第四章　生物信息资源

袁猛，章张，张子丁，胡德华

个人简介

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

内容提纲

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

章节分工

第一节 生物数据库简介

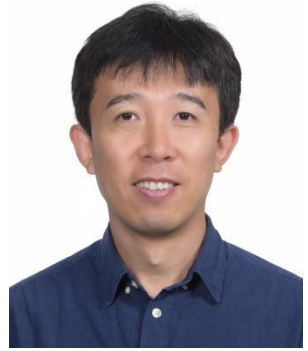
袁猛

华中农业大学

第二节 国际主要数据中心

陈梅丽

陈婷婷



章张

国家生物信息中心

第三节 国际重要生物数据库

史文聿

张子丁

中国农业大学

第四节 生物数据库发展趋势

胡德华

中南大学

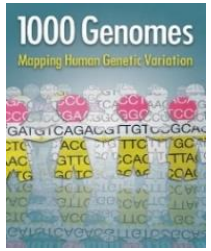
第一节 生物数据库简介

生命科学进入大数据时代

大数据驱动
研究范式转变

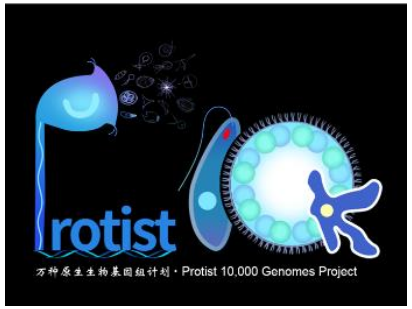
大科学计划
科学数据累积

91 PB
数据量*

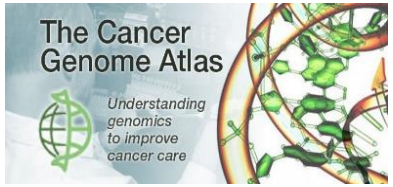


千人基因组计划
2008—2015

人类细胞图谱
2016—

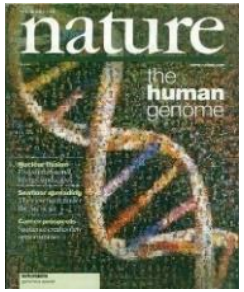


万种原生生物基因组计划
2019—



TCGA
2006-2018

万种鸟类基因组计划
2015—



人类基因组计划
1990—2001

ENCODE
2003—

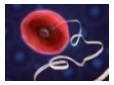
大熊猫基因组计划
2008—2009

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

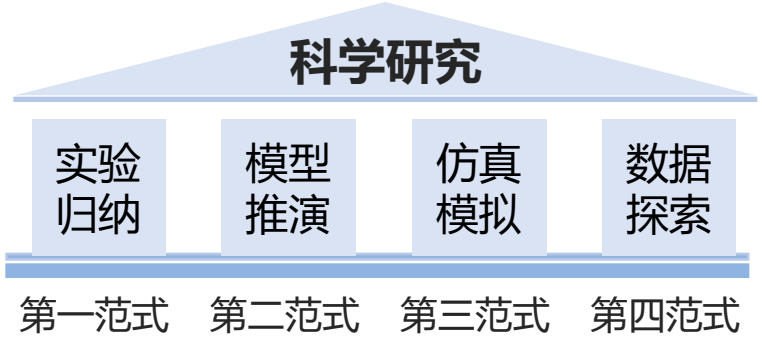
* NCBI SRA (as of July 2024)

生物大数据是国家重要战略资源

驱动生命科学研究范式变革

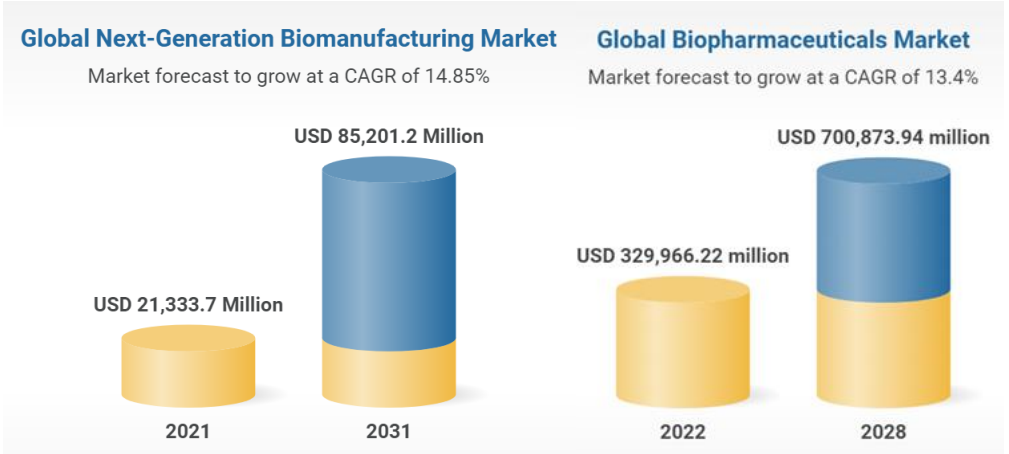
AlphaFold
Nature 2018

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

生物数据库是生物数据的重要载体

生物数据

- 收 集
- 整 理
- 储 存
- 管 理
- 审 编
- 应 用

生物数据库

- 访 问
- 检 索
- 管 理
- 更 新

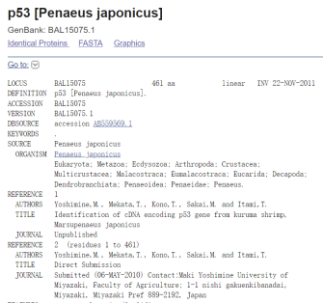
- 国家重要战略资源
- 科技发展基础设施

- 数据信息资源保藏
- 生物科技安全根基

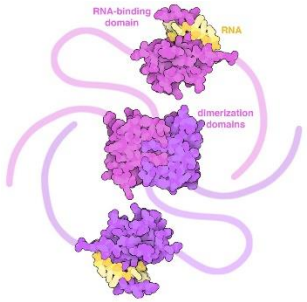
生物数据库特点

生物数据库

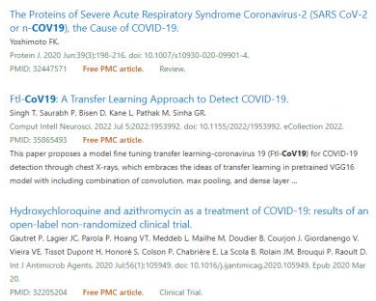
- **种类繁多：**序列、结构、文献、代谢、表达谱、基因组等
- **多层级：**数据库 → 信息库 → 知识库
- **交叉链接：**原始数据，分析结果，文献知识
- **“用户至上”+“数据至上”：**递交-存储-审核-质控-发布-共享



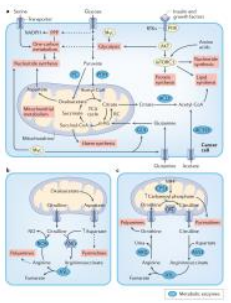
序列



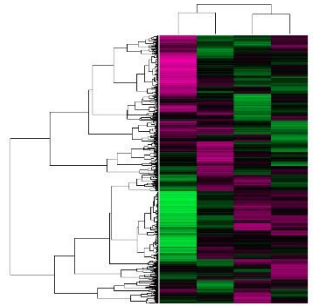
结构



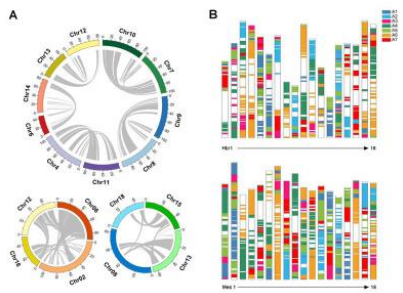
文献



代谢



表达谱



基因组

生物数据库分类

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

生物数据库期刊

- *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*

生物数据库简介

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

国外主要生物信息中心概况

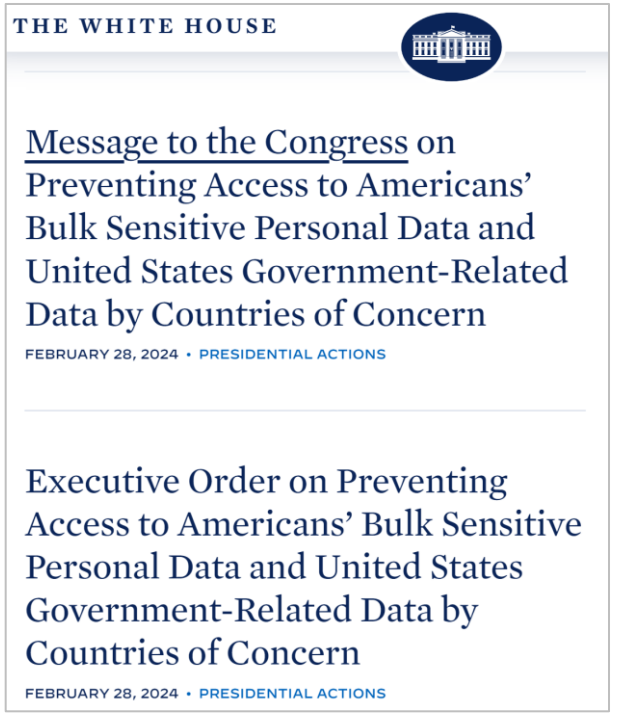
中心	美国国家生物技术信息中心（NCBI）	欧洲生物信息研究所（EBI）	日本DNA数据库（DDBJ）
成立时间	1988年	1994年	1986年
人员规模（人）	~ 700	~ 800	~ 50
年度经费（万美元）	15,375	8,820	834
经费来源	美国政府	英国政府	日本文部省
日访问量（万次）	2100	2700	<10

我国生物信息大数据面临严峻挑战

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

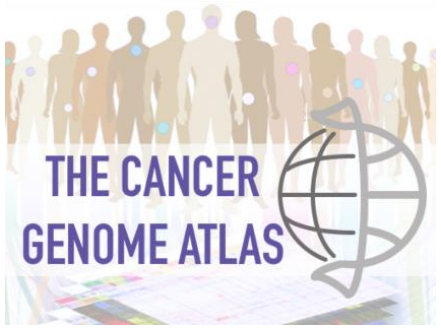

THE WHITE HOUSE

Message to the Congress on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern

FEBRUARY 28, 2024 • PRESIDENTIAL ACTIONS

Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern

FEBRUARY 28, 2024 • PRESIDENTIAL ACTIONS



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

国家生物信息中心建设历程

1999年6月
郝柏林《建议尽快组建国家级的生物医学信息中心》

2014年11月
中科院学部《关于中国“国家生物医学信息中心”的调研报告与建议》

2016年2月
北京基因组所“**生命与健康大数据中心**”正式成立

2017年10月
国家生物信息中心建设方案

2018年8月
国家生物信息中心建设方案

2019年6月
批复成立 **国家基因组科学数据中心**

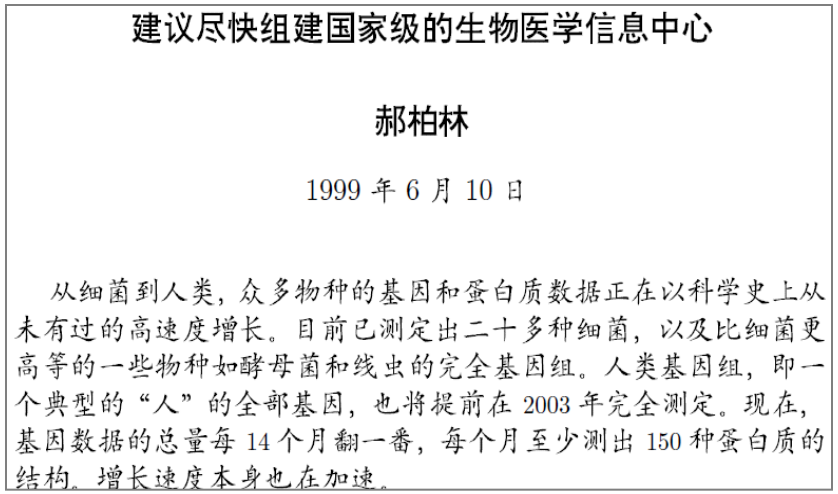
2019年8月
国家生物信息中心建设方案

2019年11月
中编办批复加挂 **国家生物信息中心**

2022年1月
国家发改委批复 **CNCB项目建议书**

2024年4月
CNCB基建开工

1999 ..... 2014 2016 2018 2020 2022 2024

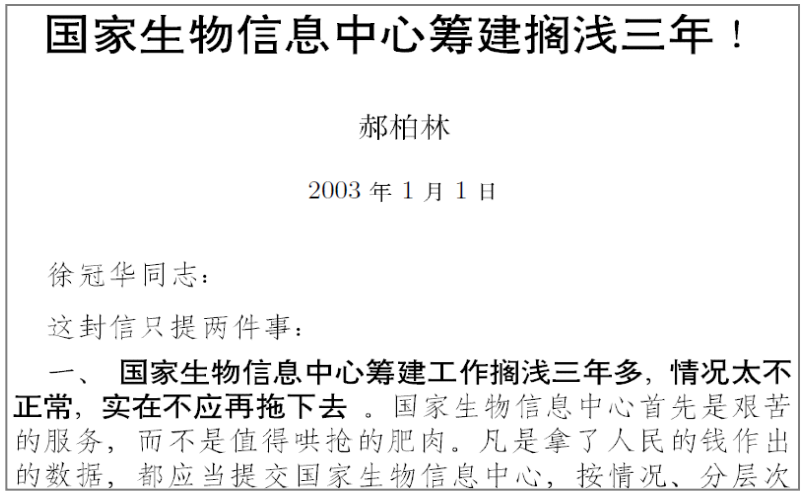


建议尽快组建国家级的生物医学信息中心

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！

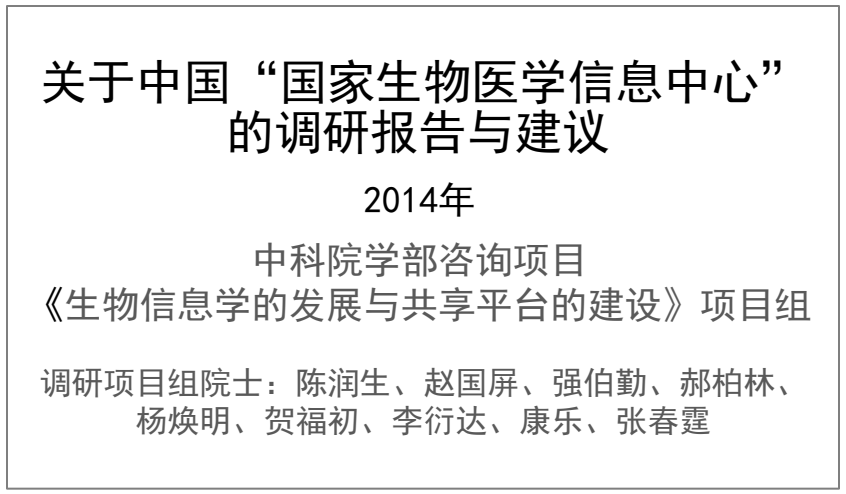
郝柏林

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

科学数据管理

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

中华人民共和国中央人民政府
www.gov.cn
EN

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

生物数据安全是国家安全的重要组成

- 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。

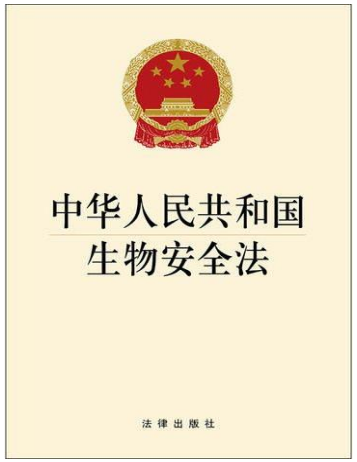
1996年7月《科学》杂志称，国外某大学要在中国开展遗传疾病合作，用到两亿样本，由国外制药公司支持的多个项目已在进行中，包括对600万中国人进行哮喘基因筛选。

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。

生物数据全链条安全管理和开发利用

2018年3月17日

2019年7月1日



2021年4月15日

2021年9月1日

2021年11月1日

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

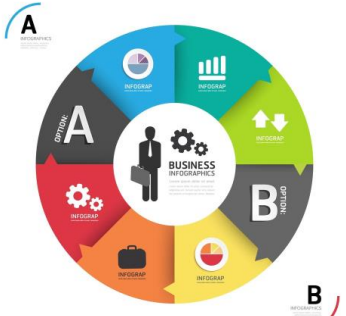
生物数据全链条安全管理和开发利用

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

国际主要数据中心

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

国际核苷酸序列数据库联盟

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

国际期刊对数据递交的要求

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

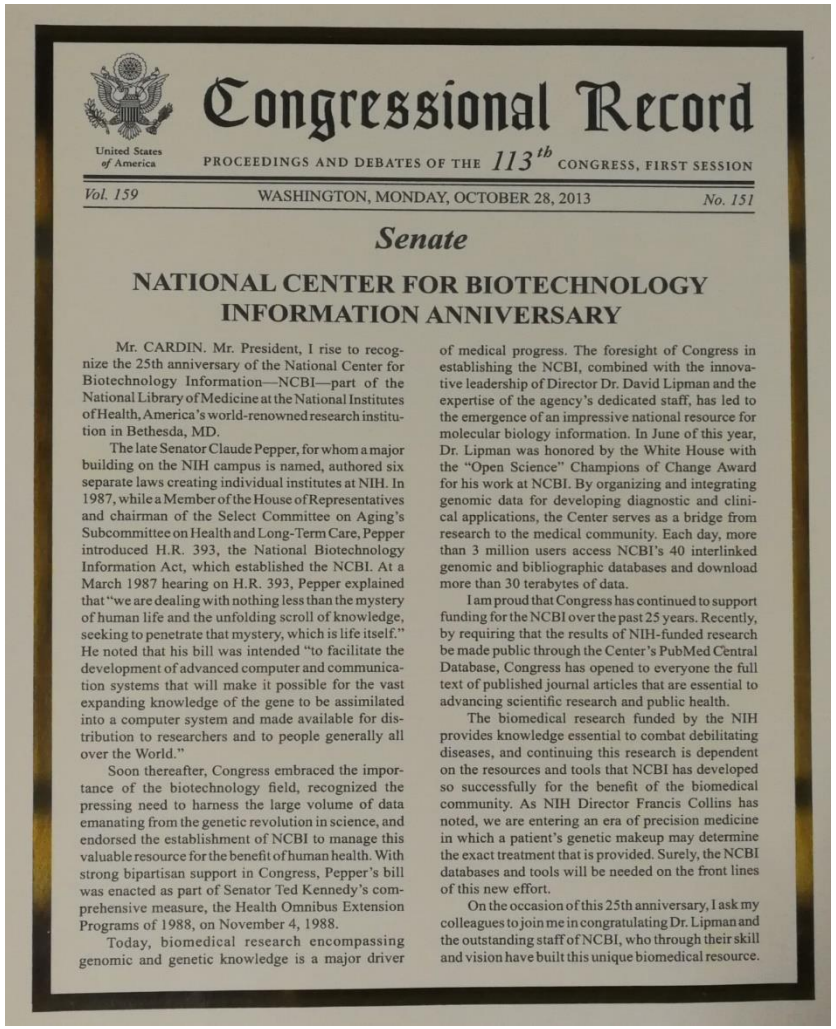
DNA and RNA Sequences: Genbank/European Nucleotide Archive (ENA)/DDBJ, Protein DataBank, UniProt

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.

美国政府给予NCBI大力支持



Congressional Record

United States of America

PROCEEDINGS AND DEBATES OF THE 113th CONGRESS, FIRST SESSION

Vol. 159 WASHINGTON, MONDAY, OCTOBER 28, 2013 No. 151

Senate

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION ANNIVERSARY

Mr. CARDIN. Mr. President, I rise to recognize the 25th anniversary of the National Center for Biotechnology Information—NCBI—part of the National Library of Medicine at the National Institutes of Health, America's world-renowned research institution in Bethesda, MD.

The late Senator Claude Pepper, for whom a major building on the NIH campus is named, authored six separate laws creating individual institutes at NIH. In 1987, while a Member of the House of Representatives and chairman of the Select Committee on Aging's Subcommittee on Health and Long-Term Care, Pepper introduced H.R. 393, the National Biotechnology Information Act, which established the NCBI. At a March 1987 hearing on H.R. 393, Pepper explained that "we are dealing with nothing less than the mystery of human life and the unfolding scroll of knowledge, seeking to penetrate that mystery, which is life itself." He noted that his bill was intended "to facilitate the development of advanced computer and communication systems that will make it possible for the vast expanding knowledge of the gene to be assimilated into a computer system and made available for distribution to researchers and to people generally all over the World."

Soon thereafter, Congress embraced the importance of the biotechnology field, recognized the pressing need to harness the large volume of data emanating from the genetic revolution in science, and endorsed the establishment of NCBI to manage this valuable resource for the benefit of human health. With strong bipartisan support in Congress, Pepper's bill was enacted as part of Senator Ted Kennedy's comprehensive measure, the Health Omnibus Extension Programs of 1988, on November 4, 1988.

Today, biomedical research encompassing genomic and genetic knowledge is a major driver of medical progress. The foresight of Congress in establishing the NCBI, combined with the innovative leadership of Director Dr. David Lipman and the expertise of the agency's dedicated staff, has led to the emergence of an impressive national resource for molecular biology information. In June of this year, Dr. Lipman was honored by the White House with the "Open Science" Champions of Change Award for his work at NCBI. By organizing and integrating genomic data for developing diagnostic and clinical applications, the Center serves as a bridge from research to the medical community. Each day, more than 3 million users access NCBI's 40 interlinked genomic and bibliographic databases and download more than 30 terabytes of data.

I am proud that Congress has continued to support funding for the NCBI over the past 25 years. Recently, by requiring that the results of NIH-funded research be made public through the Center's PubMed Central Database, Congress has opened to everyone the full text of published journal articles that are essential to advancing scientific research and public health.

The biomedical research funded by the NIH provides knowledge essential to combat debilitating diseases, and continuing this research is dependent on the resources and tools that NCBI has developed so successfully for the benefit of the biomedical community. As NIH Director Francis Collins has noted, we are entering an era of precision medicine in which a patient's genetic makeup may determine the exact treatment that is provided. Surely, the NCBI databases and tools will be needed on the front lines of this new effort.

On the occasion of this 25th anniversary, I ask my colleagues to join me in congratulating Dr. Lipman and the outstanding staff of NCBI, who through their skill and vision have built this unique biomedical resource.

1988年，美国国会立法成立NCBI

NCBI News

NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION
National Library of Medicine
National Institutes of Health

August 1997

Vice President Launches PubMed, Lauds Free MEDLINE Access

"MEDLINE...will henceforth be available free to the American people." With those words, Vice President Al Gore inaugurated the PubMed search system at a Capitol Hill press conference on June 26. PubMed, which provides Web access to the National Library of Medicine's (NLM) database of the biomedical journal literature, MEDLINE, was heralded by Senator Tom Harkin (IA) as "...the model of a smart, creative government initiative." The Vice President viewed free access to MEDLINE as consistent with the Clinton administration's other "empowerment" initiatives stating, "This development...may do more to reform and improve the quality of health care in the United States than anything else we've done in a long time."

Searching PubMed

PubMed grew out of NCBI's Entrez project which, since 1992, has offered a subset of MEDLINE records related to molecular biology. In addition to encompassing all of MEDLINE and PreMEDLINE, PubMed retains Entrez's ability to use one article as a "seed" to find other similar articles. By traversing the **See Related Articles'** links, a user can find articles similar in concept with speed and precision. PubMed expands upon Entrez by linking MEDLINE articles to full-text Web sites maintained by publishers. Currently, 95 journals are linked to PubMed, including Cell, Journal of Biological Chemistry, Journal of Cell Biology, New England Journal of Medicine, and Science. Access to publishers' Web sites may require subscriptions or registration.

PubMed Options

PubMed offers the option to search MEDLINE or any of NCBI's molecular biology databases. Users can select from a variety of search fields, including but not limited to: text words, author names, and journal titles. A MEDLINE citation for which there is a corresponding online, full-text article will have a button at the top of the abstract page that links to the publisher's Web site. Additional links point to

Continued on page 2

NCBI Director David Lipman (far left) coaches Vice President Gore (seated) as he searches PubMed. NIH Director Harold Varmus (center) and NLM Director Donald Lindberg (far right) look on.

In This Issue

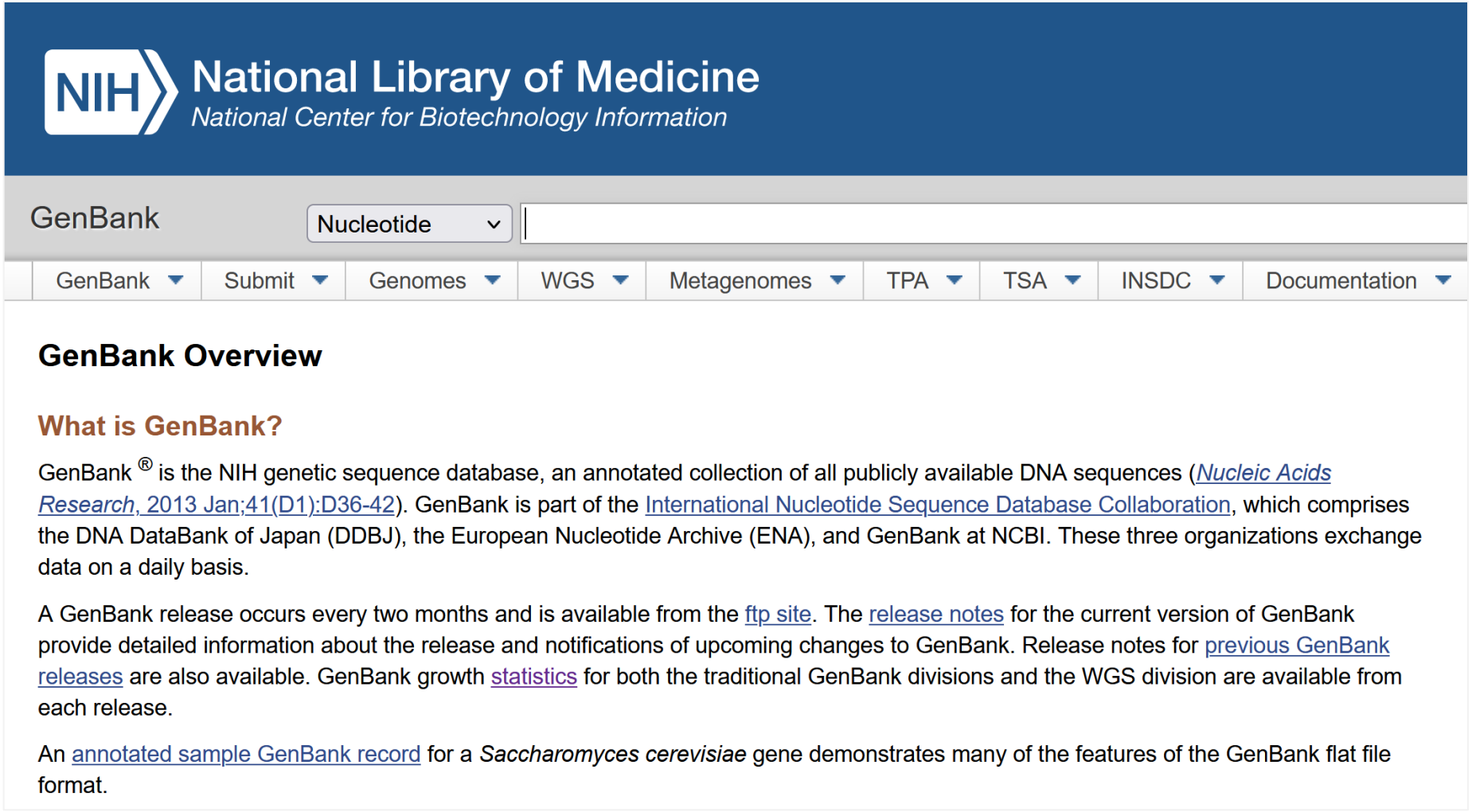
1997年，美国副总统戈尔启动PubMed

NCBI代表性资源

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 dbGap dbVar	变异数据库	1999 2006 2009	https://www.ncbi.nlm.nih.gov/snp/ https://www.ncbi.nlm.nih.gov/gap/ 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/

基因序列数据库GenBank



- **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

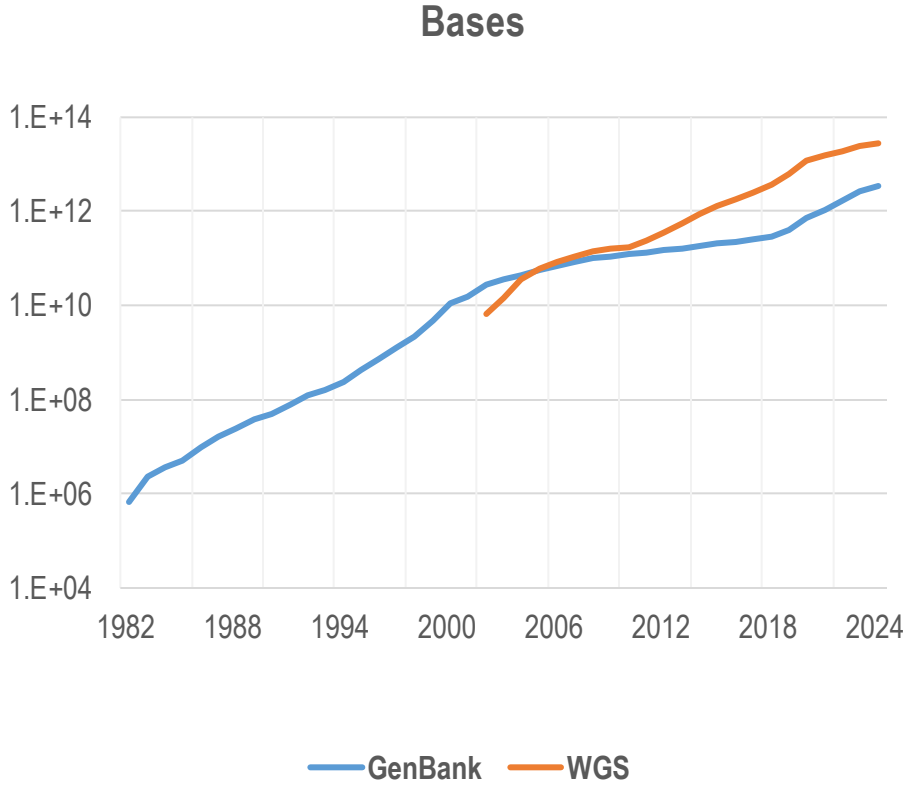
GenBank收集注释所有公开可利用的核苷酸序列和蛋白质序列，是全球最有影响力的生物领域数据库之一

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

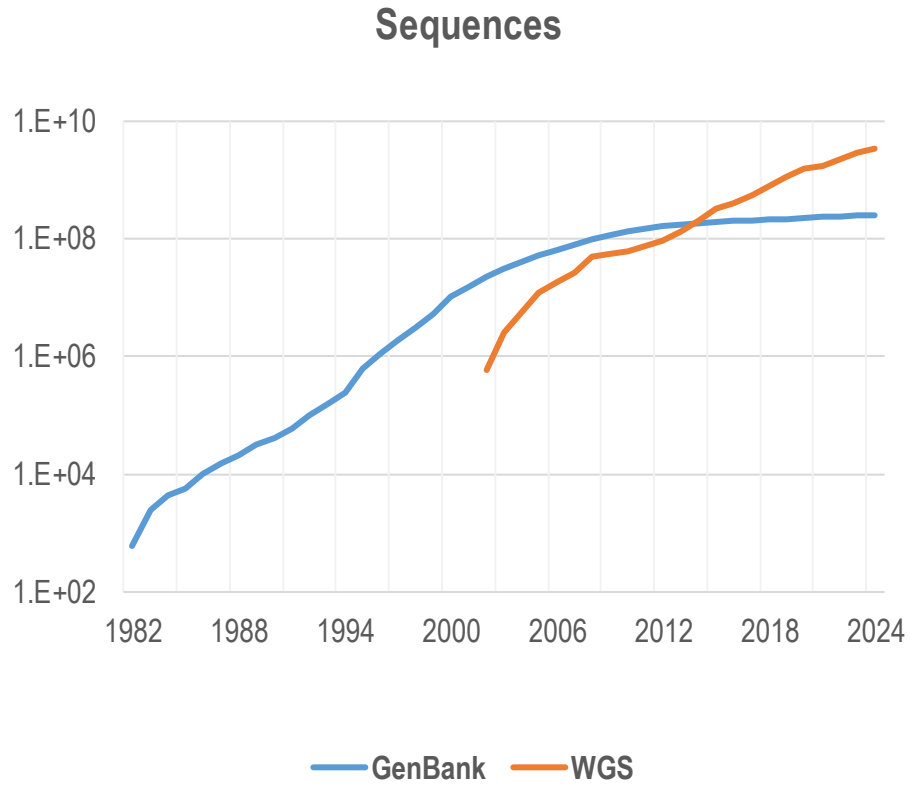
GenBank数据增长

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

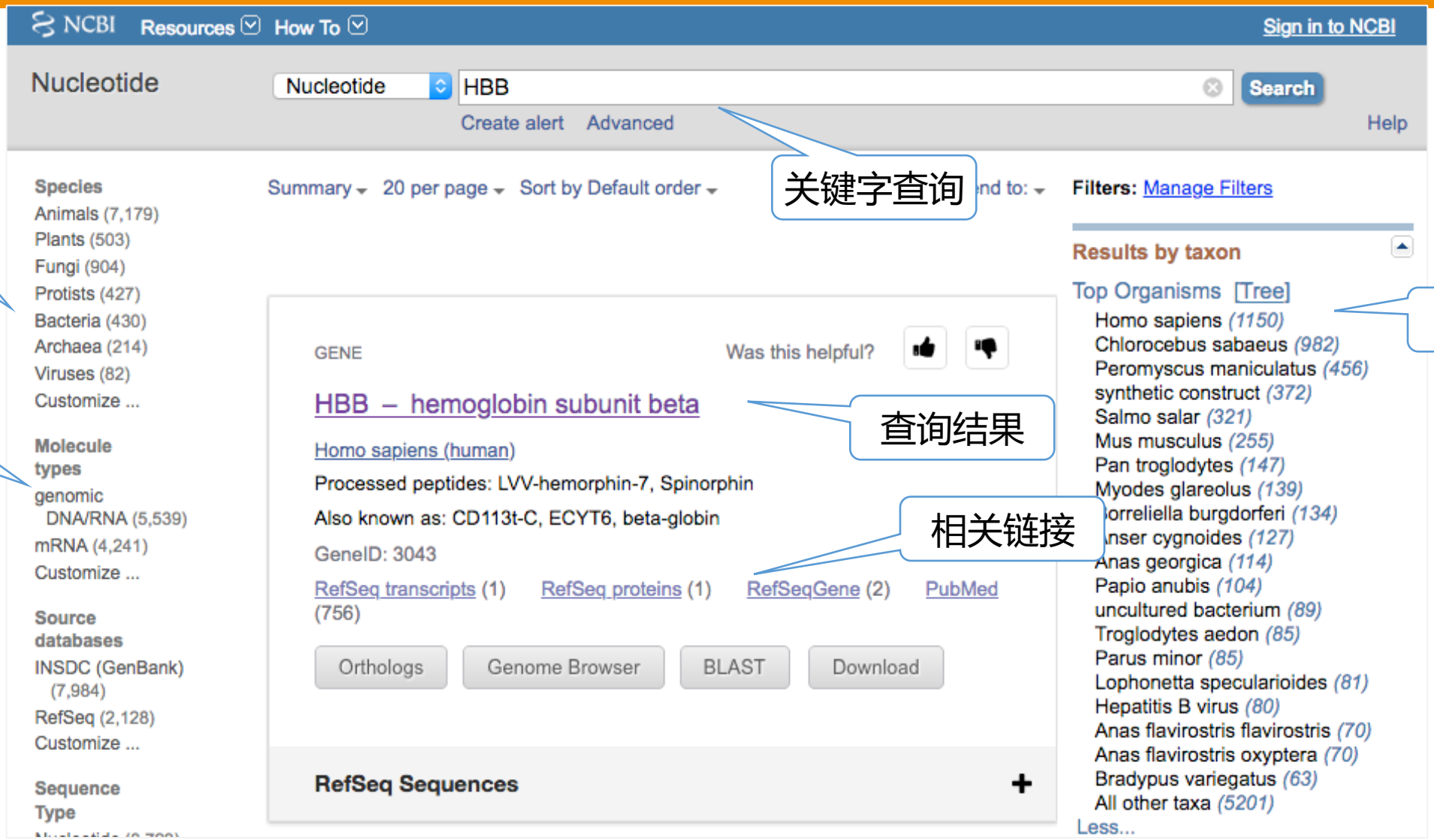


GenBank：~2.16亿条
WGS：~12.07亿条



https://www.ncbi.nlm.nih.gov/genbank/statistics/

GenBank数据库搜索



过滤选择

数据类型

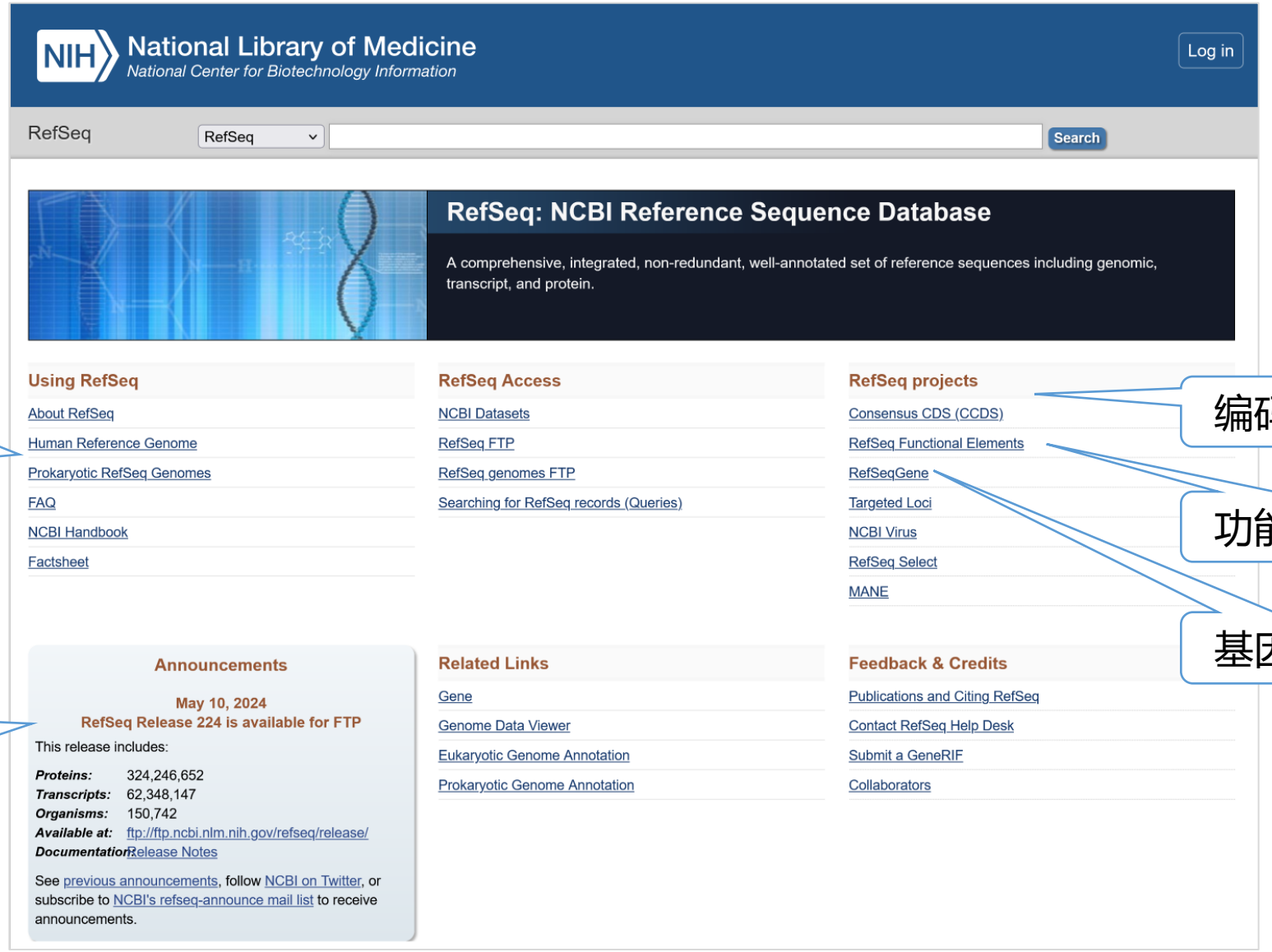
关键字查询

查询结果

相关链接

物种过滤

参考序列数据库RefSeq



人类参考基因组

数据统计信息

编码序列

功能元件

基因集合

https://www.ncbi.nlm.nih.gov/refseq/

参考序列数据库RefSeq

NCBI Resources How To

Nucleotide Nucleotide Advanced

GenBank

Homo sapiens hemoglobin subunit beta (HBB), mRNA

NCBI Reference Sequence: NM_000518.5

FASTA Graphics

RefSeq编号

```
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

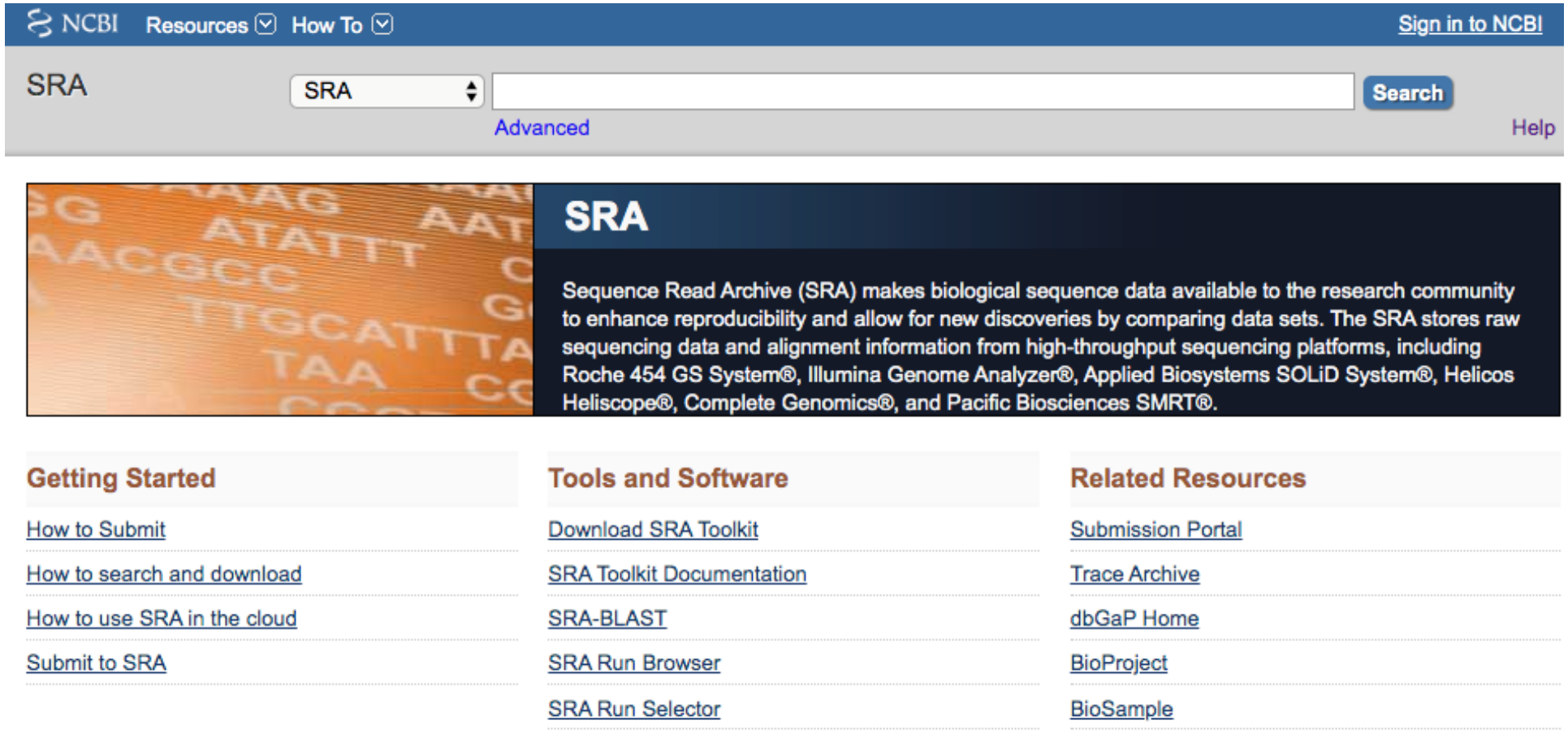
```
                     UniProtKB/Swiss-Prot (P68871.2); other site"
     misc_feature    483..488
                     /gene="HBB"
                     /gene_synonym="beta-globin; CD113t-C; ECYT6"
                     /experiment="experimental evidence, no additional details
                     recorded"
                     /note="(Microbial infection) Cleavage, by N.americanus
                     apr-2. {ECO:0000269|PubMed:12552433}; propagated from
                     UniProtKB/Swiss-Prot (P68871.2); other site"
     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"
     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"
     exon            143..365
                     /gene="HBB"
                     /gene_synonym="beta-globin; CD113t-C; ECYT6"
                     /inference="alignment:Splign:2.1.0"
     exon            366..628
                     /gene="HBB"
                     /gene_synonym="beta-globin; CD113t-C; ECYT6"
                     /inference="alignment:Splign:2.1.0"
ORIGIN
        1 acatttgctt ctgacacaac tgtgttcact agcaacctca aacagacacc atggtgcatc
       61 tgactcctga ggagaagtct gccgttactg ccctgtgggg caaggtgaac gtggatgaag
      121 ttggtggtga ggccctgggc aggctgctgg tggtctaccc ttggacccag aggttctttg
      181 agtcctttgg ggatctgtcc actcctgatg ctgttatggg caaccctaag gtgaaggctc
      241 atggcaagaa agtgctcggt gcctttagtg atggcctggc tcacctggac aacctcaagg
      301 gcacctttgc cacactgagt gagctgcact gtgacaagct gcacgtggat cctgagaact
      361 tcaggctcct gggcaacgtg ctggtctgtg tgctggccca tcactttggc aaagaattca
      421 ccccaccagt gcaggctgcc tatcagaaag tggtggctgg tgtggctaat gccctggccc
      481 acaagtatca ctaagctcgc tttcttgctg tccaatttct attaaaggtt cctttgttcc
      541 ctaagtccaa ctactaaact gggggatatt atgaagggcc ttgagcatct ggattctgcc
      601 taataaaaaa catttatttt cattgcaa
//
```

基因结构信息

序列信息

https://www.ncbi.nlm.nih.gov/nuccore/NM_000518.5

序列读段归档库Sequence Read Archive



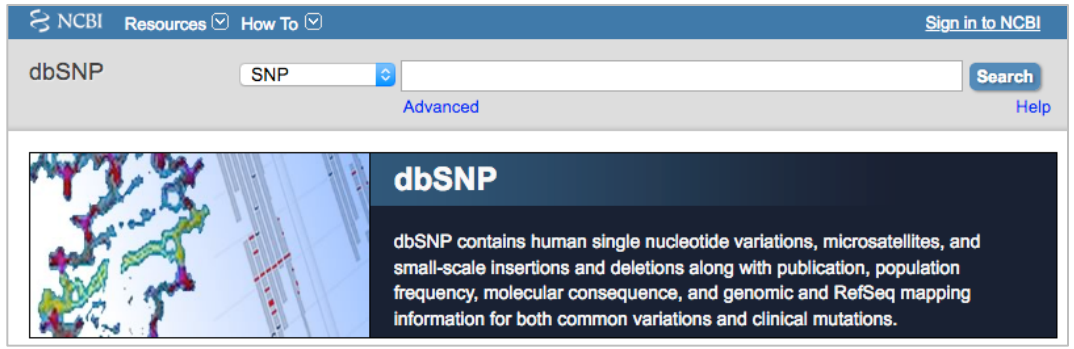
高通量测序原始数据提交共享与数据编号获取

学界共识
文章投稿

截止2024年7月
数据总量91PB
1P=1000T，1T=1000G

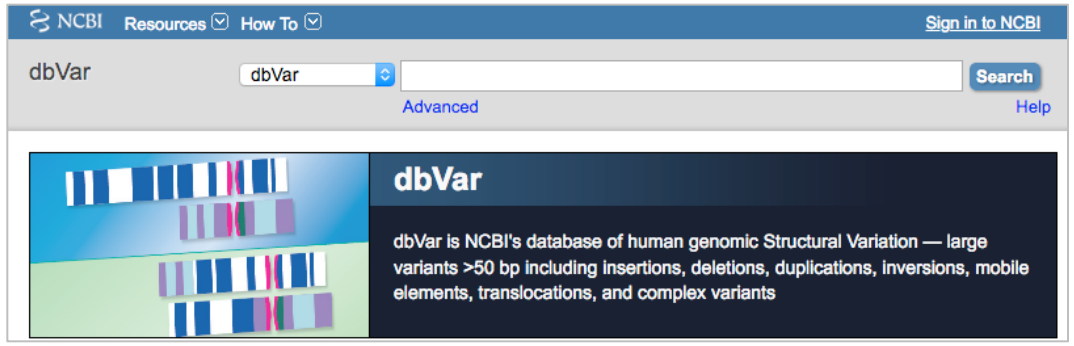
https://www.ncbi.nlm.nih.gov/sra/docs/sragrowth/

变异数据库：dbSNP、dbVar、dbGap



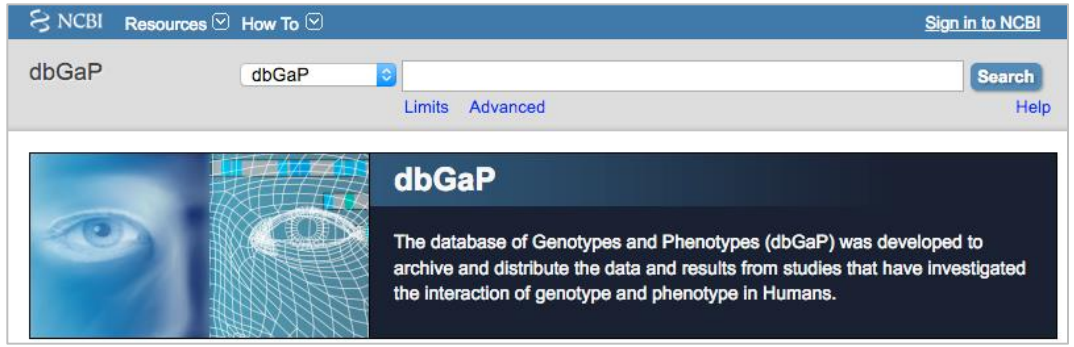
dbSNP：单核苷酸变异、微卫星、小插入或删除

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



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：基因变异与表型关联信息

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

dbSNP：HBB基因上的SNP突变

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）

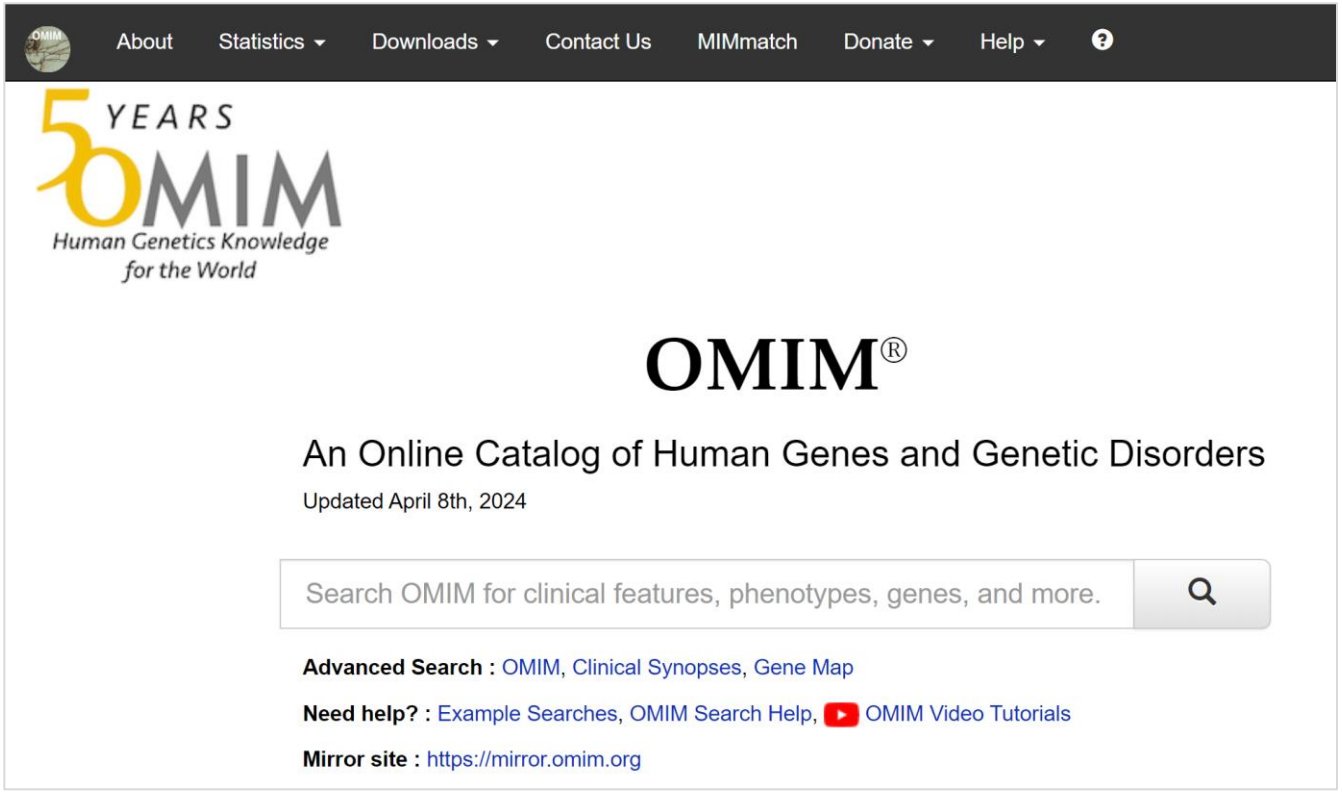
Gene: HBB, hemoglobin subunit beta (minus strand)

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

在线人类孟德尔遗传数据库OMIM



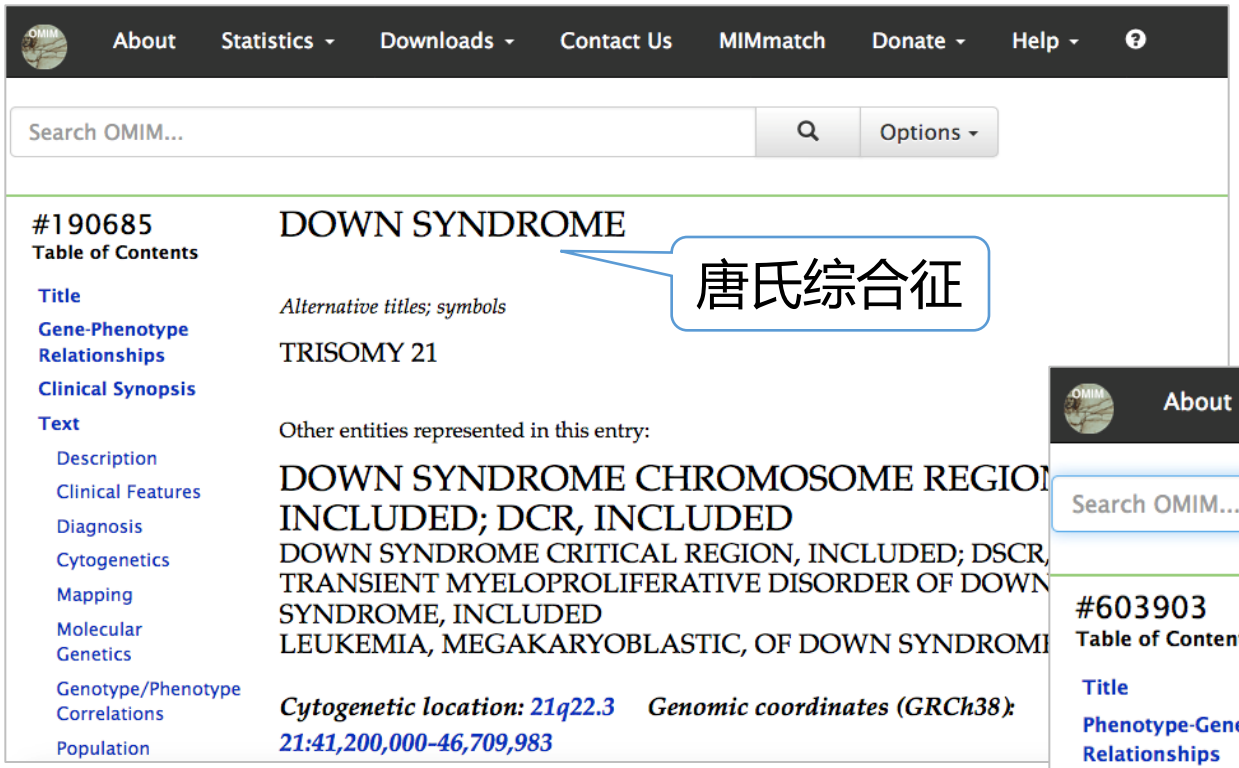
人类基因和遗传疾病数据库，是医学遗传学最权威的百科全书，被誉为医学遗传学界的“《圣经》”，包括所有已知的遗传病、遗传决定的性状及其基因，以及各种疾病的临床特征、诊断、鉴别诊断、治疗与预防等

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

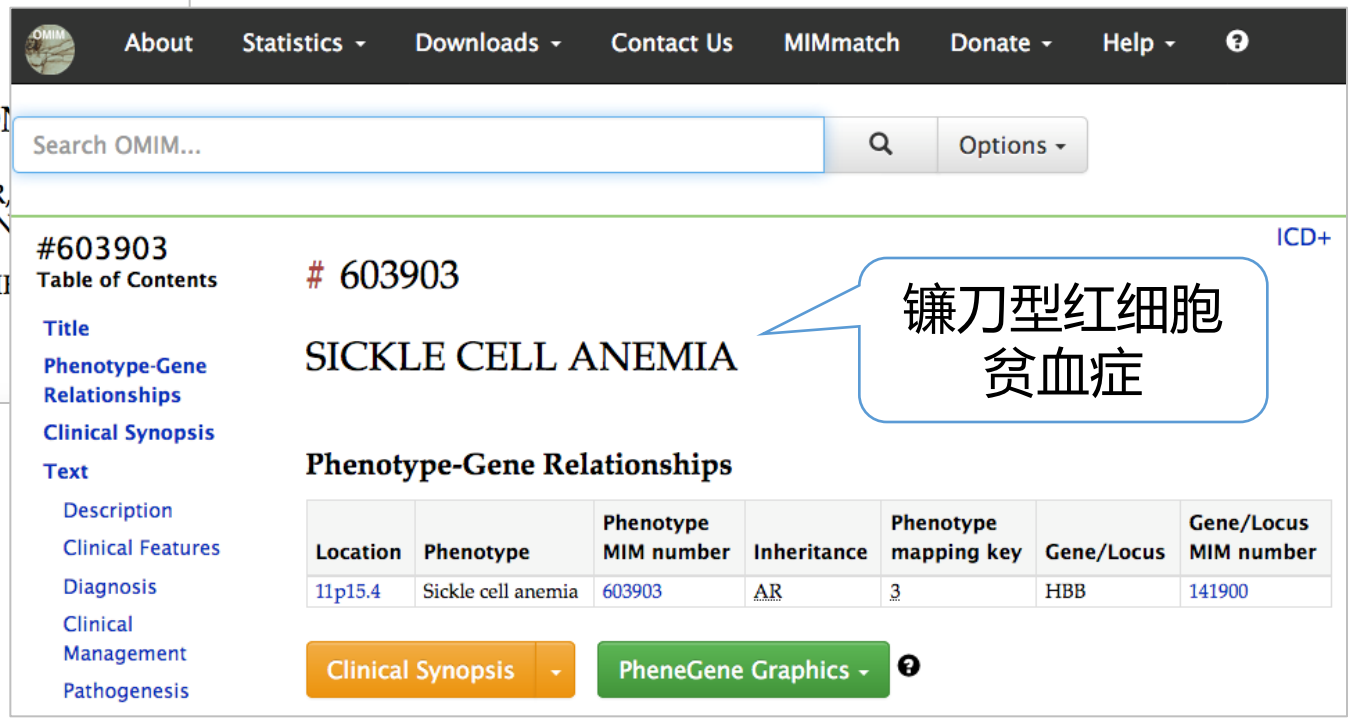
27338条记录数

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

OMIM数据库检索



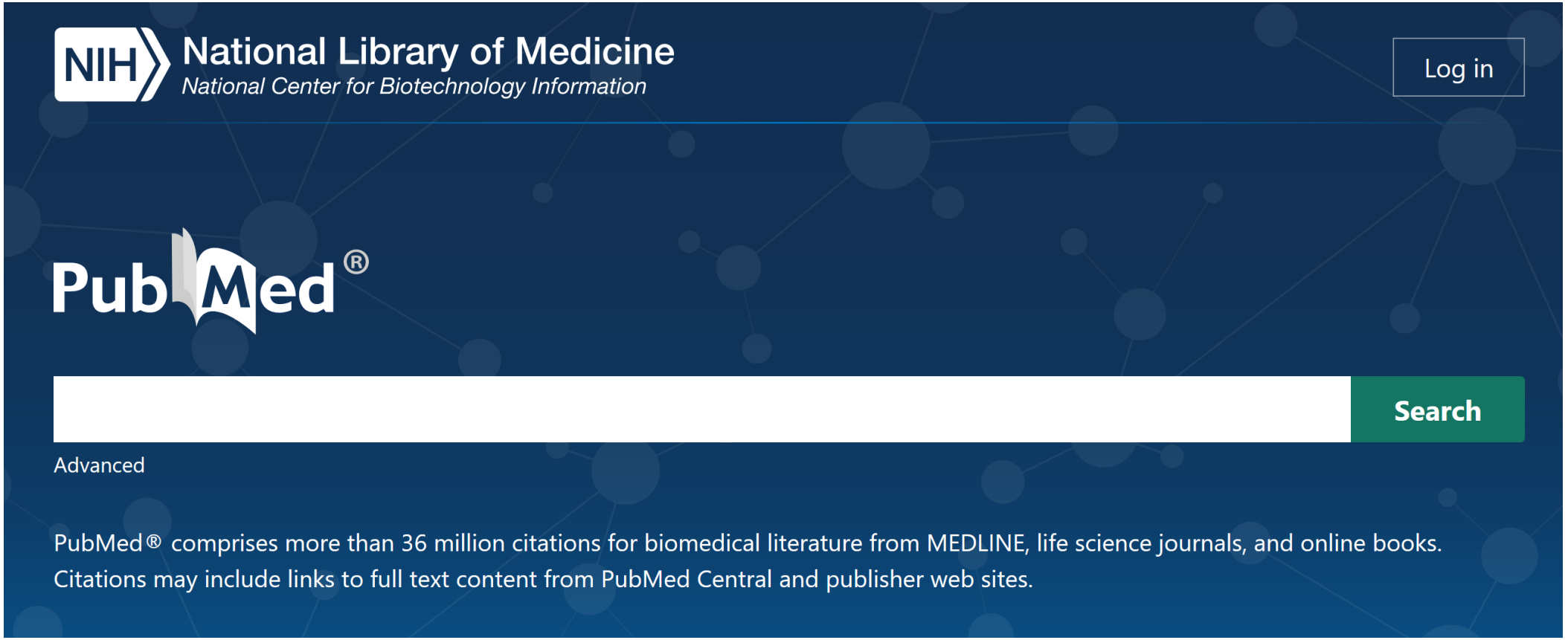
记录编号规则：
1开头，染色体显性遗传
2开头，染色体隐性遗传
3开头，X连锁； 4开头，Y连锁
5开头，线粒体；6开头，染色体



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

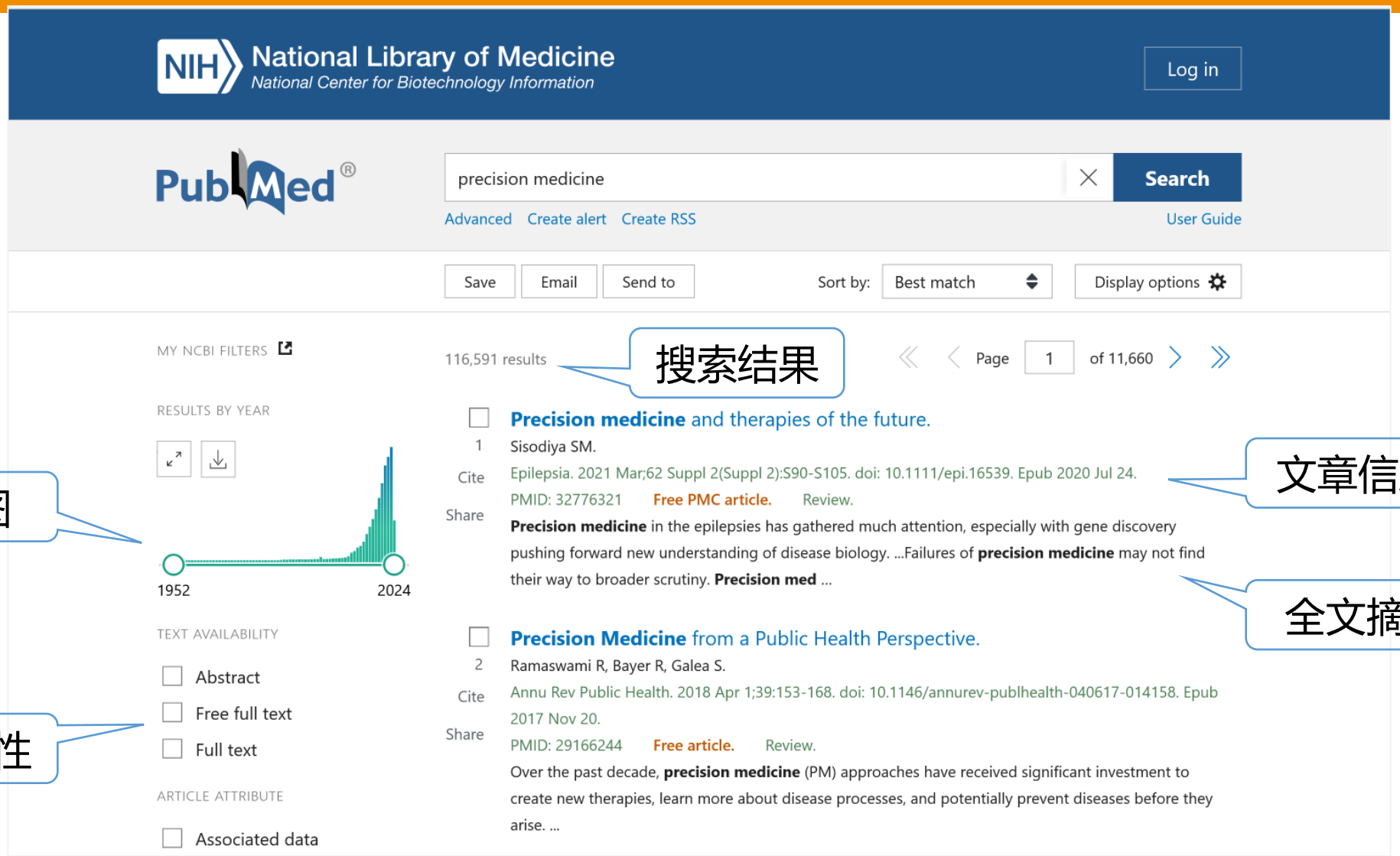
https://www.omim.org/entry/603903

生物医学文献库PubMed/PMC



- 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

PubMed关键字搜索



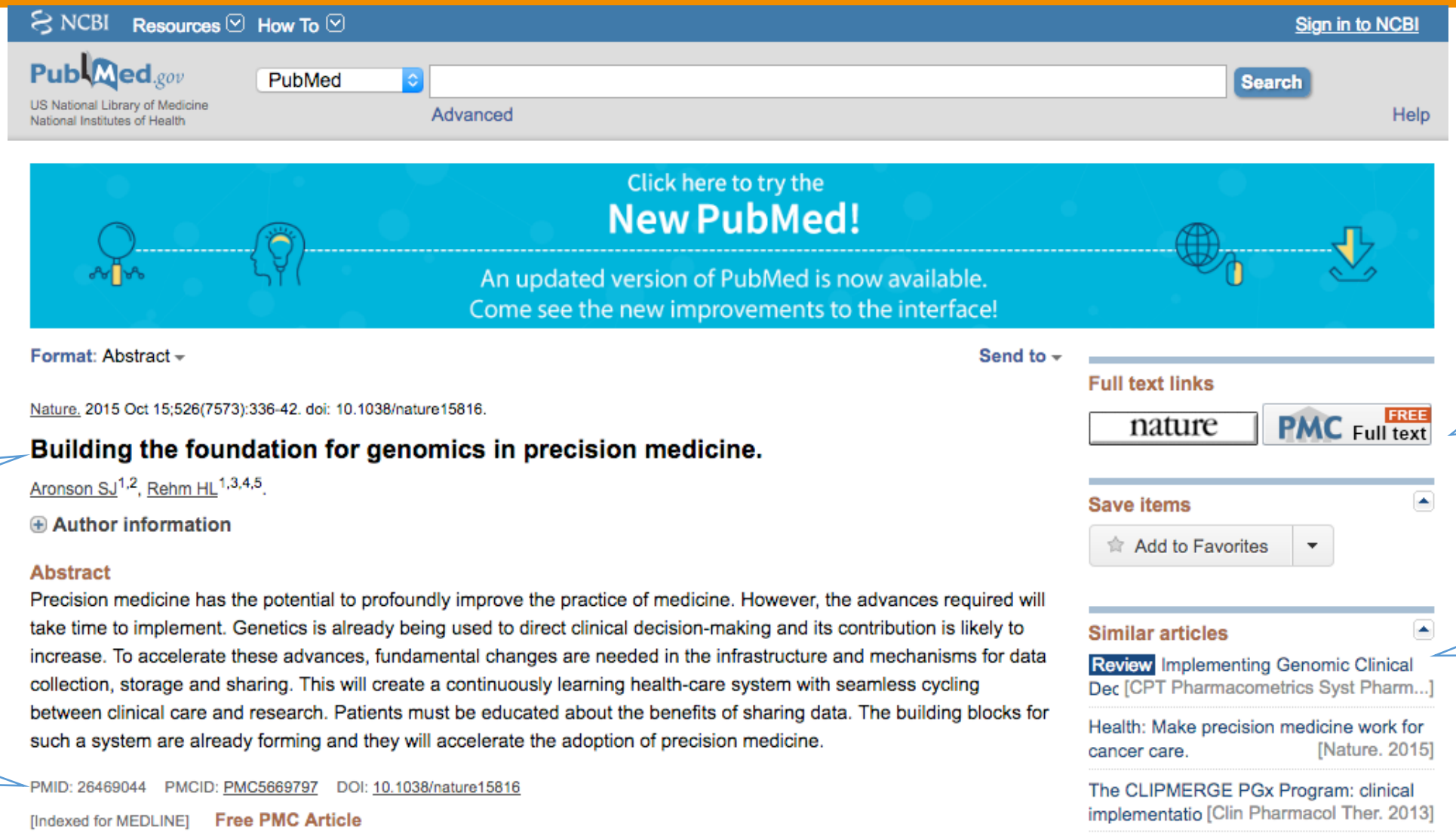
趋势图

论文属性

文章信息及其PMID

全文摘要

PubMed学术论文信息



论文信息

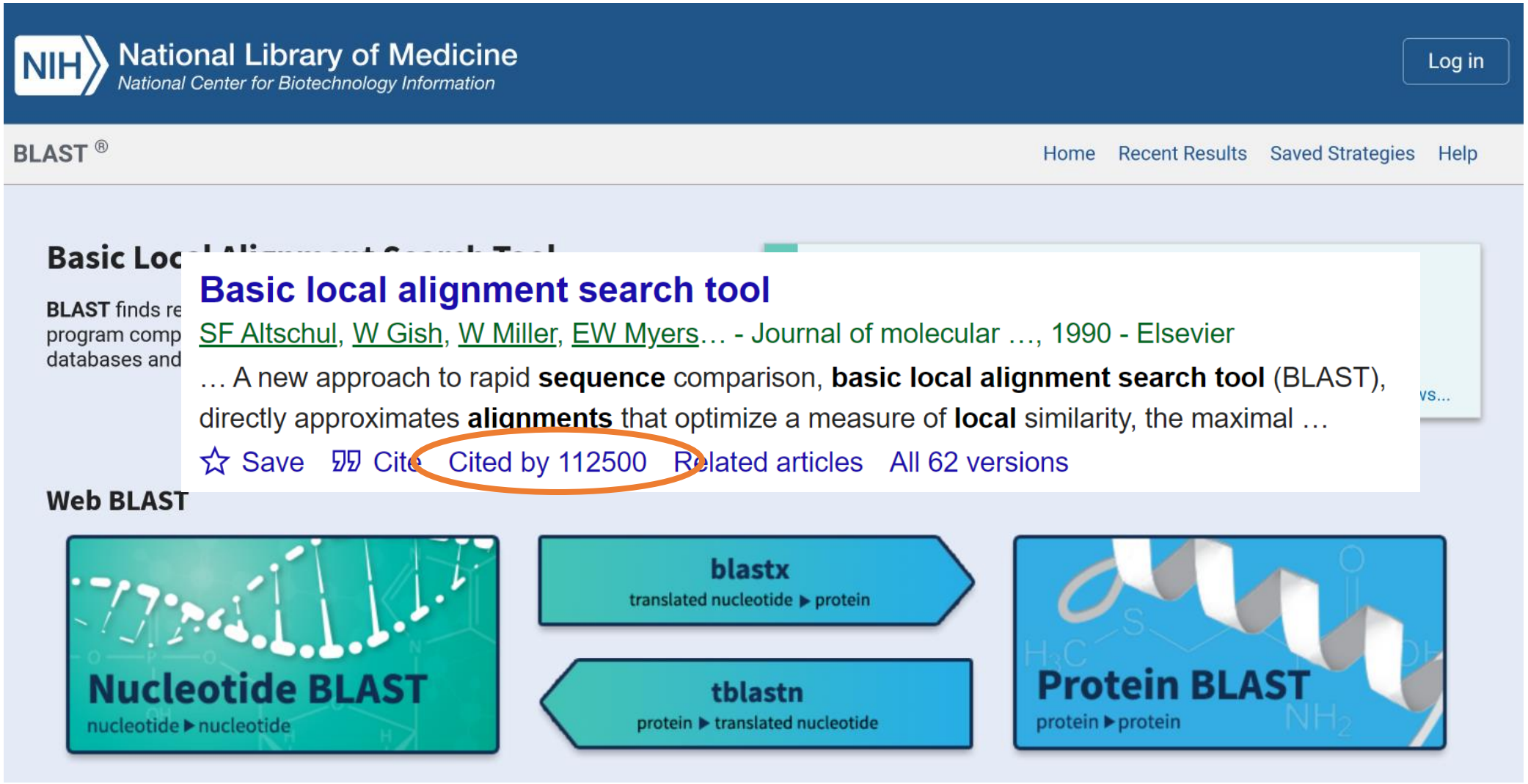
全文链接

相关论文

PMID

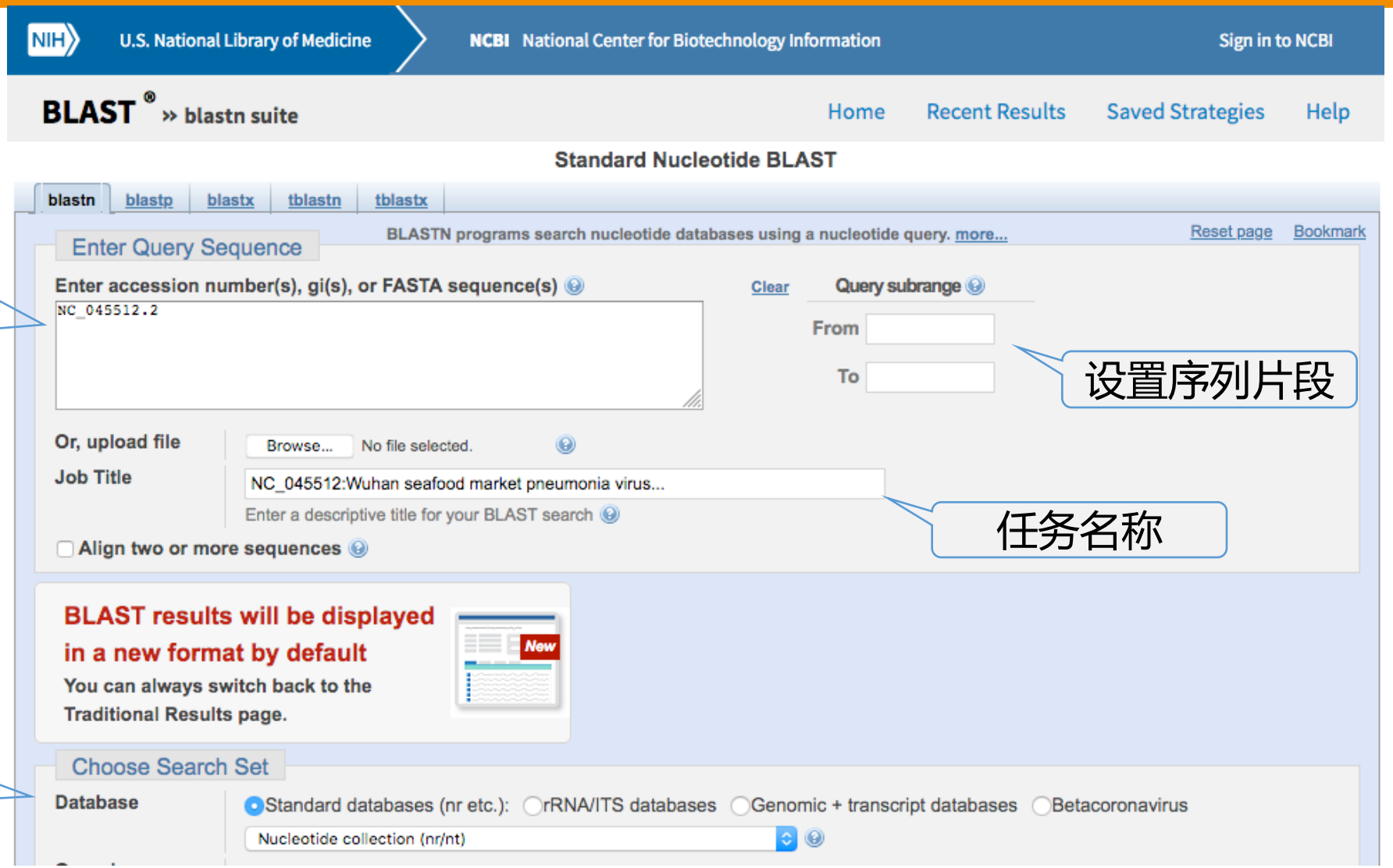
https://www.ncbi.nlm.nih.gov/pubmed/26469044

在线序列比对分析工具BLAST



https://blast.ncbi.nlm.nih.gov

BLAST在线比对



输入序列或序列ID号

设置序列片段

任务名称

比对库

新型冠状病毒基因组 (NC_045512.2)

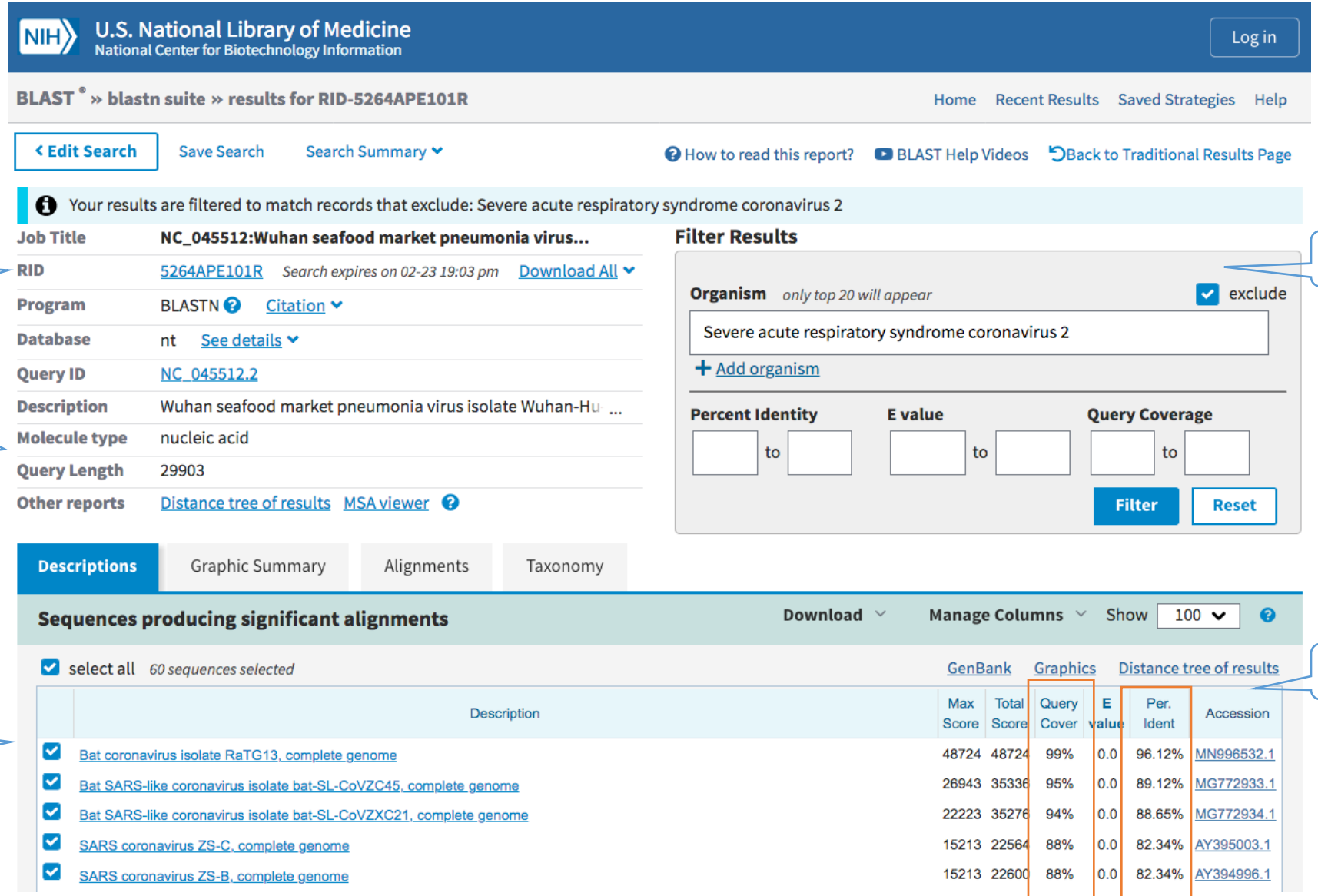
RefSeq编号

基因组长度
病毒类型

BLAST分析

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

BLAST比对结果



任务ID

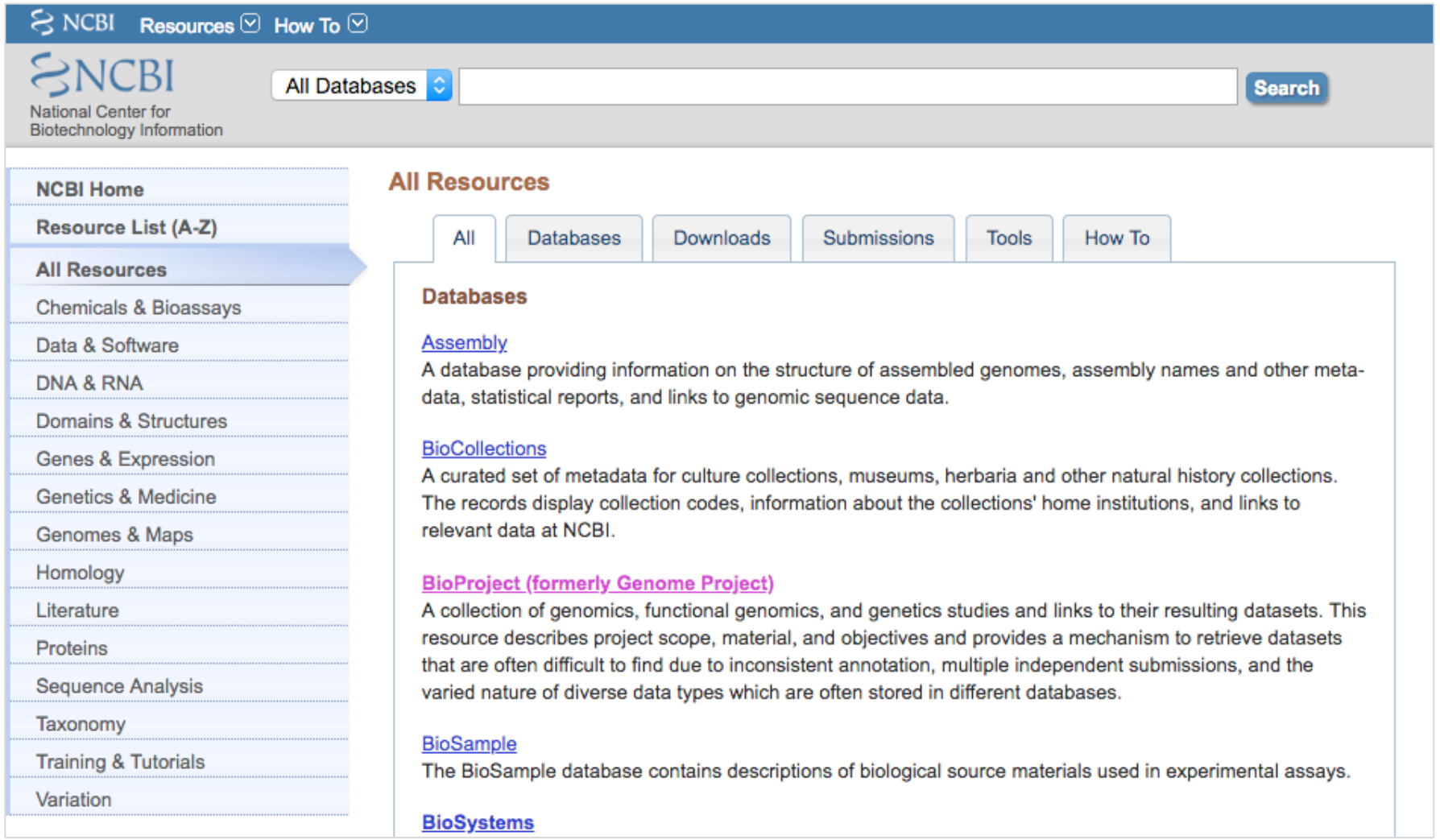
序列类型和长度

比对结果

过滤条件

比对参数

NCBI数据资源总览



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

第四章 生物储库与生物信息学研究所EBI

EBI代表性资源

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

基因组数据库Ensembl

http://www.ensembl.org

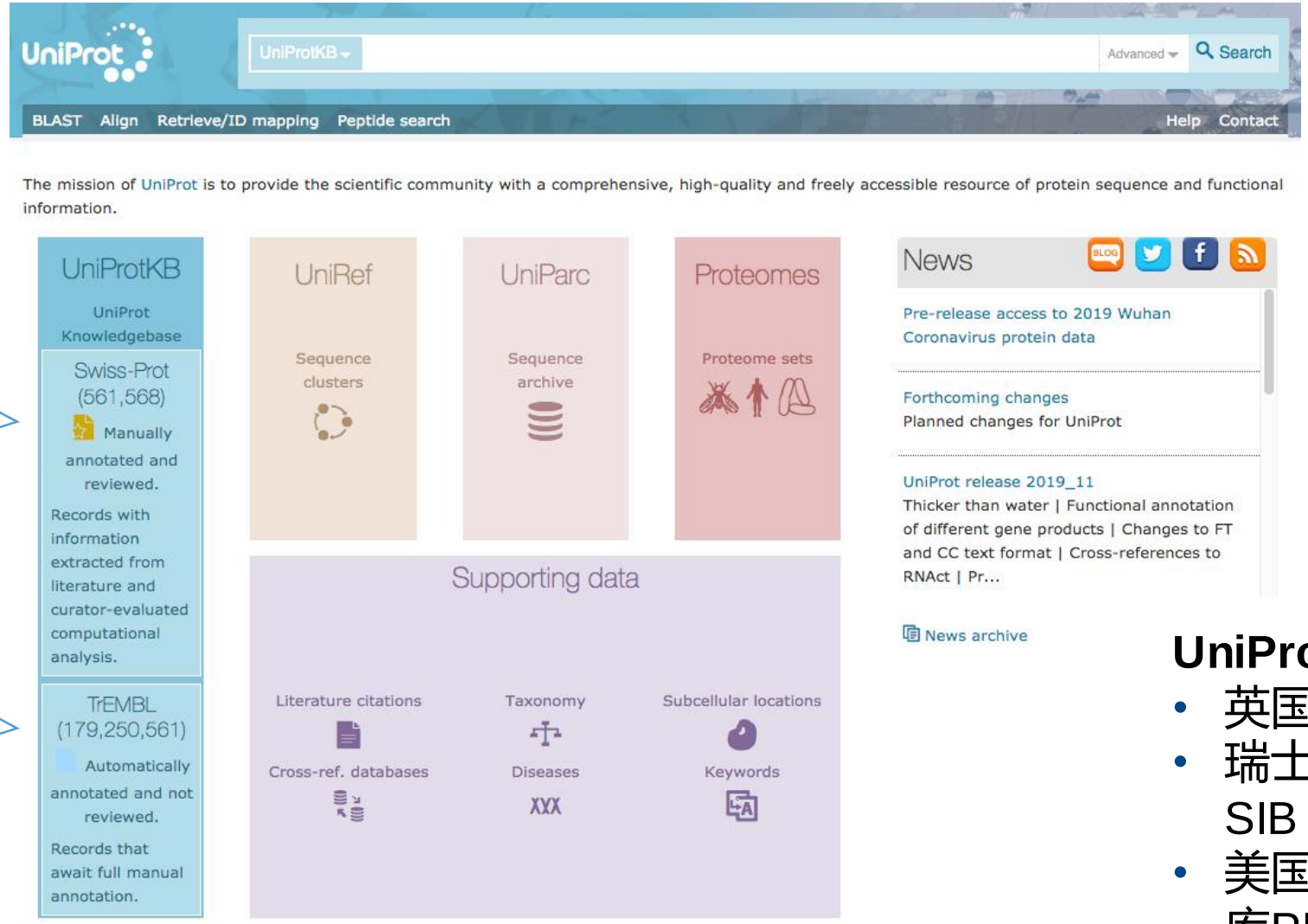
基因组数据库Ensembl

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 *Homo sapiens*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
Y	Mouse *Mus musculus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
Y	Zebrafish *Danio rerio*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
	Abingdon island giant tortoise *Chelonoidis abingdonii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
	African ostrich *Struthio camelus australis*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
	Agassiz's desert tortoise *Gopherus agassizii*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
	Algerian mouse *Mus spretus*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF
	Alpaca *Vicugna pacos*	FASTA	FASTA	FASTA	FASTA	FASTA	EMBL	GenBank	GTF GFF3	TSV RDF JSON	MySQL	GVF

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

蛋白质数据库UniProt



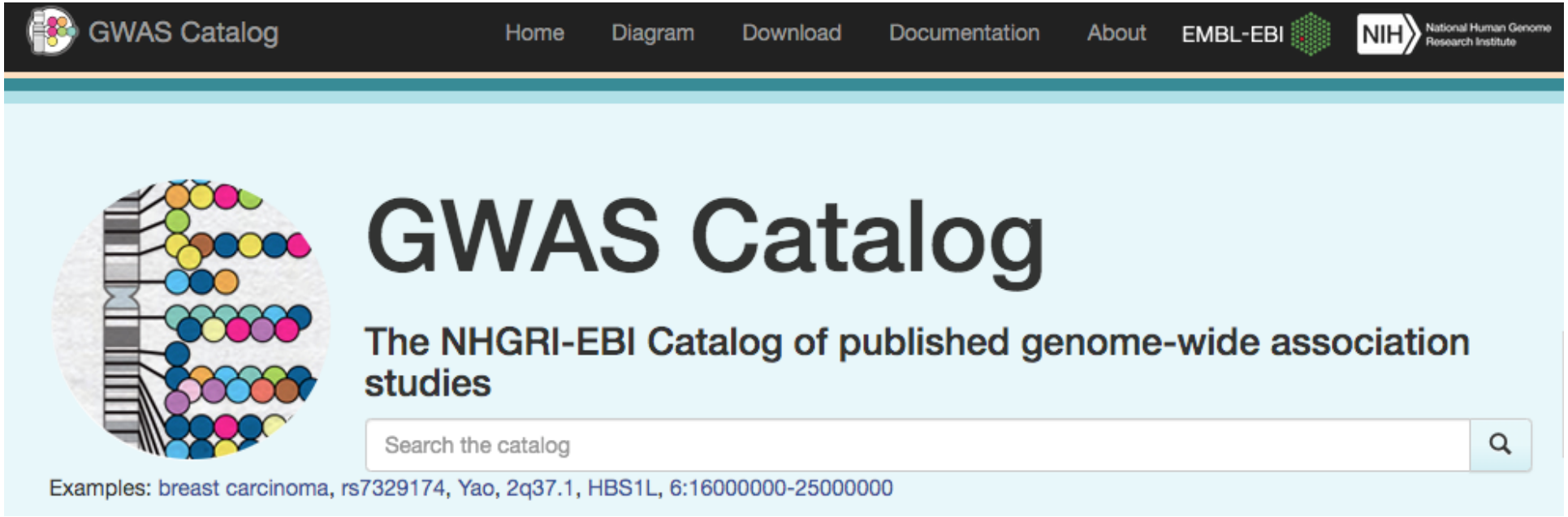
经过人工质控审核

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）

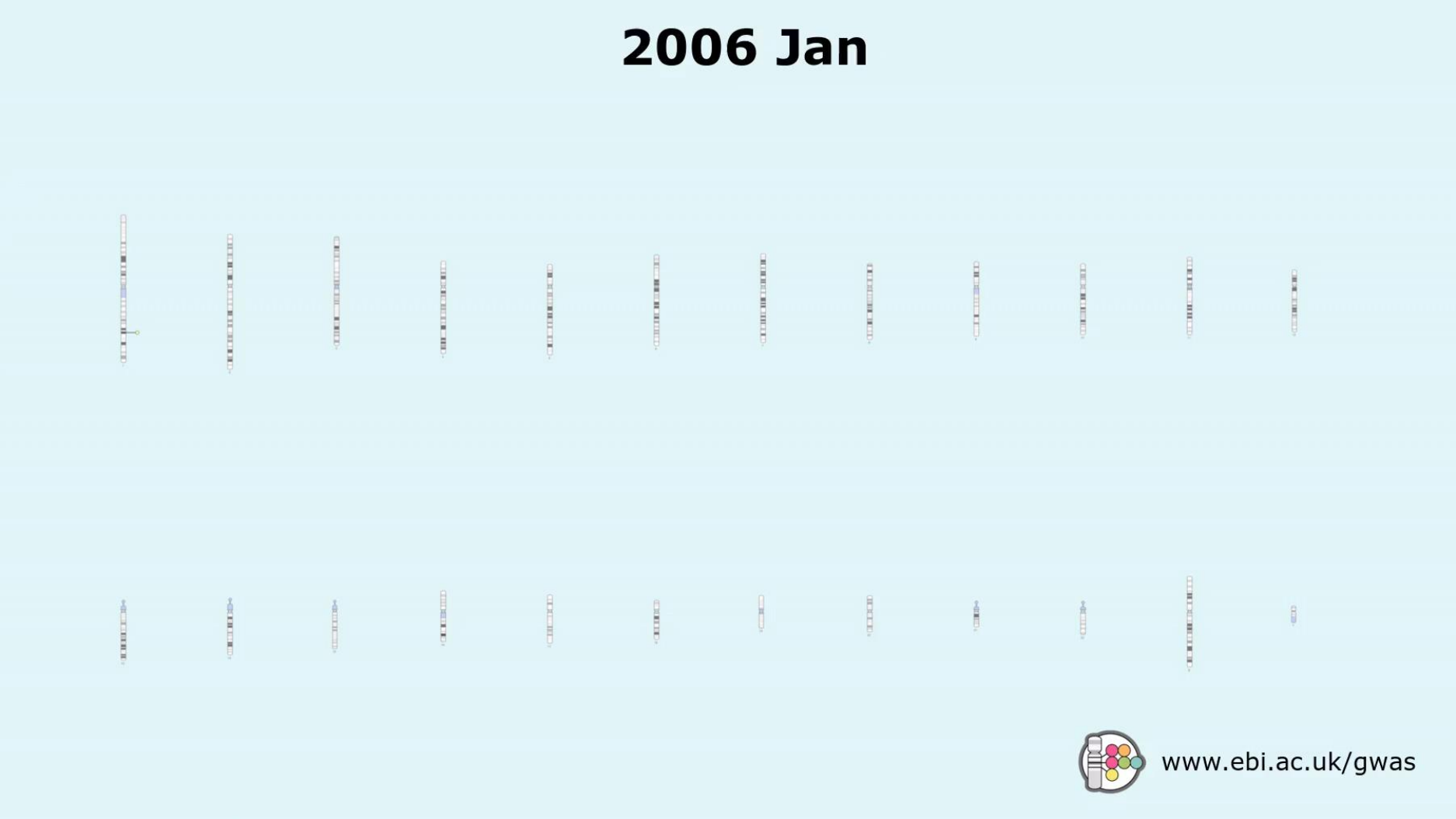
全基因组关联分析知识库GWAS Catalog



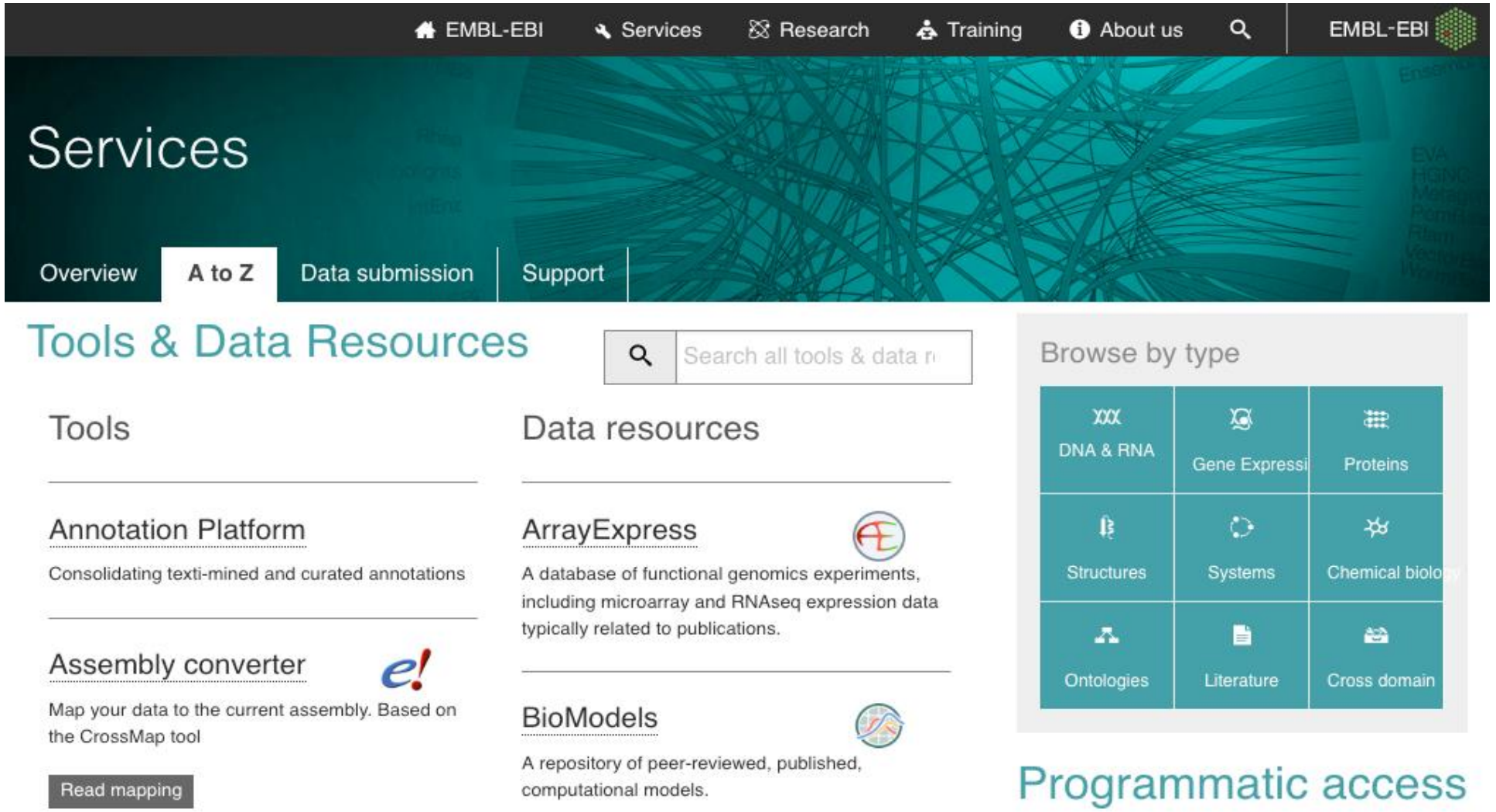
- 基于文献资料采用人工质控和审编的方法，提供高质量的基因型与表型关联信息
- 2008年由美国NIH下属的国家人类基因组研究所（National Human Genome Research Institute）创建，2010年与EBI合作开发
- 数据统计：>5000篇文献，>17万条关联信息

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

GWAS Catalog知识条目增长趋势



EBI数据和工具资源总览



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

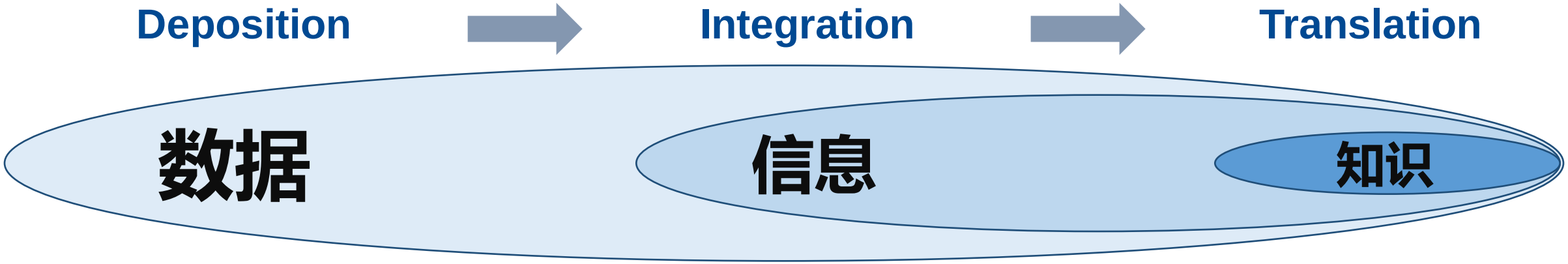
中国国家生物信息中心CNCB-NGDC

CNCB-NGDC代表性资源

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

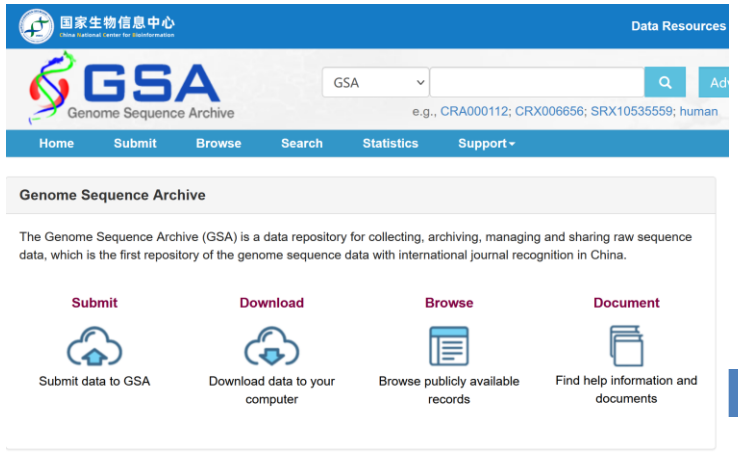
多层次数据资源体系



- **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

原始组学数据归档库GSA



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

2015年首次获国际期刊认可

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

Shaoping Ling[a,1], Zheng Hu[a,1], Zuyu Yang[a,1], Fang Yang[a,1], Yawei Li[a], Pei Lin[b], Ke Chen[a], Lili Dong[a], Lihua Cao[a], Yong Tao[a], Lingtong Hao[a], Qingjian Chen[b], Qiang Gong[a], Dafei Wu[a], Wenjie Li[a], Wenming Zhao[a], Xiuyun Tian[c], Chunyi Hao[c], Eric A. Hungate[d], Daniel V. T. Catenacci[e], Richard R. Hudson[f], Wen-Hsiung Li[g,2], Xuemei Lu[a,2], and Chung-I Wu[a,b,f,2]

[a]Key Laboratory of Genomics and Precision Medicine, China Gastrointestinal Cancer Research Center, Beijing Institute of Genomics, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Beijing 100101, People's Republic of China; [b]State Key Laboratory of Biocontrol, College of Ecology and Evolution, Sun Yat-Sen University, Guangzhou 510275, People's Republic of China; [c]Key Laboratory of Carcinogenesis and Translational Research, Peking University Cancer Hospital, Beijing 100142, People's Republic of China; [d]Department of Pediatrics, University of Chicago, Chicago, IL 60637; [e]Department of Medicine, Section of Hematology/Oncology, University of Chicago, Chicago, IL 60637; [f]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637; and [g]Biodiversity Research Center, Academia Sinica, Taipei 11529, Taiwan

Contributed by Wen-Hsiung Li, October 10, 2015 (sent for review September 14, 2015; reviewed by Takashi Gojobori and Jianzhi Zhang)

The prevailing view that the evolution of cells in a tumor is driven by Darwinian selection has never been rigorously tested. Because selection greatly affects the level of intratumor genetic diversity, it is important to assess whether intratumor evolution follows the Darwinian or the non-Darwinian mode of evolution. To provide the statistical power, many regions in a single tumor need to be sampled and analyzed much more extensively than has been attempted in previous intratumor studies. Here, from a hepatocellular carcinoma (HCC) tumor, we evaluated multiregional samples from the tumor, using either whole-exome sequencing (WES) (n = 23 samples) or genotyping (n = 286) under both the infinite-site and infinite-allele models of population genetics. In addition to the many single-nucleotide variations (SNVs) present in all samples, there were 35 "polymorphic" SNVs among samples. High genetic diversity was evident as the 23 WES samples defined 20 unique cell clones. With all 286 samples genotyped, clonal diversity agreed well with the non-Darwinian model with no evidence of positive Darwinian selection. Under the non-Darwinian model, M_{ALL} (the number of coding region mutations in the entire tumor) was estimated to be greater than 100 million in this tumor. DNA sequences reveal local diversities in small patches of cells and validate the estimation. In contrast, the genetic diversity under a Darwinian model would generally be orders of magnitude smaller. Because the level of genetic diversity will have implications on therapeutic resistance, non-Darwinian evolution should be heeded in cancer treatments even for microscopic tumors.

intratumor heterogeneity | genetic diversity | neutral evolution | cancer evolution | natural selection

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.

实现“零”的突破

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**).

人类遗传组学数据归档库GSA-Human

国家生物信息中心

Data Resources | Computing Analysis | Data Network | Standards

GSA for Human
Genome Sequence Archive

GSA for Humar | find a GSA-Human accession

e.g., HRA000087; HRA000150

中文 English

Home | Submit | Browse | Search | DAC | Statistics | Documentation | Policy | Login | Register

Home / Browse

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

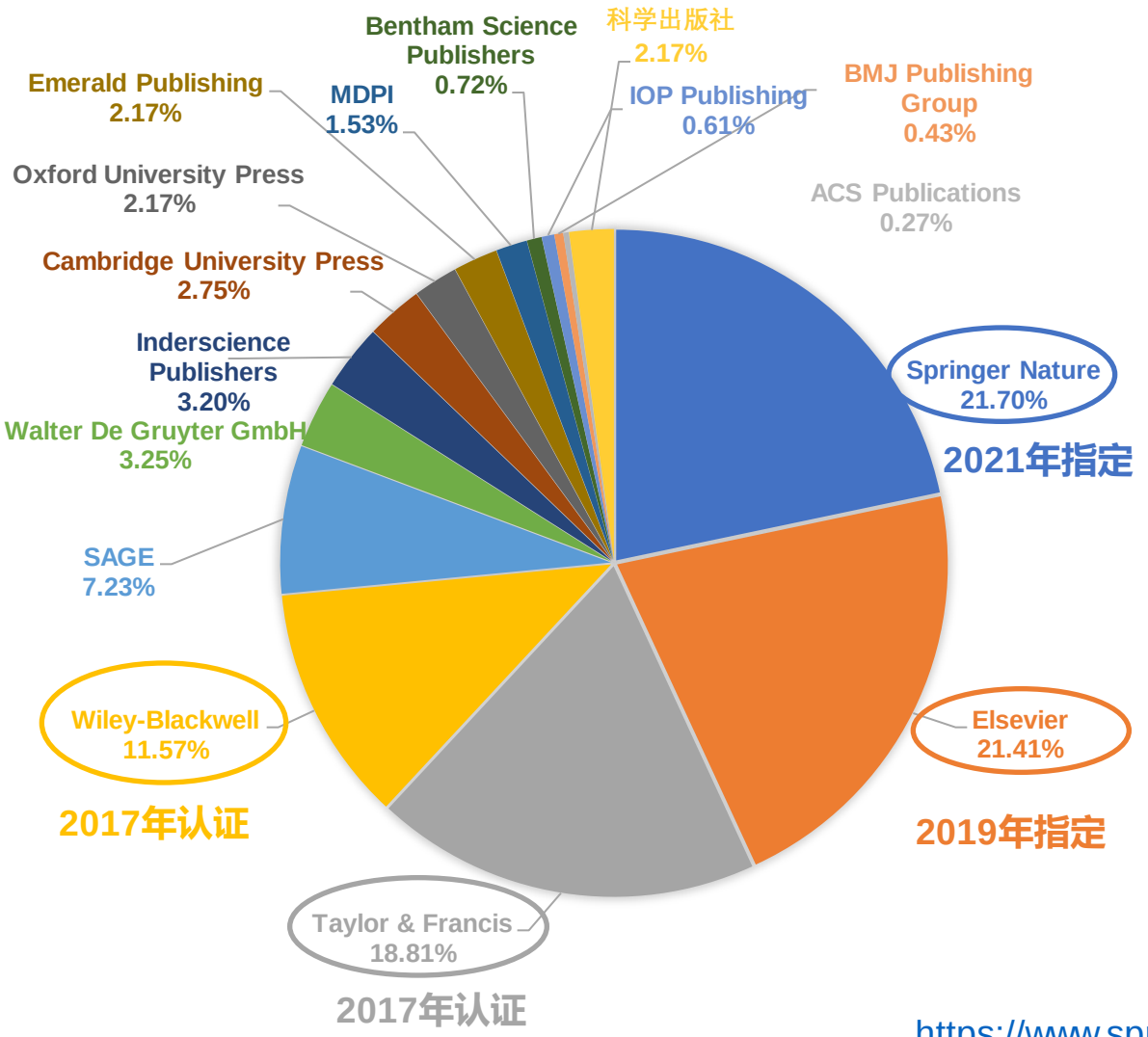
5374
Studies

551386
Individuals

791053
Samples

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

GSA被国际出版集团认可



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	Repositories
DNA sequence data*	Any INSDC member repository
RNA sequence data*	Genome Sequence Archive (GSA)
Genome assembly data*	
Genetic variation data	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.

https://www.springernature.com/gp/authors/research-data-policy/repositories-bio/

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

备份数据

备份编号

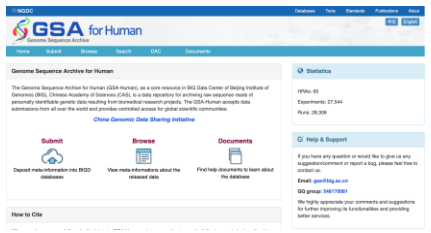
获得备案号

科研人员

科技部人遗管理办

提交数据

获得数据发布编号



CNCB-NGDC人遗数据管理平台

2022年7月18日

优化

CNCB-NGDC–**备份、发布**

提交数据

备份编号与发布编号

备份号

备案号

科研人员

备案信息注册

获得备案号

科技部信息中心 - **备案审查**

人类遗传资源信息管理备份一体化平台

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 > 通知通告

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

实现人类遗传资源信息管理、备份、发布与共享一体化

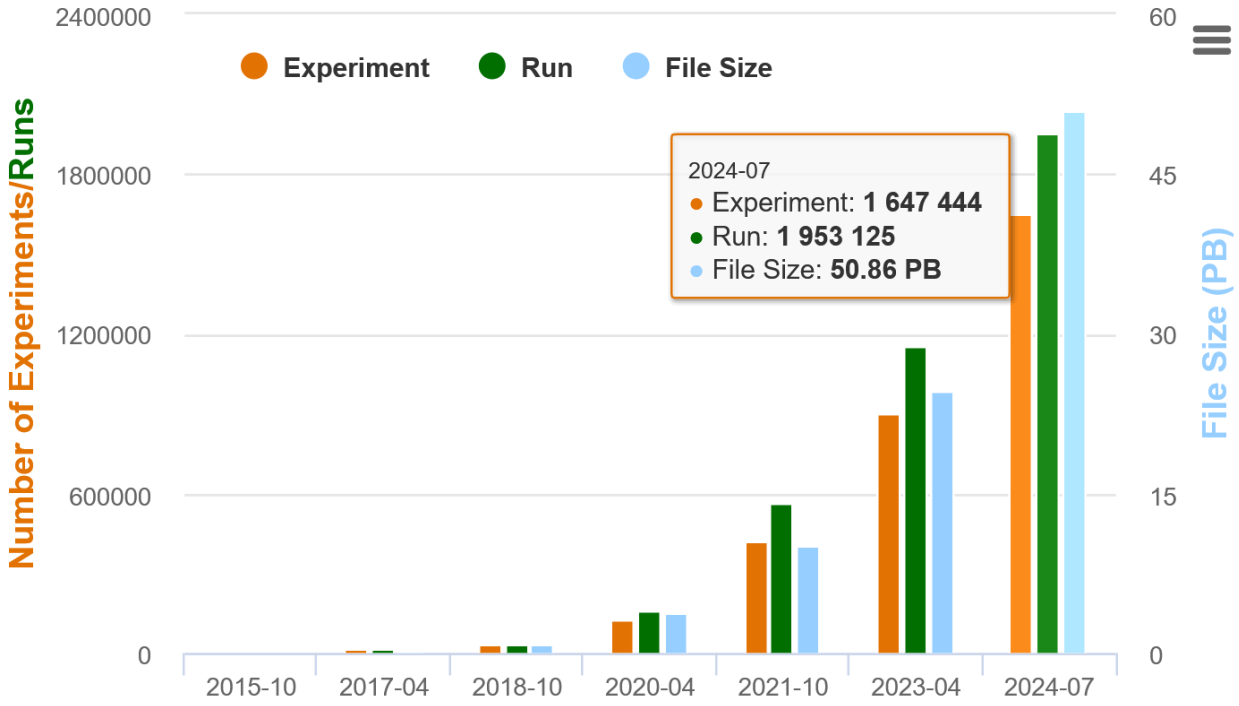
https://ngdc.cncb.ac.cn/hgrip/

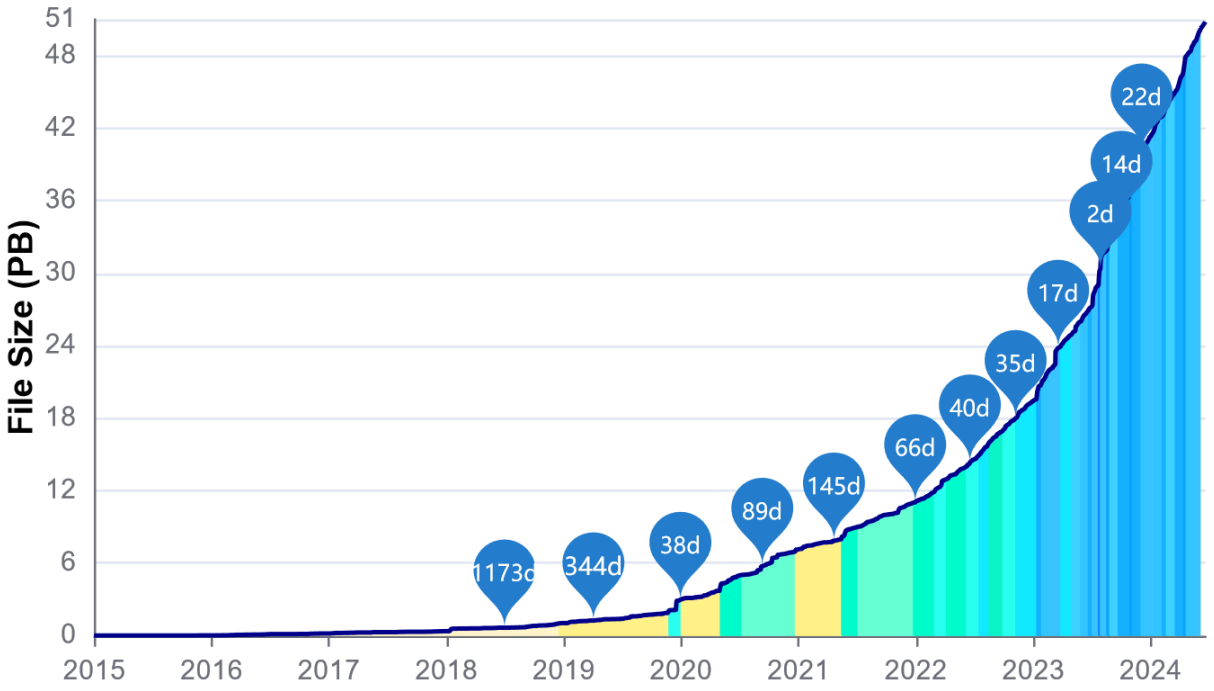
支撑管理科学项目数据汇交归档

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

数据统一汇交和安全管理

用户递交数据量：50.86 PB

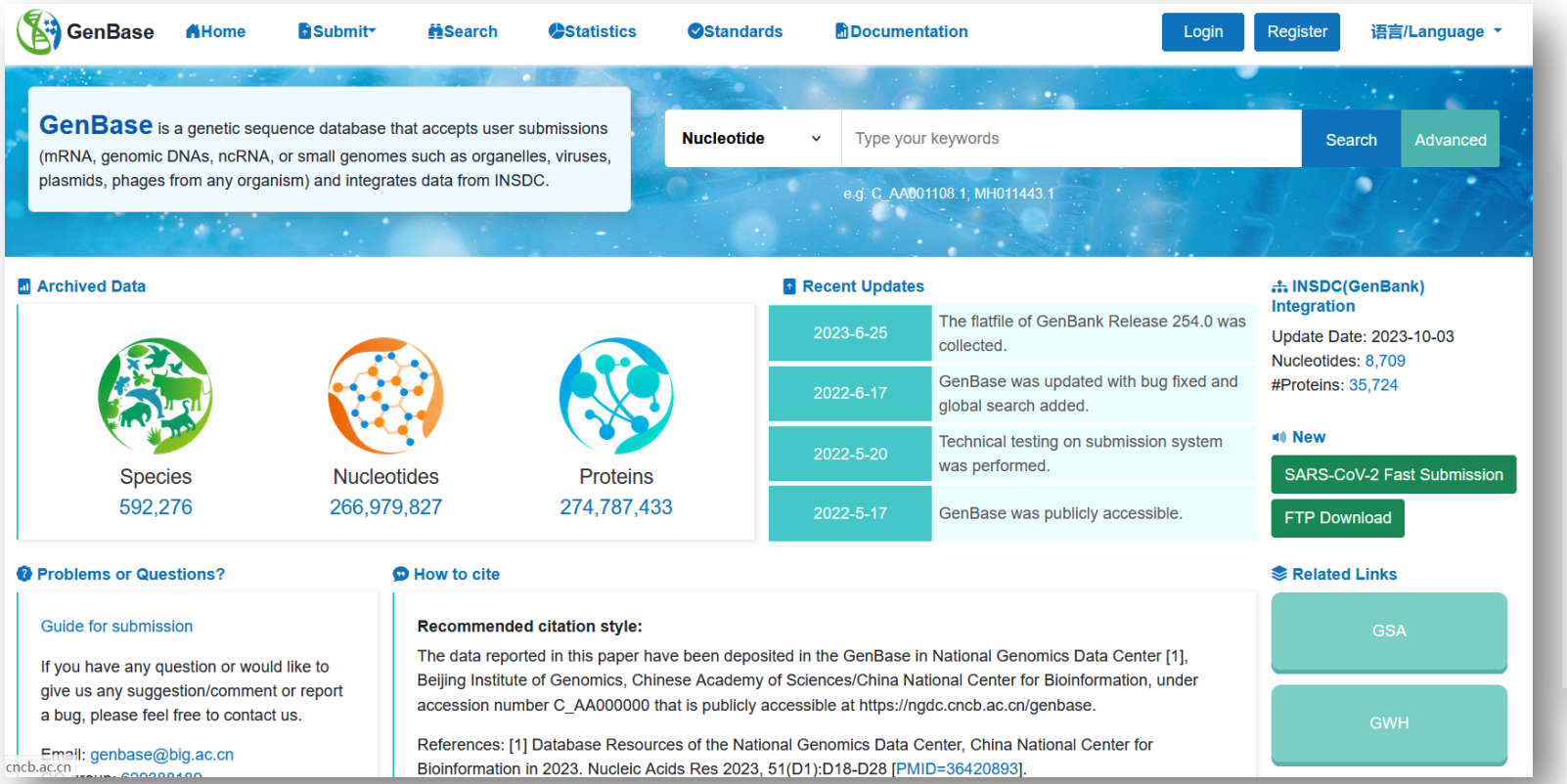




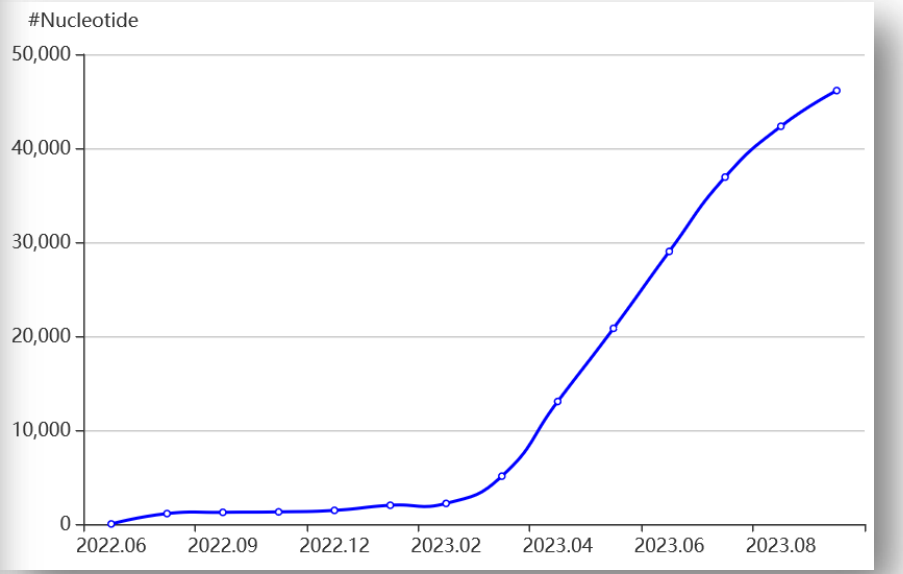
Genomics Proteomics & Bioinformatics (2017, 2021)

https://ngdc.cncb.ac.cn/gsa/statistics

基因序列数据库GenBase



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

基因组序列库Genome Warehouse

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.

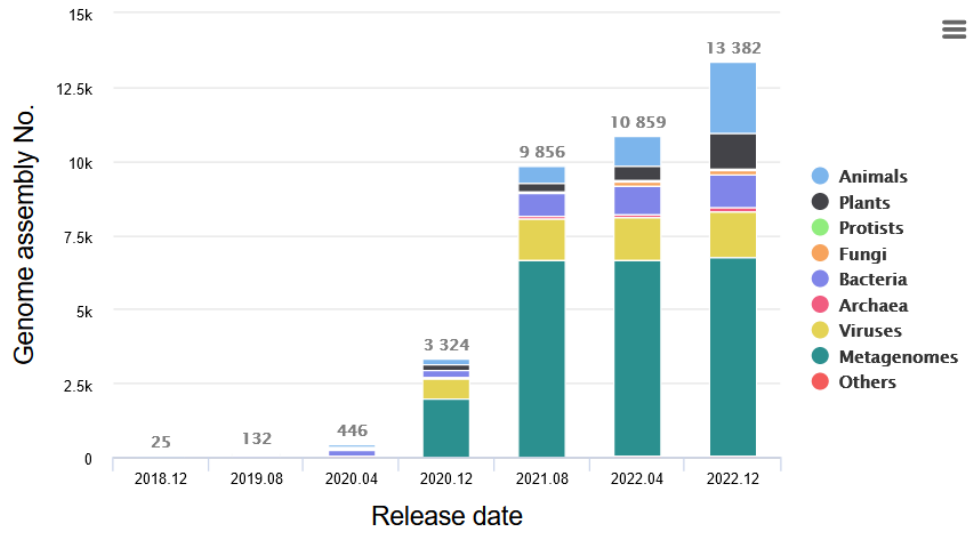
Submit
Deposit meta-information into GWH databases

Download
Transfer GWH data to your computer

Browse
View genome information about the released data

Documentation
Find help documents to learn more about GWH

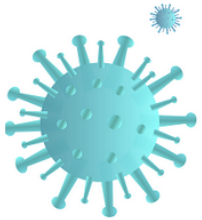
Data Growth

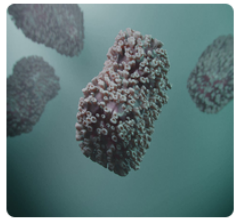


Statistics

- **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

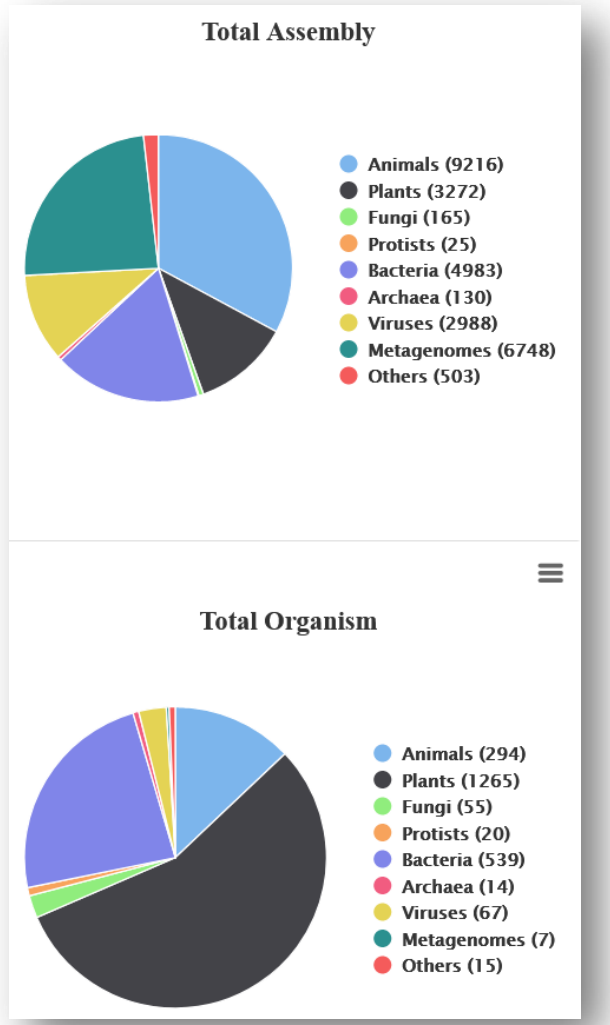

RCoV19


MPoxVR

GWH-supported Deposition

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

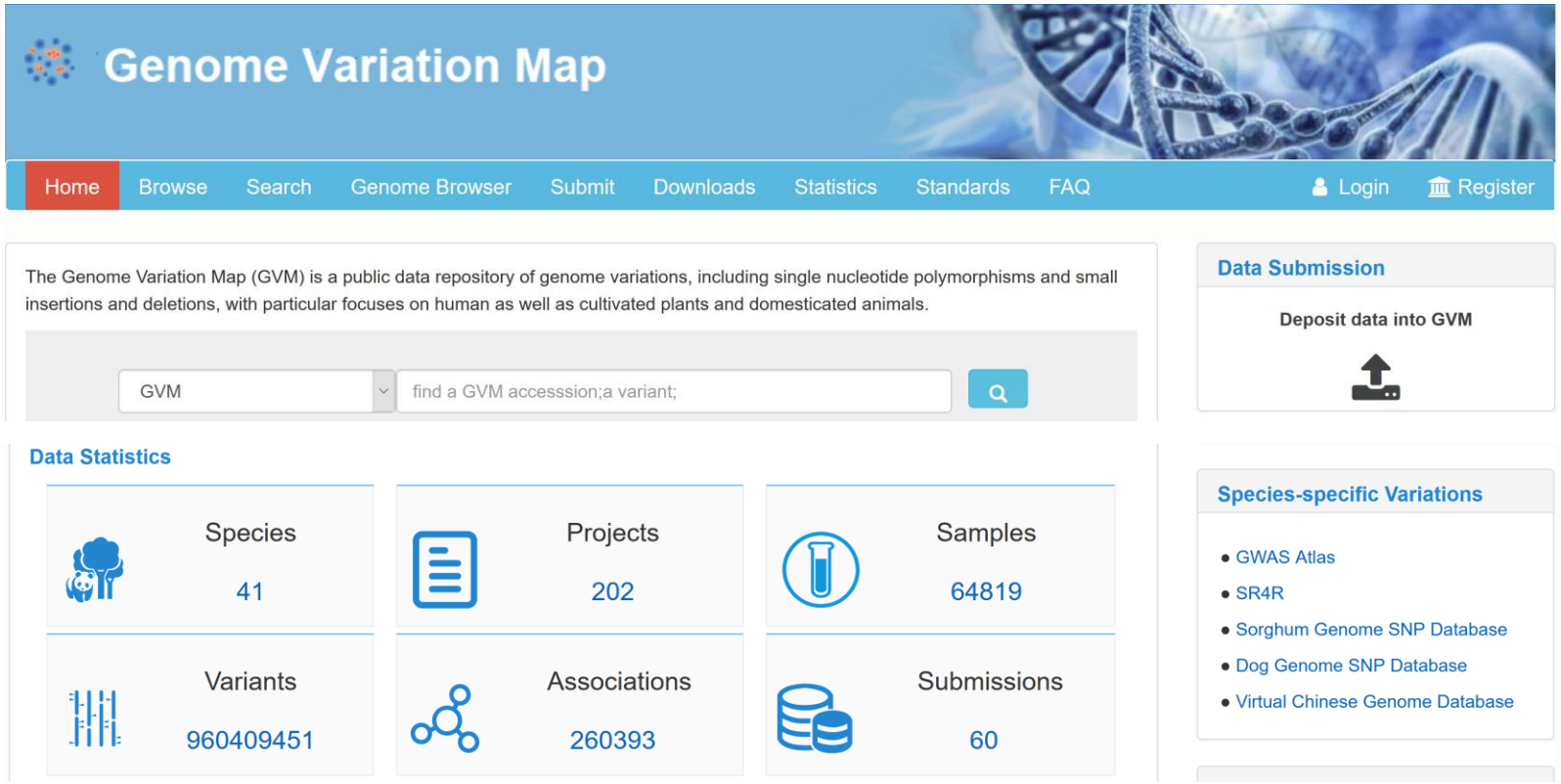
NSR National Science Review | Molecular Plant | New Phytologist



Geno Proto Bioinfo (2021)

http://ngdc.cncb.ac.cn/gwh

基因组变异库Genome Variation Map



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

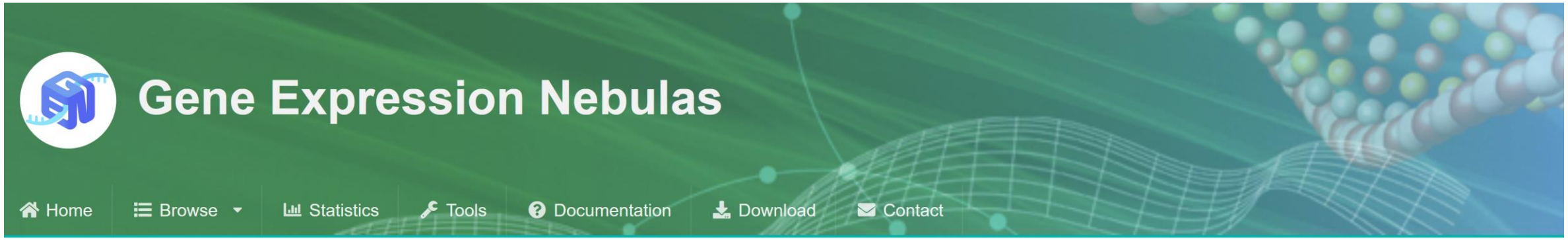
人、牛、鸡、狗
鸭、大熊猫、虎鲸
家猪、绵羊

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

基因表达库Gene Expression Nebulas



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	Glycine max	Glycine soja	Oryza sativa	Sorghum bicolor	Triticum aestivum
141 projects	16 projects	1 projects	25 projects	5 projects	3 projects
25619 experiments	499 experiments	7 experiments	945 experiments	462 experiments	78 experiments

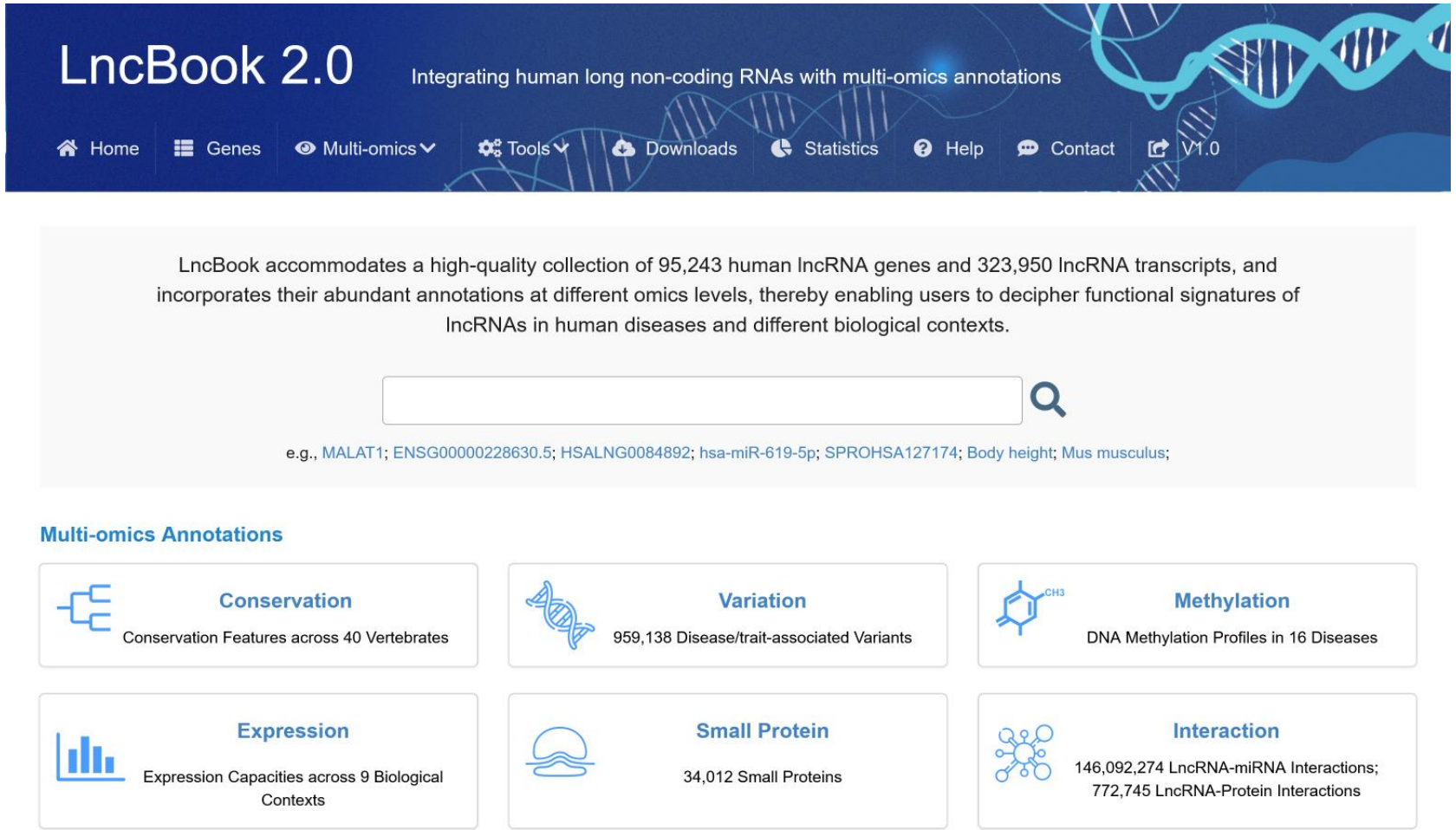
Featured Projects

The latest RNA-seq research data on COVID-19.

- PRJCA002326: Transcriptome of patients infected with SARS-CoV-2
- PRJNA615032: Transcriptional response to SARS-CoV-2 infection
- PRJNA631753: Spectrum of Viral Load and Host Response Seen in Autopsies of SARS-CoV-2 Infected Lungs

Briefings in Bioinformatics (2016) *Nucleic Acids Research* (2018, 2022) https://ngdc.cncb.ac.cn/gen

人类长非编码RNA知识库LncBook



RNAcentral

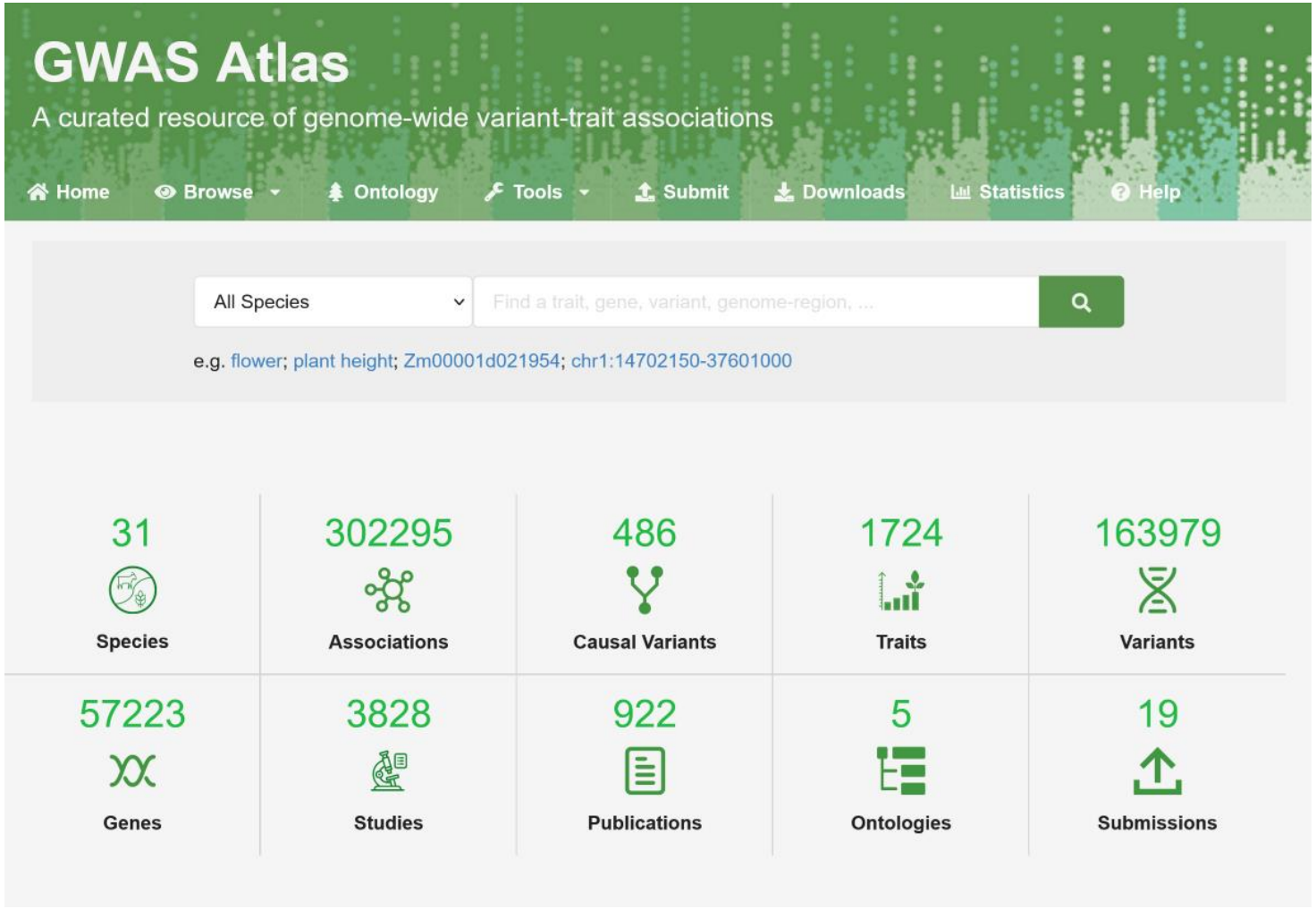
2016年7月加入国际RNAcentral联盟

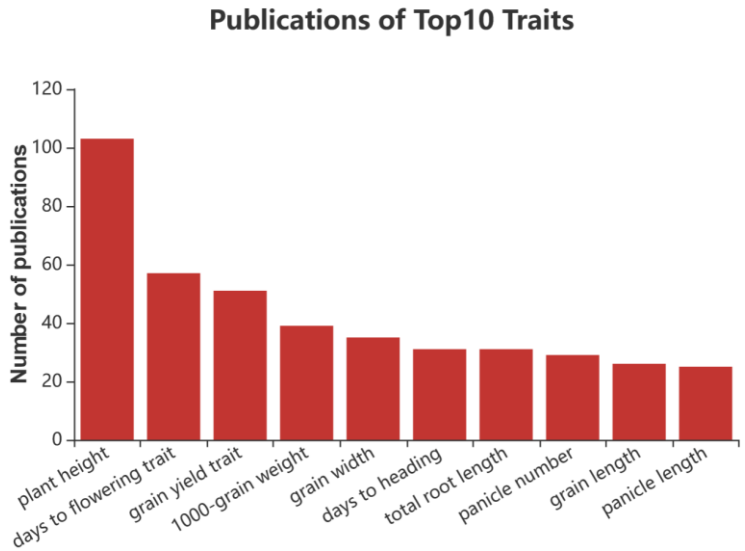
http://ngdc.cncb.ac.cn/lncbook

Bioinformatics (2019)

Nucleic Acids Res (2015, 2019, 2022)

全基因组关联分析知识库GWAS Atlas





Oryza sativa
Rice

163479 Associations
461 Traits

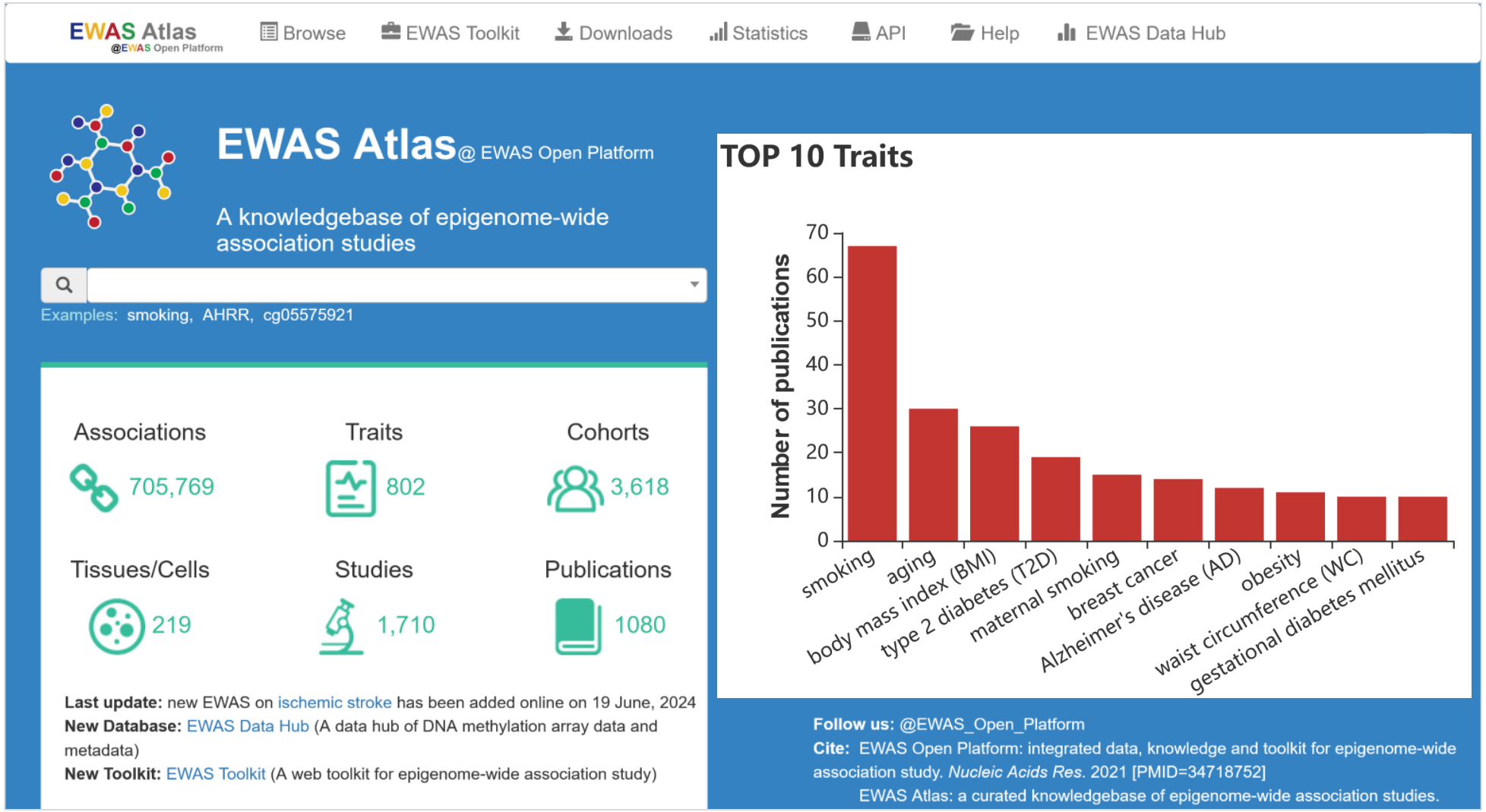
Sorghum bicolor
Sorghum

8829 Associations
151 Traits

http://ngdc.cncb.ac.cn/gwas

Nucleic Acids Res (2020, 2023)

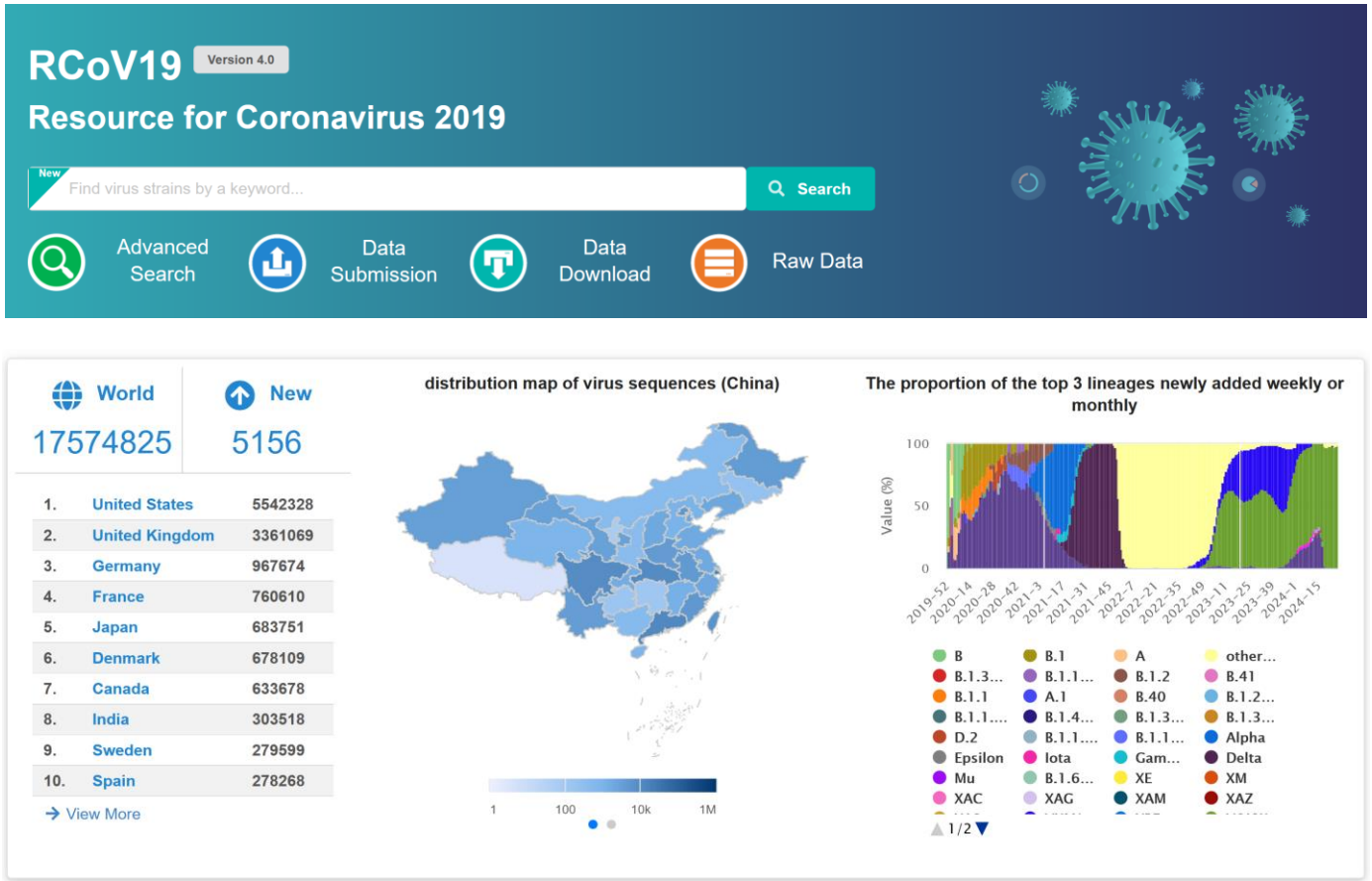
全表观组关联分析知识库EWAS Atlas



Nucleic Acids Res (2019, 2021)

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

新冠病毒信息库



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%

Yi Chuan 2020; *Zoological Research* 2020; *Genomics Proteomics Bioinformatics* 2020; *Nucleic Acids Research* 2021

序列质控审编

1、# of Ns
- N<15→G
- N>15→R

3、# of Gap(G)
- G<=2
- G > 2

5、Density of variants
- D < 0.25→G
- D>=0.25→R

2、# of degenerate bases (X)
- X<=50→G
- X > 50→R

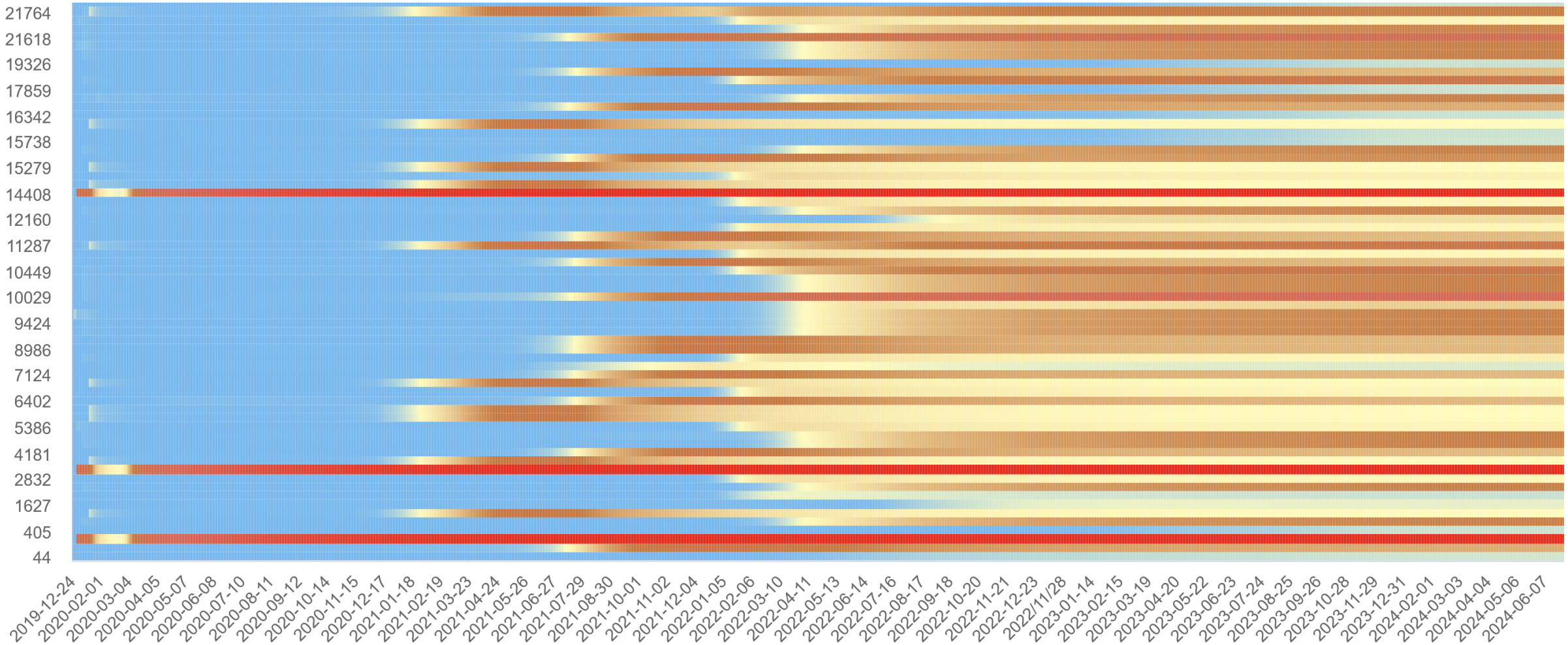
4、# of variants (S)
- S<=Exp →G
- S > Exp →R

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	17472053	17574825	3340	13072	9613
CORONAVIRUS SEQUENCES	ISOLATES	SARS-COV-2 SEQUENCES	SUBMITTING LABS	ORIGINATING LABS	LOCATIONS

新冠基因组变异动态图谱

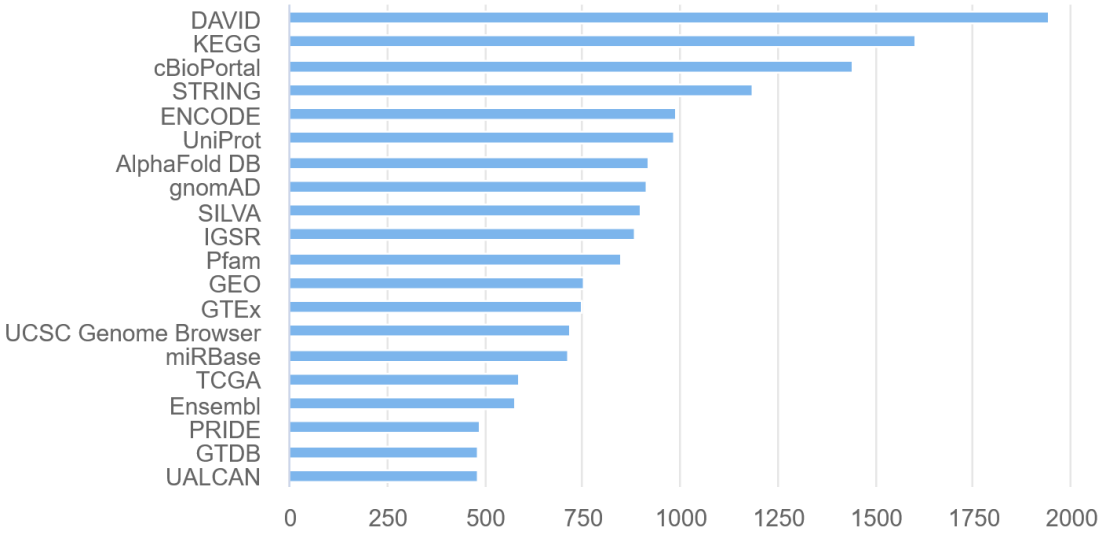


https://ngdc.cncb.ac.cn/ncov/variation/heatmap

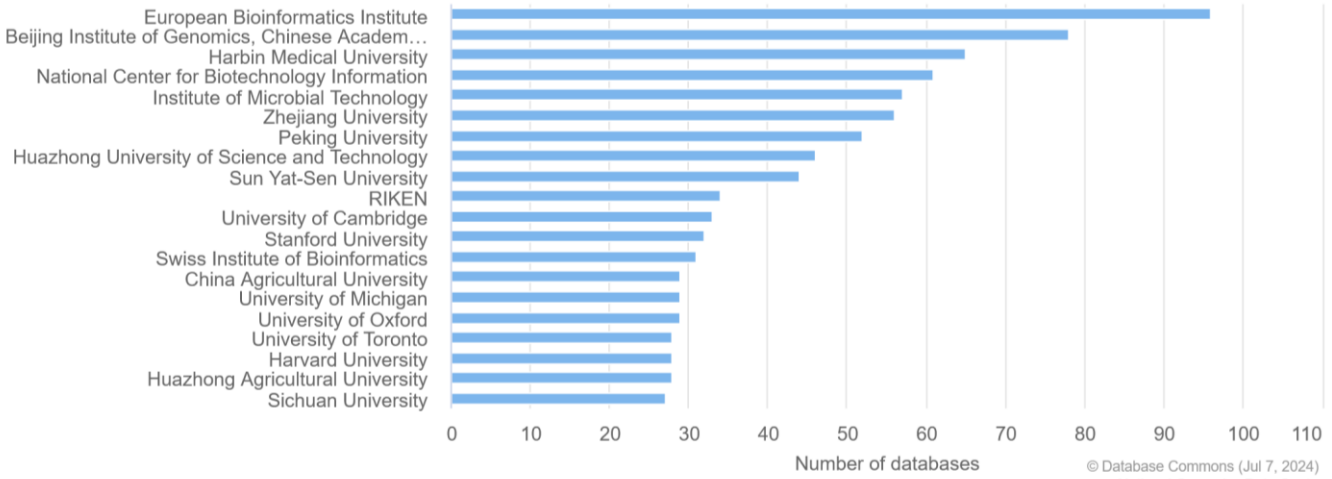
全球生物数据库目录Database Commons

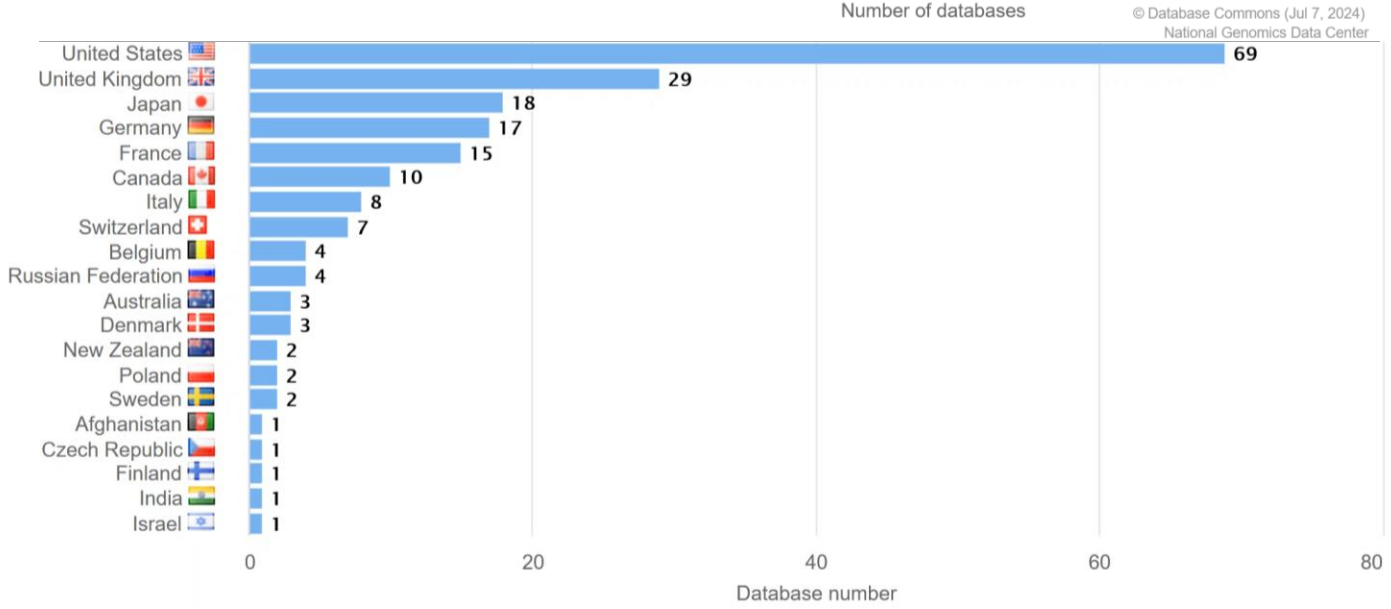
6392	13	1647
DATABASES	CATEGORIES	SPECIES
9875	2142	76
PUBLICATIONS	INSTITUTIONS	COUNTRIES / REGIONS

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



Top 20 institutions by database count





https://ngdc.cncb.ac.cn/databasecommons

跨库搜索引擎BIG Search

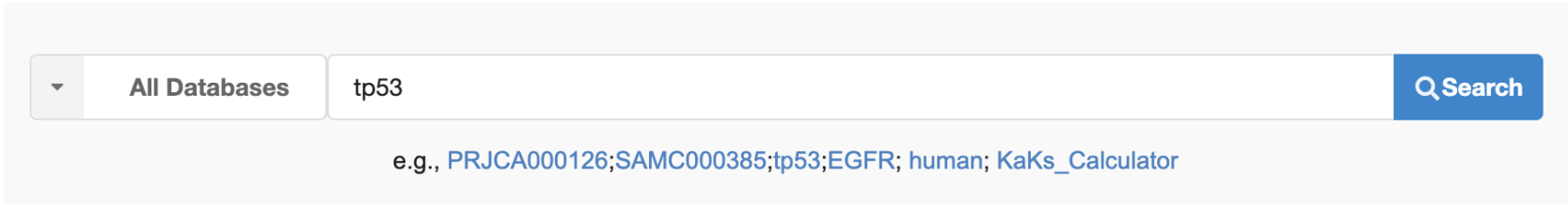
索引记录数：14.77亿　　索引大小：1.5TB

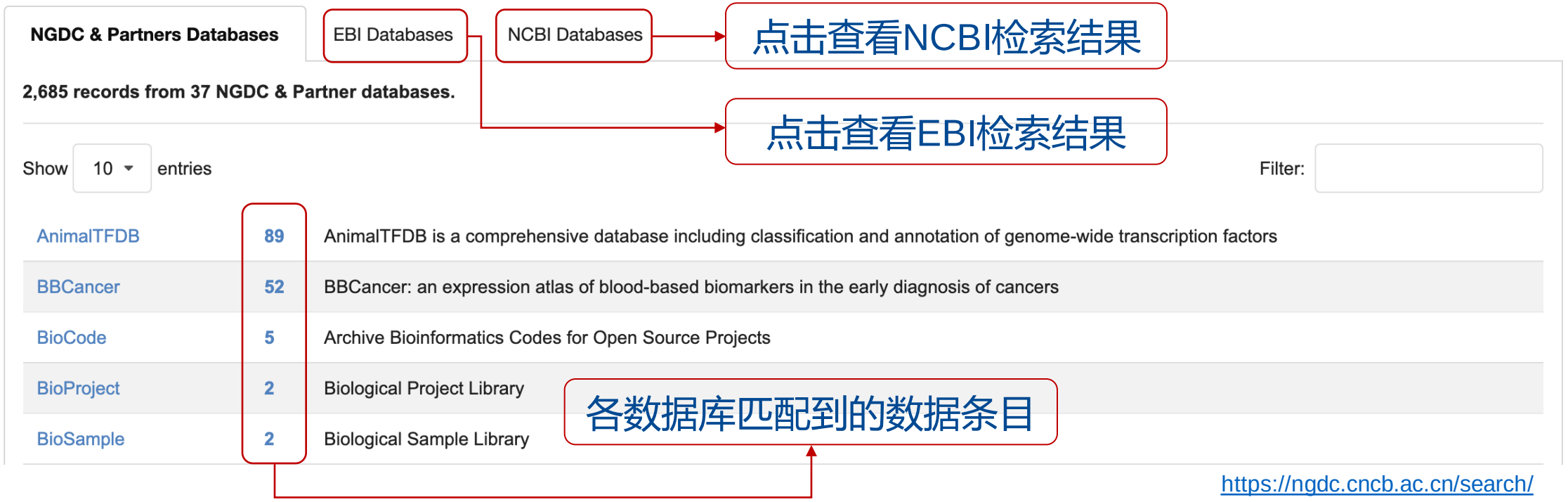
CancerSEA Cancer Single-cell State Atlas
PceRBase
GenTree The Time Tree of Genes Along the Evolution History
AnimalTFDB2.0
CellMarker
NERAM
FLMD
iUUCD
SEGreg
PlantTFDB
BBCancer
lnCAR
RhesusBase
RMVar!
dbPSP
GTDC Glycosyltransferases Database
SEECancer
EPSD
dbPSP
LncRNADisease
lncRNASNP2
DoriC
DEG
DiseaseEnhancer
DrLLPS
dbPAF
iEKPD
dbPPT
hTFtarget
eLMSG An eLibrary of Microbial Systematics and Genomics
miRNASNP

一站式跨库搜索引擎BIG Search

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.



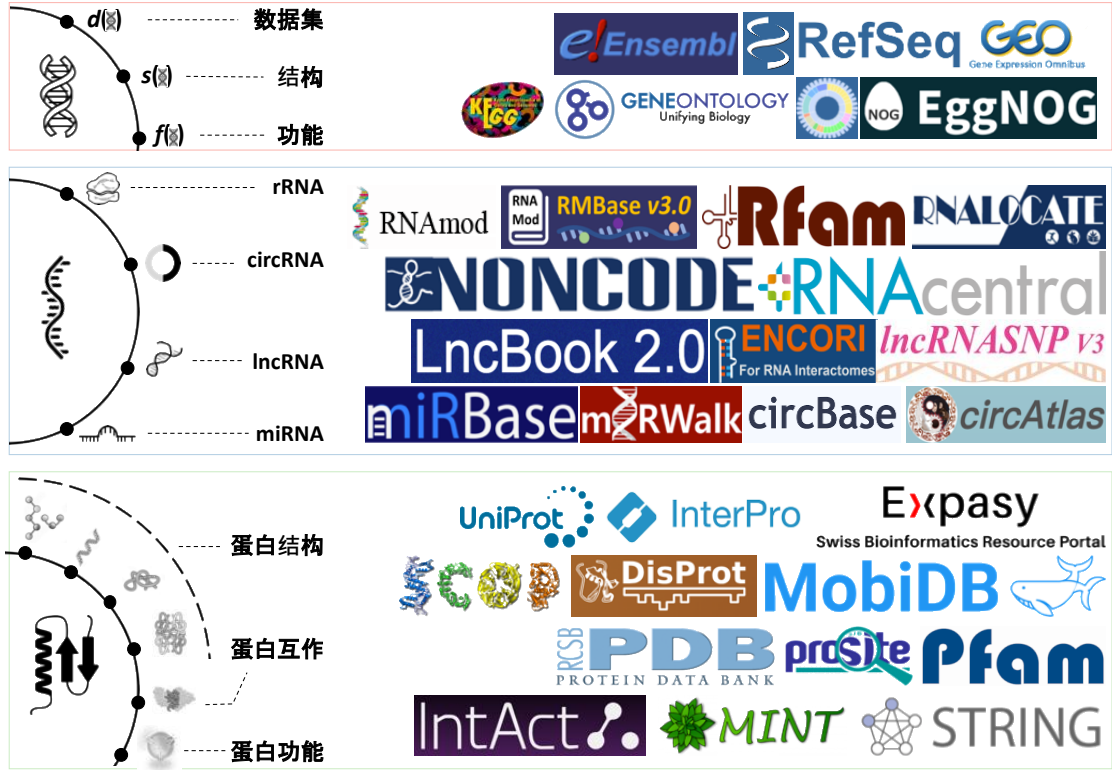


https://ngdc.cncb.ac.cn/search/

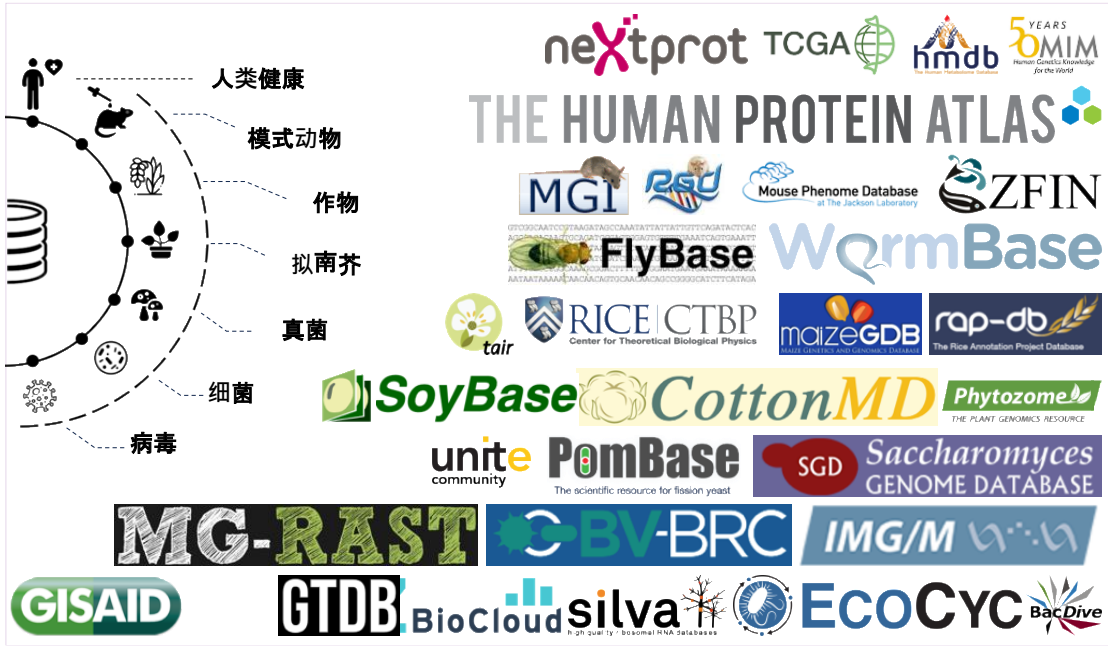
第三节 国际重要生物数据库

代表性生物数据库

生物大分子数据库（DNA、RNA、Protein）



物种专题数据库



其他类型数据库

生物大分子数据库

DNA

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

RNA

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
- lncRNASNP 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

Protein

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

生物大分子数据库（DNA）

DNA

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

RNA

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
- lncRNASNP 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

Protein

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

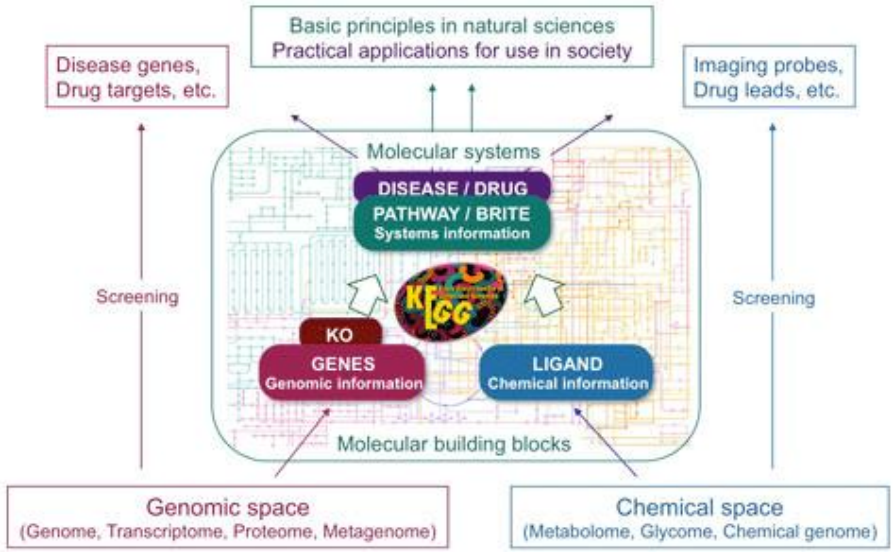
- Pfam 蛋白质家族与结构域分类数据库
- PROSITE 结构域与功能位点数据库

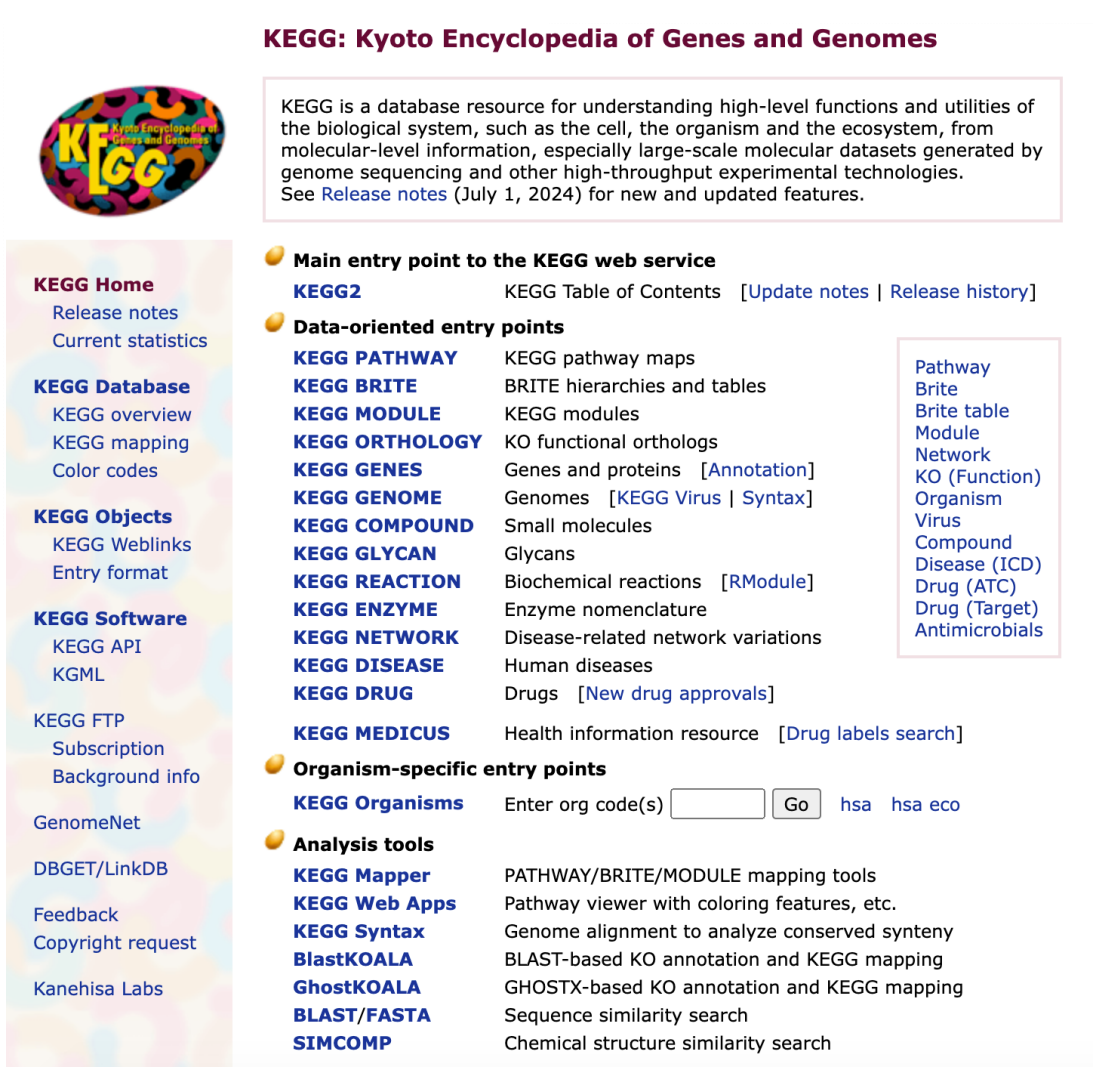
生物大分子数据库（KEGG）

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数据库首页（2024.7）

生物大分子数据库（KEGG）

KEGG **Escherichia coli K-12 MG1655**

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

Search genes: Go Clear

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; Gammaproteobac
Brite	KEGG organisms [BR:br08601] KEGG organisms in the NCBI taxonomy [BR:b KEGG organisms in taxonomic ranks [BR:br08
Data source	RefSeq (Assembly: GCF_000005845.2 Comple BioProject: 57779
Original DB	Wisconsin, Pasteur, ECOCYC, ASAP
Comment	Colonizes the lower gut of animals, and, as a Maintained as a laboratory strain with minima bacteriophage lambda and F plasmid by ultrav
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

gy (KO) - 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

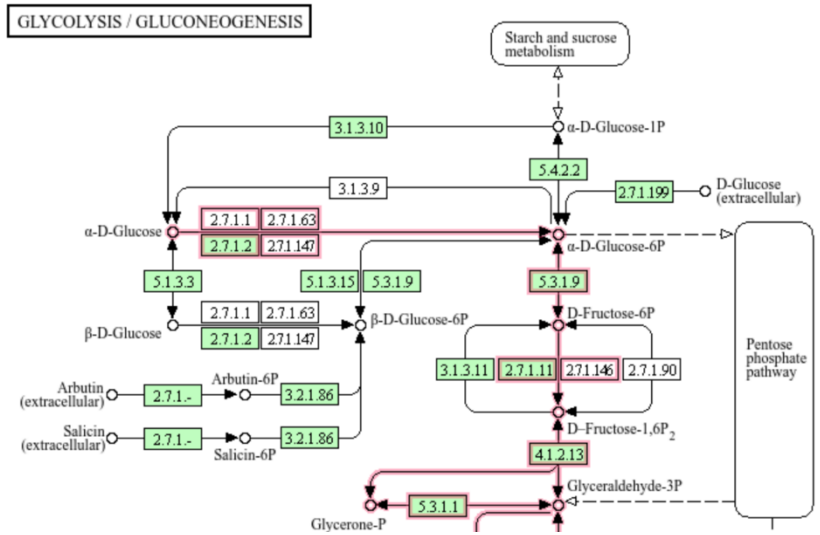
Download json | Copy URL | Help]

09100 Metabolism

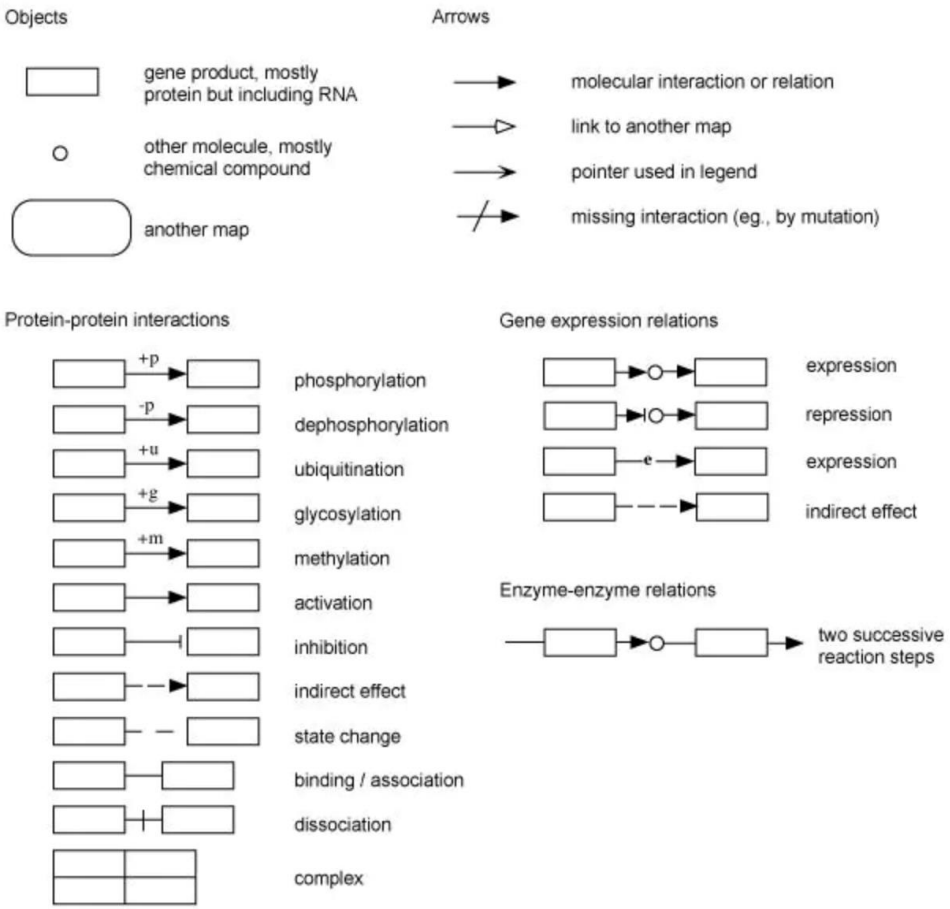
09101 Carbohydrate metabolism

00010 Glycolysis / Gluconeogenesis [PATH:eco00010]

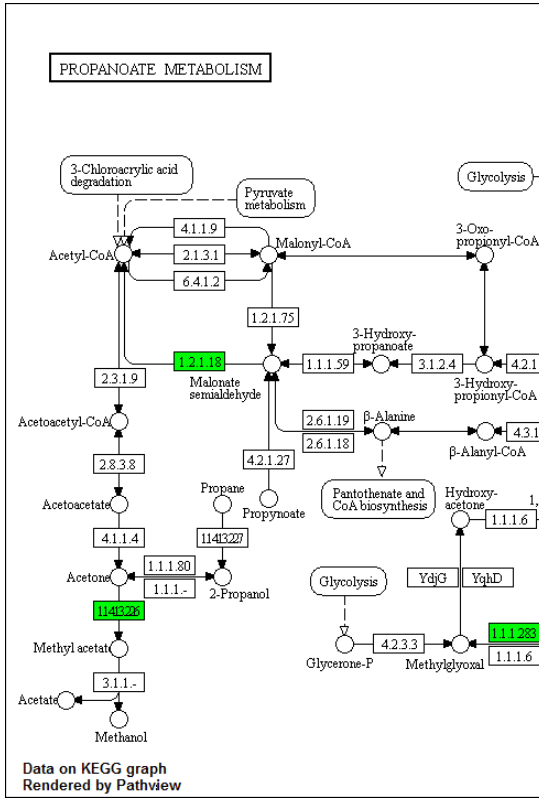
b2388 glk; glucokinase	K00845 glk; glucokinase [EC:2.7.1.2]
b4025 pgi; glucose-6-phosphate isomerase	K01810 GPI; glucose-6-phosphate isomerase [EC:5.3.1.9]
b3916 pfkA; 6-phosphofructokinase 1	K00850 pfkA; 6-phosphofructokinase 1 [EC:2.7.1.11]
b1723 pfkB; 6-phosphofructokinase 2	K16370 pfkB; 6-phosphofructokinase 2 [EC:2.7.1.11]
b4232 fbp; fructose-1,6-bisphosphatase 1	K03841 FBP; fructose-1,6-bisphosphatase I [EC:3.1.3.11]
b3925 glpX; fructose-1,6-bisphosphatase 2	K02446 glpX; fructose-1,6-bisphosphatase II [EC:3.1.3.11]
b2930 yggF; fructose 1,6-bisphosphatase YggF	K02446 glpX; fructose-1,6-bisphosphatase II [EC:3.1.3.11]
b2097 fbaB; fructose-bisphosphate aldolase class I	K11645 fbaB; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]
b2925 fbaA; fructose-bisphosphate aldolase class II	K01624 FBA; fructose-bisphosphate aldolase, class II [EC:4.1.2.13]
b3919 tpiA; triose-phosphate isomerase	K01803 TPI; triosephosphate isomerase (TIM) [EC:5.3.1.1]
b1779 gapA; glyceraldehyde-3-phosphate dehydrogenase A	K00134 GAPDH; glyceraldehyde 3-phosphate dehydrogenase (phosphorylating) [EC:1.2.
b2926 pgk; phosphoglycerate kinase	K00927 PGK; phosphoglycerate kinase [EC:2.7.2.3]
b0755 gpmA; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase	K01834 PGAM; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.
b4395 ytjC; putative phosphatase	K15634 gpmB; 2,3-bisphosphoglycerate-dependent phosphoglycerate mutase [EC:5.4.2.
b3612 gpmM; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase	K15633 gpmI; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase [EC:5.4.
b2779 eno; enolase	K01689 ENO1_2_3; enolase 1/2/3 [EC:4.2.1.11]
b1676 pykF; pyruvate kinase I	K00873 PK; pyruvate kinase [EC:2.7.1.40]
b1854 pykA; pyruvate kinase II	K00873 PK; pyruvate kinase [EC:2.7.1.40]
b1702 ppsA; phosphoenolpyruvate synthetase	K01007 pps; pyruvate, water dikinase [EC:2.7.9.2]
b0114 aceE; pyruvate dehydrogenase E1 component	K00163 aceE; pyruvate dehydrogenase E1 component [EC:1.2.4.1]
b0115 aceF; pyruvate dehydrogenase, E2 subunit	K00627 DLAT; pyruvate dehydrogenase E2 component (dihydrolipoyllysine-residue ace
b0116 lpd; lipoamide dehydrogenase	K00382 DLD; dihydrolipoyl dehydrogenase [EC:1.8.1.4]
b1378 ydbK; putative pyruvate-flavodoxin oxidoreductase	K03737 por; pyruvate-ferredoxin/flavodoxin oxidoreductase [EC:1.2.7.1 1.2.7.-]



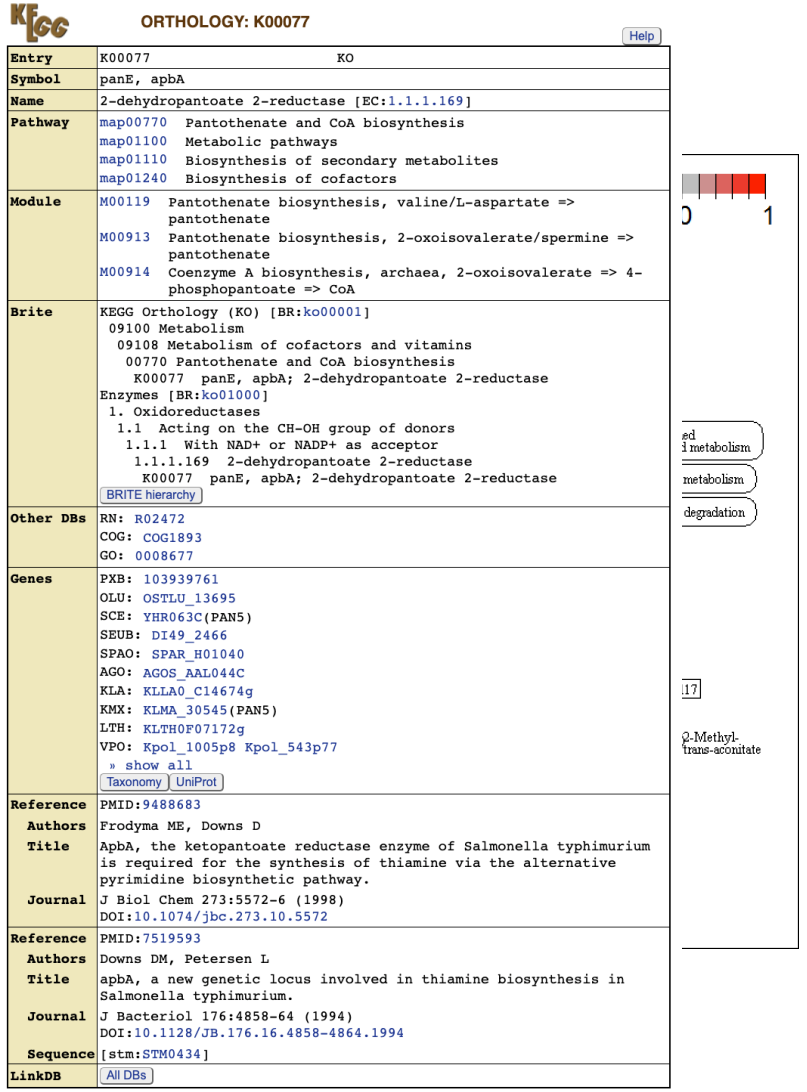
生物大分子数据库（KEGG）



KEGG通路图图例



KEGG多数据库关联体系



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: Kpol_1005p8 Kpol_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

生物大分子数据库

DNA

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

RNA

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
- lncRNASNP 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

Protein

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

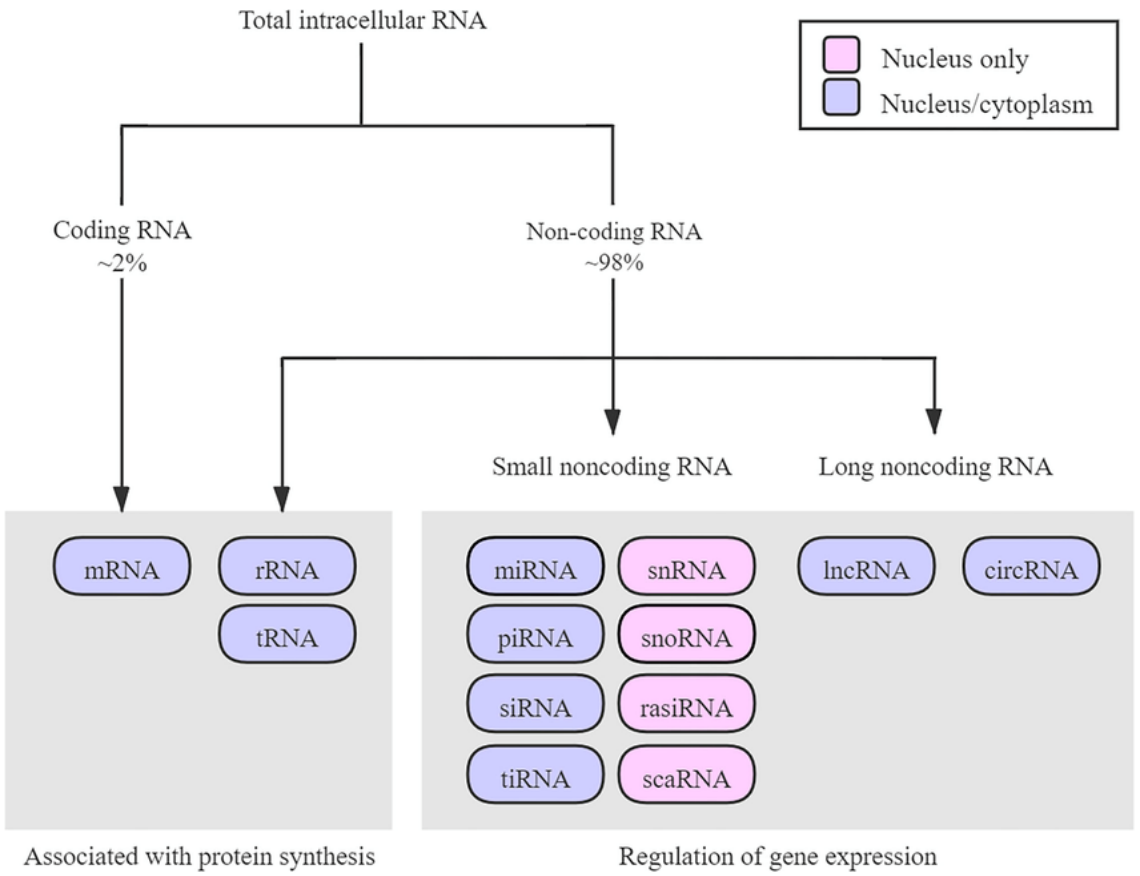
互作

- 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

生物大分子数据库（RNA）



Frontiers in Neurology (2021)

数据库名	运营情况	序列	circBaseID	host gene	miRNA	translatio	m6A	RBP	Conserve	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		*	*						*
pigCircNet	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		*	*		*				*

生物大分子数据库（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

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

RNA

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
- lncRNASNP 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

Protein

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

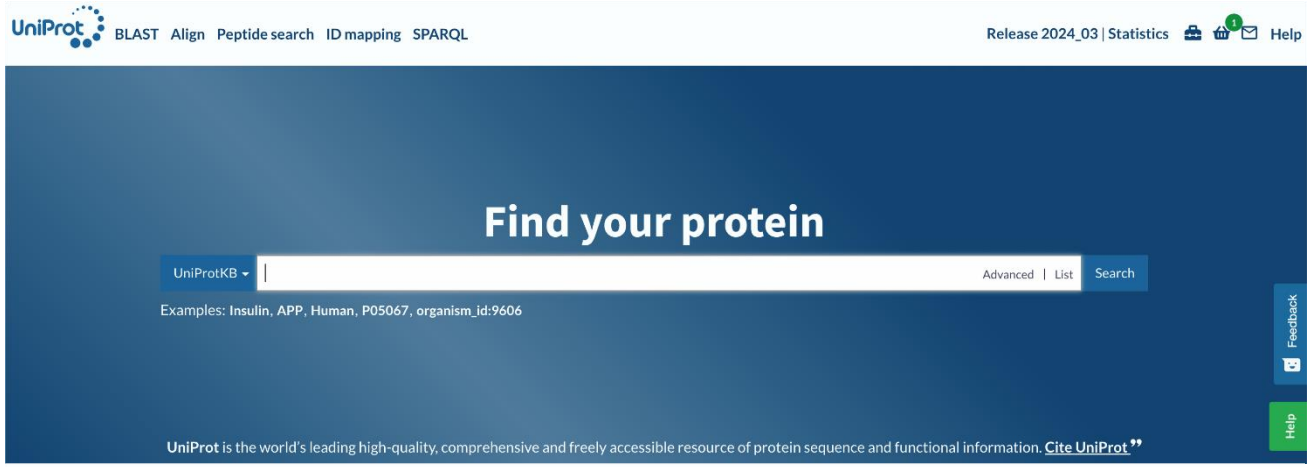
空间结构
- PDB 综合结构数据库
- AlphaFoldDB/ESM Atlas 预测结构

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

生物大分子数据库（蛋白质家族与结构域）

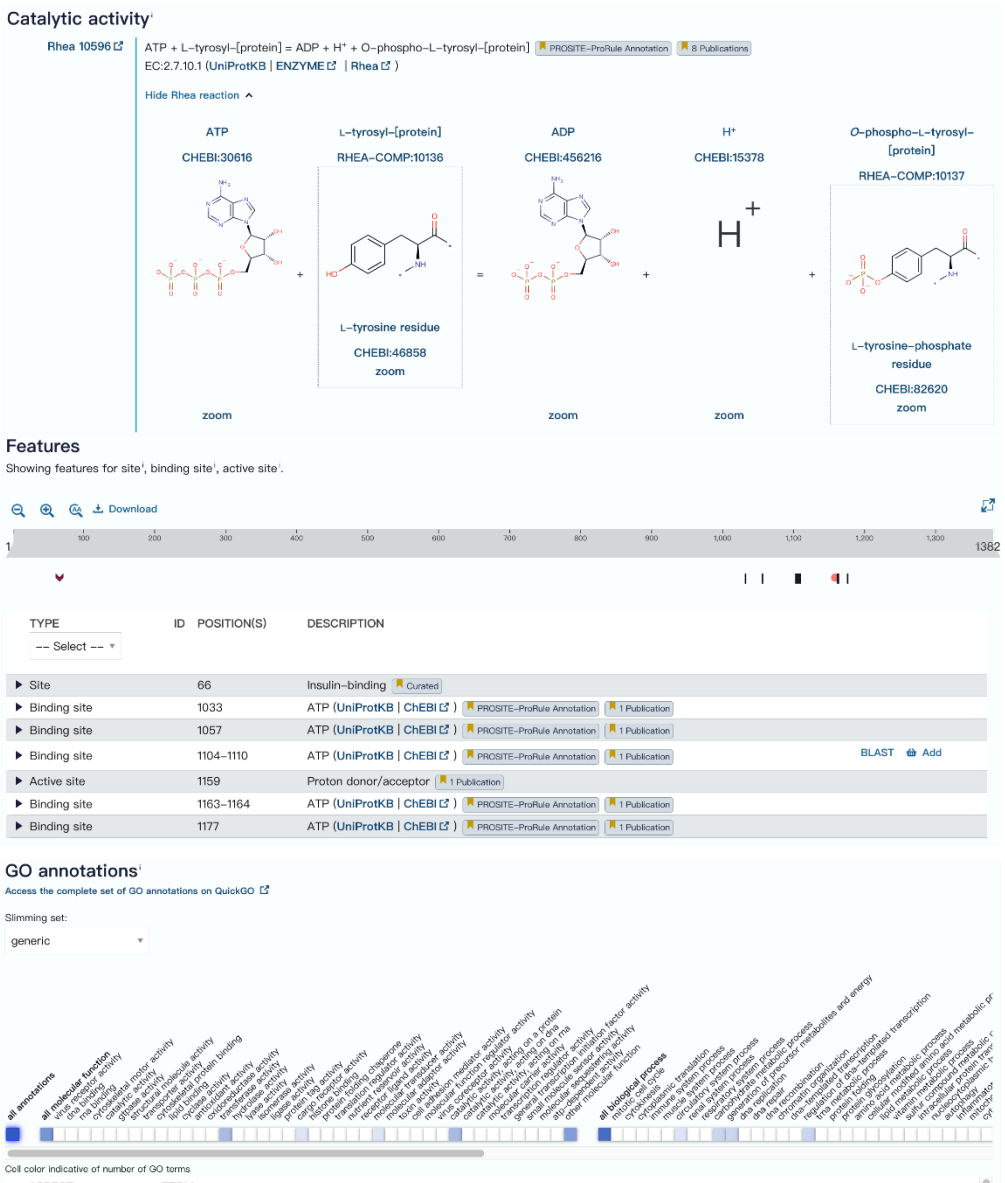


UniProtKB 301,086 results

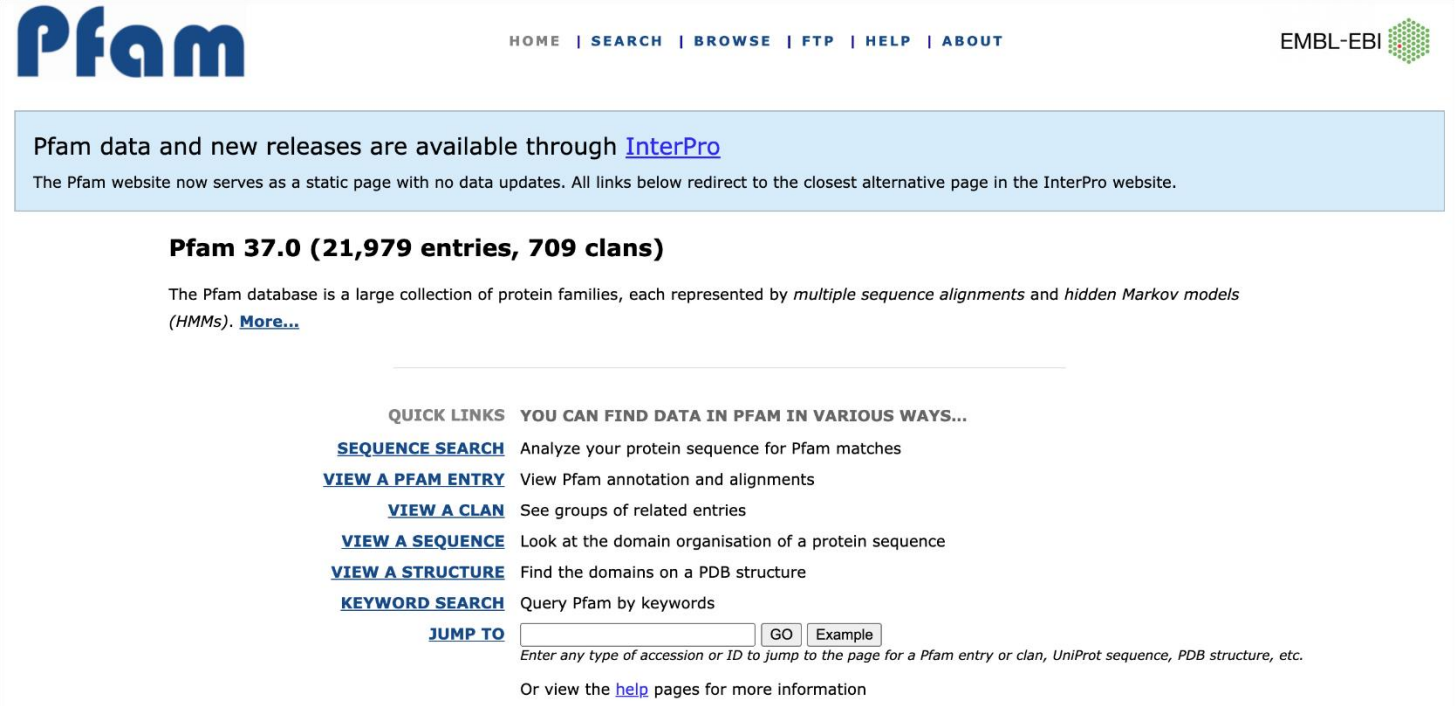
BLAST Align Map IDs Download Add View: Cards ○ Table ◉ Customize columns Share

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

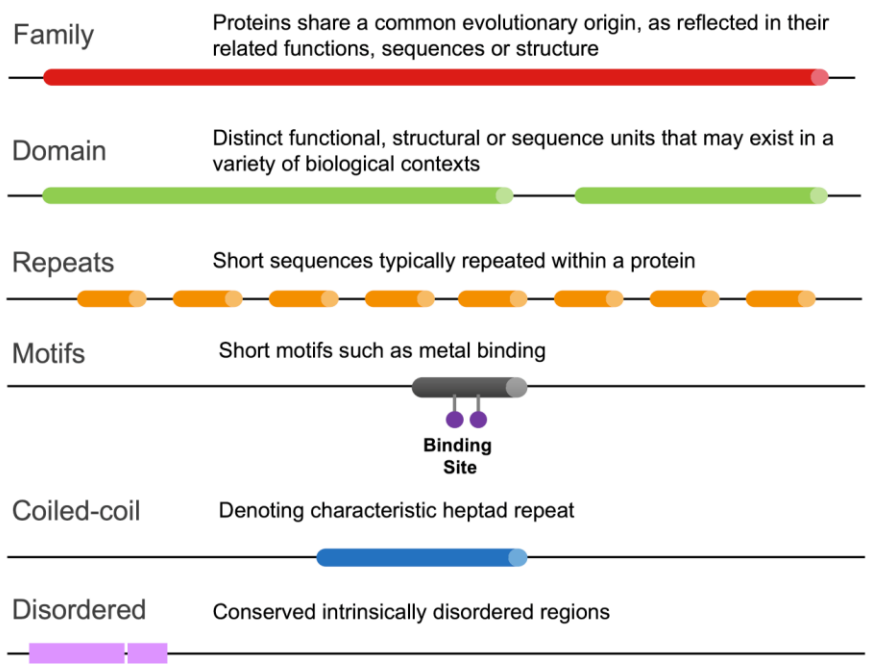
UniProt数据库主页与检索页



生物大分子数据库（蛋白质家族与结构域）

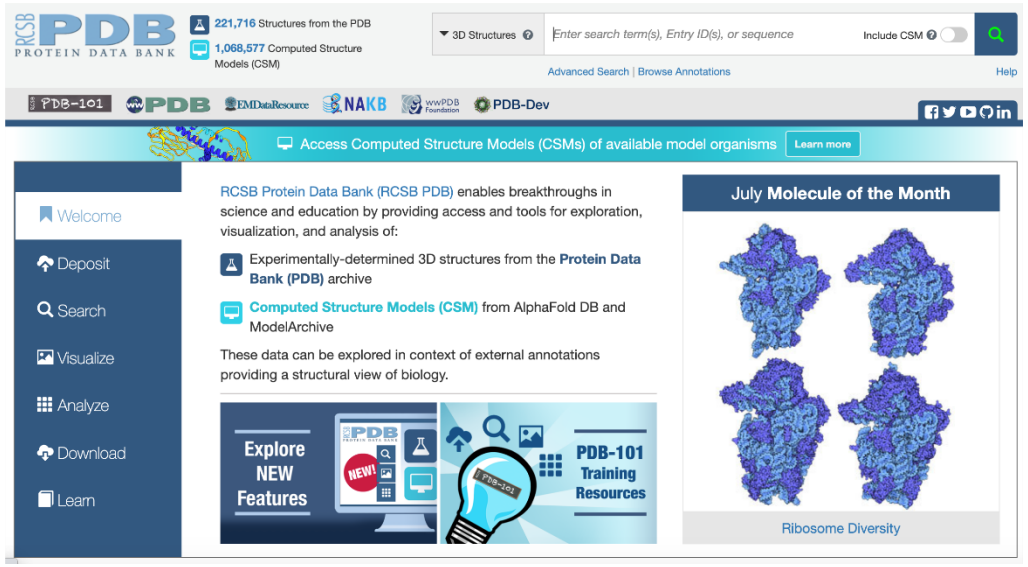


Pfam（已被整合至InterPro数据库中）数据库主页（2024.7）

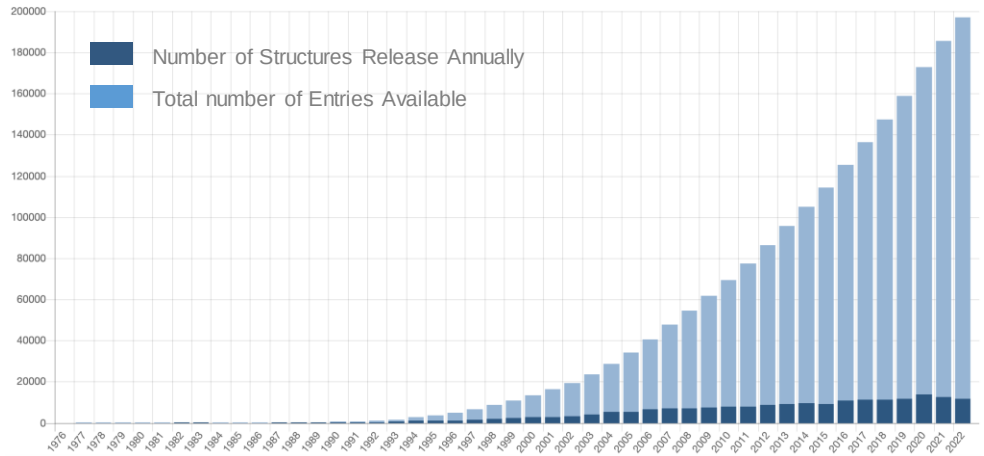


Pfam蛋白质注释内容

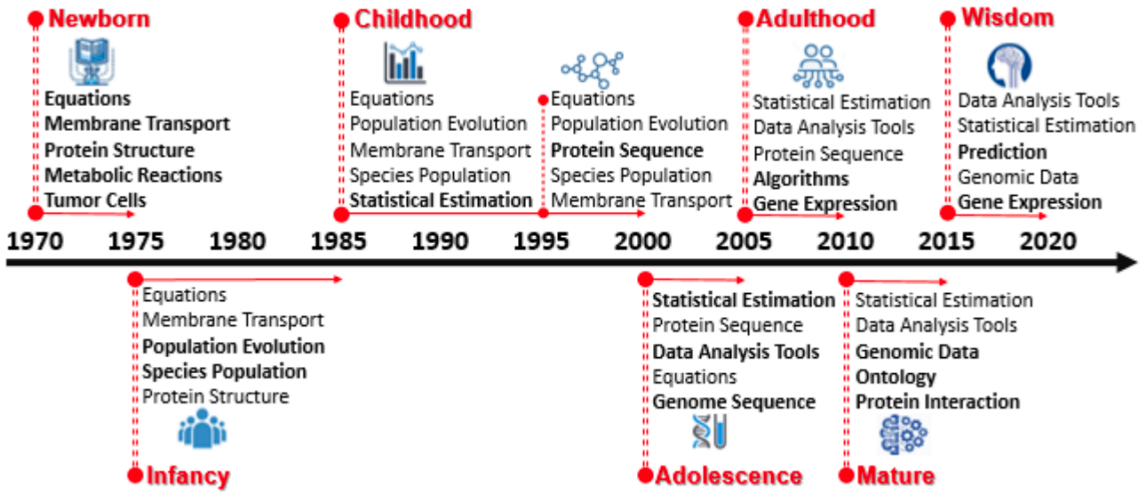
生物大分子数据库（蛋白质结构）



PDB数据库首页（2024.7）



蛋白数据库数据量变化情况

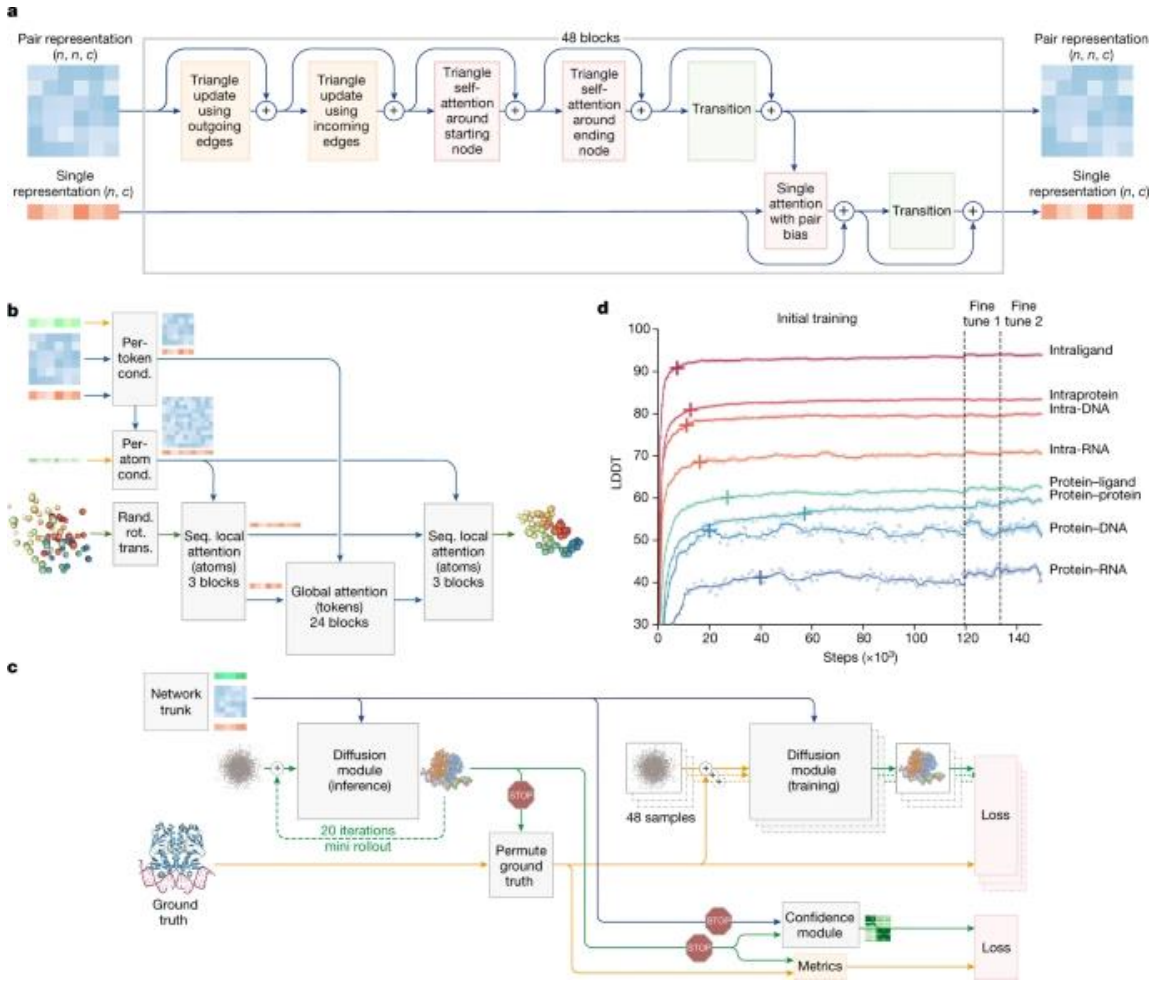


生物信息学七阶段与不同阶段的热点话题

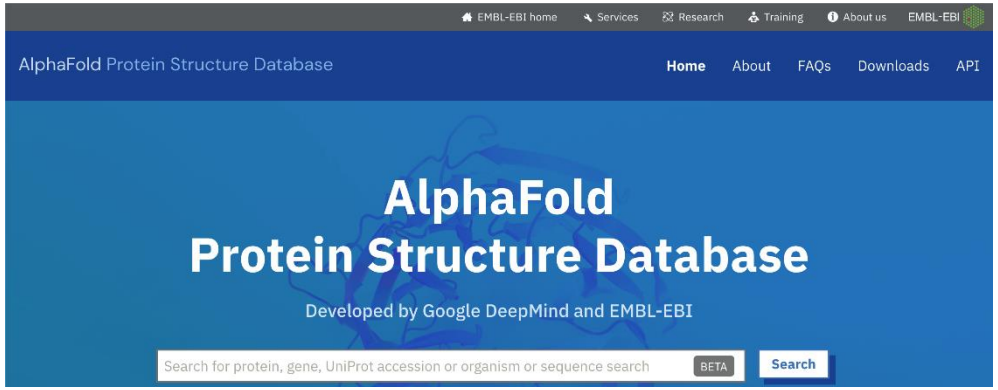
Statistical Estimation	**7.59**	Ontology	4.56	Membrane Transport	4.04
Equations	7.50	**Prediction**	**4.41**	Genome Sequence	3.93
Data Analysis Tools	**6.87**	Protein Interaction	4.38	Species Population	3.73
Algorithms	**5.17**	Genomic Data	4.32	Signaling Regulatory	3.38
Protein Sequence	4.99	Protein 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)

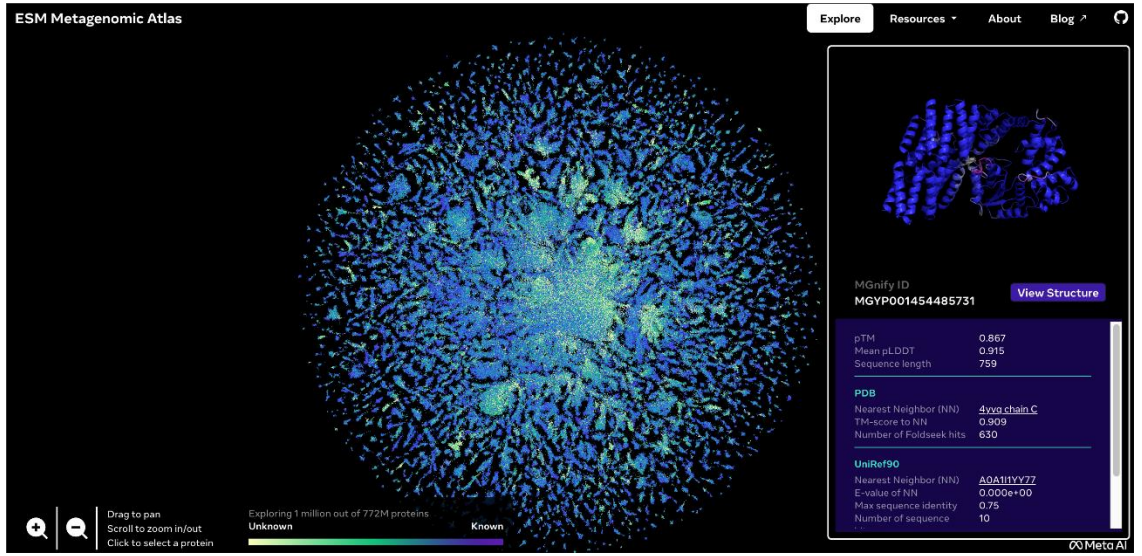
生物大分子数据库（蛋白质结构）



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

物种专题数据库

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

生物数据库发展现状、挑战与趋势

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

生物数据库：高质量数据，高质量AI语料库

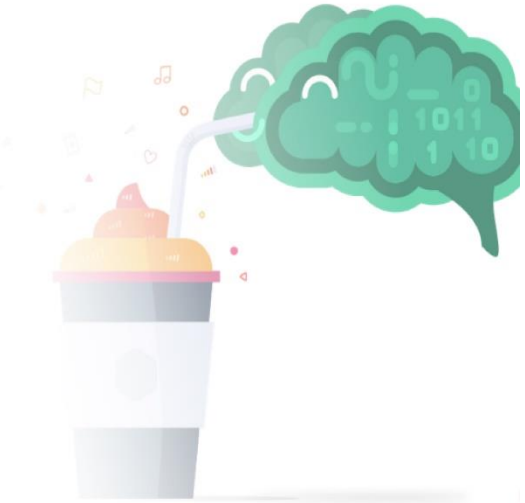
THOUGHT LEADERS

Importance of Data Quality in AI Implementation

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

PredictHQ

Data is to AI as Food is to Humans



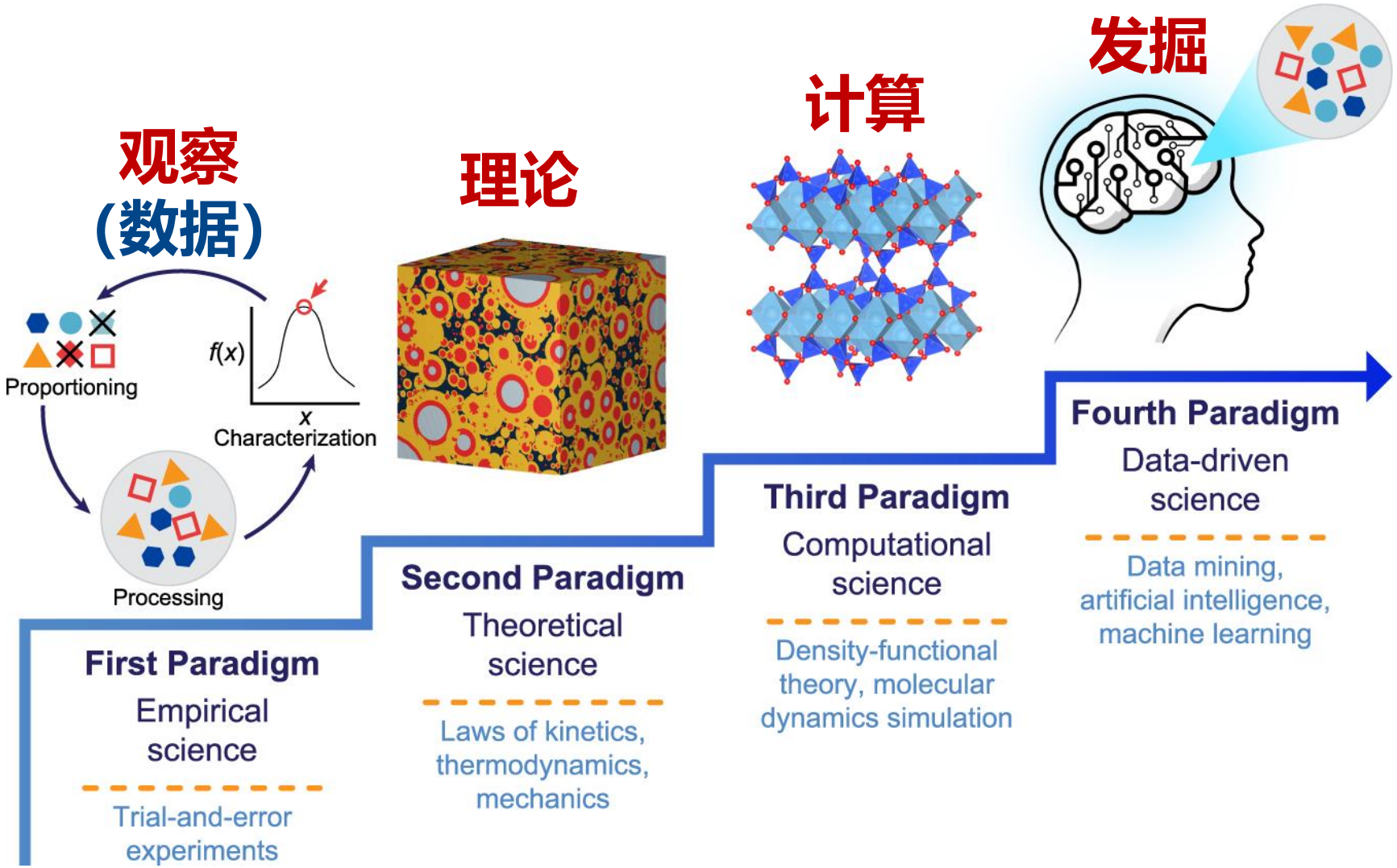
高质量数据：AI-ready语料库

- 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/

生命科学主要在哪个范式阶段?



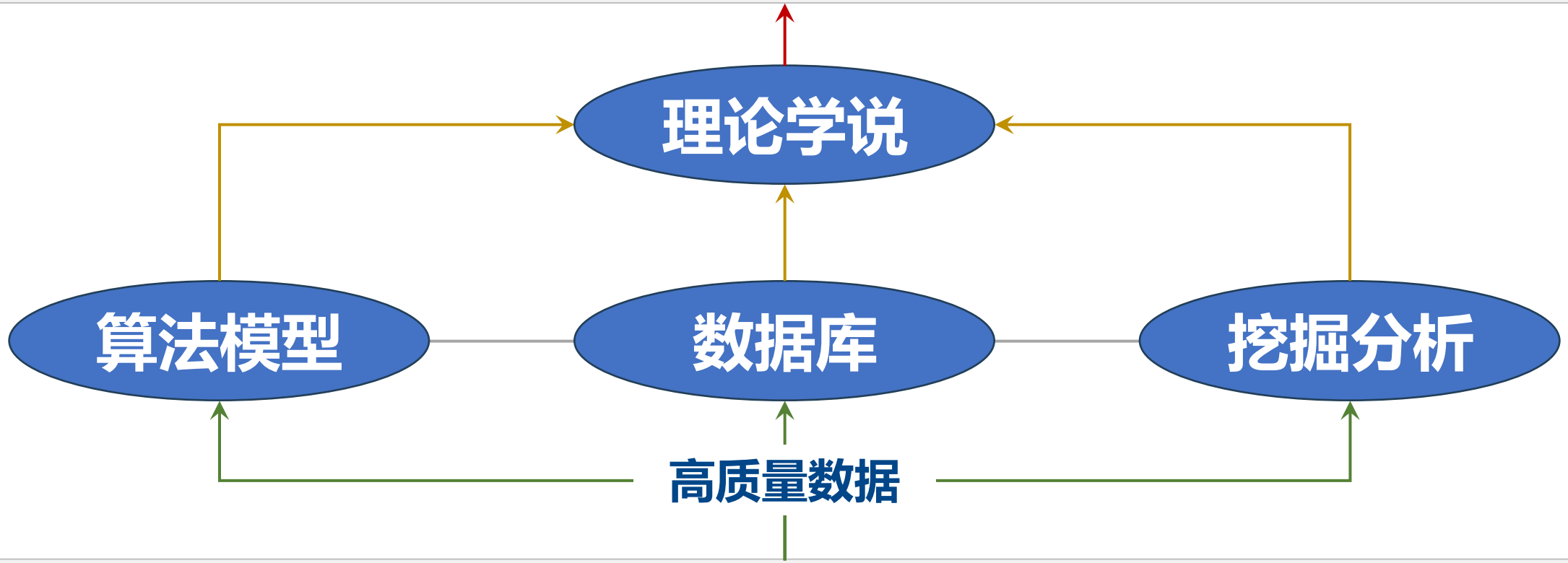
npj Comput Mater 2022

生物学重要理论：历久弥坚，历史沉淀

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

生物信息学：从数据到理论研究范式变革

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

总结

- 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**

共享约束需知：使用条款

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！