

BIG Search

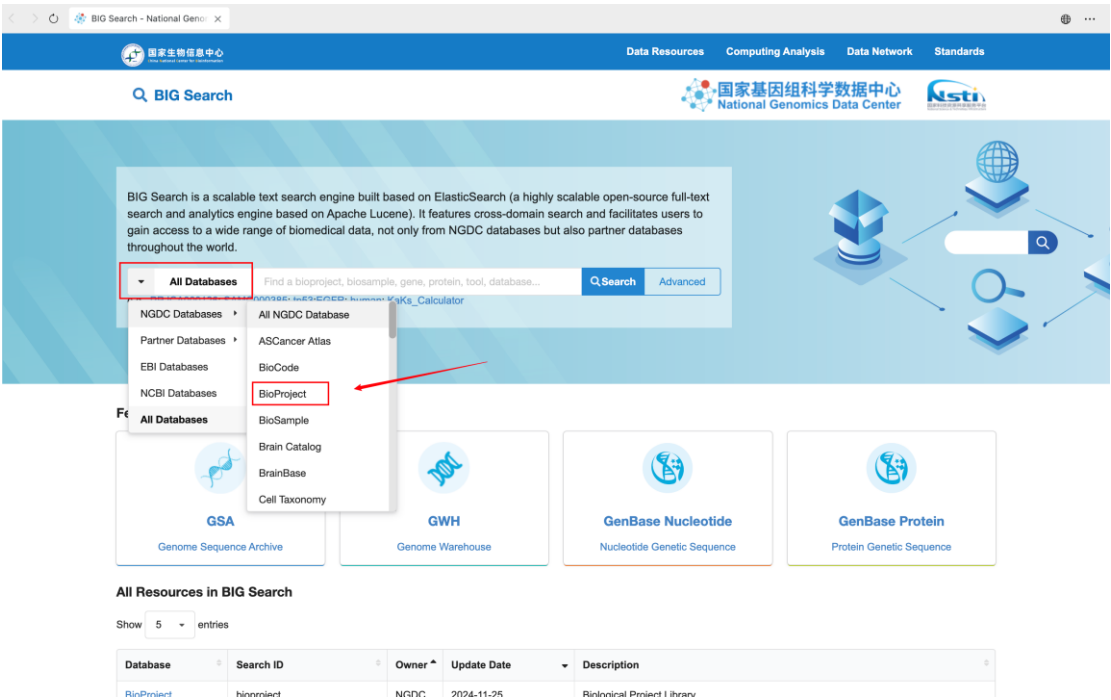
BIG Search 跨库检索服务（BIG Search）是通过建立面向多层次生物数据的通用索引标准，基于 Elasticsearch 等前沿信息技术，实现数据高效可扩展的一站式搜索引擎。目前 BIG Search 已整合中心和华中科技大学、四川大学、北京大学、中国医学科学院、华中农业大学、哈尔滨医科大学、中科院北京生命科学研究院等多个合作结构的 140 余个数据资源库，集成数据索引 15.4 亿条，索引存储达到 1.6TB，同时还整合了 EBI 和 NCBI 的数据资源，能够实现高性能、灵活的全文索引和检索，帮助用户更有效、快速、准确地获取知识。

一、检索

系统首页检索入口

（1）具体数据库检索

用户可以预设检索数据库，并输入关键词进行指定数据库资源检索。



点击 search 按钮后会跳转至对应数据库的检索结果页面，用户可以在左侧筛选框设置详细的筛选条件点击 Filter 按钮筛选。右侧可以浏览检索列表，点击对应名称查看详情，用户也可以选择多个所需的数据资源，点击右上角 Save 按钮进行数据下载。

Search Results for **single cell** > **BioProject**

Sort by: Best match ☐ F ☐ 1F

Filter Reset

Project Source

- ☐ SRA (20322)
- ☐ GSA (2070)
- ☐ ERA (607)
- ☐ DRA (177)

Data Types

- ☐ Single cell sequencing (697)
- ☐ Transcriptome or Gene expression (318)
- ☐ Transcriptome or Gene expression Single cell sequencing (150)
- ☐ Transcriptome or Gene expression Raw sequence reads (103)
- ☐ Raw sequence reads (102)
- ☐ Transcriptome or Gene expression Raw sequence reads Single cell sequencing (94)
- ☐ Raw sequence reads Single cell sequencing (64)

Found: 1 to 10 of 31,388 entries.

PRJCA010092

Accession: PRJCA010092

Title: Single-cell transcriptome of PAM

Species:

Data Type: BioProject

Description: Single-cell transcriptome of pulmonary alveolar microlithiasis

Basic Information: PRJCA010092; Single cell sequencing; Medical

Project Source: GSA

Samples:

GSAs:

Data Types: Single cell sequencing

Scope: Single cell

Relevance: Medical

Update Date: 2024-11-25 04:00:00

PRJCA013345

Show 10 entries

Search details

single cell

Save

(2) 全局检索

用户可以直接在检索框中输入关键词，点击 **search** 按钮进行全局资源检索。

检索后系统会对关键词进行全局检索。用户可以在左侧选择数据库的归属，页面主体会列举出不同数据库中与该关键词相关的具体数据库和该数据库中对应的资源数量，点击具体 **Database** 名称可以跳转至该数据库的首页，点击 **Records Number** 具体数据可以跳转至该数据中的检索结果列表。

Search Results for **single cell** > **All Databases**

17,879,995 records from 29 NGDC databases.

Show 10 entries

Filter:

Database	Records Number	Description
AScancer Atlas	2	A comprehensive knowledgebase of alternative splicing in human cancers
BioCode	161	Archive Bioinformatics Codes for Open Source Projects
BioProject	31,388	Biological Project Library
BioSample	186,202	Biological Sample Library
Cell Taxonomy	80	Cell Taxonomy is a curated repository of cell types with multifaceted characterization.
eLMSG	2	An eLibrary of Microbial Systematics and Genomics
Genbase Nucleotide	983,788	a collection of nucleotide sequences from several sources
Genbase Protein	14,548,317	a collection of protein sequences from several sources
Gene Expression Nebulas	752	A data portal of transcriptomic profiles across multiple species

NGDC Databases (17,879,995)

Partner Databases (300,400)

EBI Databases (3,267,018)

NCBI Databases (14,969,535)

AlphaFold (0)

(3) 高级检索

点击 **Advanced** 按钮可进行高级检索。高级检索页面用户可以通过 **Choose Resource** 筛选检索数据库，在 **Build Query** 可以拼接关键词匹配的数据字段，例如选择“**Entry Title**”点击右侧 **AND** 按钮会拼接基于 **Title** 的关键词查询。如果有多个筛选条件可通过右侧按钮进行“**AND**”、

“OR”、“NOT” 拼接。

Advanced Search

Choose Resource
AScancer Atlas

Build Query
Entry Title: gene
rna AND title:"gene"
+ AND

Search **Clear**

NGDC
National Genomics Data Center
86-10-84097298
ngdc@big.ac.cn
Policies and Disclaimers

Research & Resources
Databases
Tools
Standards
Topics & Projects
Publications
Annual Reports

Featured
Database Commons
GenBase
GSA
GWH
GVM
GEN
RCov19

Alliance & Collaboration
BHBD
P10K
Partners
Collaborations
Funding
Contact Us
Join Us

Conference & Outreach
Conferences
Education

About
About NGDC
People
Mission
Advisory Board
Organizational Structure
History

© 2024 China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences

二、Big Search 中的特色资源库，点击可快速跳转至资源库

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.

Featured Resources

- GSA**
Genome Sequence Archive
- GWH**
Genome Warehouse
- GenBase Nucleotide**
Nucleotide Genetic Sequence
- GenBase Protein**
Protein Genetic Sequence

All Resources in BIG Search

Show 5 entries

Database	Search ID	Owner	Update Date	Description
BioProject	bioproject	NGDC	2024-11-25	Biological Project Library
GSA	gsa	NGDC	2024-11-25	Genome Sequence Archive
GSA for Human	hra	NGDC	2024-11-25	Genome Sequence Archive for Human
OMIX	omix	NGDC	2024-11-25	OMIX

三、Big search 覆盖的数据资源浏览

点击“show entries”可以设置每页展示的数据库数量，默认为 5 个。在数据库列表点击具体的数据库名称可跳转至详情页面。

ngdc.cncb.ac.cn

BIG Search - National Genomics...

Search ASCancer Atlas - BIG Se...

Search BioCode - BIG Search

Search BioProject - BIG Search

Search BioProject - BIG Search

Search ASCancer Atlas - BIG Se...

国家生物信息中心

Data Resources

Computing Analysis

Data Network

Standards

BIG Search

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

KSTU

AScancer Atlas

title:"gene" NOT title:"human"

Q Search

Advanced

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

Search Results for **title:"gene" NOT title:"human"** > **AScancer Atlas**

Sort by: Best match ☒ IF ☐ 1F

Save

Filter

Reset

☐ Cancer Name

☐ Colon Cancer (39)

☐ Ovarian Cancer (11)

☐ Hepatocellular Carcinoma (10)

☐ Breast Cancer (4)

☐ Non-small Cell Lung Cancer (3)

☐ Spinal Muscular Atrophy (2)

☐ Cervical Cancer (1)

☐ Colorectal Cancer (1)

☐ Glioblastoma (1)

☐ Kidney Cancer (1)

☐ Melanoma (1)

☐ Nasopharyngeal Carcinoma (1)

☐ Gene Name

☐ BHMT (5)

☐ BHMT2 (5)

Found: 1 to 10 of 75 entries.

Show 10 entries

SMN2_chr5_+_ES_69366468:69366578:69372348:69372401:69372846:69373419

Accession: SMN2_chr5_+_ES_69366468:69366578:69372348:69372401:69372846:69373419

Title: Regulation of Survival Motor Neuron Gene Expression by Calcium Signaling

Species: Homo sapiens

Data Type: as_event

Description: Spinal muscular atrophy (SMA) is caused by homozygous survival of motor neurons 1 (SMN1) gene deletion, leaving a duplicate gene, SMN2, as the sole source of SMN protein. However, a...

Basic Information:

Event ID: SMN2_chr5_+_ES_69366468:69366578:69372348:69372401:69372846:69373419

Cancer Name: Spinal Muscular Atrophy

Gene Name: SMN2

Ensembl ID: ENSG00000205571

Canonical Transcript ID: SMN2-001

Ct Ensembl ID: ENST00000380743

Ct RefSeq ID: NM_017411

Splice Variant ID: -

Search details

title:"gene" NOT title:"human"