“免疫力数字解码”重大研究计划

国家生物信息中心 CNCB

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

跨尺度免疫力大数据资源平台（IDP）

人类遗传原始组学数据

汇交流程介绍

CONTENTS

1 数据汇交流程

- 人类遗传原始测序数据 GSA-Human

2 汇交注意事项

3 汇交管理可视化

4 资源材料

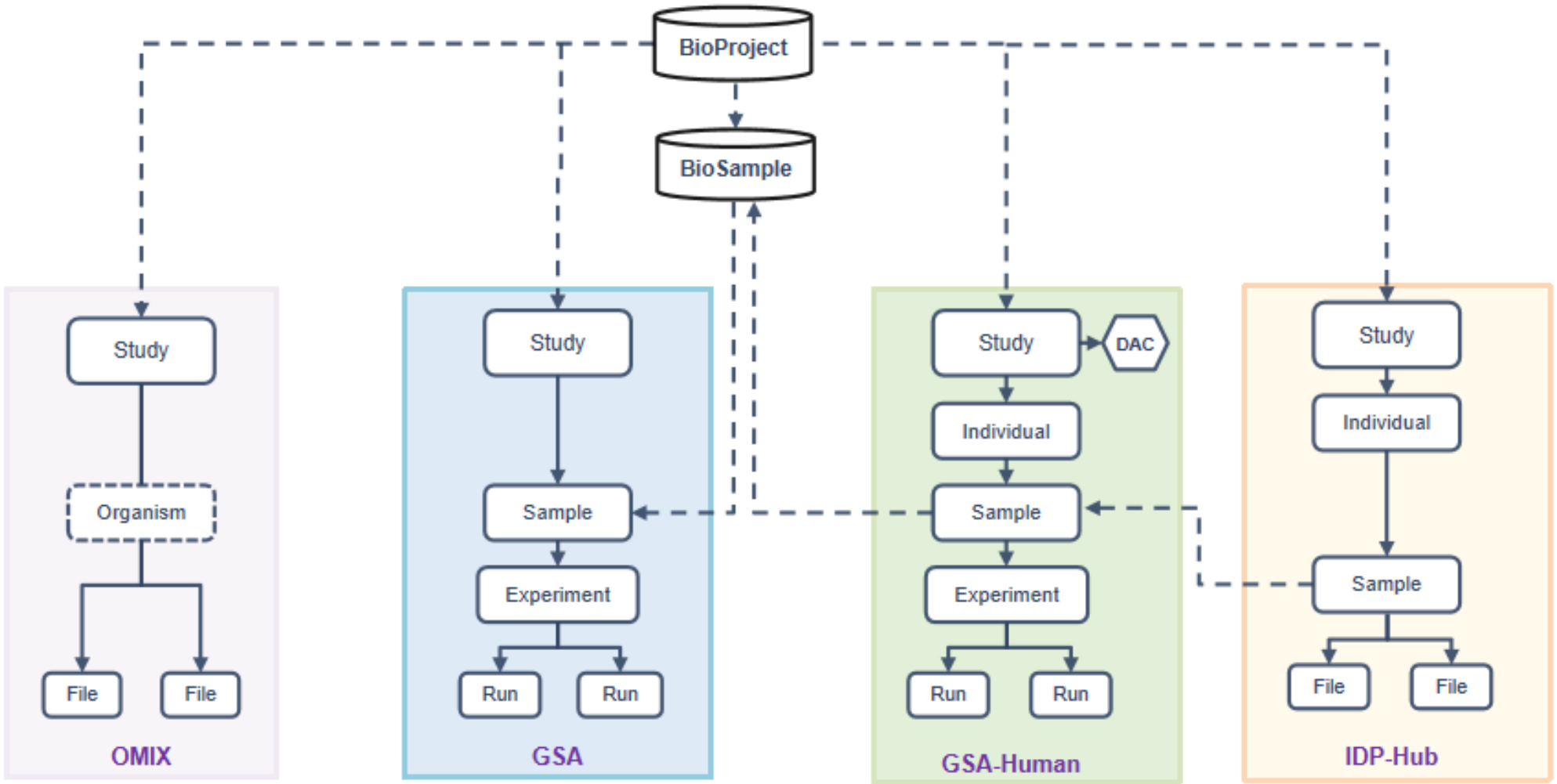
汇交数据入口

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

汇交数据入口

不同模态的数据分别点击对应图片即可跳转到递交页面

多模态数据汇交架构图



拓展型数据

非人遗数据

人遗数据

临床数据

（如果涉及多模态数据，在提交时编号一定要用生成的）

CONTENTS

1 数据汇交流程

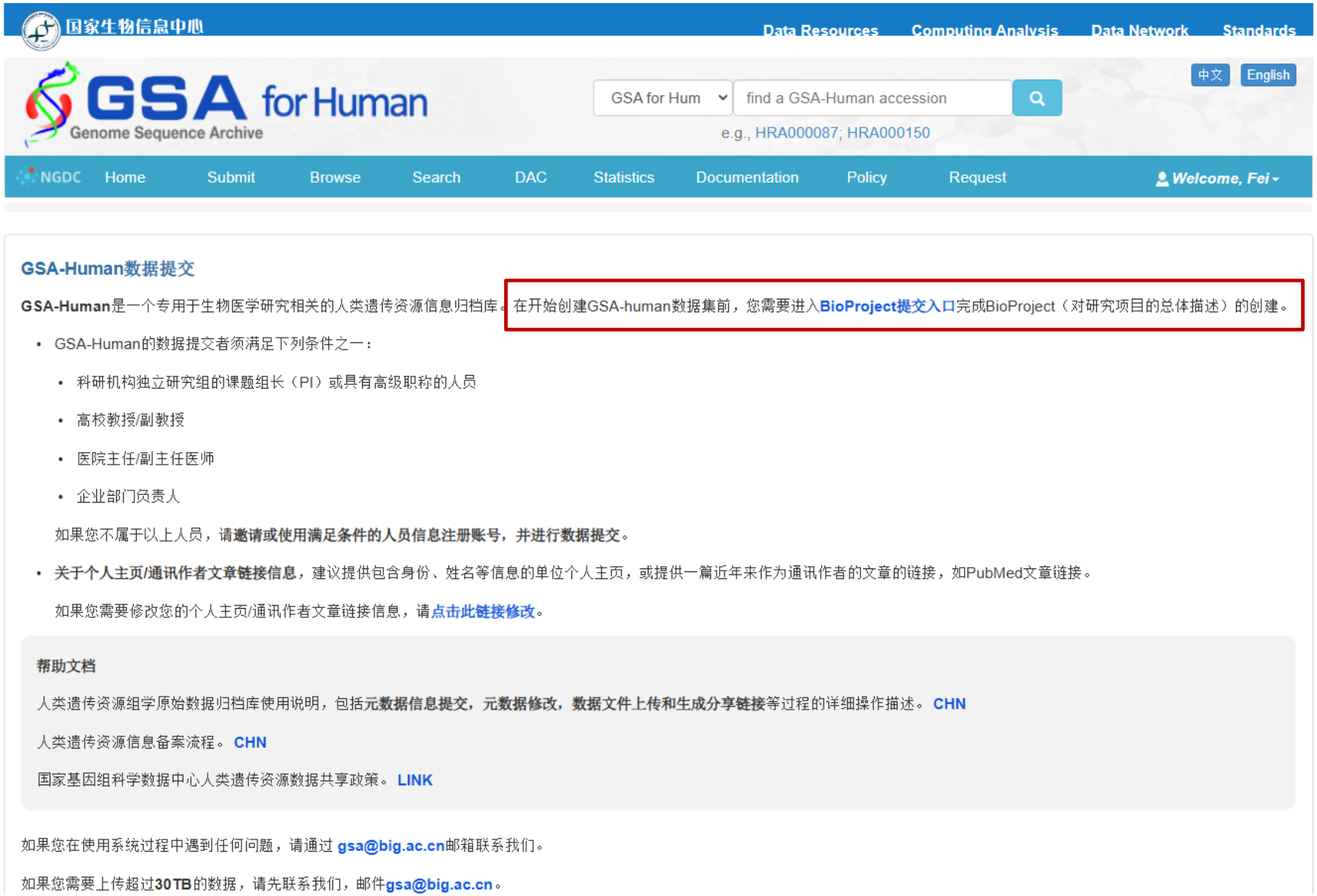
- 人类遗传原始测序数据 GSA-Human

2 汇交注意事项

3 汇交管理可视化

4 资源材料

进入GSA-Human提交介绍页面



https://ngdc.cncb.ac.cn/gsa-human/submit/hra/submit

BioProject：创建提交

BIG Sub
BIG Submisson Portal

中文 English

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

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
PRJCA003418	subPRO004980	GSA_TEST_20200909	2021-09-09	finished	confidential	修改 删除

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

BioProject：基本信息

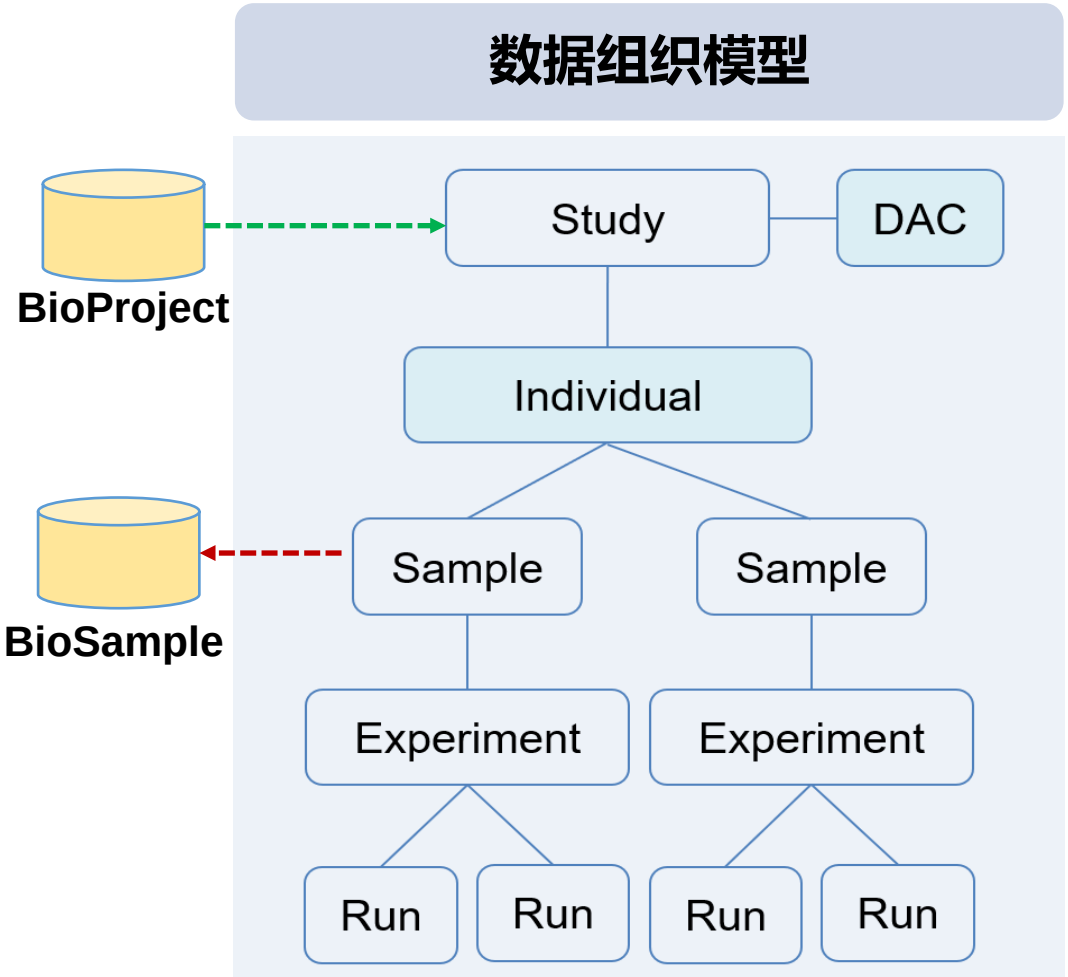
★此两处是**汇交进度关联**的重要依据，请务必填写正确！

BioProject：出版信息

- 文章发表时，请按照正确的引用方式引用数据集编号
- 文章发表后，请返回填写PMID或DOI在该部分，方便关联成果产出。同时请把发表文章的全部信息—作者、题目、期刊、刊号、页数、日期信息发送到邮箱gsa@cncb.ac.cn

GSA-Human提交：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)	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

数据需符合指南中数据质控标准

1 bulk TCR-seq/BCR-seq

- 外周血单样本clonotypes平均值不低于100,000种，组织样本clonotypes平均值不低于10,000种，重链和轻链各不少于1,000,000读数，单样本测序量不低于2Gbp （其中有效clonotype的定义为：能够组装出全长氨基酸序列且不包含终止密码子，具有明确的胚系V和J基因信息，同时CDR3区域含有典型的保守序列)

2 bulk RNA-Seq

- DV200不低于50%，检测覆盖不低于50M reads

3 WES

- 单碱基覆盖率不低于100X，检测盒覆盖不低于32MB

数据需符合指南中数据质控标准

4 scRNA-seq

- 测序指标（有效barcodes不低于75%，有效UMIs不低于75%，RNA read Q30不低于75%）；细胞指标（单个样本细胞数不少于5000个，平均每个细胞的测序read pairs数大于20000，单个细胞基因中位数大于1000）

5 scTCR-seq/scBCR-seq

- 平均每个细胞所测得的VDJ文库read pairs数不低于5000，RNA read Q30 不低于75%，有效barcodes不低于75%；细胞指标（3000-80000个细胞/芯片）

6 scATAC-seq

- 测序指标（测序read pairs不小于25000，有效barcodes不低于75%，barcode Q30大于65%，read1/2 Q30大于65%，样本指数的Q30大于90%）；细胞指标（单个样本细胞数不少于5000个，高质量测序片段大于40%）

建议各项目完成数据质控后，符合的上述标准的数据再进行汇交

原始组学数据递交-数据准备

口原始序列文件整理，并计算文件的MD5码

- 格式正确的序列文件（如: gz压缩格式的fastq文件， xxx.fastq.gz）
- MD5码计算方式：
  - 在Linux 系统下，可运行md5sum 命令计算
  - 在MacOs系统下，可运行md5命令计算
  - 在Windows系统下，可借助第三方工具计算，如： WinMD5Free (http://www.WinMD5.com)

口元数据整理

- 数据的描述性信息
- Sample / Experiment / Run

原始组学数据递交-数据汇交环境

浏览器建议

- Firefox
- Google Chrome
- IE（V11以上）
- Safari

FTP客户端软件

- FileZillla （https://filezilla-project.org/）
- Aspera Connect客户端软件(https://www.ibm.com/aspera/connect/)

GSA-Human提交：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

GSA-Human提交：数据提交流程

GSA-Human提交：个人信息补充

①

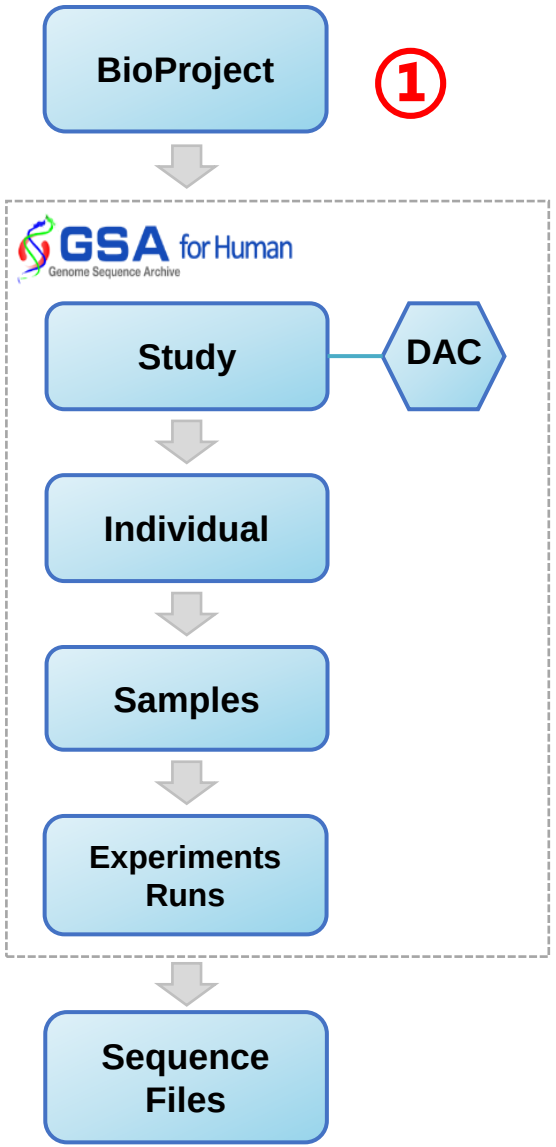
登录后，需要完善个人信息：

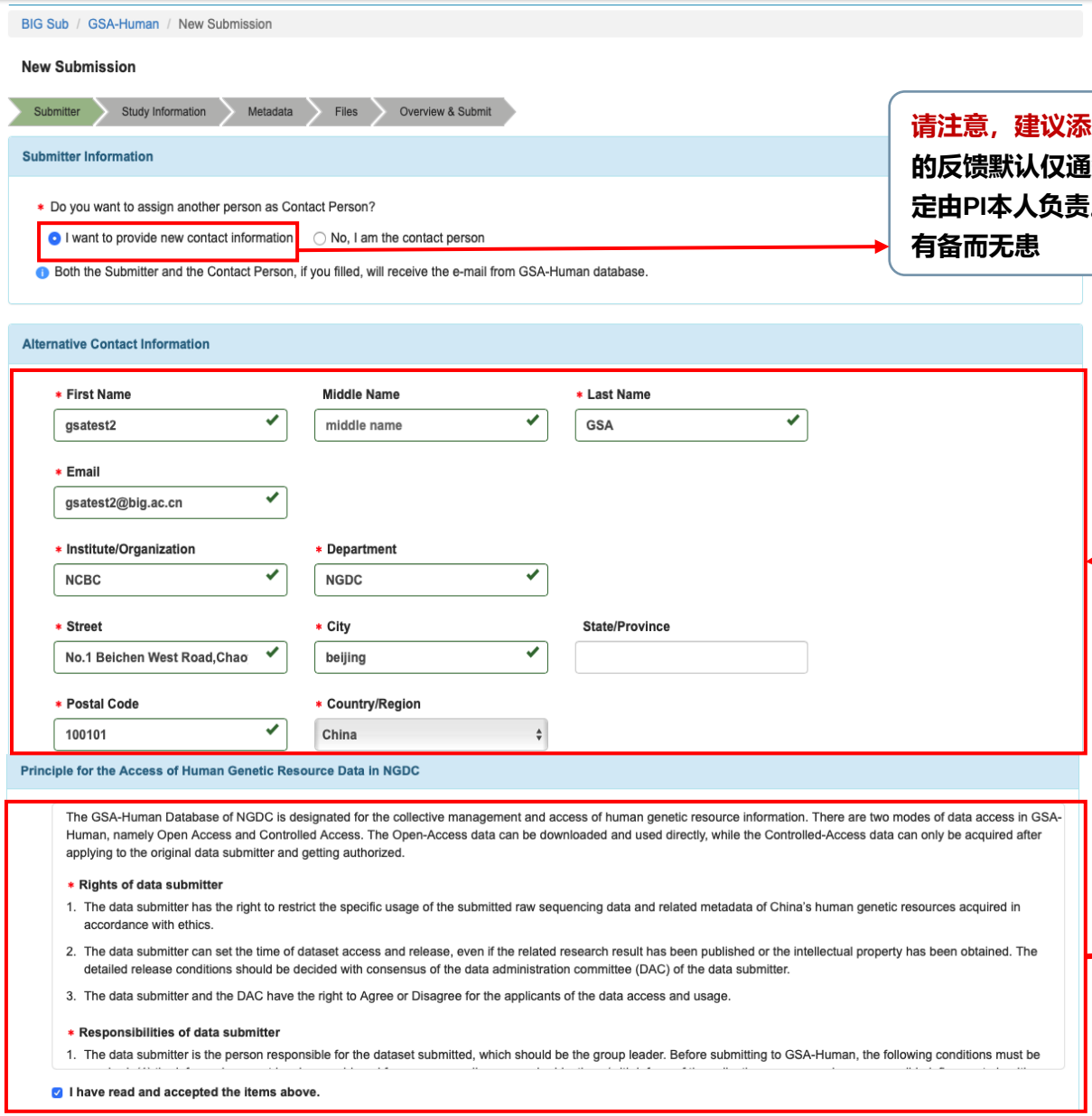
补充职称、个人主页/通讯文章链接、单位主页信息

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

GSA-Human提交：数据提交者信息

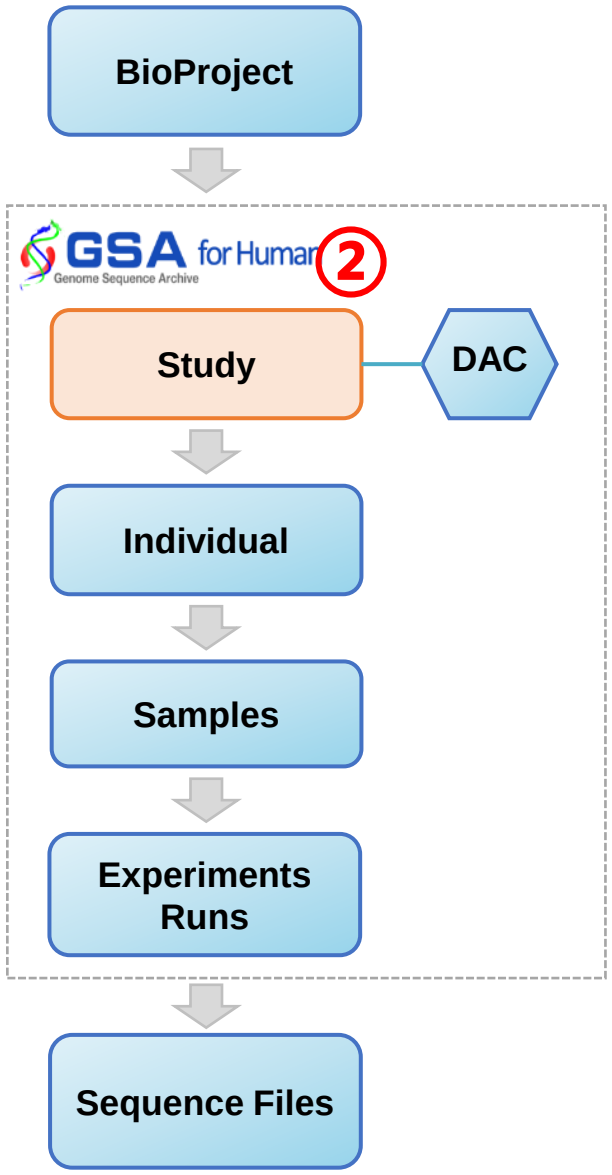


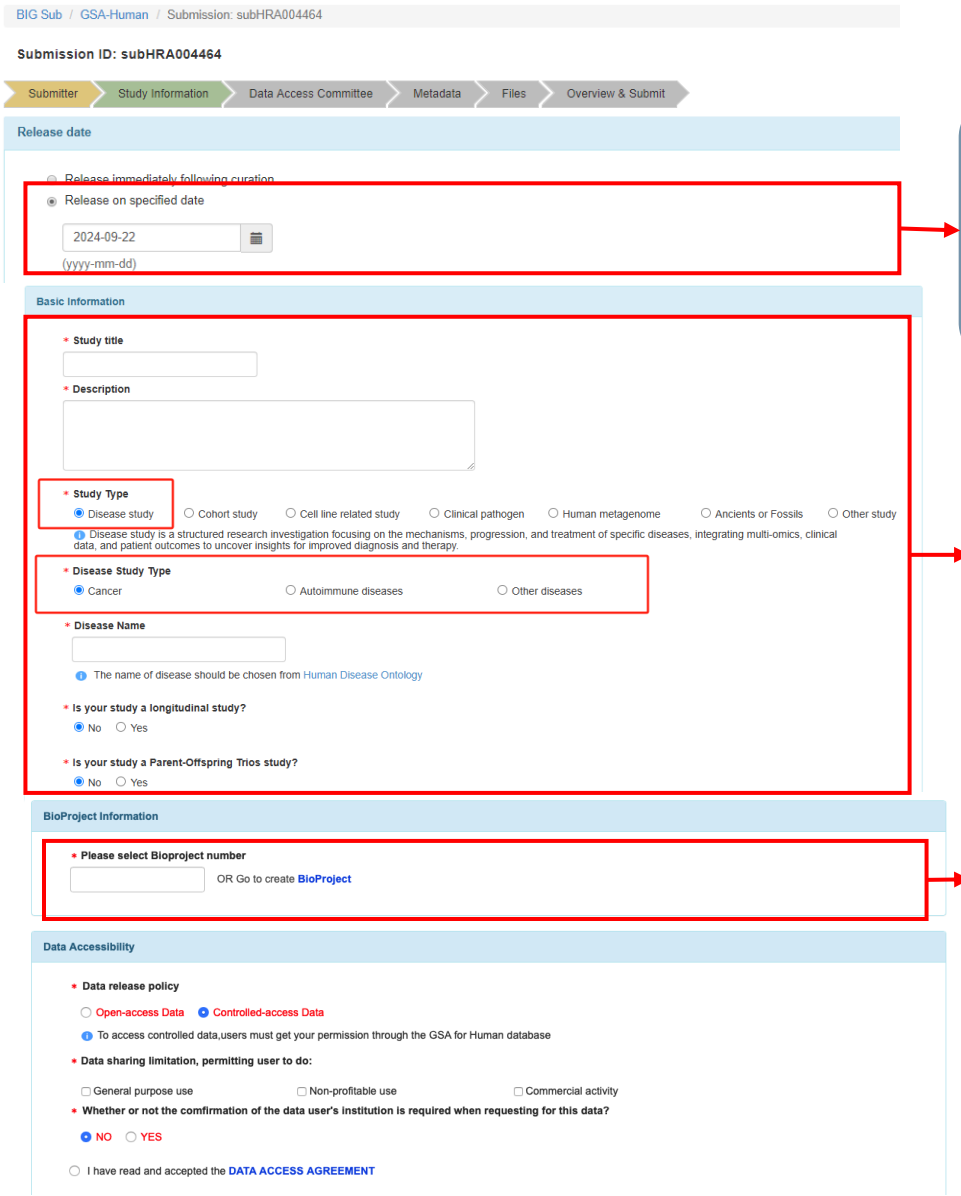


请注意，建议添加提交联系人信息。如果不添加，数据集问题的反馈默认仅通知注册账号，而往往数据提交和问题处理不一定由PI本人负责。添加后两个人都会收到通知信息，多一个人有备而无患

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





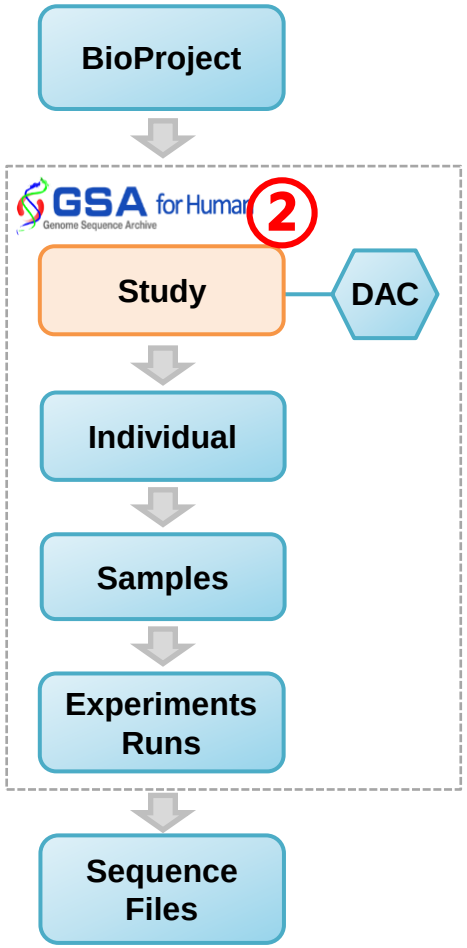
发布时间：可选**两年之内**的任意时间作为发布时间；只要数据未公开，可随时修改该时间

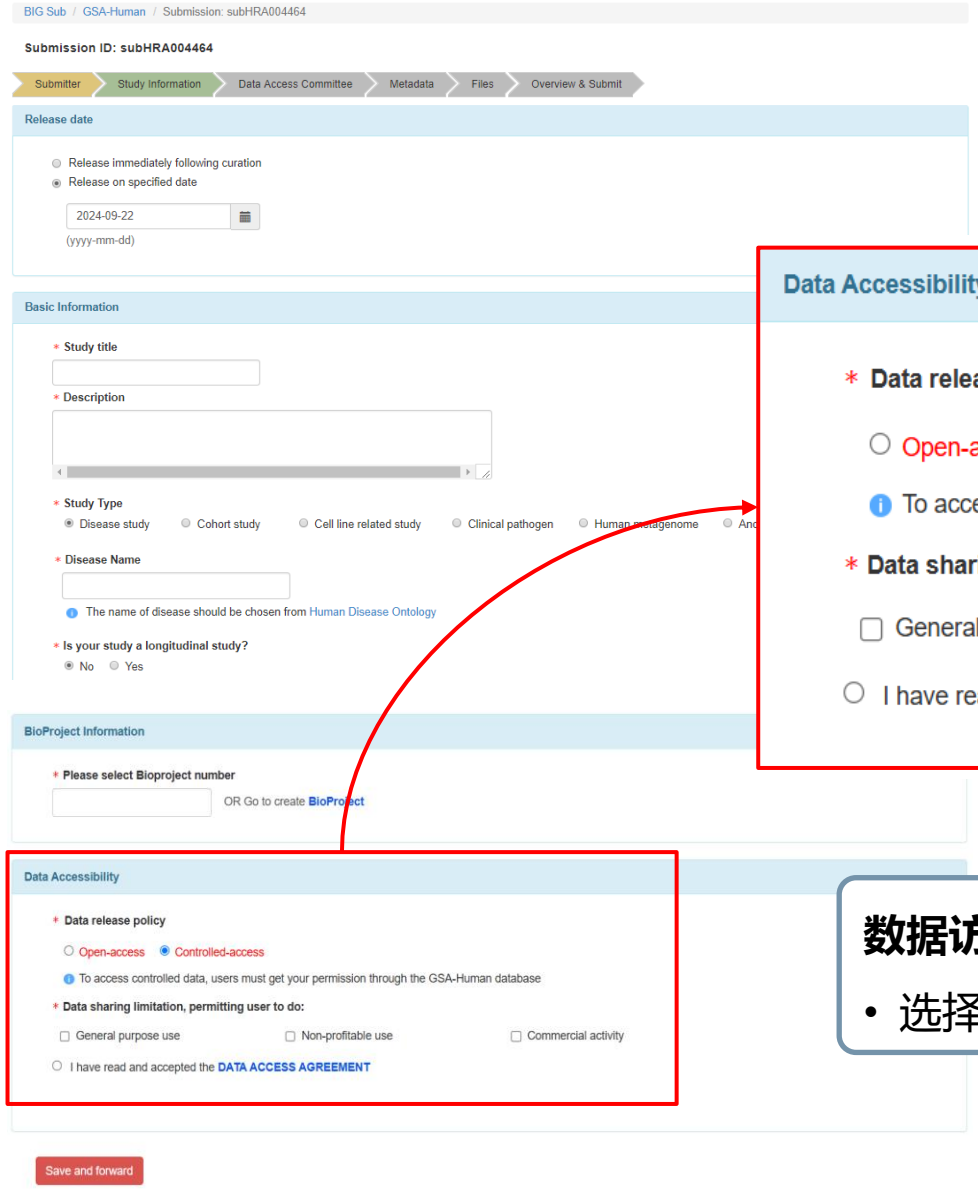
基本信息：选择“Disease study”和具体疾病类型及其相关信息

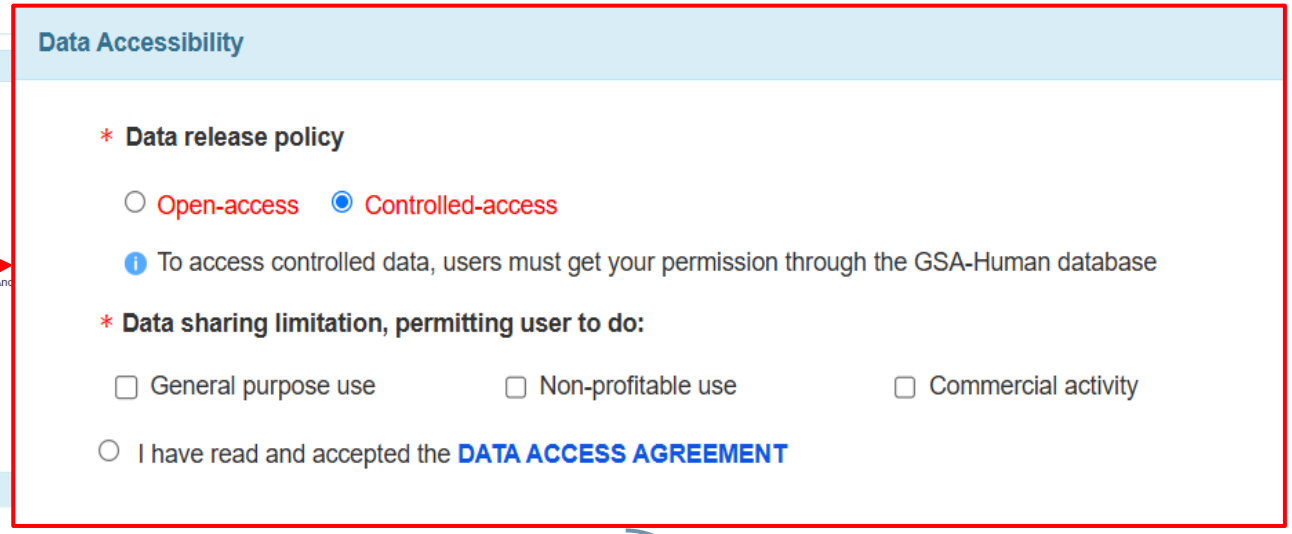
疾病名称请与汇交计划填写一致

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



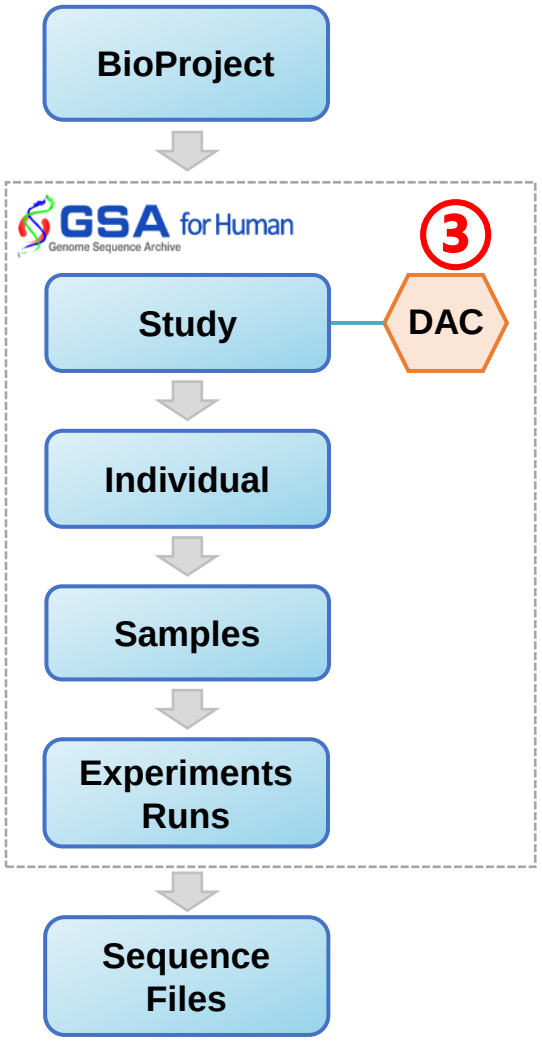


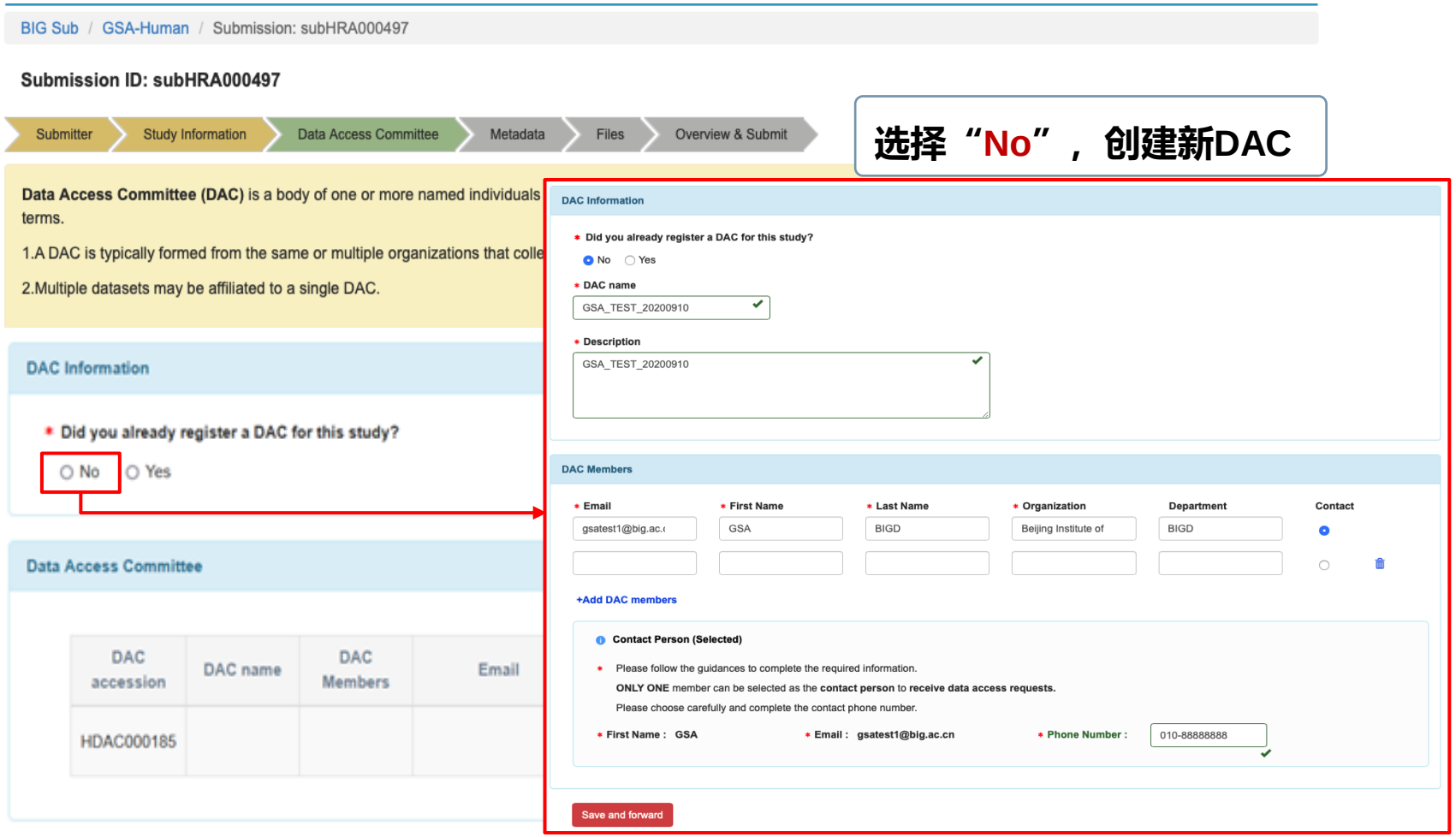


数据访问方式

- 选择数据访问方式（受控访问或开放访问）与使用限制

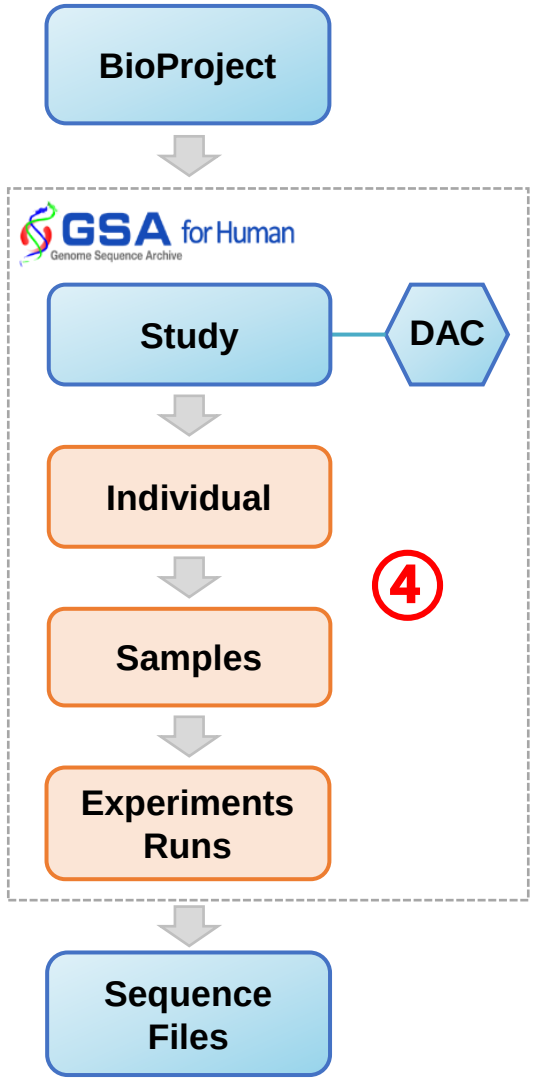
GSA-Human提交：数据管理委员会（DAC）信息

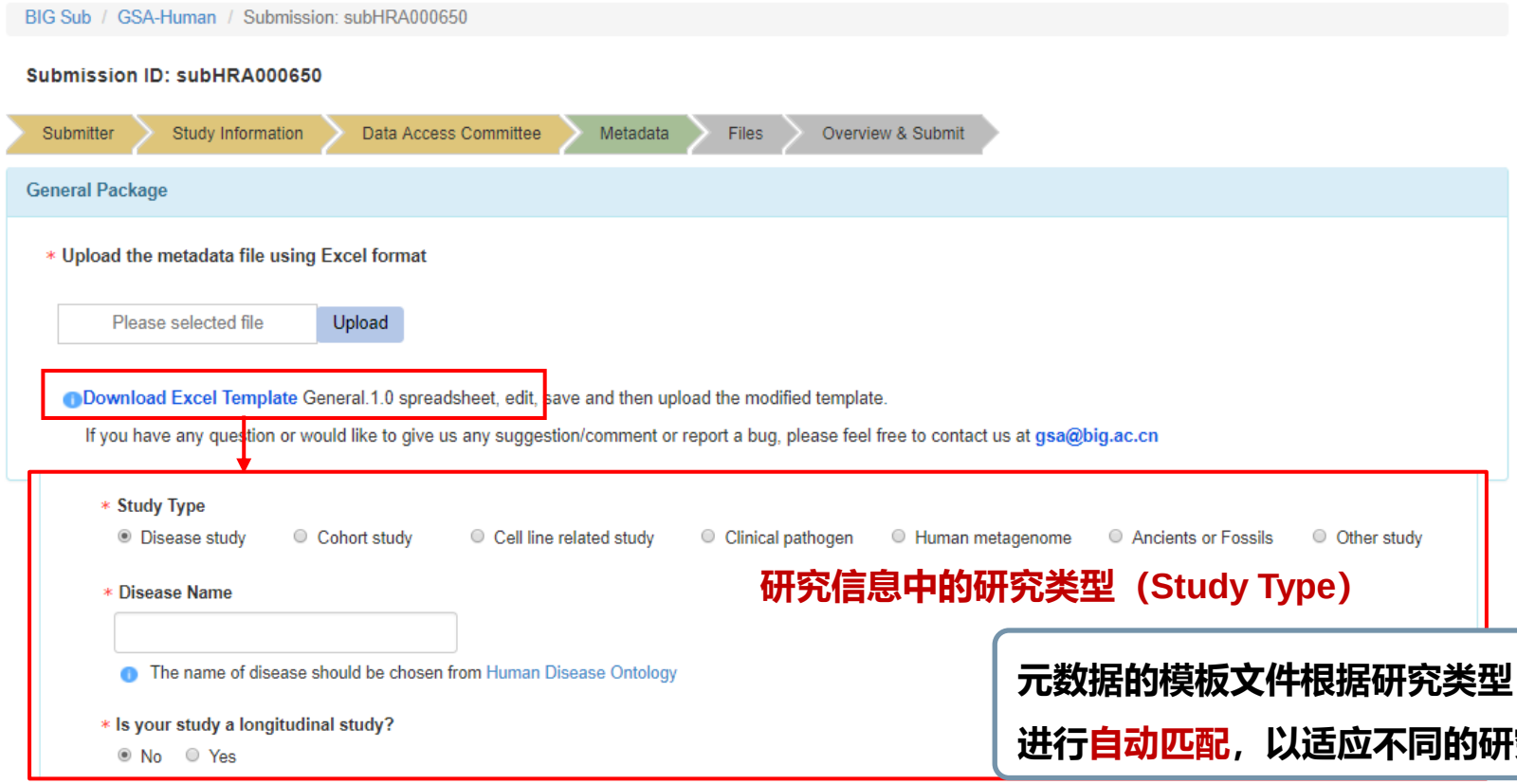




如果数据访问方式使用“受控访问 (Controlled-access)”，需建立数据管理委员会（DAC）并指定DAC联系人，负责数据发布后的申请审批与授权

GSA-Human提交：元数据信息



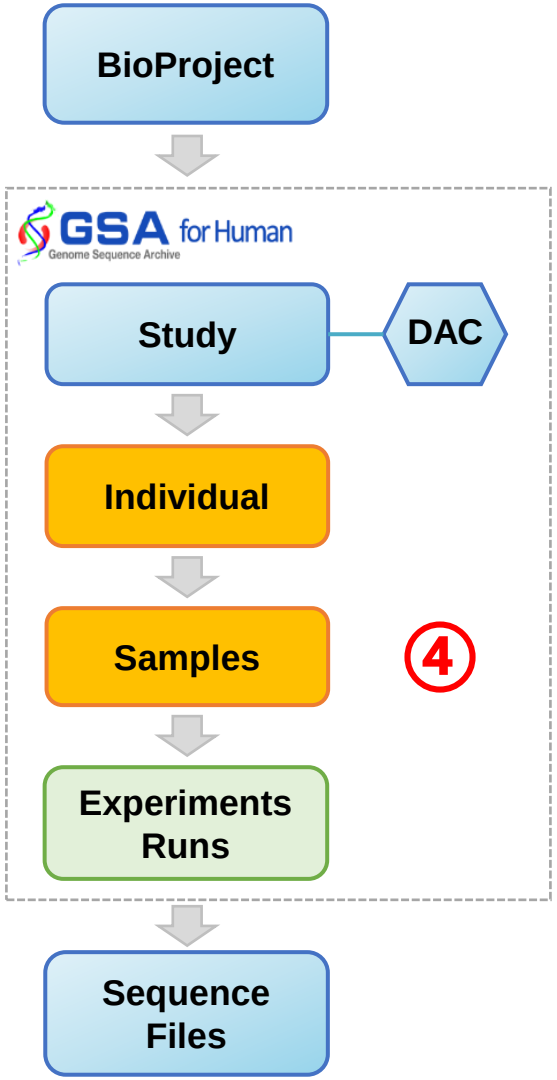


研究信息中的研究类型（Study Type）

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

注：除必填字段以外，IDP项目需提供表格（Individual）中的具体临床信息（Demographic, Diagnosis, Exposure, Most Releated Family History, Follow-Up, Treatment, Pathology Detail）

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 HRI000001, already exists in GSA for Human database
YELLOW columns have drop-down menues that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.
BLUE fields require for individual in Control set of disease study. Please fill in when the individual has disease information. The name of disease should be chosen from Human Disease Ontology(https://disease-ontology.org/)
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	12	13
*ID	*individual_name	*gender	existed_individual_accession	*experiment_group	disease_name	height (m)	body_weight (kg)	body_mass_index(BMI)	province	occupation	birthplace	rac
D1												

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 menues 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

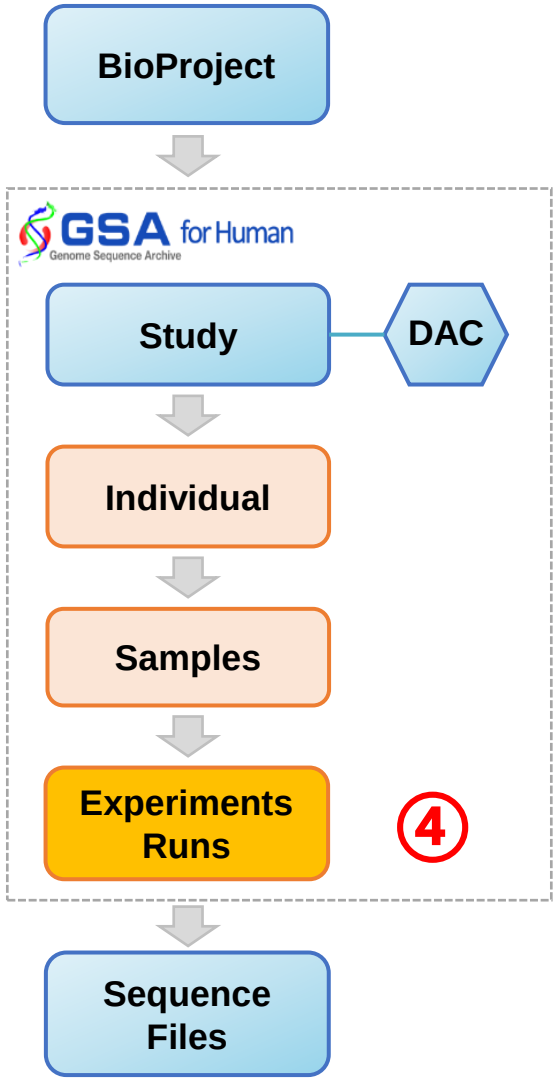
Sample 表格

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

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 menues 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 HeliScope, 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 mate1(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。

常见问题2：建库策略与来源的正确选择
例如是单细胞转录组，请在Strategy选择RNA-Seq、Source选择TRANSCRIPTOMIC SINGLE CELL，并确保与汇交计划填写一致

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 r
YELLOW columns have drop-down menues 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

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							

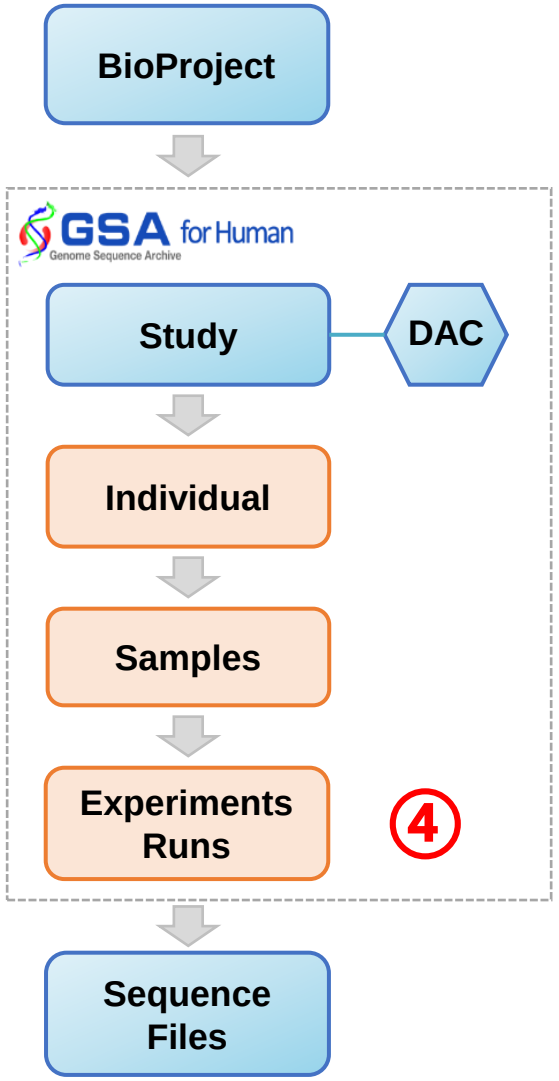
9	10	11	12
Reference file name	MD5 for reference file	Assembly Name or Accession	Assembly Accession URL

Run 表格

常见问题3：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4：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



- 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（721553402或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

BioProject

GSA for Human
Genome Sequence Archive

Study — DAC

Individual

Samples

Experiments
Runs

Sequence
Files

⑤

BIG Sub / GSA-Human / Submission: subHRA000497

Submission ID: subHRA000497

Submitter | Study Information | Data Access Committee | Metadata | Files | Overview & Submit

File upload mode

Notice:

Each file **must be listed 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 overview 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 client to upload files. FTP Username and Password are same to the BIG Sub account.

You may upload your files to our FTP site using an FTP client software (such as the **FileZilla Client**) or using the FTP command.

Address: ftp://submit.big.ac.cn
Username: Same as you login the BIG Sub
Password: Same as you login the BIG Sub

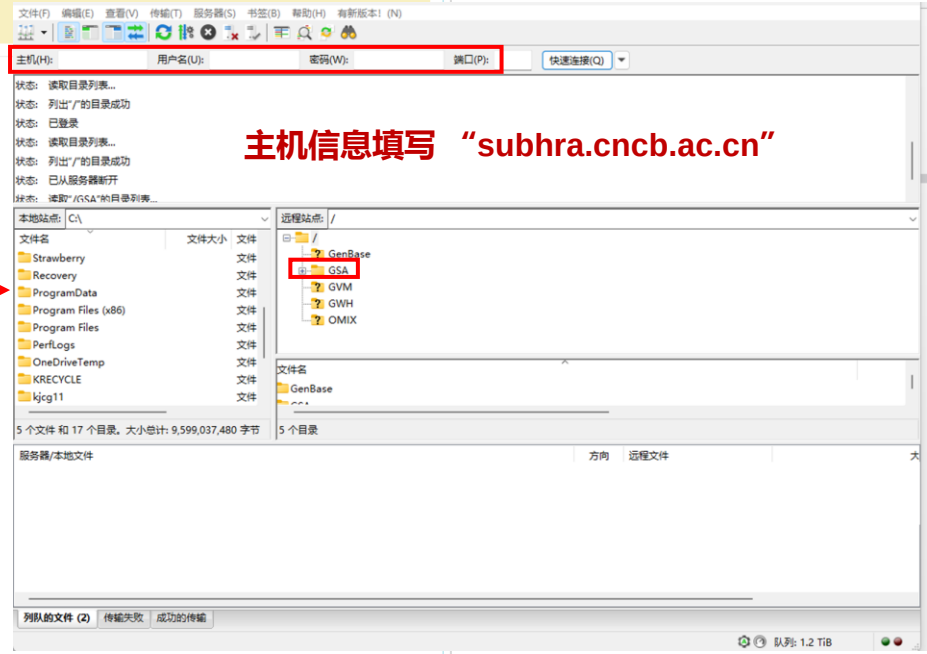
After you connected to the FTP server, please **navigate to the GSA directory and upload your files here. DO NOT upload files to the root directory.**

NOTE: Please use the **binary 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 ftp command, **type the binary command before the mput command**.

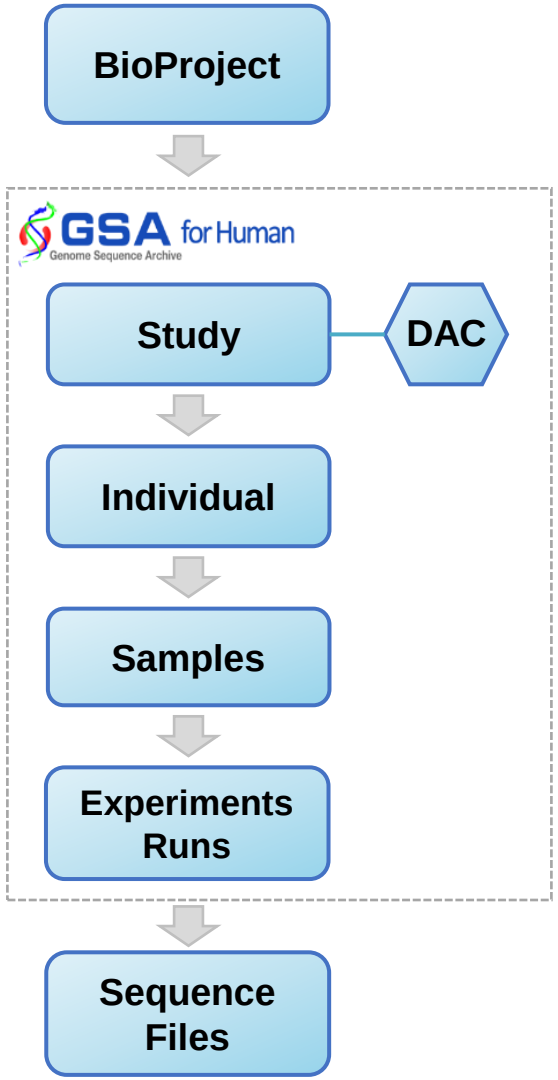
Aspera Command Line

Use Aspera Command Line to upload files.

Save and forward



GSA-Human提交：概览 & 提交



BIG Sub / GSA-Human / Submission: subHRA000497

Submission: subHRA000497

Submitter | Study Information | Data Access Committee | Metadata | Files | Overview & Submit

Study Information

Study Title	GSA_TEST_20200910
Description	GSA_TEST_20200910
Data Accessibility	Controlled access
Data use limitation	Non-profitable use
Data access agreement document	DAA-for-GSA-Human.pdf
BioProject Accession	PRJCA003418
Study Type	Disease Study
Disease Name	immune system cancer
Release date	2022-09-10

Submitter Information

Submitter	BIGD GSA gsatest1@big.ac.cn
Organization	Beijing Institute of Genomics, Chinese Academy of Sciences
Department	BIGD
Country/Region	China
Postal code	100101

Data Access Comittee

Accession	Name	DAC members	Email	Organization	Department	Office phone	Is contact person
HDAC000225	DAC_TEST_GSATEST1	BIGD GSA	gsatest1@big.ac.cn	Beijing Institute of Genomics, Chinese Academy of Sciences	BIGD	010-84097340	Yes
		wang yanqing	wangyanqing@big.ac.cn	BIG	BIGD		No

Metadata Information

Metadata file: General_1.0_us-GSATEST.xlsx 2020-09-10

Individuals & Samples

Individual Name	Gender	experiment_group	disease_name	ethnicity	samples
GSATEST1_2	male	case set	acute myeloid leukemia	Han Chinese	samples(1)

Experiments & Runs

Experiment Title	Platform	Sample Name	Selection
GSATEST1_2_S1_E1	Illumina HiSeq X Ten	GSATEST1_2_S1	cDNA
Run Alias	**Run Data File Information**		**File Type**
GSATEST1_2_S1_E1_L001	File: human_f1.fq.gz File: human_r2.fq.gz		fastq

Total Items: 1 | Items of 1 - 1 | Page size 10 | Page 1/1 | << First | Last >> | Jump To 1 GO

Submit

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

- **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（** 721553402或548170081 **）交流，或通过电子邮件**gsa@cncb.ac.cn**联系我们**
- **数据归档后，将分配唯一的序列号（Accession）并更新至项目汇交进度中**

GSA-Human提交：数据引用

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]

CONTENTS

1 数据汇交流程

- 人类遗传原始测序数据 GSA-Human

2 汇交注意事项

3 汇交管理可视化

4 资源材料

IDP-Hub数据汇交的注意事项

1 项目信息准确关联

- 汇交时，务必准确填写并关联**项目批准编号**，此编号是数据与项目绑定的唯一标识。**若缺失或错误，将导致数据无法通过接口同步至汇交管理系统（IDP-Hub），影响进度跟踪与成果归属**

2 临床数据汇交规范

- 临床数据必须按模板格式汇交。血常规、尿常规等检测结果文件需要转换为标准格式，支持在模板表里添加字段，但不可以直接上传校验报告

3 数据质量前置审核

- 秉持“**质量过关，方可汇交**”。请依据指南中的数据标准（https://ngdc.cncb.ac.cn/idp/introduction#std）设计项目测序数据，完成测序后，请先自行完成数据质控标准的校验，确保数据的有效性和高质量性后再进行汇交

IDP-Hub数据汇交的注意事项

4 项目信息准确关联

- 对于分批次汇交的同类型的数据，后续批次再追加数据时，**无需新建数据集ID**，直接在原来的数据集编号下追加新样本即可

疾病类型	汇交数据类型	计划汇交样本数量	样本数汇交进度	计划汇交数据量(TB)	数据量汇交进度	数据集
搜索	搜索	搜索	搜索	搜索	搜索	搜索
Healthy	scRNA-seq	200	12869.00% (25738)	7.6	1087.03% (82.61)	HRA008649 HRA009319 HRA008986 HRA008322 HRA008896 HRA008620 HRA008619 HRA010469 HRA010657 HRA010797 HRA011429 HRA013386 CRA022299 CRA022086 CRA028526 CRA029766

无需新建数据集！！！！在第一个数据集下追加样本即可

IDP-Hub成果信息的注意事项

为确保您汇交数据的学术影响力得到准确追踪与展示，请遵循以下规范关联发表文章：

1 文章标记数据集ID

- 只有在发表文章中明确标注了本数据集编号（GSA / GSA-Human / OMIX 数据编号），才会被系统认定为数据集关联的文献（并不是指Funding模块填写的项目资助号）

Data availability

The raw sequencing data generated in this study have been deposited in the Genome Sequence Archive[76] at the National Genomics Data Center[77], China National Center for Bioinformation/Beijing Institute of Genomics, Chinese Academy of Sciences (GSA-Human) under accession code HRA008519 (https://ngdc.cncb.ac.cn/gsa-human/). Access to these

2 关联发表文献

- 文章发表后，可随时登录系统，在对应数据集页面**手动补录**相关文献信息（如DOI、标题）

CONTENTS

1 数据汇交流程

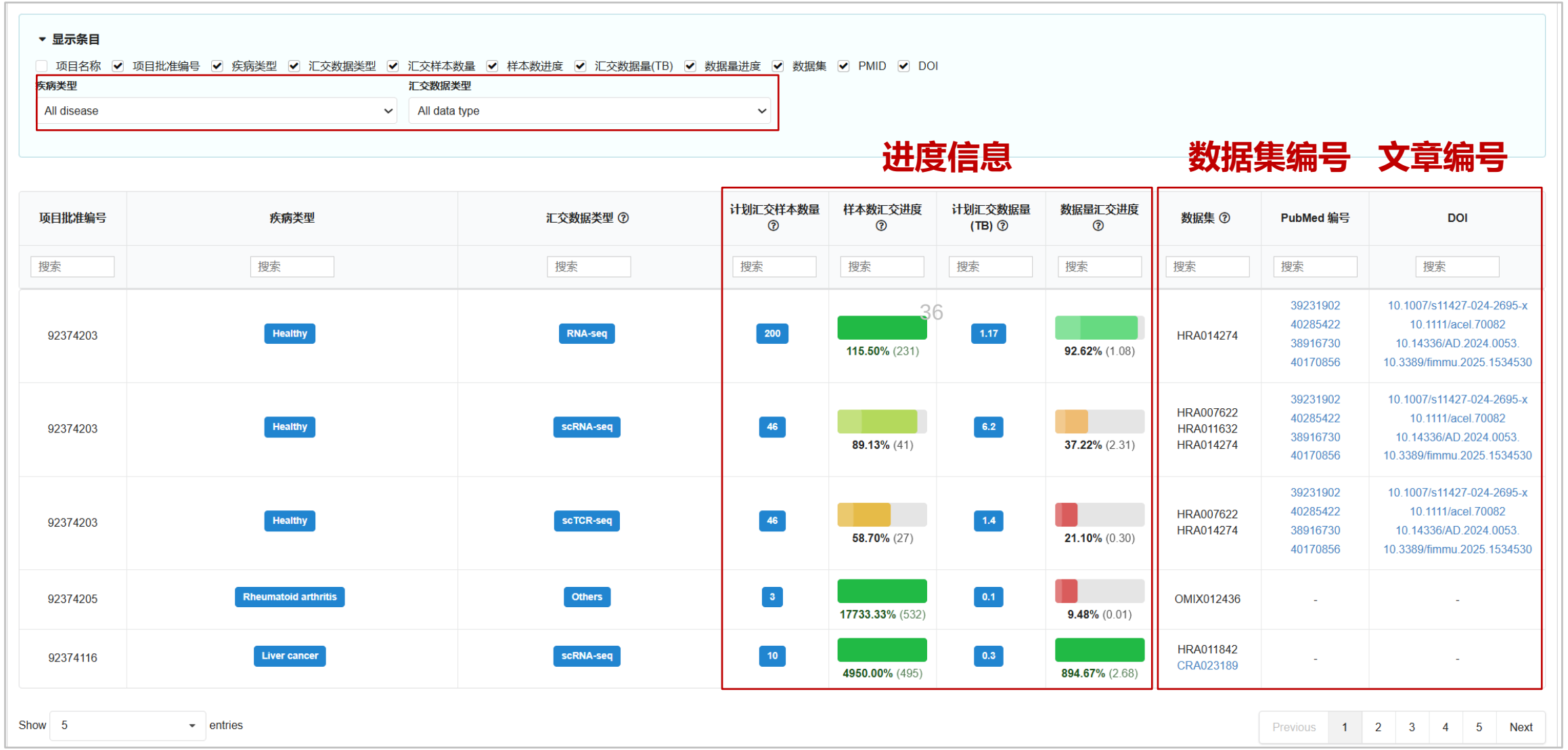
- 人类遗传原始测序数据 GSA-Human

2 汇交注意事项

3 汇交管理可视化

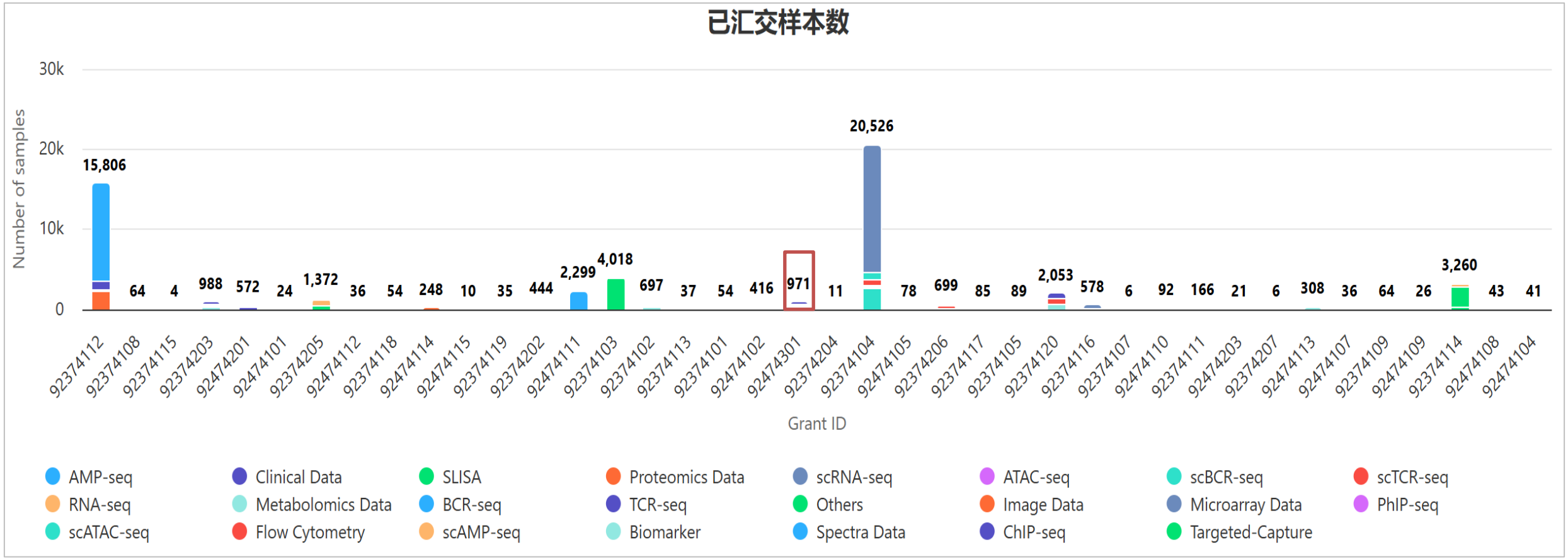
4 资源材料

IDP数据汇交进度的实时可视化



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

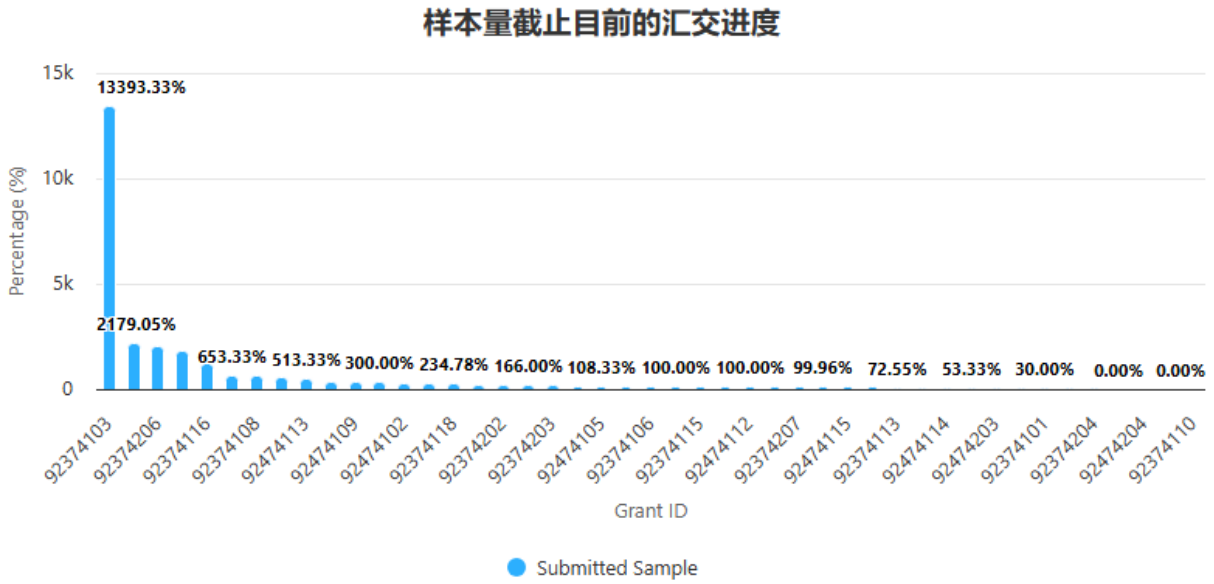
IDP各项目已汇交样本数统计

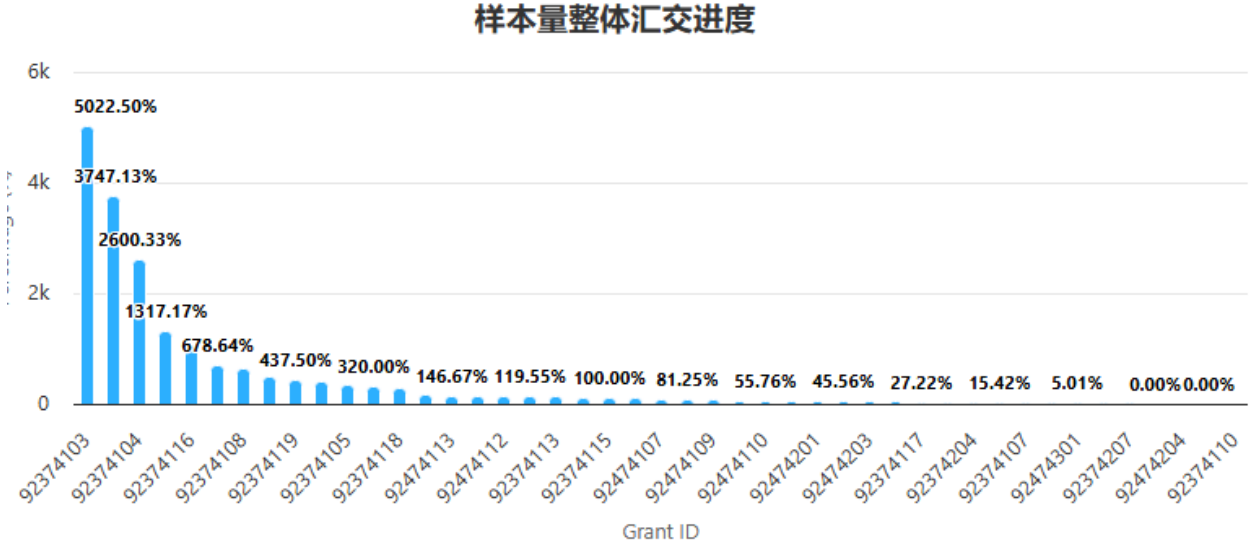


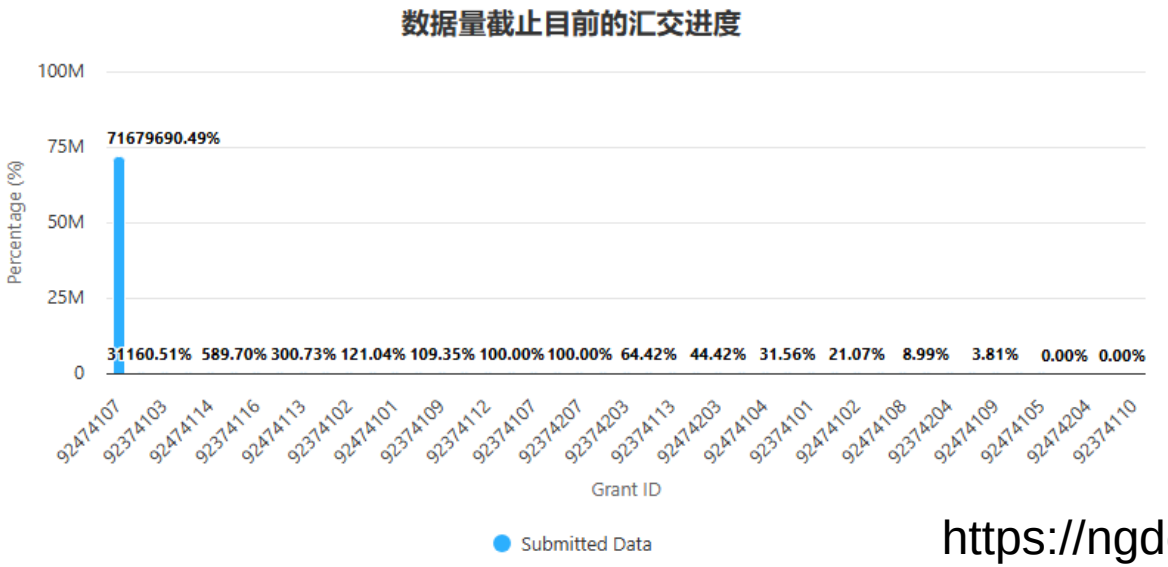
该可视化结果图表在IDP首页展示，可供各课题在年度及中期各类汇报中下载和使用

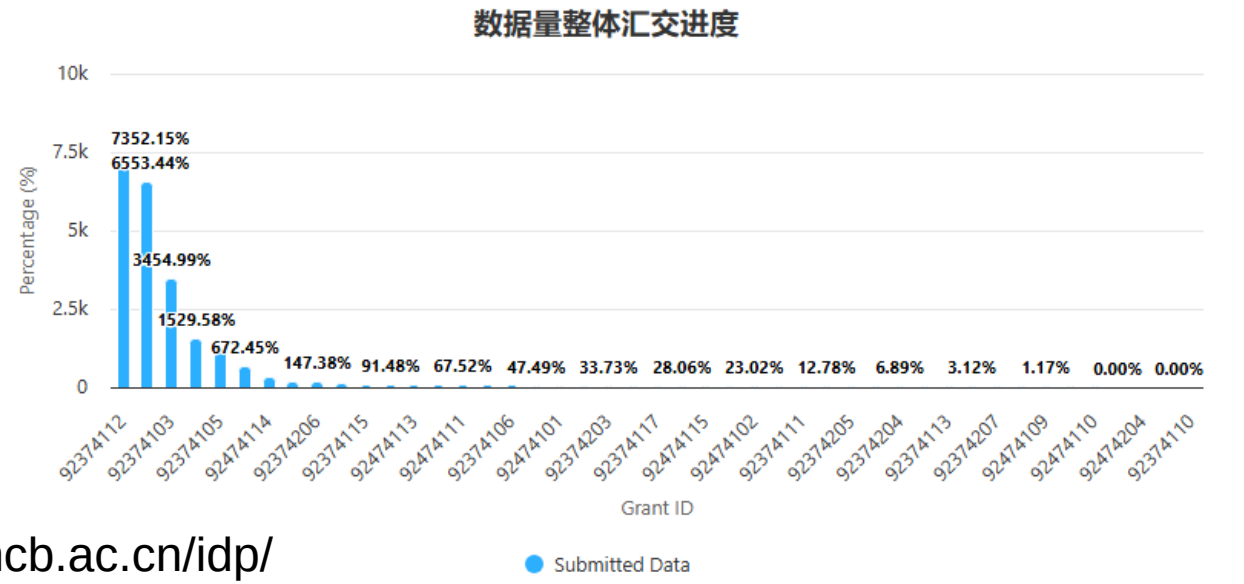
https://ngdc.cncb.ac.cn/idp/

IDP数据汇交进度的实时可视化









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

CONTENTS

1 数据汇交流程

- 人类遗传原始测序数据 GSA-Human

2 汇交注意事项

3 汇交管理可视化

4 资源材料

资源资料

- **免疫力重大研究计划（IDP）的数据标准** https://ngdc.cncb.ac.cn/idp/introduction#std
- **IDP数据汇交计划书和承诺书模板** https://ngdc.cncb.ac.cn/idp/introduction
- **IDP多模态数据汇交模板及流程** https://ngdc.cncb.ac.cn/idp/submit

- **人类遗传资源信息备案流程** https://ngdc.cncb.ac.cn/hgrip/help

  https://ngdc.cncb.ac.cn/gsa-human/document/Principle_of_Accessing_Human_Genetic_Resource_Data_in_NGDC_V1.pdf
- **人遗相关的原始数据递交细节使用说明 （GSA-Human）**

  https://ngdc.cncb.ac.cn/gsa-human/document/GSA-Human_Submission_Guide_1.1.cn.pdf

- **非人遗相关的原始数据递交细节使用说明 （GSA）**

  https://ngdc.cncb.ac.cn/gsub/document/GSA-GSA_Submission_Guide_2.2.cn.pdf

谢　谢

欢迎随时与我们联系

如在数据汇交中遇到任何问题，或需要进一步协助，请随时通过以下方式与我们沟通。

我们将全程提供支持，助力您顺利完成数据汇交

汇交交流QQ群：721553402

IDP专用邮箱：idp@big.ac.cn