



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

# 组学原始数据汇交与共享

王彦青

2024年10月26日

# 提纲



## GSA Family介绍



## GSA数据提交及共享



## GSA-Human数据提交及共享

# 组学原始数据管理体系 (GSA Family)



为组学原始数据统一汇交、存储、管理、共享的公共平台，保障国家数据安全



# 组学原始数据归档库GSA



组学原始数据汇交、存储、管理、发布与归档系统

equivalent

SRA  
Sequence  
Read Archive  
@NCBI

DRA  
DDBJ Read  
Archive  
@DDBJ

ENA  
European  
Nucleotide  
Archive  
@EBI

International Nucleotide Sequence Database  
Collaboration (INSDC)

The International Genome Sequence archive databases

Wang, Yanqing et al. *Genomics Proteomics Bioinformatics*, 2017  
Chen, Tingting et al. *Genomics Proteomics Bioinformatics*, 2021

# 国际数据整合

整合国际组学数据，实现SRA数据的本地镜像、同步和共享



# GSA-Human: 人类遗传资源组学原始数据管理

The screenshot shows the GSA-Human database interface. At the top, there are tabs for Data Resources, Computing Analytics, Data Network, and Standards. Below that is a search bar with 'All Databases' and 'Find a GSA-Human dataset' options. The main content area includes:

- 人类遗传资源组学原始数据提交指南:** A section with a 'Study' icon, listing:
  - 人类遗传资源组学原始数据归档 (GSA-Human) GSA-Human平台，专门用于人类遗传资源组学数据归档。GSA-Human的主要特征包括：
    - 专用于人类遗传资源组学数据归档。
    - 提供数据存储和访问服务。
    - 提供全面的数据请求服务。
- 提交:** A button to submit data.
- 浏览:** A button to view datasets.
- 申请数据:** A button to apply for data.
- 数据统计:** A chart showing the number of datasets by year (2018-2021).
- 统计数据:** A table with study details:

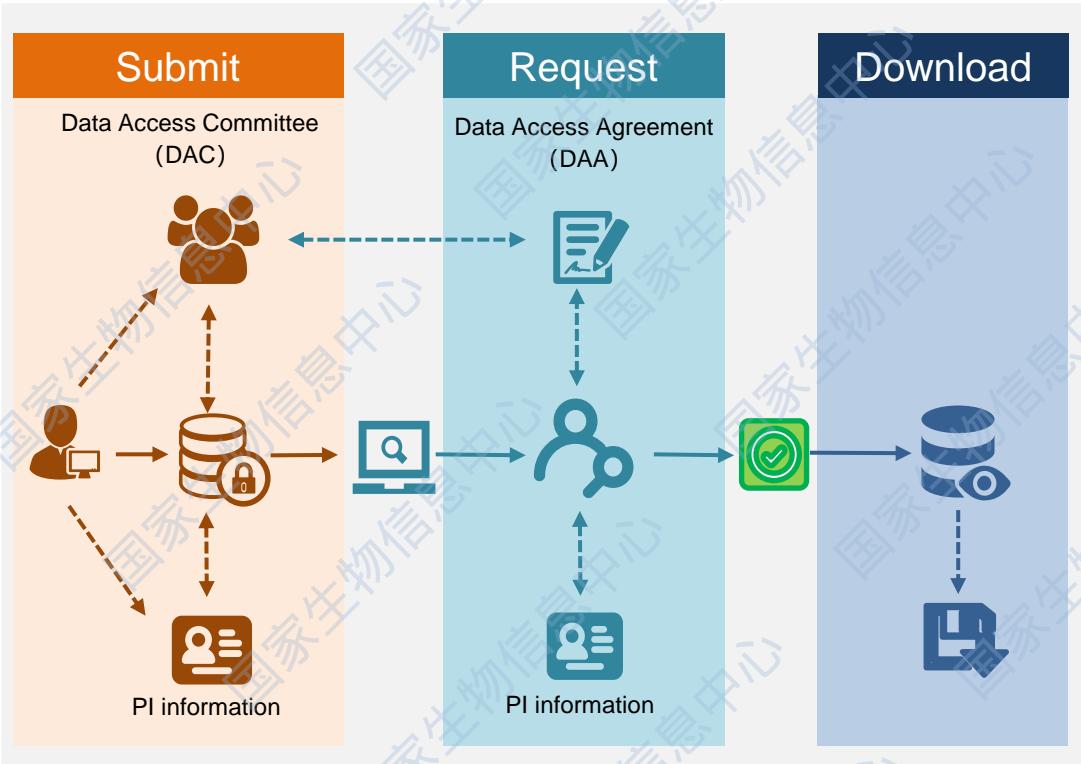
研究 (Study)	4811
个体 (Individual)	87729
样本 (Sample)	849727
实验 (Experiment)	65472
评估情况 (Rate)	89.17%
- 帮助和支持:** Contact information: Email: gsa@genomics.org.cn, QQ群: 948110661, 电话: +86 (10) 8406-7942. A note: If you have any questions about GSA-Human's features, or if you find any problems, please feel free to contact us. We will respond as soon as possible.
- 最新文章摘要:** A list of recent publications:
  - H1A000094: Concurrent de novo MAGeCK modeling and...
  - H1A000093: Tcf7 cell-sorted TWISTAK provides...
  - H1A000092: Single-cell sequencing of white blood...

Chen, Tingting et al. *Genomics Proteomics Bioinformatics*, 2021

Zhang, Sisi et al. *Yi Chuan*, 2021

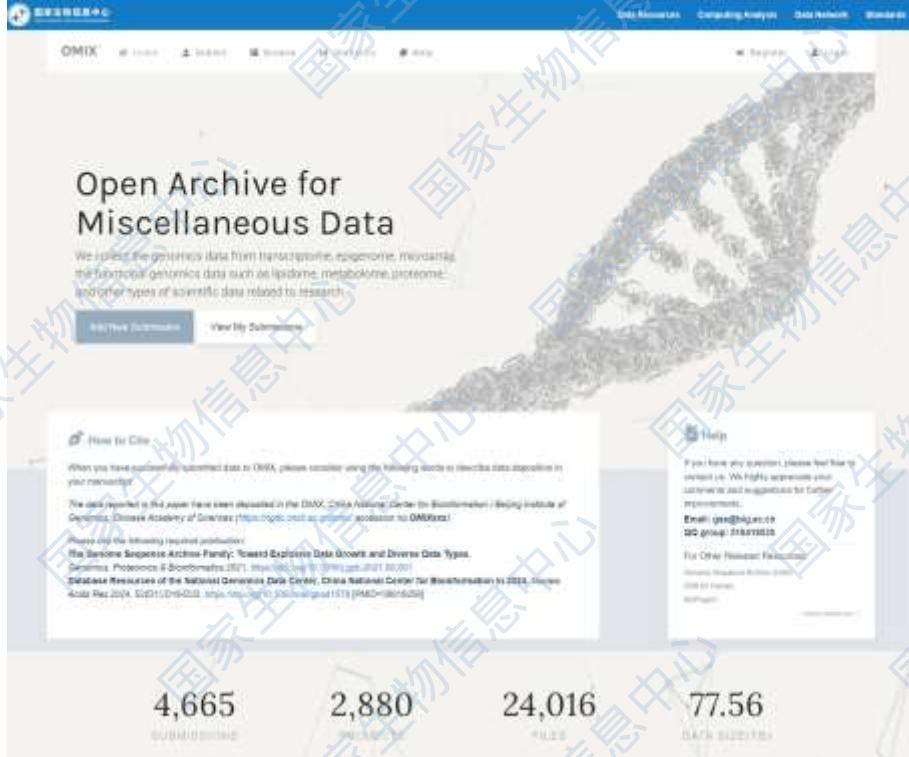
- 聚焦**人类遗传资源组学原始数据**汇交、归档、管理、可控共享
  - 对标NCBI的dbGap和EBI的EGA
  - 对接国家人遗数据管理
  - 管理>58万人数据
  - 归档数据量**38.13PB**
  - 支持**883篇**文章发表

# GSA-Human系统核心功能要素



- 支持**5种**研究类型数据汇交
- 支持**2种**文件上传方式
  - Aspera和FTP
- 支持**2种**访问方式
  - 受控访问 (Controlled-access)
  - 公开访问 (Open-access)
- 受控访问采用申请审核制
  - 数据管理委员会 (Data Access Committee, DAC)
  - 审核数据访问申请
  - 授予访问权限
- PI账号提交和申请数据：数据责任人

# 多元数据归档库 OMIX



汇交、存储、共享 来自生物与医学  
领域研究的多元数据

BMC Series

SPRINGER NATURE

PLOS

frontiers

MDPI



Taylor & Francis Group  
an informa business

WILEY

SAGE journals



ELSEVIER

OXFORD  
UNIVERSITY PRESS

CelPress

ACS Publications

Most Trusted. Most Cited. Most Read.

AACR  
American Association  
for Cancer Research



AMERICAN  
SOCIETY FOR  
MICROBIOLOGY



Cold  
Spring  
Harbor  
Laboratory

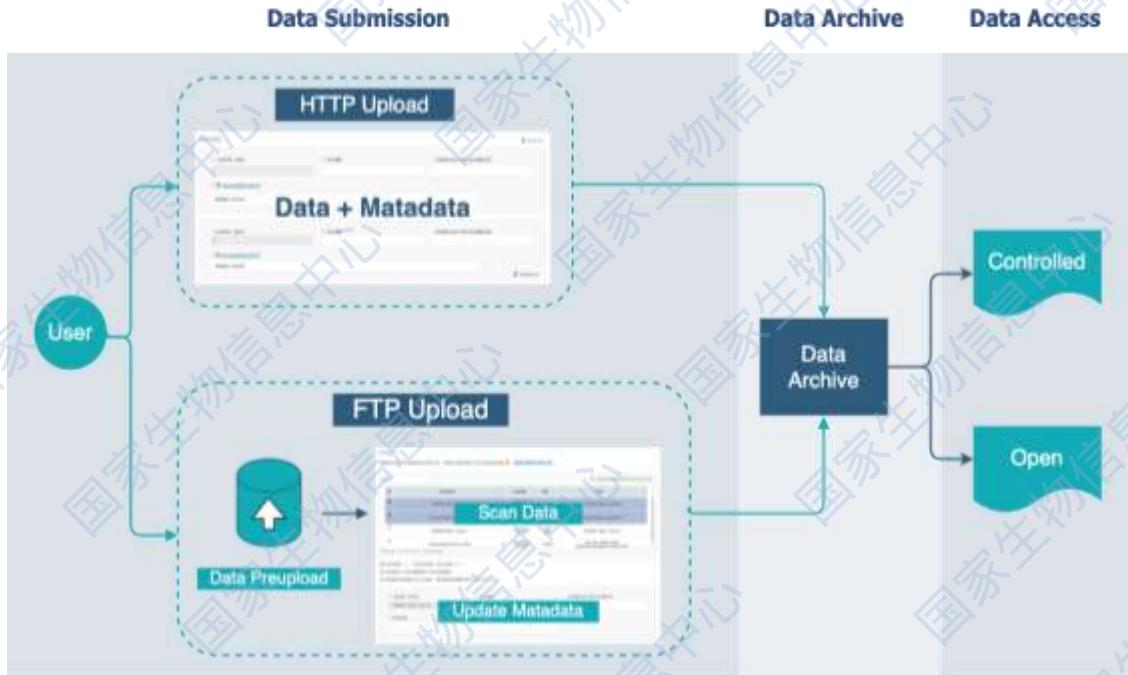
- 已通过 FAIRSharing 和 re3data.org 认证
- 获得 40 个出版集团的 250 个期刊认可，支撑文章 604 篇



CNCB-NGDC

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

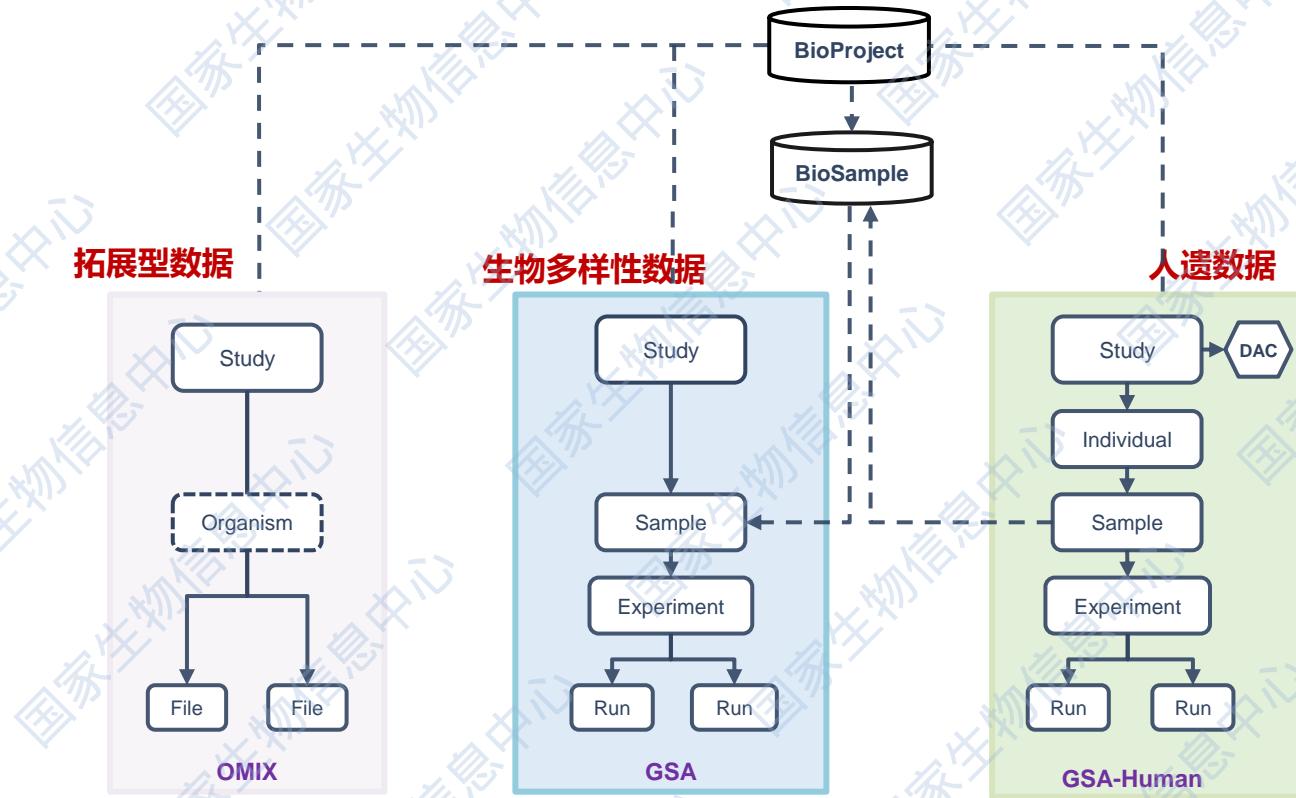
# 应用灵活可扩展的管理模式，支撑多模态数据汇交共享



- 支持**10个大类, 31个小类**, 超过**70种**格式数据汇交
- 支持**2种**数据访问策略
  - 开放访问
  - 受控访问** (仅对人遗数据)
- 支持**2种**文件上传方式
  - HTTP在线上传
  - FTP预上传 (>1 TB 可协助上传)

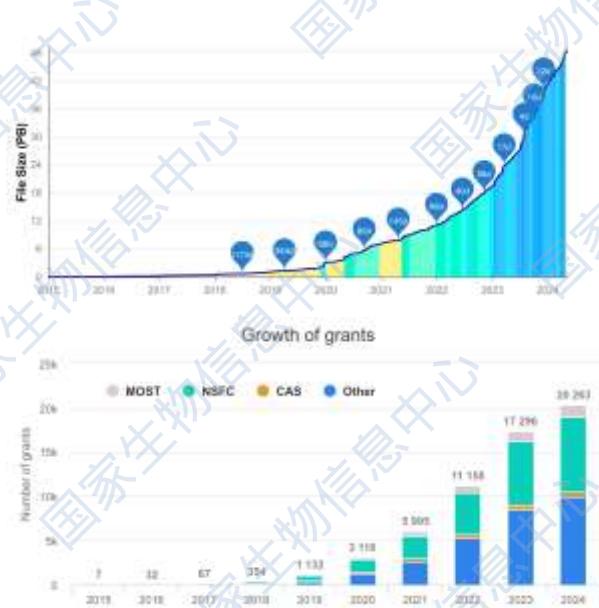


# GSA Family数据模型



# 运行效果

- GSA Family得到生命领域几乎所有国际主流期刊认可
- 汇交数据量超 56 PB，支撑发表文章3900 篇
- 汇交数据量**每12个月翻一番**，超过“摩尔定律”增速，**增长1PB数据用时最短4天**



>120万 独立IP

>17亿 次下载

>190 国家/地区

>9PB 下载量

# 获得国际著名出版商认可

**SPINGER NATURE**

Authors

Research data

## Research data policy



### Biological sciences repository examples

› Imaging

› Nucleic acid sequence and 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

Any INSDC member repository

Genome Sequence Archive (GSA)

dbSNP (human variations less than 50bp)  
clinvar (human variations greater than 50bp)  
European Variation Archive (EVA) (all species)

Genomic Sequence Archive for Human (Human variation)

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

**ELSEVIER**

About Elsevier Products & Solutions Services

Home > Journals > Data in Brief > Policies and Guidelines > Public repositories to store and find data

## Data in Brief

Submit your Paper View Article

Guide for authors Track your paper

### Public repositories to store and find data

2. Specialized Repositories are recommended. Please note that this list is not exhaustive. We encourage the use of Re3data and FAIRsharing to find additional specialized repositories in a specific discipline.

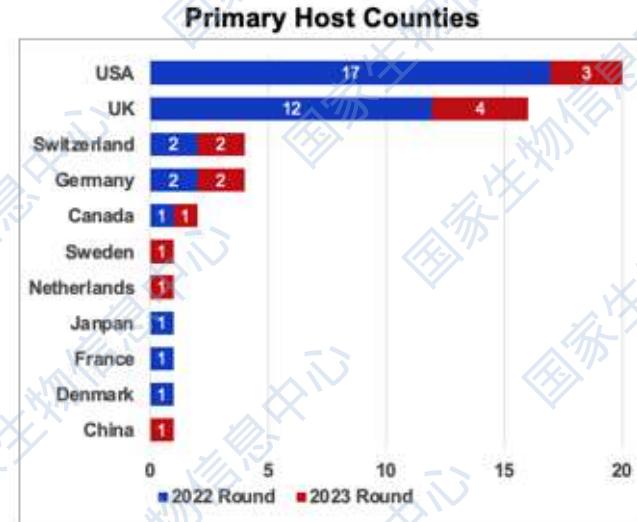
Discipline	Type of data	Repository name	Cost
Genetic variation data	Raw sequencing data, genome assemblies, annotated sequences, and sample metadata	INSDC repositories Genome Sequence Archive (GSA)	Free
	dbSNP dbVar ClinVar European Variation Archive (EVA)	Genome Sequence Archive for Human	Free

<https://www.journals.elsevier.com/data-in-brief/policies-and-guidelines/public-repositories-to-store-and-find-data>

# GSA入选全球核心数据资源 (GCBR)

GSA入选由国际生物数据联盟 (Global Biodata Coalition, GBC) 发起的**全球核心生物数据资源** (GCBR)，是我国目前**唯一入选**的数据库

The screenshot shows the Global Biodata Coalition website. At the top, there is a logo for "GLOBAL BIODATA COALITION". Below it, a navigation bar includes "NAME", "OVERVIEW", "WHAT WE DO", "MEMBERSHIP", and "NEWS & RESOURCES". Under "NAME", it lists "Host Countries: Primary, Additional". A section for "gromAD" shows "USA" under "Primary". The "OVERVIEW" section contains a detailed description of gromAD, mentioning its goal of aggregating and harmonizing genome sequencing data from various projects. Below this, a red-bordered box highlights the "GSA: Genome Sequence Archive" entry, which describes it as a data repository for collecting, archiving, managing, and sharing raw genome sequence data. Another red-bordered box highlights the "GWAS Catalog" entry, which provides a database of SNP-trait associations. The bottom of the page has links for "UK, USA" and "GWAS Catalog".



# 全面保障我国人类遗传资源信息安全管理与共享应用

承担国家人类遗传资源信息汇交备份任务，遵循“人遗条例”，制定专业化数据访问机制“申请审核制”，开发人类遗传资源信息备份管理与发布共享的一体化平台

中华人民共和国科学技术部

国科办函社〔2022〕316号

科技部办公厅关于委托中国科学院北京基因组研究所（国家生物信息中心）承担国家人类遗传资源信息管理备份工作的函

中科院办公厅公开：

为简化国家人类遗传资源信息备案流程，确保备份备案与发表共享数据的一致性，按照人类遗传资源信息汇交合规、共享与再利用、经科技部研究，决定委托中国科学院北京基因组研究所（国家生物信息中心）承担国家人类遗传资源信息管理备份任务。

委托的主要任务包括建设现有的人类遗传资源信息备份平台（<https://202.108.211.75/>）和国家生物信息中心人类遗传资源管理。

人类遗传资源信息  
管理备份平台

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



促进人遗数据资源安全共享

已生成备份号: 3335
已备份数据集: 6388
汇交数据总量: 2.29 PB
已发布受控数据集: 1298
总申请次数: 1285
获得授权次数: 477 (境外102)
下载数据量: 574.9 TB



国家互联网应急中心

国家基因组科学数据中心 - 汇交共享、备份

科技部人遗管理办 - 备案审查



# 提纲

—

—

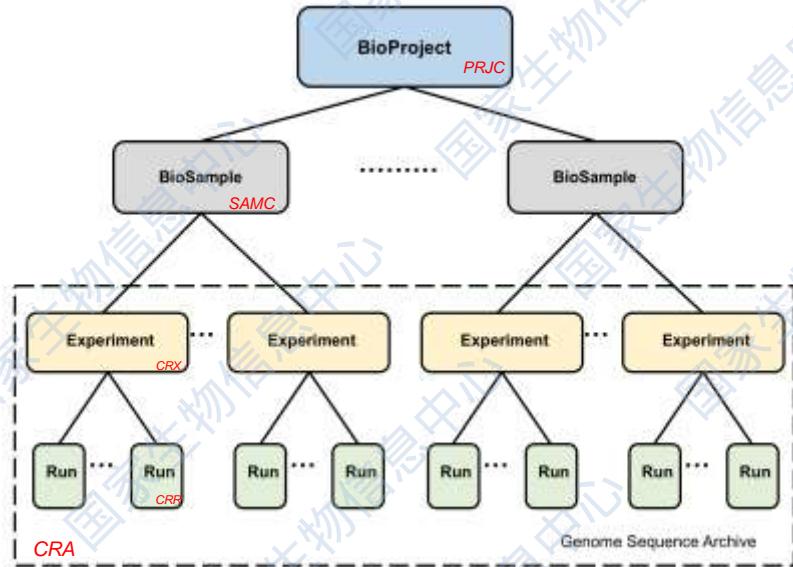
—

## GSA Family介绍

## GSA数据提交及共享

## GSA-Human数据提交及共享

# GSA数据模型及支持的测序平台



数据模型与数据标准：参考INSDC标准

Platform	Data File Type
ILLUMINA	FASTQ, BAM
PACBIO_SMRT	HDF5, BAM, FASTQ
Oxford Nanopore	FASTQ, BAM, Fast5
COMPLETE_GENOMICS	Complete_Genomics_native, FASTQ, BAM
BGISEQ	FASTQ, BAM
BIONANO	BNX
ABI_SOLID	FASTA/QUAL, FASTQ, BAM
...	...

# 支持的样本类型

## Pathogen

用于与公共卫生相关的病原体样本。

- Clinical or host-associated

- Environmental, food or other

## Microbe

用于能具体写出物种名称的微生物样本，但不包括真菌或病毒样本。宏基因组数据选择的Metagenome/Environmental Sample (GSC MIMS unsupported)或Metagenome/Environmental Sample (GSC MIMS compliant)

## Animal

用于模式生物或动物的多组织样本或细胞系样本，如大脑、小鼠、黑猩猩、恒河猴、食蟹、食肉、两栖类或其他哺乳动物数据。

## Human

人类遗传资源相关研究请提交到[GSA for Human](#)数据库。

新冠病毒相关样本，首选此类型。

\*注意：如果想要受控处理(Controlled-access)，可以提交到GSA-Human

## Plant

用于植物样本或植物细胞系样本。

人类型相关样本，请提交到GSA-Human。

## Virus

用于没有引起病害的病毒样本，该团体归类为Clinical or host-associated pathogen

## Metagenome/Environmental Sample (GSC MIMS unsupported)

适用于不适用于Metagenome/Environmental Sample (GSC MIMS compliant) 的宏基因组生物样本数据。

不适用于Metagenome/Environmental GSC MIMS类型的宏基因组样本类型，请选此类型

## Metagenome/Environmental Sample (GSC MIMS compliant)

适用于宏基因组生物样本数据。下设三个小类分别接收来自人类胃肠道、土壤和水相关宏基因数据，即human-gut, soil, water

- human-gut

- soil

- water



# 数据递交入口: BIG-Sub

The screenshot shows the homepage of the BIG Sub submission portal. At the top, there is a navigation bar with links for 'Datasets', 'Tools', 'Standards', 'Publications', and 'About'. Below the navigation bar, there are two red-bordered boxes: '登录入口' (Login) and '注册入口' (Register). The main content area features a section titled '请选择您要提交的数据类型' (Please select the data type you want to submit) with a brief introduction. To the right of this is a section titled '科技计划项目科学数据汇交' (Scientific data submission for scientific projects). Below these sections are several cards representing different data types:

- BioProject**: 项目与共享生物医学研究项目数据库。
- BioSample**: 项目与共享实验材料的生物样本信息数据库。
- Genome Sequence Archive**: 组学序列数据库。存储、管理与共享基因组。
- GSA for Human**: 人类基因组学资源数据库。
- Genome Warehouse**: 各种物种的基因组序列和注释数据库。
- NucBank**: 核酸和蛋白质序列标注数据库。
- OMIX**: 蛋白质相互作用数据库。单蛋白表达蛋白组学。代谢途径与信号通路数据库。
- Genomes Variation Map**: 海洋微生物变异数据库。微生物遗传学数据库。
- BioCode**: 基因开放性信息软件工具数据库。
- Database Commons**: 生物数据库集合：分类学、检索系统。

At the bottom right of the page, there is a URL: <https://ngdc.cncb.ac.cn/gsub>.

# 注册并激活账号

Welcome to register for an account of BIGD

Register information

Account Login Information

Email:

Password:

Confirm Password:

Personal Information

First Name:

Middle Name:

Last Name:

E-mail Address:

City:

State / Province:

Postal Code:

Country / Region:  China

Institutional Information

Institute / Organization:

Department:

Laboratory:

Title / Position:

Research Area:

Check Code:  eath

**注意：由于需要通过邮件接收激活链接，请确保该邮箱属于自己并能够登录进去点击激活链接！否则账号无法激活和登录！同时请合理设置垃圾邮件判断策略，检查激活邮件是否被邮件系统转移到垃圾邮箱中，或者被邮件系统拒收！**

**注意：密码必须同时包含大写字母、小写字母、数字、特殊字符，且长度在8~30位之间！**

## 激活邮件

### SSO Account Activation

发件人: bigd-admin <bigd-admin@bigd.ac.cn>  
时间: 2017年10月10日(星期三) 下午2:50  
收件人:

Dear user,

Please click the URL below to active your account within 48 hours or you will need to register again.

<http://sso.bigd.ac.cn/register/active.action?mailAddr=token=2f6934f45681c87a93752662c8784c22>

Thanks.

BIGD SSO ADMIN

2017-10-10 15:05:03

Note:This email is sent by the system automatically,please do not reply this directly..

点击链接，激活账号

# 登录系统

CNCB-NGDC

Databases Tools

NGDC Central Authentication Service

Enter your Username and Password

Email: gsatest1@big.ac.cn

Password: [Forgot password?](#)

Check code: 36bb 

Keep me signed in

[Login](#) [Reset](#) [Register](#)

Central Authentication Service

---

BioProject

BioSample

BioCode

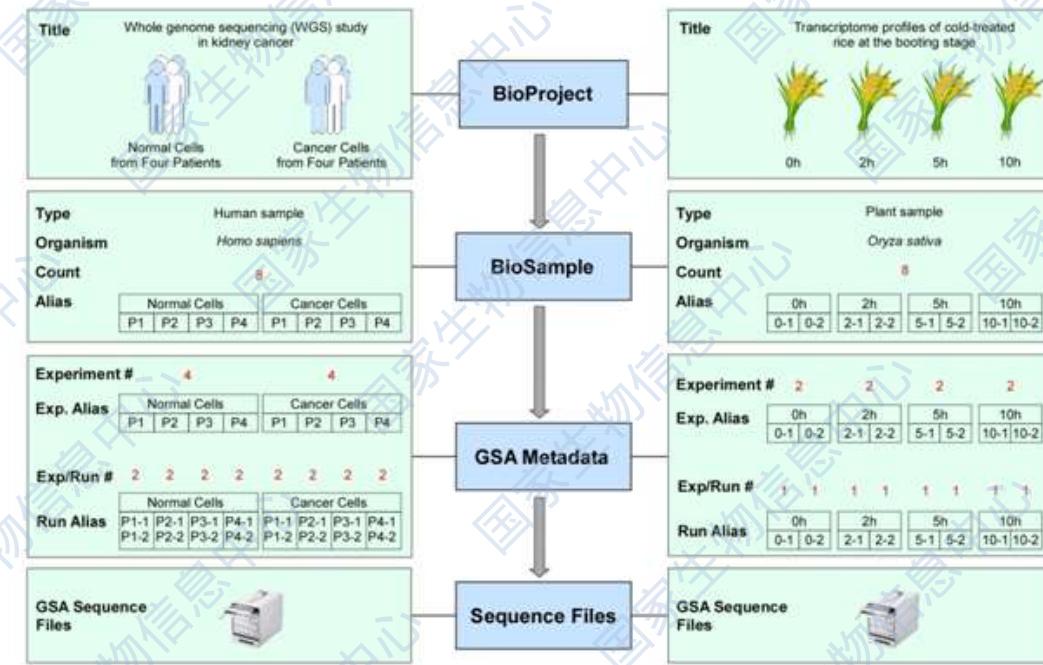
GSA for Human

Genome Sequence Archive (GSA)

Genome WareHouse (GWH)

Genome Variation Map (GVM)

# GSA数据递交流程



# 数据递交 - 创建BioProject

The screenshot shows the homepage of the BIG Sub Submission Portal. At the top, there is a navigation bar with links for Home, Documentation, Login, and Register. Below the navigation bar, a section titled "请选择您要提交的数据类型" (Please select the data type you want to submit) displays several options:

- Genome Sequence Archive (GSA)**: 精学原始数据汇交、存储、管理与共享系统。
- GSA for Human**: 人类遗传资源组学原始数据库。
- Genome Warehouse (GWH)**: 面向多物种全基因组数据的归档数据库。
- BioProject**: 收集与共享生物学研究项目信息的资源库。 This option is highlighted with a red box and a red arrow points to it from the left.
- BioSample**: 收集与共享实验相关的生物样本信息的资源库。
- Genome Variation Map (GVM)**: 基因组序列变异信息汇交、管理与检索的资源库。
- BioCode**: 整合开源生物信息软件工具的数据库。
- Database Commons**: 生物数据库收集、分类、检索系统。

At the bottom right of the page, there is a URL: <https://ngdc.cncb.ac.cn/gsub>.

# 创建BioProject – 提交列表

**BIG Sub**  
BIG Submission Portal

Home Documentation Welcome, GSA -

BIG Sub / BioProject

BioProject 是对研究项目的总体描述。一个BioProject可包含多个生物学样本（BioSample）。

**新建BioProject**

项目编号	提交编号	项目标题	发布日期	提交状态	操作	
PRJCA003418	subPRO004980	GSA_TEST_20200909	2021-09-09	finished	confidential	<a href="#">修改</a> <a href="#">删除</a>
PRJCA002492	subPRO003649	TEST_1_GSATEST1	2020-04-30	finished	confidential	<a href="#">修改</a> <a href="#">删除</a>
PRJCA001010	subPRO001500	project2	2019-12-31	finished	confidential	<a href="#">修改</a> <a href="#">删除</a>
PRJCA001009	subPRO001499	project-test1	2019-12-31	checked OK	Public	

[BioProject Quick Start Guide](#) [BioProject使用说明](#)

# 创建BioProject – 提交者信息

Bio Sub / BioProject / New BioProject

01 提交者信息 02 基本信息 03 项目类型 04 出版信息 05 署名/各提交

提交者信息

\* 姓名: GSA  
中间名: middle name  
姓氏: BIGD  
\* 邮箱: gsateast1@big.ac.cn  
备用邮箱: secondary email  
\* 单位: Beijing Institute of Genomics, Ch  
单位网址: http://www.big.ac.cn  
\* 部门: BIGD  
手机:  
传真:  
\* 街道: No.1 Beichen West Road, Chaoyi  
\* 城市: Beijing  
\* 省:   
\* 邮编: 100101  
\* 国家/地区: China  
  
**保存并进入下一项**

# 创建BioProject – 基本信息

BIG Sub - BioProject New BioProject

01 项目背景 02 基本信息 03 项目类型 04 出版信息 05 相关文件

基本信息

发表日期:

审核通过后即可发布 (推荐)  
 指定日期发布

2023-10-31

发表项目:

项目标题:

项目说明:

项目说明:

涉及领域:

# 创建BioProject – 基本信息

项目资金来源

• 机构  项目类别  • 项目批准号  项目名称

项目相关网站链接

网站信息描述  网站地址

关联项目

项目编号  相关描述信息

外部数据库编号

项目编号  数据库名称

# 创建BioProject – 项目类型

BIG Sub | BioProject | New BioProject

01 提交者信息 02 样本信息 03 项目类型 04 出版信息 05 真实性验证

### 项目类型

\* 项目数据类型

- Whole genome sequencing
- Clone ends
- Epigenomics
- Exome
- Map
- Metagenome
- Phenotype or Genotype
- Random survey
- Targeted Locus (Loci)
- Transcriptome or Gene expression
- Variation
- Genome sequencing and assembly
- Raw sequence reads
- Genome sequencing
- Assembly
- Metagenomic assembly
- Targeted loci cultured
- Targeted loci environmental
- Other

\* 样本范围

Monoisolate

样本范围选择提示

**Monoisolate:** 单个动物、培养细胞系、近交群体（或可能是从集合样本中产生单个基因型群体）。注意此项适用范围较窄，不为优选项，请谨慎选择。

**Multisolate:** 同一物种的多基因型的个体样本集；

**Multi-species:** 多物种样本集；

**Environment:** 环境样本集，多用于宏基因组研究；

**Synthetic:** 在实验室里制造/合成的样本集；

**Single cell:** 单个细胞测序样本集；

**Other:** 其他样本集/无法归入以上类别的特殊样本集。

保存并进入下一项

# 创建BioProject – 出版信息

BIG Sub / BioProject / New BioProject

01 递交者信息    02 基本信息    03 项目类型    04 出版信息    05 概览 & 提交

出版信息

PubMed ID  OR DOI

+ Add another publication

保存并进入下一项

# 创建BioProject – 概览 & 提交

BIG Sub / BioProject / New BioProject

01 提交者信息    02 基本信息    03 项目类型    04 出版信息    05 概览 & 提交

概况信息

提交者信息

提交者	GSA BIGD <a href="mailto:gstatest1@big.ac.cn">gstatest1@big.ac.cn</a>
单位	Beijing Institute of Genomics, Chinese Academy of Sciences
部门	BIGD
国家/地区	China
地址	No.1 Beichen West Road, Chaoyang District Beijing
邮编	100101

项目类型

项目数据类型	Whole genome sequencing
样本范围	Multisample

基本信息

项目标题	TEST_20201015
涉及领域	Medical
项目说明	TEST_20201015
发布日期	2021-10-31

**提交**

# 创建BioProject – 完成提交

BIG Sub / BioProject

BioProject 是对研究项目的总体描述。一个BioProject可包含多个生物学样本（BioSample）。

新建BioProject

生成的BioProject编号，  
记住这个编号，后边提交GSA信息的时候会用到

[BioProject Quick Start Guide](#) [BioProject使用说明](#)

项目编号	提交编号	项目标题	发布日期	提交状态		操作	
PRJCA003665	subPRO005344	TEST_20201015	2021-10-31	finished	confidential	<button>修改</button>	<button>删除</button>
PRJCA003418	subPRO004980	GSA_TEST_20200909	2021-09-09	finished	confidential	<button>修改</button>	<button>删除</button>
PRJCA002492	subPRO003649	TEST_1_GSATEST1	2020-04-30	finished	confidential	<button>修改</button>	<button>删除</button>
PRJCA001010	subPRO001500	project2	2019-12-31	finished	confidential	<button>修改</button>	<button>删除</button>
PRJCA001009	subPRO001499	project-test1	2019-12-31	checked OK	Public		

# 数据递交 – 创建GSA

GSA递交入口

BIG Sub Quick Start Guide BIG Sub 使用说明

请选择您要提交的数据类型

生物数据提交系统(BIG Submission, BIG Sub)是国家基因组科学数据中心生物数据统一入口，为用户提供一站式数据递交服务。

**Genome Sequence Archive** 相关数据汇交、存储、管理与共享系统

**BioProject** 收集与共享生物学研究项目信息的资源库

**BioSample** 收集与共享实验相关的生物样本信息的资源库

**GSA for Human** 人类遗传资源基因型与表型信息

**Gene Expression** 基因表达数据的汇交、存储与分析

**NucBank** 核酸和蛋白序列数据库与分析

**OMIX** 表型和基因表达谱、甲基化谱、蛋白质、代谢组学等数据的资源库

**Genome Variation Map** 基因组序列变异信息汇交、管理与检索的资源库

**EcoCode** 整合开源生物信息软件工具的数据仓库

**Database Commons** 生物数据库聚集、分类、检索系统

科技计划项目科学数据汇交

科学项目数据汇交服务系统  
提供汇交计划提文、实时数据归档情况查询、汇交证明出具等服务  
**SDAC**

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

## 创建GSA – 提交列表

国家生物信息中心

Home Documentation

BIG Sub GSA

GSA 是组学原始数据汇交、存储、管理与共享系统。在开始创建GSA数据集前，您需要进入 [BioProject提交入口](#) 完成 BioProject（对研究项目的总体描述）的创建。

提示：

- 人类遗传资源相关组学原始数据请提交到GSA for human数据库。
- 同一个项目（BioProject）的数据应归档到同一个GSA数据集中；若 BioProject 包含多种样本类型（Sample type），如 [PRJCA001357](#)，您需要按样本类型创建多个GSA数据集，并将相关类型数据归档到对应的GSA数据集中。
- 如果您在数据上传过程中遇到有问题或发现任何系统报错，请通过 [gsa@big.ac.cn](mailto:gsa@big.ac.cn) 邮箱联系我们。

[如何共享和发布GSA数据集](#)

[新建 GSA](#) 点击新建GSA提交

GSA Quick Start Guide GSA使用说明

GSA编号	提交编号	GSA标签	发布日期	提交状态	操作
Unassigned	subCRA002947		2020-04-06	Unfinished at the General Info step Confidential	<a href="#">新建</a>
CRA001076	subCRA001069	GSA-test	2019-12-31	Checked OK Public	<a href="#">新建</a>

## 创建GSA - 提交者信息

提交者信息

* 姓名 GSA	中间名 middle name	* 姓氏 BIGD
* 邮箱 greatest@big.ac.cn	备用邮箱 secondary email	
* 单位 Beijing Institute of Genomics, Chi	单位网址 http://www.big.ac.cn	* 部门 BIGD
手机	传真	
* 地址 No.1 Beichen West Road, Chaoyang	* 城市 Beijing	* 省 北京市
* 邮编 100101	* 国家/地区 China	

**保存并进入下一项** **保存并进入下一步**

# 创建GSA - 基本信息

BIG Sub / GSA / Submission: subCRA010552

01  
提交者信息

02  
基本信息

03  
元数据信息

04  
文件

05  
概览 & 提交

基本信息

发布日期

- 审核通过后即可发布
- 指定日期发布

2024-05-26



(yyyy-mm-dd)

选择审核通过后立即发布，或者指定日期发布。  
发布日期必须是从提交起两年之内的日期

\* 发布策略和免责声明

1. 您可以根据需求设定“发布日期”，在该日期之前，GSA保证数据不公开；
2. 如果引用这些数据和该accession号的文章先于您设定的发布时间而发表，我们将根据文章的发表时间来发布该数据；否则GSA将根据您设定的发布日期而发布该数据；
3. 一旦文章发表，数据可以发布，请把已发表文章的全部信息-作者、题目、期刊、刊号、页数、日期信息发送到该邮箱: [gsa@big.ac.cn](mailto:gsa@big.ac.cn)
4. “发布日期”可以在GSA提交系统内进行修改 ([https://ngdc.cncb.ac.cn/gsub/submit/gsa/\[substitute your GSA submission number\]/finishedOverview](https://ngdc.cncb.ac.cn/gsub/submit/gsa/[substitute your GSA submission number]/finishedOverview))

- I accept it.
- I don't accept it.

## 创建GSA - 基本信息

**标题和描述信息**

\* 标题

\* 描述信息

**项目信息**

\* 请选择项目编号  
 OR Go to create BioProject

选择已有的BioProject编号，或者先创建BioProject

**样本信息**

未创建GSA相关的BioSample信息  
 已经创建好GSA相关的BioSample信息  
 如果您还未创建GSA相关的BioSample，请选择“未创建GSA相关的BioSample信息”，在本次提交流程中创建BioSample(s)  
 如果您已创建好GSA相关的BioSample，请选择“已经创建GSA相关的BioSample信息”

选择是否已创建BioSample

**保存并进入下一项**



# 创建GSA - 样本类型

BIG Sub GSA Submission: subCRA004261

01 提交信息 02 项目信息 03 样本类型 04 样本属性 05 宏基因组 06 文件 07 确认 & 提交

## 样本类型

**Pathogen**  
用于与公共卫生相关的病原体样本

Clinical or host-associated  
 Environmental, food or other  
 Microbe  
用于能具体写出物种名称的微生物样本，但不包括致病菌或病毒样本。宏基因组数据建议选择的Metagenome/Environmental Sample (GSC MIMS unsupported)或Metagenome/Environmental Sample (GSC MIMS compliant)

Animal  
用于模式生物或动物的多细胞样本或细胞系样本，如大鼠、小鼠、果蝇、线虫、鱼类、鸟类或其他哺乳动物数据

Plant  
用于植物样本或植物细胞系样本

Virus  
用于所有与疾病无关的病毒样本，病原体归类为Clinical or host-associated pathogen

**Metagenome/Environmental Sample (GSC MIMS unsupported)**  
目前用于不适合Metagenome/Environmental Sample (GSC MIMS compliant) 的宏基因组生物样本数据

**Metagenome/Environmental Sample (GSC MIMS compliant)**  
用于宏基因组生物样本数据，下设三个类别分别接收来自人类胃肠道、土壤和水相关宏基因组数据，即human-gut, soil, water

human-gut  
 soil  
 water

**保存并进入下一项**

# 创建GSA - 样本属性

BIG Sub → GSA / Submission subCRA004261

01 样本信息 02 样本类型 03 样本属性 04 样本属性 05 元数据信息 06 文件 07 提交与提交

临床或病原微生物样本属性信息

\* 上传BioSample批量提交文件  
请选择文件 上传 点击下载Sample模板文件，填写完成后上传

① BioSample批量提交模板文件 [Pathogen\\_cl.cn.xlsx](#)，完成填写并检查无误后上传。  
BioSample批量提交示例。详见 [e.g. Pathogen\\_cl.cn.xlsx](#)。  
更多帮助，请查看 [Help](#)。

示例文件。注意：不能编辑、上传此文件

本表仅适用于 Clinical or host-associated pathogen 类型的样本  
“绿色标题栏”为必填项  
“灰色标题栏”为选填项，如果无法提供，可以为空  
“黄色区域”有下拉菜单

organism (第5列) 和 host (第11列) 请填写物种学名，具体名称可使用 NCBI 的 Taxonomy 数据库进行搜索和验证 (<https://www.ncbi.nlm.nih.gov/taxonomy>)

1	2	3	4	5	6	7	8	9
sample_name	public_description	project_accession	sample_title	organism	isolate	strain	collected_by	collection_date
10	11	12	13	14	15	16	17	18
geographic_location	host	host_disease	isolation_source	latitude_longitude	culture_collection	genotype	host_ages	host_age_unit
								year

\*注意：BioSample name需要跟experiment 表格里的sample\_name保持一致

[pathogen\\_cl.cn.xlsx](#)

# 创建GSA - 样本属性

临床或病原微生物样本属性信息

\* 已上传的BioSample批量提交文件  
Pathogen\_cl.cn-12.xlsx 33KB 删除 校验

① 请点击“校验”按钮对批量提交文件进行在线审核。若文件审核通过，请点击“保存并进入下一项”完成BioSample批量提交；若文件审核未通过，请删除已上传文件并修改后重新上传。

↓

临床或病原微生物样本属性信息

\* 已上传的BioSample批量提交文件  
Pathogen\_cl.cn-12.xlsx 33KB 删除 校验 ERROR

下载报告文件: error.txt

Error in Sample sheet  
row 11, column 10: "geographic\_location" ("China: Hubei: Wuhan") Formal error. You can use a colon to separate the country or ocean from more detailed information about the location, eg "Canada: Vancouver" or "Germany: halfway down Zugspitze, Alps"

↓

临床或病原微生物样本属性信息

\* 已上传的BioSample批量提交文件  
Pathogen\_cl.cn-12.xlsx 33KB 删除 校验 ✓ Checked OK.

① 请点击“校验”按钮对批量提交文件进行在线审核。若文件审核通过，请点击“保存并进入下一项”完成BioSample批量提交；若文件审核未通过，请删除已上传文件并修改后重新上传。

保存并进入下一项

校验错误，删除已上传的表格，按提示修改后重新上传并校验

# 创建GSA - 序列元数据信息

请上传GSA元数据模板文件

请选择样本

GSA批量提交模板文件 [GSA\\_Template.cn.xlsx](#) 完成填写并检查无误后上传。  
GSA批量提交示例，详见 [e.g.GSA\\_Template.cn.xlsx](#)

更多帮助，请查看 [Help](#)。

保存并进入下一环

点击下载GSA元数据模板文件，填写完成后上传

点击下载GSA元数据示例文件。注意：不能编辑、上传此文件

# 元数据表格: Experiment/Run

GSA批量表格由两部分组成Experiment和Run表，本表用于收集Experiment信息

“绿色标题列”为必填项

“灰色标题列”为选填项，如果无法提供，可以为空

“蓝色标题列”为Paired数据必填

“黄色区域”有下拉菜单

当“Platform (第6列) 为Helicos HeliScope时。Planned number of cycles必填

注意：1. 不要随意删除或插入词条列！

2. 在您使用Excel“自动填充功能”时需谨慎，以防出

3. 为节省您批量上传表格的审核时间，请您在提交

1	2	3	4	5	6	7
ID	*Experiment title	BioProject accession	BioSample name	BioSample accession	Platform	*Library Construction / Experimental Design
E1						
E2						

\*注意： BioSample name需要跟sample表格里的sample\_name保持一致

GSA批量表格由两部分组成Experiment和Run表，本表用于收集Run信息

“绿色标题列”为必填项

“灰色标题列”为选填项，如果无法提供，可以为空

“蓝色标题列”为FASTQ格式Paired数据必填项，FASTQ格式只接收.bz2或.gz压缩，不接收其它格式的压缩文件

“紫色标题列”为BAM格式，且不是来自PacBio sequel或Ion Torrent系列测序平台需填项；如果您想把参考序列传到GSA数据库，请您填写Reference file name和MD5 for reference file (第10和11)

“黄色区域”有下拉菜单

1	2	3	4	5	6	7
ID	*Run title	BioProject accession	*Experiment ID	*Run data file type	*File name 1	*MD5 checksum 1
R1						
R2						

GSA\_Template.cn.xlsx

# 创建GSA - 序列元数据信息



# 创建GSA - 序列文件上传方式

The screenshot shows the 'File Upload Method' section of the GSA submission interface. It includes instructions for file naming and compression, a note about using client software, and a warning about using binary mode.

**建议使用客户端软件上传文件  
注意：请使用二进制模式上传**

**文件名、MD5码需要跟 metadata表格里填写的一致**

**保存并进入下一项**

The screenshot shows a Windows File Explorer window with a folder structure for GSA submission. It includes files named 'ef\_123456.gff' and 'ef\_123456.gz' with their respective file sizes.

# 创建GSA – 序列文件上传方式

The screenshot shows a configuration page for 'File Upload Method'. It includes sections for 'General' (uploading files with SRS and MDS metadata matching GSA Metadata schema type), 'Selected File Upload Method' (FTP), and 'Advanced Command Line' (selected). The 'Advanced Command Line' section contains a command line with a red oval around the path [path/to/keyfile]. A callout box next to it says: 'Please copy the command line from the page and replace the path in the command line with the local path of the secret key file.' Another red oval surrounds the 'Get the key file' link, with a callout box below it: 'Secret key file address, click to download secret key file and replace the [path/to/keyfile] in the command line with the local path of the downloaded file.' At the bottom left is a red button: 'Save and Go to Next Step'.

注意：本页面不是提交的最后一步。  
请选择文件上传方式后，点击进入下一页，浏览并保存提交

请拷贝页面上的命令行，并替换命令行中的相关路径

秘钥文件地址，点击下载秘钥文件  
并将命令行中的[path/to/keyfile]替换为下载的本地路径

保存并进入下一项

文件上传方法

说明  
上传的文件的生作者和MDS的必须跟在GSA Metadata schema type的一致。  
上传的文件名必须是唯一的。  
Info文件必须压缩后上传，目前限制文件名.gzip 和 .bz2格式的压缩文件。

\* 选择文件上传方式

FTP  
使用FTP账户密码进行上传工作。FTP账号密码与SRS上的完全一样。

Advanced Command Line  
使用Advanced Command Line上传文件。  
帮助：使用命令行上传文件，类似于Windows命令行。  

```
java -jar GSAtoolkit.jar -jbossUser=YourUser -jbossPass=YourPass -keyfile=PathToYourKeyfile -secretkeyfile=PathToYourSecretKeyfile -resource=PathToYourResource
```

Where:  
[path/to/keyfile]:  
Microsoft Windows: C:\Program Files\Java\jdk1.8.0\_111\bin\java -jar GSAtoolkit.jar -jbossUser=YourUser -jbossPass=YourPass -keyfile=PathToYourKeyfile -secretkeyfile=PathToYourSecretKeyfile -resource=PathToYourResource  
Mac OS X: /Applications/Java/jdk1.8.0\_111/bin/java -jar GSAtoolkit.jar -jbossUser=YourUser -jbossPass=YourPass -keyfile=PathToYourKeyfile -secretkeyfile=PathToYourSecretKeyfile -resource=PathToYourResource  
Linux: /usr/lib/jvm/jdk1.8.0\_111/bin/java -jar GSAtoolkit.jar -jbossUser=YourUser -jbossPass=YourPass -keyfile=PathToYourKeyfile -secretkeyfile=PathToYourSecretKeyfile -resource=PathToYourResource  
注：如果秘钥文件不在当前目录，请将该参数修改为秘钥文件的完整路径。  
注：如果秘钥文件在其他目录，请将该参数修改为秘钥文件的完整路径。  
Out the key file  
说明：如果一个项目对一个测试子项目可以使用GSA的相同密钥作为子项目密钥，通过将这个子项目作为一个临时文件夹，将多子项目密钥放在同一个父项目下的临时文件夹后，将项目重新打包。

# 创建GSA - 概览 & 提交

The screenshot shows the GSA submission interface with the following sections:

- GSA基本信息 (Basic Information):**
  - 提交编号: subGSA04291
  - 标题: TEST\_20201015
  - 描述: TEST\_20201015
  - 发表日期: 2021-10-31
- 递交者信息 (Submitter Information):**
  - 递交者: GSA BIRD  
gaoxif@igb.ac.cn
  - 单位: Beijing Institute of Genomics, Chinese Academy of Sciences
  - 部门: BIGO
  - 国家/地区: China
  - 地址: No.1 Beichen West Road, Chongming District, Beijing
  - 邮编: 100101
- 检测者信息 (Detector Information):**
  - 样本类型: Clinical or host-associated pathogen
  - Biosample检测文件: Pathogen\_GSA-test.f2.xlsx
  - 元数据文件: GSA\_Template\_for\_gsa-test.xlsx
- 实验设计 (Experiment Design):**

实验名称	测序平台	样本名称	生物名称	实验选择
experiment	Illumina HiSeq 2000	Sample1	SARS-like respiratory syndrome coronavirus 2	unspecified
RunID名			Raw序列文件处理流程	文件类型
File	File: s1_1.fastq File: s1_2.fastq			fastq

**提交按钮 (Submit Button):** A red button labeled "提交" (Submit) is highlighted with an orange arrow pointing to the text "检查无误后, 点击“提交”, 完成本次提交" (Check after confirming, click "Submit" to complete this submission).

# 创建GSA – 元数据提交完成

GSA编号	提交编号	GSA标题	创建日期	发布日期	提交状态	操作
Unassigned	subCRA004261	TEST_20201015	2020-10-15	2021-10-31	<i>Checking detail</i> Confidential	<button>删除</button>
Unassigned	subCRA004036	TEST_20200915	2020-09-15	2020-09-30	<i>Checking detail</i> Confidential	<button>删除</button>
Unassigned	subCRA003988	GSA_TEST_20200910	2020-09-10	2022-11-08	<b>⚠️ Unfinished at the OverView step</b> Confidential	<button>删除</button>
Unassigned	subCRA002947		2020-04-06	2020-04-06	<b>⚠️ Unfinished at the General Info step</b> Confidential	<button>删除</button>
CRA001076	subCRA001069	GSA-test	2018-08-29	2025-02-24	<b>Checked OK</b> Confidential	<button>立即发布</button> <button>分享</button>

# 创建GSA – 数据审核状态

BIG Sub / GSA / Submission: subCRA004261

## GSA基本信息

GSA 提交信息: subCRA004261 / Release Date : 2021-10-31 / Project : PRJCA003665

状态 1 Runs are checked failed [detail](#)

标题

TEST\_20201015

发布日期

2021-10-31

[更新](#)

**File not found:** 如果元信息刚审核通过，需要耐心等待后台关联审核数据；如果元信息审核通过很久，需自行检查数据是否上传到GSA目录下，文件名是否一致。如果不一致，可点run编号后的edit修改一致。Filezilla上传数据的用户也可以自行修改数据文件名称

**md5 is inconsistent:** 自行核对本地文件的md5与元信息填写中是否一致，如果填写有误，点run编号后面的edit修改。如果确认填写无误，则是文件上传过程有问题导致，选择二进制重新上传文件

**processed error:** 数据审核报错，务必联系管理员反馈报错信息

**processed succeed:** 数据审核没有问题，全部成功后等归档就可以收到编号

# 数据发布

Public

Data are fully available by anyone

Confidential

Data are not available publicly through any means, data owners should provide the date release date

\* 发布日期  
2018-05-24   
(yyyy-mm-dd)

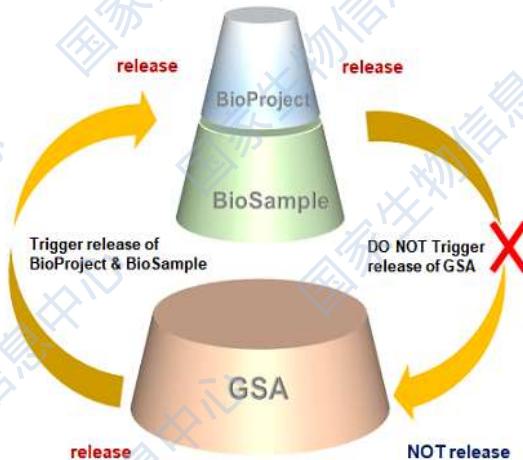
- 发布策略和免责声明
- 1. 你提交的数据可以保留在一定时间内不发布；
- 2. 发布时间可以在GSA提交系统内进行修改：[http://bigd.big.ac.cn/gsa/submit/gsa?substitute your GSA accession number\)/contents](http://bigd.big.ac.cn/gsa/submit/gsa?substitute your GSA accession number)/contents)
- 3. 如果引用这些数据和该accession号的文章免于您设定的发布时间而发表，我们将根据文章的发表时间来发布该数据；否则GSA将根据您设定的发布日期而发布该数据；
- 4. 一旦文章发表，数据可以发布，请把已发表文章的全部信息—作者，题目，期刊，刊号，页数，日期信息发送到该邮箱：GSA@big.ac.cn
- I accept it.  I don't accept it.

新建 GSA

GSA编号	提交编号	GSA标题	发布日期	提交状态	操作
CRA000532	SubCRA000585	Human dataset test	2020-12-31	Checked OK Confidential	<input type="button" value="立即发布"/> <input type="button" value="驳回"/>

数据发布条件：

- 到达用户设定的发布日期
- 文章见刊



# 设置共享链接

The figure shows two screenshots of a GSA submission interface. The top screenshot displays a table with columns: Accession, Submission ID, Title, Release date, Status, and Operation. A red box highlights the 'Share' button in the Operation column for a row with CRA000620 as the Accession. The bottom screenshot shows the same table after sharing, where the 'Operation' column now includes a 'Shared URL' field containing a temporary link and a 'Cancel Share' button.

Review Link Opinion

Accession: CRA000620  
Set review link expire time  
1 month ▾

Create data access link

NOTE: Reviewers can access the sequence files through this link, please tick the checkbox to generate this link as required.

Cancel Submit

注：此链接为临时链接，可以将该链接分享给编辑和审稿人，方便其查看数据，但为了您的数据安全请不要将此链接对外公布。数据共享结束后，请点击“Cancel share”按钮，取消数据共享。

# 数据引用

## 如何引用GSA?

当您成功提交数据到GSA并通过审核后，请在您要发表的论文中添加如下语句：

*The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2022), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA: CRAxxxxxx) that are publicly accessible at <https://ngdc.cncb.ac.cn/gsa>.*

请按照以下格式引用我们的文章：

- **The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types.** *Genomics, Proteomics & Bioinformatics* 2021, 19(4):578-583. <https://doi.org/10.1016/j.gpb.2021.08.001> [PMID=34400360] 

[Endnote文件下载](#)

- **Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024.** *Nucleic Acids Res* 2024, Jan 5;52(D1):D18-D32. <https://doi.org/10.1093/nar/gkad1078> [PMID=38018256] 

[Endnote文件下载](#)



# 数据访问

The screenshot shows the GSA search interface. A red box highlights the search bar where 'GSA' is selected and 'CRAD01962' is entered. Below the search bar, a message indicates '200 items found' and provides a link to 'View all'. A red arrow points from the 'Send to:' dropdown in the search results table to the 'Choose Destination' dialog box.

**Search result:**

Rank	Title	Accession	Platform name	Strategy name	Source name	Selection name
1	Spleen-2-S-IP	CRX0010012	ILLUMINA	RNASeq	TRANSCRIPTOMIC	PCR
2	Heart-1-S-IP	CRX0498994	ILLUMINA	RNASeq	TRANSCRIPTOMIC	PCR
3	Lung-2-S-Input	CRX0262013	ILLUMINA	RNASeq	TRANSCRIPTOMIC	PCR
4	Lung-1-S-IP	CRX0262311	ILLUMINA	RNASeq	TRANSCRIPTOMIC	PCR
5	Liver-1-S-Input	CRX0498909	ILLUMINA	RNASeq	TRANSCRIPTOMIC	PCR

**Choose Destination**

File

Download 40 items.  
Format  
RunInfo

Create Files



**GSA** Genome Sequence Archive

Search result:

Total items: 5804304 | Items 1 of 20 | Page 1 / 290216 | First | Next | Last | Jump To: 1 Go

**Send to :**

**Choose Destination**

- File
- Download 5804385 items.
- Format
- Accession List
- RunInfo
- DataSet Accession List

1.	Title: Illumina HiSeq 1000 paired end sequencing of SAMD00082794	Accession: DRX088646	Platform name: ILLUMINA	Source name: METAGENOMIC	Strategy name: Genome	Selection name: RANDOM
2.	Title: Illumina HiSeq 1000 paired end sequencing of SAMD00082790	Accession: DRX088642	Platform name: ILLUMINA	Source name: METAGENOMIC	Strategy name: Genome	Selection name: RANDOM
3.	Title: Illumina HiSeq 1000 paired end sequencing of SAMD00082792	Accession: DRX088644	Platform name: ILLUMINA	Source name: METAGENOMIC	Strategy name: Genome	Selection name: RANDOM
4.	Title: Illumina HiSeq 1000 paired end sequencing of SAMD00082791	Accession: DRX088643	Platform name: ILLUMINA	Source name: METAGENOMIC	Strategy name: Genome	Selection name: RANDOM

# 数据访问

Home / GSA / CRA003018

## CRA003018 基本信息

标题: ZEA affects porcine ovarian granulosa cells  
项目编号: PRJCA003148 /  
发布日期: 2020-09-09  
文件个数: 27  
文件大小: 69.51 GB

点击下载元数据

数据下载

元数据信息: [CRA003018.xlsx](#)

文件: [HTTPS: https://download.cncb.ac.cn/gsa/CRA003018](https://download.cncb.ac.cn/gsa/CRA003018) [FTP: ftp://download.big.ac.cn/gsa/CRA003018](ftp://download.big.ac.cn/gsa/CRA003018)

EgdeTurbo (测试) 文件下载 Aspera命令行 帮助

● 提示: HTTP下载速度有限, 推荐使用Egde Turbo或FTP客户端(比如 FileZilla Client) 下载数据。EgdeTurbo 支持Linux命令行、Windows/Mac平台的Chrome、Edge和Firefox浏览器。

多种文件下载方式

## Experiments 和 Runs 信息

实验编号	实验名称	生物名称	测序平台	样品编号
CRX130439	S3_3_miRNA	Sus scrofa	Illumina NovaSeq 6000	SAMC206797
Run编号	Run别名		Run序列文件信息	
CRR156624	S3_3_miRNA		File: CRR156624.fq.gz	
CRX130438	S3_2_miRNA	Sus scrofa	Illumina NovaSeq 6000	SAMC206796
Run编号	Run别名		Run序列文件信息	
CRR156623	S3_2_miRNA		File: CRR156623.fq.gz	



# 提纲

一

二

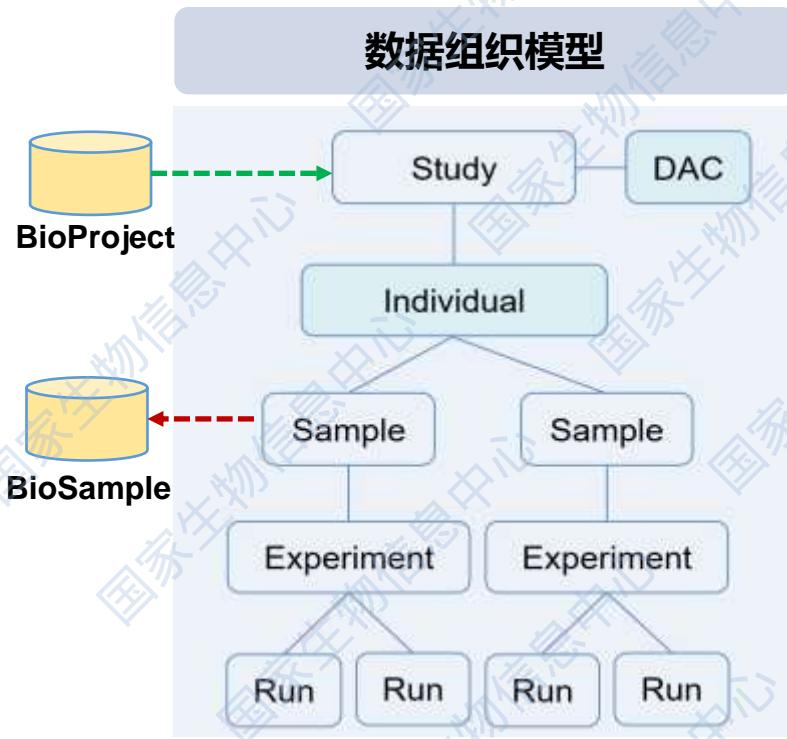
三

## GSA Family介绍

## GSA数据提交及共享

## GSA-Human数据提交及共享

# 数据模型和支持的数据类型



以individual为核心，组织元信息与数据文件

研究类型	适用元数据模板
疾病研究 (Disease study)	肿瘤研究 (Cancer) Tumor_1.0_us.xlsx
	自身免疫研究 (Autoimmune diseases) Autoimmune_1.0_us.xlsx
	其他疾病研究 (Other diseases) General_1.0.xlsx
队列研究 (Cohort study)	General_1.0.xlsx
细胞系相关研究 (Cell line related study)	Cellline_1.0.xlsx
临床病原体研究 (Clinical pathogen)	Clinical_pathogen_1.0.xlsx
人体相关宏基因组研究 (Human associated metagenome)	Human_associated_metagenome_1.0.xlsx
测序平台	文件格式
ILLUMINA	FASTQ, BAM
PACBIO_SMRT	HDF5, BAM, FASTQ
Oxford Nanopore	FASTQ, BAM, Fast5
COMPLETE_GENOMICS	Complete_Genomics_native, FASTQ, BAM
BGISEQ	FASTQ, BAM
BIONANO	BNX
ABI_SOLID	FASTA/QUAL, FASTQ, BAM
...	...

更多数据格式信息: [https://ngdc.cncb.ac.cn/gsa/support/standardsGsa#file\\_1](https://ngdc.cncb.ac.cn/gsa/support/standardsGsa#file_1)

# 数据提交须知

- 数据提交必须使用PI账号 – 数据责任人
  - 科研机构独立研究组的课题组长 (PI) 或具有高级职称的人员
  - 高校教授/副教授
  - 医院主任/副主任医师
  - 企业部门负责人
- 元数据信息必须脱敏，即不能包含受试者隐私信息
- 提交时设置数据访问方式 (用户自行设定)
  - 受控访问 (默认)
  - 公开访问
- 受控访问数据，须提供DAC (数据管理委员会) 及成员信息，并指定DAC联系人
- 对外共享数据集前，应根据《中华人民共和国人类遗传资源管理条例》的规定，获得相关数据集在“科学技术部政务服务平台”中“开放使用”类别的备案号

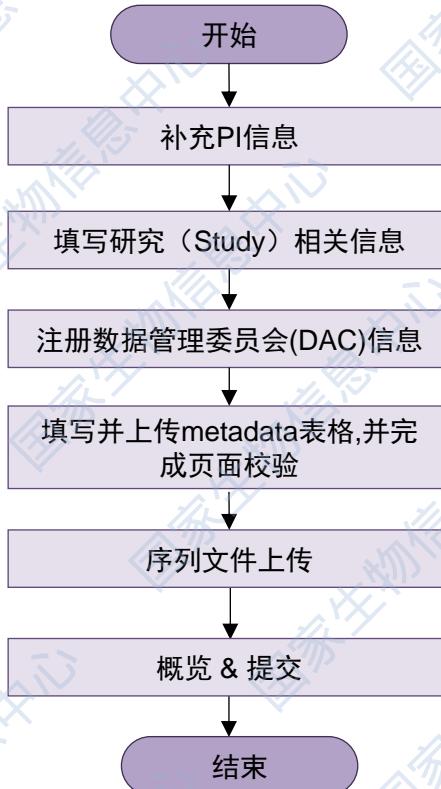
[https://ngdc.cncb.ac.cn/gsa-human/document/Principle\\_of\\_Accessing\\_Human\\_Genetic\\_Resource\\_Data\\_in\\_NGDC\\_V1.pdf](https://ngdc.cncb.ac.cn/gsa-human/document/Principle_of_Accessing_Human_Genetic_Resource_Data_in_NGDC_V1.pdf)

# 受控访问数据的“申请-审核”模式



遗传, 2021

# 数据提交流程



## 个人信息补充



登录入口

10



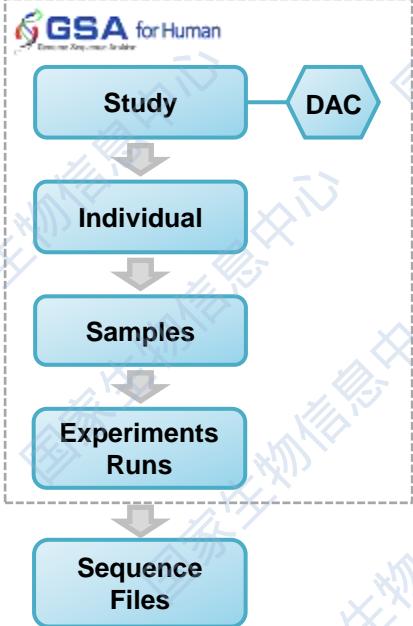
**登录后，需要完善个人信息：**

## 补充职称、个人主页/通讯文章链接、单位 主页信息

**个人主页上的邮箱信息最好与账号邮箱一致  
不符合条件的账号进行数据提交，将会审批不通过**

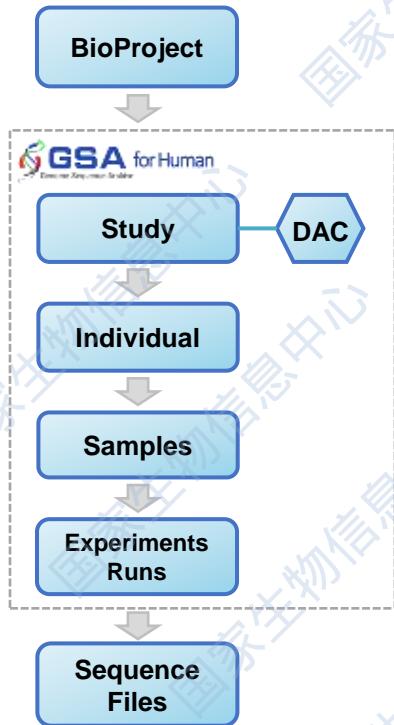
# 新建GSA-Human提交

BioProject



The screenshot shows the 'GSA for Human' website interface. At the top, there are links for 'Data Resources', 'Computing Analysis', 'Data Network', and 'Standards'. Below the header, a search bar contains 'GSA for Human' and 'Submit a GSA-Human submission'. A user 'Welcome, Zheng' is logged in. The main content area is titled 'GSA-Human数据提交' (GSA-Human Submission). It explains that GSA-Human is a database for human genome sequence data. To submit, users must log in via BioProject or use the BioProject submission interface. It lists requirements for study leaders (PIs) or project managers, including high-level descriptions and specific individuals. It also covers authorship and copyright issues. A '帮助文档' (Help Document) section provides links to CHN (Human Genome Nomenclature), CHM (Human Genome Mutation Database), and CNA (National Center for Biotechnology Information). A note about multiple submissions is present. A red box highlights the '新建GSA-Human' (Create New GSA-Human) button at the bottom left. To its right, a callout box says '进入GSA-Human提交列表页面, 创建一个新的提交' (Enter the GSA-Human submission list page, create a new submission). The bottom of the page shows a table with columns for '提交状态' (Submission Status), '临时类型' (Temporary Type), and '操作' (Operation).

# GSA-Human提交：数据提交者信息



New Submission

bioRxiv - USA-Human - New Submission

bioRxiv - USA-Human - New Submission

New Submission

Step 1: Information Step 2: Human Step 3: File Step 4: Summary & Submit

Contact Information

Do you want to assign another person as Contact Person?

I want to add one more contact information

From the Submitter and the Contact Person, if you click, we'll receive the email from GSA-Human database.

Add another Contact Information:

First Name: genetell  
Middle Name: Kristine  
Last Name: DAC

Email: genetell@genetell.org.cn

Institute/Organization: NIBIB

Department: NSBDC

Street: No.1 Neimeng West Road, Chao

City: Beijing

State/Province: China

Postal Code: 100181

Country: China

Protocol for the Release of Human Genetic Material (HGRN)

The GSA-Human Database (GSA) is designed for the collective management and access of human genetic resource information. There are two types of data access in GSA: Human Genetic Access and Controlled Access. The Open-Access data can be downloaded and used directly, while the Controlled-access data can only be accessed after applying to the relevant data authorities and getting authorized.

I agree to the following:

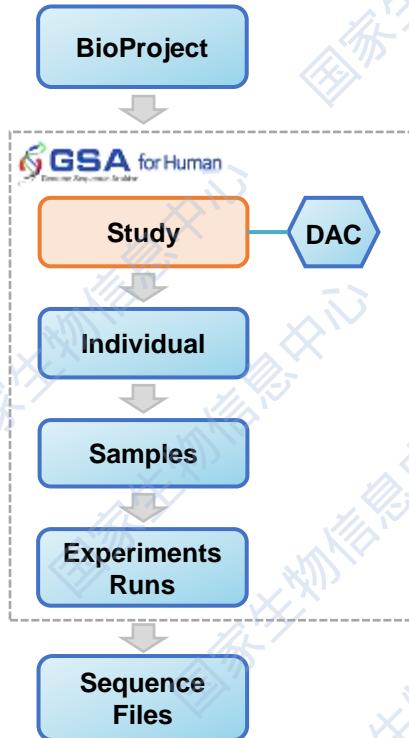
- I agree to the following:
- This study authorizer has the right to extract fine-grained usage of the submitted raw sequencing data and related metadata on GSA-Human genetic resources acquired in accordance with the relevant data authorities and getting authorized.
- This study authorizer has the right to use the fine-grained usage of the submitted raw sequencing data and related metadata on GSA-Human genetic resources acquired in accordance with the relevant data authorities and getting authorized.
- This study authorizer has the right to withdraw his/her right of dataset access and release, even if the related research result has been published or the intellectual property has been released. The related dataset contributor should be consulted with consensus of the data authorizer and the GSA-Human database.
- This study authorizer and the GSA-Human database have the right to agree on a license for the application of the data access and release.
- Non-transferability of data authorizer
- This study authorizer is the person responsible for the dataset submitted, which should be the group leader. Before submitting to GSA-Human, the following conditions must be met:
  - I have read and accepted the items above.

Save and Continue

请注意，建议添加提交联系人信息。如果不添加，数据集问题的反馈默认仅通知注册账号，而往往数据提交和问题处理不一定由PI本人负责。添加后两个人都会收到通知信息，多一个人有备而无患

进入下一步前，务必请先阅读并接受我们的数据管理规则

# GSA-Human提交：研究信息



The screenshot shows the GSA for Human study submission interface. A red box highlights the 'Release Information' section, which includes fields for 'Release date' (set to '2016-01-01') and 'Study type' (selected as 'Oncology'). Another red box highlights the 'BioProject' field, which contains the identifier 'PRJ-HU-0000000000'. A third red box highlights the 'Data availability' section, which contains a detailed policy statement and a checkbox for 'I have read this explanation of the data use & distribution required when requesting for this data!'. At the bottom, a red box highlights the 'Save and Next' button.

**发布时间：**可选**两年之内的**任意时间作为发布时间；  
只要数据未公开，可随时修改该时间

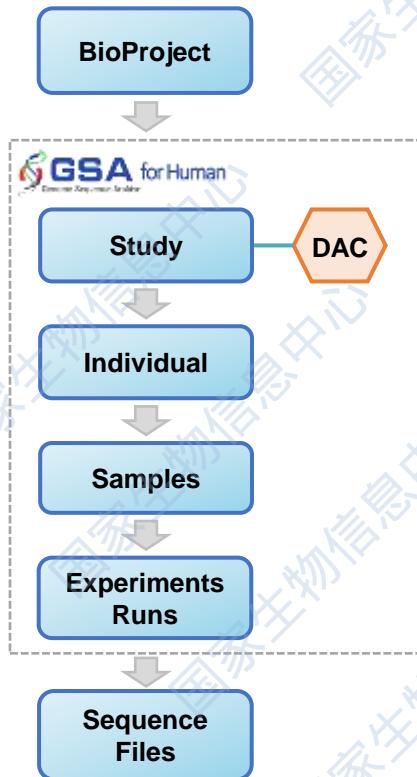
**基本信息：**按照研究情况，选择研究类型及其相关信息

**BioProject信息：**填写已有BioProject编号，或者先  
跳转至BioProject新建

# GSA-Human提交：研究信息



# GSA-Human提交：数据管理委员会 (DAC) 信息



Submission ID: subIRRA000497

Data Access Committee (DAC) is a body of one or more named individuals who are responsible for data release to external requestors based on consent and/or National Research Ethics Review.

A DAC is typically formed from the same or multiple organizations that collected the samples and generated any associated analyses.

Multiple datasets may be affiliated to a single DAC.

**DAC Information**

\* Did you already register a DAC for this study?

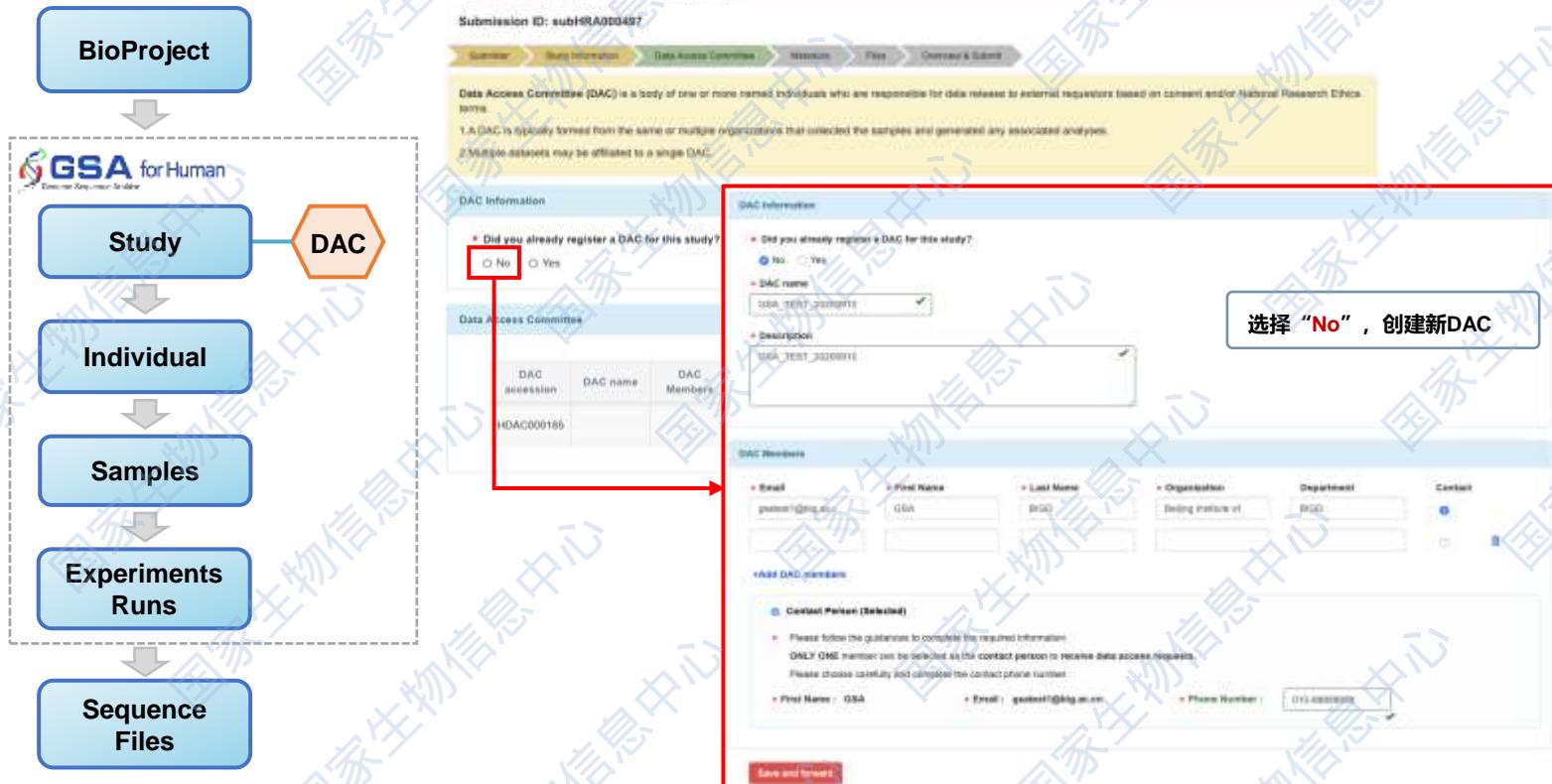
No  Yes

**Data Access Committee**

DAC accession	DAC name	DAC Members	Email	Organization	Department	Is Contact Person	Contact Phone
HDAC000166				Beijing Institute of Genomics (BIG)	BIG Data Center	Yes	

如果数据访问方式使用“受控访问 (Controlled-access)”，需建立数据管理委员会 (DAC) 并指定**DAC联系人**，负责数据发布后的申请审批与授权

# GSA-Human提交：数据管理委员会 (DAC) 信息



# GSA-Human提交：元数据信息

**BioProject**

The diagram illustrates the GSA-Human submission process. It begins with a **BioProject**, which then branches into **Study**, **Individual**, **Samples**, **Experiments Runs**, and **Sequence Files**. The **Study** path leads to the GSA-Human submission interface, which is shown in a screenshot below.

**GSA for Human**  
Submit a Biological Sample

Submission ID: subHRA000680

Submitter > Study Information > Data Access Overview > Metadata > File > Overview & Submit

General Package

Upload the metadata file using Excel format.

Please select file  Upoad

[Download Excel Template General 1.0 spreadsheet](#) edit, save and then upload the modified template.

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us at [gsa@big.ac.cn](mailto:gsa@big.ac.cn).

**研究信息中的研究类型 (Study Type)**

**研究类型**      **适用元数据模板**

疾病研究 (Disease study)	General_1.0.xlsx
队列研究 (Cohort study)	General_1.0.xlsx
细胞系相关研究 (Cell line related study)	Cellline_1.0.xlsx
临床病原体研究 (Clinical pathogen)	Clinical_pathogen_1.0.xlsx
人体相关宏基因组研究 (Human associated metagenome)	Human_associated_metagenome_1.0.xlsx

元数据的模板文件根据研究类型 (Study Type) 进行自动匹配，以适应不同的研究方向需求

# GSA-Human提交：元数据信息

以General\_1.0.xlsx为例：主要用于 疾病研究 和 队列研究



## 常见问题1：对应关系

1:N

单个个体可以多对应多个样本，请**务必注意每个样本所属关系**。这是最容易出现问题的地方。  
因为个体为核心，**一旦出错，所有元信息都会受到影响**。

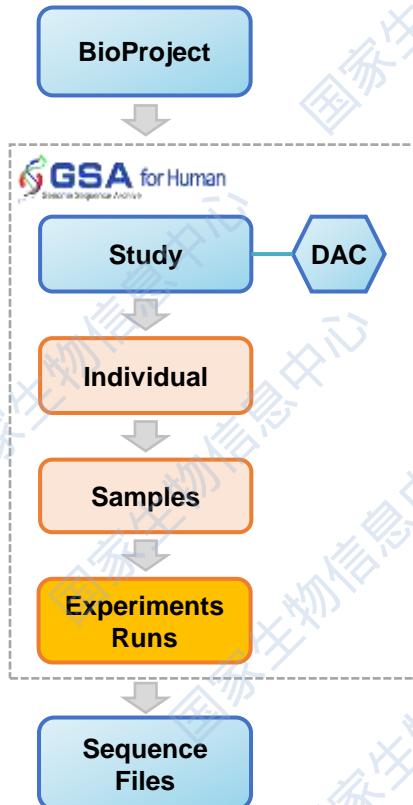
**常见问题2：敏感信息问题**  
individual和sample的名称中不可  
包括受试个体名字拼音全称或简称

**提示：**请确保信息的**完整度与准确性**。完整的信息可提供全面的背景  
和上下文，对于**管理员审核、审稿人审阅及未来数据使用**有重要帮助。  
如果已有列无法详细描述，可**自定义添加相关信息**。



# GSA-Human提交：元数据信息

以General\_1.0.xlsx为例：主要用于 疾病研究 和 队列研究



Experiment 表格									
ID	Experiment title	sample ID	Platform	Library Construction / Experimental Design	Library name	Strategy	Source	Selection	Layout
E1									
11	Read length for mate1(bp)	12	Read length for mate2(bp)	13	Insert size (bp)	14	Nominal size (bp)	15	Nominal standard deviation (bp)
16	Planned number of cycles								

## 常见问题1：对应关系

1:N

如果单个样本采用多种测序策略，则样本对应多个Experiment；Experiment与Run之间对应关系也为一对多。

Run 表格							
ID	Run title	Experiment ID	Run data file type	File name 1	MDS checksum 1	File name 2	MDS checksum 2
R1							
9	Reference file name	10	MD5 for reference file	11	Assembly Name or Accession	12	Assembly Accession URL

## 常见问题2：敏感信息问题

Experiment和Run title不可包括患者名字拼音全称或简称；文件的名称也不可包含此类信息

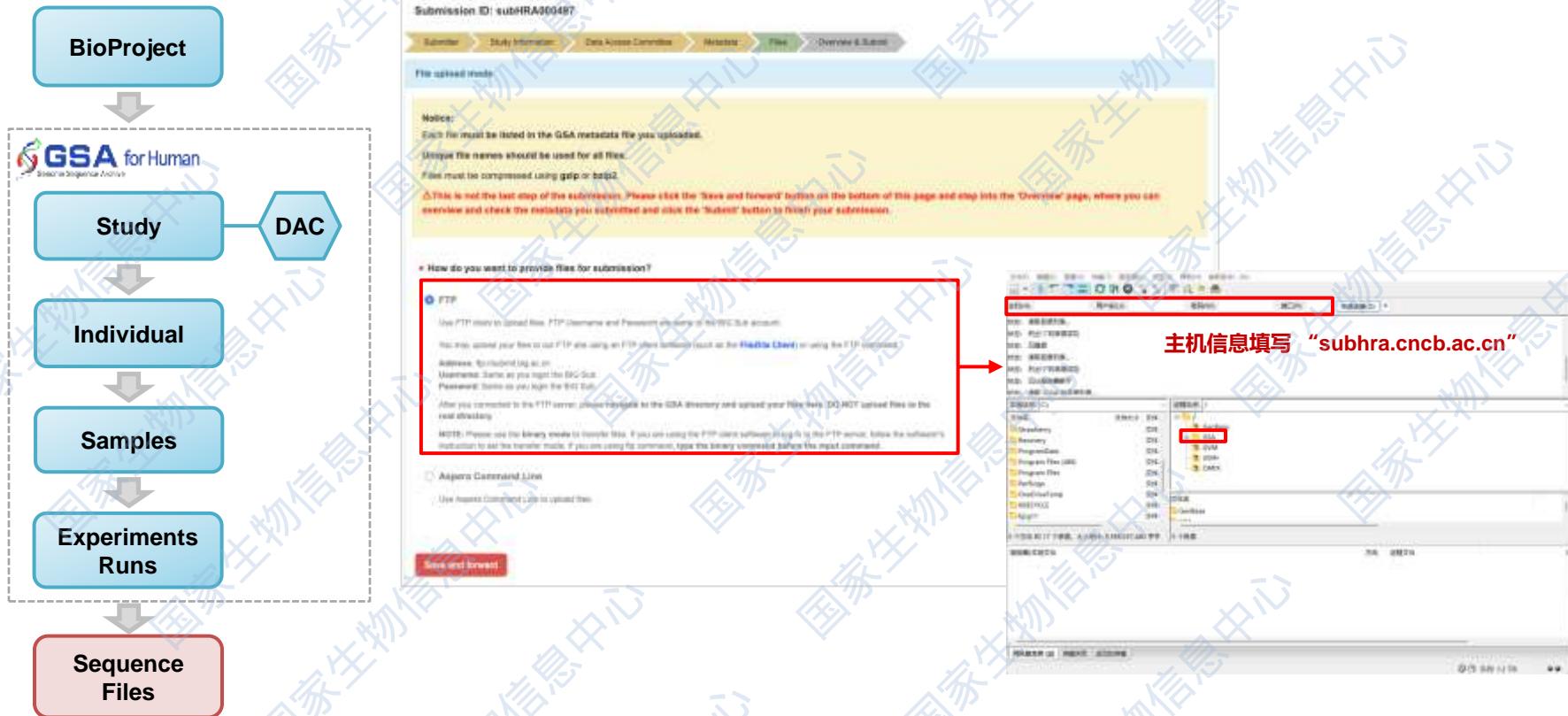
## 常见问题3：文件名称一致问题

文件名称必须与上传文件完全一致。如果不一致，后期质控时将系统无法定位，从而影响文件处理与归档效率

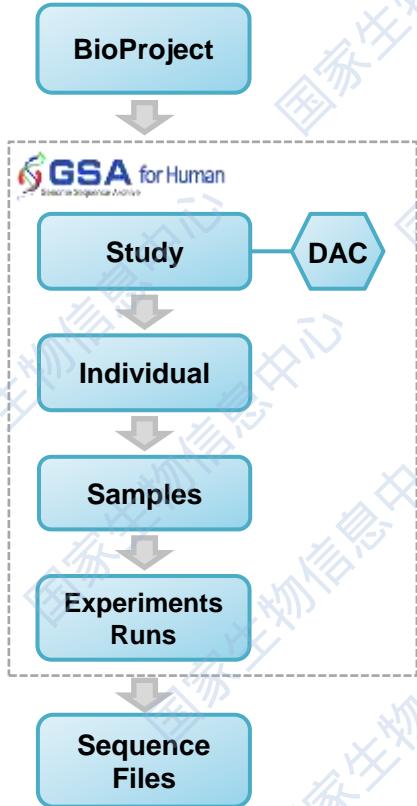
# GSA-Human提交：元数据信息



# GSA-Human提交：序列文件上传方式



# GSA-Human提交：概览 & 提交



The screenshot shows the GSA-Human submission interface. At the top, it displays 'Study ID: GSE11827' and 'Submitter: gsa@NCI-GDC'. Below this, the 'Study Information' section includes fields for 'Study Title' (GSA\_1001\_2000001), 'Data type' (Human genome), and 'Data access' (Public access). The 'Individual Information' section lists 'Individual' (Individual 1) and 'Individual 2' (Individual 2). The 'Data Access Control' section shows 'Accession' (GSE11827) and 'Name' (GSA\_1001\_2000001). The 'Experiment Information' section details 'Experiment ID' (GSE11827\_1), 'Instrument' (Illumina HiSeq X Ten), and 'Platform' (Illumina HiSeq X Ten). The 'Sample Information' section lists 'Sample ID' (GSE11827\_1\_01\_01) and 'Sample Name' (GSE11827\_1\_01\_01).



请务必点击“提交”，提交后进入数据审核流程，请耐心等反馈信息

# GSA-Human提交：等待审核

## 数据审核流程

### 第一阶段：人工审核

账号的身份认证  
是否包含隐私信息  
关联关系是否合理  
信息的合理与完整

元数据信息审核周期：  
**1-2个工作日**

### 第二阶段：数据关联与质控

数据文件关联  
文件格式检查  
质量信息分析

文件处理周期与文件  
**大小与数量** 有关

Accession	Title	Submission ID	Release Date	Package	Status	Access(Request)	Operation
Unassigned	aaaaaaa	subHRA000350	2021-09-30	General	<i>Checking detail Confidential</i>	Controlled(0)	<span>Delete</span> <span>Update</span>
HRA000165	test	subHRA000201	2024-03-22	Cell Line	<i>Checked OK Confidential</i>	Open	<span>ReleaseNow</span> <span>Share</span> <span>Update</span>

- 数据审核与文件处理的状态可通过网页进行查看。如遇到任何问题，欢迎加入我们的QQ群（548170081）交流，或通过电子邮件[gsa@big.ac.cn](mailto:gsa@big.ac.cn)联系我们
- 数据归档后，将分配唯一的序列号（Accession）。

# 设置共享链接与数据引用

The screenshot illustrates the workflow for sharing data. It starts with a 'New Submission' table where a submission has been made with Accession HRA000049. A red box highlights this accession number. An arrow points down to another 'New Submission' table, which shows the same entry but with a 'Share Now' button highlighted in red. A large downward arrow indicates the progression. To the right, a 'Review Link Options' dialog box is shown, also with a red border around it. This dialog box contains fields for 'Release Date' (set to '3 months') and 'Public data access link' (checkbox checked). A note at the bottom states: 'NOTE: Researchers can access this sequence through the URL above if they check the box to generate the link as required.' A 'Cancel' button and a green 'OK' button are at the bottom.

New Submission							
Accession	Submission ID	Release Date	Package	Status	Access	Request	Operation
HRA000049	subHRA000200	2022-04-17	General	Checked OK Confidential	Controlled	0	<button>Reassess Now</button> <button>Share Now</button>

New Submission							
Accession	Submission ID	Release Date	Package	Status	Access	Request	Operation
HRA000049	subHRA000200	2022-04-17	General	Checked OK Confidential	Controlled	0	<button>Reassess Now</button> <div style="border: 1px solid red; padding: 2px;">Shared URL: <a href="https://mgdc.cncb.ac.cn/gsa-human/">https://mgdc.cncb.ac.cn/gsa-human/</a></div> <button>Cancel Share</button>

- 用户可自行设定共享链接，决定其有效期以及是否允许获得者查看数据
- 注意，此链接不应对外公布，亦不应在任何文章中引用

当您成功提交数据到GSA-Human并通过审核后，请在您要发表的论文中添加如下语句：

*The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2021) in National Genomics Data Center (Nucleic Acids Res 2024), China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human: [HRAxxxxxx](#)) that are publicly accessible at <https://mgdc.cncb.ac.cn/gsa-human>.*

请按照以下格式引用我们的文章：

- The Genome Sequence Archive Family: Toward Explosive Data Growth and Diverse Data Types. *Genomics, Proteomics & Bioinformatics* 2021, 19(4):578-583 <https://doi.org/10.1016/j.gpb.2021.08.001> [PMID=34400360]
- Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. *Nucleic Acids Res* 2024, Jan 5;52(D1):D18-D32. <https://doi.org/10.1093/nar/gkad1078> [PMID=38018256]

# GSA-Human数据访问权限

访问类型	数据情况	访问权限分类	元信息查看	元信息下载	序列文件下载
公开访问数据	• 符合开放共享条件*	可获取	全部	全部	可下载
	• 不符合开放共享条件	不可获取	部分	部分	不能
受控访问数据	• 符合开放共享条件*	可获取	全部	全部	可下载
	• 已进行数据申请	可获取	部分	部分	不能
	• DAC审核通过				
	• 符合开放共享条件*	可获取	部分	部分	不能
	• 未进行数据申请或DAC审核未通过	不可获取	部分	部分	不能
	• 不符合开放共享条件				

\*注：符合开放共享条件，是指数据已经获得科技部备案，或者数据为国外数据，或者数据为临床病原体、古人类或化石、商用细胞系数据

The screenshot shows a table with columns: 编号 (ID), 样品 (Sample), 质谱库编号 (Library ID), 个体编号 (Individual ID), 样本编号 (Sample ID), and 实验编号 (Experiment ID). The first row is highlighted with a red box.

编号	样品	质谱库编号	个体编号	样本编号	实验编号
HHR194327	PY12	HHR194327	HHR194327	HHR194327	HHR194327

元数据信息部分可看

This screenshot shows a detailed view of a sample entry. It includes fields for Sample ID (HHR194327), Individual ID (PY12), Library ID (HHR194327), Sample ID (HHR194327), Experiment ID (HHR194327), and various sequencing details like raw data files (raw1.fq, raw2.fq) and processed data files (basecall.fq, fastq.fq).

Sample ID	HHR194327	Individual ID	PY12	Library ID	HHR194327
Sample ID	HHR194327	Individual ID	HHR194327	Library ID	HHR194327
Experiment ID	HHR194327	Individual ID	HHR194327	Library ID	HHR194327
Raw Data	raw1.fq raw2.fq	Basecall	basecall.fq	Fastq	fastq.fq

元数据信息全部可看

# GSA-Human数据浏览与获取

The screenshot shows a table of datasets from the GSA-Human platform. Each row contains information such as dataset ID, title, owner, status, and download links. A red box highlights the first dataset, "HRV202046", which is marked as "Open". Another red box highlights the "Open" status column across the entire table.

ID	Title	Owner	Status	Links
HRV202046	Singapore International Reference Database of preprint medical research papers related to SARS-CoV-2 infection	National Clinical Research Council for Infectious Diseases	Controlled	360 116 226 2024-01-11
HRV202047	lncRNAseq of liver cancer	Peking University Health Sciences Center	Controlled	146 98 80 2024-05-10
HRV202048	miRNA-seq of gastric cancer	Institute of Cell Biology, Cognition and Brain Sciences	Controlled	140 94 81 2024-03-20
HRV202049	Global Characterization of CD4+ Immature CD45 Low Cells in Peripheral Blood and Spleen Tissues of COVID-19, asymptomatic, and SARS-CoV-2 positive Rehabilitation Patients by Single-cell Sequencing	Fudan University Medical College, Chinese Academy of Medical Sciences & Peking Union Medical College	Controlled	123 98 82 2024-03-10
HRV202050	Discriminating and tracking Unmeasured variables via systematic CRISPR screens	Macquarie University	Open	2024-05-18
HRV202051	2024-05-08 2024-05-08	Wuhan University	Open	2024-05-18

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

This screenshot shows the detailed view for the dataset "HRV202046". It includes a summary section with the dataset's name, URL, and access status, followed by a large panel displaying various data visualization tools like histograms and scatter plots.

可获取栏中访问方式为“Open”的数据集，任何人都可在线查看完整信息并下载数据文件

# GSA-Human数据浏览与获取

The screenshot shows a list of datasets in the GSA-Human interface. Each dataset entry includes a red-bordered box around the '可获取' (Accessible) column, which contains a green button labeled '访问' (Access). An arrow points from this column to the detailed view shown in the next panel.

ID	Name	Type	状态	大小 (GB)	更新时间	操作
HUA000001	National Clinical Research Center for Infectious Diseases	HEC000001	Controlled	100	2024-05-10	
HUA000002	Beijing University Health Science Center	HEC000002	Controlled	140	2024-05-10	
HUA000003	mRNA-seq of glioma cancer	HEC000003	Controlled	160	2024-05-10	
HUA000004	Initial Characterization of CTCs in Tumors Cells Based on Phenotypic Features of ALKBH, program and ACTA2, and Their Relationship with Patients by Single-cell Sequencing	HEC000004	Controlled	180	2024-05-10	
HUA000005	Description and tracking of immunotherapy response via systematic CRISPR screens	HEC000005	Open	200	2024-05-10	
HUA000006	2DMA-seq Data of combined Cdk4/2, Myb and Cdk7 gene edited cell lines	HEC000006	Open	200	2024-05-10	

The screenshot shows a detailed view of a dataset access request. A red box highlights the '申请审核' (Review Application) section, which contains a red arrow pointing to the '访问方式' (Access Method) dropdown menu set to 'Control'. Another red box highlights the '数据申请' (Data Application) section, which states '仅允许PI申请下载 (与数据提交者相同的条件)' (Only PI can download data (with the same conditions as the data submitter)).

• 可获取栏中**访问方式为“Control”**的数据集，申请审核制访问

• 数据申请：仅允许PI申请下载（与数据提交者相同的条件）

• 申请审核

- 系统管理员将对申请者身份进行形式审查
- DAC对数据申请做最终审核，并决定是否授权

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

# 受控数据申请

Basic Information > Research information > Study information > Overview > Agreement

## Basic Information

\* Request title  
input short title

填写本次申请的标题

## PI Information

Email	chenlt@big.ac.cn
First Name*	Tingting
Middle Name	
Last Name	Chen
Country/Region	China
Institute/Organization	Beijing Institute of Genomics, Chinese Academy of Sciences <small>中国科学院生物化学与细胞生物学研究所</small>
Department	BIG Data Center
Title/Position	

**Save and forward**

# 受控数据申请

New Request

Request Study: HRA000067

Title: Multi-omics Characterization of Gastric Cancer Evolution under Neoadjuvant Chemotherapy

Basic information

Research information

Study information

Overview

Agreement

收集数据申请者利用该数据集拟开展的研究的相关信息

Research information

\* Research Activity Title

a short title for the research activity

研究活动名称

Choose research period

2 years

研究周期

\* Research purpose

研究的目的

\* Methods & techniques

会使用到的方法和技术

Save and forward

# 受控数据申请

New Request  
Request Study: HRA000067  
Title: Multi-omics Characterization of Gastric Cancer Evolution under Neoadjuvant Chemotherapy

Basic information Research Information Study Information Overview Agreement

## Requested study

### Study Accession

HRA000067

### Study Title

Multi-omics Characterization of Gastric Cancer Evolution under

### Uses and Limitations

General purpose use

### Whether or not the confirmation of the data user's institution is required?

NO

### Data Access Agreement

1 / 6

研究团队。

I have read the data access agreement, and I agree it

请仔细阅读数据访问协议 (Data Access Agreement)

Save and forward

确认申请的数据集的信息和数据提交者的要求

# 受控数据申请

## New Request

Request Study: HRA000067

Title: Multi-omics Characterization of Gastric Cancer Evolution under Neoadjuvant Chemotherapy

Basic information Research information Study information Overview Agreement

提供对本次数据申请信息的整体概览

### Overview of the request

#### Basic Information

Request title

Test study

#### PI Information

Email chentt@big.ac.cn

First Name Tingting

Middle Name

Last Name Chen

Country/Region

Institute/Organization Beijing Institute of Genomics, Chinese Academy of Sciences

Department BIG Data Center

Title/Position

# 受控数据申请

## New Request

Request Study: HRA000067

Title: Multi-omics Characterization of Gastric Cancer Evolution under Neoadjuvant Chemotherapy

Basic information

Research information

Study information

Overview

Agreement

提供对本次数据申请信息的整体概览

### Overview of the request

#### Basic Information

Request title

Test study

#### PI Information

Email

chentt@big.ac.cn

First Name

Tingting

Middle Name

Last Name

Chen

Country/Region

Beijing Institute of Genomics, Chinese Academy of Sciences

Institute/Organization

BIG Data Center

Department

Title/Position

# 受控数据申请

New Request  
Request Study: HRA000067  
Title: Multi-Omics Characterization of Gastric Cancer Evolution under Neoadjuvant Chemotherapy

Basic information Researcher information Study information Overview Agreement

Agreement.

I have read the following GSA-Human data policy, and I agree it.

确认协议内容

1 Rights of data user

- 1) The data user can make applications for the data needed through the GSA-Human system.
- 2) The data user can download the data in certain time after obtaining the approval of accessing the dataset.
- 3) The data user has the intellectual property of the subsequent research result based on the dataset.
- 4) The data user can download the approved data by tools like ftp or Aspera, which is designated by GSA-Human.

2 Responsibilities of data user

- 1) The data user should use the downloaded dataset in the promised research scope of application. The dataset is restricted to research purpose, and can only be used for the research group and its research collaborators. Dataset distribution to other person is not allowed and cannot be used to identify individuals.
- 2) The data user should verify the quality, content and scientific validity of the downloaded dataset by his/her own.
- 3) If the data user publish a paper using the downloaded dataset from GSA-Human, the data accession number must be indicated in the paper with the note that the data can be acquired from GSA-Human.

Choose signature type

Online signature

Offline signature

上传电子签名图片  
使用鼠标签名  
自动生成签名

Save and forward

**签署数据访问协议：**  
数据申请者可通过**在线生成签名或离线手写签名**两种方式签署数据访问协议

**注意：**如果数据提交者要求数据申请者提供单位盖章，只能通过离线方式签署协议

# 受控数据申请

Home / My Request

My Request List

New Request

Request Number	Request ID	Request Title	Request Date	Study Accession	Status	Expire Date	Operation
HREQ000133	req000000306	Request download HRA000067 dataset	2020-07-09	HRA000067	Finished   Waiting for Curation	Use expire 2022-07-03	<a href="#">Cancel</a> <a href="#">Modify</a> <a href="#">Review History</a>

My Request List

New Request

Request Number	Request ID	Request Title	Request Date	Study Accession	Status	Expire Date	Operation
HREQ000133	req000000470	test2	2020-07-02	HRA000219	Approved download	Use expire 2022-07-02 Download expire 2020-08-01	<a href="#">Review History</a>

Showing 1 to 2 of 2 entries

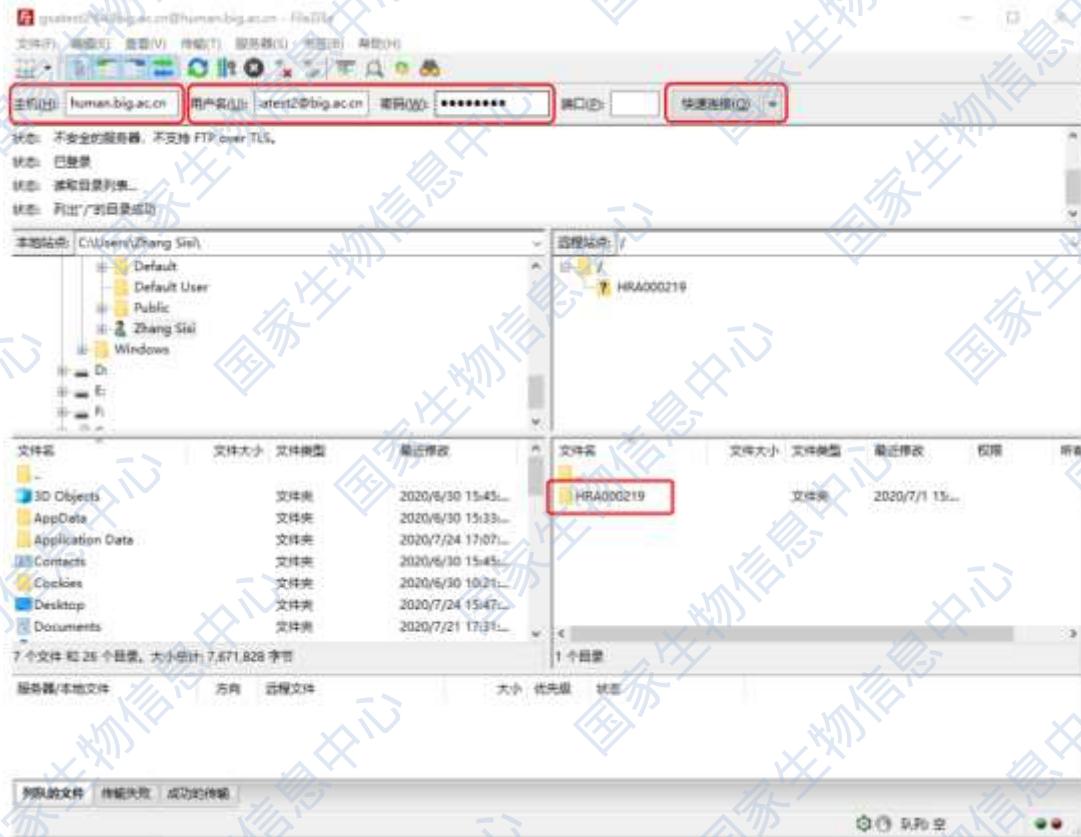
Search:

Previous 1 Next

完成申请提交后，列表中状态（Status）变为 Finished | Waiting for Curation。

- DAC 审核通过，申请的状态变为“Approved Download”。
- DAC 审核不通过，状态变为“Rejected”。可点击“Review History”在线查看反馈意见。

# 受控数据申请-数据下载



# 用户服务



Email: gsa@big.ac.cn  
QQ Group: 548170081