



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

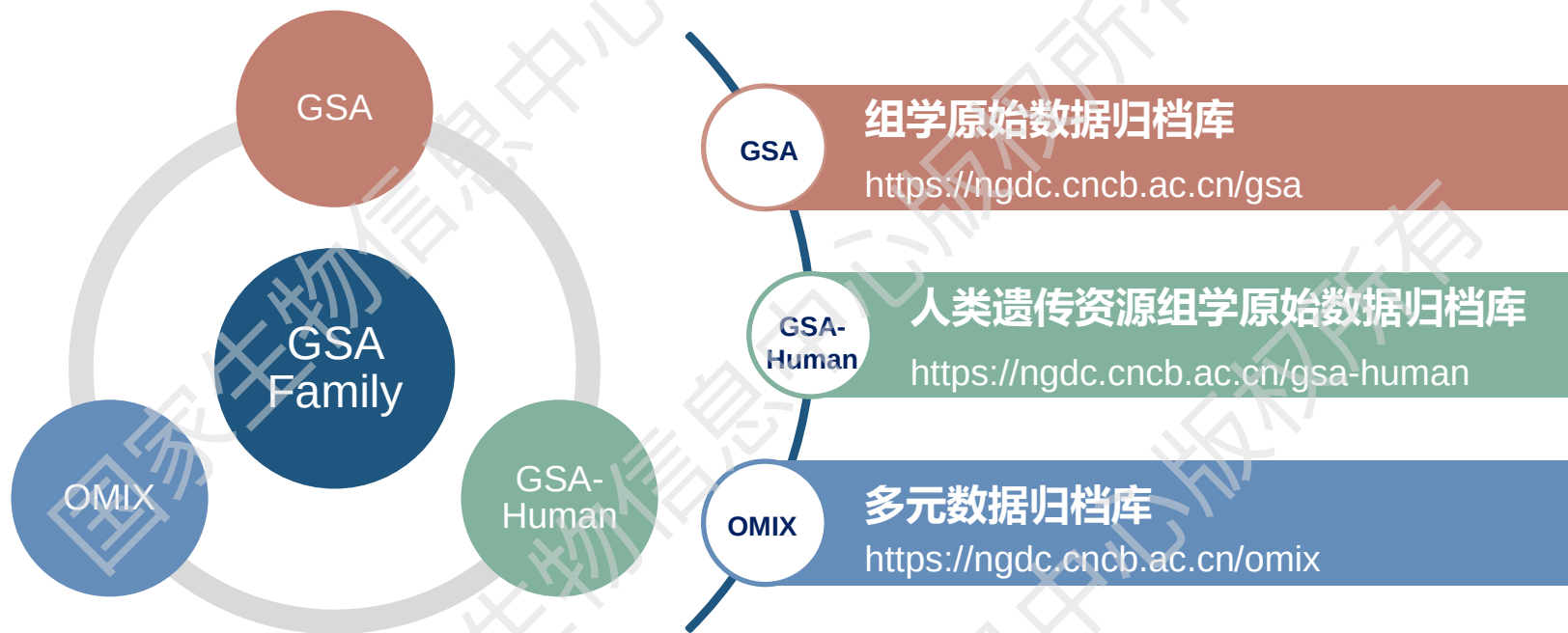
王彦青

2025年5月9日

提纲

- 一 GSA Family介绍
- 二 GSA数据提交及共享
- 三 GSA-Human数据提交及共享
- 四 OMIX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



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

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

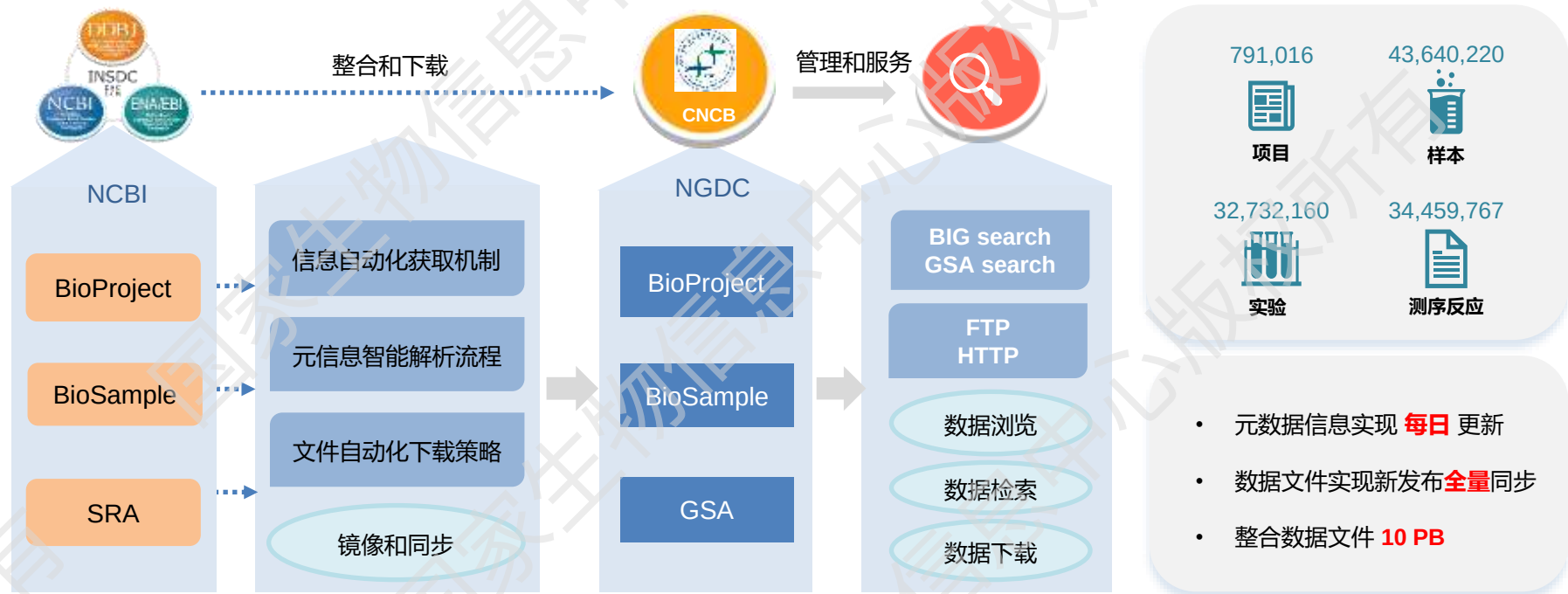
ENA
European
Nucleotide
Archive
@EBI

International Nucleotide Sequence Database
Collaboration (INSDC)

The International Genome Sequence archive databases

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



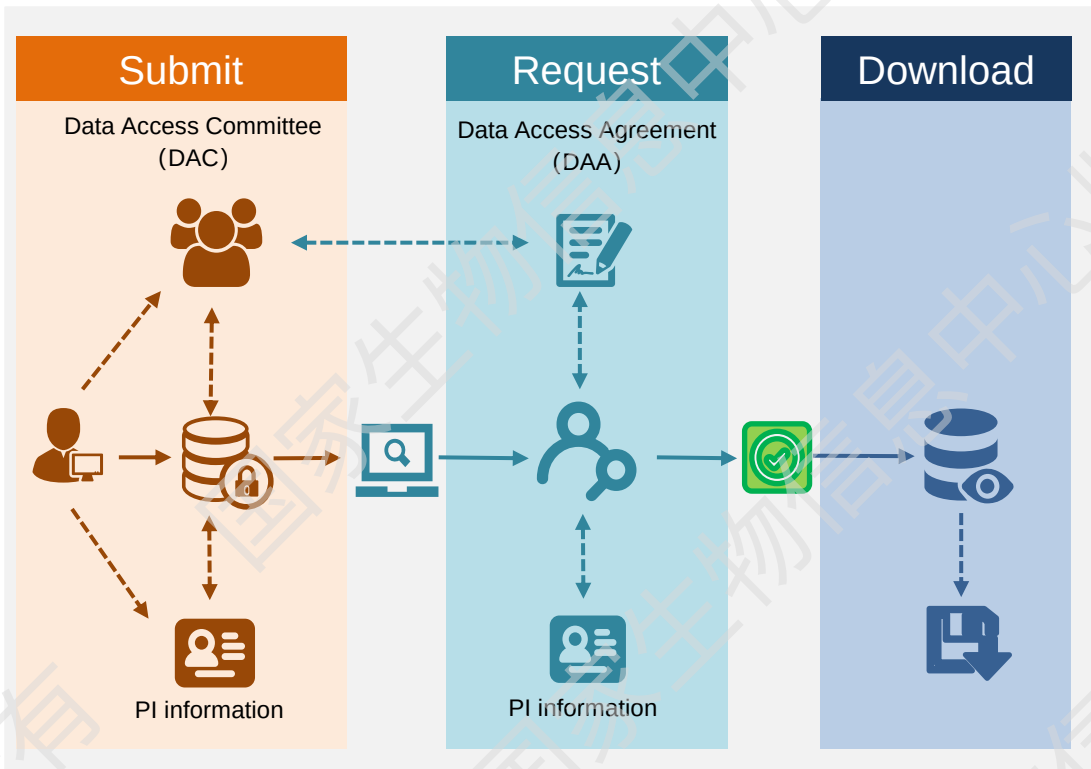
GSA-Human: 人类遗传资源组学原始数据管理



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>72万人数据
 - 归档数据量**50 PB**
 - 支持**1038篇**文章发表

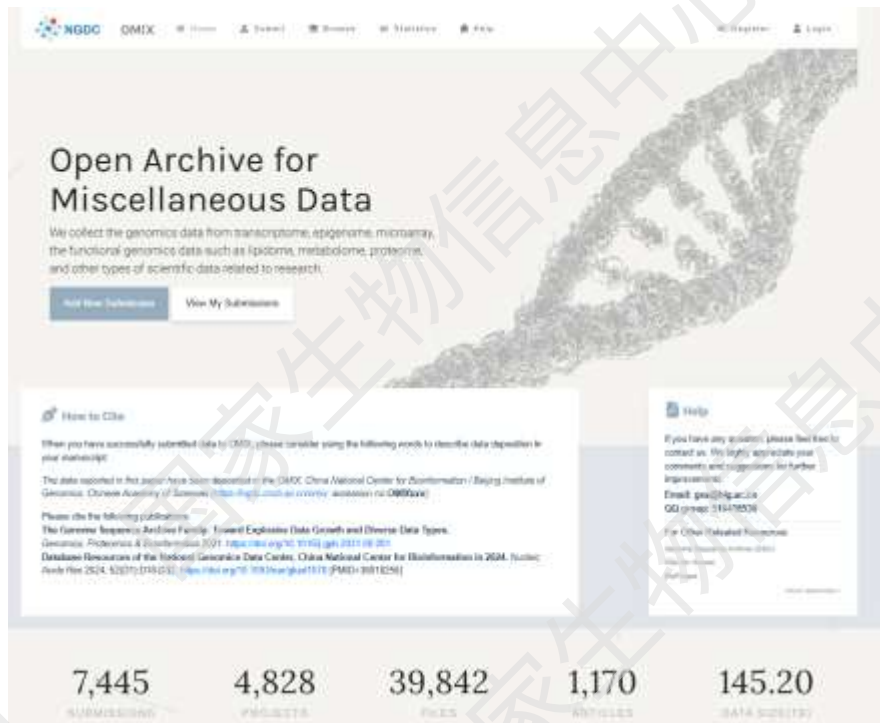
GSA-Human系统核心功能要素



- 支持**8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



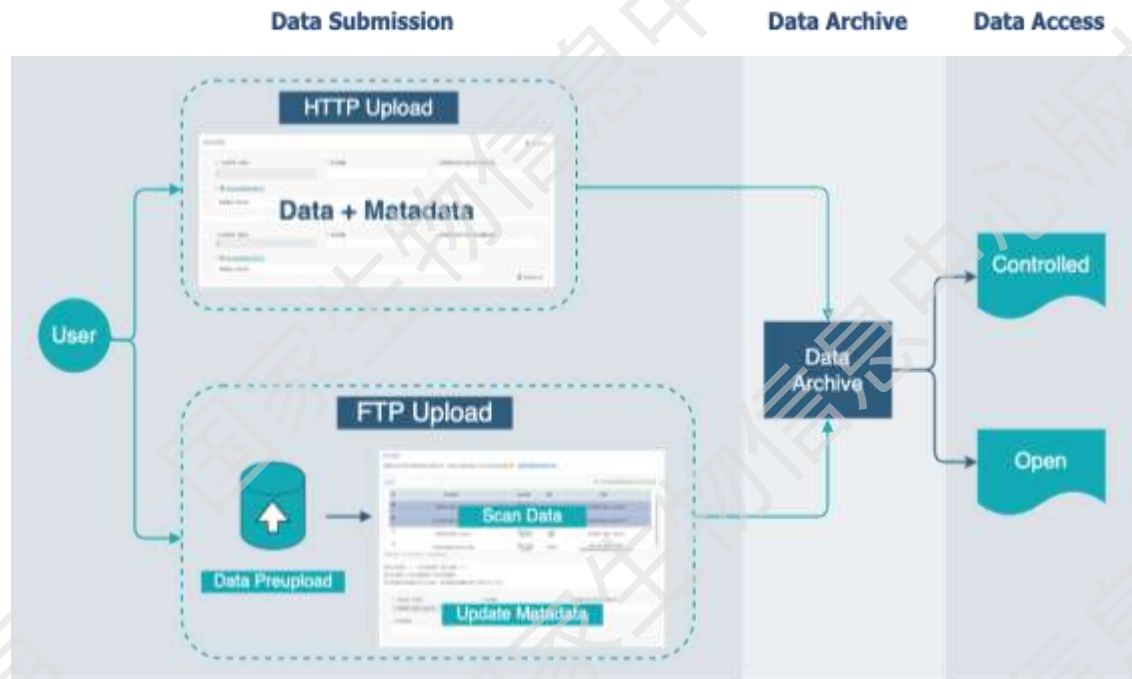
Chen, Tingting et al. *Genomics Proteomics Bioinformatics*, 2021



- 已通过 [FAIRSharing](https://fairsharing.org) 和 re3data.org 认证
- 获得 369 个国际期刊认可，支撑文章 1170 篇



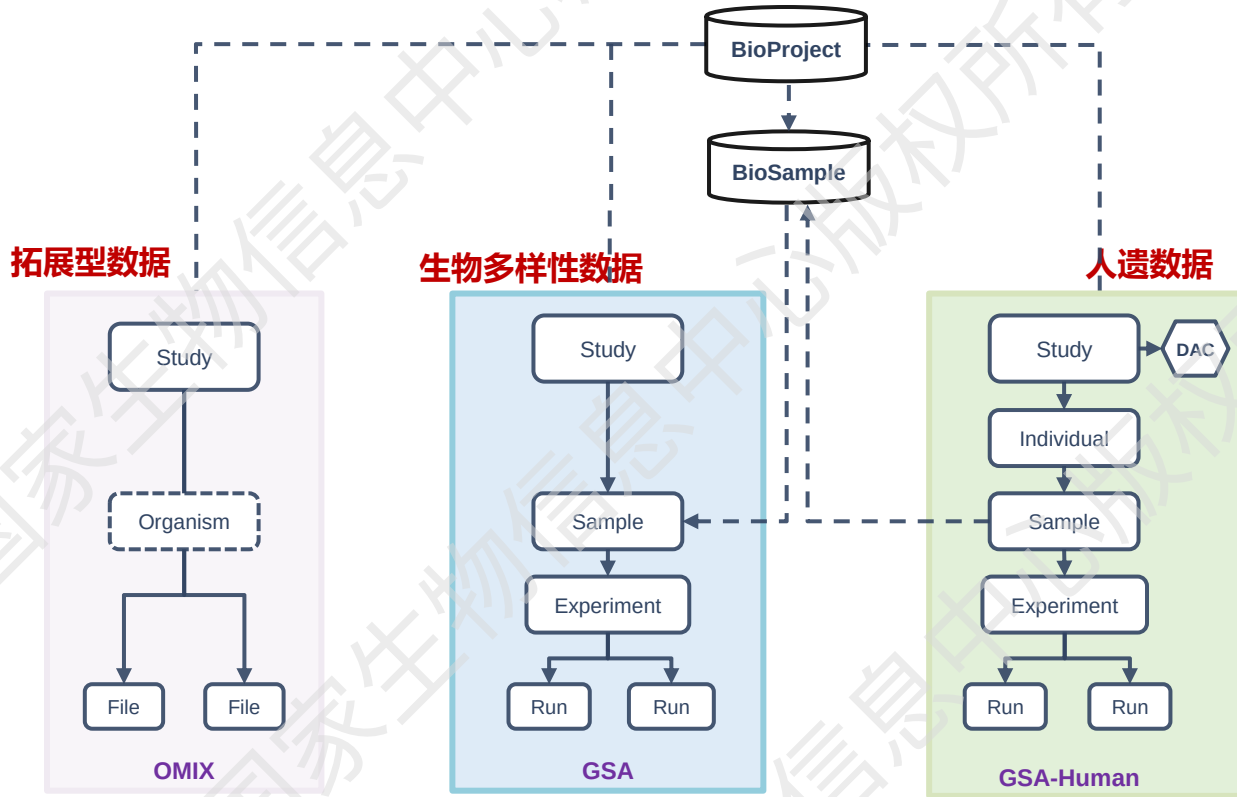
应用灵活可扩展的管理模式，支撑多模态数据汇交共享



- 支持**10**个大类，**33**个小类，超过**100**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**)

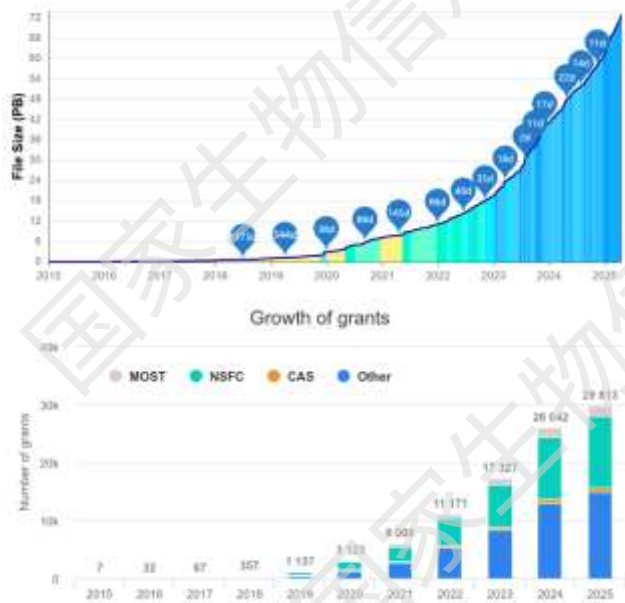


GSA Family数据模型



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 **73 PB**，支撑发表文章**4500 余篇**
- ❑ 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2天**



>180万 独立IP
>700万 次下载

> 200 国家/地区
> 19PB 下载量

获得国际著名出版商认可



Research data policy

- Data availability statements
- Data repository guidance
- Sensitive data
- Data policy FAQs
- Research data helpdesk

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

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

Genetic variation data

Repositories

Any INSOC member repository

Genome Sequence Archive (GSA)

dbSNP (human variations less than 50bp)
dbVar (human variations greater than 50bp)

European Variation Archive (EVA) (all species)
Genome Sequence Archive for Human (human variations)

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



Data in Brief

Submit your Paper

View Articles

A Guide for authors

Track your paper

Public repositories to store and find data

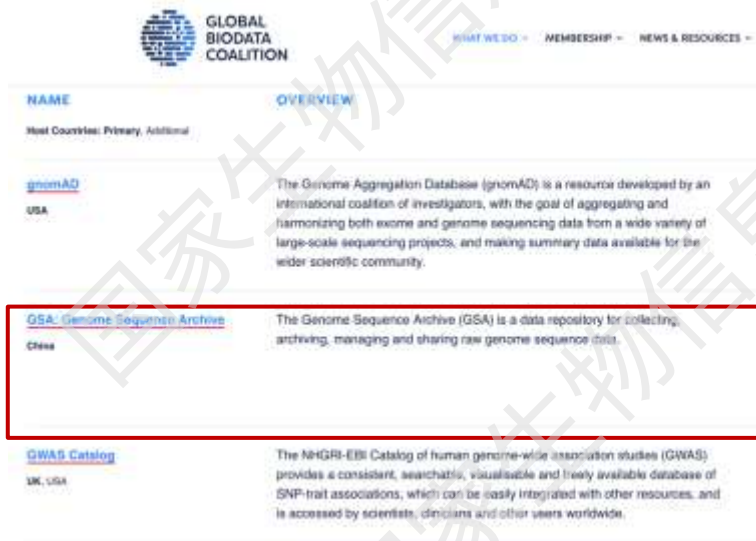
2. Specialized Repositories are recommended. Please note that this list is not exhaustive. We encourage the use of Re3data and FAIRsharing to find additional specialized repositories in a specific discipline.

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



GLOBAL BIODATA COALITION

NAME OVERVIEW

Host Countries: Primary, Additional

[gnomAD](#)

USA

The Genome Aggregation Database (gnomAD) is a resource developed by an international coalition of investigators, with the goal of aggregating and harmonizing both exome and genome sequencing data from a wide variety of large-scale sequencing projects, and making summary data available for the wider scientific community.

GSA: Genome Sequence Archive

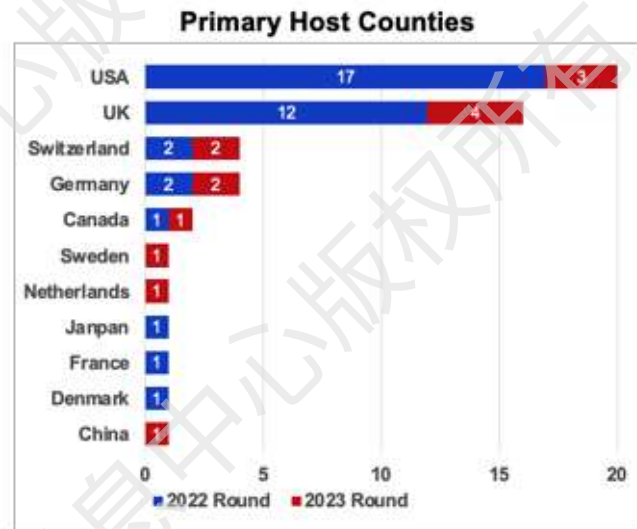
China

The Genome Sequence Archive (GSA) is a data repository for collecting, archiving, managing and sharing raw genome sequence data.

[GWAS Catalog](#)

UK, USA

The NHGRI-EBI Catalog of human genome-wide association studies (GWAS) provides a consistent, searchable, visualisable and freely available database of SNP-trait associations, which can be easily integrated with other resources, and is accessed by scientists, clinicians and other users worldwide.



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

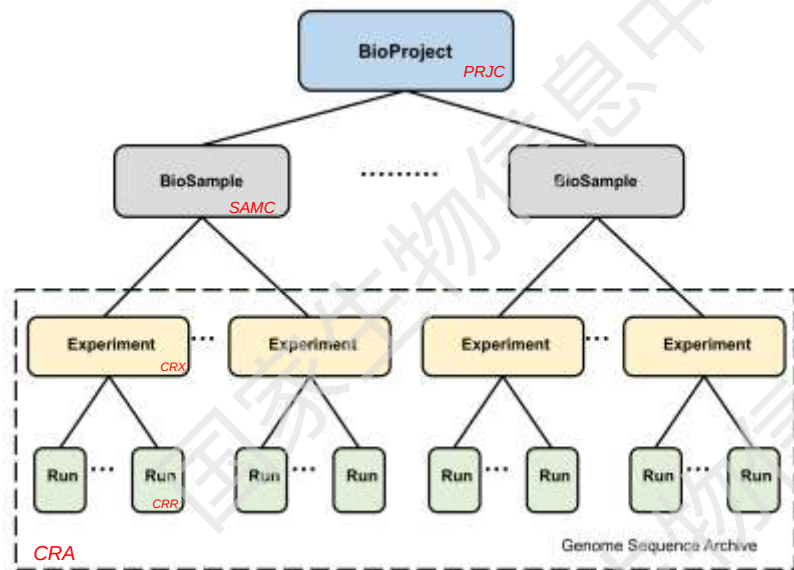
-1-



提纲

- 一 GSA Family介绍
- 二 GSA数据提交及共享**
- 三 GSA-Human数据提交及共享
- 四 OMIX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。但不包括致病微生物样本。在基因组数据提交时选择Metagenome/Environmental Sample (GSC MIMS unsupported)或Metagenome/Environmental Sample (GSC MIMS compliant)。

☐ Animal

用于模式生物或多细胞生物的多细胞样本或细胞系样本。

☐ Model organism or animal

用于模式生物或多细胞生物的多细胞样本或细胞系样本。如大鼠、小鼠、蜜蜂、斑马鱼、果蝇或其他哺乳动物数据。

☐ Animal for TCM syndrome

用于中医证候动物模型研究。

☒ Human

人类遗传资源相关组学原始数据提交到GSA for Human数据库。

☒ Plant

用于植物样本或植物细胞系样本。

☐ Virus

用于所有与疾病相关的病毒样本。病毒样本应归类为Clinical or host-associated pathogen。

☐ Metagenome/Environmental Sample (GSC MIMS unsupported)

目前用于不适用于Metagenome/Environmental Sample (GSC MIMS compliant)的宏基因组样本数据。物种名必须是一个metagenome，请予以未分类的序列开始，字头以“metagenome”结尾。

☐ Metagenome/Environmental Sample (GSC MIMS compliant)

用于宏基因组样本数据。下设三个小组分别接收来自人类肠道、土壤和水环境宏基因组数据。即human-gut, soil, water。物种名必须是一个metagenome，请予以未分类的序列开始，字头以“metagenome”结尾。

☐ human-gut

☐ soil

☐ water

<https://www.gensc.org/pages/standards-intro.html>

病原体相关（如新冠病毒）样本，首选此类型。

人类型相关样本，请提交到GSA-Human。

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

数据递交入口: BIG-Sub



<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 *

Email Address *

City *

State / Province *

Postal Code *

Country / Region *

Institutional Information

Institute / Organization *

Department *

Laboratory *

Title / Position *

Research Area *

Check Code *

Submit

Reset

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@big.ac.cn>
时 间: 2017年10月10日(星期二) 下午2:50
收件人: 用户

Dear user,

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

<http://sso.big.ac.cn/register/active.action?mailAddr=...&token=7f6934f45681c87a93757662c8784c22>

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

DatabasesTools

 **NGDC** Central Authentication Service

Enter your Username and Password

Email

gsatest1@big.ac.cn

Password

[Forgot password?](#)

Check code

36bb

36bb

☐ Keep me signed in

Login

Reset

Register

Central Authentication Service

[BioProject](#)

[BioSample](#)

[BioCode](#)

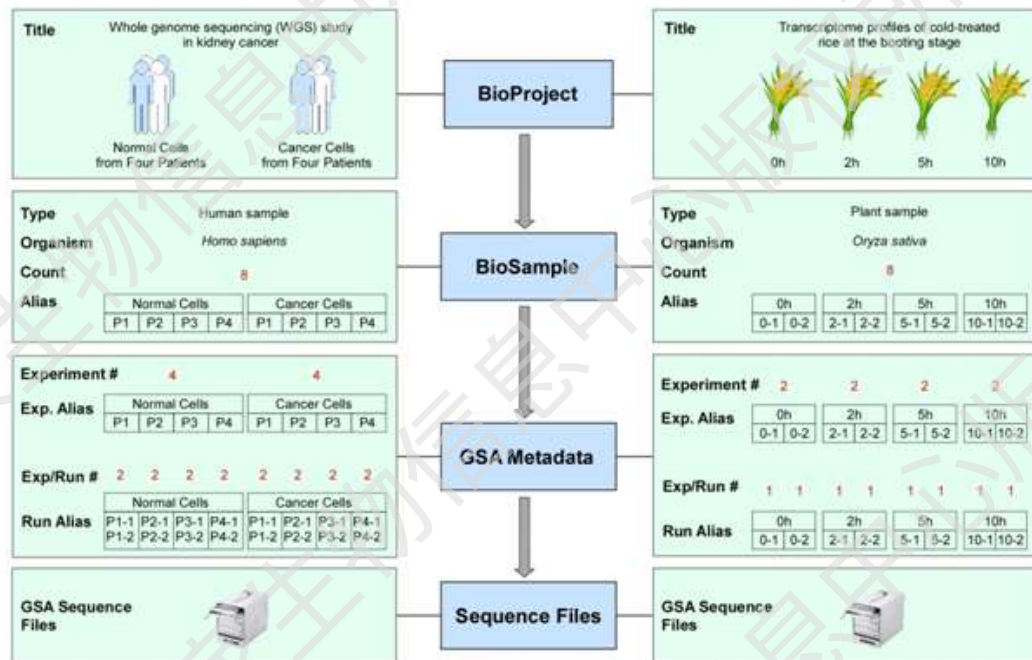
[GSA for Human](#)

[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 BIG Sub Submission Portal interface. At the top, there's a navigation bar with links for Databases, Tools, Standards, Publications, and About. Below this is the BIG Sub logo and the text 'BIG Submission Portal'. A secondary navigation bar includes Home, Documentation, Login, and Register. The main content area is titled '请选择您要提交的数据类型' (Please select the type of data you want to submit). Below this title, there's a paragraph explaining the system: '生物数据递交系统(BIG Submission, BIG Sub)是国家基因组科学数据中心生物数据统一汇入口, 为用户提供一站式数据递交服务。' (The Biological Data Submission System (BIG Submission, BIG Sub) is the unified submission portal for biological data at the National Genomics Science Data Center, providing users with a one-stop data submission service). The main area contains several colored boxes representing different data types: Genome Sequence Archive (GSA), GSA for Human, Genome Warehouse (GWH), BioProject, BioSample, Genome Variation Map (GVM), BioCode, and Database Commons. The BioProject box is highlighted with a red border and a red callout box pointing to it. The callout box contains the text 'BioProject 递交入口' (BioProject Submission Entry). The URL <https://ngdc.cncb.ac.cn/gsub> is displayed at the bottom right of the screenshot.

CNCB-NGDC

Databases Tools Standards Publications About

BIG Sub
BIG Submission Portal

Home Documentation Login Register

请选择您要提交的数据类型

BIG Sub Quick Start Guide BIG Sub 使用指南

生物数据递交系统(BIG Submission, BIG Sub)是国家基因组科学数据中心生物数据统一汇入口, 为用户提供一站式数据递交服务。

Genome Sequence Archive (GSA)
组学原始数据汇交、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

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

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

BioProject
递交入口

创建BioProject – 提交列表

**BIG Sub**
BIG Submission Portal

中文English

HomeDocumentation

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

BioProject Quick Start GuideBioProject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
PRJCA003418	subPRO004980	GSA_TEST_20200909	2021-09-09	finished	confidential	修改 删除
PRJCA002492	subPRO003649	TEST_1_GSATEST1	2020-04-30	finished	confidential	修改 删除
PRJCA001010	subPRO001500	project2	2019-12-31	finished	confidential	修改 删除
PRJCA001009	subPRO001499	project-test1	2019-12-31	checked OK	Public	

创建BioProject – 提交者信息

BiG Sub / BioProject / New BioProject

01 提交者信息 02 样本信息 03 项目类型 04 关联信息 05 数据提交

提交者信息

• 名	中间名	• 姓氏
GSA	middle name	BIGD
• 邮箱	备用邮箱	
gsatest1@big.ac.cn	secondary email	
• 单位	单位网址	• 部门
Beijing Institute of Genomics, CH	http://www.big.ac.cn	BIGD
手机	传真	
• 街道	• 城市	• 州/省
No.1 Beichen West Road, Chaoyi	Beijing	
• 邮编	• 国家/地区	
100101	China	

保存并进入下一项

创建BioProject – 基本信息

01 提交项目信息 02 基本信息 03 项目类型 04 出版信息 05 提交与提交

基本信息

发布日期:

☐ 审核通过后即可发布 (推荐)

☒ 指定日期发布

2021-10-31

(yyyy-mm-dd)

发布项目

项目标题

涉及领域

项目说明

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

[Add another project](#)

项目相关网站链接

网站信息描述 网站地址

[Add another](#)

关联项目

项目编号 相关描述信息

[Add another Accession](#)

外部数据库编号

项目编号 数据库名称

[Add another Accession](#)

保存并进入下一步

创建BioProject – 项目类型

BIG Sub BioProject New BioProject

01 提交表信息 02 基本信息 03 项目类型 04 出版信息 05 预览 & 提交

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（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
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

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
	gsatest1@big.ac.cn
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	Multisolate

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) 。

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

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

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	修改 删除
PRJCA003418	subPRO004980	GSA_TEST_20200909	2021-09-09	finished	confidential	修改 删除
PRJCA002492	subPRO003649	TEST_1_GSATEST1	2020-04-30	finished	confidential	修改 删除
PRJCA001010	subPRO001500	project2	2019-12-31	finished	confidential	修改 删除
PRJCA001009	subPRO001499	project-test1	2019-12-31	checked OK	Public	

数据递交 – 创建GSA



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

创建GSA – 提交列表

[Home](#) [Documentation](#) Welcome, GSA-

[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）的数据应归到同一个GSA数据集下；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到对应的GSA数据集下。
- 如果您在数据上传过程中遇到问题或发现任何系统报错，请通过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标题	发布日期	提交状态	操作
Unassigned	subCRA002947		2020-04-06	Unfinished at the General Info step Confidential	删除
CRA001076	subCRA001069	GSA-test	2019-12-31	Checked OK Public	

创建GSA - 提交者信息

Home Documentation Welcome, GSA -

BIG Sub GSA New GSA

01 提交者信息 02 项目信息 03 元数据信息 04 文件 05 提交与提交

提交者信息

• 名	中间名	• 姓氏
GSA	middle name	BIGD
• 邮箱	备用邮箱	
gsa1@big.ac.cn	secondary email	
• 单位	单位网址	• 部门
Beijing Institute of Genomics, CH	http://www.big.ac.cn	BIGD
• 手机	传真	
• 街道	• 城市	• 州/省
No.1 Beichen West Road, Chaoyi	Beijing	
• 邮编	• 国家/地区	
100101	China	

保存并进入下一步

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
预览 & 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. 如果引用这些数据与该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
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

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

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
用于模式生物或多细胞样本或细胞系样本，如大鼠、小鼠、果蝇、线虫、鱼类、两栖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，病原体应归类为Clinical or host-associated pathogen

☐ Metagenome/Environmental Sample (GSC MIMS unsupported)
目前用于不适用于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，下设三个小类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

[illegible]

CNCB-NGDC

创建GSA - 样本属性

临床病原微生物样本属性信息

• 已上传的BioSample批量提交文件

Pathogen_clcn-12.xlsx 33KB 删除 校验

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再重新上传。



临床病原微生物样本属性信息

• 已上传的BioSample批量提交文件

Pathogen_clcn-12.xlsx 33KB 删除 校验 ✖ ERROR:

下载错误文件: error.txt

Error in: Sample sheet
row 11, column 10: "geographic_location" ("China: Hubei: Wuhan") Format 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"

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



临床病原微生物样本属性信息

• 已上传的BioSample批量提交文件

Pathogen_clcn-12.xlsx 33KB 删除 校验 ✔ Checked OK.

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再重新上传。

保存并进入下一项

创建GSA - 序列元数据信息

BIG Sub GSA Submission: subCRA004261

01 提交信息 02 基本设置 03 样本类型 04 样本属性 05 元数据信息 06 文件 07 提交 & 提交

GSA元数据信息

• 请上传GSA批量提交文件

请选择文件

• 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 [s.g.GSA_Template.cn.xlsx](#).
更多帮助, 请查看 [Help](#).

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 HeliScopeII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

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,

3. 为节省您批量上传表格的审核时间, 请您

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

BIG Sub GSA Submission: subCRA004261

01 提交者信息 02 样本信息 03 样本类型 04 样本属性 05 元数据信息 06 文件 07 预览 & 提交

GSA元数据信息

• 已上传的GSA批量提交文件

GSA_Template.cn_gsafest.xlsx 31KB Checked OK.

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GSA批量提交；若文件审核未通过，请删除已上传文件并修改后再重新上传。

创建GSA - 序列文件上传方式

BIG Sub : GSA Submissions: subCRA002988

01 提交新数据 02 数据管理 03 样本管理 04 样本属性 05 子组管理 06 文件 07 提交 & 提交

文件上传方式

提示:
上传的文件文件名MD5码必须跟您在GSA Metadata表格填写的一致。
上传的文件名必须是唯一的。
FASTQ文件必须在原始上传。目前我们只接收gzip 或 fastq2格式的文件。

请选择文件上传方式

☒ FTP

使用FTP客户端软件上传文件。FTP账号跟您BIG Sub账号一致。
您可以通过客户端软件使用WinZilla使用FTP命令上传文件。
Address: ftp://submit.big.ac.cn
Username: Same as you login the BIG Sub
Password: Same as you login the BIG Sub
进入FTP后, 请进入GSA目录并把你文件上传到该目录下。请不要把文件上传到根目录下。该种做法会随服务器扫描不到您的文件。
注意: 请使用二进制模式上传文件。如果使用的是FTP客户端软件, 请参考软件的说明文档来设置传输模式。如果使用的是FTP命令上传文件, 请在输入input命令之前, 先输入binary命令, 来设置二进制传输模式。

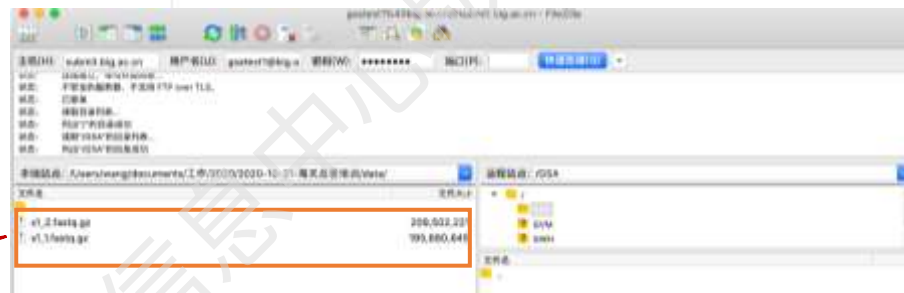
☐ Aspera Command Line

使用Aspera命令上传文件。

保存并进入下一项

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



文件上传方式

提示:

上传的文件名和MD5的必须跟您在GSA Metadata编辑器填写的一致。

上传的文件名必须遵循一例。

Asky文件必须上传到, 目前我们只接受getty 或 test2格式的视频文件。

• 请选择文件上传方式

FTP

遵循FTP客户端协议上传文件。FTP账号跟您GSA账号是一样的。

Aspera Connect Local

遵循Aspera Connect本地上传方式。

您可以用以下命令来上传文件。

`asperaconnect-PS3001 -path/to/keyfile -D1 -file -d -p path/to/asperaconnectapp/bin/asperaconnectkeyfile.exe`

Where:

`[path/to/keyfile]`:
 Microsoft Windows: C:\Program Files\Aspera\Aspera Connect\bin\asperaconnectkeyfile.exe
 Linux: /usr/bin/asperaconnectkeyfile

`[path/to/asperaconnectapp/bin/asperaconnectkeyfile.exe]`:
 Microsoft Windows: C:\Program Files\Aspera\Aspera Connect\bin\asperaconnectkeyfile.exe
 Linux: /usr/bin/asperaconnectkeyfile

Get the Key File

请为每一个文件夹创建一个新的子目录以存放所有的数据文件。建议创建一个文件夹。该文件夹应该包含以下的所有文件。该文件夹应该包含以下的所有文件。

由于我们上传文件的目录结构, 应该包含所有数据文件的子目录。

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

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

创建GSA - 概览 & 提交

项目名称: GSA / Submission: subGSA004261

01 提交信息 02 基本设置 03 样本信息 04 样本属性 05 实验设置 06 注释 07 提交 & 提交

GSA基本信息

提交编号	subGSA004261
标题	TEST_20201015
描述信息	TEST_20201015
发布日期	2021-10-31

提交者信息

提交者	GSA BIGD
单位	geninfo@big.ac.cn 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

元数据信息

样本类型	Clinical or host-associated pathogen
BlockSample数据文件	Pathogen_s1-s3.fastq
元数据文件	GSA_SampleSheet_manifest.xlsx

实验名称	测序平台	样本名称	生物名称	实验选择
experiment1	Illumina HiSeq 2500	Sample1	Severe acute respiratory syndrome coronavirus 2	untagged
Run名称		Run序列文件处理信息		文件类型
Run1	File: s1_1.fastq.gz File: s1_2.fastq.gz			fastq

Page 1 of 1 Page 1 of 1 Page 1 of 1 Page 1 of 1 Page 1 of 1 Page 1 of 1

提交 检查无误后，点击“提交”，完成本次提交

创建GSA – 元数据提交完成

GSA提交编号, 注意不能把这个号写到文章里

点击查看数据处理情况

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	Checking detail Confidential	删除
Unassigned	subCRA004036	TEST_20200915	2020-09-15	2020-09-30	Checking detail Confidential	删除
Unassigned	subCRA003988	GSA_TEST_20200910	2020-09-10	2022-11-08	⚠️ Unfinished at the OverView step Confidential	删除
Unassigned	subCRA002947		2020-04-06	2020-04-06	⚠️ Unfinished at the General Info step Confidential	删除
CRA001076	subCRA001069	GSA-test	2018-08-29	2025-02-24	Checked OK Confidential	立即发布 分享

一个审核成功的提交, 分配的GSA编号, 文章里使用此编号

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的导致，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 data release date

数据发布条件:

- 到达用户设定的发布日期
- 引用此数据的文章发表

* 发布日期

2018-05-24

(yyyy-mm-dd)

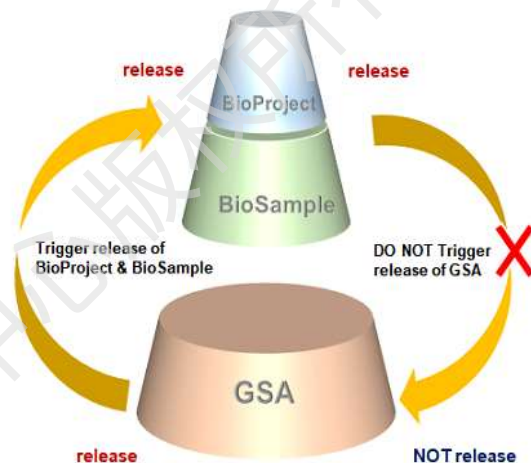
* 发布策略和免责声明

- 你提交的数据可以保留在一定时间内不发布;
- 发布时可以在GSA提交系统进行修改: ([http://bigd.big.ac.cn/gsub/submit/gsa/substitute your GSA accession number/confianta](http://bigd.big.ac.cn/gsub/submit/gsa/substitute%20your%20GSA%20accession%20number/confianta));
- 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来发布该数据;
- 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	subCRA000595	Human dataset test	2020-12-31	Checked OK Confidential	数据发布 删除



设置共享链接

Create GSA		Batch Submission		GSA Quick Start Guide (US) GSA Quick Start Guide (CN)		
Accession	Submission ID	Title	Release date	Status	Operation	
CRA000620	subCRA000620	GSA-test	2019-12-31	Checked OK Confidential	ReleaseNow Share	



Create GSA		Batch Submission		GSA Quick Start Guide (US) GSA Quick Start Guide (CN)		
Accession	Submission ID	Title	Release date	Status	Operation	
CRA000620	subCRA000620	GSA-test	2019-12-31	Checked OK Confidential	ReleaseNow Shared URL: http://bigd.big.ac.cn/gsa/s/VmVW41CA Cancel Share	

Review Link Opinion

Accession: CRA000620

Set review link expire time

1 month

勾选，生成数据
下载链接

☒ 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

Home Preview

CRA001076 基本信息

CRA 提交信息: CRA001076 / GSA-test / release time: 2025-02-24

数据下载

文件: <https://share.cnki.ac.cn/7HeHqPAJ>

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

GSA Genome Sequence Archive

Search: GSA CRA001962

Search result:

Search items: 40 / Items: 1-40 / Page: 1 / 3 / First Next Last / Jump To: 1 / Go

☐	Title: Spleen-2-5-IP Accession: CRX050012 Platform name: ILLUMINA Strategy name: RNASeq	Source name: TRANSCRIPTOMIC Selection name: PCR
☐	Title: Heart-1-5-IP Accession: CRX049994 Platform name: ILLUMINA Strategy name: RNASeq	Source name: TRANSCRIPTOMIC Selection name: PCR
☐	Title: Lung-2-5-Input Accession: CRX050013 Platform name: ILLUMINA Strategy name: RNASeq	Source name: TRANSCRIPTOMIC Selection name: PCR
☐	Title: Lung-1-4-IP Accession: CRX050011 Platform name: ILLUMINA Strategy name: RNASeq	Source name: TRANSCRIPTOMIC Selection name: PCR
☐	Title: Liver-1-5-Input Accession: CRX049999 Platform name: ILLUMINA Strategy name: RNASeq	Source name: TRANSCRIPTOMIC Selection name: PCR

Send to : ▼

Choose Destination

☒ File

Download 40 items.

Format

RunInfo

Create Files



Center

- INSDC (5715855)
- NGDC (88449)

Downloadable

- Non-downloadable (4282444)
- Downloadable (1521725)

File type

- ab1 (20)
- sff (299)
- sra (5715855)
- oxfordnanopore_native (5)
- pacbio_sequel_native (5)
- fastq (85969)
- fasta (334)
- bam (1814)

Platform

- unspecified (36681)
- ILLUMINA (5270548)

Search result:

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

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

Send to : ▼

Send to : ▼

Choose Destination

☒ File

Download 5804385 items.
Format

☒ Accession List

RunInfo

DataSet Accession List

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: 2020-09-08
文件个数: 27 文件大小: 69.51 GB

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) 推荐: [EdgeTurbo](#) (测试) [文件下载](#) Aspera命令行: [帮助](#)

FTP: <ftp://download.bjg.ac.cn/gsa/CRA003018>

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

Experiments/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		

点击下载元数据

多种文件下载方式



提纲

- 一 GSA Family介绍
- 二 GSA数据提交及共享
- 三 **GSA-Human数据提交及共享**
- 四 OMIX数据提交及共享

人类遗传资源相关法律法规

2019年7月

《中华人民共和国人类遗传资源管理条例》

人类遗传资源信息向外国组织……机构提供或者开放使用的，应当**向国务院科学技术行政部门备案并提交信息备份**

2021年4月

《中华人民共和国生物安全法》

加强对我国人类遗传资源和生物资源采集、保藏、利用、对外提供等活动的管理和监督，**保障人类遗传资源和生物资源安全**

2023年7月

《人类遗传资源管理条例实施细则》

加强人类遗传资源管理信息化建设，……推进实时动态管理，实现人类遗传资源管理信息可追溯、可查询

人类遗传资源信息管理的适用范围

2019年《中华人民共和国人类遗传资源管理条例》		2023年《人类遗传资源管理条例实施细则》	
异	业务名	信息备案	事先报告
	管理范围	人类遗传资源材料是指含有人体基因组、基因等遗传物质的器官、组织、细胞等遗传材料。人类遗传资源信息是指利用人类遗传资源材料产生的数据等信息资料	人类遗传资源信息包括基因、基因组、转录组、表观组及ctDNA等核酸类生物标志物等数据信息，以及与此数据相关的疾病、人种等关联信息， 其他不含人类遗传资源信息类型无需申报
	具体类别	临床数据、影像数据、生物标志物数据、基因数据、蛋白数据、代谢数据	全基因组测序相关、全基因组芯片相关、全外显子组测序相关、基因组靶向测序相关、表观组测序相关、转录组测序相关数据
同	申办人	中方单位	
	安全审查范围	1. 重要遗传家系的人类遗传资源信息 2. 特定地区的人类遗传资源信息 3. 人数大于500例的外显子组测序、基因组测序信息资源 4. 可能影响我国公众健康、国家和社会公共利益的其他情形	

注：尿液、粪便、血清、血浆不纳入材料管理，出境不需要申报；但是信息进行对外提供或开放共享时需要进行事先报告管理

承接国家人类遗传资源信息管理备份平台运维

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：

为简化国家人类遗传资源信息备份备案流程，确保备份备案与发表共享数据的一致性，推动人类遗传资源信息合法合规开放共享与再利用，经科技律研室，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工作。

委托的主要任务包括整合现有的人类遗传资源信息备份平台（<https://202.108.211.73/>）和国家生物信息中心人类遗传资源管理

— 1 —



- ❑ 优化人类遗传资源信息备份流程
- ❑ 确保备份、备案数据与发表共享数据一致性
- ❑ 推动人类遗传资源信息合法合规开放共享与再利用

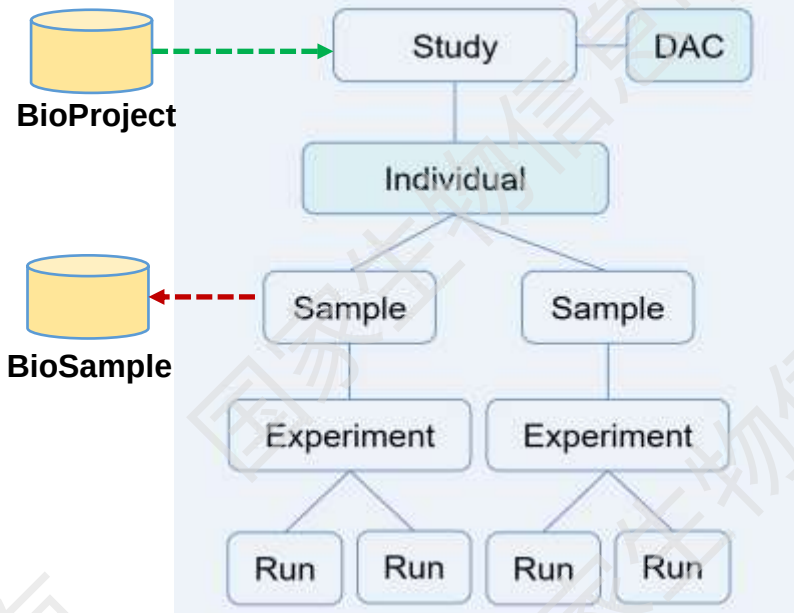
人类遗传资源数据的管理策略

- **全基因组测序**：包括全基因组原始测序数据、全基因组组装数据、全基因组变异数据、全基因组分析与统计数据；
- **全基因组芯片**：包括全基因组芯片原始数据、全基因组变异芯片数据、全基因组芯片相关统计或分析数据；
- **全外显子组测序**：包括全外显子组原始测序数据、全外显子组变异数据、全外显子组分析与统计数据；
- **基因组靶向测序**：包括基因组靶向原始测序数据、核酸类生物标志物、基因组靶向分析与统计数据；
- **表观组测序**：包括表观组原始测序数据、核酸类生物标志物、表观组分析与统计数据；
- **转录组测序**：包括转录组原始测序数据、核酸类生物标志物、转录组分析与统计数据。 《人类遗传资源管理条例实施细则》



数据模型和支持的数据类型

数据组织模型



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)	实体肿瘤研究 (Solid tumor)	Solid_Tumor.1.0_us.xlsx
	血液肿瘤研究 (Hematologic tumor)	Hematologic_Tumor_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.cnbc.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

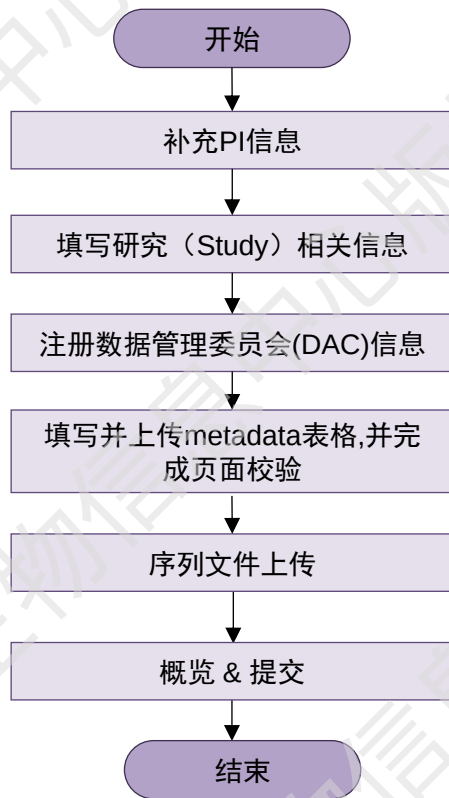
https://ngdc.cncb.ac.cn/gsa-human/document/Principle_of_Accessing_Human_Genetic_Resource_Data_in_NGDC_V1.pdf

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



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

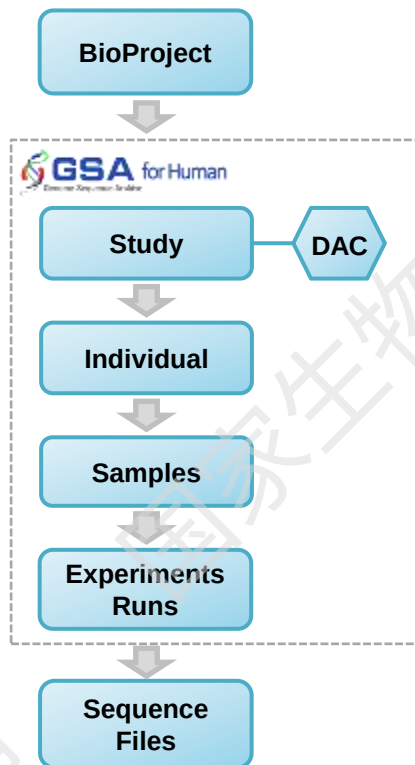


登录后，需要完善个人信息：

补充职称、个人主页/通讯文章链接、单位主页信息

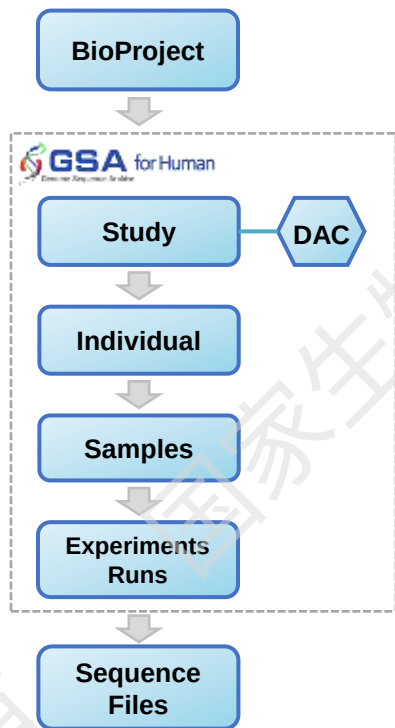
个人主页上的邮箱信息最好与账号邮箱一致
不符合条件的账号进行数据提交，将会审批
不通过

新建GSA-Human提交



The screenshot shows the GSA for Human website interface. The header includes navigation links like 'Data Resources', 'Computing Analysis', 'Data Network', and 'Standards'. The main content area is titled 'GSA-Human数据提交' (GSA-Human Data Submission). It provides instructions on how to submit data, including a list of eligible users (e.g., principal investigators, senior researchers, etc.) and a list of supported file formats (e.g., FASTQ, BAM, etc.). A red box highlights the 'New Submission' button. A text box with an arrow points to this button, containing the text: '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).

GSA-Human提交：数据提交者信息

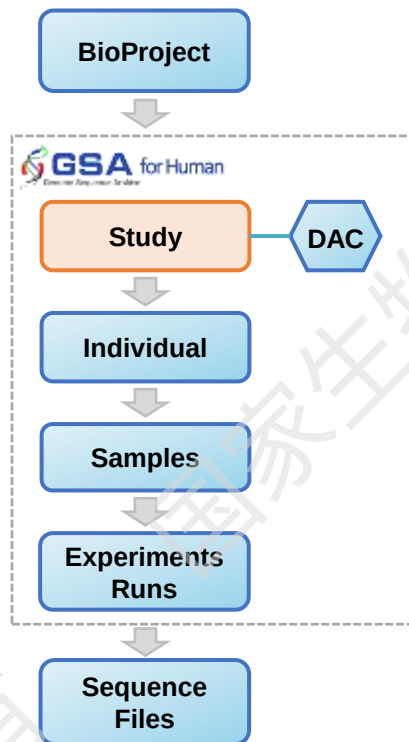


The screenshot shows the 'New Submission' form on the GSA-Human website. The form is divided into several sections: 'Subscriber Information', 'Mechanism Consent Information', and 'Privacy Policy'. The 'Subscriber Information' section includes fields for 'First Name', 'Middle Name', 'Last Name', 'Email', 'Institution/Organization', 'Department', 'Street', 'City', 'State/Province', 'Postal Code', and 'Country/Region'. The 'Mechanism Consent Information' section includes a checkbox for 'I want to provide my contact information' and a checkbox for 'I have read and accepted the terms above'. The 'Privacy Policy' section contains a detailed statement about the GSA-Human Database and its data management policies.

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-Human submission form. The form is divided into several sections, with three sections highlighted by red boxes and arrows pointing to explanatory text boxes on the right:

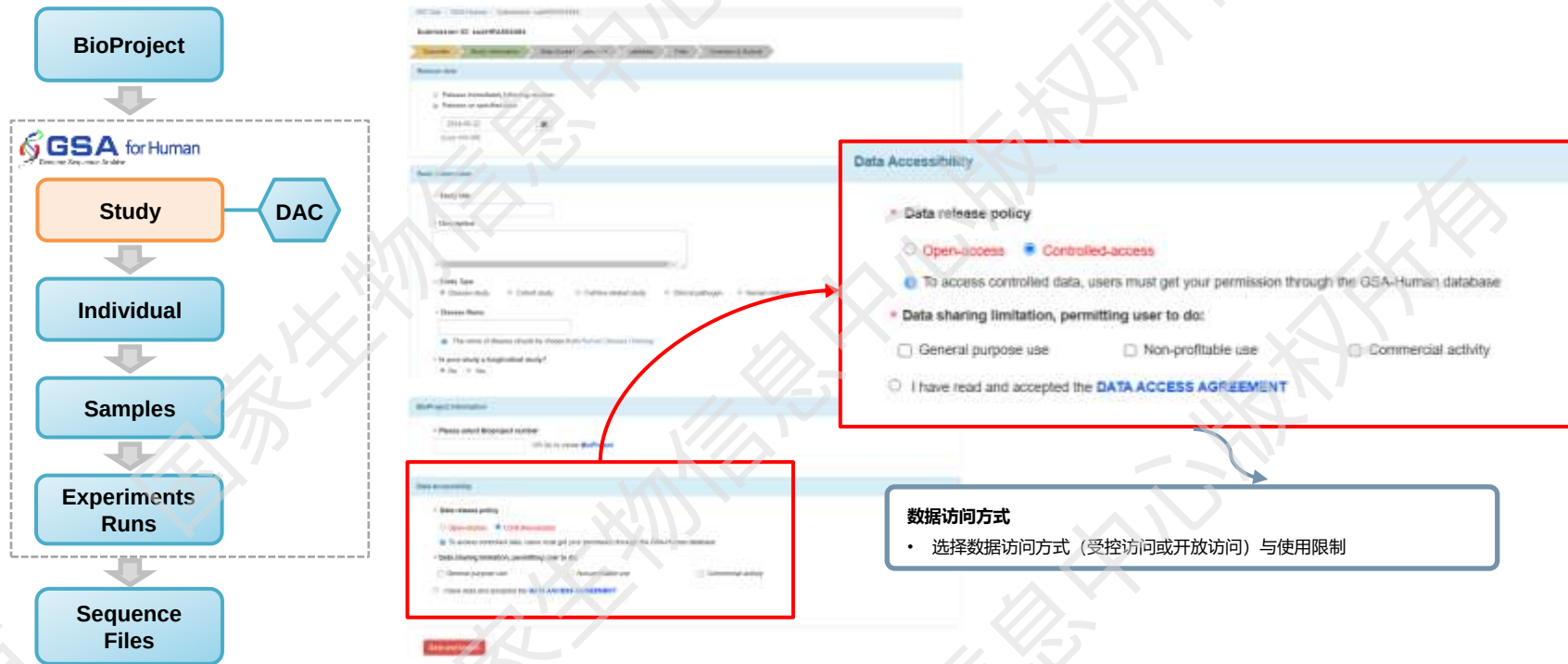
- Release date:** A section with a date picker set to 2019-05-02. An arrow points to the text: "发布时间：可选两年之内的任意时间作为发布时间；可以设置一次延期发布".
- Study information:** A section with a title field, a description field, and a study type dropdown menu. An arrow points to the text: "基本信息：按照研究情况，选择研究类型及其相关信息".
- Study related information:** A section with a field for the parent study ID. An arrow points to the text: "BioProject信息：填写已有BioProject编号，或者先跳转至BioProject新建".

发布时间：可选两年之内的任意时间作为发布时间；
可以设置一次延期发布

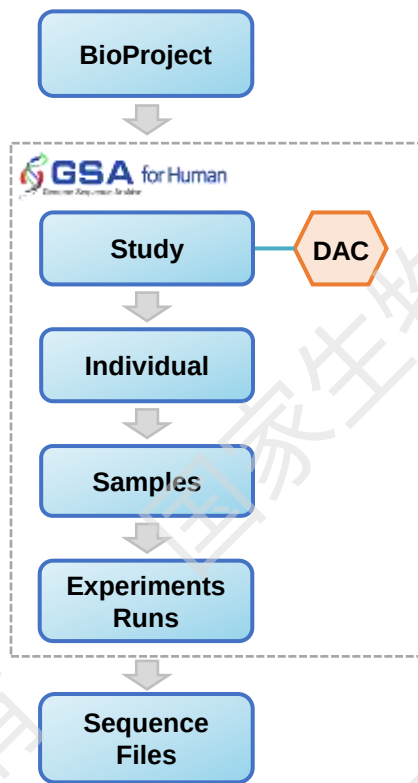
基本信息：按照研究情况，选择研究类型及其相关信息

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



SIS Site: GSA-Human | Submission: subHRA000497

Submission ID: subHRA000497

Submitter Study Introduction Data Access Committee Release Files Overview & Submit

Data Access Committee (DAC) is a body of one or more named individuals who are responsible for data release to external requesters based on consent and/or National Research Ethics laws.
1. A DAC is typically formed from the same or multiple organizations that collected the samples and generated any associated analyses.
2. Multiple datasets may be affiliated to a single DAC.

DAC Information

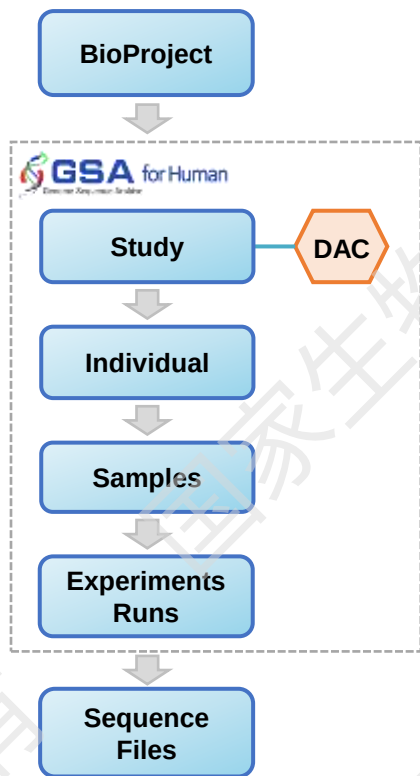
• Did you already register a DAC for this study?
☐ No ☐ Yes

Data Access Committee

DAC accession	DAC name	DAC Members	Email	Organization	Department	Is Contact Person	Contact Phone
HDAC000186				Beijing Institute of Genomics (BIG)	BIG Data Center	Yes	

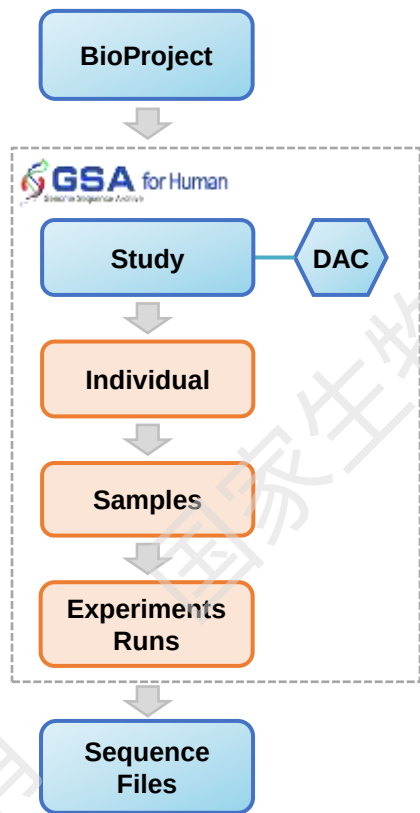
如果数据访问方式使用“受控访问 (Controlled-access)”，需建立数据管理委员会 (DAC) 并指定DAC联系人，负责数据发布后的申请审批与授权

GSA-Human提交：数据管理委员会（DAC）信息



The screenshot shows the 'DAC Information' section of the GSA-Human submission form. The form is titled 'GSA-Human : Submission subHRA003497' and 'Submission ID: subHRA003497'. The 'DAC Information' section includes a question: 'Did you already register a DAC for this study?' with radio buttons for 'No' and 'Yes'. The 'No' option is selected. Below this, there is a table for 'Data Access Committee' with columns 'DAC accession', 'DAC name', and 'DAC Members'. The table contains one entry: 'HDAC000186'. To the right of the table, there is a text box with the text '选择“No”，创建新DAC' (Select 'No', create new DAC). Below the table, there is a section for 'DAC Information' with fields for 'DAC name' (GSA_TEST_20200110) and 'Description' (GSA_TEST_20200110). Below this, there is a section for 'DAC Processors' with fields for 'Email', 'First Name', 'Last Name', 'Organization', 'Department', and 'Contact'. Below this, there is a section for 'Add DAC members' with a 'Contact Person (Selected)' section. The 'Contact Person (Selected)' section includes a note: 'Please follow the guidelines to complete the required information. ONLY GNC member can be selected as the contact person to receive data access requests. Please choose carefully and complete the contact phone number.' Below this, there are fields for 'First Name' (GSA), 'Email' (gus@ncbi.nlm.nih.gov), and 'Phone Number' (010-60000000). At the bottom, there is a 'Save and forward' button.

GSA-Human提交：元数据信息



The screenshot shows the GSA-Human submission interface. At the top, it displays 'Submission ID: subHRA000650'. Below this is a navigation bar with tabs: 'Submitter', 'Study Information', 'Data Accession Committee', 'Metadata', 'Files', and 'Overview & Submit'. The 'General Package' section is active, showing instructions to 'Upload the metadata file using Excel format'. A red box highlights the 'Download Excel Template General 1.0 spreadsheet' link. Below this, the 'Study Type' section is highlighted with a red box, showing options for 'Disease study', 'Cohort study', 'Cell line related study', 'Clinical pathogen', 'Human metagenome', 'Ancients or Fossils', and 'Other study'. The 'Disease Name' field is also visible, with a note that the name should be chosen from Human Disease Ontology.

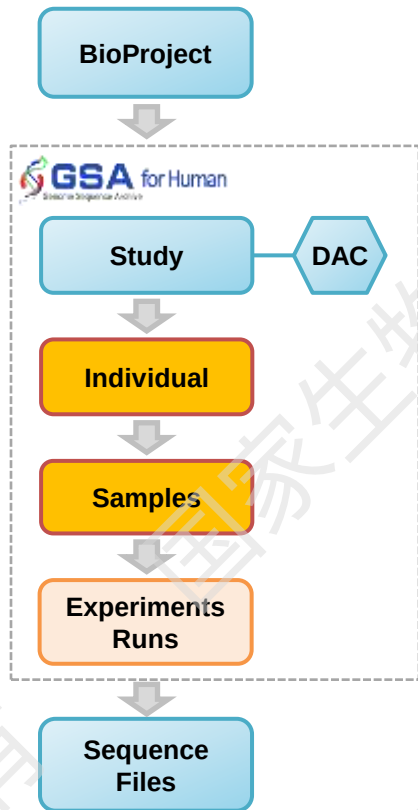
研究信息中的研究类型 (Study Type)

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



This sheet is used for individual information.
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.
GREY fields are optional. Leave optional fields empty if no information is available.
ORANGE fields require that the individual accession number, like HR000001, already exists in GSA for Human database.
YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.

CAUTION: DO NOT delete any columns! But you can insert the custom attributes at the end of all column as needed!
Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before sending to us.

1	2	3	4	5	6	7	8	9	10	11
ID	individual_name	gender	existing_individual_accession	experiment_group	disease_name	disease_stage	treatment	survival_time	is_smoker	is_drinker
D1										

12	13	14	15	16	17	18	19	20
height_cm	body_weight_kg	body_mass_index(BMI)	ethnicity	province	occupation	race	birthplace	custom_attributes_1

自定义字段

Individual 表格

1:N

常见问题1：对应关系

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。

This sheet is used for Sample information.
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.
GREY fields are optional. Leave optional fields empty if no information is available.
YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.

CAUTION: DO NOT delete any columns! But you can insert the custom attributes at the end of all column as needed!
Be aware that Excel may automatically apply formatting to your data.
In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before sending to us.

1	2	3	4	5	6	7	8	9	10	11
ID	sample_name	tissue	individual_ID	age	age_unit	public_description	sample_title	collection_date	biomaterial_provider	culture_collection
S1										

12	13	14	15
karyotype	phenotype	population	custom_attributes_1

自定义字段

Sample 表格

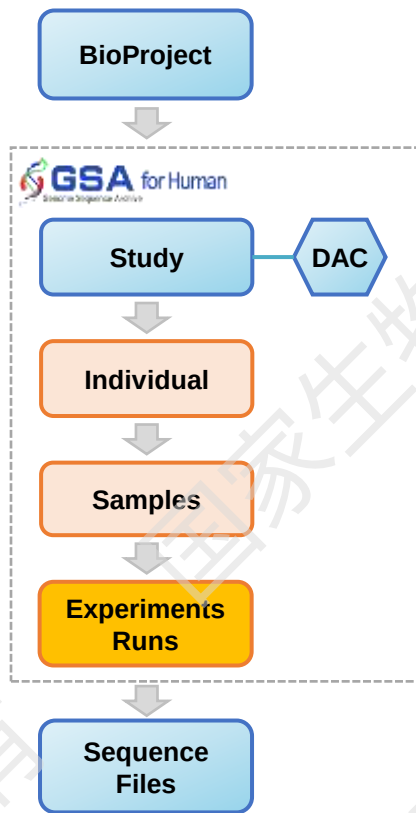
常见问题2：敏感信息问题

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：请确保信息的完整性与准确性。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**

This sheet is used for Experiment information.
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.
GREY fields are optional. Leave optional fields empty if no information is available.
BLUE fields require for paired-end data only.
YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.
When the Platform (column 4) is Helicos HelScope, the Planned number of cycles (column 16) is required.
CAUTION: DO NOT delete or insert column!
Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before sending to us.

1	2	3	4	5	6	7	8	9	10
ID	Experiment title	Sample ID	Platform	Library Construction / Experimental Design	Library name	Strategy	Source	Selection	Layout
E1									

11	12	13	14	15	16
Read length for mate 1(bp)	Read length for mate 2(bp)	Insert size (bp)	Nominal size (bp)	Nominal standard deviation (bp)	Planned number of cycles

Experiment 表格

1:N

常见问题1：对应关系

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。

This sheet is used for Run information and needs to be created after Experiment sheet.
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.
BLUE fields require for paired-end data only.
PURPLE When your Run data file type (column 4) select bam, if you want to submit your reference file to our FTP Site, you need to fill in the Reference file name and MD5 for reference file (columns 9-10). If your Run data file type is not bam, you do not need to fill in the Reference file name and MD5 for reference file.
YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.
CAUTION: DO NOT delete or insert column!
Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before sending to us.

1	2	3	4	5	6	7	8
ID	Run title	Experiment ID	Run data file type	File name 1	MD5 checksum 1	File name 2	MD5 checksum 2
R1							

9	10	11	12
Reference file name	MD5 for reference file	Assembly Name or Accession	Assembly Accession URL

Run 表格

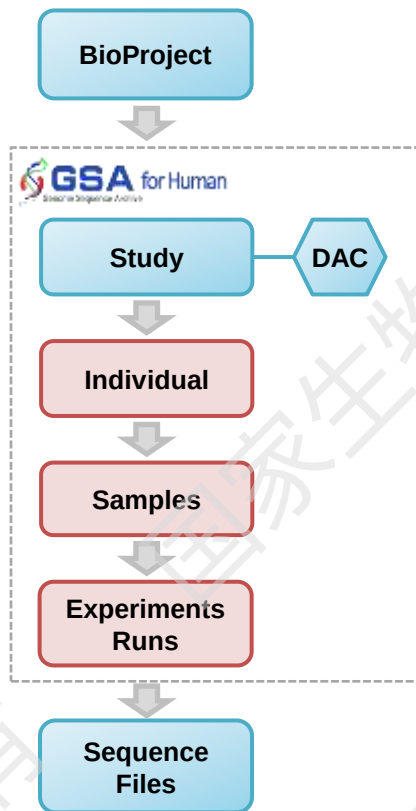
常见问题2：敏感信息问题

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

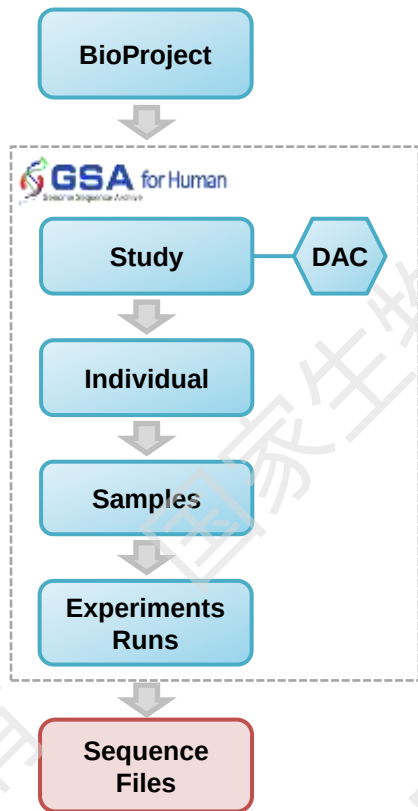
GSA-Human提交：元数据信息



The screenshot shows the GSA-Human submission interface. The submission ID is subHRA000550. The interface includes tabs for 'Submitter', 'Study information', 'Data Accession Codes', 'Metadata', 'Files', and 'Overview & Submit'. The 'Metadata' tab is selected, showing the 'General Package' section. The instructions state: 'Upload the metadata file using Excel format'. A red box highlights the 'Upload' button, with the text '上传填写完成的表格' (Upload the completed form). Below this, there is a link to 'Download Excel Template General 1.0 spreadsheet, edit, save and then upload the modified template.' and a note: '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'. A red arrow points down to the next step, which shows the 'Check' button highlighted with a red box and the text '点击进行“格式”校验' (Click to perform "format" check). Below this, another red arrow points down to the final step, which shows the 'Save and forward' button highlighted with a red box and the text '校验通过后，点击“Save and forward”进入下一步' (After passing the check, click "Save and forward" to enter the next step).

- 若文件审核未通过，请按系统提示修改信息后重新上传校验
- 如遇到任何问题，欢迎加入QQ群 (548170081) 交流，或通过电子邮件 gsa@big.ac.cn 联系我们。

GSA-Human提交：序列文件上传方式



The screenshot shows the GSA-Human submission interface. The submission ID is subHRA000487. The 'File' tab is selected. The 'File upload mode' section contains instructions: 'Each file must be added in the GSA metadata file you uploaded. Unique file names should be used for all files. Files must be compressed using gzip or bzip2. This is not the last step of the submission. Please click the 'Save and forward' button on the bottom of this page and step into the 'Overview' page, where you can preview and check the metadata you submitted and click the 'Submit' button to finish your submission.'

How do you want to provide files for submission?

☒ FTP

Use FTP only to upload files. FTP Username and Password are same as the GSA Sub account.

You may upload your files to our FTP site using an FTP client software (such as the [FileZilla Client](#)) using the FTP command:

Address: ftp.cnrcb.cn
Username: Same as you register the GSA Sub.
Password: Same as you login the GSA Sub.

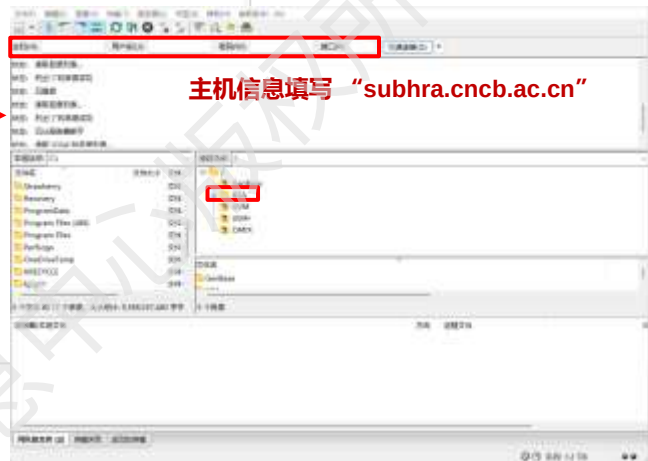
After you connected to the FTP server, please navigate to the GSA directory and upload your files here. DO NOT upload files to the root directory.

NOTE: Please use the library mode to transfer files. If you are using the FTP client software to log in to the FTP server, follow the software's instruction to set the transfer mode. If you are using the command line, please the library command before the input command.

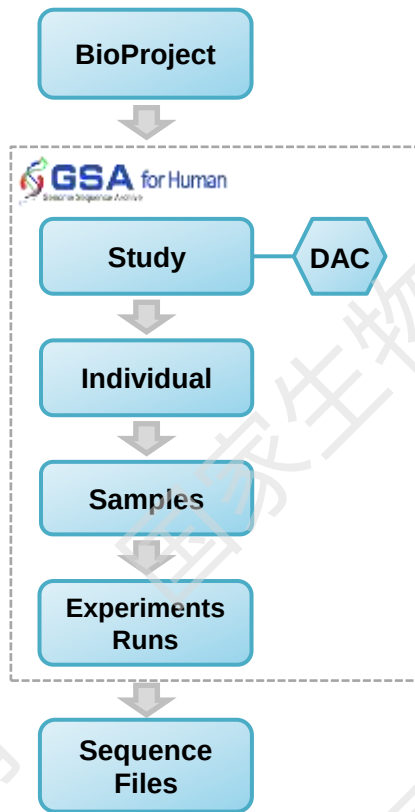
☐ Aspera Command Line

Use Aspera Command Line to upload files.

Save and forward



GSA-Human提交：概览 & 提交



The screenshot shows the GSA-Human submission form. The form is divided into several sections: 'Study Information', 'Sample Information', 'Data Accession Codes', 'Metadata Information', and 'Experiments & Files'. The 'Study Information' section includes fields for 'Study Title', 'Study ID', 'Study Type', 'Accession ID', and 'Accession Date'. The 'Sample Information' section includes fields for 'Sample Title', 'Sample ID', 'Sample Type', 'Accession ID', and 'Accession Date'. The 'Data Accession Codes' section includes a table with columns for 'Accession', 'Name', 'Email', 'Organization', 'Department', 'Accession ID', and 'Accession Date'. The 'Metadata Information' section includes a table with columns for 'Individual Name', 'Accession ID', 'Accession Date', 'Accession Type', 'Accession Status', and 'Accession Date'. The 'Experiments & Files' section includes a table with columns for 'Experiment Name', 'Platform', 'Sample Name', 'Accession ID', 'Accession Date', 'Accession Type', and 'Accession Date'. The 'Submit' button is highlighted in red.



请务必点击“提交”，提交后进入数据审核流程，请耐心等待反馈信息

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	aaaaaa	subHRA000350	2021-09-30	General	Checking detail Confidential	Controlled(0)	<button>Delete</button> <button>Update</button>
HRA000165	test	subHRA000201	2024-03-22	Cell Line	Checked OK Confidential	Open	<button>ReleaseNow</button> <button>Share</button> <button>Update</button>

- 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

分配的编号

Accession	Submission ID	Release Date	Package	Status	Access	Request	Operation
HRA000049	subHRA000206	2022-04-17	General	Checked OK Confidential	Controlled	0	Release Now Share

↓

Accession	Submission ID	Release Date	Package	Status	Access	Request	Operation
HRA000049	subHRA000206	2022-04-17	General	Checked OK Confidential	Controlled	0	Release Now Shared URL: http://gsa.big.ac.cn/gsa-human/s-ns000000 Cancel Share

Review User Options

Accession: HRA000049

Set share link expiry time:

☐ Enable data access link

NOTE: Researchers who access the sequence file through the link, please tick the checkbox to generate the link as required.

- 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://ngdc.cncb.ac.cn/gsa-human>.

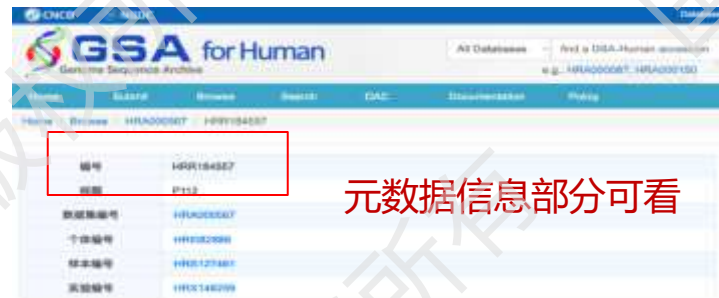
请按照以下格式引用我们的文章：

- **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]

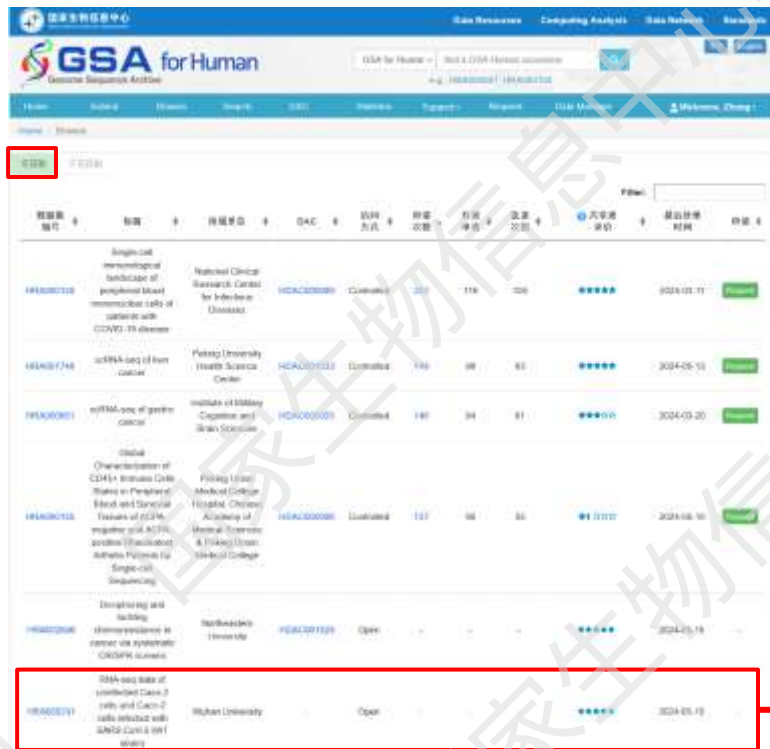
GSA-Human数据访问权限

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* - 已进行数据申请 - DAC审核通过	可获取	全部	全部	可下载
	- 符合开放共享条件* - 未进行数据申请或DAC审核未通过	可获取	部分	部分	不能
	不符合开放共享条件	不可获取	部分	部分	不能

*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



GSA-Human数据浏览与获取



数据集编号	数据集名称	项目单位	数据类型	测序方法	测序深度	测序平台	测序日期	数据集状态	数据集描述
HSR000001	Genomic and transcriptomic landscape of peripheral blood mononuclear cells of patients with COVID-19 disease	National Clinical Research Center for Infectious Diseases	Genomic	RNA-seq	118	118	2020-01-11	Available	
HSR000002	scRNA-seq of liver cancer	Peking University Health Science Center	Genomic	RNA-seq	118	118	2020-06-18	Available	
HSR000003	scRNA-seq of gastric cancer	Institute of Cell Biology, Capital Medical University	Genomic	RNA-seq	118	118	2020-03-20	Available	
HSR000004	Genomic and transcriptomic landscape of CD4+ T helper cells from peripheral blood and spleen of patients with COVID-19 disease	Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College	Genomic	RNA-seq	118	118	2020-01-11	Available	
HSR000005	Transcriptomic and genomic landscape of human liver cancer via single-cell RNA-seq	Harbin Medical University	Genomic	RNA-seq	118	118	2020-01-11	Available	
HSR000006	scRNA-seq data of peripheral blood mononuclear cells and liver cancer cells infected with SARS-CoV-2	Wuhan University	Genomic	RNA-seq	118	118	2020-01-11	Available	

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



数据集编号: HSR000001
主键: <https://ngdc.cnbc.ac.cn/gsa-human/HSR000001>
访问方式: [Open](#)

数据集描述: HSR000001: Genomic and transcriptomic landscape of peripheral blood mononuclear cells of patients with COVID-19 disease

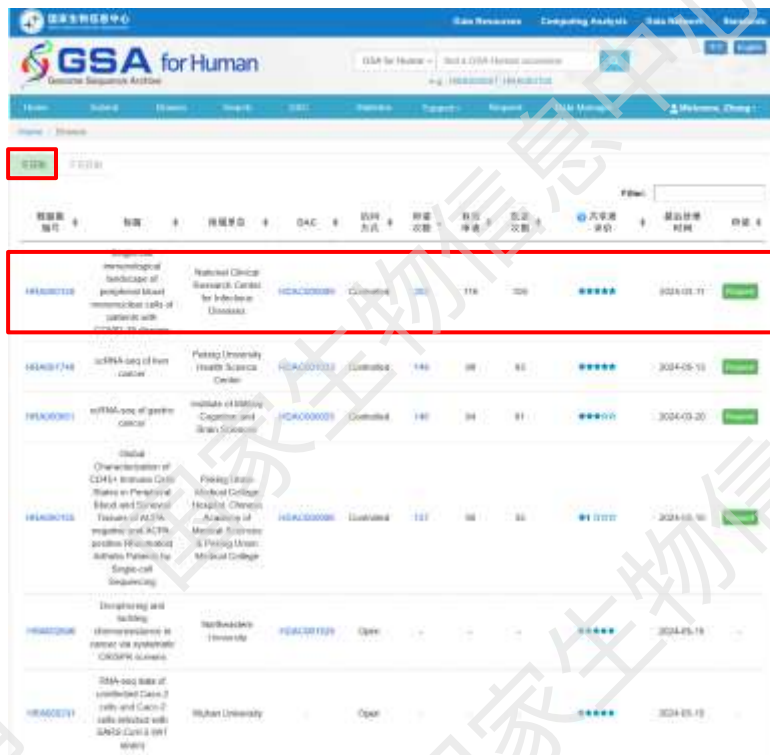
数据集状态: Available

数据集描述: HSR000001: Genomic and transcriptomic landscape of peripheral blood mononuclear cells of patients with COVID-19 disease

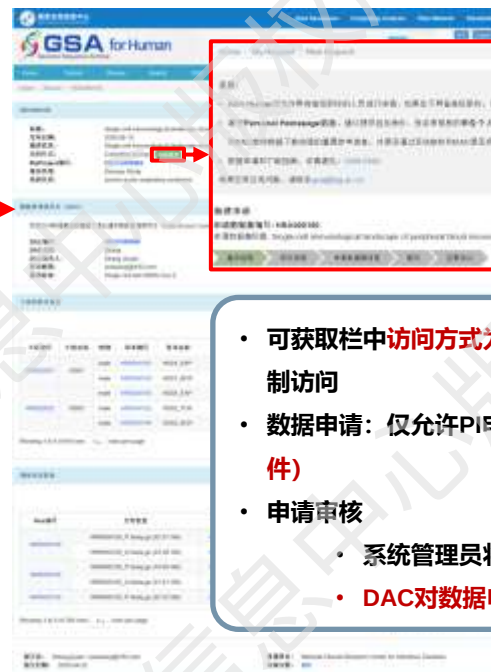
数据集描述: HSR000001: Genomic and transcriptomic landscape of peripheral blood mononuclear cells of patients with COVID-19 disease

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



数据集编号	标题	项目编号	DAC	访问方式	数据格式	数据量	数据大小	数据日期	数据状态
HS000001	Immunological landscape of peripheral blood mononuclear cells of patients with COVID-19	National Clinical Research Center for Infectious Diseases	HS000001	Controlled	200	118	100	2024-01-11	Available
HS000002	scRNA-seq of liver cancer	Peking University Health Science Center	HS000002	Controlled	140	88	80	2024-06-18	Available
HS000003	scRNA-seq of gastric cancer	Institute of Basic Medicine, Capital and Brain Sciences	HS000003	Controlled	140	88	81	2024-03-20	Available
HS000004	Single-cell characterization of CD4+ Treg cells in peripheral blood and synovial tissues of RA patients and ACRA patients (Heterologous autologous plasma) for single-cell sequencing	Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College	HS000004	Controlled	100	88	80	2024-05-10	Available
HS000005	Immunizing and blocking chemoprevention in cancer via synthetic CRISPR screens	Harvard Medical School	HS000005	Open	-	-	-	2024-05-18	Available
HS000006	RNA-seq data of uninfected and infected cells and cells infected with SARS-CoV-2	Wuhan University	HS000006	Open	-	-	-	2024-05-18	Available



数据集编号	标题	项目编号	DAC	访问方式	数据格式	数据量	数据大小	数据日期	数据状态
HS000001	Immunological landscape of peripheral blood mononuclear cells of patients with COVID-19	National Clinical Research Center for Infectious Diseases	HS000001	Controlled	200	118	100	2024-01-11	Available

- 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.cnbc.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	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	

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

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

收集数据申请者利用该数据集拟开展的研究的相关信息

受控数据申请

New Request

Request Study: HRA000067

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

确认申请的数据集的信息和数据提交者的要求

Basic Information Research Information Study Information Overview Agreement

Requested study

* Study Accession

HRA000067

Study Title

Multi-omics Characterization of Gastric Cancer Evolution unde

Uses and Limitations

General purpose use

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

NO

* Data Access Agreement



请仔细阅读数据访问协议 (Data Access Agreement)

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

提供对本次数据申请信息的整体概览

Basic Information Research Information Study Information Overview Agreement

Overview of the request

Basic Information

Request title
Test study

PI information

Email	chenlt@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**

Agreement

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

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 ☒ Online signature

☐ Offline signature ☐ Upload a signature picture (recommend size: 120px * 50px)

☐ Generate signature by using mouse

☐ Create a signature automatically

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

确认协议内容

上传电子签名图片
使用鼠标签名
自动生成签名

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	Cancel Modify Review History

↓

My Request List

[New Request](#)

Search:

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	Review History

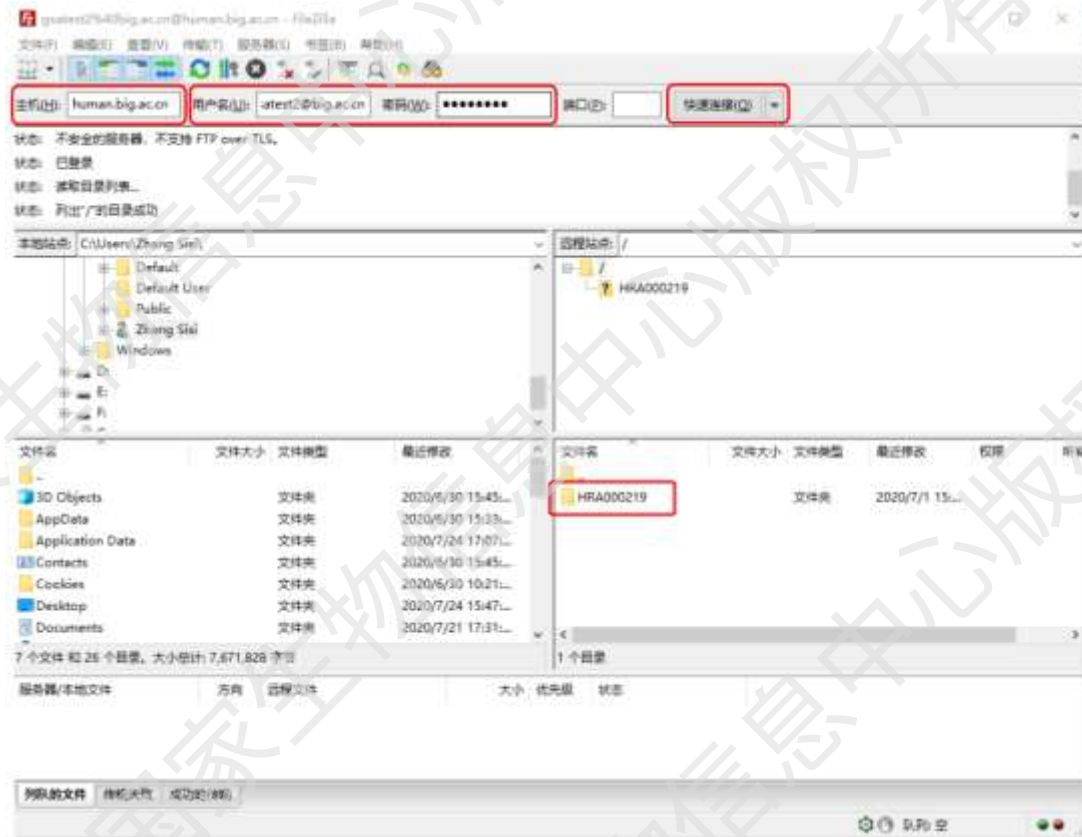
Showing 1 to 2 of 2 entries

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



提纲

- 一 GSA Family介绍
- 二 GSA数据提交及共享
- 三 GSA-Human数据提交及共享
- 四 **OMIX数据提交及共享**

数据类型范围

一级菜单		二级菜单	
基因数据	Genomics Data	Gene Expression Profiling by NGS	基因表达图谱
		Non-coding RNA Profiling by NGS	非编码RNA图谱
		Genome Binding/Occupancy Profiling by NGS	基因组结合/占用分析图谱
		Chromatin Accessibility Profiling by NGS	染色质可及性分析图谱
		Methylation Profiling by NGS	甲基化图谱
芯片数据	Microarray Data	Other Type of Genomic Data	其它类型基因数据
		Gene Expression Profiling by Array	基因表达芯片
		Non-coding RNA Profiling by Array	非编码RNA芯片
		Genome binding/occupancy profiling by Array	基因组结合/占用芯片
		Methylation profiling by Array	基因组甲基化芯片
蛋白组数据	Proteomic Data	Proteomic Data by Mass Spectrometry (MS)	蛋白质谱数据
		Protein 3D Structure Data	蛋白三维结构数据
		Other Type of Proteomic Data	其他类型蛋白数据
代谢组数据	Metabolome Data	Metabolome Data by Mass Spectrometry (MS)	代谢组质谱数据
		Lipidome Data by Mass Spectrometry (MS)	脂质组质谱数据
		Other Type of Metabolome Data	其它类型代谢数据
临床数据	Clinical information	Demographic Data	人口学资料
		Clinical Research data	临床研究性数据
		Other Type of Clinical information	其它类型临床数据
影像数据	Imaging data	Magnetic Resonance Imaging (MRI)	磁共振成像数据
		Other Type of Image Data	其它类型图像数据

推荐的数据格式

类型	优先推荐格式	不推荐格式
压缩文件	gz, zip, tar, 单个压缩文件≥50 GB, 请务必使用zip压缩	不建议提交rar
NGS分析后数据	txt, tsv, csv, mtx, bed	bedgraph, bigwig等二进制文件
芯片数据	原始数据: cel, idat, txt, csv, tsv, chp等	
代谢质谱数据	原始下机数据: D,.BAF,.FID,.YEP,.WIFF等 分析数据: mzML, mzXML, xml, txt等	
蛋白质谱数据	原始数据: D,.BAF,.FID,.YEP,.WIFF等 峰值数据: mzML, mzXML, mgf等 结果数据: mzIdentML, mzTab; 搜库结果文件: xml, pepxml, protxml, txt	
图像数据	JPEG (jpg, jpeg); TIFF (.tif, .tiff); MRC	
影像数据	MP4, avi	
电子表格/统计数据	xls, csv, txt	

文件上传：在线上传



The screenshot shows a multi-step form for uploading data. The steps and their corresponding annotations are:

- Step 1: Data Info** - Annotation: 数据集基本信息 (Dataset Basic Information)
- Step 2: BioProject Info** - Annotation: 关联BioProject (Associate BioProject)
- Step 3: Release Date** - Annotation: 发布时间 (Release Time)
- Step 4: Files** - Annotation: 点击添加文件 (Click to add file)

At the bottom right of the form, there is a button labeled: 支持添加多个 (Support adding multiple).

- **在线提交：**数据总量 < 100 MB
- **数据集发布时间：**
 - a. 可选五年之内的任意时间作为发布时间
 - b. 只要数据未公开，可随时修改该时间
 - c. 如果系统检测到包含OMIX数据集编号的文章已发表，会自动触发释放数据集
- **说明文档：**[OMIX使用说明](#)

文件上传：FTP上传



数据浏览与获取

OMIX

Home

Analysis

Genotype

Variant

View all released data of OMIX

Search: OMIX

Released Data: 486 records

Table

12

Columns

OMIX ID	Project	Title	Organization	Access Type	Release Time	Operation
OMIX026	PRJCA000308	Polystyrene spectrum in the peritoneal lavage fluid	Shanghai Institute of Materia Medica	Open	2008-12-23	 
OMIX020	PRJCA000408	Digoxin polymers	China Biotech	Open	2017-05-06	 
OMIX021	PRJCA000308	Protein content of digoxin after normalisation	China Biotech	Open	2017-05-06	 
OMIX022	PRJCA000308	HPLC	Protein Science	Controlled	2012-04-01	 
OMIX024	PRJCA000308	Proteinase K (PK) assay	Protein Science	Controlled	2012-04-01	 
OMIX025	PRJCA000308	High-resolution protein mass spectrometry of pancreatic islet cells	Protein Science	Open	2017-05-07	 
OMIX021	PRJCA000308	A test of normal protein reference sequence	Protein Science	Open	2017-05-07	 
OMIX026	PRJCA000308	Maximum likelihood protein structure prediction of pancreatic islet cells	Multi-omics	Open	2017-05-07	 
OMIX020	PRJCA000308	High-resolution protein mass spectrometry of pancreatic islet cells	Multi-omics	Open	2017-05-07	 
OMIX021	PRJCA000308	Proteinase K (PK) assay	Protein Science	Controlled	2012-04-01	 

Showing 486 of 486 (1 of 0) records

Previous

1

2

3

4

5

6

7

8

9

10

Next

ID#	Project	Site	Contract	Access Type	Release Date	Download
1064118	FLC0000000	Protein crystal structure of the apoferritin domain (PDB ID: 1A8N)	Researcher access	Open	2007-01-01	 
1064118	FLC0000000	Long range structure	Researcher access	Open	2011-01-01	  FTP Download

公开访问数据集，可自由选择使用FTP或HTTP方式获取数据

OMIX

Manuscript Information System

Applicant Information

Name: [Redacted]

Email: [Redacted]

Address: [Redacted]

Research Purpose, Methods and Techniques

Title: [Redacted]

Abstract: [Redacted]

Keywords: [Redacted]

申请者基本信息

研究目的，方法和技术等

受控访问数据，采用申请
审核制，审核通过后，方
可获取所需数据

研究目的, 方法和技术等

用户服务



Email: gsa@big.ac.cn

QQ Group: 548170081 (GSA)

516418530 (OMIX)



THANKS