [Tree]

- Severe acute respiratory syndrome coronavirus 2 (1215321)
- Klebsiella pneumoniae (6633)
- Homo sapiens (1380)
- Human coronavirus 229E (124)
- [Candida] auris (75)
- All other taxa (261)

[More...](#)

Find related data

Database: Select[Find items](#)

Search details

"Severe acute respiratory syndrome coronavirus 2" [Organism] OR COVID-19 [All Fields]

[Search](#)[See more...](#)

Recent activity

[Turn Off](#) [Clear](#)

REFERENCE GENOME

Was this helpful?

[Severe acute respiratory syndrome coronavirus 2 \(SARS-CoV-2\) reference genome](#)[Severe acute respiratory syndrome coronavirus 2 \(Host: human,vertebrates\)](#)

ssRNA(+)

RefSeq: NC_045512.2

[NCBI Virus](#) [RefSeq genome \(1\)](#) [RefSeq Proteins \(38\)](#) [NCBI SARS-CoV-2 resources](#)

Assembly and annotation statistics

+

④ 结果呈现 <https://www.ncbi.nlm.nih.gov/nuccore/?term=COVID-19>

⑤ 展示COVID-19的全基因组序列（GenBank格式）

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

COVID-19 Information

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | Español

GenBank ▾ Send to: ▾ Change region shown ▾

Customize view ▾ Analyze this sequence ▾

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

NCBI Virus

Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information ▾

Assembly

BioProject

Protein

PubMed

Taxonomy

Full text in PMC

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

FASTA Graphics

Go to: ▾

LOCUS NC_045512 29903 bp ss-RNA linear VRL 18-JUL-2020

DEFINITION Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome.

ACCESSION NC_045512

VERSION NC_045512.2

DBLINK BioProject: PRJNA485481

KEYWORDS RefSeq.

SOURCE Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2)

ORGANISM [Severe acute respiratory syndrome coronavirus 2](#)

Viruses; Riboviria; Orthornavirae; Pisuviricota; Pisoniviricetes; Nidovirales; Cornidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.

REFERENCE 1 (bases 1 to 29903)

AUTHORS Wu, F., Zhao, S., Yu, B., Chen, Y.M., Wang, W., Song, Z.G., Hu, Y., Tao, Z.W., Tian, J.H., Pei, Y.Y., Yuan, M.L., Zhang, Y.L., Dai, F.H., Liu, Y., Wang, Q.M., Zheng, J.J., Xu, L., Holmes, E.C. and Zhang, Y.Z.

TITLE A new coronavirus associated with human respiratory disease in China

JOURNAL Nature 579 (7798), 265-269 (2020)

PUBMED [32015508](#)

REMARK Erratum: [Nature. 2020 Apr;580(7803):E7. PMID: 32296181]

REFERENCE 2 (bases 13476 to 13503)

AUTHORS Baranov, P.V., Henderson, C.M., Anderson, C.B., Gesteland, R.F., Atkins, J.F. and Howard, M.T.

<https://www.ncbi.nlm.nih.gov/nuccore/1798174254>

⑥ 展示COVID-19的全基因组序列（FASTA格式）

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Advanced Search Help

COVID-19 Information
[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)

FASTA Send to: Change region shown

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2
[GenBank](#) [Graphics](#)

>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

ATTAAGGTTTACCTTCCCAGGAACAAACCAACTTCGATCTTGTAGATCTGTTCTCTAAA
CGAACCTTAAATCTGTTGGCTGTCACCTGGCTCATGCTTAGTGCACTCAGCAGTATAATTAAAC
TAATTACTGCGITGACAGGCCACAGTAATCGCTATCTCTGCAGGCTGCTTACGGTTCCGCTCGTG
TTGCAGCCGATCATCAGCACATCTAGGTTTCGTCGGGTG TGACCGAAAGGTAAGATGGAGAGCCTGTC
CCTGTTCAACCGAAAAACACCGTCCAACCTACTTGCCTGTTACAGGTTCCGGACGTGCTGGTAC
GTGGCTTGGAGACTCCGTGGAGGGTCTTATCAGAGGCACGTCAACATCTAAAGATGGCACTTGTG
CTTAGTAGAAGTTGAAAAAGGGTTTGCGCTCAACTTGAAACAGCCATGTGTTCATCAAACGTTCCGGAT
GCTCGAACTGCACCTCATGGTCATGTTATGGTTGAGCTGGTAGCAGAACTCGAAGGCATTCACTGGTC
GTAGTGGTGAGACACTTGTGTCCTTGTCCTCATGGCGAAATACCCAGTGGCTTACCGCAAGGTTCT
TCTCGTAAGAACCGTAATAAAGGACTGGTGGCCATAGTTACGGCGCCATCAAGTCATTGACTTA
GGCAGACGAGCTGGCACTGATCTATGAAGATTTCAGAAACACTTAAACATAGCAGTGGTG
TTACCCGTAACTCATGGTGACTTAAACGGGGCATACACTCGCTATGTCGATAACAACCTCTGTG
CCCTGATGGCTACCCCTGACTGCAATTAAAGACCTTCTAGCACGTGCTGGTAAAGCTTATGCACTTGTG
TCCGAACAACATGGGACTTATTGACACTAAAGAGGGGTGTATACTCTGCTGCGTGAACATGAGCATGAAATTG
CTTGGTACACGGAAACGTTCTGAAAAGAGCTATGGAATTGCGACACCTTTGAAATTAAATTGGCAAAAGAA
ATTGACACCTTCAATGGGAATGTCAAATTGTTATTTCCCTAAATCCATAATCAAGACTATTC
CCAAGGGTTGAAAAAGAAAAAGCTTGATGGCTTATGGGTAGAATTGCACTGCTCATCCAGTTGGCTCAC
CAAATGAAATGCAACCAAAATGTCGCTTCAACTCTCATGAAAGTGTGATCATTTGTTGAAACCTTCATGGCA
GACGGCGATTGTTAAAGGCCACTTGTGCAAGTTGTCGACTGAGAAATTGACTAAAGAAGGTGCCACT
ACTTGTTGTTACTTACCCCCAAATGCTGTTGTTAAATTATTGTCGCACTGTCACAATTCTGAGAAGTAG
GACCTGAGCATGTTGCGAATACCATAATGAAATCTGGCTTGAACAGGATTCTGTAAGGGTGGCTCG
CACTATTGCTTGGAGGCTGTGTTGTTCTTATGTTGTTGCGATAACAAAGTGTCGCTATTGGGTICCA
CGTGCTAGCGCTAACATAGGTGTAACCATACAGGTGTTGTTGGAGAAGGTCCGAAGGTCTTAATGACA

Customize view

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features
Find in this Sequence

NCBI Virus
Retrieve, view, and download SARS-CoV-2 coronavirus genomic and protein sequences.

Related information
Assembly
BioProject
Protein
PubMed
Taxonomy

7 展示COVID-19的全基因组序列（Graphics格式）

Graphics ▾

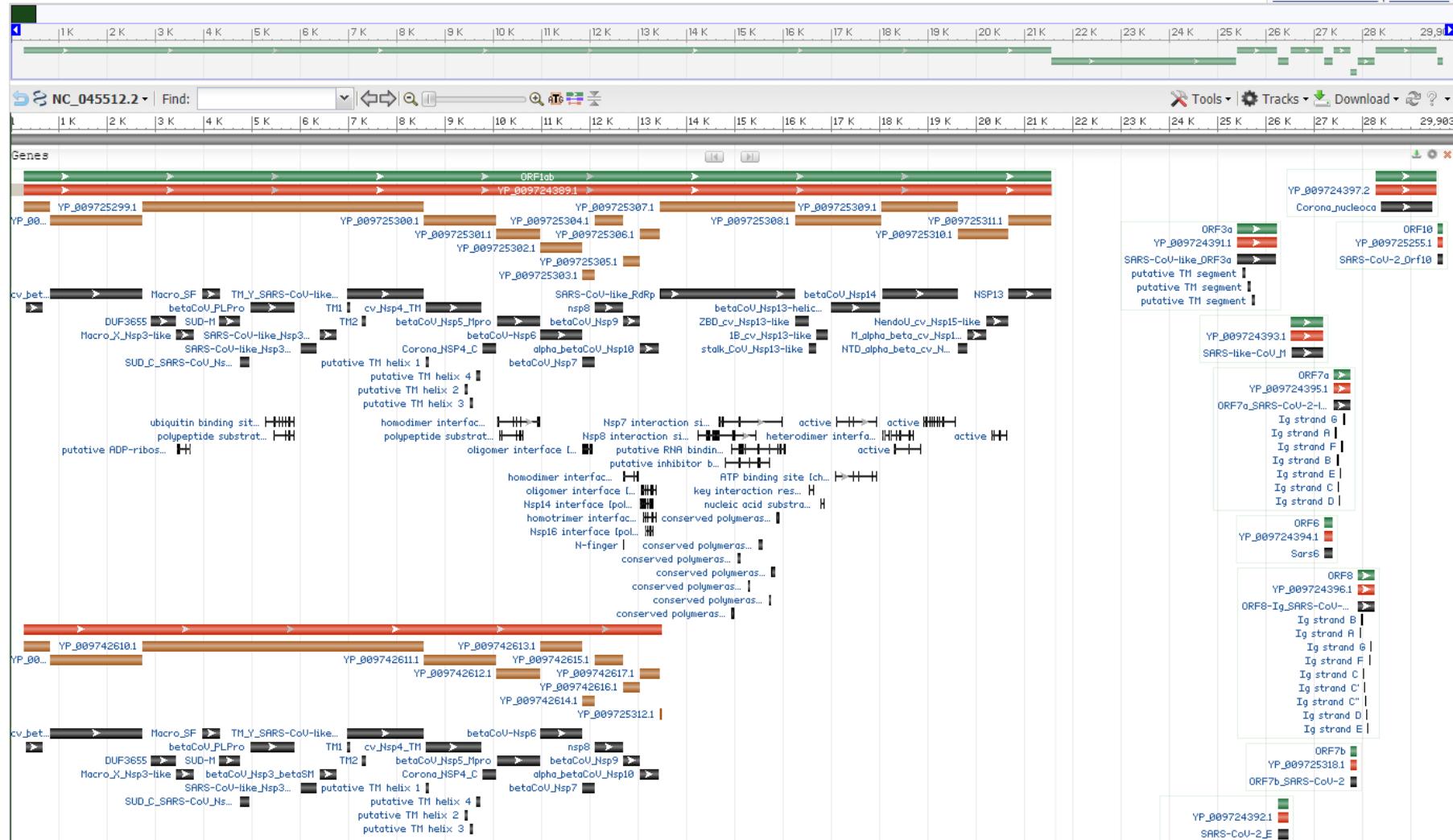
Send to: ▾

Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

GenBank FASTA

[Link To This View](#) | [Feedback](#)



【实例2】

对“人类beta-珠蛋白cDNA序列”进行相似性分析。

(提示: *NM_000518.5*; 工具: 在线BLAST)

步骤:

- 1、先检索出cDNA核苷酸序列
- 2、用FASTA格式显示
- 3、将FASTA格式全选后黏贴与BLAST检索框内
- 4、选择分析所用的数据库
- 5、点击BLAST按钮进行分析



National Center for
Biotechnology Information

[All Databases](#)[Search](#)

COVID-19 Information



[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#) | [Español](#)



UNITE

A new NIH initiative to end structural racism and achieve racial equity in the biomedical research enterprise.

[Click here](#)[LEARN MORE](#)

NCBI Home

Resource List (A-Z)

[All Resources](#)[Chemicals & Bioassays](#)[Data & Software](#)[DNA & RNA](#)[Domains & Structures](#)[Genes & Expression](#)[Genetics & Medicine](#)[Genomes & Maps](#)[Homology](#)[Literature](#)[Proteins](#)[Sequence Analysis](#)[Taxonomy](#)[Training & Tutorials](#)[Variation](#)

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit

Deposit data or manuscripts
into NCBI databases



Download

Transfer NCBI data to your
computer



Learn

Find help documents, attend a
class or watch a tutorial



Develop

Use NCBI APIs and code
libraries to build applications



Analyze

Identify an NCBI tool for your
data analysis task



Research

Explore NCBI research and
collaborative projects



Popular Resources

[PubMed](#)[Bookshelf](#)[PubMed Central](#)[BLAST](#)[Nucleotide](#)[Genome](#)[SNP](#)[Gene](#)[Protein](#)[PubChem](#)

NCBI News & Blog

NCBI on YouTube: ClinVar API, check
data with GaPTools, get genetic context
with Sequence Viewer

27 Aug 2021

Every so often we gather our most

GenBank release 245.0

26 Aug 2021

GenBank release 245.0 (8/18/2021) is
now available on the NCBI FTP site. This
release has 15.31 trillion bases and 249

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

N
E
W
S

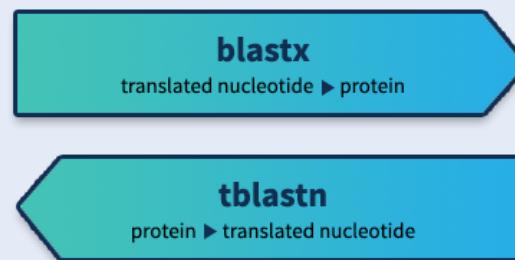
BLAST+ 2.12.0 is here!

We have made some improvements to how BLAST multi-threads and the amount of memory required by makeblastdb.

Tue, 13 Jul 2021 12:00:00 EST

[More BLAST news...](#)

Web BLAST



BLAST Genomes

[Search](#)[Human](#)[Mouse](#)[Rat](#)[Microbes](#)

Standalone and API BLAST

[Download BLAST](#)

Get BLAST databases and executables

[Use BLAST API](#)

Call BLAST from your application

[Use BLAST in the cloud](#)

Start an instance at a cloud provider

blastn

blastp

blastx

tblastn

tblastx

Standard Nucleotide BLAST

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)Query subrange [?](#)From To

Or, upload file

 sequence.fasta [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#) Align two or more sequences [?](#)

Choose Search Set

Database

 Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus [?](#)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 sequencesLimit to
Optional
Entrez Query
Optional Sequences from type material Enter an Entrez query to limit search [?](#)

Program Selection

Optimize for

 Highly similar sequences (megablast)
 More dissimilar sequences (discontiguous megablast)
 Somewhat similar sequences (blastn)Choose a BLAST algorithm [?](#)

BLAST

Search **database Nucleotide collection (nr/nt)** using **Megablast (Optimize for highly similar sequences)** Show results in a new window

[Edit Search](#)[Save Search](#)[Search Summary](#) ▾[How to read this report?](#)[BLAST Help Videos](#)[Back to Traditional Results Page](#)

Job Title NM_000518.5 Homo sapiens hemoglobin subunit

RID JVW9TRDW01R Search expires on 09-01 17:07 pm
[Download All](#) ▾Program BLASTN ? [Citation](#) ▾Database nt [See details](#) ▾

Query ID lcl|Query_329291

Description NM_000518.5 Homo sapiens hemoglobin subunit beta (HbB) mRNA

Molecule type dna

Query Length 628

Other reports [Distance tree of results](#) [MSA viewer](#) ?[Descriptions](#)[Graphic Summary](#)[Alignments](#)

Sequences producing significant alignments

 select all 100 sequences selected

① 序列信息

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin subunit beta (HBB).mRNA	Homo sapiens	1160	1160	100%	0.0	100.00%	628	NM_000518.5
<input checked="" type="checkbox"/>	PREDICTED: Pan troglodytes hemoglobin subunit beta (HBB).mRNA		1155	1155	100%	0.0	99.84%	748	XM_508242.4
<input checked="" type="checkbox"/>	PREDICTED: Pan paniscus hemoglobin subunit beta (LOC100976462).mRNA		1149	1149	100%	0.0	99.68%	646	XM_003819029.3
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin_beta.mRNA (cDNA clone MGC:14540 IMAGE:4292125).cor		1142	1142	99%	0.0	99.52%	658	BC007075.1
<input checked="" type="checkbox"/>	Human messenger RNA for beta-globin		1140	1140	99%	0.0	99.52%	626	V00497.1
<input checked="" type="checkbox"/>	PREDICTED: Gorilla gorilla gorilla hemoglobin subunit beta (LOC101126932).mRNA	Gorilla gorilla gorilla	1133	1133	100%	0.0	99.20%	753	XM_019036164.1
<input checked="" type="checkbox"/>	Homo sapiens hemoglobin beta mRNA, complete cds		1133	1133	99%	0.0	99.36%	647	AY509193.1
<input checked="" type="checkbox"/>	PREDICTED: Pongo abelii hemoglobin subunit beta (HBB).mRNA		1098	1098	99%	0.0	98.25%	627	XM_002822127.4
<input checked="" type="checkbox"/>	PREDICTED: Nomascus leucogenys hemoglobin subunit beta (HBB).mRNA		1088	1088	100%	0.0	97.93%	753	XM_004090649.3

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

to

E value

to

Query Coverage

to

② Score:
比对得分，值越高
相似性越高④ E value: 期望值,
<1E-50(极可信),
0.01可信

Download New Select columns Show 100 ?

③ Query Cover:
所比对序列与目标
序列的覆盖度⑤ Identities:
一致性，两个序列
有多少是一样

[Edit Search](#)[Save Search](#)[Search Summary](#) ▾[How to read this report?](#)[BLAST Help Videos](#)[Back to Traditional Results Page](#)

Job Title NM_000518.5 Homo sapiens hemoglobin subunit

RID JVW9TRDW01R Search expires on 09-01 17:07 pm

[Download All](#) ▾Program BLASTN ? [Citation](#) ▾Database nt [See details](#) ▾

Query ID lcl|Query_329291

Description NM_000518.5 Homo sapiens hemoglobin subunit beta (HBB) mRNA

Molecule type dna

Query Length 628

Other reports [Distance tree of results](#) [MSA viewer](#) ?

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

 to

E value

 to

Query Coverage

 to [Filter](#)[Reset](#)

Descriptions

Graphic Summary

Alignments

Taxonomy

Alignment view

Pairwise



CDS feature

[Restore defaults](#)

Download ▾

100 sequences selected [?](#)[Download](#) ▾ [GenBank](#) [Graphics](#)[▼ Next](#) [▲ Previous](#) [◀ Descriptions](#)

Homo sapiens hemoglobin subunit beta (HBB), mRNA

Sequence ID: [NM_000518.5](#) Length: 628 Number of Matches: 1Range 1: 1 to 628 [GenBank](#) [Graphics](#)[▼ Next Match](#) [▲ Previous Match](#)

Score	Expect	Identities	Gaps	Strand
1160 bits(628)	0.0	628/628(100%)	0/628(0%)	Plus/Plus

Query 1 ACATTGCTTCTGACACAACGTGTTCACTAGCAACCTCAAACAGACACCATGGTCATC 60

Sbjct 1 ACATTGCTTCTGACACAACGTGTTCACTAGCAACCTCAAACAGACACCATGGTCATC 60

Query 61 TGACTCCTGAGGAGAACTCTGCCGTTACTGCCCTGTGGGCAAGGTGAACCTGGATCAAG 120

Sbjct 61 TGACTCCTGAGGAGAACTCTGCCGTTACTGCCCTGTGGGCAAGGTGAACCTGGATCAAG 120

Related Information

[Gene](#) - associated gene details[PubChem BioAssay](#) - bioactivity screening[Genome Data Viewer](#) - aligned genomic context

<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone IMAGE:100002185 for hypothetical protein (HBB gene)	synthetic construct	811	811	70%	0.0	99.55%	483	AM393351.1
<input checked="" type="checkbox"/> Human ORFeome Gateway entry vector pENTR223-HBB, complete sequence	Human ORFeo...	809	809	70%	0.0	99.55%	3231	LT736709.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone ccsbBroadEn_06353 HBB gene, encodes complete protein	synthetic construct	809	809	70%	0.0	99.55%	573	KJ896959.1
<input checked="" type="checkbox"/> Homo sapiens full open reading frame cDNA clone RZPD0834E0633D for gene HBB, hemoglobin_beta, complete cds	Homo sapiens	809	809	70%	0.0	99.77%	441	CR541913.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone FLH028849_01X hemoglobin beta (HBB) mRNA, complete cds	synthetic construct	809	809	70%	0.0	99.55%	444	AY890157.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone FLH130860_01L hemoglobin beta (HBB) mRNA, partial cds	synthetic construct	809	809	70%	0.0	99.77%	444	AY894014.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone FLH028845_01L hemoglobin beta (HBB) mRNA, partial cds	synthetic construct	806	806	70%	0.0	99.55%	444	AY892640.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone CCSBhM_00010626 HBB (HBB) mRNA, encodes complete protein	synthetic construct	804	804	70%	0.0	99.32%	573	KR710229.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone CCSBhM_00010601 HBB (HBB) mRNA, encodes complete protein	synthetic construct	804	804	70%	0.0	99.32%	573	KR710228.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone CCSBhM_00010525 HBB (HBB) mRNA, encodes complete protein	synthetic construct	804	804	70%	0.0	99.32%	573	KR710227.1
<input checked="" type="checkbox"/> Synthetic construct Homo sapiens clone CCSBhM_00010498 HBB (HBB) mRNA, encodes complete protein	synthetic construct	804	804	70%	0.0	99.32%	573	KR710226.1
<input checked="" type="checkbox"/> PREDICTED: Mandrillus leucophaeus hemoglobin subunit beta (LOC105535916), transcript variant X1, mRNA	Mandrillus leuco...	791	1027	99%	0.0	95.90%	756	XM_011975165.1
<input checked="" type="checkbox"/> Symphalangus syndactylus hemoglobin subunit beta (HBB) gene, complete cds	Symphalangus ...	782	782	70%	0.0	98.42%	444	MH382900.1
<input checked="" type="checkbox"/> PREDICTED: Theropithecus gelada hemoglobin subunit beta (LOC112607021), transcript variant X1, mRNA	Theropithecus g...	780	999	99%	0.0	95.50%	756	XM_025358025.1
<input checked="" type="checkbox"/> TPA: Pongo abelii GLNA1 gene for globin A1	Pongo abelii	776	776	70%	0.0	98.20%	444	LT548116.1
<input checked="" type="checkbox"/> Macaca mulatta hemoglobin subunit beta (HBB), mRNA	Macaca mulatta	771	771	75%	0.0	96.19%	471	NM_001164428.1
<input checked="" type="checkbox"/> PREDICTED: Carlito syrichta hemoglobin subunit beta (LOC103254684), mRNA	Carlito syrichta	769	769	99%	0.0	88.94%	750	XM_008052701.1
<input checked="" type="checkbox"/> Cercopithecus wolffii hemoglobin subunit beta (HBB) gene, complete cds	Cercopithecus ...	749	749	70%	0.0	97.08%	444	MH382906.1
<input checked="" type="checkbox"/> Erythrocebus patas hemoglobin subunit beta (HBB) gene, complete cds	Erythrocebus pa...	749	749	70%	0.0	97.08%	444	MH382898.1
<input checked="" type="checkbox"/> PREDICTED: Hylobates moloch hemoglobin subunit delta (HBD), mRNA	Hylobates moloch	745	745	77%	0.0	94.26%	785	XM_032166807.1
<input checked="" type="checkbox"/> PREDICTED: Nomascus leucogenys hemoglobin subunit delta (LOC100580435), mRNA	Nomascus leuc...	745	745	77%	0.0	94.26%	765	XM_003254822.3
<input checked="" type="checkbox"/> Allenopithecus nigroviridis hemoglobin subunit beta (HBB) gene, complete cds	Allenopithecus n...	743	743	70%	0.0	96.85%	444	MH382901.1
<input checked="" type="checkbox"/> Miopithecus talapoin hemoglobin subunit beta (HBB) gene, complete cds	Miopithecus tala...	732	732	70%	0.0	96.40%	444	MH382902.1
<input checked="" type="checkbox"/> Homo sapiens hemoglobin subunit delta (HBD), mRNA	Homo sapiens	723	723	77%	0.0	93.44%	620	NM_000519.4
<input checked="" type="checkbox"/> Cercocebus atys hemoglobin subunit beta (LOC105578869), mRNA	Cercocebus atys	723	723	71%	0.0	95.77%	448	NM_001305959.1
<input checked="" type="checkbox"/> Homo sapiens hemoglobin_delta, mRNA (cDNA clone MGC:88275 IMAGE:30418964), complete cds	Homo sapiens	723	723	77%	0.0	93.44%	644	BC070282.1
<input checked="" type="checkbox"/> Homo sapiens hemoglobin_delta, mRNA (cDNA clone MGC:96894 IMAGE:7262103), complete cds	Homo sapiens	723	723	77%	0.0	93.44%	544	BC069307.1

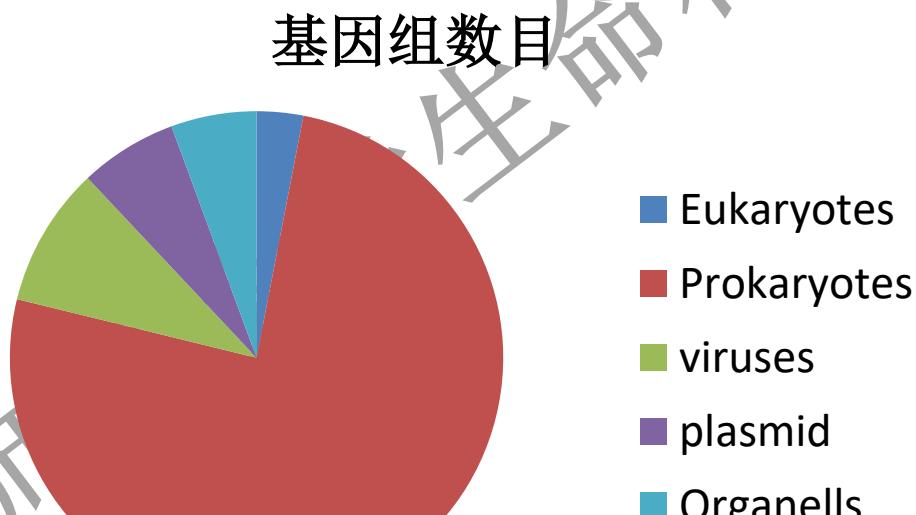


重要概念辨析 (identity, similarity, homology)

- 同源性是定性推断（表示序列同源与否！），而一致性进而相似性则是数量推断，用于描述序列的相关性程度。
- 一致性用以表示两条氨基酸（或核苷酸）序列发生变化的程度，常用百分比表示。
- 相似性是指两条蛋白质（或核苷酸）序列中相同和相似残基对所占的百分比之和。
- 基本假定：如果两条DNA（或蛋白质）序列的比对结果分数较高，那么它们就被定义为同源DNA（或蛋白质）。
- 几点特别说明：
 - ① 当两条序列同源时，它们的序列之间常常有显著的相似性；同源蛋白质几乎总是在三维结构上具有显著的相似性。
 - ② 两个分子即使没有统计学上显著的氨基酸或核苷酸，它们也可能是同源的。
 - ③ 对蛋白质而言，考虑一致性比相似性更有用，因为相似性的计算取决于不同氨基酸残基之间相似程度的定义方式。

【实例3】基因组数据库展示

在NCBI genome数据库中，目前共收录了大量已公开的物种基因组序列等信息。



All Databases

PubMed

Nucleotide

Protein

Genome

Structure

PMC

Search

Genome Project

for

Go

Clear

Limits

Preview/Index

History

Clipboard

Details

About Entrez

Entrez Genome
Project

Home

Overview

Help

Statistics

Sequencing Centers

Submitting

Project Submissions

Project Instructions

General Genome

Submissions

Feature Tables

Bacterial Genome

Submissions

Whole Genome

Shotgun Sequences

Related Resources

DOE Projects

DOE SAI Survey

Genome News

Network

Genomes OnLine

Database

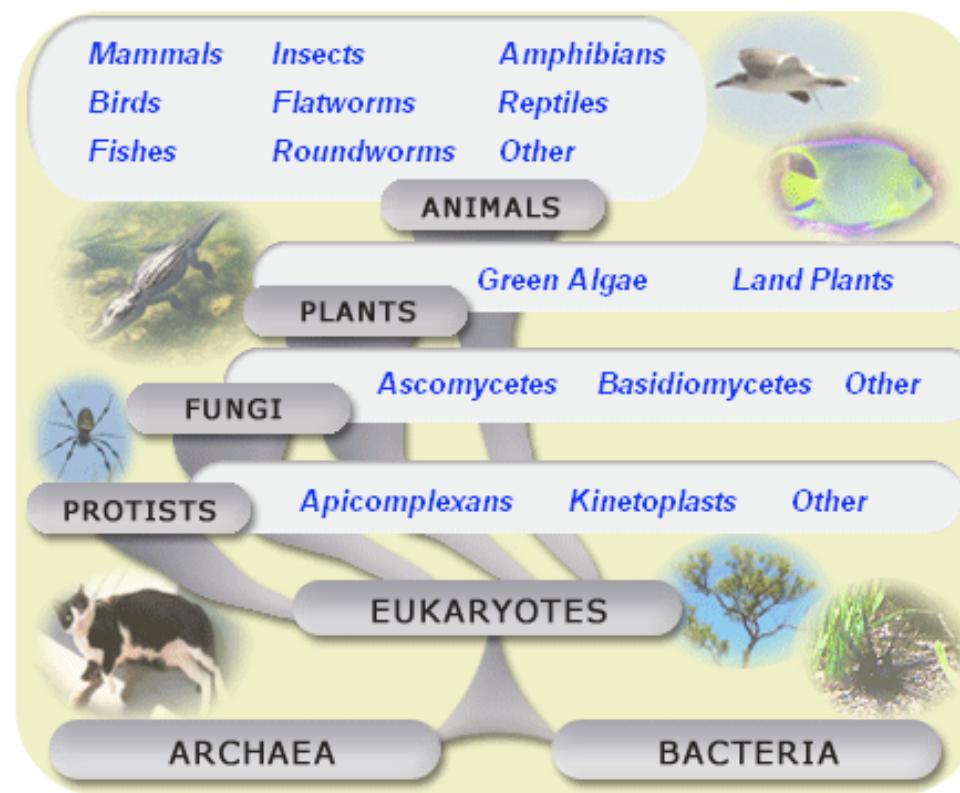
IntlGenome

NHGRI Projects

NIHID Projects

Welcome to the NCBI Entrez Genome Project database.

This searchable database is a collection of complete and incomplete large-scale sequencing, assembly, annotation, and mapping projects for cellular organisms. The database is organized into organism-specific overviews that function as portals from which all projects in the database pertaining to that organism can be browsed and retrieved. [Read more...](#)

**New!****NCBI Res..**[Entrez Gene](#)gene-related
information[Entrez Genome](#)sequence and map data
from whole genomes[Environmental Projects](#)metagenomic-specific
genome projects[Eukaryotic Projects](#)eukaryotic-specific
genome projects[Genomic Biology](#)organism-specific
links[Prokaryotic Projects](#)prokaryotic-specific
genome projects[Organellar Genomes](#)organellar reference sequences
and tools[Plant Genomes](#)major plant
genome projects[RefSeq](#)the reference
sequence project[Viral Genomes](#)viral reference sequences
and tools[WGS Sequences](#)whole genome
shotgun sequences

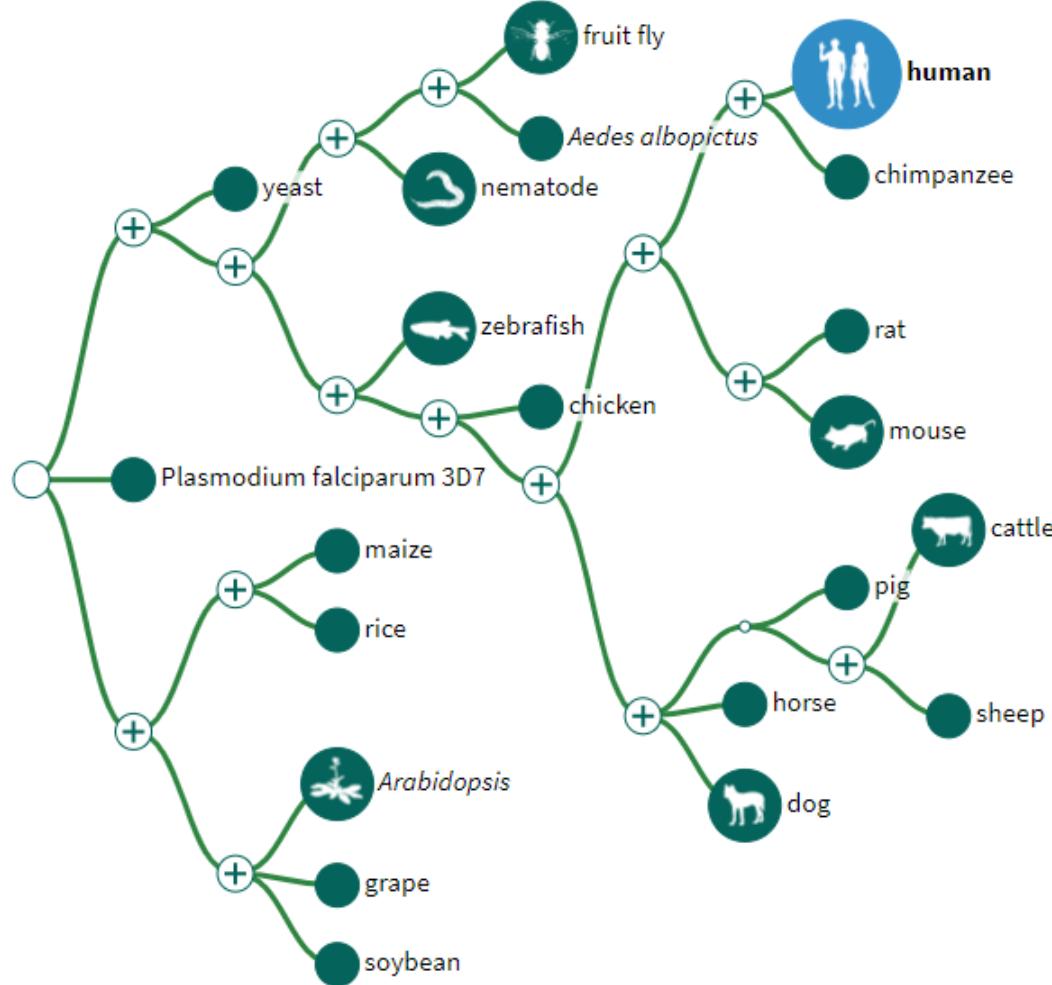
Genome Data Viewer

Switch view



Search organisms

Homo sapiens (human)



Homo sapiens (human)



Search in genome

Location, gene or phenotype



Examples: TXLNA, chr1:32178000-32200000, DNA repair

Assembly

GRCh38.p13

Browse genome

BLAST genome

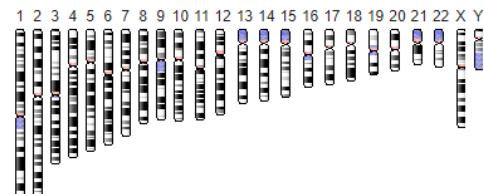
Download via NCBI Datasets

Assembly details

Name	GRCh38.p13
RefSeq accession	GCF_000001405.39
GenBank accession	GCA_000001405.28
Submitter	Genome Reference Consortium
Level	Chromosome
Category	Reference genome
Replaced by	GCF_000001405.25

Annotation details

Annotation Release 109 i
Release date May 16, 2021





Genome Projects

Properties of Eukaryotic Genome Sequencing Projects

[Organism info](#)
[Complete genomes](#)
[Genomes in progress](#)

organism group:

Tools legend: **T** - TaxTable; **P** - ProtTable; **C** - COG Table; **D** - 3-D neighbors; **L** - BLAST; **S** - CDD search; **F** - FTP; **R** - Publications.

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

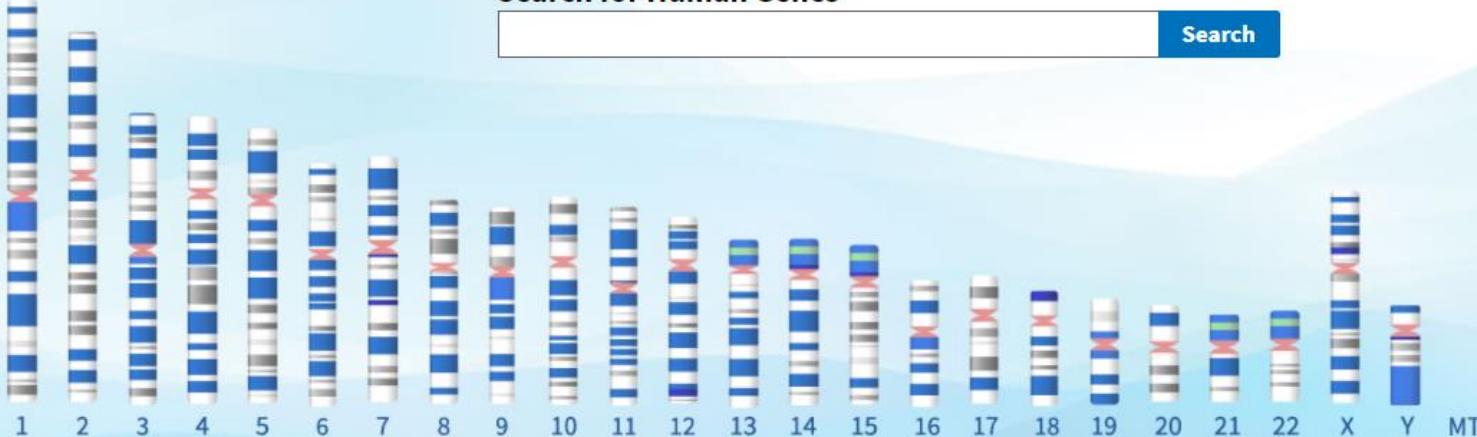
294 Complete Microbial Genomes selected: [A] - 25, [B] - 269

Organism	King	Group	* Size	GC	#chr	#plsm	GenBank	RefSeq	Released	Center	Tools
Acinetobacter sp. ADP1	B	Gammaproteobacteria	3.6	40.4	1		CR543861	NC_005966	10/30/2004	Genoscope	T P C D L S F R
Aeropyrum pernix K1	A	Crenarchaeota	1.67	67	1		BA000002	NC_000854	06/26/1999	NITE	T P C D L S F R
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59	2	2	AE008688	NC_003304	12/18/2001	University of Washington	T P C D L S F R
Agrobacterium tumefaciens str. C58	B	Alphaproteobacteria	5.67	59	2	2	AE007869	NC_003062	12/18/2001	Cereon	T P C D L S F R
Anabaena variabilis ATCC 29413	B	Cyanobacteria	7.07	41.4	1	3	CP000117	NC_007413	03/10/2004	DOE Joint Genome Institute	T P D L S F
Anaplasma marginale str. St. Maries	B	Alphaproteobacteria	1.2	49.8	1		CP000030	NC_004842	12/08/2004	Washington State University	T P C D L S F R
Aquifex aeolicus VF3	B	Aquificae	1.59	43	1	1	AE000657	NC_000918	04/16/1998	DIVERSA	T P C D L S F R
Archaeoglobus fulgidus DSM 4304	A	Euryarchaeota	2.18	46	1		AE000782	NC_000917	12/06/1997	TIGR	T P C D L S F R
Azoarcus sp. EbN1	B	Betaproteobacteria	4.73	65.1	1	2	CR555306	NC_006513	11/20/2004	Max Planck Institute	T P C D L S F R
Bacillus anthracis str. 'Ames Ancestor'	B	Firmicutes	5.5	35.2	1	2	AE017334	NC_007530	05/20/2004	TIGR	T P C D L S F
Bacillus anthracis str. Ames	B	Firmicutes	5.23	35	1		AE016879	NC_003997	05/02/2003	TIGR	T P C D L S F R
Bacillus anthracis str. Sterne	B	Firmicutes	5.23	35.4	1		AE017225	NC_005945	06/24/2004	DOE Joint Genome Institute	T P C D L S F
Bacillus cereus ATCC 10987	B	Firmicutes	5.43	38	1	1	AE017194	NC_003909	02/13/2004	TIGR	T P C D L S F R
Bacillus cereus ATCC 14579	B	Firmicutes	5.43	38	1	1	AE016877	NC_004722	05/02/2003	Institut National de la Recherche Agronomique (INRA)	T P C D L S F R
Bacillus cereus E33L	B	Firmicutes	5.84	35.4	1	5	CP000001	NC_006274	09/16/2004	DOE Joint Genome Institute	T P C D L S F

Human Genome Resources at NCBI

[Download](#)[Browse](#)[View](#)[Learn](#)

Search for Human Genes

[Search](#)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y MT

Select a chromosome to access the [Genome Data Viewer](#)

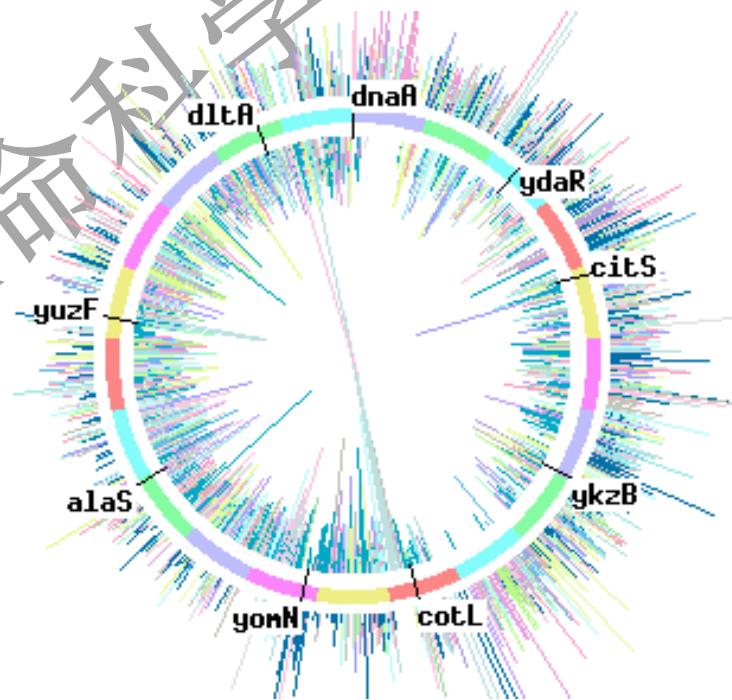


Download

[GRCh38](#)[GRCh37](#)[Reference Genome Sequence](#)[Fasta](#)[Fasta](#)[RefSeq Reference Genome Annotation](#)[gff3](#)[gff3](#)[RefSeq Transcripts](#)[Fasta](#)[Fasta](#)[RefSeq Proteins](#)[Fasta](#)[Fasta](#)



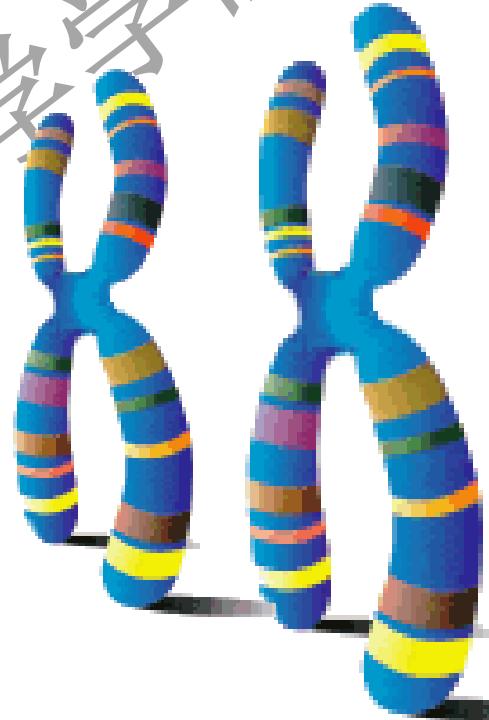
- Full chromosomal sequences are provided
- Genes are annotated
- The annotation can be shown graphically and linked to sequence records



Complex Genomes

- Sequences are provided complete or we help assemble
- Heavy annotation:
transcript regions & ORFs, sequence variations &
markers, clones, ESTs, etc.
- The annotation can be shown graphically and linked
to other databases using the MapViewer

Genes,



A database for retrieval and analysis of karyotype data:
Cancer Chromosomes

以昆虫基因组序列为为例，目前已经公开全基因序列的有538种，其中组

装到染色体、Contig和Scaffold的分别有42、52和444种。



Overview (39517): Eukaryotes (6572); Prokaryotes (163344); Viruses (19778); Plasmids (13790); Organelles (12082)

Filters (1)

Group	<input type="checkbox"/> Animals (538)
Subgroup	<input type="checkbox"/> Amphibians (7) <input type="checkbox"/> Apicomplexans (211) <input type="checkbox"/> Ascomycetes (2,633) <input type="checkbox"/> Basidiomycetes (682) <input type="checkbox"/> Birds (145) <input type="checkbox"/> Fishes (224) <input type="checkbox"/> Flatworms (32) <input type="checkbox"/> Green Algae (57) <input checked="" type="checkbox"/> Insects (538) <input type="checkbox"/> Kinetoplasts (87) <input type="checkbox"/> more
Assembly level	<input type="checkbox"/> Chromosome (42) <input type="checkbox"/> Contig (52) <input type="checkbox"/> Scaffold (444)
Partial	<input type="checkbox"/> All (538) <input type="checkbox"/> Exclude partial (527) <input type="checkbox"/> Include partial only (11)
Anomalous	<input type="checkbox"/> All (538) <input type="checkbox"/> Exclude anomalous (538)
Host	<input type="checkbox"/> plants (5)
RefSeq category	<input type="checkbox"/> reference (1) <input type="checkbox"/> representative (311)
Organism Name	filter by Scientific Name

Download

View 1 - 50 of 538															
#	Organism Name	Organism Groups	Strain	BioSample	BioProject	Assembly	Level	Size(Mb)	GC%	Replicons	WG:	Scaffold	CDS	Release Da	FTP
1	Achilarus lyciades	Eukaryota;Animals;Insects		SAMN05958461	PRJNA35192	GCA_002930495.1	(●)	566.83	18.10	mitochondrion MT: CM009487.1	MOOZ01	47,369		14-Feb-2018	G
2	Acromyrmex echinatior	Eukaryota;Animals;Insects		SAMN02953789	PRJNA62733	GCA_000204515.1	(●)	295.94	34.00		AEVX01	4,339	20,241	14-Apr-2011	R G
3	Acyrthosiphon pisum	Eukaryota;Animals;Insects	LSR1	SAMN00000061	PRJNA13657	GCA_000142985.2	(●)	541.69	31.20	mitochondrion MT: NC_011594.1/FJ411411.1 chromosome 1: NC_035107.1/CM008043.1 chromosome 2: NC_035108.1/CM008044.1 chromosome 3: NC_035109.1/CM008045.1 Show all 4 replicons	ABLFO2	23,925	27,999	01-Apr-2008	R G
4	Aedes aegypti	Eukaryota;Animals;Insects	LVP_AGWG	SAMN07177802	PRJNA39211	GCA_002204515.1	(●)	1,278.73	38.17		NIGP01	2,310	28,317	15-Jun-2017	R G
5	Aedes aegypti	Eukaryota;Animals;Insects	Liverpool	SAMN02953616	PRJNA12434	GCA_000004015.3	(●)	1,383.97	38.80	mitochondrion MT: EU352212.1	AAE02	4,757	17,400	11-Feb-2005	R G
6	Aedes aegypti	Eukaryota;Animals;Insects	BV_Aedes	SAMN03284760	PRJNA26839	GCA_001014885.1	(●)	744.60	28.00		JXPU01	223,039		28-May-2015	G
7	Aedes albopictus	Eukaryota;Animals;Insects	C6/36	SAMN05908721	PRJNA357111	GCA_001876365.2	(●)	2,247.31	40.40	mitochondrion MT: NC_006817.1/AY072044.1	MNAFO2	2,435	42,912	10-Nov-2016	R G
8	Aedes albopictus	Eukaryota;Animals;Insects	Foshan	SAMN03265380	PRJNA27027	GCA_001444175.2	(●)	1,923.48	40.70		JXUM01	154,782	17,146	29-Sep-2015	G
9	Aedes albopictus	Eukaryota;Animals;Insects	Rimini	SAMN03853745	PRJNA289461	GCA_001574995.1	(●)	1,341.04	40.40		LMAW01	3,342,920		22-Dec-2015	G
10	Aethina tumida	Eukaryota;Animals;Insects	BRL-Maryland	SAMN06036204	PRJNA36127	GCA_001937115.1	(●)	234.34	30.00		MRBJ01	3,063	17,463	05-Jan-2017	R G
11	Agrilus planipennis	Eukaryota;Animals;Insects	EAB-ADULT	SAMN02439908	PRJNA34347	GCA_000699045.2	(●)	353.07	36.00	mitochondrion MT: NC_030758.1/KT363854.1	JENH02	3,613	22,159	06-Jun-2014	G
12	Aleochara bilineata	Eukaryota;Animals;Insects	Rennes	SAMN06611084	PRJNA37816	GCA_003054995.1	(●)	85.90	-		NBZA01	33,003		17-Apr-2018	G
13	Amphinemura sulcicollis	Eukaryota;Animals;Insects	Dwyryd	SAMN04568201	PRJNA315681	GCA_001676325.1	(●)	271.93	41.80		LVVV01	432,491		24-Jun-2016	G
14	Amyelois transitella	Eukaryota;Animals;Insects	UIUC subculture of SPIRL-1966	SAMN03009087	PRJNA29202	GCA_001186105.1	(●)	406.47	37.60		LACK01	7,301	18,472	21-Jul-2015	R G
15	Anopheles albimanus	Eukaryota;Animals;Insects	ALBI9_A	SAMN02953839	PRJNA19156	GCA_000349125.2	(●)	173.34	49.25	chromosome 2L: CM008152.1 chromosome 2R: CM008153.1 chromosome 3L: CM008154.1 Show all 5 replicons	APCK01	236		25-Mar-2013	G
16	Anopheles aquasalis	Eukaryota;Animals;Insects	PFPP-2014	SAMN07206858	PRJNA38975	GCA_002846955.1	(●)	162.94	-		NJHH01	16,504		21-Dec-2017	G
17	Anopheles arabiensis	Eukaryota;Animals;Insects	DONG5_A	SAMN02953842	PRJNA19156	GCA_000349185.1	(●)	246.57	45.30		APCN01	1,214		25-Mar-2013	G
18	Anopheles atroparvus	Eukaryota;Animals;Insects	E BRO	SAMN01087926	PRJNA67233	GCA_000473505.1	(●)	224.29	46.80		AXCP01	1,371		30-Sep-2013	G
19	Anopheles christyi	Eukaryota;Animals;Insects	ACHKN1017	SAMN02953841	PRJNA19156	GCA_000349165.1	(●)	172.66	42.80		APCM01	30,369		25-Mar-2013	G

【实例4】了解NCBI新发布的数据库Datasets

The screenshot shows the GitHub repository page for 'ncbi/datasets'. At the top, there's a large title 'ncbi/datasets' and a blue NCBI logo with a stylized 'S' and the word 'NCBI'. Below the title, a description states 'NCBI Datasets is an experimental resource for finding and building datasets'. The repository statistics are listed: 3 contributors, 11 issues, 68 stars, and 18 forks. A GitHub icon is also present. A banner at the bottom of the page reads 'NCBI Insights'.

**NCBI Datasets now provides
downloads of gene data for more
than 30 thousand organisms**

My NCBI Password Retirement

- [Details](#)
- [Frequently Asked Questions](#)
- [My NCBI Login Transition Tips](#)



This is an NCBI Labs experiment. Learn more.

NCBI

NLM

NIH



menu

Search NCBI

SEARCH

NCBI Datasets

BETA

NCBI Datasets is a new resource that lets you easily gather data from across NCBI databases. Find and download gene, transcript, protein and genome sequences, annotation and metadata.

What's new

[NCBI Insights](#) JULY 14, 2021

Introducing the new NCBI Datasets Genomes page

The updated NCBI Datasets Genomes page now has genome data for all domains of life, including ...

[NCBI Insights](#) JUNE 22, 2021

June 30 Webinar: Using NCBI Datasets to download sequence and annotation for genomes and genes

Join us on June 30, 2021 at 12PM eastern time to learn how to use the ...

[MORE NEWS](#)

[NCBI Insights](#) APRIL 20, 2021

New NCBI Datasets home and documentation pages provide easier access

NCBI Datasets, the new set of services for downloading genome assembly and annotation data (previous Datasets ...

Datasets数据库重构了基因组数据的部署方式

Genomes Quickstart

Browse and download genome data using our Genome page. Genome data is also available using our command-line tool and API. Genome data includes genome, transcript and protein sequences, genome annotation and metadata.

[BROWSE GENOMES](#)

Popular species

Homo sapiens

human

Rattus norvegicus

rat

Mus musculus

mouse

Drosophila melanogaster

fruit fly

Arabidopsis thaliana

thale cress

Danio rerio

zebrafish

Bos taurus

cow

Gallus gallus

chicken

Oryza sativa

rice

How to

[Get genome metadata](#)

[Download a genome data package](#)

[Download large genome data packages](#)



menu MENU

Search NCBI

SEARCH

Genomes – NCBI Datasets BETA

Download a genome dataset including genome, transcript and protein sequence, annotation and a data report

TAXONOMIC NAME

search

Please enter a taxon name or select from popular species

Homo sapiens
human

Mus musculus
mouse

Arabidopsis thaliana
thale cress

Rattus norvegicus
rat

Drosophila melanogaster
fruit fly

Danio rerio
zebrafish

Bos taurus
cow

Gallus gallus
chicken

Oryza sativa
rice

① 命令行下载

② 压缩格式的数据包

③ 集成化的编程语言接口



Documentation

Quickstart guides

Command line

How-to guides

Data packages

Programming languages

Python

R

Reference

Documentation / Programming languages

Supported programming languages

Datasets supports all languages via its RESTful API, accessible using the [OpenAPI v3 spec](#).

Python

Python-related resources for NCBI Datasets

R

R-related resources for NCBI Datasets

Generated August 30, 2021

【实例5】Structure数据库展示

NCBI Structure数据库即生物大分子三维数据库（MMDB），包含来自X-ray晶体学和三维结构的实验数据。MMDB的数据从蛋白质结构数据库PDB获得，但对结构的阐述较PDB更为详细，更多的是比较结构的相似性及亲缘关系。

NCBI Resources How To Sign in to NCBI

Structure Structure Advanced Search Help



Using Structure 使用帮助等

Structure Tools 本库工具 More Resources 其他资源

Macromolecular Resources Overview PDB

CBLAST Protein

Help Cn3D CDD

News IBIS PubChem

FTP VAST NCBI Structure Group Resources & Research

**Cn3d
阅读器**

- 1
- 2
- 3
- 4

**Install
安装方法一**

NCBI > Structure Home > 3D Macromolecular Structures > Cn3D

NCBI Structure

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains PubChem BioSystems

Cn3D macromolecular structure viewer

About Cn3D Windows 安装方法二 Tutorial 学习指南

Windows 安装方法二 Tutorial 学习指南

Highlights

New Features in Cn3D 4.3.1:

- View superpositions of structures that have similar molecular complexes, as identified by the newly released VAST+ (an enhanced version of the existing Vector Alignment Search Tool). The VAST+ help document provides additional details about the tool and examples of how it can be used to learn more about proteins.
- Cn3D 4.3.1 uses the MIME type: application/xnd.ncbi.cnb3d. (Up to version 4.3, Cn3D used the MIME type chemica/ncbi-asn1-binary.)

New Features in Cn3D 4.3:

- View biological units and crystallographic symmetry from MMDB

WD40 - Cn3D 4.3

File View Select Style Window Cn3D Help

WD40 Descriptive Items

Name: WD40

WD40: WD40 domain, found in a cover a wide variety of functions modules in signal transduction, protein assembly; typically contains a



下半页

This example is a curated CO of the WD40 domain, which is a multi-functional 7-fold beta propeller. Cn3D is showing a representative protein structure, the family alignment, and annotation panels with information about annotated features of this protein family. Highlighted in both structure and sequence windows are the conserved residues in a pattern characteristic to this domain.

Download Cn3D 4.3.1 Now!

**Download Cn3d
方法三**

Structure > 3D Macromolecular Structures > Cn3D > Downloading Cn3D for Windows

NCBI Structure

HOME SEARCH GUIDE Structure Home 3D Macromolecular Structures Conserved Domains PubChem BioSystems

Cn3D macromolecular structure viewer

ABOUT TUTORIAL FAQ INSTALL PUBLICATIONS NEWS RESOURCES DISCOVER

Downloading Cn3D for Windows

Download the Cn3D 4.3.1 installer here:
http://ftp.ncbi.nlm.nih.gov/cn3d/Cn3D-4.3.1_setup.exe

点击下载、安装

Close all internet browsers and then double-click on the .exe file.

The installer should automatically configure the common web browsers to launch Cn3D from NCBI web pages. If for some reason this does not work, see the instructions for manually configuring your browser.

New Features in Cn3D 4.0:

- Re-written in C++ to improve performance and extensibility
- Improved OpenGL rendering speed
- New user interface using wxWidgets
- Complete alignment editing system, along with algorithms for constructing new alignments
- Extensive structure and alignment annotation features
- Detailed style control
- High-resolution image export
- Built-in help system

蛋白质结构与药物设计研究的利器--PDB数据库

RCSB PDB Deposit Search Visualize Analyze Download Learn More Documentation Careers MyPDB ▾

RCSB PDB PROTEIN DATA BANK 181535 Biological Macromolecular Structures Enabling Breakthroughs in Research and Education

Enter search terms or PDB ID(s). Advanced Search | Browse Annotations Help

PDB-101 Worldwide Protein Data Bank EMDDataResource NDB Worldwide Protein Data Bank Foundation

Celebrating 50 YEARS OF Protein Data Bank

Developers: Join the RCSB PDB Team Explore Open Positions

Welcome

Deposit

Search

Visualize

Analyze

Download

Learn

A Structural View of Biology

This resource is powered by the Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies that helps students and researchers understand all aspects of biomedicine and agriculture, from protein synthesis to health and disease.

As a member of the wwPDB, the RCSB PDB curates and annotates PDB data. The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

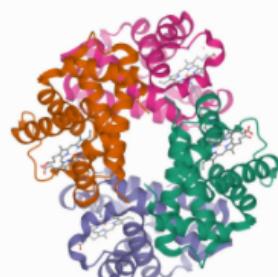
COVID-19 CORONAVIRUS Resources

Celebrating 50 Years

August Molecule of the Month

Ribonuclease P

<https://www.rcsb.org/>



2DN2

1.25 Å resolution crystal structure of human hemoglobin in the deoxy form

Park, S.-Y., Yokoyama, T., Shibayama, N., Shiro, Y., Tame, J.R.

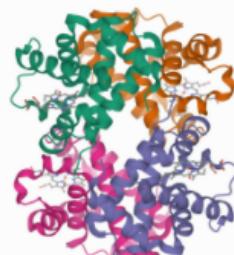
(2006) J Mol Biol 360: 690-701

Released 2006-05-09**Method** X-RAY DIFFRACTION 1.25 Å**Organisms** Homo sapiens**Macromolecule** Hemoglobin alpha subunit (protein)

Hemoglobin beta subunit (protein)

Unique Ligands HEM Download File View File

直接下载文件

 3D View 3D View

2DN3

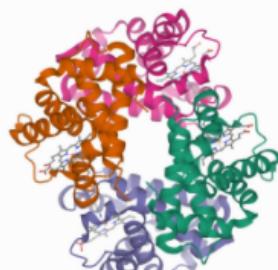
1.25 Å resolution crystal structure of human hemoglobin in the carbonmonoxy form

Park, S.-Y., Yokoyama, T., Shibayama, N., Shiro, Y., Tame, J.R.

(2006) J Mol Biol 360: 690-701

Released 2006-05-09**Method** X-RAY DIFFRACTION 1.25 Å**Organisms** Homo sapiens**Macromolecule** Hemoglobin alpha subunit (protein)

Hemoglobin beta subunit (protein)

Unique Ligands CMO, HEM Download File View File在线浏览及下
载

2HHE

OXYGEN AFFINITY MODULATION BY THE N-TERMINI OF THE BETA CHAINS IN HUMAN AND BOVINE HEMOGLOBIN

Gilliland, G.L., Pechik, I., Fronticelli, C., Ji, X.

(1994) J Biol Chem 269: 23965-23969

Released 1995-01-26**Method** X-RAY DIFFRACTION 2.2 Å**Organisms** Homo sapiens**Macromolecule** HEMOGLOBIN (DEOXY) (ALPHA CHAIN) (protein)

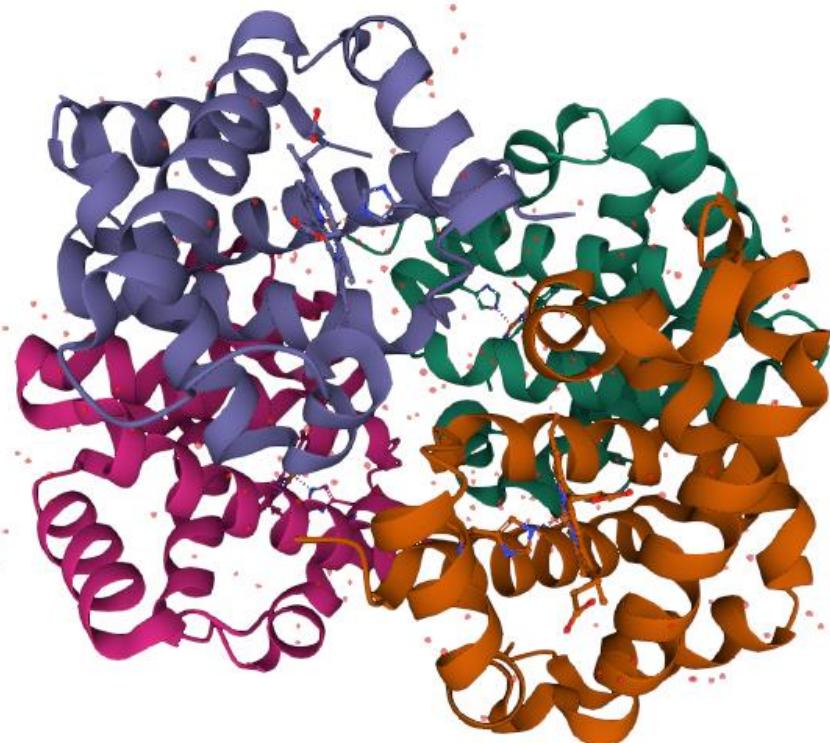
HEMOGLOBIN (DEOXY) (BETA CHAIN) (protein)

Unique Ligands HEM Download File View File

2DN2

1.25A resolution crystal structure of human hemoglobin in the deoxy form

Sequence of	2DN2 1.25A r...	Chain	#	1: Hemoglobi...	#	A #
1	VLSPADKINV 11 KAAANGKVGLH 21 AGEYGAEAL 31 RMFLSFPTTK 41 TYFPHFDSLH 51 GSAQVKGHGK 61 KVADALTNAV 71 AHVDDMPNAL 81 SALSDLHAKH 91 LRVDPVNFKL 101 LS	HCLLVTLA 111 AHLPAEFTPA 121 VHASLDKFIA 131 SVSTVLTSKY 141 R				

Display Files Download Files

FASTA Sequence

PDB Format

PDB Format (gz)

PDBx/mmCIF Format

PDBx/mmCIF Format (gz)

PDBML/XML Format (gz)

Biological Assembly 1

Structure Factors (CIF)

Structure Factors (CIF - gz)

Validation Full PDF

Validation XML

fo-fc Map (DSN6)

2fo-fc Map (DSN6)

Map Coefficients (MTZ format)

Water

Ball & Stick

Unit Cell P 1 21 1

Density

Assembly Symmetry

Export Animation

选择特定格式文件进行下载

【实例6】GEO数据库展示

在NCBI GEO数据库中，目前共收录了大量已公开的基因表达谱数据（RNA表达水平数据）。

The screenshot shows the NCBI GEO homepage. At the top, there's a navigation bar with links for NCBI Resources, How To, Sign in to NCBI, and menu items for GEO Home, Documentation, Query & Browse, and Email GEO. Below the header, the title "Gene Expression Omnibus" is displayed, along with a brief description of what GEO is: "GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles." To the right of the description is the GEO logo and a search bar labeled "Keyword or GEO Accession" with a "Search" button. On the left, there's a sidebar titled "Getting Started" with links to Overview, FAQ, About GEO DataSets, About GEO Profiles, About GEO2R Analysis, How to Construct a Query, and How to Download Data. Next to it is a "Tools" section with links to various resources like Search for Studies at GEO DataSets, Search for Gene Expression at GEO Profiles, and Analyze a Study with GEO2R. On the right, a large section titled "Browse Content" is highlighted with a red dashed border. It contains statistics: DataSets: 4348, Series: 160420, Platforms: 22575, and Samples: 4625052. Below this are links to MIAME Standards, Citing and Linking to GEO, Guidelines for Reviewers, and GEO Publications.

Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

Browse Content

- Repository Browser
- DataSets: 4348
- Series: 160420
- Platforms: 22575
- Samples: 4625052

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines

- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications

截止2022/01/01的统计数据



GEO

Gene Expression Omnibus

HOME | SEARCH | SITE MAP

NCBI Handbook Chapter

NAR 2002 Paper

FAQ

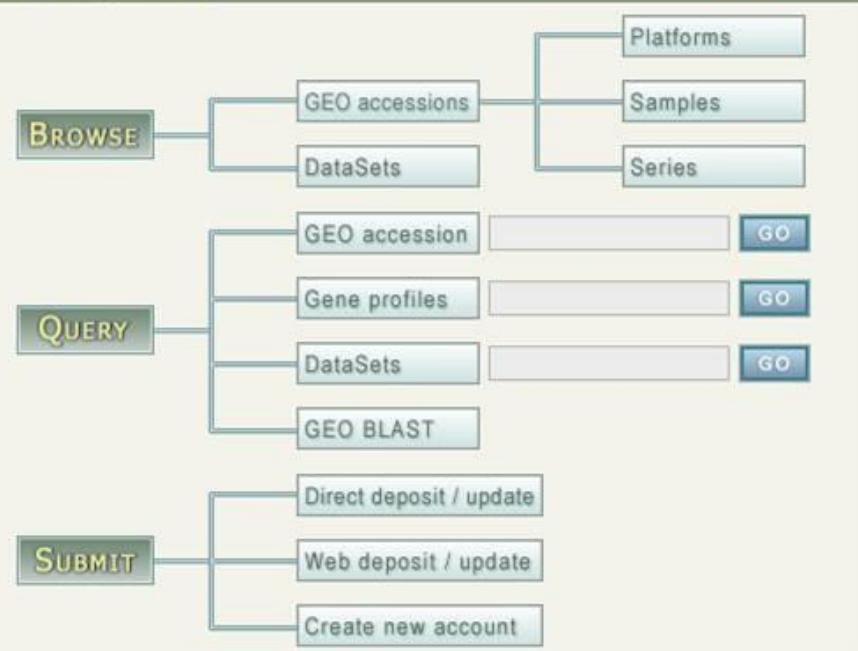
MIAME

Email GEO

NCBI > GEO

The **Gene Expression Omnibus** is a high-throughput gene expression / molecular abundance data repository, as well as a curated, online resource for gene expression data browsing, query and retrieval. GEO became operational in July 2000.

GEO navigation



GEO navigation help window. Mouse over flow chart for information.

public data

GPL Platforms 847
GSM Samples 23380
GSE Series 1221
Total 25448

Sep 23 2004

Site contents

Documentation

[Overview](#) | [FAQ](#)
[Web deposit brief](#)
[Batch deposit guide](#)
[SOFT examples](#)
[Linking & citing](#)
[GDS Clustering](#)
[GEO announce list](#)
[Data disclaimer](#)
[GEO staff](#)

Query & Browse

[DataSet browser](#)
[Repository browser](#)
[SAGEmap](#)
[FTP site](#)
[GEO Profiles](#)
[GEO Datasets](#)

Deposit & Update

[Web deposit](#)
[Direct deposit](#)
[New account](#)

External Tools

coming soon

Submit and update data

Query the database:

- gene identifiers
- field information
- sequence

Browse datasets

Download data

Retrieve GEO accession

Scope:

In:

view:

Depositors only

User :

Password :

Unlogged



*Submitted by
Manufacturer**

*Submitted by
Experimentalists*

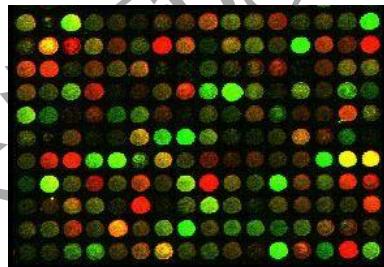
*Curated by
NCBI*

GPL
Platform
descriptions



GSM

Raw/processed
spot intensities
from a single
slide/chip
“a single experiment”



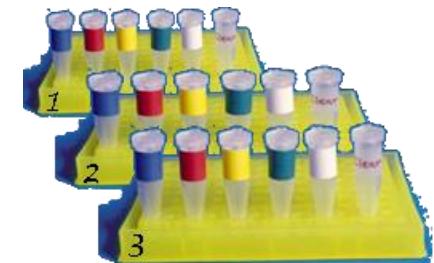
GSE

Grouping of
slide/chip data
“a single experiment”



GDS

Grouping of
experiments



Entrez GEO

Entrez GEO Datasets



Common Species in GEO database



Organism	Series	Platforms	Samples
<i>Homo sapiens</i>	70,455	5,897	2,383,630
<i>Mus musculus</i>	49,767	2,576	1,321,462
<i>Rattus norvegicus</i>	4,669	638	112,170
<i>Drosophila melanogaster</i>	4,220	391	100,249
<i>Arabidopsis thaliana</i>	4,850	405	74,384
<i>Saccharomyces cerevisiae</i>	3,256	625	73,025
<i>Macaca mulatta</i>	595	73	25,881
<i>Sus scrofa</i>	1,021	158	23,852
<i>Caenorhabditis elegans</i>	1,992	215	21,441
<i>Bos taurus</i>	1,081	204	19,847
<i>Gallus gallus</i>	894	150	14,827
<i>Oryza sativa</i>	953	203	14,242
<i>Zea mays</i>	532	120	14,407
<i>Escherichia coli</i>	890	176	9,873
<i>Glycine max</i>	318	58	8,938
<i>Xenopus laevis</i>	242	38	2,872

【实例7】化合物数据资源

Entrez PubChem

The screenshot shows the NCBI PubChem homepage. At the top, there is a navigation bar with links to PubMed, Entrez, PubChem, OMIM, Books, TaxBrowser, and Entrez Structure. Below the navigation bar is a search bar containing the text "Search Entrez PubChem Compound for zidovudine". To the right of the search bar is a "Go" button. On the left side of the page, there is a sidebar with links to "PubChem Substance", "PubChem Compound", "PubChem BioAssay", and "Research". The main content area features a purple header "The NCBI PubChem Project" followed by several paragraphs of text describing the database's purpose, search capabilities, and organization.

PubChem Substance

PubChem's Substance Database

PubChem Compound

PubChem's Compound Database

PubChem BioAssay

PubChem's Bioactivity Database

Research

Research topics and staff

PubChem

National Library of Medicine NLM

PubMed Entrez PubChem OMIM Books TaxBrowser Entrez Structure

Search Entrez PubChem Compound for **zidovudine** Go

▶ The NCBI PubChem Project

PubChem contains the chemical structures of small organic molecules and information on their biological activities. It is intended to support the Molecular Libraries and Imaging component of the [NIH Roadmap Initiative](#).

PubChem's chemical structure database may be searched on the basis of descriptive terms, chemical properties, and structural similarity. When possible, PubChem's chemical structure records are linked to other NCBI databases. These include the [PubMed](#) scientific literature database, for example, and NCBI's [protein 3D structure database](#). PubChem also contains the results of high-throughput biological screening experiments.

PubChem is organized as three linked databases within the [Entrez/PubMed](#) information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. More information about using each component database may be found by following the links to each:



HOME SEARCH SITE MAP

PubMed

Entrez

Structure

GenBank

PubChem

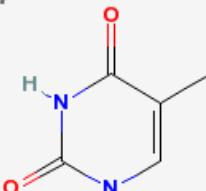
Help

Search

PubChem Compound

GO

Compound Summary:

Medical Subject Annotations: (Total: 4) [?](#)Display: [Next 1](#) | [All](#)CID: 35370 [?](#)Substances: [?](#)All: 238 Links
Same: 33 Links
Mixture: 205 LinksBioActivity: 4 Links [?](#)**Zidovudine**

A dideoxynucleoside compound in which the 3'-hydroxy group on the sugar moiety has been replaced by an azido group. This modification prevents the formation of phosphodiester linkages which are needed for the completion of nucleic acid chains. The compound is a potent inhibitor of HIV replication, acting as a chain-terminator of viral DNA during reverse transcription. It improves immunologic function, partially reverses the HIV-induced neurological dysfunction, and improves certain other clinical abnormalities associated with AIDS. Its principal toxic effect is dose-dependent suppression of bone marrow function.

[Show MeSH Tree Structure](#)

Pharmacological Action:

[Antimetabolites](#)
[Reverse Transcriptase Inhibitors](#)
[Anti-HIV Agents](#)
Depositor-Supplied Synonyms: (Total: 144) [?](#)Properties Computed from Structure: [?](#)

Molecular Weight: 267.242 g/mol

Molecular Formula: C₁₀H₁₃N₅O₄

Hydrogen Bond Donor Count: 2

Hydrogen Bond Acceptor Count: 6

Rotatable Bond Count: 3

Tautomer Count: 3

[zidovudine](#)
[Retrovir](#)
[Azidothymidine](#)
[Combivir](#)
[Trizivir](#)
[Compound S](#)
[Propolis+AZT](#)
[Retrovir \(TN\)](#)
[Zidovudinum \[Latin\]](#)
[Zidovudina \[Spanish\]](#)
Display: [Next 10](#) | [All](#) | Sort: [Weight](#) Descriptors Computed from Structure: [?](#)

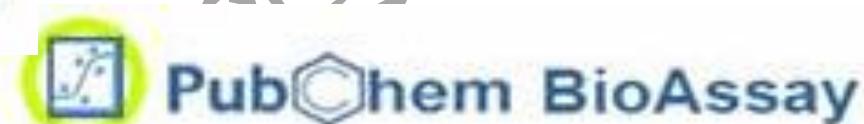
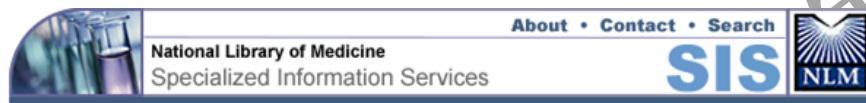
IUPAC Name: 1-[4-azido-5-(hydroxymethyl)oxolan-2-yl]-5-methyl-pyrimidine-2,4-dione

Isomeric SMILES: CC1=CN(C(=O)NC1=O)[C@H]2C[C@@H]([C@H](O2)CO)N=[N+]=[N-]

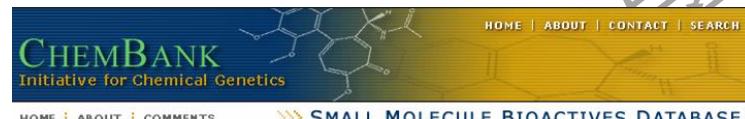
Canonical SMILES: CC1=CN(C(=O)NC1=O)C2CC(C(O2)CO)N=[N+]=[N-]

InChI: InChI=1/C10H13N5O4/c1-5-3-15(10(18)12-9(5)17)8-2-6(13-14-11)7(4-16)19-8/h3,6-8,16H,2,4H2,1H3,(H,12,17,18)/t6-,7+,8+/m0/s1/f/h12H [?](#)

PubChem: Compound, Substance, BioAssay



Developmental
Therapeutics Program
NCI/NIH

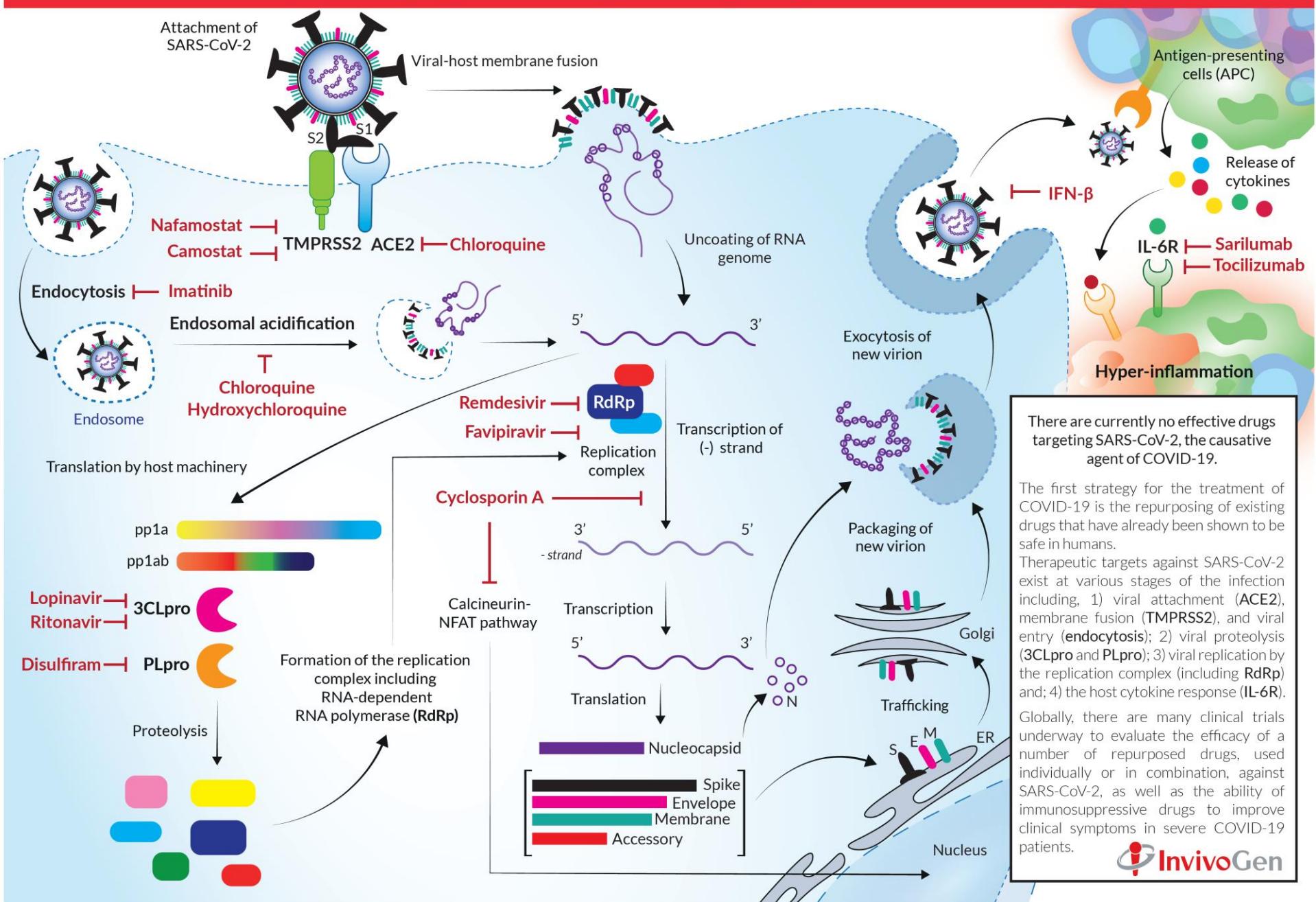


KEGG - Table of Contents



and more...

Repurposing approved drugs for targeting SARS-CoV-2



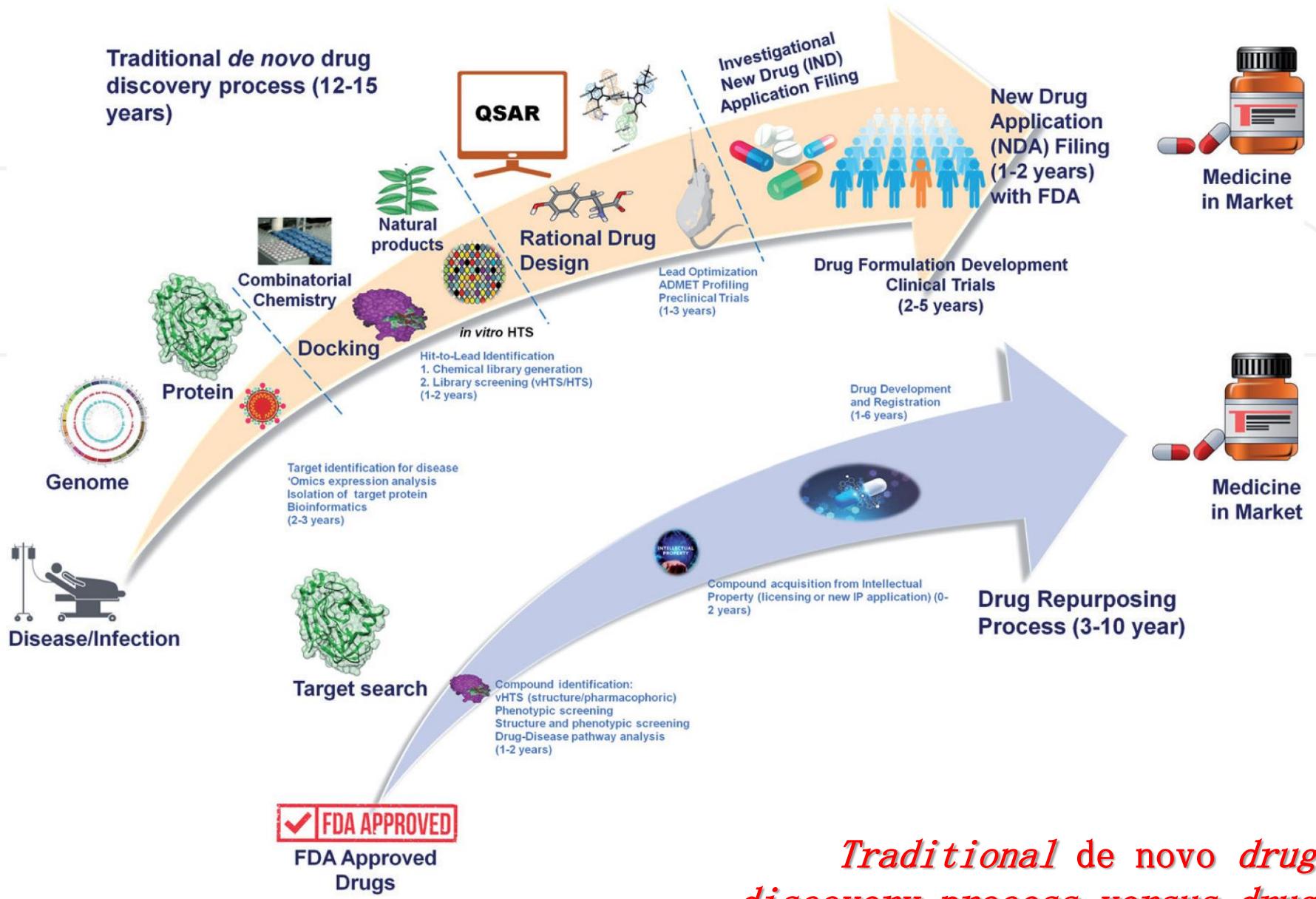
There are currently no effective drugs targeting SARS-CoV-2, the causative agent of COVID-19.

The first strategy for the treatment of COVID-19 is the repurposing of existing drugs that have already been shown to be safe in humans.

Therapeutic targets against SARS-CoV-2 exist at various stages of the infection including, 1) viral attachment (**ACE2**), membrane fusion (**TMRSS2**), and viral entry (**endocytosis**); 2) viral proteolysis (**3CLpro** and **PLpro**); 3) viral replication by the replication complex (including **RdRp**) and; 4) the host cytokine response (**IL-6R**).

Globally, there are many clinical trials underway to evaluate the efficacy of a number of repurposed drugs, used individually or in combination, against SARS-CoV-2, as well as the ability of immunosuppressive drugs to improve clinical symptoms in severe COVID-19 patients.

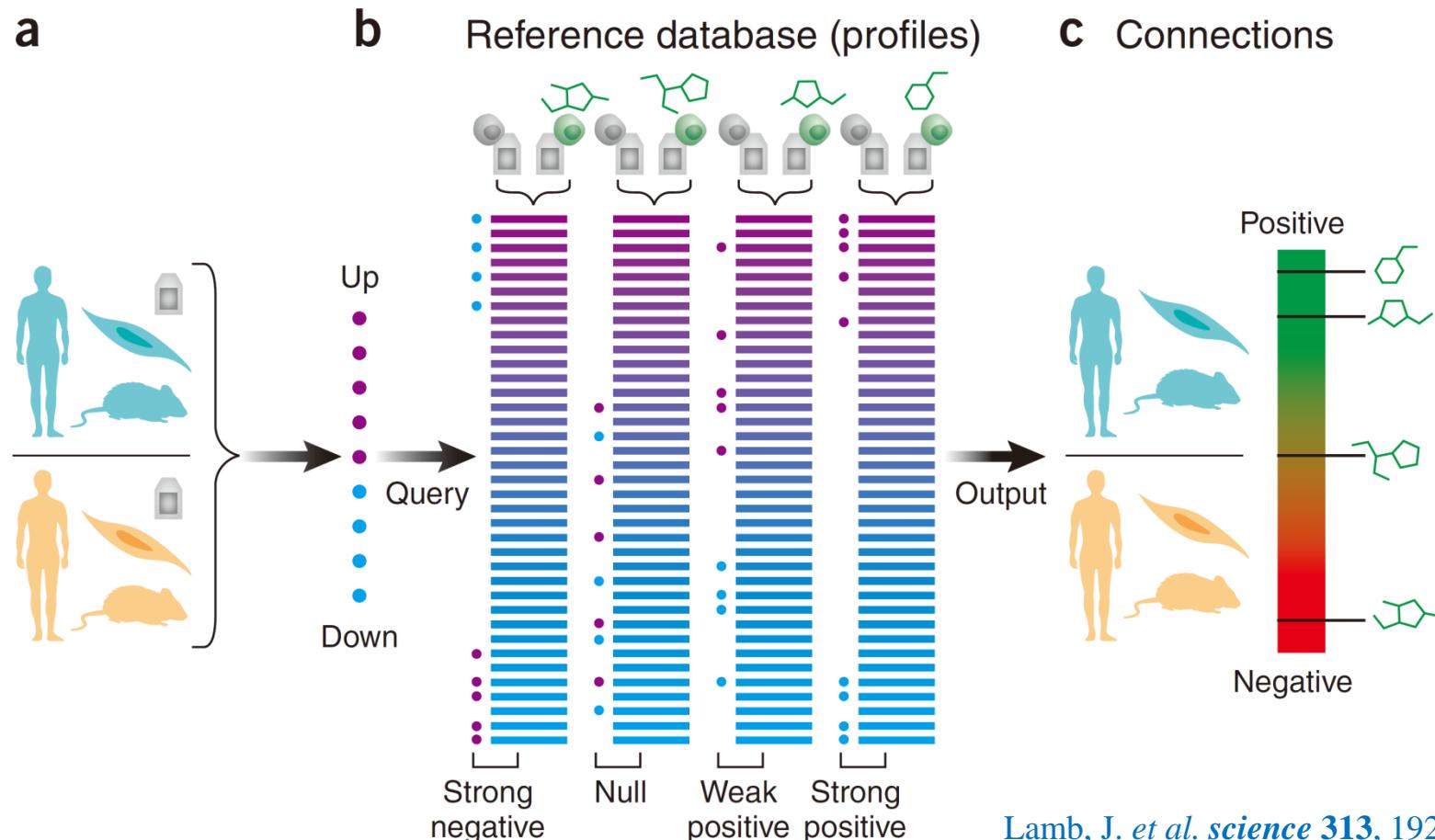
Traditional de novo drug discovery process (12-15 years)



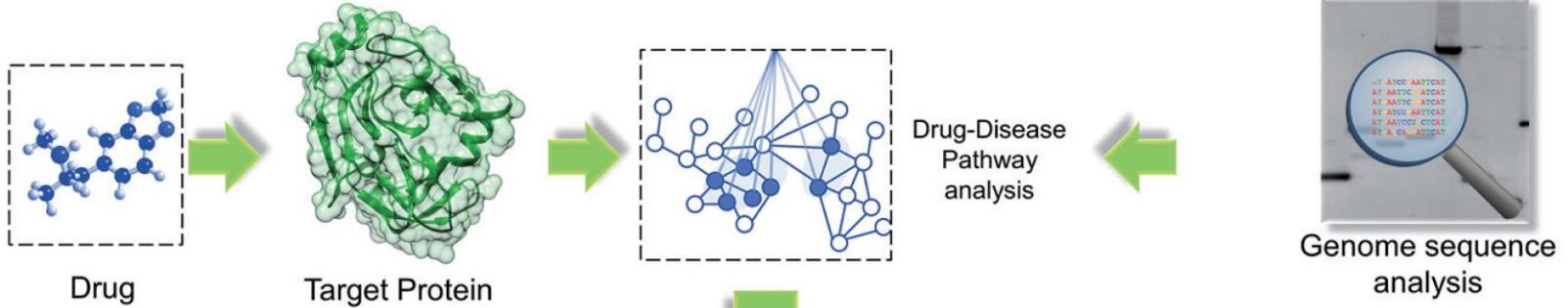
Traditional de novo drug discovery process versus drug repurposing process

CMap: a new tool for drug discovery

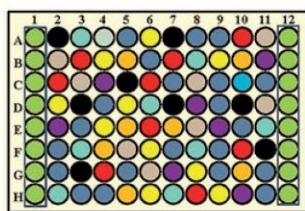
Connectivity Map: *using gene-expression signatures to connect small molecules, genes, and disease*



- Lamb, J. et al. *science* **313**, 1929-1935 (2006).
- Lamb, J. The. *Nature reviews cancer* **7**, 54 (2007).
- Subramanian, A. et al. *Cell* **171**, 1437-1452. e17 (2017).



Phenotype-Based Drug Repurposing



Lead identification

High throughput screening

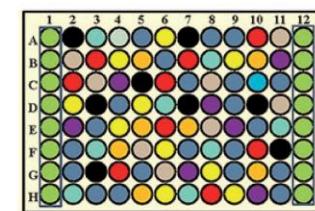
Identified leads

Pharmacophore Modeling
Based on previous Dataset

Poly-Pharmacology
Analysis

Drug repurposed for
New Indication and
Treatment

Structure and Phenotype-Based Drug Repurposing



Lead identification

High throughput screening

Microbial Protein
Target(s)

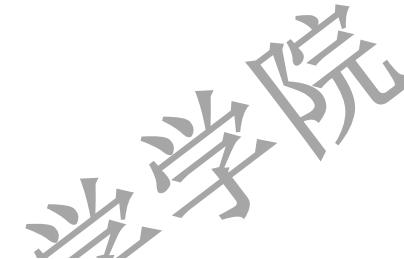
Virtual High
throughput screening

1861 FDA approved drugs
Experimental Knowledge-Based
Drug
Repositioning Database
(EK-DRD)
<http://www.idruglab.com/drdb/index.php>.

Drug repurposed for
new indication and
treatment

*Strategies for
drug
repurposing*

【实例8】从NCBI数据库大批量获取数据（借助FTP）



NCBI

FTP site

PubMed Entrez BLAST OMIM Books TaxBrowser Structure

Search Nucleotide for

Major resources available by ftp (<ftp.ncbi.nih.gov>):

- ▶ [BLAST Basic Local Alignment Search Tool](#)
For stand-alone sequence comparison software.
- ▶ [Cn3D](#)
Stand-alone software for viewing structures in three dimensions.
- ▶ [Data Repository](#)
Download collections of contributed molecular biology data.
- ▶ [GenBank](#)
Download the full release database or daily updates.

Note: there are mirror sites for GenBank files at the San Diego SuperCompute Center (genbank.sdsc.edu/pub) and at Indiana University (bio-mirror.net/biomirror/genbank).
- ▶ [Genome Assembly/Annotation Projects](#)
Download complete genomes/chromosomes, contigs and reference sequence mRNAs and proteins.

▶ [NCBI Toolbox](#)

NCBI software tools for building bioinformatics resources.

▶ [RefSeq](#)

Download curated RefSeq full release or daily updates.

▶ [Sequin](#)

Stand-alone GenBank sequence submission software.

▶ [SKY/CGH](#)

Download the SKY and CGH database.

▶ [dbSNP](#)

Download the SNP database.

▶ [Taxonomy](#)

Download data files from the Taxonomy database.

▶ [UniGene](#)

Download data files from the UniGene datasets.

▶ [UniSTS](#)

Download data files from the UniSTS resource.



NCBI Toolbox: In-house source code useful for incorporating NCBI-like functionality into their programs.

Three main parts: Data Model, Data Encoding and Programming Libraries.

- Examples: BLAST, Cn3D, Sequin, Data format conversion scripts

<http://www.ncbi.nlm.nih.gov/IEB/ToolBox/index.cgi>

E-Utilities: Guidelines for Entrez “URL calls” used to access data.

Designed for use in scripts.

- Examples: ESearch, EPost, ESummary, EFetch and ELink

http://www.ncbi.nih.gov/entrez/query/static/eutils_help.html

Caution: Overuse may result in blocked IPs!

Software	Description
<i>rentrez</i>	An R package for the NCBI eUtils API
<i>biodbNcbi</i>	a library for connecting to NCBI Databases
<i>ExpressionAtlas</i>	Download datasets from EMBL-EBI Expression Atlas
<i>gwascat</i>	representing and modeling data in the EMBL-EBI GWAS catalog
<i>prideR</i>	obtain data from the EMBL-EBI PRIDE Archive and PRIDE Cluster)
<i>BiomaRt</i>	quick, easy and powerful way to access BioMart right from your R software terminal
<i>biofiles</i>	An Interface for GenBank/GenPept Flat Files
<i>GEOquery</i>	
<i>SRAdb</i>	A compilation of metadata from NCBI SRA and tools
<i>ensemblDb</i>	Utilities to create and use Ensembl-based annotation databases
<i>proActiv</i>	Estimate Promoter Activity from RNA-Seq data
<i>TSRchitect</i>	Promoter identification from large-scale TSS profiling data
<i>CAGER</i>	Analysis of CAGE (Cap Analysis of Gene Expression) sequencing data for precise mapping of transcription start sites and promoterome mining
<i>UniprotR</i>	Retrieving and visualizing protein sequence and functional information from Universal Protein Resource
<i>AssessORF</i>	Assess Gene Predictions Using Proteomics and Evolutionary Conservation
<i>LedPred</i>	Learning from DNA to Predict Enhancers
<i>ORFhunteR</i>	Predict open reading frames in nucleotide sequences
<i>ORFik</i>	analysis of transcript and translation features through manipulation of sequence data and NGS data like Ribo-Seq, RNA-Seq, TCP-Seq and CAGE

第3节：EMBL数据库与数据资源

The screenshot shows the EMBL-EBI website homepage. At the top, there's a navigation bar with links to EMBL-EBI, Services, Research, Training, About us, and a green circular logo. A banner on the right displays statistics: "18% webpage Usage s~ wNb !n]erfaceM /h 2wh61" and "More about EMBL-EBI's impact in our annual report > Data from 2016". The main content area features a large search bar with placeholder text "Find a gene, protein or chemical" and a search icon. Below the search bar, there are three main sections: "Our unique Search service helps you explore dozens of biological data resources. More about EBI Search >" (with a link to "blast keratin bII1"), "Find a tool for your data analysis. Find a tool >" (with a link to "Deposit data >"), and "Share your scientific data with the world. Deposit data >". The background of the page features a green globe graphic.

We are EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI) is part of EMBL, Europe's flagship laboratory for the life sciences. More about EMBL-EBI and our impact. >

Training

Access a wealth of world-leading training in bioinformatics and scientific service provision, regardless of your career stage or sector >

Latest news

Research highlights, service updates and more



18 Sep 2018

LifeLab - free events highlight discovery on your doorstep

Data resources

Explore our open data resources to enrich your research. Browse data, perform analyses or share your own results. >

Industry

Explore our knowledge-exchange Industry Programme and take part in translational partnerships and projects >

Our events

Thurs 27th Sep | Seminar

Deep Learning Double Bill - Dr Antony Rix & Dr Tom Whitehead

In this double bill, they will talk about the potential use of AI in a range of life science applications and when using sparse data.

Tues 2nd Oct - Fri 5th | Course

Introduction to Next Generation Sequencing

This course will provide an introduction to the technology, data analysis, tools and resources for next generation sequencing (NGS) data.

Wed 3rd Oct | Course

Ensembl Browser Workshop, Rabat, 3 October 2018

The Ensembl project at www.ensembl.org provides a comprehensive and integrated source of annotation of mainly

Research

Find out about our research groups, postdoctoral schemes and PhD Programme >

ELIXIR

We support, as an ELIXIR node, the coordination of biological data provision throughout Europe >



- 欧洲生物信息学研究所(**European Bioinformatics Institute**)
- 1994年建于英国剑桥,前身是德国海德堡的欧洲分子生物学实验室的信息服务部门
- EBI接收了原来**EMBL**数据库的管理和维护
- 是欧洲分子生物学网(EMBnet)的一个特别节点
- <http://www.ebi.ac.uk/> (主页)
- <http://www2.ebi.ac.uk/> (工具)
- <http://www3.ebi.ac.uk/> (服务)

RESEARCH UNITS

- › Cell Biology and Biophysics
- › Developmental Biology
- › Directors' Research
- › Genome Biology
- › Structural and Computational Biology
- › Research at other EMBL Sites

SEMINARS

EMBL AND THE TARA OCEAN
FOUNDATION

VENTURE CAPITAL

INTERDISCIPLINARY RESEARCH

- › Bioinformatics at EMBL
- › Chemistry at EMBL
- › Physics and Engineering at EMBL
- › Mathematics and Statistics at EMBL
- › EMBL Centres

FACULTY

PARTNERSHIPS

LIFE SCIENCE ALLIANCE

TECHNOLOGY TRANSFER

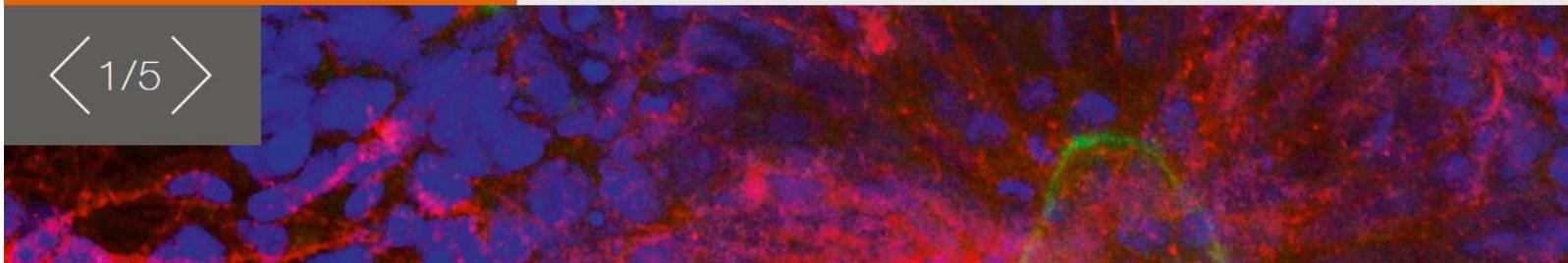


EMBL International PhD
Programme:
Winter recruitment 2020

TRAINING

READ MORE >

< 1/5 >





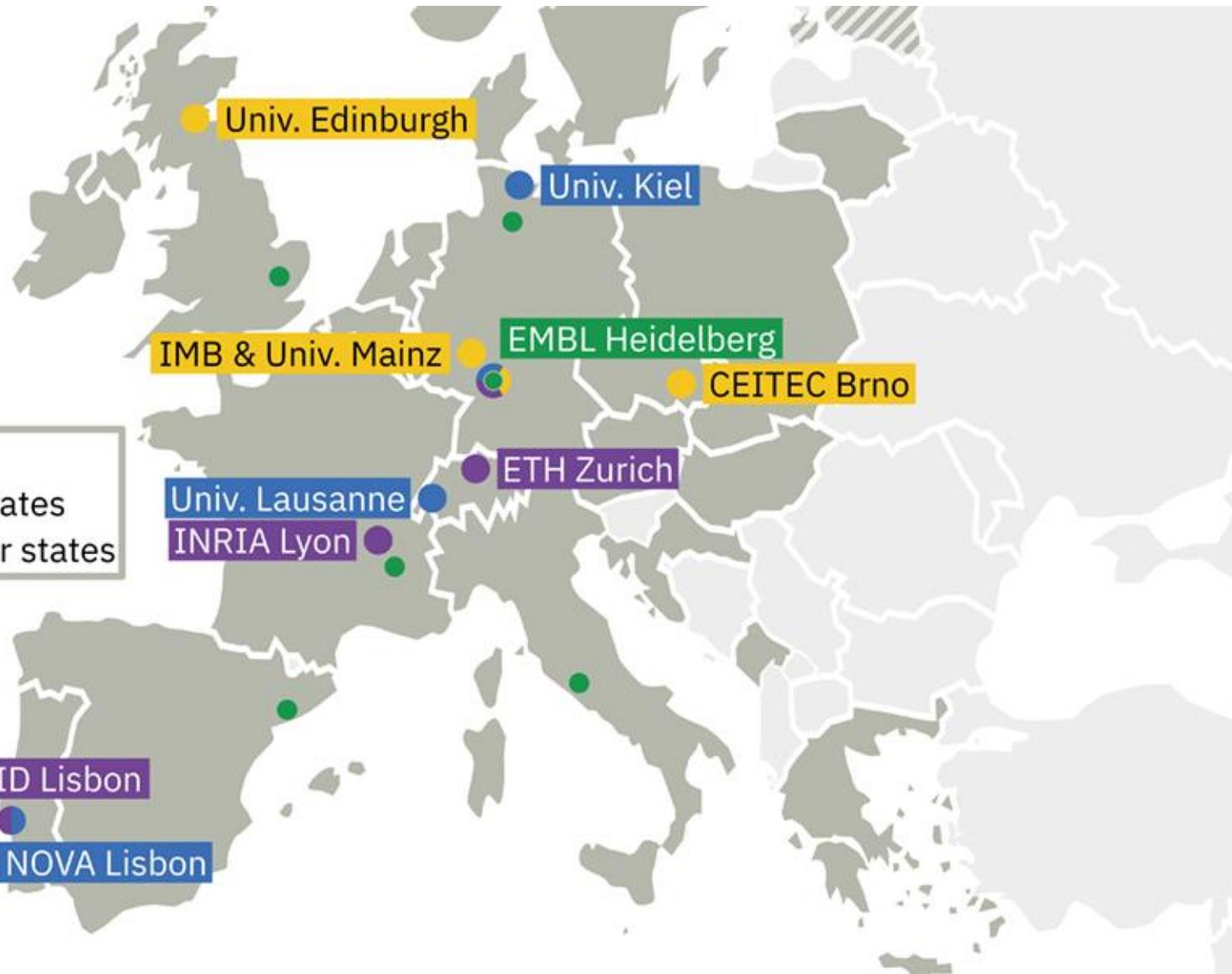
EMBL的六处的节点

Locations



Twinning projects
OLISSIPO
SymbNET
INTEG-RNA

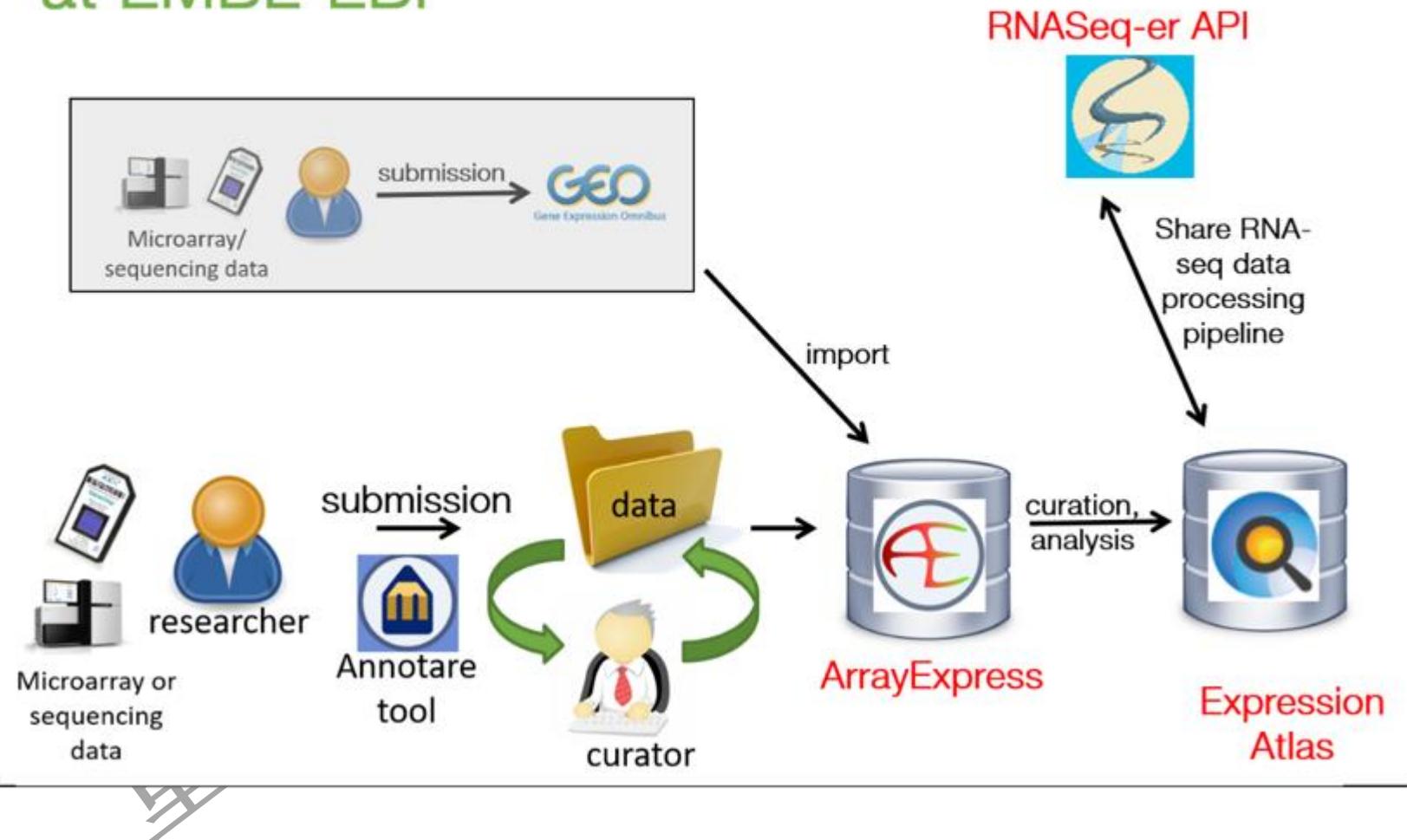
- EMBL sites
- EMBL member states
- Prospect member states



EU funding Archives | EMBL



Summary of functional genomics resources at EMBL-EBI





Ensembl

Genome browser, API and database,
providing access to reference genome annotation



UniProt

A comprehensive resource for protein
sequence and functional annotation.



PDBe

The European resource for the collection,
organisation and dissemination of 3D structural data (from PDB
and EMDB) on biological macromolecules and their complexes.



Europe PMC

A database to search the worldwide life
sciences literature



Expression Atlas

An added-value database that shows which
genes/proteins are expressed under which conditions, and how
expression differs between conditions.



ChEMBL

An open data resource of binding, functional
and ADMET bioactivity data.



[See all data resources ➤](#)

<https://www.ebi.ac.uk/services/all>

第4节：UCSC基因组浏览器与数据资源

UCSC Genome Browser



UCSC Genome Browser是由University of California Santa Cruz (UCSC)创立和维护的，该站点包含有人类、小鼠和大鼠等多个物种的基因组草图，并提供一系列的网页分析工具。(可以在任何尺度上快速查询和显示基因组内容。)

The screenshot shows the UCSC Genome Browser Gateway homepage. At the top, there are logos for the University of California Santa Cruz and UCSC. Below them is the title "Genome Browser Gateway". A navigation bar includes links for Home, Genomes, Genome Browser, Tools, Mirrors, Downloads, My Data, Help, and About Us. The main interface has two main sections: "Browse>Select Species" on the left and "Find Position" on the right. Under "Browse>Select Species", there is a "POPULAR SPECIES" section with icons for Human, Mouse, Rat, Fruitfly, Worm, and Yeast, each with a red arrow pointing to it. Below this is a search bar with the placeholder "Enter species or common name". Under "Find Position", there is a dropdown menu set to "Human Assembly Dec. 2013 (GRCh38/hg38)" with a red arrow pointing to it. Below the dropdown is a "Position/Search Term" input field containing "Enter position, gene symbol or search terms" and a "GO" button. A red arrow points to the input field. At the bottom left, there is a "REPRESENTED SPECIES" section with a small icon.

如果想将测序的reads对于到染色体上，可以用UCSc Genome Browser或IGV.

Organization of Genomic Data

sequence

Annotation Tracks

Genome backbone: base position number

chromosome band

sts sites

gap locations

known genes

predicted genes

*Links out to
more data*

microarray/expression data

evolutionary conservation

SNPs

repeated regions

more...

A Sample of the UCSC Genome Browser

UCSC Genome Browser on Human Mar. 2006 Assembly

move <<< << < > >> zoom in 1.5x

position/search chr17:7,512,445-7,531,642

Annotation Tracks

Human Gene TP53 Description and Page Index

Description: tumor protein p53
Alternate Gene Symbols: AB082923, AF307851, BC003596, DQ191317, DQ286964, K03199, P53
Representative RefSeq: NM_000546 Protein: P04637 (aka P53_HUMAN)

RefSeq Summary: Tumor protein p53, a nuclear protein, plays an essential role in the regulation of cell cycle, specifically in the transition from G0 to G1. It is found in very low levels in normal cells, however, in a variety of transformed cell lines, it is expressed in high amounts, and believed to contribute to transformation and malignancy. p53 is a DNA-binding protein containing DNA-binding, oligomerization and transcription activation domains. It is postulated to bind as a tetramer to a p53-binding site and activate expression of downstream genes that inhibit growth and/or invasion, and thus function as a tumor suppressor. Mutants of p53 that frequently occur in a number of different human cancers fail to bind DNA, and therefore do not activate the p53 pathway.

Vertebrate Multiz Alignment & PhastCons Conservation (28 Species)

Capitalize coding exons based on RefSeq Genes show all bases

Place cursor over species for alignment detail. Click on 'B' to list variants.

Components not displayed:
Cat Shrew Cow Elephant GuineaPig Medaka Tetraodon

Alignment block 1 of 168 in window, 7512445 - 7512484, 40 bps

B	D	Human	cac-ccctcagcac---acagggtggcag-caaaattttttggta
B	D	Rhesus	NN
B	D	Mouse	aac-cttc-aaaaccataatgtataaa-aaaaattttttttgtta
B	D	Dog	cccgccggacac---acacatgttag-taaagtttttttgtta
B	D	Horse	cat-cctgcagcac---acagggtgttag-aaaaattttttttgtta
B	D	Armadillo	cgc-tcttgagacac---acagggtgttag-aaaaatgtttttgtt-
B	D	Opossum	
B	D	Platypus	

Simple Nucleotide Polymorphisms (dbSNP build 128)

dbSNP build 128 rs1042522

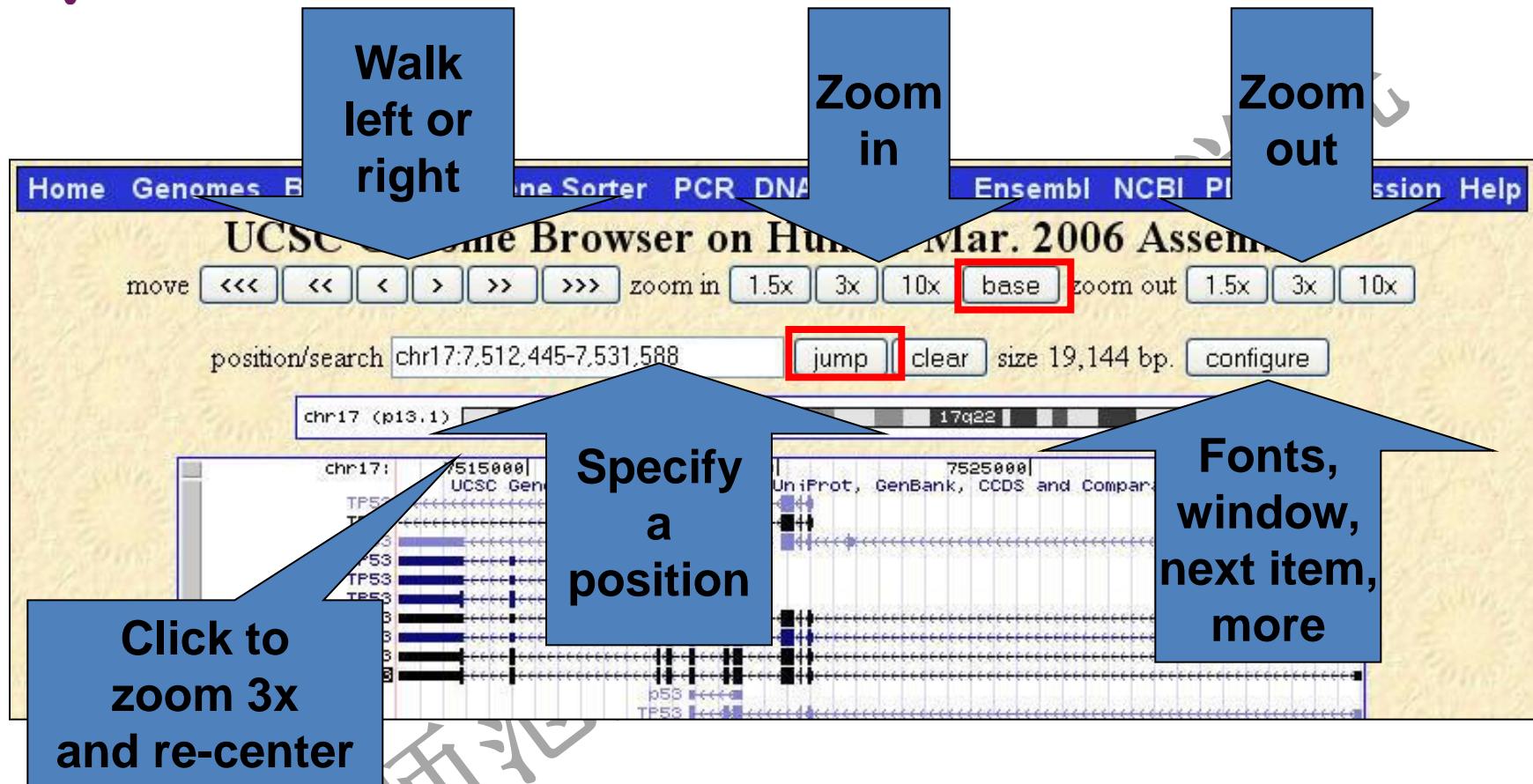
dbSNP: rs1042522
Position: chr17:7520197-7520197
Build: 17p13.1
Genomic Size: 1
View DNA for this feature

Summary: C>C/G (chimp allele displayed first, then '>', then human alleles)
Strand: -
Observed: C/G
Reference allele: C
Chimp allele: C Chimp strand: - Chimp position: chr17:7857509-7857509
Macaque allele: C Macaque strand: - Macaque position: chr16:7407095-7407095

Class: single
Validation: by-frequency,by-2hit-2allele,by-hapmap
Function: cds-reference,missense
Molecule Type: genomic
Average Heterozygosity: 0.492 +/- 0.065
Weight: 1

HapMap SNP

Options for Changing Images: Upper Section



- Next item: Change your view or location with controls at the top
- Use “base” to get right down to the nucleotides
- Configure: to change font, window size, more...
 - m, next exon navigation assistance can be turned on

IGV is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

The screenshot shows the IGV website's homepage. At the top left is the IGV logo and navigation links for Home, Downloads, and Documents. Below this is a sidebar with links for Hosted Genomes, FAQ, IGV User Guide, File Formats, Release Notes, IGV for iPad, Credits, Contact, and a search bar. The main content area features a large title "Integrative Genomics Viewer" with a background image of a genome browser interface. Below the title are sections for Overview, Downloads, and Funding, each with descriptive text and icons.

Home

Integrative Genomics Viewer

Overview

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. It supports a wide variety of data types, including array-based and next-generation sequence data, and genomic annotations.

Citing IGV

To cite your use of IGV in your publication:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011)

Downloads

Download the IGV desktop application and igvtools.

Funding

Development of IGV is made possible by funding from the National Cancer Institute, the National Institute of General Medical Sciences of the National Institutes of Health, and the Starr Cancer Consortium.

IGV participates in the [GenomeSpace](#) initiative, which is funded by the [National Human Genome Research Institute](#).



重庆师范大学
CHONG QING NORMAL UNIVERSIYT

Thanks for your attention!

Acknowledgement

College of Life Sciences, Chongqing Normal University
2022, Chongqing of P. R. C