

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。

The screenshot shows the BIG Search interface. At the top, there is a navigation bar with tabs for Data Resources, Computing Analysis, Data Network, and Standards. Below the navigation bar is a search bar with the placeholder "Find a bioproject, biosample, gene, protein, tool, database...". To the right of the search bar are "Search" and "Advanced" buttons. A dropdown menu titled "All Databases" is open, showing categories: NGDC Databases, Partner Databases, EBI Databases, and NCBI Databases. Under "All Databases", a sub-menu for "BioProject" is highlighted with a red box and a red arrow pointing to it. Other options in this sub-menu include BioSample, Brain Catalog, BrainBase, and Cell Taxonomy. To the right of the search area, there is a graphic illustrating the search process with a globe, a magnifying glass, and books. Below the search area, there are three cards for GSA (Genome Sequence Archive), GWH (Genome Warehouse), and GenBase Nucleotide. At the bottom, there is a table titled "All Resources in BIG Search" with columns for Database, Search ID, Owner, Update Date, and Description. One entry is visible: RunProject, hinnmact, NGDC, 2024-11-25, Biographical Project Library.

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。

The screenshot shows the search results for 'single cell' under the 'BioProject' category. The results list 31,388 entries. A specific entry, PRJCA010092, is highlighted with a red box. The page includes filters for 'Project Source' (SRA, GSA, ERA, DRA) and 'Data Types' (Single cell sequencing, Transcriptome or Gene expression, etc.). A 'Save' button is located in the top right corner.

(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。

The screenshot shows the search results for 'single cell' across all databases. It lists 17,879,995 records from 29 NGDC databases. A specific entry, BioProject, is highlighted with a red box. The page includes a sidebar for 'All Databases' and a 'Records Number' column.

Database	Records Number	Description
ASCCancer 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

(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**”或“**NOT**”操作。

“OR”、“NOT” 拼接。

Advanced Search

Choose Resource
ASCancer Atlas

Build Query
Entry Title gene

rna AND title:"gene"

+ AND

Search Clear

Research & Resources

Featured

Alliance & Collaboration

Conference & Outreach

About

National Genomics Data Center

86-10-84097298

ngdc@big.ac.cn

Policies and Disclaimers

Database Commons

GenBase

BHBD

P10K

GSA

Partners

GWH

Collaborations

GVM

Funding

GEN

Contact Us

Join Us

RCoV19

Conferences

Education

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.

Select Databases gene

PRJCA000126; SAMC000385; tp53;EGFR; human; KaKs_Calculator

Featured Resources

GSA

GWH

GenBase Nucleotide

GenBase Protein

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。

The screenshot shows the BIG Search interface on a web browser. At the top, there are 'Featured Resources' for GSA, GWH, GenBase Nucleotide, and GenBase Protein. Below this is a table titled 'All Resources in BIG Search' with columns: Database, Search ID, Owner, Update Date, and Description. The table lists BioProject, GSA, GSA for Human, OMIX, and BioSample entries. A red box highlights the 'Show 5 entries' dropdown, and another red box highlights the page navigation buttons at the bottom right. Below the table, a section titled 'Query syntax' contains a table with operators (AND, OR, NOT, "", *, ()), their explanations, examples, and descriptions.

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
BioSample	biosample	NGDC	2024-11-23	Biological Sample Library

Showing 1 to 5 of 147 entries

Operator	Explain	Usage	Example	Description
AND	In addition to	term1 AND term2	human AND rice	Search entries where both <i>human</i> and <i>rice</i> occur
OR	Equivalence	term1 OR term2	human OR rice	Search entries where either <i>human</i> or <i>rice</i> occur

四、查询语法

系统支持 AND、OR、NOT 等多种查询语法，并且查询语法可拼接使用，用户可以在检索模块通过查询语法进行高级检索，例如：title：“gene” NOT title：“human”会筛选标题中含有“gene”并且不含有“human”的数据资源

The screenshot shows the BIG Search interface on a web browser. At the top, there are multiple tabs for different search modules. Below the tabs is a table with columns: Human, term, NGDC, 2024-11-25, and Genome Sequence Archive for Human. The table lists OMIX and BioSample entries. A red box highlights the 'Show 5 of 147 entries' dropdown, and another red box highlights the page navigation buttons at the bottom right. Below the table, a section titled 'Query syntax' contains a table with operators (AND, OR, NOT, "", *, (), ()^n), their explanations, examples, and descriptions.

Human	term	NGDC	2024-11-25	Genome Sequence Archive for Human
OMIX	omix	NGDC	2024-11-25	OMIX
BioSample	biosample	NGDC	2024-11-23	Biological Sample Library

Showing 1 to 5 of 147 entries

Operator	Explain	Usage	Example	Description
AND	In addition to	term1 AND term2	human AND rice	Search entries where both <i>human</i> and <i>rice</i> occur
OR	Equivalence	term1 OR term2	human OR rice	Search entries where either <i>human</i> or <i>rice</i> occur
NOT	Exclusion	term1 NOT term2	tp53 NOT human	Search entries contain <i>tp53</i> but not <i>human</i>
" "	Exact match	"term1 term2"	"gene expression"	Search entries contain the exact phrase <i>gene expression</i>
*	Wildcard	term*	trans*	Search entries start with <i>trans</i>
()	Grouping	(term)	(human OR rice) AND expression	

The screenshot shows the NGDC BIG Search interface. The search bar at the top has the query "title:'gene' NOT title:'human'" entered. Below the search bar, there is a note: "e.g., PRJCA000126; SAMC000385; tp53;EGFR; human; KaKa_Calculator". The main area displays search results for ASCancer Atlas. A sidebar on the left lists filters for "Cancer Name" and "Gene Name". The "Cancer Name" filter includes options like 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), and Nasopharyngeal Carcinoma (1). The "Gene Name" filter includes options like BHMT (5) and BHMT2 (5). The results table shows one entry: SMN2_chr5_+_ES_69366468:69366578:69372348:69372401:69372846:69373419. The entry details include: 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 Results for **title:"gene" NOT title:"human"** > **ASCancer Atlas**

Sort by: Best match IF 1F

Save

Found: 1 to 10 of 75 entries.

Show 10 entries

title:"gene" NOT title:"human"

- 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)

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: -.