

第四章 生物信息资源

袁猛, 章张, 张子丁, 胡德华

章张

2011年至今
2019年至今



中国科学院北京基因组研究所（国家生物信息中心）

BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

2016年至今



中国科学院大学

University of Chinese Academy of Sciences

2019年至今



国家基因组科学数据中心 National Genomics Data Center

基因组学

2020年至今
本科生课程
36学时



基因组学 - 中国科学院大学 2024



基因组学 - 中国科学院大学 2023



B站课程视频

第一节 生物数据库简介

第二节 国际主要数据中心

第三节 国际重要生物数据库

第四节 生物数据库发展趋势

第一节 生物数据库简介



袁猛

华中农业大学

第三节 国际重要生物数据库



史文聿

中国农业大学



张子丁

第二节 国际主要数据中心

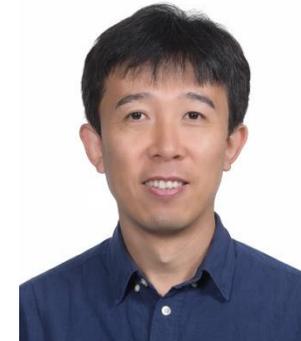


陈梅丽



陈婷婷

国家生物信息中心



章张

第四节 生物数据库发展趋势

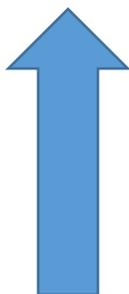


胡德华

中南大学

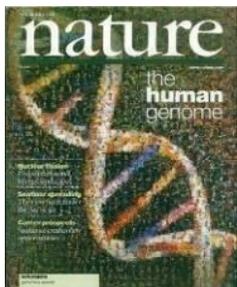
第一节 生物数据库简介

大数据驱动
研究范式转变

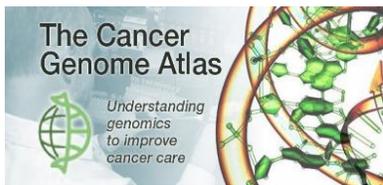


大科学计划
科学数据累积

91 PB
数据量*



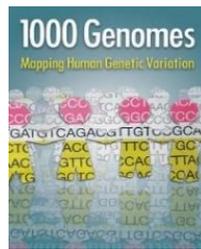
人类基因组计划
1990—2001



TCGA
2006-2018



ENCODE
2003—



千人基因组计划
2008—2015



万种鸟类基因组计划
2015—



大熊猫基因组计划
2008—2009



HUMAN
CELL
ATLAS

人类细胞图谱
2016—



万种原生生物基因组计划
2019—



地球生物基因组计划
2018—

一代测序技术



二代测序技术



三代测序技术



单细胞测序技术



测序技术

* NCBI SRA (as of July 2024)

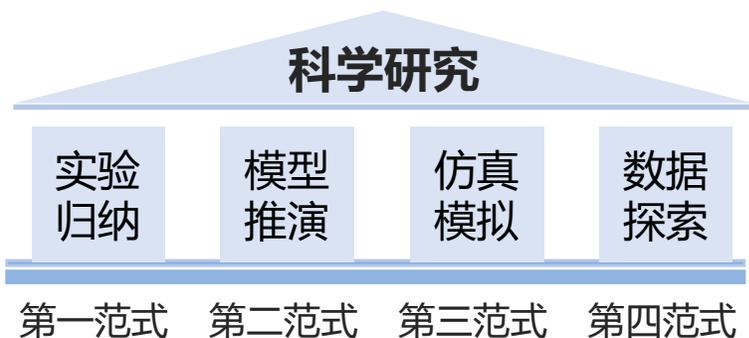
驱动生命科学研究范式变革



AlphaFold
Nature 2018

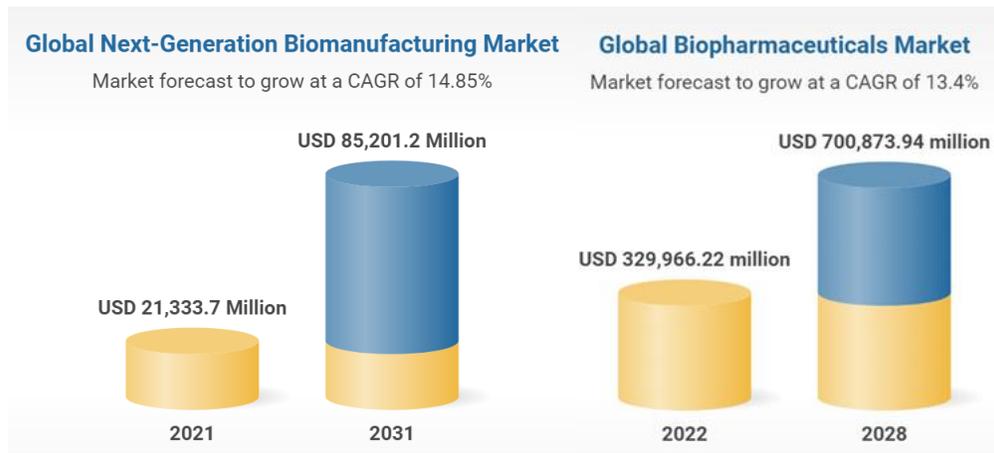


GeneFormer
Nature 2023



促进生物经济发展新质生产力

生物制造

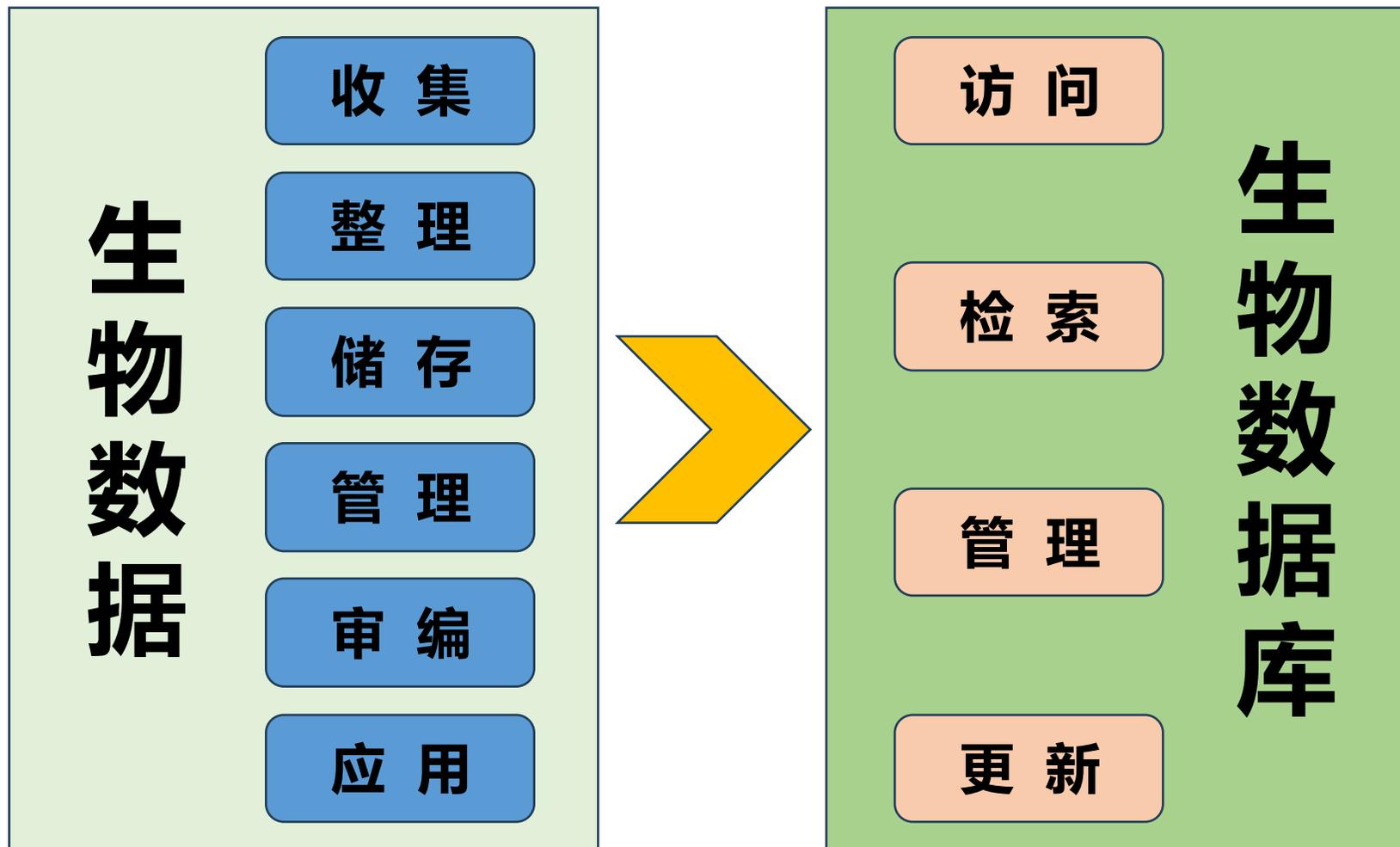


生物医药

2028年全球生物经济市场超过**7800亿美元**
(根据Research and Markets统计)



2025年
我国生物经济总量达
22万亿元



- 国家重要战略资源
- 科技发展基础设施



- 数据信息资源保藏
- 生物科技安全根基

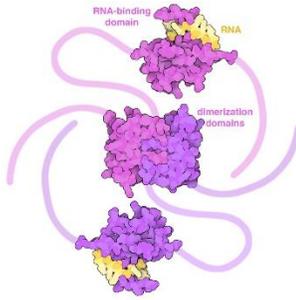
生物数据库

- **种类繁多**：序列、结构、文献、代谢、表达谱、基因组等
- **多层次**：数据库 → 信息库 → 知识库
- **交叉链接**：原始数据，分析结果，文献知识
- **“用户至上” + “数据至上”**：递交-存储-审核-质控-发布-共享

p53 [Penaeus japonicus]
GenBank: BAL15075.1
Identical Proteins: [FASTA](#) [GenBank](#)

Go to: [⊕](#)

LOCUS BAL15075 461 aa linear INV 22-NOV-2011
DEFINITION p53 [Penaeus japonicus].
ACCESSION BAL15075
VERSION BAL15075.1
SOURCE accession: [020909.1](#)
KEYWORDS
SOURCE Penaeus japonicus
ORGANISM Penaeus japonicus
Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Crustacea;
Mollusca; Malacostraca; Eumalacostraca; Eucarida; Decapoda;
Decapodae; Penaeoidea; Penaeidae; Penaeus.
REFERENCE
1 Yoshino M, Wakata T, Enno T, Sekai M and Itami T.
TITLE Identification of cDNA encoding p53 gene from kuruma shrimp.
JOURNAL *Maripenaeus Japonicus*
REFERENCE 2 (revidus 1 to 461)
1 Yoshino M, Wakata T, Enno T, Sekai M and Itami T.
TITLE Direct Submission
JOURNAL Submitted (06-MAY-2010) Contact: Maki Yoshino, Faculty of
Agriculture, 1-1-1 Sakai, Mihoshi, Miyazaki University,
Miyazaki, Miyazaki Pref 889-2192, Japan

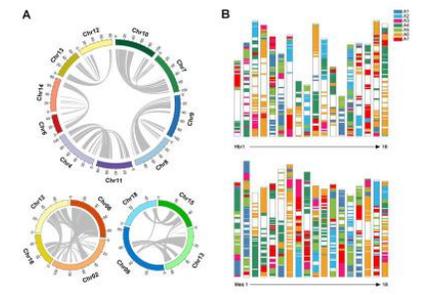
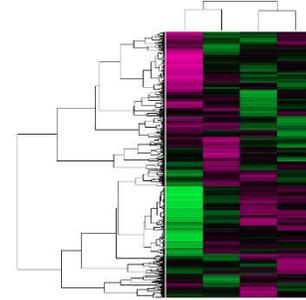
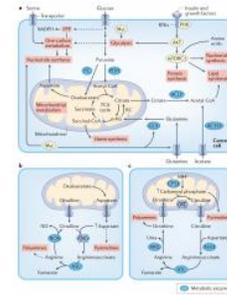


The Proteins of Severe Acute Respiratory Syndrome Coronavirus-2 (SARS CoV-2 or n-CoV19), the Cause of COVID-19:
Yoshimoto FK.
Protein J. 2020 Jun;39(3):198-216. doi: 10.1007/s10930-020-09901-4.
PMID: 32447571 Free PMC article. Review.

Fil-CoV19: A Transfer Learning Approach to Detect COVID-19.
Singh T, Saurabh P, Bisen D, Kane L, Pathak M, Sinha GR.
Comput Intell Neurosci. 2022 Jul;2022:1953992. doi: 10.1155/2022/1953992. eCollection 2022.
PMID: 3585453 Free PMC article.

This paper proposes a model fine tuning transfer learning-coronavirus 19 (Fi-CoV19) for COVID-19 detection through chest X-rays, which embraces the ideas of transfer learning in pretrained VGG16 model with including combination of convolution, max pooling, and dense layer ...

Hydroxychloroquine and azithromycin as a treatment of COVID-19: results of an open-label non-randomized clinical trial.
Gautret P, Lagier JC, Parola P, Hoang VT, Meddeb L, Mailhe M, Doudier B, Coujon J, Giordanengo V, Vieira VE, Tissot Dupont H, Honoré S, Colton P, Chabrière E, La Scola B, Rolain JM, Broutin P, Raouf D, Int J Antimicrob Agents. 2020 Jul;56(1):105948. doi: 10.1016/j.ijantimicag.2020.105948. Epub 2020 Mar 20.
PMID: 32205204 Free PMC article. Clinical Trial.



序列

结构

文献

代谢

表达谱

基因组

◆ 生物大分子数据库

- 综合型数据库
- DNA数据库
- RNA数据库
- 蛋白质数据库

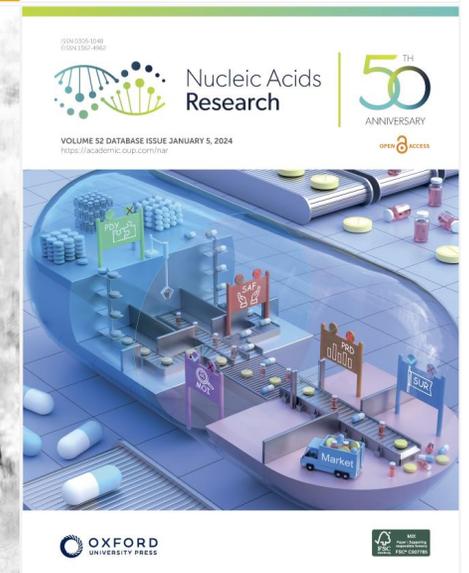
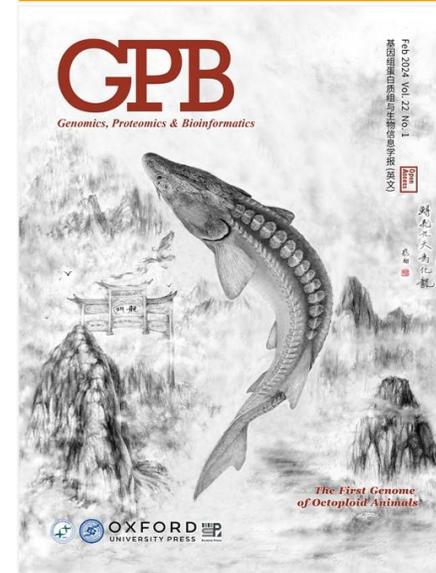
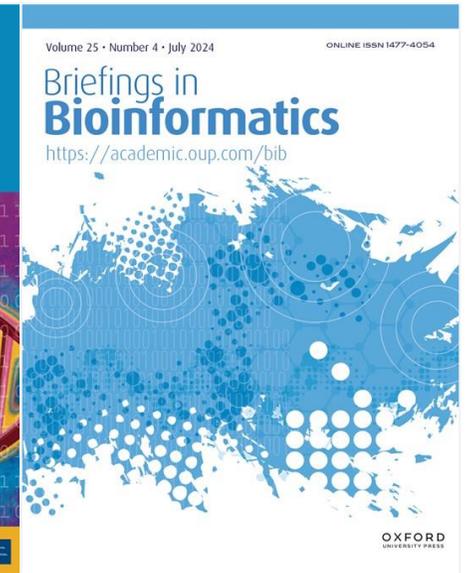
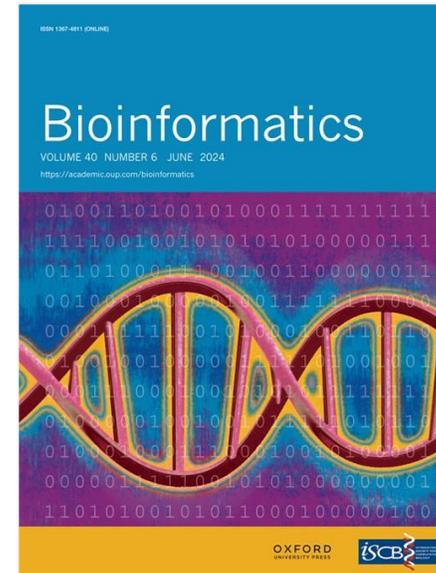
◆ 研究对象为主的数据库

- 病毒数据库
- 原核生物数据库
- 真核生物数据库
- 人与人类疾病类数据库
- 动物与动物模型数据库
- 植物数据库
- 微生物相关数据库
- 其它物种数据库

◆ 生物学其他数据库

- 生物互作数据库
- 细胞通讯数据库
- 代谢数据库
- 生物反应数据库
- 单细胞数据库
- 转录因子结合位点数据库
- 非编码RNA数据库

- ***Bioinformatics***
- ***Bioinformatics Advances***
- ***Briefings in Bioinformatics***
- ***BMC Bioinformatics***
- ***Current Bioinformatics***
- ***Database***
- ***Genomics Proteomics & Bioinformatics***
- ***Journal of Bioinformatics and Computational Biology***
- ***Molecular Plant***
- ***Nucleic Acids Research***
- ***NAR Genomics and Bioinformatics***
- ***Nature Genetics***



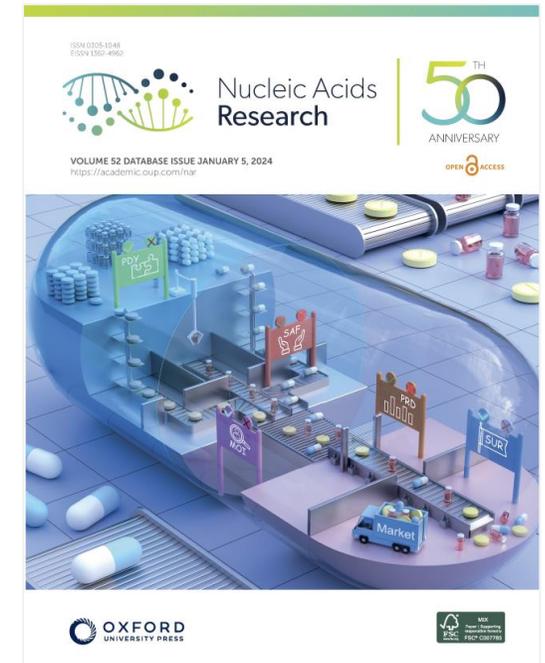
Nucleic Acids Research

◆ 数据库专辑 (1994~)

- 大型综合数据库
- 核苷酸序列、结构、调控类数据库
- 蛋白质序列、结构、结构域类数据库
- 代谢、信号途径、酶类数据库
- 病毒、细菌、原生动物、真菌类数据库
- 人类基因组、模式生物、比较基因组类数据库
- 基因组变异、疾病、药物类数据库
- 植物类数据库
- 其它类数据库

◆ 网络服务器专辑 (2003~)

- 计算机相关
- DNA
- 教育
- 表达
- 人类基因组
- 文献
- 模式生物
- 其他分子
- 蛋白质
- RNA
- 序列比较



第二节 国际主要数据中心

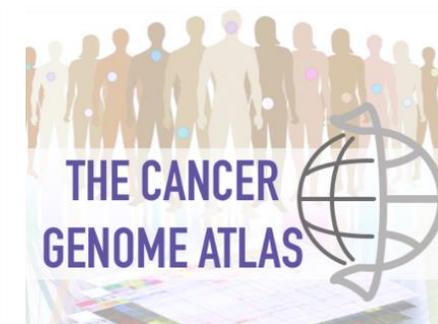
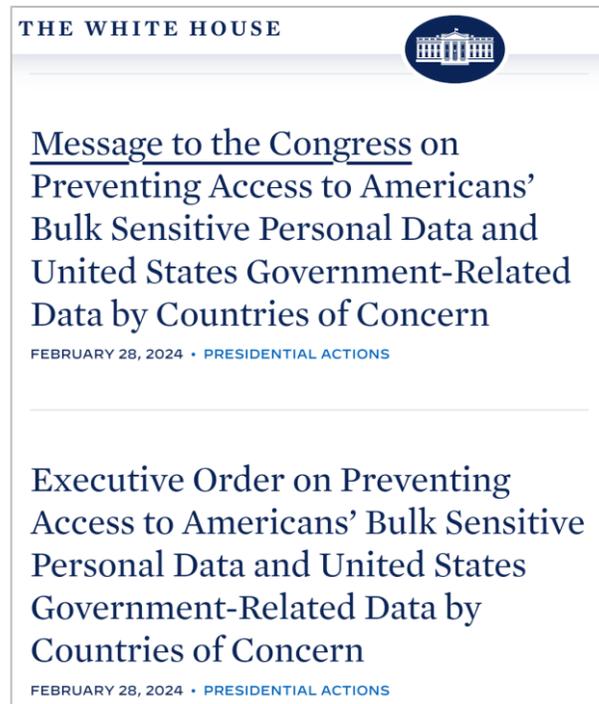
中心	美国国家生物技术信息中心 (NCBI)	欧洲生物信息研究所 (EBI)	日本DNA数据库 (DDBJ)
成立时间	1988年	1994年	1986年
人员规模 (人)	~ 700	~ 800	~ 50
年度经费 (万美元)	15,375	8,820	834
经费来源	美国政府	英国政府	日本文部省
日访问量 (万次)	2100	2700	<10

数据流失 数据主权 数据安全



- ❑ 欧美日垄断全球生物数据
- ❑ 全球生物数据中国贡献超过40%
- ❑ 依赖国外数据库，缺乏自主管理

数据受限 数据脱钩 数据冷战



All of Us
RESEARCH PROGRAM

biobank^{uk}

测序大国，数据弱国，加快自主安全可控的数据资源体系建设



1999

2014

2016

2018

2020

2022

2024

建议尽快组建国家级的生物医学信息中心

郝柏林

1999年6月10日

从细菌到人类，众多物种的基因和蛋白质数据正在以科学史上从未有过的高速度增长。目前已测定出二十多种细菌，以及比细菌更高等的一些物种如酵母菌和线虫的完全基因组。人类基因组，即一个典型的“人”的全部基因，也将提前在2003年完全测定。现在，基因数据的总量每14个月翻一番，每个月至少测出150种蛋白质的结构。增长速度本身也在加速。

国家生物信息中心筹建搁浅三年！

郝柏林

2003年1月1日

徐冠华同志：

这封信只提两件事：

一、**国家生物信息中心筹建工作搁浅三年多，情况太不正常，实在不应再拖下去。**国家生物信息中心首先是艰苦的服务，而不是值得哄抢的肥肉。凡是拿了人民的钱作出的数据，都应当提交国家生物信息中心，按情况、分层次

关于中国“国家生物医学信息中心”的调研报告与建议

2014年

中科院学部咨询项目

《生物信息学的发展与共享平台的建设》项目组

调研项目组院士：陈润生、赵国屏、强伯勤、郝柏林、杨焕明、贺福初、李衍达、康乐、张春霆

国务院办公厅印发《科学数据管理办法》

国务院办公厅印发《科学数据管理办法》（以下简称《办法》）

进一步加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好地为国家科技创新、经济社会发展和国家安全提供支撑

科学数据是国家科技创新发展和经济社会发展的重要基础性战略资源

《办法》明确了我国科学数据管理的

总体原则、主要职责、数据采集汇交与保存、共享利用、保密与安全等方面内容，着重从五个方面提出了具体管理措施



新华社发（朱禹制图）

https://www.gov.cn/xinwen/2018-04/02/content_5279353.htm



首页 > 政策 > 解读

用国家资金获得的科学数据必须上交

2018-04-15 21:37 来源：光明日报

近日，我国首次在国家层面出台《科学数据管理办法》（以下简称“办法”），大力推进科学数据资源的开放共享，特别是国家科技计划项目产生的数据，要求进行强制性汇交，否则项目不予验收。

科学数据为什么要强制上交？科研工作者以后如何上交科学数据？记者采访了科技部相关负责人和有关专家

强制上交科学数据能弥补基础研究的短板，国家财政经费投入产生的科学数据必须上交

https://www.gov.cn/zhengce/2018-04/15/content_5285589.htm

- 1998年：人类遗传资源管理暂行办法
- 2018年：科学数据管理办法
 - 科学数据管理遵循**分级管理、安全可控、充分利用**的原则，明确责任主体，加强能力建设，促进开放共享。
 - 政府预算资金资助的各级科技计划项目所形成的科学数据，应由项目牵头单位**汇交到相关科学数据中心**。
- 2019年：中华人民共和国人类遗传资源管理条例
 - 人类遗传资源：包括人类**遗传资源材料**和人类**遗传资源信息**。
 - 人类遗传资源信息**备份、备案及安全审查机制**。
- 2020年：中华人民共和国生物安全法
 - 国家对我国人类遗传资源和生物资源**享有主权**。
- 2021年：中华人民共和国数据安全法
 - 国家**统筹发展和安全**，坚持以数据开发利用和产业发展促进数据安全，以数据安全保障数据开发利用和产业发展。

1996年7月《科学》杂志称，国外某大学要在中国开展遗传疾病合作，用到两亿样本，由国外制药公司支持的多个项目已在进行中，包括对600万中国人进行哮喘基因筛选。

人类遗传资源管理条例实施细则安全审查的情形包括：
(一) 重要遗传家系的人类遗传资源信息；
(二) 特定地区的人类遗传资源信息；
(三) 人数大于**500**例的外显子组测序、基因组测序信息资源；
(四) 可能影响我国公众健康、国家安全和社会公共利益的其他情形。

国务院办公厅印发《科学数据管理办法》

国务院办公厅印发《科学数据管理办法》（以下简称《办法》），进一步加强和规范科学数据管理，保障科学数据安全，提高开放共享水平，更好地为国家科技创新、经济社会发展 and 国家安全提供支撑。

《办法》明确了我国科学数据管理的总体原则、主要职责、数据采集汇交与保存、共享利用、保密与安全等方面内容，着重从五个方面提出了具体管理措施

- 一、明确各方职责分工，强化法人单位主体责任，明确主管部门职责，体现“谁拥有、谁负责”、“谁开放、谁受益”
- 二、按照“分级分类管理，确保安全可靠”的原则，主管部门和法人单位依法确定科学数据的密级及开放条件，加强科学数据共享和使用的监管
- 三、加强知识产权保护，对科学数据使用者和生产者的行为进行规范，体现对科学数据知识产权的尊重
- 四、要求科技计划项目产生的科学数据进行强制性汇交，并通过科学数据中心进行规范管理和长期保存，加强数据积累和开放共享
- 五、提出法人单位要在岗位设置、绩效收入、职称评定等方面建立激励机制，加强科学数据管理能力建设

新华社发（朱禹制图）

2018年3月17日

**中华人民共和国
人类遗传资源管理条例**

中国法制出版社

2019年7月1日

**中华人民共和国
生物安全法**

法律出版社

2021年4月15日

**中华人民共和国
数据安全法**

法律出版社

2021年9月1日

**中华人民共和国
个人信息保护法**

法律出版社

2021年11月1日

**国家数据局
揭牌**

2023年10月25日

协调推进数据基础制度建设
统筹数据资源整合共享和开发利用
统筹推进数字中国、数字经济、
数字社会规划和建设等

2023年10月25日

坚持总体国家安全观

科学
数据

采集

汇交

存储

审核

共享

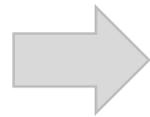
生物数据全链条安全管理和开发利用

面向我国人民生命健康、生物数据与信息安全等国家重大战略需求，建立生物信息大数据**汇交存储、安全管理、开放共享与整合挖掘**研究体系，研发大数据**前沿交叉与转化应用**的新方法和新技术，支撑公益性科学研究和产业创新发展

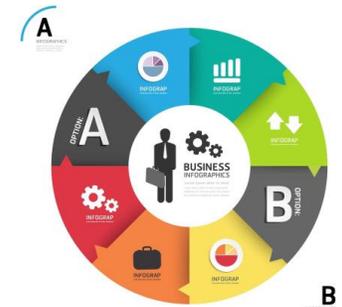
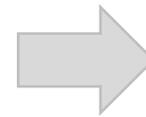
Translating big data into big discoveries



存储与汇交
Deposition



整合与挖掘
Integration



转化与应用
Translation

简称	全称	成立年份	主要数据资源
美国 NCBI	美国国家生物技术信息中心 National Center for Biotechnology Information www.ncbi.nlm.nih.gov	1988	GenBank (1982), SRA, GEO, RefSeq, HomoloGene, dbSNP, dbVar, ClinVar, CDD, Protein Clusters, PubMed/PMC, BLAST, Entrez, OMIM, Taxonomy
欧洲 EMBL- EBI	欧洲生物信息学研究所 European Bioinformatics Institute www.ebi.ac.uk	1994	ENA (1982), EVA, Ensembl, UniProt, InAct, Expression Atlas, Rfam, Pfam, GWAS Catalog, Europe PMC, EBI Search
日本 DDBJ	日本DNA数据库 DNA Data Bank of Japan www.ddbj.nig.ac.jp	1986	DDBJ, DRA, GEA, JGA
中国 CNCB- NGDC	国家生物信息中心 China National Center for Bioinformation National Genomics Data Center ngdc.cncb.ac.cn	2019	GSA, GenBase, GWH, GVM, GEN, MethBank, GWAS Atlas, EWAS Atlas, TWAS Atlas, NONCODE, PGG, LncBook, IC4R, iDog, Database Commons, BIG Search

美国国家生物技术信息中心NCBI

每天实时更新
同步交换数据



全球生物数据
汇交整合共享

The **International Nucleotide Sequence Database Collaboration (INSDC)** consists of a joint effort to collect and disseminate databases containing DNA and RNA sequences. It involves the following computerized databases: **DNA Data Bank of Japan** (Japan), **GenBank** (USA) and the **European Nucleotide Archive** (UK).

<http://www.insdc.org>

Distribution of Materials and Data

One of the terms and conditions of publishing with Cell Press is that authors be willing to distribute any materials and protocols used in the published experiments to qualified researchers for their own use. Materials include but are not limited to cells, DNA, antibodies, reagents, organisms, and mouse strains or, if necessary, the relevant ES cells. These must be made available with minimal restrictions and in a timely manner, but it is acceptable to request reasonable payment to cover the cost of maintenance and transport of materials. If there are restrictions to the availability of any materials, data, or information, these must be disclosed in the cover letter and in the STAR Methods section of the manuscript at the time of submission.

Data sets must be made freely available to readers from the date of publication and must be provided to editors and peer reviewers at submission for the purposes of evaluating the manuscript. In addition, we offer the opportunity for authors to make underlying data not included in the paper itself or deposited in a database available to the scientific community by posting them on [Mendeley Data](#) and then including a link in the published paper. For more detailed instructions, [click here](#).

For the following types of data, submission of the full data set to a community-endorsed, public repository is mandatory. Accession numbers must be provided in the paper (see "Database Linking" below for specific formatting instructions). Examples of appropriate public repositories are listed below.

[DNA and Protein Sequences](#)

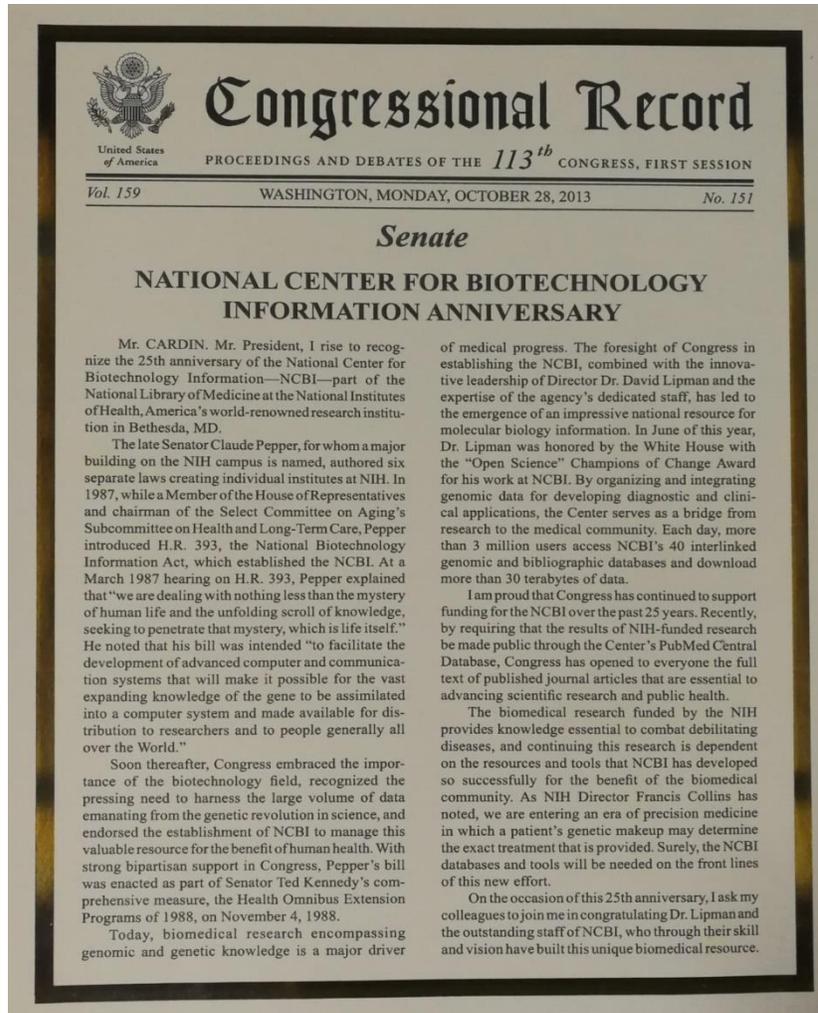
Protein Sequences: [Uniprot](#)

DNA and RNA Sequences: [Genbank/European Nucleotide Archive \(ENA\)/DDBJ, Protein DataBank, UniProt](#)

DNA Sequencing Data (traces and short reads): [NCBI Trace and Short-Read Archive, ENA's Sequence Read Archive](#)

Deep Sequencing Data: Deposit in [GEO](#) or [ArrayExpress](#) upon submission to the journal

The sequences of all RNAi, antisense, and morpholino probes must be included in the paper or deposited in a public database with the accession number provided in the paper.



1988年，美国国会立法成立NCBI



1997年，美国副总统戈尔启动PubMed

NCBI归属于美国国家卫生研究院（National Institutes of Health）的国家医学图书馆（National Library of Medicine），位于美国马里兰州的贝塞斯达，建立于1988年

简称	全称	成立年份	数据介绍
GenBank	基因序列数据库	1982	https://www.ncbi.nlm.nih.gov/genbank/
RefSeq	参考序列数据库	1986	https://www.ncbi.nlm.nih.gov/refseq/
SRA	序列读段归档库	2007	https://www.ncbi.nlm.nih.gov/Traces/sra/
dbSNP	变异数据库	1999	https://www.ncbi.nlm.nih.gov/snp/
dbGap		2006	https://www.ncbi.nlm.nih.gov/gap/
dbVar		2009	https://www.ncbi.nlm.nih.gov/dbvar/
OMIM	在线人类孟德尔遗传数据库	1966	https://www.omim.org
Taxonomy	物种分类数据库	1991	https://www.ncbi.nlm.nih.gov/taxonomy
PubMed	生物医学文献库	1996	https://www.ncbi.nlm.nih.gov/pubmed/
BLAST	在线序列比对分析工具	1990	https://blast.ncbi.nlm.nih.gov/

NIH National Library of Medicine
National Center for Biotechnology Information

GenBank

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Documentation

GenBank Overview

What is GenBank?

GenBank[®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences ([Nucleic Acids Research, 2013 Jan;41\(D1\):D36-42](#)). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

- 国家实验室背景**
 started in 1982 by Walter Goad (nuclear physicist) at Los Alamos National Laboratory (established in 1943).
- 常年持续更新**
 Release 261, June 2024

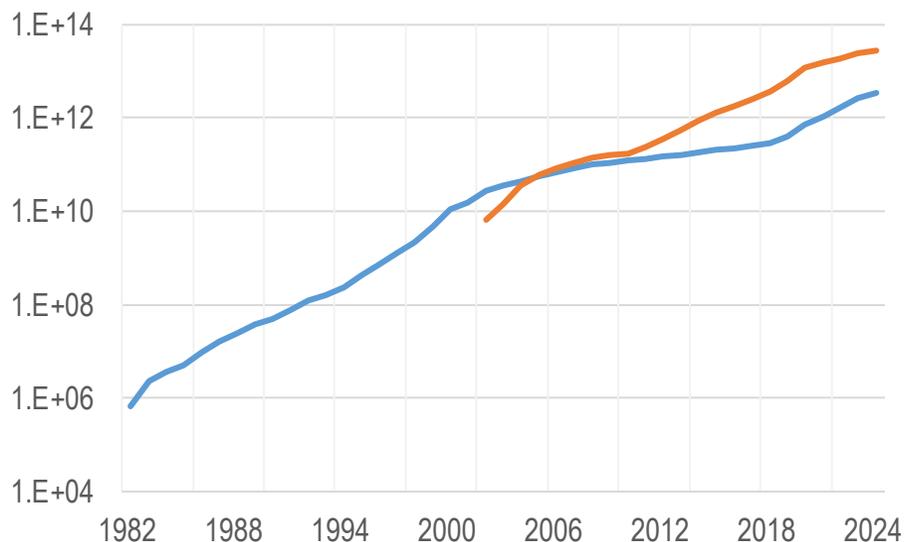
GenBank收集注释所有公开可利用的核苷酸序列和蛋白质序列，是全球最有影响力的生物领域数据库之一

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

GenBank数据量以每18个月翻一番的速度持续指数增长

GenBank: ~3.3万亿bp
WGS: ~27万亿bp

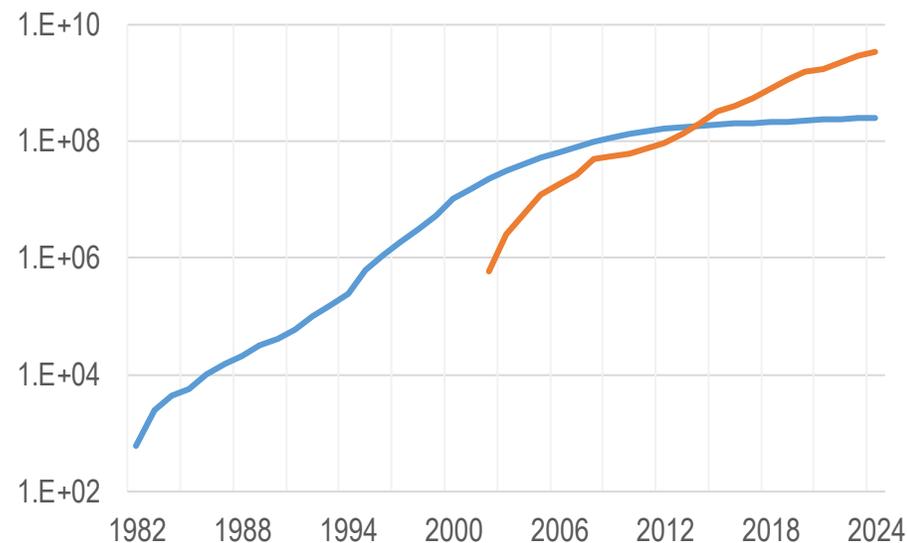
Bases



— GenBank — WGS

GenBank: ~2.16亿条
WGS: ~12.07亿条

Sequences



— GenBank — WGS

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide HBB Search

Create alert Advanced Help

Summary 20 per page Sort by Default order

Filters: Manage Filters

Species

- Animals (7,179)
- Plants (503)
- Fungi (904)
- Protists (427)
- Bacteria (430)
- Archaea (214)
- Viruses (82)
- Customize ...

Molecule types

- genomic
- DNA/RNA (5,539)
- mRNA (4,241)
- Customize ...

Source databases

- INSDC (GenBank) (7,984)
- RefSeq (2,128)
- Customize ...

Sequence Type

Species: Animals (7,179) Plants (503) Fungi (904) Protists (427) Bacteria (430) Archaea (214) Viruses (82) Customize ...

Molecule types: genomic DNA/RNA (5,539) mRNA (4,241) Customize ...

Source databases: INSDC (GenBank) (7,984) RefSeq (2,128) Customize ...

Sequence Type: Nucleotide (10,700)

Results by taxon

Top Organisms [Tree]

- Homo sapiens (1150)
- Chlorocebus sabaeus (982)
- Peromyscus maniculatus (456)
- synthetic construct (372)
- Salmo salar (321)
- Mus musculus (255)
- Pan troglodytes (147)
- Myodes glareolus (139)
- Correliella burgdorferi (134)
- Anser cygnoides (127)
- Anas georgica (114)
- Papio anubis (104)
- uncultured bacterium (89)
- Troglodytes aedon (85)
- Parus minor (85)
- Lophonetta specularioides (81)
- Hepatitis B virus (80)
- Anas flavirostris flavirostris (70)
- Anas flavirostris oxyptera (70)
- Bradypus variegatus (63)
- All other taxa (5201)

Less...

GENE

Was this helpful?

HBB – hemoglobin subunit beta

[Homo sapiens \(human\)](#)

Processed peptides: LVV-hemorphin-7, Spinorphin

Also known as: CD113t-C, ECYT6, beta-globin

GeneID: 3043

[RefSeq transcripts \(1\)](#) [RefSeq proteins \(1\)](#) [RefSeqGene \(2\)](#) [PubMed \(756\)](#)

Orthologs Genome Browser BLAST Download

RefSeq Sequences +

过滤选择

关键字查询

物种过滤

查询结果

相关链接

数据类型

The screenshot shows the RefSeq website interface. At the top, there is the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. Below this is a search bar with 'RefSeq' selected in a dropdown menu and a 'Search' button. The main heading is 'RefSeq: NCBI Reference Sequence Database' with a subtext: 'A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.' The page is divided into several columns of links and information. On the left, there is a 'Using RefSeq' section with links like 'About RefSeq', 'Human Reference Genome', and 'FAQ'. In the middle, there is a 'RefSeq Access' section with links like 'NCBI Datasets' and 'RefSeq FTP'. On the right, there is a 'RefSeq projects' section with links like 'Consensus CDS (CCDS)' and 'RefSeq Functional Elements'. At the bottom left, there is an 'Announcements' section with a recent release for May 10, 2024, listing statistics for proteins, transcripts, and organisms. At the bottom middle, there is a 'Related Links' section with links like 'Gene' and 'Genome Data Viewer'. At the bottom right, there is a 'Feedback & Credits' section with links like 'Publications and Citing RefSeq' and 'Contact RefSeq Help Desk'. Three callout boxes are overlaid on the image: one on the left pointing to 'Human Reference Genome', one on the right pointing to 'RefSeq Functional Elements', and one at the bottom right pointing to 'RefSeqGene'.

人类参考
基因组

编码序列

功能元件

基因集合

数据统
计信息

NCBI Resources How To

Nucleotide Advanced

GenBank

Homo sapiens hemoglobin subunit beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.5

RefSeq编号

[FASTA](#) [Graphics](#)

LOCUS NM_000518 628 bp mRNA linear PRI 21-JAN-2020

DEFINITION Homo sapiens hemoglobin subunit beta (HBB), mRNA.

ACCESSION NM_000518

VERSION NM_000518.5

KEYWORDS RefSeq; MANE Select.

SOURCE Homo sapiens (human)

ORGANISM [Homo sapiens](#)

Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.

REFERENCE 1 (bases 1 to 628)

AUTHORS Qadah T and Jamal MS.

TITLE Computational Analysis of Protein Structure Changes as a Result of Nondeletion Insertion Mutations in Human beta-Globin Gene Suggests Possible Cause of beta-Thalassemia

JOURNAL Biomed Res Int 2019, 9210841 (2019)

PUBMED [31275994](#)

REMARK GeneRIF: an attempt was made to investigate nondeletion mutations in the HBB.

Publication Status: Online-Only

REFERENCE 2 (bases 1 to 628)

AUTHORS Xiong H and Chen S.

TITLE First Description of Hb San Diego (HBB: c.328G>A) in a Chinese

Family with Congenital Erythrocytosis

JOURNAL Hemoglobin 43 (2), 126-128 (2019)

PUBMED [31304856](#)

REMARK GeneRIF: identified a heterozygous beta-globin gene (Hb San Diego or HBB: c.328G>A) in exon 3 as a causative germline mutation in a Chinese family with congenital erythrocytosis.

功能描述

文献信息

```
misc_feature UniProtKB/Swiss-Prot (P68871.2); other site"
483..488
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C; ECYT6"
/experiment="experimental evidence, no additional details
recorded"
/note="(Microbial infection) Cleavage, by N.americanus
apr-2. {ECO:0000269|PubMed:12552433}; propagated from
UniProtKB/Swiss-Prot (P68871.2); other site"
483..485
misc_feature
483..485
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C; ECYT6"
/experiment="experimental evidence, no additional details
recorded"
/note="N-linked (Glc) (glycation) lysine, alternate.
{ECO:0000269|PubMed:7358733}; propagated from
UniProtKB/Swiss-Prot (P68871.2); glycosylation site"
483..485
misc_feature
483..485
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C; ECYT6"
/experiment="experimental evidence, no additional details
recorded"
/note="N6-acetyllysine, alternate.
{ECO:0000269|PubMed:4531009}; propagated from
UniProtKB/Swiss-Prot (P68871.2); acetylation site"
143..365
exon
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C; ECYT6"
/inference="alignment:Splign:2.1.0"
366..628
exon
/gene="HBB"
/gene_synonym="beta-globin; CD113t-C; ECYT6"
/inference="alignment:Splign:2.1.0"
```

基因结构
信息

```
ORIGIN
1 acatttgctt ctgacacaac tgtgttcaact agcaacctca aacagacacc atgggtgcatc
61 tgactcctga ggagaagtct gccgttactg ccctgtgggg caagtggaac gtggatgaag
121 ttgtgtgtga ggcctggggc aggctgctgg tggctctacc ttggaccacag aggttctttg
181 agtcctttgg ggatctgtcc actcctgatg ctgttatggg caaccctaag gtgaaggctc
241 atggcaagaa agtgcctcgg gcttttagtg atggcctggc tcacctggac aacctcaagg
301 gcacctttgc cacactgagt gagctgcaact gtgacaagct gcacgtggat cctgagaact
361 tcaggctcct gggcaacgtg ctggtctgtg tgctggccca tcactttggc aaagaattca
421 ccccaccagt gcaggctgcc tatcagaaaag tgggtggctgg tgtggctaat gccctggccc
481 acaagtatca ctaagctcgc tttcttgctg tccaatttct ataaagggtt cctttgttcc
541 ctaagtccaa ctactaaact ggggatatt atgaagggcc ttgagcatct ggattctgcc
601 taataaaaaa catttttttt cattgcaa
```

序列信息

NCBI Resources How To Sign in to NCBI

SRA SRA Search Help

Advanced



SRA

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Getting Started

[How to Submit](#)

[How to search and download](#)

[How to use SRA in the cloud](#)

[Submit to SRA](#)

Tools and Software

[Download SRA Toolkit](#)

[SRA Toolkit Documentation](#)

[SRA-BLAST](#)

[SRA Run Browser](#)

[SRA Run Selector](#)

Related Resources

[Submission Portal](#)

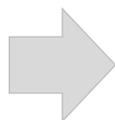
[Trace Archive](#)

[dbGaP Home](#)

[BioProject](#)

[BioSample](#)

高通量测序原始数据提交共享与数据编号获取



学界共识
文章投稿



截止2024年7月
数据总量91PB
1P=1000T, 1T=1000G

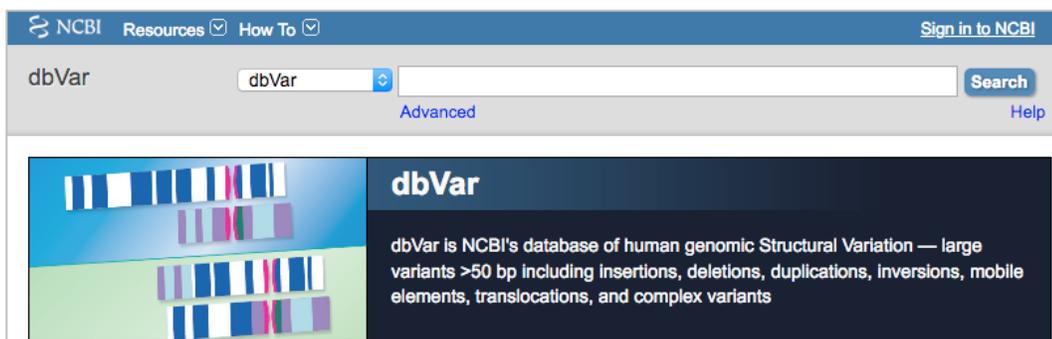


dbSNP

dbSNP contains human single nucleotide variations, microsatellites, and small-scale insertions and deletions along with publication, population frequency, molecular consequence, and genomic and RefSeq mapping information for both common variations and clinical mutations.

dbSNP: 单核苷酸变异、
微卫星、小插入或删除

<https://www.ncbi.nlm.nih.gov/snp/>

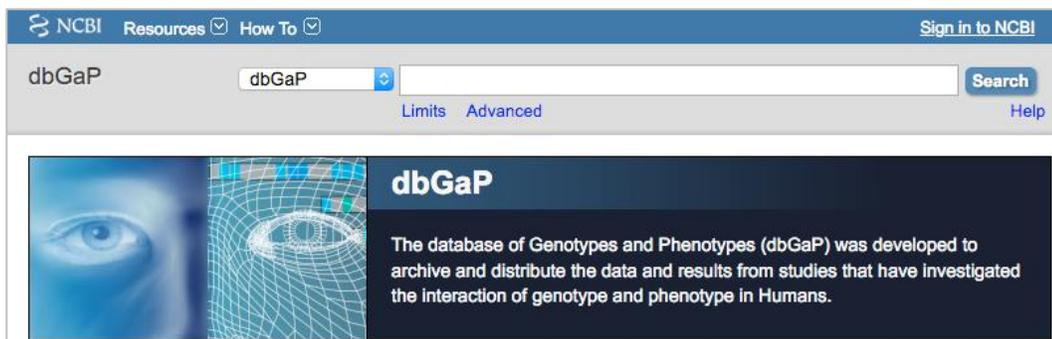


dbVar

dbVar is NCBI's database of human genomic Structural Variation — large variants >50 bp including insertions, deletions, duplications, inversions, mobile elements, translocations, and complex variants

dbVar: 结构变异数据库, 大于
50bp变异, 包括插入、删除、复制、
倒位、易位

<https://www.ncbi.nlm.nih.gov/dbvar/>



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

dbGap: 基因变异与表型关联信息

<https://www.ncbi.nlm.nih.gov/gap/>

rs334

Current Build 153

Released July 9, 2019

Organism *Homo sapiens*

Position chr11:5227002 (GRCh38.p12) [📍](#)
The anchor position for this RefSNP. Includes all nucleotides potentially affected by this change, thus it can differ from HGVS, which is right-shifted. See [here](#) for details.

Alleles T>A / T>C / T>G

Variation Type SNV Single Nucleotide Variation

Frequency A=0.00348 (874/251180, GnomAD_exome)
A=0.00438 (532/121340, ExAC)
A=0.0115 (362/31400, GnomAD) ([+ 1 more](#))

Clinical Significance Reported in [ClinVar](#)

Gene : Consequence HBB : Missense Variant

Publications 101 citations

Genomic View [See rs on genome](#)

镰刀型红细胞
贫血症 (Sickle
Cell Anemia)

Gene: [HBB](#), hemoglobin subunit beta (minus strand)

Molecule type	Change	Amino acid[Codon]	SO Term
HBB transcript	NM_000518.5:c.20A>T	E [GAG] > V [GTG]	Coding Sequence Variant
hemoglobin subunit beta	NP_000509.1:p.Glu7Val	E (Glu) > V (Val)	Missense Variant
HBB transcript	NM_000518.5:c.20A>G	E [GAG] > G [GGG]	Coding Sequence Variant
hemoglobin subunit beta	NP_000509.1:p.Glu7Gly	E (Glu) > G (Gly)	Missense Variant
HBB transcript	NM_000518.5:c.20A>C	E [GAG] > A [GCG]	Coding Sequence Variant
hemoglobin subunit beta	NP_000509.1:p.Glu7Ala	E (Glu) > A (Ala)	Missense Variant

编号规则:
Submitted SNP (ss)
Reference SNP (rs)

<https://www.ncbi.nlm.nih.gov/snp/rs334>

人类基因和遗传疾病数据库，是医学遗传学最权威的百科全书，被誉为医学遗传学界的“《圣经》”，包括所有已知的遗传病、遗传决定的性状及其基因，以及各种疾病的临床特征、诊断、鉴别诊断、治疗与预防等

27338条记录数

MIM Number Prefix	Autosomal	X Linked	Y Linked	Mitochondrial	Totals
Gene description *	16,389	770	51	37	17,247
Gene and phenotype, combined +	21	0	0	0	21
Phenotype description, molecular basis known #	6,405	387	5	34	6,831
Phenotype description or locus, molecular basis unknown %	1,388	109	4	0	1,501
Other, mainly phenotypes with suspected mendelian basis	1,635	100	3	0	1,738
Totals	25,838	1,366	63	71	27,338

1966年美国John Hopkins大学医学院Victor A McKusiek教授创建MIM，1985年与NLM联合开发OMIM，1987年上线，1995年至今由NCBI负责运营



#190685 DOWN SYNDROME
唐氏综合征

Alternative titles; symbols
TRISOMY 21

Other entities represented in this entry:
DOWN SYNDROME CHROMOSOME REGION 21, INCLUDED; DCR, INCLUDED
DOWN SYNDROME CRITICAL REGION, INCLUDED; DSCR,
TRANSIENT MYELOPROLIFERATIVE DISORDER OF DOWN SYNDROME, INCLUDED
LEUKEMIA, MEGAKARYOBLASTIC, OF DOWN SYNDROME

Cytogenetic location: 21q22.3 Genomic coordinates (GRCh38): 21:41,200,000-46,709,983

<https://www.omim.org/entry/190685>

<https://www.omim.org/entry/603903>

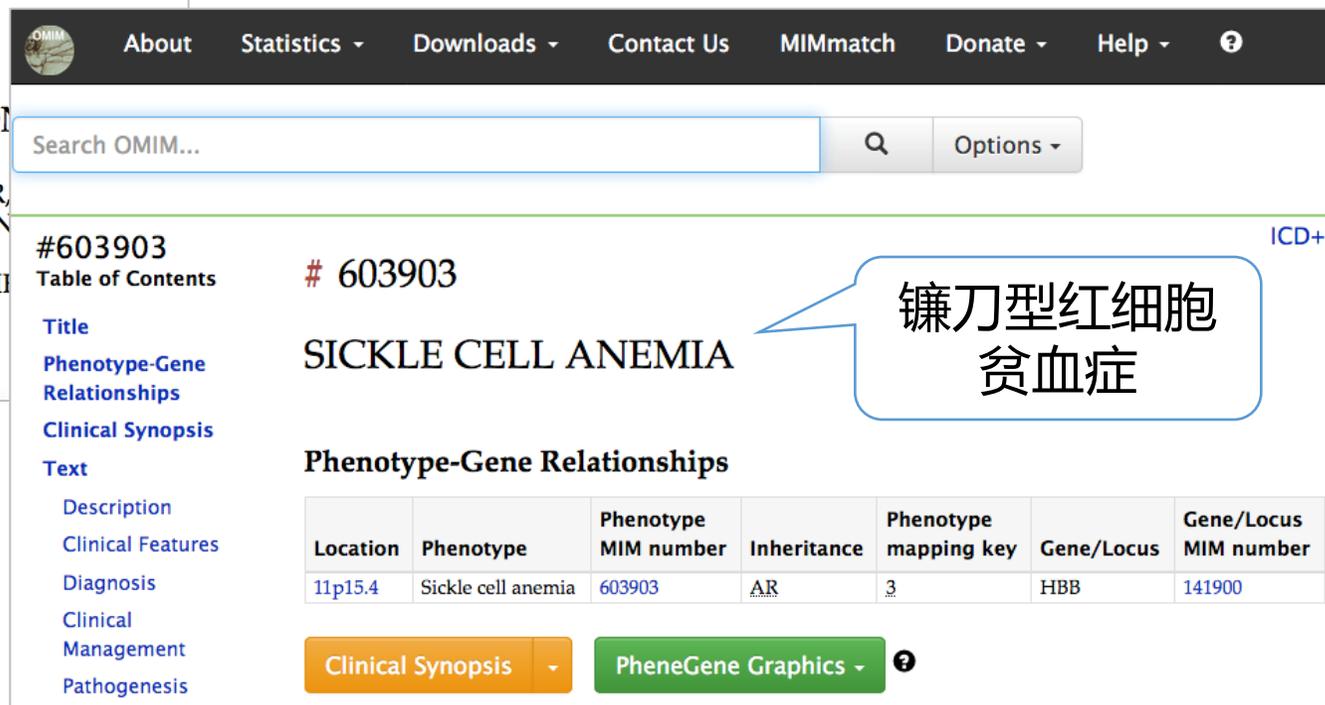
记录编号规则:

1开头, 染色体显性遗传

2开头, 染色体隐性遗传

3开头, X连锁; 4开头, Y连锁

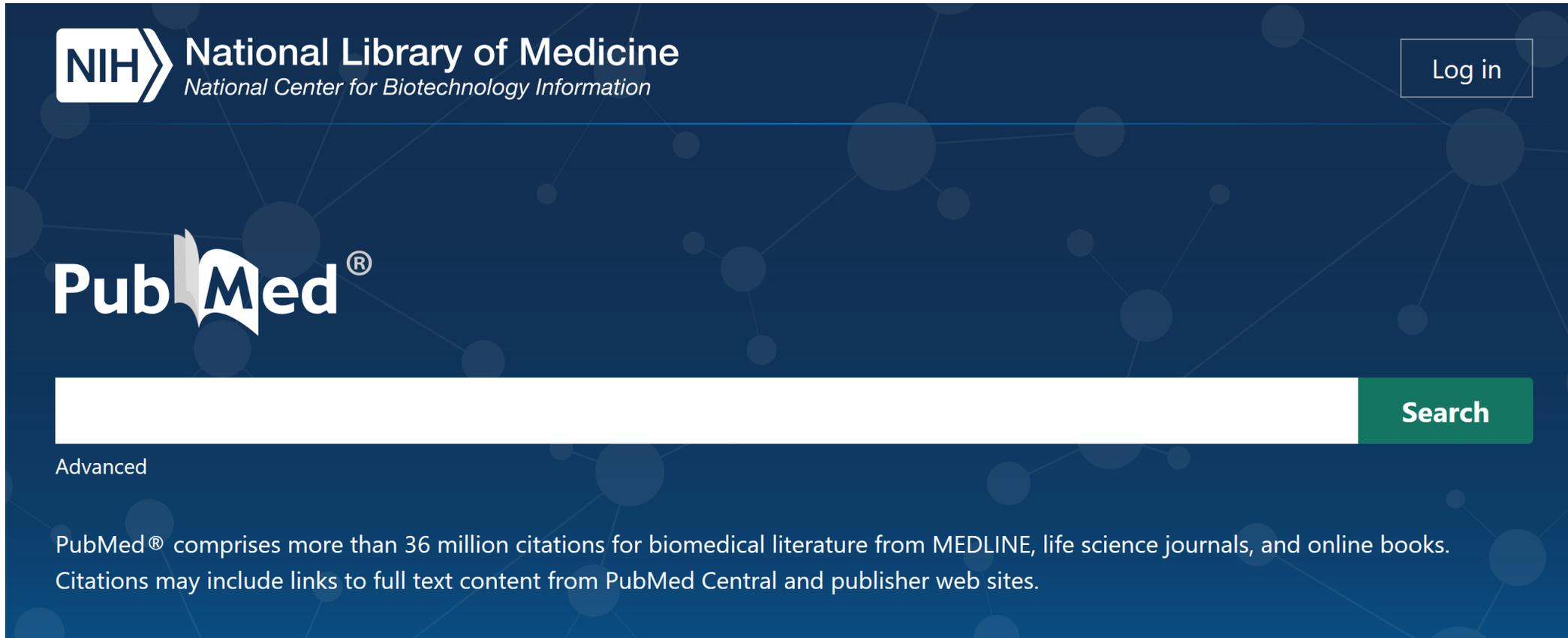
5开头, 线粒体; 6开头, 染色体



#603903 # 603903 ICD+

SICKLE CELL ANEMIA
镰刀型红细胞贫血症

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key	Gene/Locus	Gene/Locus MIM number
11p15.4	Sickle cell anemia	603903	AR	3	HBB	141900



NIH National Library of Medicine
National Center for Biotechnology Information

Log in

PubMed®

Search

Advanced

PubMed® comprises more than 36 million citations for biomedical literature from MEDLINE, life science journals, and online books. Citations may include links to full text content from PubMed Central and publisher web sites.

- PubMed由美国国家医学图书馆1996年建立，提供生物医学论文以及摘要，收录文献超过3600万篇
- 2000年建立PubMed Central (PMC)，提供2874个期刊的约980万篇论文全文（截止2024年）
- NIH Public Access Policy（2008年起执行），即发表后12个月内必须全文公开

The screenshot shows the PubMed search results for the query "precision medicine". At the top, the NIH National Library of Medicine logo is visible, along with a "Log in" button. The search bar contains "precision medicine" and a "Search" button. Below the search bar, there are links for "Advanced", "Create alert", "Create RSS", and "User Guide". The results are sorted by "Best match" and there are buttons for "Save", "Email", and "Send to". The search results are displayed in a list format, with the first result being "Precision medicine and therapies of the future." by Sisodiya SM. The second result is "Precision Medicine from a Public Health Perspective." by Ramaswami R, Bayer R, and Galea S. On the left side, there are filters for "MY NCBI FILTERS", "RESULTS BY YEAR" (with a bar chart showing an increase in results over time from 1952 to 2024), "TEXT AVAILABILITY" (with checkboxes for Abstract, Free full text, and Full text), and "ARTICLE ATTRIBUTE" (with a checkbox for Associated data).

趋势图

搜索结果

文章信息及其PMID

全文摘要

论文属性

NCBI Resources How To Sign in to NCBI

PubMed.gov PubMed Search Help

US National Library of Medicine National Institutes of Health Advanced

Click here to try the **New PubMed!**

An updated version of PubMed is now available. Come see the new improvements to the interface!

Format: Abstract

Send to

Nature. 2015 Oct 15;526(7573):336-42. doi: 10.1038/nature15816.

Building the foundation for genomics in precision medicine.

Aronson SJ^{1,2}, Rehm HL^{1,3,4,5}.

Author information

Abstract

Precision medicine has the potential to profoundly improve the practice of medicine. However, the advances required will take time to implement. Genetics is already being used to direct clinical decision-making and its contribution is likely to increase. To accelerate these advances, fundamental changes are needed in the infrastructure and mechanisms for data collection, storage and sharing. This will create a continuously learning health-care system with seamless cycling between clinical care and research. Patients must be educated about the benefits of sharing data. The building blocks for such a system are already forming and they will accelerate the adoption of precision medicine.

PMID: 26469044 PMCID: PMC5669797 DOI: 10.1038/nature15816

[Indexed for MEDLINE] Free PMC Article

Full text links

nature PMC Full text **FREE**

Save items

Add to Favorites

Similar articles

Review Implementing Genomic Clinical Dec [CPT Pharmacometrics Syst Pharm...]

Health: Make precision medicine work for cancer care. [Nature. 2015]

The CLIPMERGE PGx Program: clinical implementatio [Clin Pharmacol Ther. 2013]

论文信息

全文链接

PMID

相关论文

<https://www.ncbi.nlm.nih.gov/pubmed/26469044>

The screenshot shows the BLAST website interface. At the top, there is the NIH logo and the text 'National Library of Medicine National Center for Biotechnology Information'. A 'Log in' button is in the top right. Below the header, there are navigation links: 'Home', 'Recent Results', 'Saved Strategies', and 'Help'. The main content area is titled 'Basic Local Alignment Search Tool'. A text box highlights the title and a paragraph: 'BLAST finds relevant sequences in large databases and... A new approach to rapid **sequence** comparison, **basic local alignment search tool** (BLAST), directly approximates **alignments** that optimize a measure of **local** similarity, the maximal ...'. Below this, there are links for 'Save', 'Cite', 'Cited by 112500', 'Related articles', and 'All 62 versions'. Under the heading 'Web BLAST', there are three tool options: 'Nucleotide BLAST' (nucleotide to nucleotide), 'blastx' (translated nucleotide to protein), and 'tblastn' (protein to translated nucleotide). To the right of 'tblastn' is 'Protein BLAST' (protein to protein).

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® >> blastn suite Home Recent Results Saved Strategies Help

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

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

Enter Query Sequence

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

Or, upload file No file selected.

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

Align two or more sequences

BLAST results will be displayed in a new format by default
You can always switch back to the Traditional Results page.

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

输入序列或
序列ID号

设置序列片段

任务名称

比对库

The screenshot shows the NCBI GenBank entry for the Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome (NC_045512.2). The interface includes a search bar at the top, navigation links for Resources and How To, and a 'Sign in to NCBI' button. The main content area displays the sequence title, accession number, and various fields such as LOCUS, DEFINITION, ACCESSION, VERSION, DBLINK, KEYWORDS, SOURCE, ORGANISM, REFERENCE, AUTHORS, TITLE, and JOURNAL. On the right side, there are options to 'Change region shown', 'Customize view', and 'Analyze this sequence', which includes 'Run BLAST', 'Pick Primers', 'Highlight Sequence Features', and 'Find in this Sequence'. Below these are sections for 'SARS Coronavirus Resource' and 'Related information'.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Advanced Help

GenBank Send to Change region shown

Wuhan seafood market pneumonia virus isolate Wuhan-Hu-1, complete genome

NCBI Reference Sequence: NC_045512.2

[FASTA](#) [Graphics](#)

LOCUS NC_045512 29903 bp ss-RNA linear VRL 28-JAN-2020

DEFINITION Wuhan seafood market pneumonia virus 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; Nidovirales; Coronidovirineae; Coronaviridae; Orthocoronavirinae; Betacoronavirus; Sarbecovirus.

REFERENCE 1 (bases 1 to 29903)

AUTHORS Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., Song, Z.-G., 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 novel coronavirus associated with a respiratory disease in Wuhan of Hubei province, China

JOURNAL Unpublished

Analyze this sequence

Run BLAST

Pick Primers

Highlight Sequence Features

Find in this Sequence

SARS Coronavirus Resource

Retrieve, view, and download SARS coronavirus genomic and protein sequences.

Related information

Assembly

BioProject

RefSeq编号

基因组长度
病毒类型

BLAST分析

https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2/

i Your results are filtered to match records that exclude: Severe acute respiratory syndrome coronavirus 2

Job Title	NC_045512:Wuhan seafood market pneumonia virus...
RID	5264APE101R Search expires on 02-23 19:03 pm Download All
Program	BLASTN Citation
Database	nt See details
Query ID	NC_045512.2
Description	Wuhan seafood market pneumonia virus isolate Wuhan-Hu ...
Molecule type	nucleic acid
Query Length	29903
Other reports	Distance tree of results MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Severe acute respiratory syndrome coronavirus 2

[+ Add organism](#)

Percent Identity: to E value: to Query Coverage: to

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

select all 60 sequences selected [GenBank](#) [Graphics](#) [Distance tree of results](#)

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> Bat coronavirus isolate RaTG13, complete genome	48724	48724	99%	0.0	96.12%	MN996532.1
<input checked="" type="checkbox"/> Bat SARS-like coronavirus isolate bat-SL-CoVZC45, complete genome	26943	35336	95%	0.0	89.12%	MG772933.1
<input checked="" type="checkbox"/> Bat SARS-like coronavirus isolate bat-SL-CoVZXC21, complete genome	22223	35276	94%	0.0	88.65%	MG772934.1
<input checked="" type="checkbox"/> SARS coronavirus ZS-C, complete genome	15213	22564	88%	0.0	82.34%	AY395003.1
<input checked="" type="checkbox"/> SARS coronavirus ZS-B, complete genome	15213	22600	88%	0.0	82.34%	AY394996.1

任务ID

序列类型
和长度

比对结果

过滤条件

比对参数

NCBI Resources How To

NCBI National Center for Biotechnology Information

All Databases Search

All Resources

All Databases Downloads Submissions Tools How To

Databases

Assembly
A database providing information on the structure of assembled genomes, assembly names and other meta-data, statistical reports, and links to genomic sequence data.

BioCollections
A curated set of metadata for culture collections, museums, herbaria and other natural history collections. The records display collection codes, information about the collections' home institutions, and links to relevant data at NCBI.

BioProject (formerly Genome Project)
A collection of genomics, functional genomics, and genetics studies and links to their resulting datasets. This resource describes project scope, material, and objectives and provides a mechanism to retrieve datasets that are often difficult to find due to inconsistent annotation, multiple independent submissions, and the varied nature of diverse data types which are often stored in different databases.

BioSample
The BioSample database contains descriptions of biological source materials used in experimental assays.

BioSystems

<https://www.ncbi.nlm.nih.gov/guide/all/>

隶属欧洲分子生物学实验室（EMBL，1974年成立），是一个政府间国际组织学术机构，致力于提供免费生物信息资源、促进基础研究、提供培训和传播行业尖端技术，1994年成立

简称	全称	成立年份	数据介绍
ENA	European Nucleotide Archive	1982	http://www.ebi.ac.uk/ena
Ensembl	基因组数据库	1999	http://www.ensembl.org
UniProt	蛋白质数据库	2002	https://www.uniprot.org
GWAS Catalog	基因型-表型关联知识库	2008	https://www.ebi.ac.uk/gwas
Pfam	蛋白质家族数据库	1995	http://pfam.xfam.org
Rfam	RNA家族数据库	2003	http://rfam.xfam.org
TreeFam	动物基因系统发育树数据库	2006	http://www.treefam.org
IntAct	分子相互作用数据库	2004	https://www.ebi.ac.uk/intact
Reactome	信号通路数据库	2003	https://reactome.org
Expression Atlas	基因表达数据库	2010	http://www.ebi.ac.uk/gxa
Europe PMC	科研文献资料库	2011	http://europepmc.org

 [BLAST/BLAT](#) | [VEP](#) | [Tools](#) | [BioMart](#) | [Downloads](#) | [Help & Docs](#) | [Blog](#) [Login/Register](#)

Tools
[All tools](#)

BioMart >
Export custom datasets from Ensembl with this data-mining tool

BLAST/BLAT >
Search our genomes for your DNA or protein sequence

Variant Effect Predictor >
Analyse your own variants and predict the functional consequences of known and unknown variants

基因信息提取工具

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotate genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

Ensembl Release 99 (January 2020)

- Update to GENCODE 33 for human
- Update to dbSNP153 for human
- Import of updated VISTA enhancers for human and mouse
- New genomes: 10 mammals (including 2 dog breeds), 11 birds, 15 fish and 4 reptiles

Search

All species for

变异结果分析预测

About Us

- [About us](#)
- [Contact us](#)
- [Citing Ensembl](#)
- [Privacy policy](#)
- [Disclaimer](#)

Get help

- [Using this website](#)
- [Adding custom tracks](#)
- [Downloading data](#)
- [Video tutorials](#)
- [Variant Effect Predictor \(VEP\)](#)

子库

Our sister sites

- [Ensembl Bacteria](#)
- [Ensembl Fungi](#)
- [Ensembl Plants](#)
- [Ensembl Protists](#)
- [Ensembl Metazoa](#)

Follow us

-  [Blog](#)
-  [Twitter](#)
-  [Facebook](#)

Show 10 entries		Show/hide columns										
★	Species	DNA (FASTA)	cDNA (FASTA)	CDS (FASTA)	ncRNA (FASTA)	Protein sequence (FASTA)	Annotated sequence (EMBL)	Annotated sequence (GenBank)	Gene sets	Other annotations	Whole databases	Variation (GVF)
Y	Human <i>Homo sapiens</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
Y	Mouse <i>Mus musculus</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
Y	Zebrafish <i>Danio rerio</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
	Abingdon island giant tortoise <i>Chelonoidis abingdonii</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
	African ostrich <i>Struthio camelus australis</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
	Agassiz's desert tortoise <i>Gopherus agassizii</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
	Algerian mouse <i>Mus spretus</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				
	Alpaca <i>Vicugna pacos</i>	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF				

<http://www.ensembl.org/info/data/ftp/index.html>



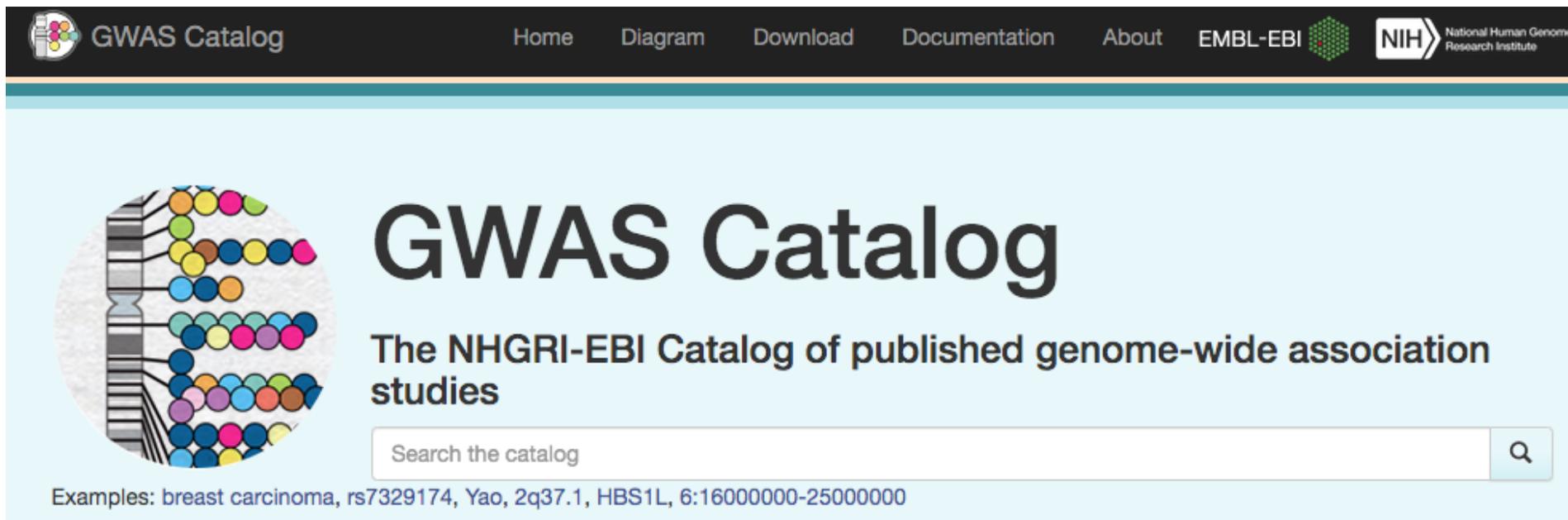
The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

经过人工
质控审核

自动化注
释未审核

UniProt联盟 (2002年)

- 英国EBI
- 瑞士生物信息研究所 SIB (1986年)
- 美国蛋白质信息资源库PIR (1965年)



GWAS Catalog

Home Diagram Download Documentation About EMBL-EBI NIH National Human Genome Research Institute

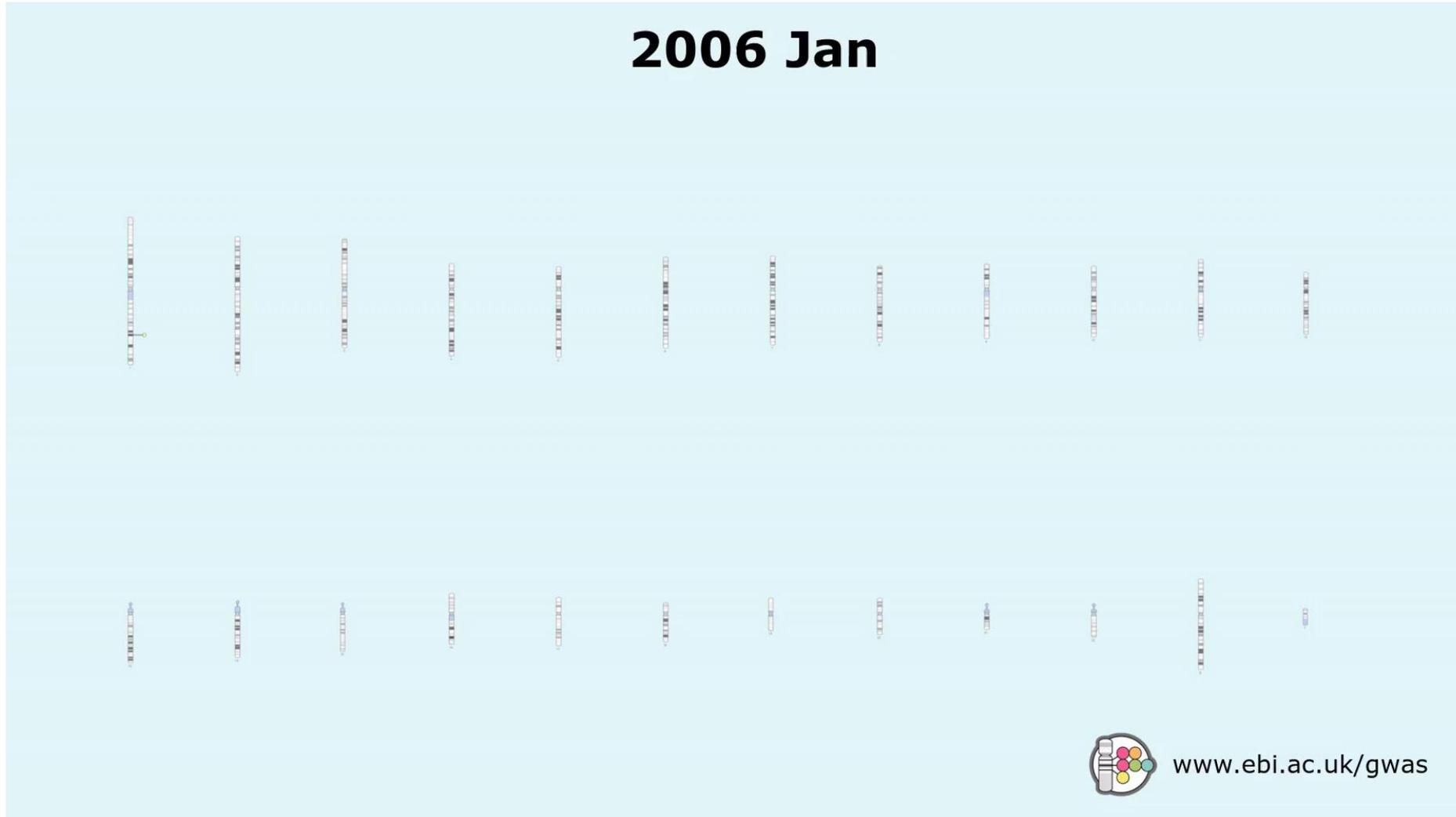
GWAS Catalog

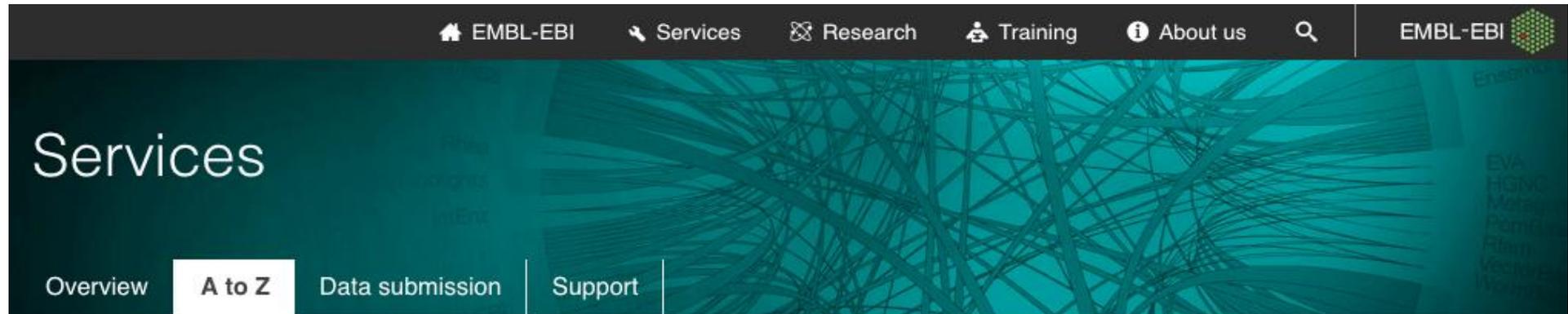
The NHGRI-EBI Catalog of published genome-wide association studies

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

- 基于文献资料采用人工质控和审编的方法，提供高质量的基因型与表型关联信息
- 2008年由美国NIH下属的国家人类基因组研究所（National Human Genome Research Institute）创建，2010年与EBI合作开发
- 数据统计：>5000篇文献，>17万条关联信息

<https://www.ebi.ac.uk/gwas>





Tools & Data Resources

Tools

Annotation Platform

Consolidating text-mined and curated annotations

Assembly converter



Map your data to the current assembly. Based on the CrossMap tool

Read mapping

Data resources

ArrayExpress



A database of functional genomics experiments, including microarray and RNAseq expression data typically related to publications.

BioModels



A repository of peer-reviewed, published, computational models.

Browse by type

DNA & RNA	Gene Expression	Proteins
Structures	Systems	Chemical biology
Ontologies	Literature	Cross domain

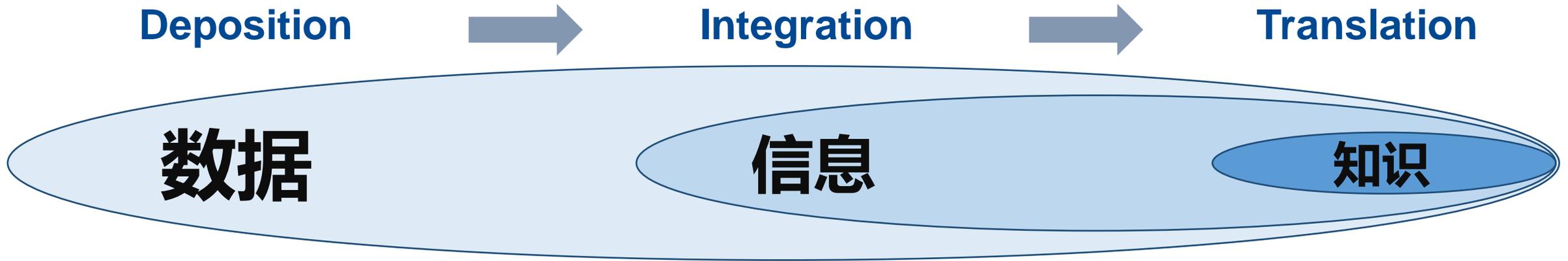
Programmatic access

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

中国国家生物信息中心CNCB-NGDC

CNCB于2019年批复加挂在中国科学院北京基因组研究所，NGDC隶属于CNCB，负责生物信息大数据统一汇交、集中存储、安全管理与开放共享，以及前沿交叉研究与转化应用等

简称	全称	年份	数据介绍
GSA	Genome Sequence Archive	2015	https://ngdc.cncb.ac.cn/gsa
GWH	Genome Warehouse	2016	https://ngdc.cncb.ac.cn/gwh
GVM	Genome Variation Map	2016	https://ngdc.cncb.ac.cn/gvm
MethBank	Methylation DataBank	2014	https://ngdc.cncb.ac.cn/methbank
GWAS Atlas	全基因组关联分析知识库	2019	https://ngdc.cncb.ac.cn/gwas
EWAS Atlas	全表观组关联分析知识库	2018	https://ngdc.cncb.ac.cn/ewas
LncBook	人类长非编码RNA知识库	2018	https://ngdc.cncb.ac.cn/lncbook
IC4R	水稻多组学数据库	2015	http://ngdc.cncb.ac.cn/ic4r
iDog	家狗多组学数据库	2018	https://ngdc.cncb.ac.cn/idog
Database Commons	全球生物医学数据库目录	2015	https://ngdc.cncb.ac.cn/databasecommons
BIG Search	跨库搜索引擎	2017	https://ngdc.cncb.ac.cn/search



➤ 数据库

- 生物项目数据库 (BioProject)
- 生物样本数据库 (BioSample)
- 生信算法工具代码库 (BioCode)
- 原始组学数据归档库 (GSA)
- 多元数据归档库 (OMIX)
- 基因序列数据库 (GenBase)
- 全球生物数据库目录 (Database Commons)
- 植物影像归档库 (OPIA)

➤ 信息库

- 基因组序列库 (Genome Warehouse)
- 基因组变异库 (Genome Variation Map)
- 基因表达库 (Gene Expression Nebulas)
- 甲基化信息库 (Methylation Bank)
- 人类长非编码RNA信息库 (LncBook)
- 癌症单细胞表达图谱库 (CancerSCEM)
- 2019新冠病毒信息库 (RCoV19)*
- 特色物种基因库 (水稻、大豆、家鸡 ChickenGTEx、家狗、绵羊等)

➤ 知识库

- 全基因组关联分析知识库 (GWAS Atlas)
- 全表观组关联分析知识库 (EWAS Atlas)
- 脑疾病知识库 (BrainBase)
- 细胞分类库 (Cell Taxonomy)
- 癌症可变剪切知识库 (ASCancer Atlas)
- 生命科学文献库 (OpenLB)

原始组学数据归档库GSA

2015年首次获国际期刊认可

实现“零”的突破

国家生物信息中心
Data Resources

GSA
Genome Sequence Archive

Home Submit Browse Search Statistics Support

Genome Sequence Archive

The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw sequence data, which is the first repository of the genome sequence data with international journal recognition in China.

Submit: Submit data to GSA
Download: Download data to your computer
Browse: Browse publicly available records
Document: Find help information and documents

Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution

Shaoping Ling¹, Zheng Hu^{2,3}, Zuyu Yang^{4,5}, Fang Yang^{6,7}, Yawei Li⁸, Pei Lin⁹, Ke Chen⁹, Lili Dong⁹, Lihua Cao⁹, Yong Tao⁹, Lingtong Hao⁹, Qingjian Chen⁹, Qiang Gong⁹, Dafei Wu⁹, Wenjie Li⁹, Wenming Zhao⁹, Xiuyun Tian⁹, Chunyi Hao⁹, Eric A. Hungate¹⁰, Daniel V. T. Catenacci¹¹, Richard R. Hudson¹², Wen-Hsiung Li¹³, Xuemei Lu¹⁴, and Chung-IWu¹⁵

Significance
A tumor comprising many cells can be compared to a natural population with many individuals. The amount of genetic diversity reflects how it has evolved and can influence its future evolution. We evaluated a single tumor by sequencing or genotyping nearly 300 regions from the tumor. When the data were analyzed by modern population genetic theory, we estimated more than 100 million coding region mutations in this unexceptional tumor. The extreme genetic diversity implies evolution under the non-Darwinian mode. In contrast, under the prevailing view of Darwinian selection, the genetic diversity would be orders of magnitude lower, because genetic diversity accrues rapidly, a high probability of drug resistance should be heeded, even in the treatment of microscopic tumors.

Author contributions: X.L. and C.-I.W. designed research; Z.Y., F.Y., K.C., D.W., W.L., and W.Z. performed experiments; S.Z., Z.H., F.Y., Y.L., P.L., L.Z., L.C., Y.L., L.H., G.C., and D.G. analyzed data; S.L., Z.H., Y.L., P.L., L.A.X., D.V.T.C., R.R.H., and C.I.W. contributed to the theory; X.L. and C.I. provided final samples; S.L., L.Z., and L.C. contributed new analytic tools; and S.L., Z.H., W.L., X.L., and C.I.W. wrote the paper.

Reviewer's T.G., King Abdullah University of Science and Technology; and J.Z., University of Michigan.

Competing financial interests: The authors declare no conflict of interest.

Data deposition: The sequence data reported in this paper have been deposited in the genome sequence archive of Beijing Institute of Genomics, Chinese Academy of Sciences, accession no. PRJCA000091.

*S.L., Z.H., Z.Y., and F.Y. contributed equally to this work.

To whom correspondence may be addressed: Email: c.i.wu@uic.edu, whli@genomics.cn, ed.zhu@bgi.ac.cn.

This article comes with supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1519561112/-DCSupplemental.

对应

中国CNCB的GSA库
遵循国际组学数据规范

美国NCBI的SRA库
英国EBI的ENA/SRA库
日本DDBJ的DRA库

Data deposition:
The sequence data reported in this paper have been deposited in the **Genome Sequence Archive (GSA)** of Beijing Institute of Genomics, Chinese Academy of Sciences, <http://gsa.big.ac.cn> (Accession no. **PRJCA000091**).

国家生物信息中心
Data Resources Computing Analysis Data Network Standards

GSA for Human
Genome Sequence Archive

Home Submit Browse Search DAC Statistics Documentation Policy Login Register

Available Unavailable

数据受控访问 & 数据管理委员会 (DAC)

Filter:

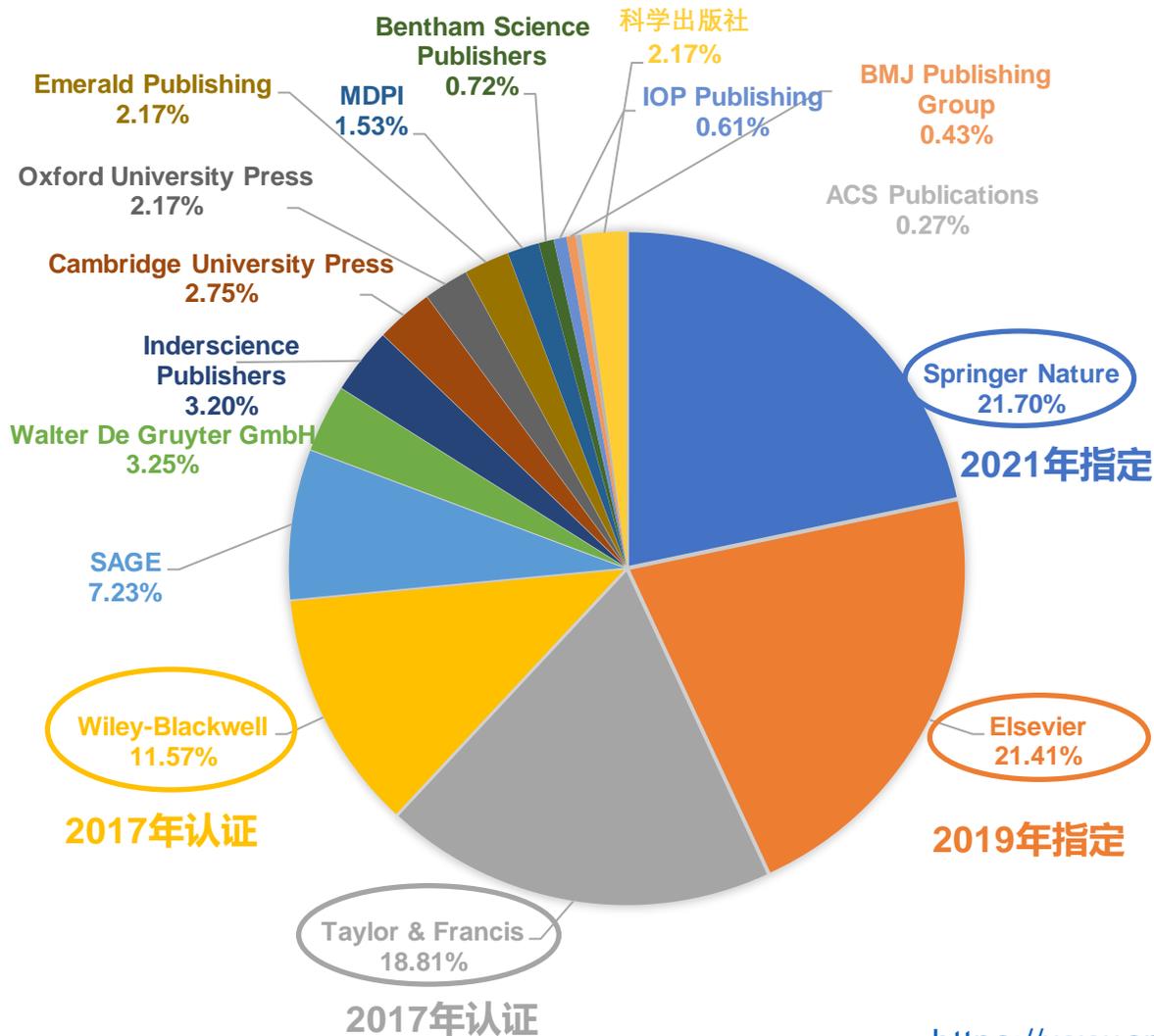
Study	Title	Organization	DAC	Access	Requests	Effective Requests	Approved	Sharing Rank	Last Processed	Request
HRA000150	Single-cell immunological landscape of peripheral blood mononuclear cells of patients with COVID-19 disease	National Clinical Research Center for Infectious Diseases	HDAC000089	Controlled	205	119	104	★★★★★	2024-06-16	Request
HRA001748	scRNA-seq of liver cancer	Peking University Health Science Center	HDAC001033	Controlled	155	104	99	★★★★★	2024-07-03	Request
HRA000051	scRNA-seq of gastric cancer	Institute of Military Cognition and Brain Sciences	HDAC000025	Controlled	150	86	63	★★★★☆	2024-07-04	Request

<https://ngdc.cncb.ac.cn/gsa-human/browse/>

**5374
Studies**

**551386
Individuals**

**791053
Samples**



SPRINGER NATURE

2021年8月



▼ Nucleic acid sequence & Omics

Nucleic acid sequence data and metadata should follow the Genome Standards Consortium (GSC) guidance, which can be browsed at [FAIRsharing GSC collection](#).

Data types

- DNA sequence data*
- RNA sequence data*
- Genome assembly data*
- Genetic variation data

Repositories

- Any INSDC member repository
- Genome Sequence Archive (GSA)**
- dbSNP (human variations less than 50bp)
- dbVar (human variations greater than 50bp)
- European Variation Archive (EVA) (all species)
- Genome Sequence Archive for Human (human variation)**

* Novel DNA sequence, novel RNA sequence, and novel genome assembly data must be deposited to repositories that are part of the International Nucleotide Sequence Collaboration (INSDC), or those which are working towards INSDC inclusion (included in the table), unless there are privacy or ethics restrictions that prevent open sharing of such data. Novel DNA sequence, novel RNA sequence, and novel genome assembly data may in addition be deposited to any other repository (including regional or national repositories) as required.

人类遗传资源信息备份、备案、发布流程

国家互联网应急中心



科研人员

备份数据

备份编号

获得备案号

2022年7月18日

优化

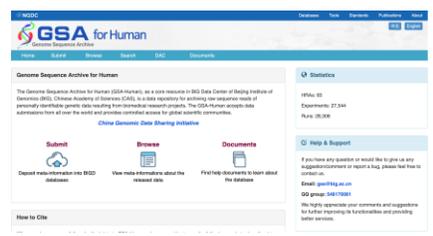
提交数据
备份编号与发布编号



科研人员

备案信息注册
获得备案号

提交数据
获得数据发布编号



CNCB-NGDC人遗数据管理平台

CNCB-NGDC-备份、发布



备份号

备案号

科技部信息中心 - 备案审查



中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

全站 请输入关键字 搜索

首页 组织机构 信息公开 科技政策 政务服务 党建工作 公众参与 专题专栏

当前位置: 科技部门户 > 通知公告

www.most.gov.cn

关于人类遗传资源信息备份平台迁移及更名的公告

日期: 2022年07月11日 11:09 来源: 科技部 【字号: 大 中 小】

为优化人类遗传资源信息备份及数据共享发布流程, 人类遗传资源信息备份平台 (<https://202.108.211.75>) 将于2022年7月17日18:00至2022年7月18日07:00进行迁移升级, 升级期间暂停人类遗传资源信息备份事宜。2022年7月18日07:00恢复服务, 并更改访问地址为<https://hgrip.cncb.ac.cn>或<https://ngdc.cncb.ac.cn/hgrip>, 技术咨询电话变更为010-84097816/010-84097340。

升级后, 原“人类遗传资源信息备份平台”将更名为“人类遗传资源信息管理备份平台”, 新平台将整合人类遗传资源信息管理、备份、发布与共享等功能, 实现人类遗传资源信息一体化服务。

特此公告。

中国人类遗传资源管理办公室

2022年7月11日



温馨提示
本系统自2022年7月18日7点开始服务!

人类遗传资源信息管理备份平台

中华人民共和国科学技术部

用户登录

请输入手机号

请输入密码

验证码 **8XU4**

注册 忘记密码

登录

中国科学院北京基因组研究所 (国家生物信息中心)
服务热线: 010-84097816 / 010-84097340 服务邮箱: hgrip@big.ac.cn 业务咨询热线: 010-88225151

实现人类遗传资源信息管理、备份、发布与共享一体化

<https://ngdc.cncb.ac.cn/hgrip/>



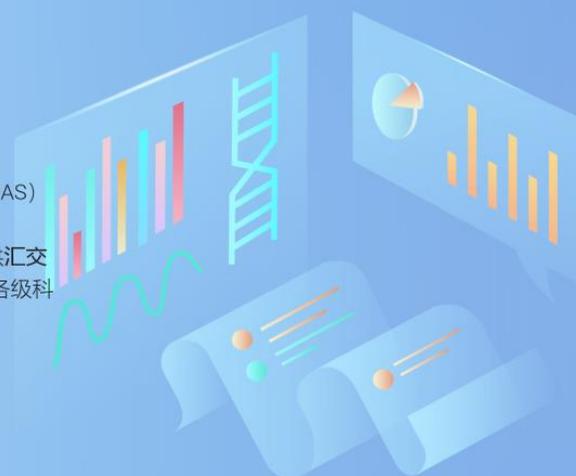
🏠 主页 ⬆️ 提交 ❓ 帮助 ➡️ 注册 👤 登录

科学项目数据汇交服务系统

科学项目数据汇交服务系统 (Scientific Data Archive System, SDAS) 作为国家基因组科学数据中心 (National Genomics Data Center, NGDC) 的汇交计划服务系统。为生物领域国家重点专项项目提供汇交计划提交、实时数据归档情况查询、汇交证明出具等服务, 帮助各级科技计划项目顺利验收。

汇交计划提交入口

科学数据提交入口



33

服务重点专项

370

审核汇交计划

174

出具汇交证明

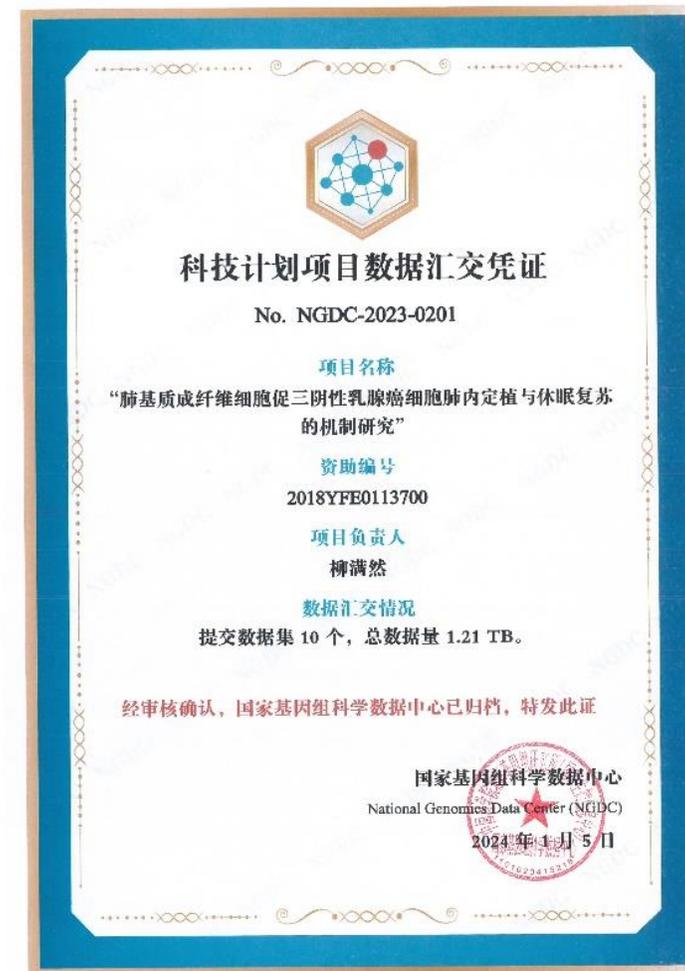
5290

归档数据集

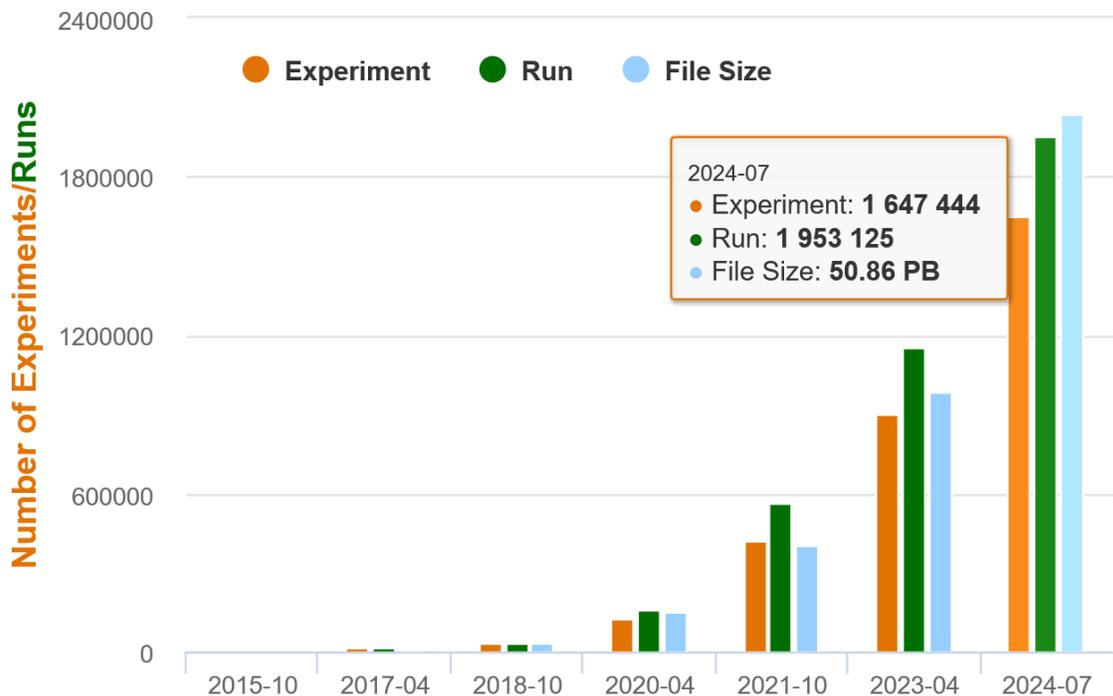
7.32

总汇交数据量 (PB)

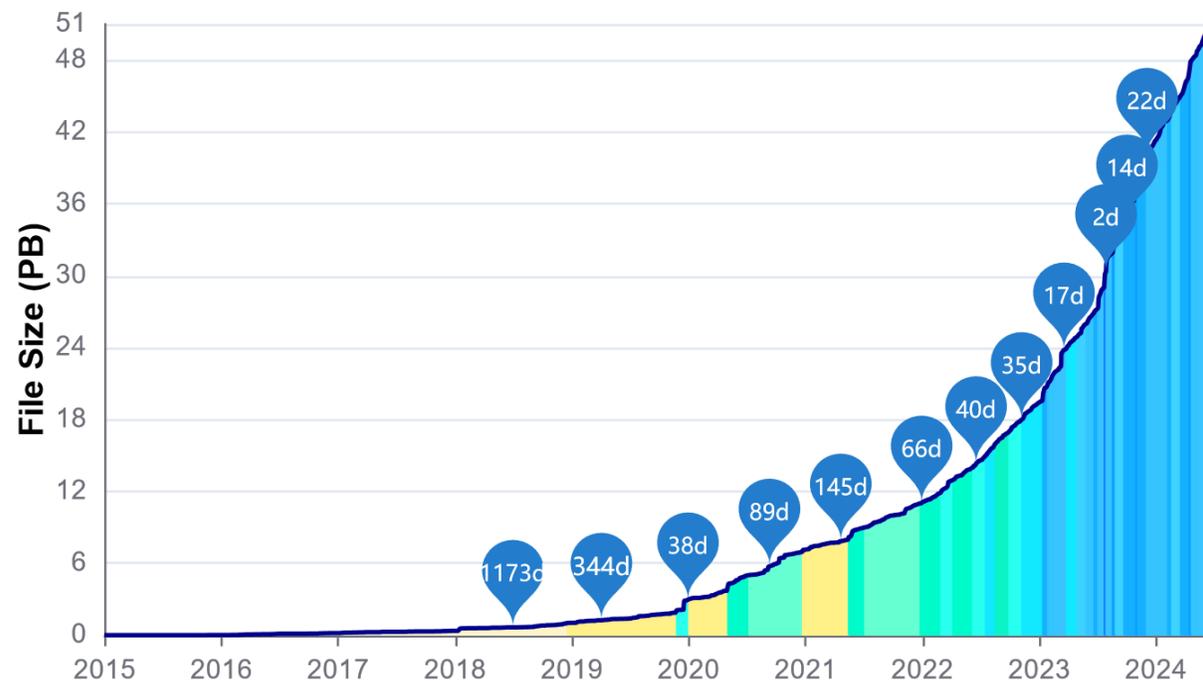
<https://ngdc.cncb.ac.cn/sdas/>



用户递交数据量: **50.86 PB**



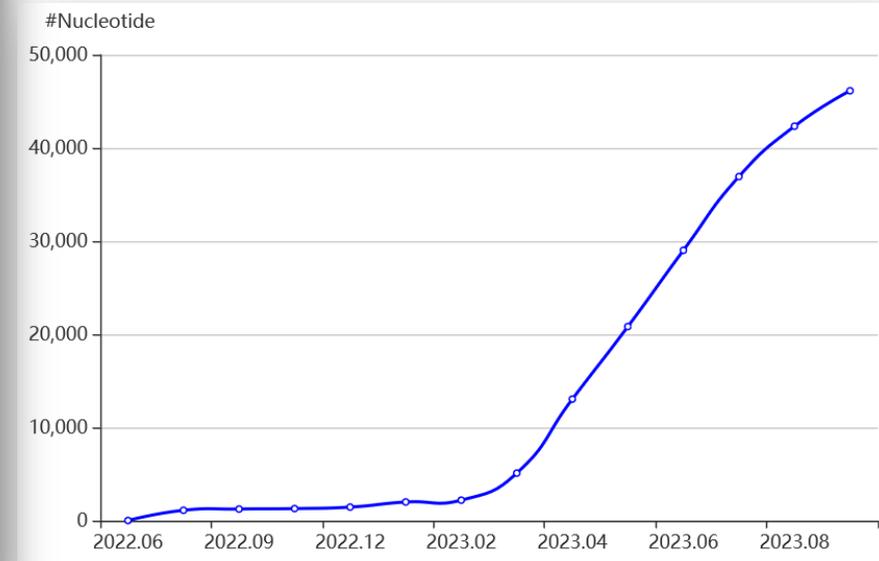
Genomics Proteomics & Bioinformatics (2017, 2021)



<https://ngdc.cnca.ac.cn/gsa/statistics>

The screenshot shows the GenBase website interface. At the top, there are navigation links: Home, Submit, Search, Statistics, Standards, and Documentation. There are also buttons for Login and Register, and a language dropdown menu. Below the navigation is a search bar with a dropdown menu set to 'Nucleotide' and a search button. A description of GenBase is provided: 'GenBase is a genetic sequence database that accepts user submissions (mRNA, genomic DNAs, ncRNA, or small genomes such as organelles, viruses, plasmids, phages from any organism) and integrates data from INSDC.' Below this, there are three main data categories: Species (592,276), Nucleotides (266,979,827), and Proteins (274,787,433). To the right, there is a 'Recent Updates' section with a table of updates from 2022-5-17 to 2023-6-25. Further right, there is an 'INSDC (GenBank) Integration' section with update date, nucleotide count, and protein count. Below that are 'New' buttons for 'SARS-CoV-2 Fast Submission' and 'FTP Download', and 'Related Links' for 'GSA' and 'GWH'. At the bottom, there are sections for 'Problems or Questions?' and 'How to cite', including a recommended citation style and references.

Direct submissions



- **GenBank Release 254.0** has been integrated, with daily updates
- **In total: 592,276** Species, **~267 mil.** Nucleotides, **~274 mil.** Proteins
- **Direct submissions: 46 k** Nucleotides, **484 k** Proteins

<http://ngdc.cncb.ac.cn/genbase>

Genome Warehouse

The Genome Warehouse (GWH) is a public repository housing genome-scale data for a wide range of species and delivering a series of web services for genome data submission, storage, release and sharing.

Submit

Deposit meta-information into GWH databases



Download

Transfer GWH data to your computer



Browse

View genome information about the released data

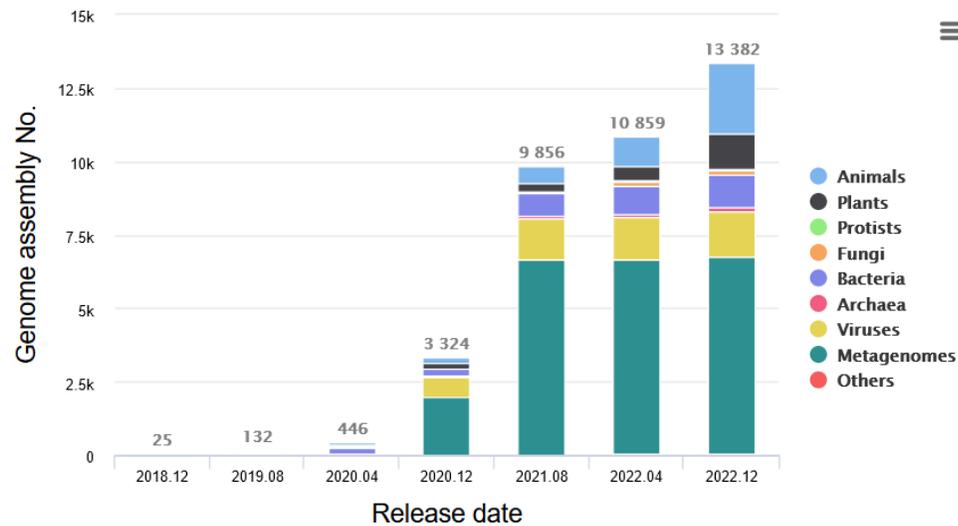


Documentation

Find help documents to learn more about GWH



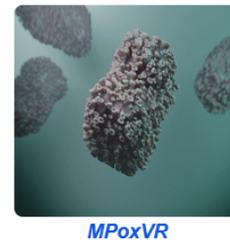
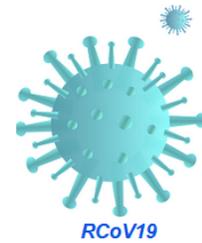
Data Growth



Statistics

- **Integrated Animals:** 61 genomes, 62 assemblies
- **Integrated Plants:** 77 genomes, 88 assemblies
- **Release of Direct Submissions (Total: 13382):** 2439 Animals; 1212 Plants; 24 Protists; 152 Fungi; 1130 Bacteria; 122 Archaea; 1543 Viruses; 6730 Metagenomes; 30 Others
- **Direct Submissions (Total: 28030):** 9216 Animals; 3272 Plants; 25 Protists; 165 Fungi; 4983 Bacteria; 130 Archaea; 2988 Viruses; 6748 Metagenomes; 503 Others

Virus Resources

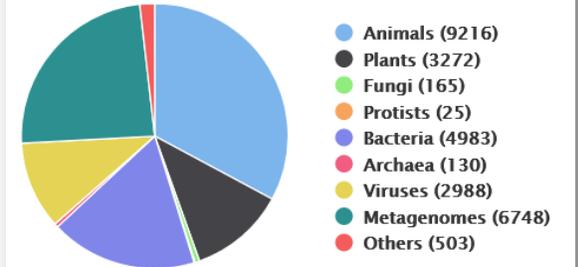


GWH-supported Deposition

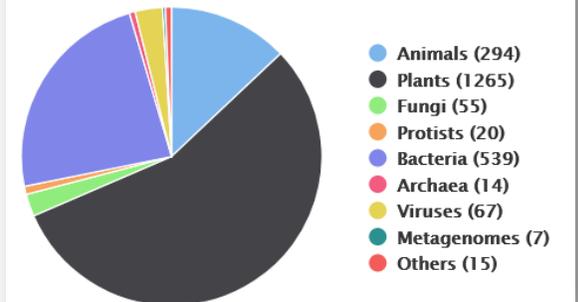
Data submissions to GWH have been reported by multiple journals, including:



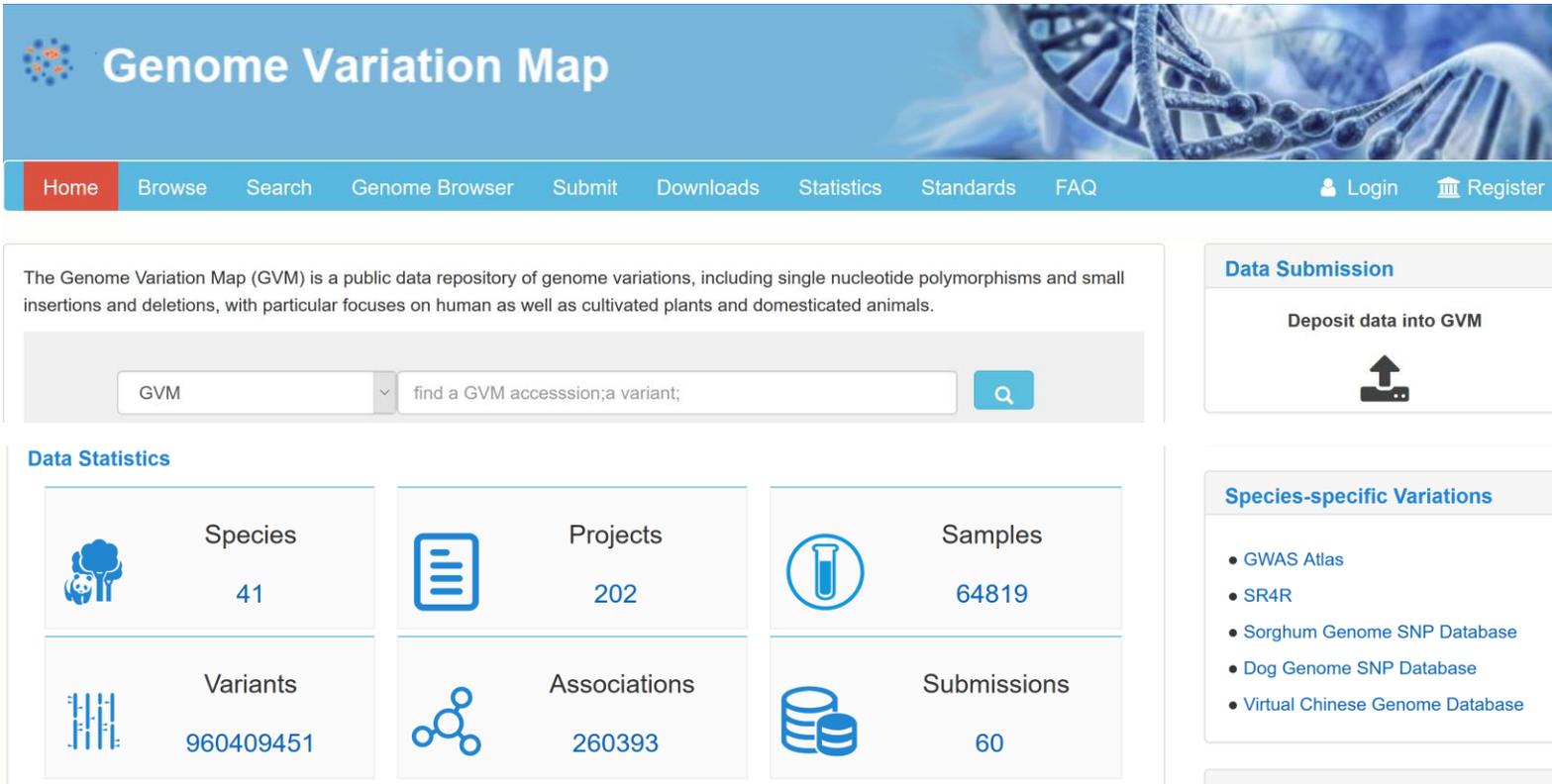
Total Assembly



Total Organism



Geno Proto Bioinfo (2021)



The Genome Variation Map (GVM) is a public data repository of genome variations, including single nucleotide polymorphisms and small insertions and deletions, with particular focuses on human as well as cultivated plants and domesticated animals.

Data Statistics

Species	Projects	Samples
41	202	64819
Variants	Associations	Submissions
960409451	260393	60

Species	Taxonomy ID	Assembly Version	Genome Size	SNP	INDEL
2019-nCoV SARS-CoV-2	2697049	MN908947.3	29.90K	16,114	820
Cannabis Cannabis sativa	3483	GCF_900626175.1	876.15M	19,855,268	4,293,684
Carrot Daucus carota	79200	Dcarota_388_v2.0	421.50M	21,396,685	5,961,358
Cassava Manihot esculenta	3983	Manihot esculenta v6.1	582.28M	27,032,954	4,872,533
Cat Felis catus	9685	Felis_catus_9.0	2.52G	38,047,892	9,316,631
Catharanthus roseus Catharanthus roseus	4058	ASM94934v1.1	522.65M	4,342,746	684,176
Cattle Bos taurus	9913	UMD_3.1	2.67G	53,609,957	6,724,343
Chicken Gallus gallus	9031	Gallus_gallus-5.0	1.23G	36,174,851	4,619,064
Common bean Phaseolus vulgaris	3885	Pvulgaris_442_v2.0	537.22M	13,876,210	3,707,970
Cotton Gossypium hirsutum	3635	Ghir.BGI	2.15G	8,669,229	4,170,512

SARS-CoV-2
SARS-CoV
MERS-CoV

人、牛、鸡、狗
鸭、大熊猫、虎鲸
家猪、绵羊

玉米、小麦、杨树
水稻、橡胶、高粱
大豆、西红柿

Nucleic Acids Research (2018, 2021)

<https://ngdc.cncb.ac.cn/gvm>



Gene Expression Nebulas (GEN) is a data portal of gene expression profiles under various conditions derived entirely from bulk and single-cell RNA-Seq data analysis in multiple species.

 Homo sapiens 141 projects 25619 experiments	 Glycine max 16 projects 499 experiments	 Glycine soja 1 projects 7 experiments	 Oryza sativa 25 projects 945 experiments	 Sorghum bicolor 5 projects 462 experiments	 Triticum aestivum 3 projects 78 experiments	 Featured Projects The latest RNA-seq research data on COVID-19. <ul style="list-style-type: none">PRJCA002326: Transcriptome of patients infected with SARS-CoV-2PRJNA615032: Transcriptional response to SARS-CoV-2 infectionPRJNA631753: Spectrum of Viral Load and Host Response Seen in Autopsies of SARS-CoV-2 Infected Lungs
---	---	---	---	--	---	--



LncBook accommodates a high-quality collection of 95,243 human lncRNA genes and 323,950 lncRNA transcripts, and incorporates their abundant annotations at different omics levels, thereby enabling users to decipher functional signatures of lncRNAs in human diseases and different biological contexts.

e.g., MALAT1; ENSG00000228630.5; HSALNG0084892; hsa-miR-619-5p; SPROHSA127174; Body height; Mus musculus;

Multi-omics Annotations

 **Conservation**
Conservation Features across 40 Vertebrates

 **Variation**
959,138 Disease/trait-associated Variants

 **Methylation**
DNA Methylation Profiles in 16 Diseases

 **Expression**
Expression Capacities across 9 Biological Contexts

 **Small Protein**
34,012 Small Proteins

 **Interaction**
146,092,274 LncRNA-miRNA Interactions;
772,745 LncRNA-Protein Interactions



2016年7月加入国际
RNAcentral联盟

<http://ngdc.cncb.ac.cn/lncbook>
Bioinformatics (2019)
Nucleic Acids Res (2015, 2019, 2022)

GWAS Atlas
A curated resource of genome-wide variant-trait associations

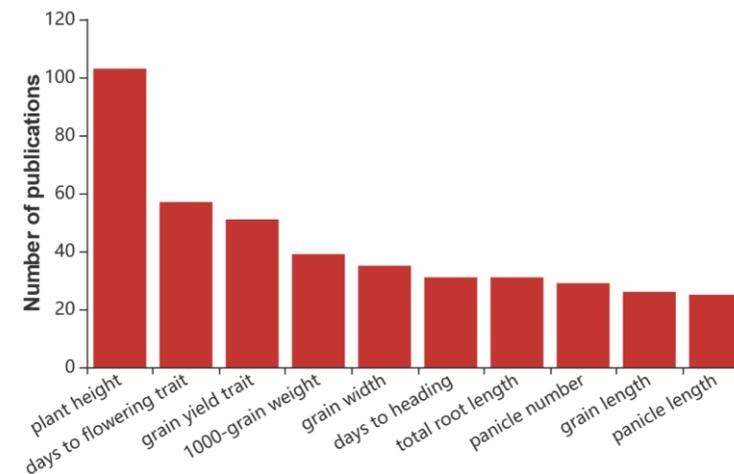
Home Browse Ontology Tools Submit Downloads Statistics Help

All Species Find a trait, gene, variant, genome-region, ...

e.g. flower; plant height; Zm00001d021954; chr1:14702150-37601000

31 Species	302295 Associations	486 Causal Variants	1724 Traits	163979 Variants
57223 Genes	3828 Studies	922 Publications	5 Ontologies	19 Submissions

Publications of Top10 Traits



Oryza sativa
Rice

163479 Associations
461 Traits



Sorghum bicolor
Sorghum

8829 Associations
151 Traits

EWAS Atlas
@EWAS Open Platform

[Browse](#)
[EWAS Toolkit](#)
[Downloads](#)
[Statistics](#)
[API](#)
[Help](#)
[EWAS Data Hub](#)



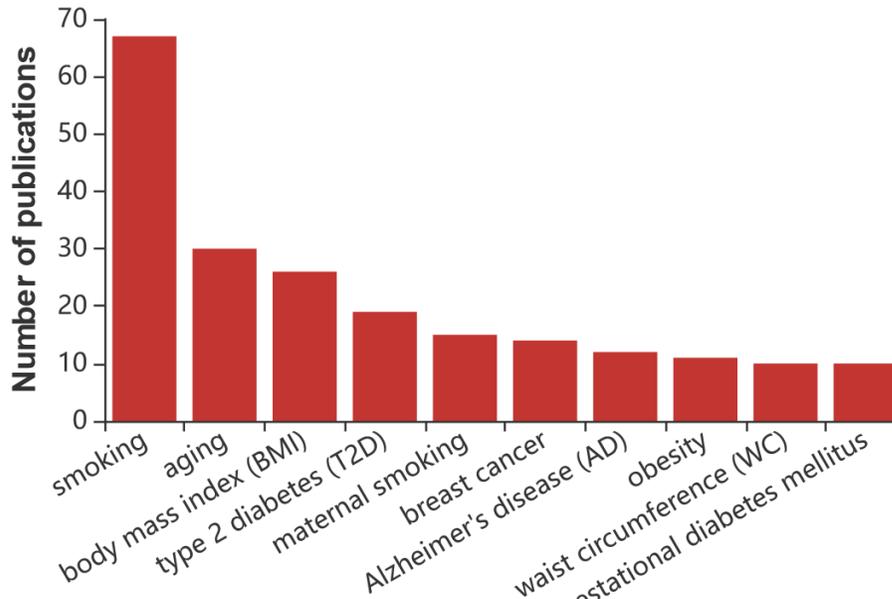
EWAS Atlas

@ EWAS Open Platform

A knowledgebase of epigenome-wide association studies

Examples: smoking, AHRR, cg05575921

TOP 10 Traits



Trait	Number of Publications
smoking	67
aging	30
body mass index (BMI)	26
type 2 diabetes (T2D)	19
maternal smoking	15
breast cancer	14
Alzheimer's disease (AD)	12
obesity	11
waist circumference (WC)	10
gestational diabetes mellitus	10

Associations



705,769

Traits



802

Cohorts



3,618

Tissues/Cells



219

Studies



1,710

Publications



1080

Last update: new EWAS on [ischemic stroke](#) has been added online on 19 June, 2024

New Database: [EWAS Data Hub](#) (A data hub of DNA methylation array data and metadata)

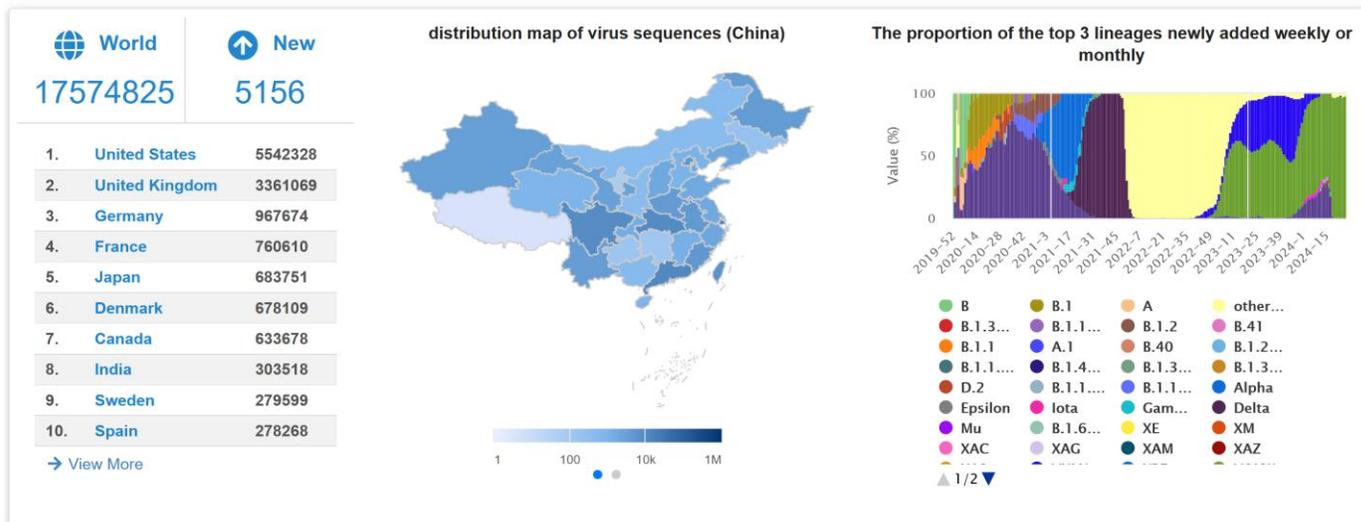
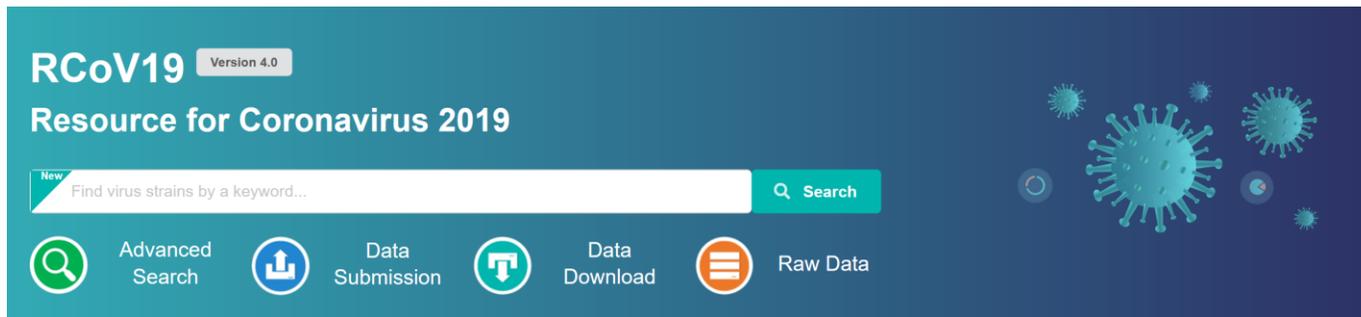
New Toolkit: [EWAS Toolkit](#) (A web toolkit for epigenome-wide association study)

Follow us: [@EWAS_Open_Platform](#)

Cite: EWAS Open Platform: integrated data, knowledge and toolkit for epigenome-wide association study. *Nucleic Acids Res.* 2021 [PMID=34718752]

EWAS Atlas: a curated knowledgebase of epigenome-wide association studies.

<http://ngdc.cncb.ac.cn/ewas>



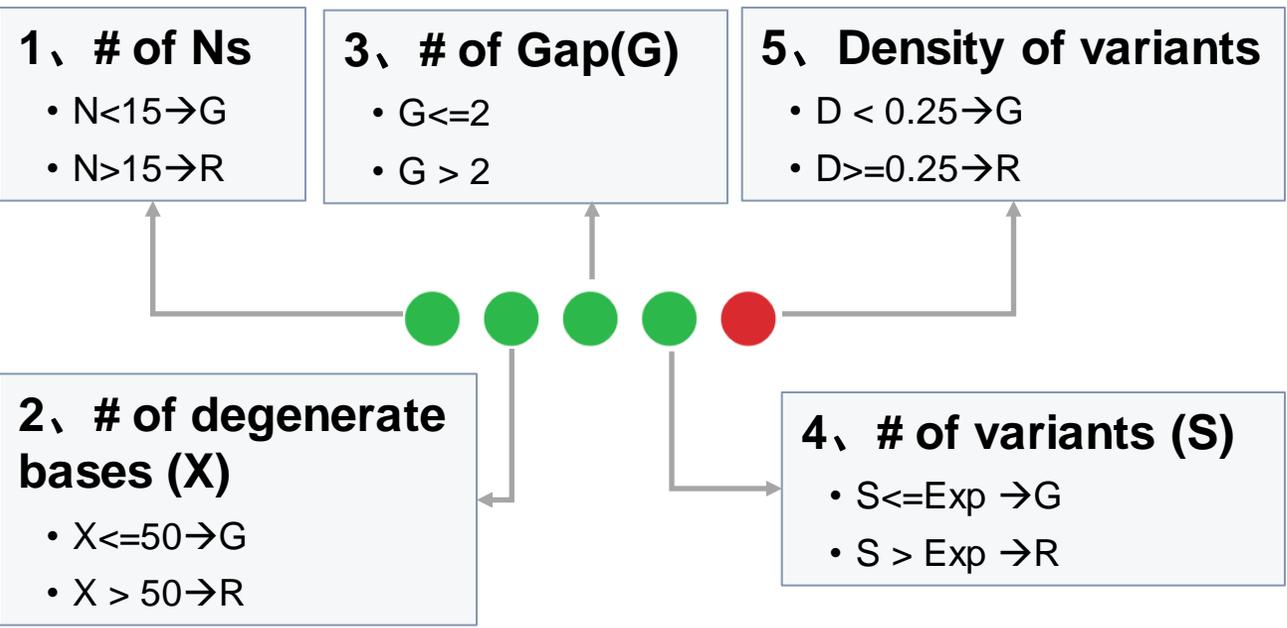
<https://ngdc.cncb.ac.cn/ncov> (as of July 6, 2024)

2020年1月22日上线，全球新冠病毒数据汇交、整合、质控、分析、监测



每日更新，服务全球

截至2024年7月6日，已收录全球**1757万**多个新冠病毒基因组序列信息，为全球**182**个国家/地区**45万**余访客提供数据服务，累计数据下载超过**202亿**条次，**国外访客占比高达60%**



Nuc.Completeness	Sequence Length	Sequence Quality	Quality Assessment
Complete	29834	High	●●●●●
Complete	29782	High	●●●●●
Complete	29782	Low	●●○○○
Complete	29782	Low	●●○○○
Complete	29782	High	●●●●●

~48% complete & high quality



50004651
CORONAVIRUS SEQUENCES

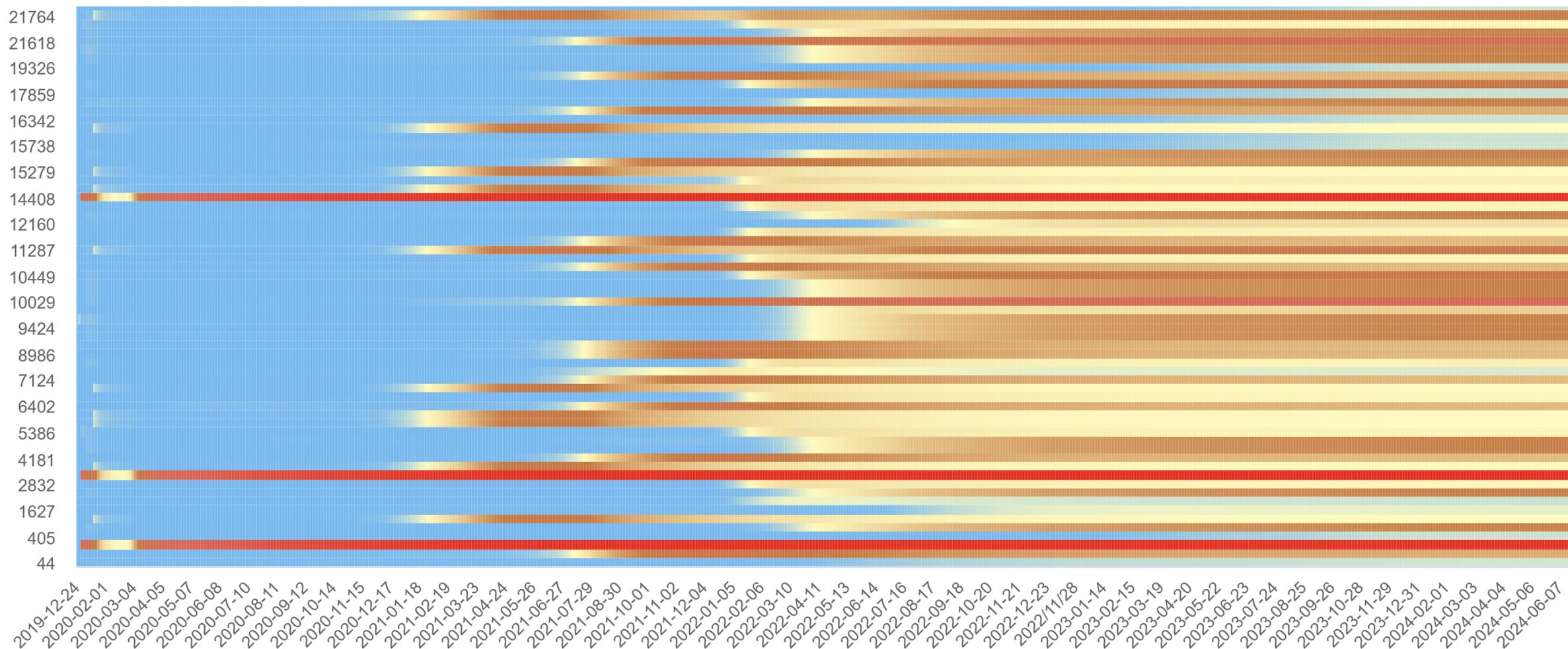
17472053
ISOLATES

17574825
SARS-COV-2 SEQUENCES

3340
SUBMITTING LABS

13072
ORIGINATING LABS

9613
LOCATIONS

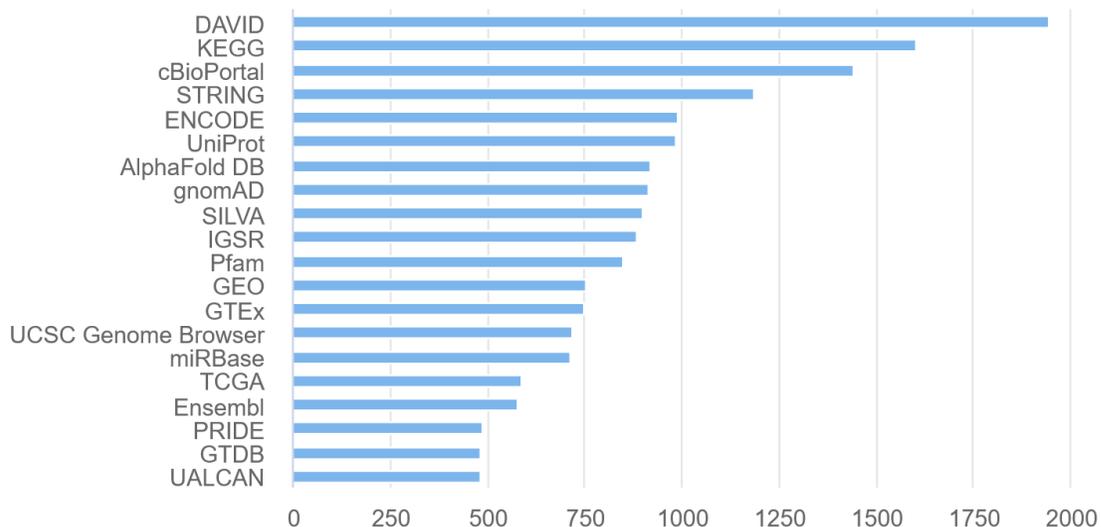


<https://ngdc.cncb.ac.cn/ncov/variation/heatmap>

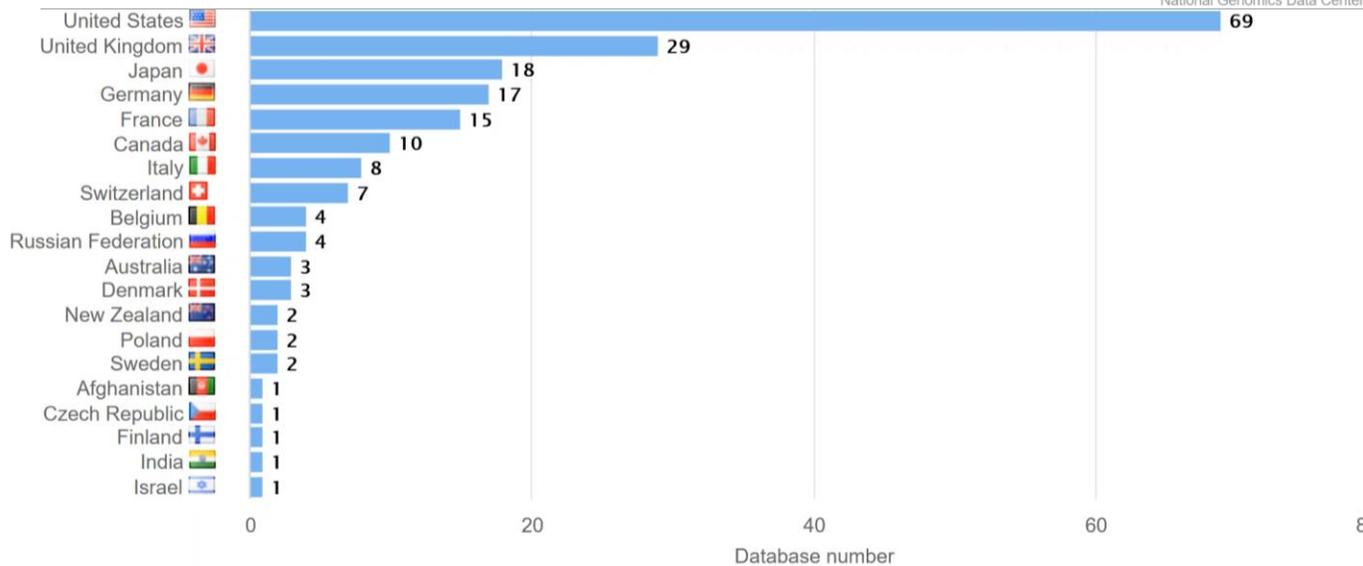
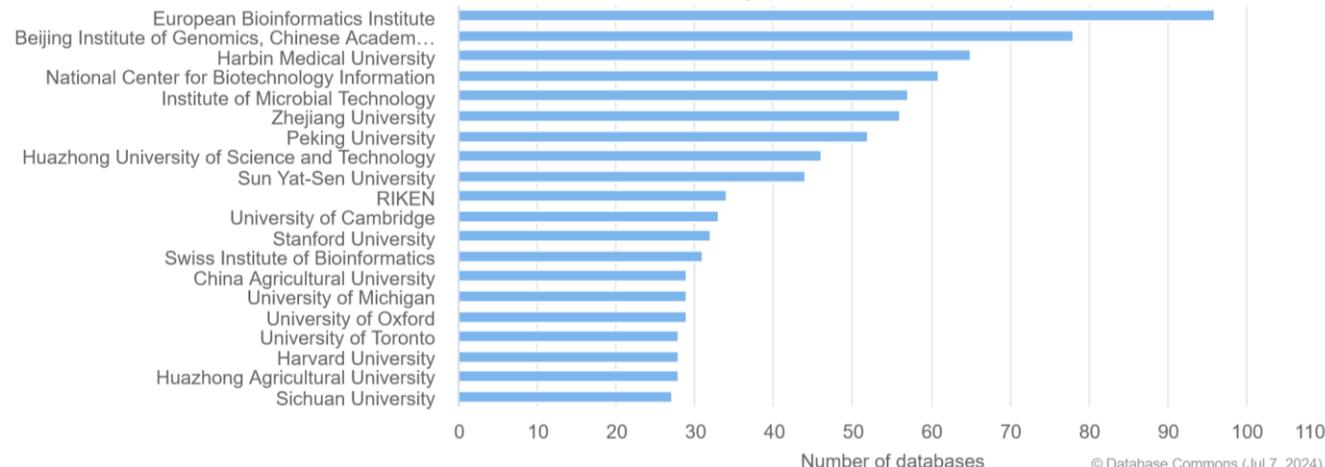
6392 DATABASES
13 CATEGORIES
1647 SPECIES

9875 PUBLICATIONS
2142 INSTITUTIONS
76 COUNTRIES / REGIONS

Top 20 databases by z-index (=citation/age)



Top 20 institutions by database count





索引记录数: 14.77亿 索引大小: 1.5TB



Q BIG Search

BIG Search is a scalable text search engine built based on Elasticsearch (a highly scalable open-source full-text search and analytics engine based on Apache Lucene). It features cross-domain search and facilitates users to gain access to a wide range of biomedical data, not only from NGDC databases but also partner databases throughout the world.

e.g., PRJCA000126;SAMC000385;tp53;EGFR; human; KaKs_Calculator

NGDC & Partners Databases

2,685 records from 37 NGDC & Partner databases.

Show entries Filter:

AnimalTFDB	89	AnimalTFDB is a comprehensive database including classification and annotation of genome-wide transcription factors
BBCancer	52	BBCancer: an expression atlas of blood-based biomarkers in the early diagnosis of cancers
BioCode	5	Archive Bioinformatics Codes for Open Source Projects
BioProject	2	Biological Project Library
BioSample	2	Biological Sample Library

各数据库匹配到的数据条目

第三节 国际重要生物数据库

生物大分子数据库 (DNA、RNA、Protein)

生物大分子数据库 (DNA、RNA、Protein)

基因组与转录组数据库: Ensembl, RefSeq, GEO, KEGG, GENE ONTOLOGY, NOG, EggNOG

RNA数据库: RNAmod, RNA Mod, Rfam, RMBASE v3.0, RNALocate, NONCODE, RNAcentral, LncBook 2.0, ENCORI, lncRNASNP v3, miRBase, mRWalk, circBase, circAtlas

蛋白质数据库: UniProt, InterPro, ExPasy (Swiss Bioinformatics Resource Portal), DisProt, MobiDB, PDB (Protein Data Bank), proSite, Pfam, IntAct, MINT, STRING

物种专题数据库

物种专题数据库

人类健康: neXtProt, TCGA, hmdb, 5 YEARS 5MIM

模式动物: THE HUMAN PROTEIN ATLAS, MGI, RGD, Mouse Phenome Database, ZFIN

作物: FlyBase, WormBase

拟南芥: tair, RICE, CTBP, maizeGDB, rap-db

真菌: SoyBase, CottonMD, Phytozome, unite community, PomBase, SGD, Saccharomyces GENOME DATABASE

细菌: MG-RAST, BV-BRC, IMG/M, GISAID, GTDB, BioCloud, silva, ECOCYC, BacDive

病毒: (无具体数据库名称)

其他类型数据库

其他类型数据库

甲基化: MethDB, MEXPRESS

启动子: OrthoDB, MethHC, EPD (Eukaryotic Promoter Database), reactome

转录子: JASPAR 2024, NAKB (Nucleic Acid Knowledgebase)

蛋白酶: BRENDA, CAZy (Carbohydrate-Active Enzymes)

基因组与功能注释

RefSeq 高质量参考基因组数据库

Ensembl 整合型基因组数据库

KEGG 功能基因与代谢通路参考数据库

GO 基因本体分类数据库

COG/KOG 原核/真核同源基因数据库

EggNOG 拓展型多物种同源基因集

基因型与表型

dbGaP 基因型与表型关联数据库

结构

NDB/3DNA DNA空间三维结构数据库

基因表达

GEO 整合型基因表达数据库

功能元件

EPD 真核生物启动子数据库

JASPAR/TRANSFAC 转录因子数据库

TRRD 转录调控区数据库

DNA

综合性数据库

Rfam 综合性RNA家族数据库

RNAmod/RMBase 转录修饰数据库

RNALocate 非编码RNA亚细胞定位

miRNA

MiRBase/microRNA.org 综合数据库

miRWalk miRNA靶点数据库

PolymiRTS miRNA靶位点多态性

lncRNA

NONCODE/LNCipedia 长非编码RNA

lncRNA SNP SNP对lncRNA的影响

lncRNADisease lncRNA疾病数据库

circRNA

CircAtlas/circBase 环形RNA数据库

rRNA

SILVA/RDP/GreenGene 核糖体RNA

RNA

综合性数据库

UniProt/InterPro 整合型蛋白数据库

ExPASy 综合性蛋白分析系统

空间结构

PDB 综合结构数据库

AlphaFoldDB/ESM Atlas 预测结构

分类与注释

CATH/SCOP 结构分类数据库

PRINTS 蛋白家族指纹图谱数据库

DisProt/MobiDB 无序蛋白数据库

互作

IntAct/MINT/BioGRID 综合性数据库

STRING 高质量种内互作数据库

结构域

Pfam 蛋白质家族与结构域分类数据库

PROSITE 结构域与功能位点数据库

Protein

基因组与功能注释

RefSeq 高质量参考基因组数据库

Ensembl 整合型基因组数据库

KEGG 功能基因与代谢通路参考数据库

GO 基因本体分类数据库

COG/KOG 原核/真核同源基因数据库

EggNOG 拓展型多物种同源基因集

基因型与表型

dbGaP 基因型与表型关联数据库

结构

NDB/3DNA DNA空间三维结构数据库

基因表达

GEO 整合型基因表达数据库

功能元件

EPD 真核生物启动子数据库

JASPAR/TRANSFAC 转录因子数据库

TRRD 转录调控区数据库

DNA

综合性数据库

Rfam 综合性RNA家族数据库

RNAmod/RMBase 转录修饰数据库

RNALocate 非编码RNA亚细胞定位

miRNA

MiRBase/microRNA.org 综合数据库

miRWalk miRNA靶点数据库

PolymiRTS miRNA靶位点多态性

lncRNA

NONCODE/LNCipedia 长非编码RNA

lncRNA SNP SNP对lncRNA的影响

lncRNADisease lncRNA疾病数据库

circRNA

CircAtlas/circBase 环形RNA数据库

rRNA

SILVA/RDP/GreenGene 核糖体RNA

RNA

综合性数据库

UniProt/InterPro 整合型蛋白数据库

ExPASy 综合性蛋白分析系统

空间结构

PDB 综合结构数据库

AlphaFoldDB/ESM Atlas 预测结构

分类与注释

CATH/SCOP 结构分类数据库

PRINTS 蛋白家族指纹图谱数据库

DisProt/MobiDB 无序蛋白数据库

互作

IntAct/MINT/BioGRID 综合性数据库

STRING 高质量种内互作数据库

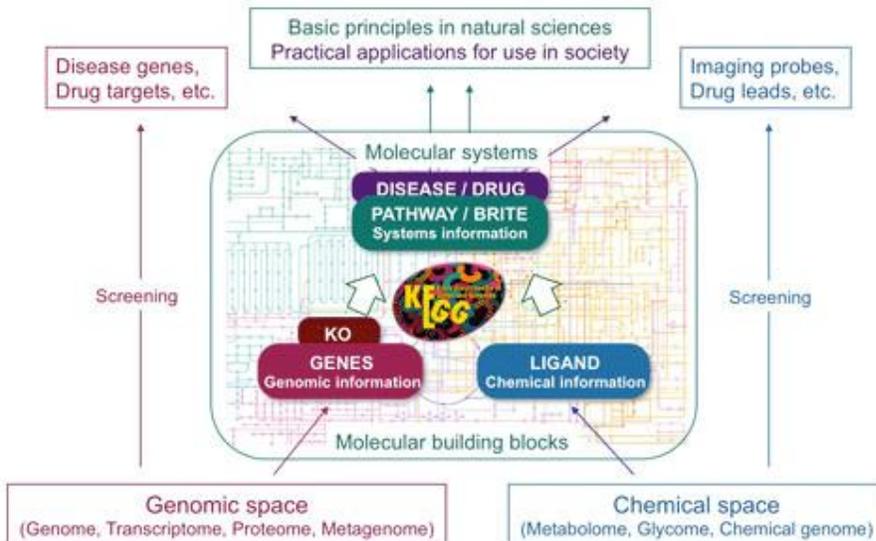
结构域

Pfam 蛋白质家族与结构域分类数据库

PROSITE 结构域与功能位点数据库

Protein

1. KEGG数据库由**金久实 (Minoru Kanehisa, 1948—)** 实验室于**1995**年建立，研究团队从公开发表的文献和其他数据库中提取基因功能数据，**标准化处理后**以统一的格式存储。
2. KEGG由多个数据库组成，最为核心的是**KEGG PATHWAY**。研究人员通过阅读大量的科学文献，**手动绘制各种代谢通路、信号通路和其他生物过程的图示**。这些图示结合了基因、蛋白质、代谢物等信息，直观地展示了生物分子间的相互作用和功能关系。
3. KEGG保持了**长期稳定的更新**，提供了直观的工具和API，方便各类研究人员使用。



- KEGG Home**
 - Release notes
 - Current statistics
- KEGG Database**
 - KEGG overview
 - KEGG mapping
 - Color codes
- KEGG Objects**
 - KEGG Weblinks
 - Entry format
- KEGG Software**
 - KEGG API
 - KGML
- KEGG FTP
 - Subscription
 - Background info
- GenomeNet
- DBGET/LinkDB
- Feedback
- Copyright request
- Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (July 1, 2024) for new and updated features.

- **Main entry point to the KEGG web service**
 - KEGG2** KEGG Table of Contents [[Update notes](#) | [Release history](#)]
- **Data-oriented entry points**
 - KEGG PATHWAY** KEGG pathway maps
 - KEGG BRITE** BRITE hierarchies and tables
 - KEGG MODULE** KEGG modules
 - KEGG ORTHOLOG** KO functional orthologs
 - KEGG GENES** Genes and proteins [[Annotation](#)]
 - KEGG GENOME** Genomes [[KEGG Virus](#) | [Syntax](#)]
 - KEGG COMPOUND** Small molecules
 - KEGG GLYCAN** Glycans
 - KEGG REACTION** Biochemical reactions [[RModule](#)]
 - KEGG ENZYME** Enzyme nomenclature
 - KEGG NETWORK** Disease-related network variations
 - KEGG DISEASE** Human diseases
 - KEGG DRUG** Drugs [[New drug approvals](#)]
- **KEGG MEDICUS** Health information resource [[Drug labels search](#)]
- **Organism-specific entry points**
 - KEGG Organisms** Enter org code(s) hsa hsa eco
- **Analysis tools**
 - KEGG Mapper** PATHWAY/BRITE/MODULE mapping tools
 - KEGG Web Apps** Pathway viewer with coloring features, etc.
 - KEGG Syntax** Genome alignment to analyze conserved synteny
 - BlastKOALA** BLAST-based KO annotation and KEGG mapping
 - GhostKOALA** GHOSTX-based KO annotation and KEGG mapping
 - BLAST/FASTA** Sequence similarity search
 - SIMCOMP** Chemical structure similarity search

- Pathway
- Brite
- Brite table
- Module
- Network
- KO (Function)
- Organism
- Virus
- Compound
- Disease (ICD)
- Drug (ATC)
- Drug (Target)
- Antimicrobials

KEGG数据库首页 (2024.7)

KEGG Escherichia coli K-12 MG1655

Genome info | Pathway map | Brite hierarchy | Module | Genome browser

Search genes:

Genome information

T number T00007
Org_code eco
Name Escherichia coli K-12 MG1655
Category Reference genome
Annotation yes
Taxonomy TAX: 511145
Lineage Bacteria; Pseudomonadota; Gammaproteobacteria
Brite KEGG organisms [BR:br08601]
 KEGG organisms in the NCBI taxonomy [BR:br08601]
 KEGG organisms in taxonomic ranks [BR:br08601]
Data source RefSeq (Assembly: GCF_000005845.2 Complete Genome Annotation) BioProject: 57779
Original DB Wisconsin, Pasteur, ECOLIBC, ASAP
Comment Colonizes the lower gut of animals, and, as a model organism, is maintained as a laboratory strain with minimal genetic modification by bacteriophage lambda and F plasmid by ultraviolet mutagenesis.
Chromosome Circular
Sequence RS: NC_000913 (GB: U00096)
Length 4641652
Statistics Number of nucleotides: 4641652
 Number of protein genes: 4278
 Number of RNA genes: 186

KEGG pathway maps

Metabolism

Global and overview maps

- 01100 Metabolic pathways
- 01110 Biosynthesis of secondary metabolites
- 01120 Microbial metabolism in diverse environments
- 01200 Carbon metabolism
- 01210 2-Oxocarboxylic acid metabolism
- 01212 Fatty acid metabolism
- 01230 Biosynthesis of amino acids
- 01232 Nucleotide metabolism
- 01250 Biosynthesis of nucleotide sugars
- 01240 Biosynthesis of cofactors
- 01220 Degradation of aromatic compounds

Carbohydrate metabolism

- 00010 Glycolysis / Gluconeogenesis
- 00020 Citrate cycle (TCA cycle)
- 00030 Pentose phosphate pathway
- 00040 Pentose and glucuronate interconversions
- 00051 Fructose and mannose metabolism
- 00052 Galactose metabolism
- 00053 Ascorbate and aldarate metabolism
- 00500 Starch and sucrose metabolism
- 00520 Amino sugar and nucleotide sugar metabolism
- 00620 Pyruvate metabolism
- 00630 Glyoxylate and dicarboxylate metabolism
- 00640 Propanoate metabolism
- 00650 Butanoate metabolism
- 00660 C5-Branched dibasic acid metabolism
- 00562 Inositol phosphate metabolism

Energy metabolism

- 00190 Oxidative phosphorylation
- 00710 Carbon fixation by Calvin cycle
- 00720 Other carbon fixation pathways
- 00680 Methane metabolism

Escherichia coli K-12 MG1655 K+ID, KEGG gene functional orthologs

第一级分类 Metabolism

第二级分类 Carbohydrate metabolism

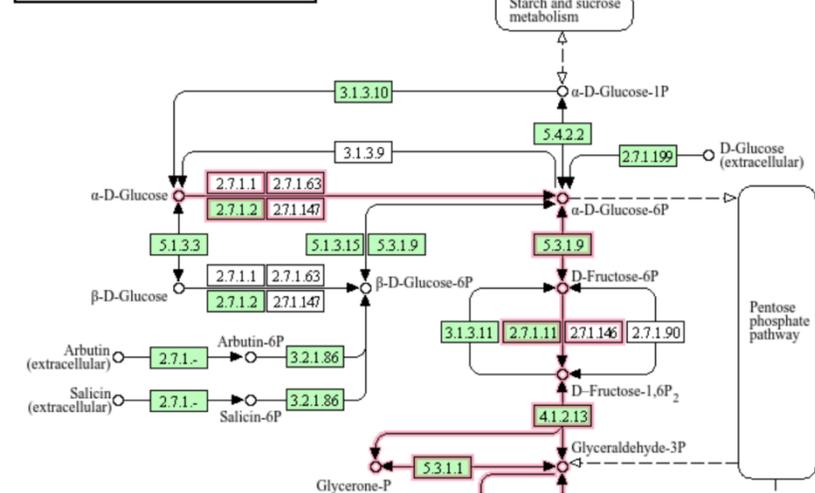
第三级分类 Glycolysis / Gluconeogenesis

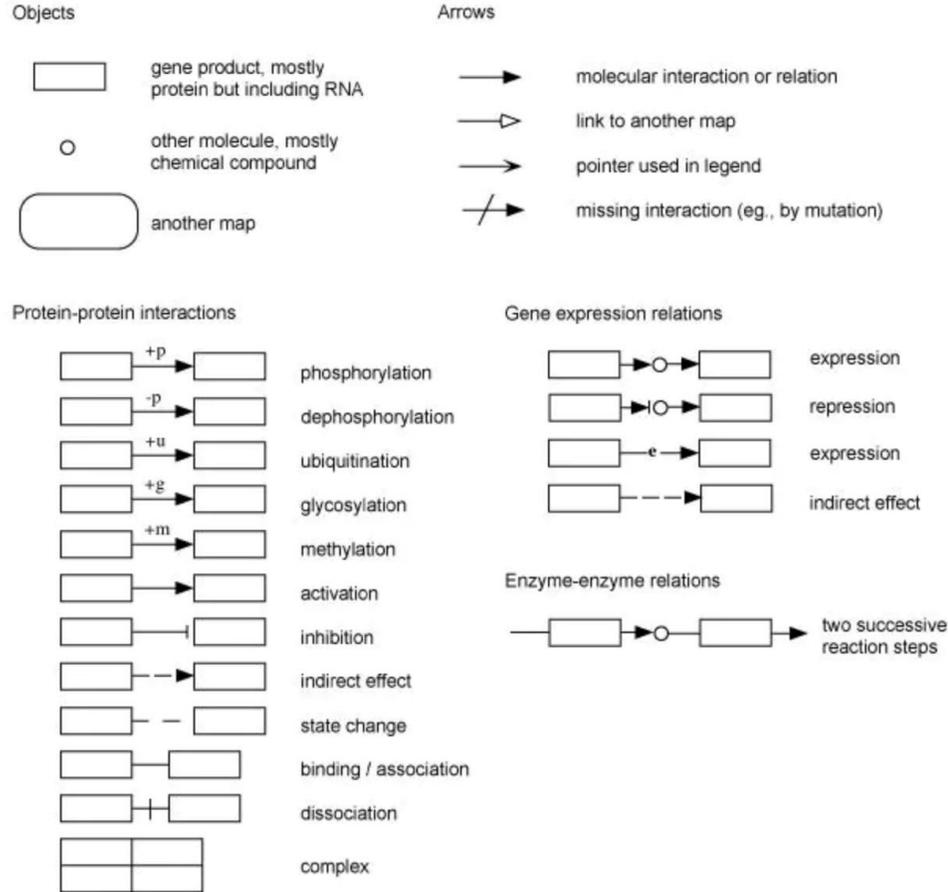
Download json | Copy URL | Help

```

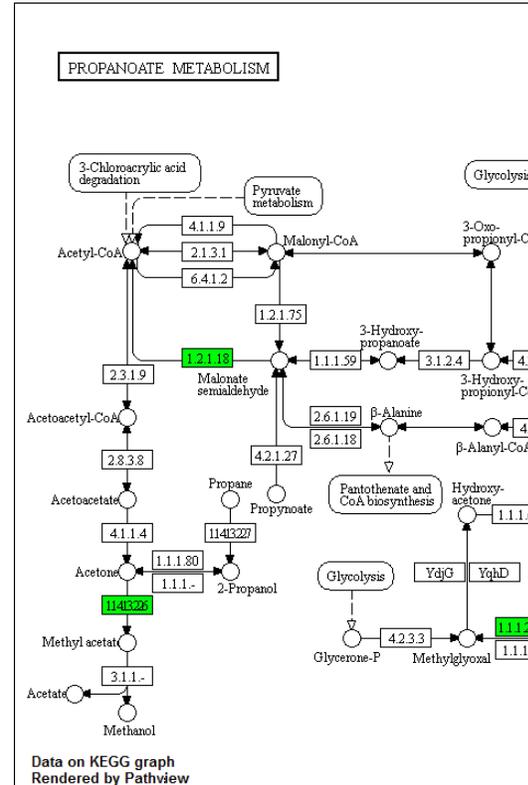
09100 Metabolism
  09101 Carbohydrate metabolism
    00010 Glycolysis / Gluconeogenesis [PATH:eco00010]
      b2388 glk; glucokinase
      b4025 pgi; glucose-6-phosphate isomerase
      b3916 pfkA; 6-phosphofructokinase 1
      b1723 pfkB; 6-phosphofructokinase 2
      b4232 fbp; fructose-1,6-bisphosphatase 1
      b3925 glpK; fructose-1,6-bisphosphatase 2
      b2930 yggF; fructose 1,6-bisphosphatase YggF
      b2097 fbaB; fructose-bisphosphate aldolase class I
      b2925 fbaA; fructose-bisphosphate aldolase class II
      b1919 tpiA; triose-phosphate isomerase
      b1779 gapA; glyceraldehyde-3-phosphate dehydrogenase A
      b2926 pgi; phosphoglycerate kinase
      b0756 gpmA; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase
      b4395 ytcC; putative phosphatase
      b3612 gpmM; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase
      b2779 eno; enolase
      b1676 pykF; pyruvate kinase I
      b1854 pykA; pyruvate kinase II
      b1702 ppsA; phosphoenolpyruvate synthase
      b0114 aceE; pyruvate dehydrogenase E1 component
      b0115 aceF; pyruvate dehydrogenase, E2 subunit
      b0116 lpd; liponamide dehydrogenase
      b1378 ydcK; putative pyruvate-flavodoxin oxidoreductase
      K00845 glk; glucokinase [EC:2.7.1.2]
      K01810 GPI; glucose-6-phosphate isomerase [EC:5.3.1.9]
      K00850 pfkA; 6-phosphofructokinase 1 [EC:2.7.1.11]
      K16370 pfkB; 6-phosphofructokinase 2 [EC:2.7.1.11]
      K03841 FBP; fructose-1,6-bisphosphatase I [EC:3.1.3.11]
      K02446 glpK; fructose-1,6-bisphosphatase II [EC:3.1.3.11]
      K02446 glpK; fructose-1,6-bisphosphatase II [EC:3.1.3.11]
      K11645 fbaB; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]
      K01624 FBA; fructose-bisphosphate aldolase, class II [EC:4.1.2.13]
      K01803 TPI; triosephosphate isomerase (TIM) [EC:5.3.1.1]
      K00134 GAPDH; glyceraldehyde 3-phosphate dehydrogenase (phosphorylating) [EC:1.2.1.12]
      K00927 PGK; phosphoglycerate kinase [EC:2.7.2.3]
      K01834 PGAM; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.2]
      K15634 gpmB; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.2]
      K15633 gpmI; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase [EC:5.4.2.2]
      K01689 EMO1_2_3; enolase 1/2/3 [EC:4.2.1.11]
      K00873 PK; pyruvate kinase [EC:2.7.1.40]
      K00873 PK; pyruvate kinase [EC:2.7.1.40]
      K01007 pps; pyruvate, water dikinase [EC:2.7.9.2]
      K00163 aceE; pyruvate dehydrogenase E1 component [EC:1.2.4.1]
      K00627 DLAT; pyruvate dehydrogenase E2 component (dihydrolipoyllysine-residue ace)
      K00382 DLD; dihydrolipoyl dehydrogenase [EC:1.8.1.4]
      K03737 por; pyruvate-ferredoxin/flavodoxin oxidoreductase [EC:1.2.7.1 1.2.7.-]
    
```

GLYCOLYSIS / GLUCONEOGENESIS





KEGG通路图图例



KEGG多数据库关联体系

KEGG		ORTHOLOGY: K00077	Help
Entry	K00077	KO	
Symbol	panE, apbA		
Name	2-dehydropantoate 2-reductase [EC:1.1.1.169]		
Pathway	map00770 Pantothenate and CoA biosynthesis map01100 Metabolic pathways map01110 Biosynthesis of secondary metabolites map01240 Biosynthesis of cofactors		
Module	M00119 Pantothenate biosynthesis, valine/L-aspartate => pantothenate M00913 Pantothenate biosynthesis, 2-oxoisovalerate/spermine => pantothenate M00914 Coenzyme A biosynthesis, archaea, 2-oxoisovalerate => 4-phosphopantoate => CoA		
Brite	KEGG Orthology (KO) [BR:ko00001] 09100 Metabolism 09108 Metabolism of cofactors and vitamins 00770 Pantothenate and CoA biosynthesis K00077 panE, apbA; 2-dehydropantoate 2-reductase Enzymes [BR:ko01000] 1. Oxidoreductases 1.1 Acting on the CH-OH group of donors 1.1.1 With NAD+ or NADP+ as acceptor 1.1.1.169 2-dehydropantoate 2-reductase K00077 panE, apbA; 2-dehydropantoate 2-reductase BRITE hierarchy		
Other DBs	RN: R02472 COG: COG1893 GO: 0008677		
Genes	PXB: 103939761 OLU: OSTLU_13695 SCE: YHR063C(PAN5) SEUB: DI49_2466 SPAO: SPAR_H01040 AGO: AGOS_AAL044C KLA: KLLA0_C14674g KMX: KLMA_30545(PAN5) LTH: KLTH0F07172g VPO: Kpo1_1005p8 Kpo1_543p77 » show all Taxonomy UniProt		
Reference	PMID:9488683		
Authors	Frodyma ME, Downs D		
Title	ApbA, the ketopantoate reductase enzyme of Salmonella typhimurium is required for the synthesis of thiamine via the alternative pyrimidine biosynthetic pathway.		
Journal	J Biol Chem 273:5572-6 (1998)		
	DOI:10.1074/jbc.273.10.5572		
Reference	PMID:7519593		
Authors	Downs DM, Petersen L		
Title	apbA, a new genetic locus involved in thiamine biosynthesis in Salmonella typhimurium.		
Journal	J Bacteriol 176:4858-64 (1994)		
	DOI:10.1128/JB.176.16.4858-4864.1994		
Sequence	[stm:STM0434]		
LinkDB	All DBs		

基因组与功能注释

RefSeq 高质量参考基因组数据库

Ensembl 整合型基因组数据库

KEGG 功能基因与代谢通路参考数据库

GO 基因本体分类数据库

COG/KOG 原核/真核同源基因数据库

EggNOG 拓展型多物种同源基因集

DNA

基因型与表型

dbGaP 基因型与表型关联数据库

结构

NDB/3DNA DNA空间三维结构数据库

基因表达

GEO 整合型基因表达数据库

功能元件

EPD 真核生物启动子数据库

JASPAR/TRANSFAC 转录因子数据库

TRRD 转录调控区数据库

综合性数据库

Rfam 综合性RNA家族数据库

RNAmod/RMBase 转录修饰数据库

RNALocate 非编码RNA亚细胞定位

RNA

miRNA

MiRBase/microRNA.org 综合数据库

miRNA靶点数据库

Polym iRTS miRNA靶位点多态性

lncRNA

NONCODE/LNCipedia 长非编码RNA

lncRNA SNP SNP对lncRNA的影响

lncRNADisease lncRNA疾病数据库

circRNA

CircAtlas/circBase 环形RNA数据库

rRNA

SILVA/RDP/GreenGene 核糖体RNA

综合性数据库

UniProt/InterPro 整合型蛋白数据库

ExPASy 综合性蛋白分析系统

空间结构

Protein

PDB 综合结构数据库

AlphaFoldDB/ESM Atlas 预测结构

分类与注释

CATH/SCOP 结构分类数据库

PRINTS 蛋白家族指纹图谱数据库

DisProt/MobiDB 无序蛋白数据库

互作

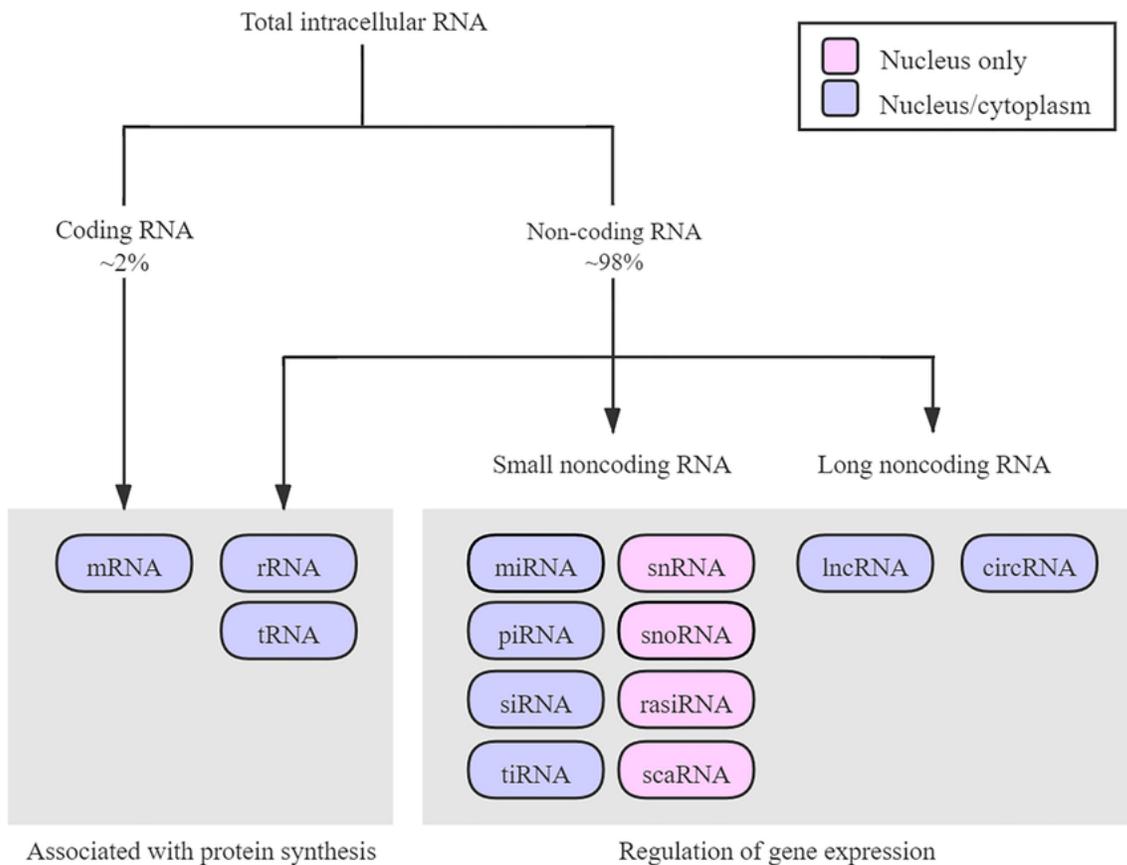
IntAct/MINT/BioGRID 综合性数据库

STRING 高质量种内互作数据库

结构域

Pfam 蛋白质家族与结构域分类数据库

PROSITE 结构域与功能位点数据库



Frontiers in Neurology (2021)

数据库名	运营情况	序列	circBaseID	host gene	miRNA	translation	m6A	RBP	Conserved	批量下载
circBase	异常	*	*	*						*
circRNADb	正常	*		*		*				*
circBank	正常	*	*	*	*	*	*		*	*
circAtlas	正常	*		*				*	*	
CircInteractome	异常		*		*			*		*
LincSNP	异常		*	*						
starBase	正常		*		*			*		*
CIRCpedia	正常	*		*						*
deepBase	正常	*		*					*	*
TSCD	正常		*	*	*			*		*
TRCirc	正常		*							
NPInter(V4)	正常		*		*			*		*
circVAR	正常		*	*						
isoCirc	正常	*	*	*						*
lncTarD	正常			*	*			*		*
AtCircDB	异常			*	*					*
GreenCircRNA	异常	*			*					
At-C-RNA	异常			*	*					*
PlantcircBase	正常	*		*	*	*		*	*	*
PlantCircNet	正常			*	*	*				*
CropCircDB	正常	*		*	*	*				*
PanCircBase	异常	*	*	*	*	*		*	*	
exoRBase	正常		*	*	*					*
piqCircNet	正常	*		*	*	*				*
VirusCircBase	正常	*		*	*					*
RNALOCATE	正常		*							*
Circ2Traits	异常			*	*					*
CircR2Disease	异常			*	*			*		
CCRDB	异常									
CircNet	正常	*		*	*	*		*	*	*
circRNADisease	正常		*	*	*	*		*	*	*
CSCD	正常	*		*	*	*		*	*	*
MiOncoCirc	正常			*	*	*				*
CircRIC	正常			*	*	*		*		
Circad	正常			*	*	*				
CircR2Cancer	正常			*	*	*				*
circExp	正常		*	*	*	*				*
circMine	正常		*	*	*	*				*
CircFunBase	正常	*	*	*	*	*		*		*
circ2GO	正常			*	*	*				*
ncEP	正常									*
IRESbase	正常	*	*	*	*	*		*	*	*
TransCirc	正常	*	*	*	*	*	*	*	*	*
riboCIRC	正常		*	*	*	*	*	*	*	*

转录与调控数据库 (10)

基因组与功能注释

RefSeq 高质量参考基因组数据库

Ensembl 整合型基因组数据库

KEGG 功能基因与代谢通路参考数据库

GO 基因本体分类数据库

COG/KOG 原核/真核同源基因数据库

EggNOG 拓展型多物种同源基因集

DNA

基因型与表型

dbGaP 基因型与表型关联数据库

结构

NDB/3DNA DNA空间三维结构数据库

基因表达

GEO 整合型基因表达数据库

功能元件

EPD 真核生物启动子数据库

JASPAR/TRAN SFAC 转录因子数据库

TRRD 转录调控区数据库

综合性数据库

Rfam 综合性RNA家族数据库

RNAmod/RMBase 转录修饰数据库

RNALocate 非编码RNA亚细胞定位

RNA

miRNA

MiRBase/microRNA.org 综合数据库

miRNAalk miRNA靶点数据库

Polym iRTS miRNA靶位点多态性

lncRNA

NONCODE/LNCipedia 长非编码RNA

lncRNA SNP SNP对lncRNA的影响

lncRNADisease lncRNA疾病数据库

circRNA

CircAtlas/circBase 环形RNA数据库

rRNA

SILVA/RDP/GreenGene 核糖体RNA

综合性数据库

UniProt/InterPro 整合型蛋白数据库

ExPASy 综合性蛋白分析系统

空间结构

Protein

PDB 综合结构数据库

AlphaFoldDB/ESM Atlas 预测结构

分类与注释

CATH/SCOP 结构分类数据库

PRINTS 蛋白家族指纹图谱数据库

DisProt/MobiDB 无序蛋白数据库

互作

IntAct/MINT/BioGRID 综合性数据库

STRING 高质量种内互作数据库

结构域

Pfam 蛋白质家族与结构域分类数据库

PROSITE 结构域与功能位点数据库

UniProt BLAST Align Peptide search ID mapping SPARQL Release 2024_03 | Statistics Help

Find your protein

UniProtKB | Advanced | List | Search

Examples: Insulin, APP, Human, P05067, organism_id:9606

UniProt is the world's leading high-quality, comprehensive and freely accessible resource of protein sequence and functional information. [Cite UniProt](#)

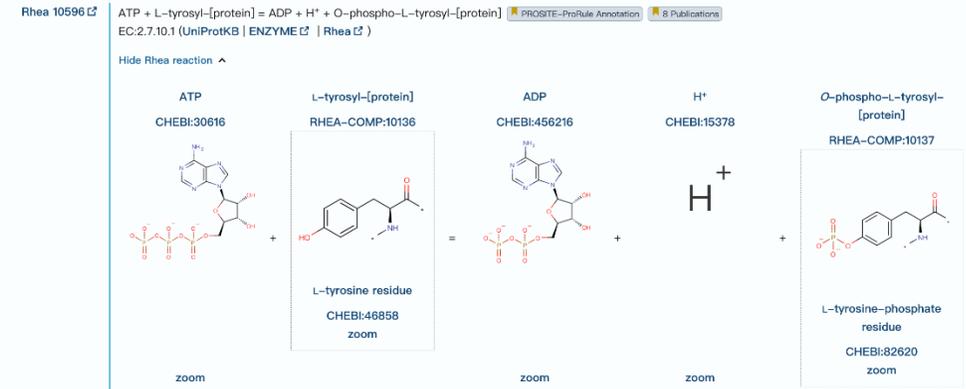
UniProtKB 301,086 results

BLAST Align Map IDs Download Add View: Cards Table Customize columns Share

Entry	Entry Name	Protein Names	Gene Names	Organism	Length
<input type="checkbox"/> P06213	INSR_HUMAN	Insulin receptor[...]	INSR	Homo sapiens (Human)	1,382 AA
<input type="checkbox"/> P14735	IDE_HUMAN	Insulin-degrading enzyme[...]	IDE	Homo sapiens (Human)	1,019 AA
<input type="checkbox"/> P01308	INS_HUMAN	Insulin[...]	INS	Homo sapiens (Human)	110 AA
<input type="checkbox"/> O73727	INS_DANRE	Insulin[...]	ins	Danio rerio (Zebrafish) (Brachydanio rerio)	108 AA
<input type="checkbox"/> P01317	INS_BOVIN	Insulin[...]	INS	Bos taurus (Bovine)	105 AA
<input type="checkbox"/> P01329	INS_CAVPO	Insulin[...]	INS	Cavia porcellus (Guinea pig)	110 AA
<input type="checkbox"/> P17715	INS_OCTDE	Insulin[...]	INS	Octodon degus (Degu) (Sciurus degus)	109 AA
<input type="checkbox"/> P01315	INS_PIG	Insulin[...]	INS	Sus scrofa (Pig)	108 AA
<input type="checkbox"/> P67970	INS_CHICK	Insulin[...]	INS	Gallus gallus (Chicken)	107 AA

UniProt数据库主页与检索页

Catalytic activity¹



Features

Showing features for site¹, binding site¹, active site¹.

Download

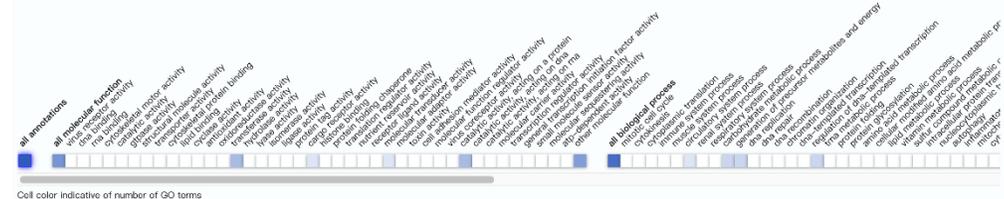
TYPE	ID	POSITION(S)	DESCRIPTION
Site	66		Insulin-binding Curated
Binding site	1033		ATP (UniProtKB ChEBI EC) PROSITE-ProRule Annotation 1 Publication
Binding site	1057		ATP (UniProtKB ChEBI EC) PROSITE-ProRule Annotation 1 Publication
Binding site	1104-1110		ATP (UniProtKB ChEBI EC) PROSITE-ProRule Annotation 1 Publication BLAST Add
Active site	1159		Proton donor/acceptor 1 Publication
Binding site	1163-1164		ATP (UniProtKB ChEBI EC) PROSITE-ProRule Annotation 1 Publication
Binding site	1177		ATP (UniProtKB ChEBI EC) PROSITE-ProRule Annotation 1 Publication

GO annotations¹

Access the complete set of GO annotations on QuickGO [EC](#)

Slimming set:

generic



Pfam

HOME | SEARCH | BROWSE | FTP | HELP | ABOUT

EMBL-EBI 

Pfam data and new releases are available through [InterPro](#)

The Pfam website now serves as a static page with no data updates. All links below redirect to the closest alternative page in the InterPro website.

Pfam 37.0 (21,979 entries, 709 clans)

The Pfam database is a large collection of protein families, each represented by *multiple sequence alignments* and *hidden Markov models (HMMs)*. [More...](#)

QUICK LINKS YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...

- [SEQUENCE SEARCH](#) Analyze your protein sequence for Pfam matches
- [VIEW A PFAM ENTRY](#) View Pfam annotation and alignments
- [VIEW A CLAN](#) See groups of related entries
- [VIEW A SEQUENCE](#) Look at the domain organisation of a protein sequence
- [VIEW A STRUCTURE](#) Find the domains on a PDB structure
- [KEYWORD SEARCH](#) Query Pfam by keywords

JUMP TO

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Family

Proteins share a common evolutionary origin, as reflected in their related functions, sequences or structure

Domain

Distinct functional, structural or sequence units that may exist in a variety of biological contexts

Repeats

Short sequences typically repeated within a protein

Motifs

Short motifs such as metal binding

Binding Site

Coiled-coil

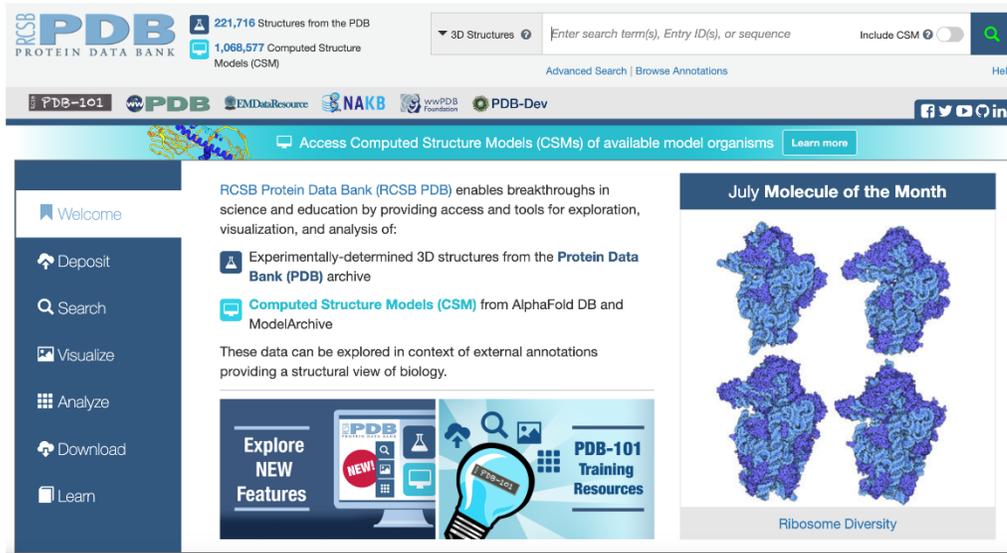
Denoting characteristic heptad repeat

Disordered

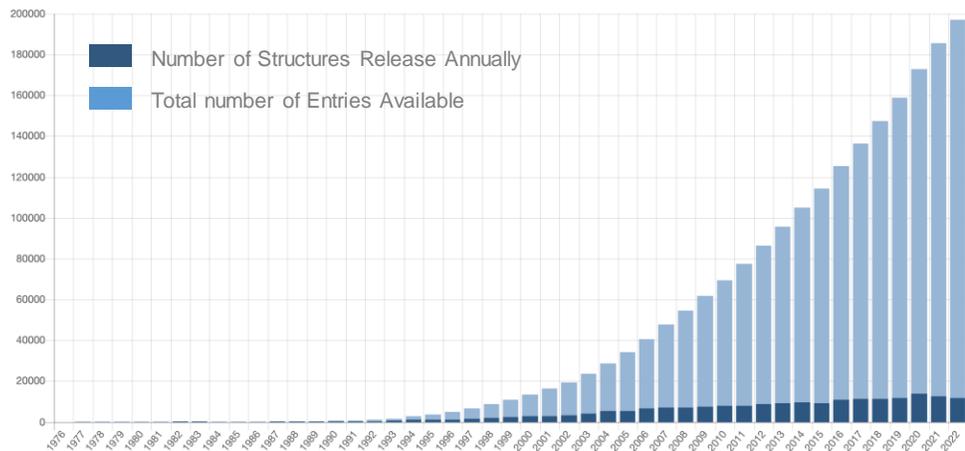
Conserved intrinsically disordered regions

Pfam (已被整合至InterPro数据库中) 数据库主页 (2024.7)

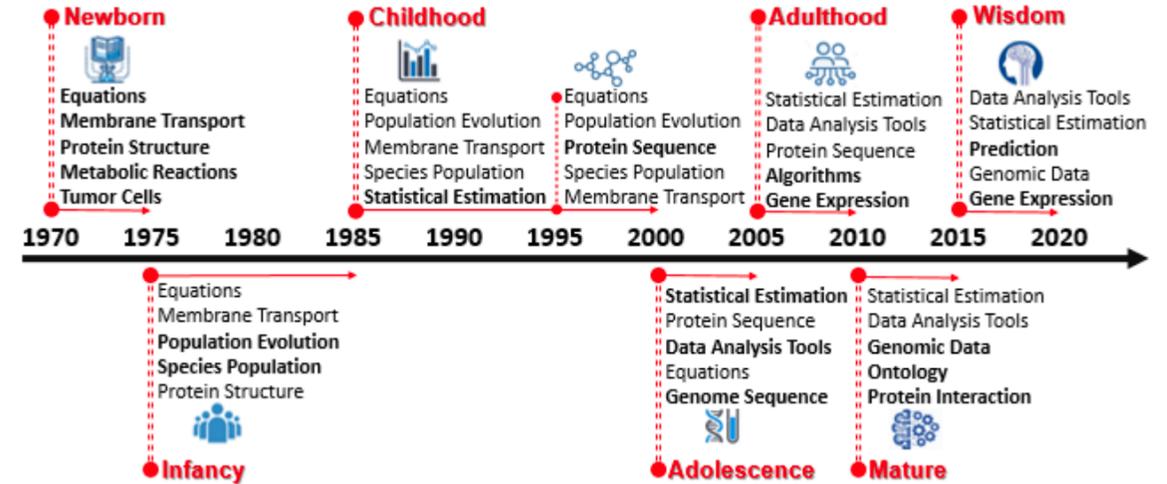
Pfam蛋白质注释内容



PDB数据库首页 (2024.7)



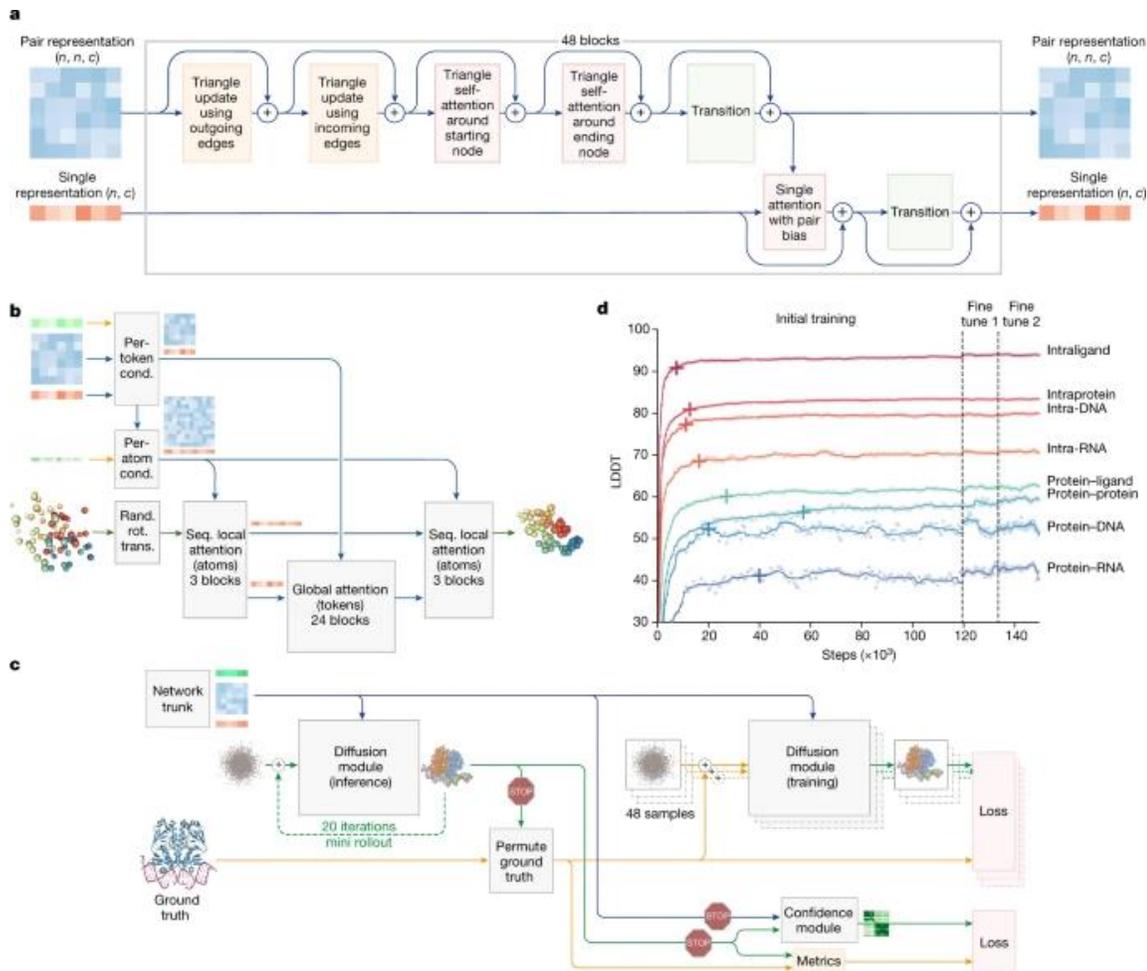
蛋白数据库数据量变化情况



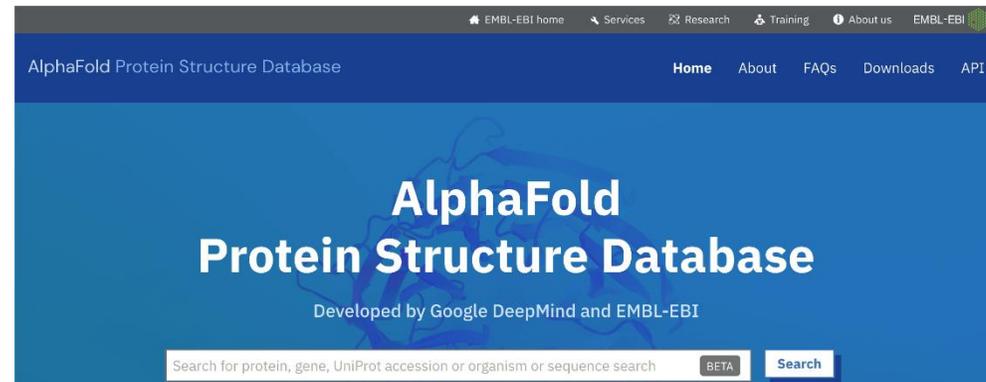
生物信息学七阶段与不同阶段的热点话题

Statistical Estimation 7.59	Ontology	4.56	Membrane Transport	4.04	
Equations	7.50	Prediction	4.41	Genome Sequence	3.93
Data Analysis Tools 6.87	Protein Interaction	4.38	Species Population	3.73	
Algorithms 5.17	Genomic Data	4.32	Signaling Regulatory	3.38	
Protein Sequence	4.99	Protein Structure	4.11	Tumor Cells	3.28
Gene Expression	4.70	Population Evolution	4.07	Infectious Diseases	3.01

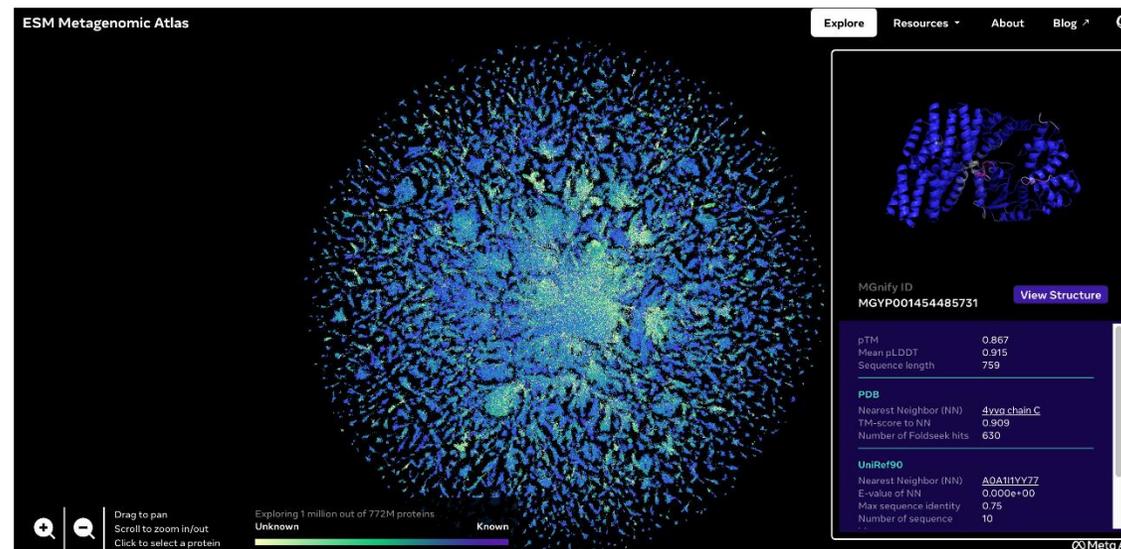
10.1109/ACCESS.2022.3160795 (2022)



蛋白质结构预测AlphaFold3架构与训练方法
Nature (2024)



AlphaFoldDB数据库首页



ESMfoldAtlas数据库展示页

物种	数据库名称	简称	更新时间 (2024年)	网址
人	The Cancer Genome Atlas	TCGA		https://cancergenome.nih.gov
人	Online Mendelian Inheritance in Man	OMIM	Updated August 17, 2023	https://omim.org/
人	Human Protein Atlas	HPA	version 23.0 2023-06-19	http://www.proteinatlas.org/
人	The Cancer Imaging Archive	TCIA		http://www.cancerimagingarchive.net
人	The Human Gene Mutation Database	HGMD	December 2021	https://www.hgmd.cf.ac.uk/
人	ClinVar	ClinVar	Aug 18, 2023	https://www.ncbi.nlm.nih.gov/clinvar/
人	Database of Genotypes and Phenotypes	dbGaP		https://www.ncbi.nlm.nih.gov/gap/
拟南芥	The Arabidopsis Information Resource	TAIR	Araport11 genome release	https://www.arabidopsis.org/
玉米	Maize Genetics and Genomics Database	MaizeGDB	Last update: August 8, 2023	https://maizegdb.org/
豆科	SoyBase	SoyBase	26 Apr 2023	https://soybase.org/
豆科	Legume information system	LIS	9-Aug-23	https://legumeinfo.org/
植物	Planteome	Planteome	Version 5.0 July 2023	https://browser.planteome.org/amigo
果蝇	FlyBase	FlyBase	FB2023_04 August 8, 2023	flybase.org
线虫	WormBase	WormBase	WS289	https://wormbase.org/
斑马鱼	The Zebrafish Information Network	ZFIN	2023	http://zfin.org/
蟾蜍	Xenopus biology and genomics resource	Xenbase	version 6.0.0	https://www.xenbase.org/xenbase/
鼠	Mouse Genome Informatics	MGI	MGI 6.22	https://www.informatics.jax.org/
微生物	Bacterial and Viral Bioinformatics Resource	BV-BRC	3.30.19a	https://www.bv-brc.org/
病毒	Virus pathogen database and analysis resource	ViPR	BV-BRC 3.30.19a	www.ViPRbrc.org
微生物	Integrated Microbial Genomes & Microbiomes	IMG/M		https://img.jgi.doe.gov/
酵母	Saccharomyces Genome Database	SGD	SGD2021-01	https://www.yeastgenome.org/
病原	Global initiative on sharing all influenza data	GISAID		gisaid.org
流感	Influenza Research Database	IRD		https://www.bv-brc.org/
噬菌体	PhagesDB	PhagesDB		http://phagesdb.org/
原核	The Bacterial Diversity Metadatabase	BacDive	20.02.2023	https://bacdive.dsmz.de
原核	Genome Taxonomy Database	GTDB	GTDB Release 214.1	https://gtdb.ecogenomic.org/
原核	EzBioCloud Database	EzTaxon-e	2023.06.29	http://www.eztaxon.org/
真菌	UNITE database	UNITE	version 9.0 2023-08-01	https://unite.ut.ee/
大肠	Encyclopedia of E. coli Genes and Metabolism	EcoCyc	version 27.0 April 12,2023	https://ecocyc.org/

第四节 生物数据库发展趋势

现状

数据库的数据容量爆炸式增长
数据库的数据多维结构差异大
数据库的知识服务创新多样化



挑战

多维数据整合与知识挖掘发现
数据资源交互与信息共建共享
数据隐私安全与知识产权保护



趋势

基于多源异构数据整合的数据存储
基于生物信息知识检索的数据交互
基于智能分析技术的生物信息挖掘



THOUGHT LEADERS

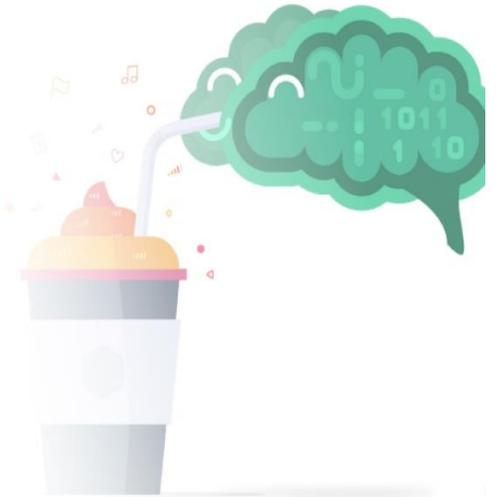
Importance of Data Quality in AI Implementation



Published 7 months ago on September 7, 2022
By Amy Groden-Morrison



Data is to AI
as Food is to
Humans

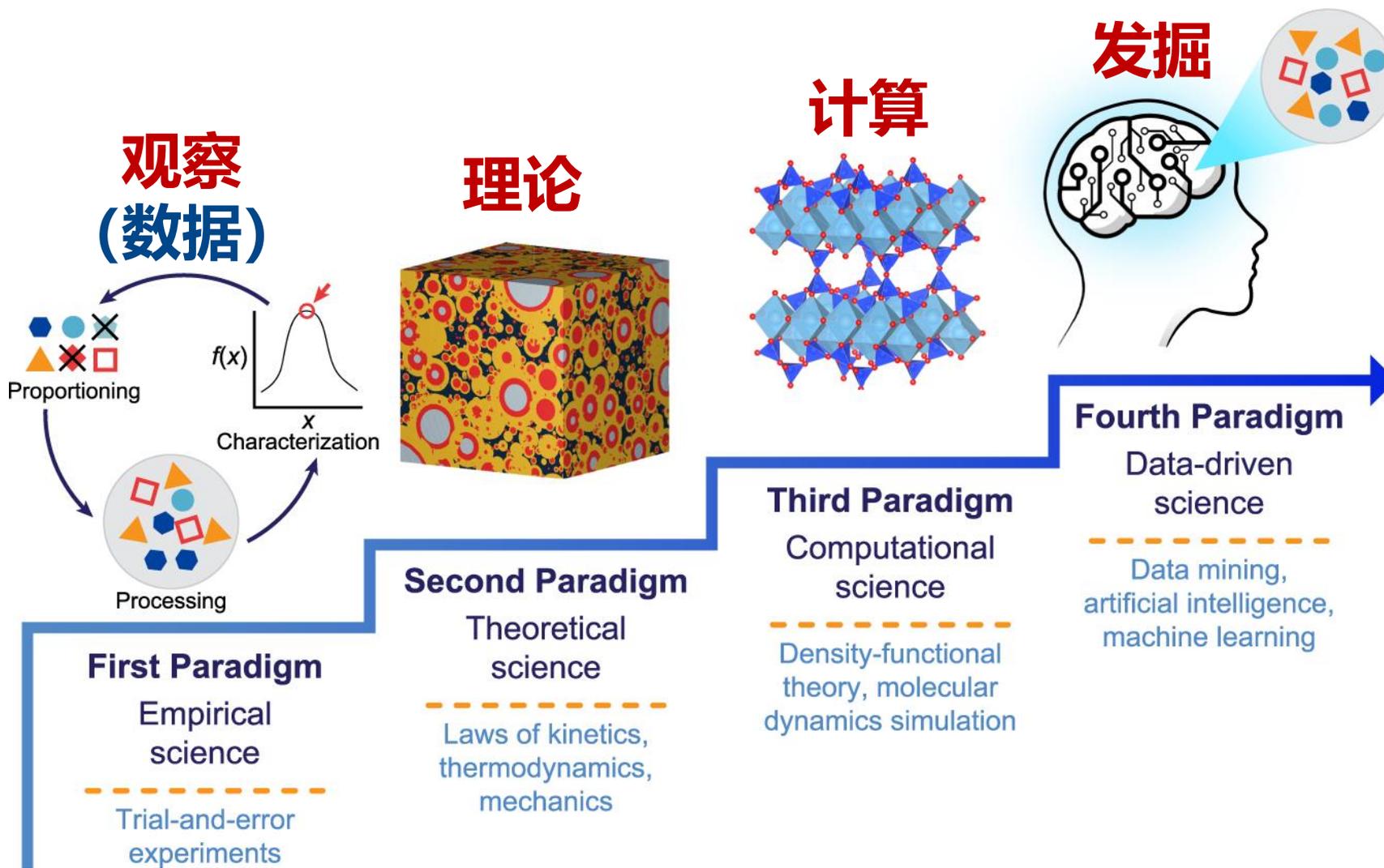


高质量数据：AI-ready语料库

- 数据质量、数据容量
- 完整性、一致性、准确性

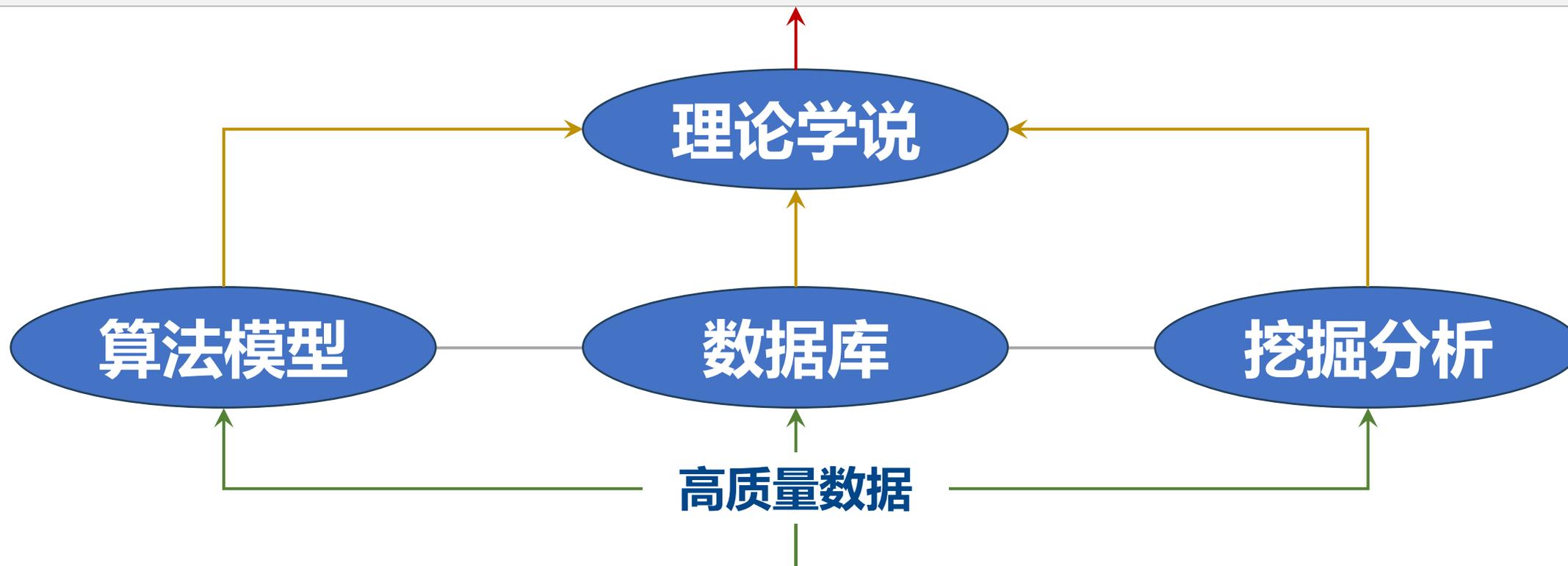
<https://devicology.com>
<https://elplanetaurbano.com>

<https://www.addwebsolution.com/blog/chatgpt-understanding-capabilities-use-cases>
<https://www.unite.ai/importance-of-data-quality-in-ai-implementation/>



进化			
拉马克学说	1809	Jean-Baptiste Lamarck	用进废退与获得性遗传
达尔文学说	1858	Charles Darwin	自然选择
中性进化理论	1968	Motoo Kimura	中性突变
内共生学说	1970	Lynn Margulis	线粒体（细菌）和叶绿体（蓝藻）起源
遗传			
孟德尔遗传定律	1865	Gregor Mendel	分离定律和自由组合定律
摩尔根定律	1911	Thomas Morgan	基因的连锁和交换定律
细胞			
细胞学说	1838、1839	Matthias Schleiden、Theodor Schwann	细胞是一切动植物的基本构造
细胞重建学说	1988	贝时璋	细胞的自组织过程
分子生物学			
四核苷酸假说	1928	Phoebus Levene	DNA组成
查伽夫定律	1952	Erwin Chargaff	$A\%=T\%$ ， $G\%=C\%$
中心法则	1957	Francis Crick	遗传信息流动

Theory: “He who loves practice without theory is like the sailor who boards ship without a rudder and compass and never knows where he may cast.” **知其然而不知其所以然。** —Leonardo da Vinci



Data: “I don't think we can get a Nobel prize by what we are doing, but the Nobel prize winners know what we are doing for.” **功成不必在我，功成必定有我。** —Alan Bleasby

- 生物数据是国家重要**战略资源**
- 生物数据库是国家生物科技发展**基础设施**
- **生物数据安全**：人类遗传资源，科学数据管理
- 国际主要数据中心：**国家意志，国家战略**
 - NCBI: GenBank、RefSeq、PubMed、OMIM等
 - EMBL-EBI: Ensembl、GWAS Catalog、UniProt等
 - CNCB-NGDC: GSA、GWH、GWAS/EWAS Atlas等
- 国际重要生物数据库
 - 生物大分子数据库 (DNA、RNA、Protein) : **PDB**  **AlphaFold**
 - 物种专题数据库
- 数据库发展趋势：**高质量数据，AI语料库，数据到理论**

- 1. 版权声明：**本PPT及其所有内容（以下简称“本PPT”）仅用于教育和教学用途，版权归属于本PPT作者。
- 2. 使用要求：**任何使用本PPT的行为均须遵守以下条件：
 - 1) 致谢和标注：**若部分或全部使用本PPT的内容，请在使用内容的适当位置标注出处，并致谢本PPT作者。
 - 2) 修改和再分发：**未经作者书面许可，不得对本PPT进行修改或再分发。
- 3. 禁止商业化使用：**严禁将本PPT用于任何形式的商业化用途，包括但不限于：
 - 1) 通过网络或其他途径进行付费使用或分发；
 - 2) 在商业培训、广告或其他商业活动中使用本PPT的内容。
- 4. 法律责任：**任何违反上述条款的行为，作者保留追究法律责任的权利，包括但不限于：
 - 1) 要求停止侵权行为；
 - 2) 追究侵权使用者的经济赔偿责任。
- 5. 其他规定：**
 - 1) 本使用条款的解释权归本PPT作者所有。
 - 2) 作者保留随时更新本使用条款的权利，更新后的条款将即时生效。

敬请各位老师批评指正!