组学原始数据归档库（GSA）使用说明

系统简介.....2

用户注册.....2

元数据信息录入.....3

第一步：创建 BioProject.....3

第二步：创建 BioSample.....8

第三步：创建 GSA.....12

数据文件上传.....16

FTP 上传.....16

协助上传.....16

同网大型机协助拷取.....16

数据信息批量上传说明.....17

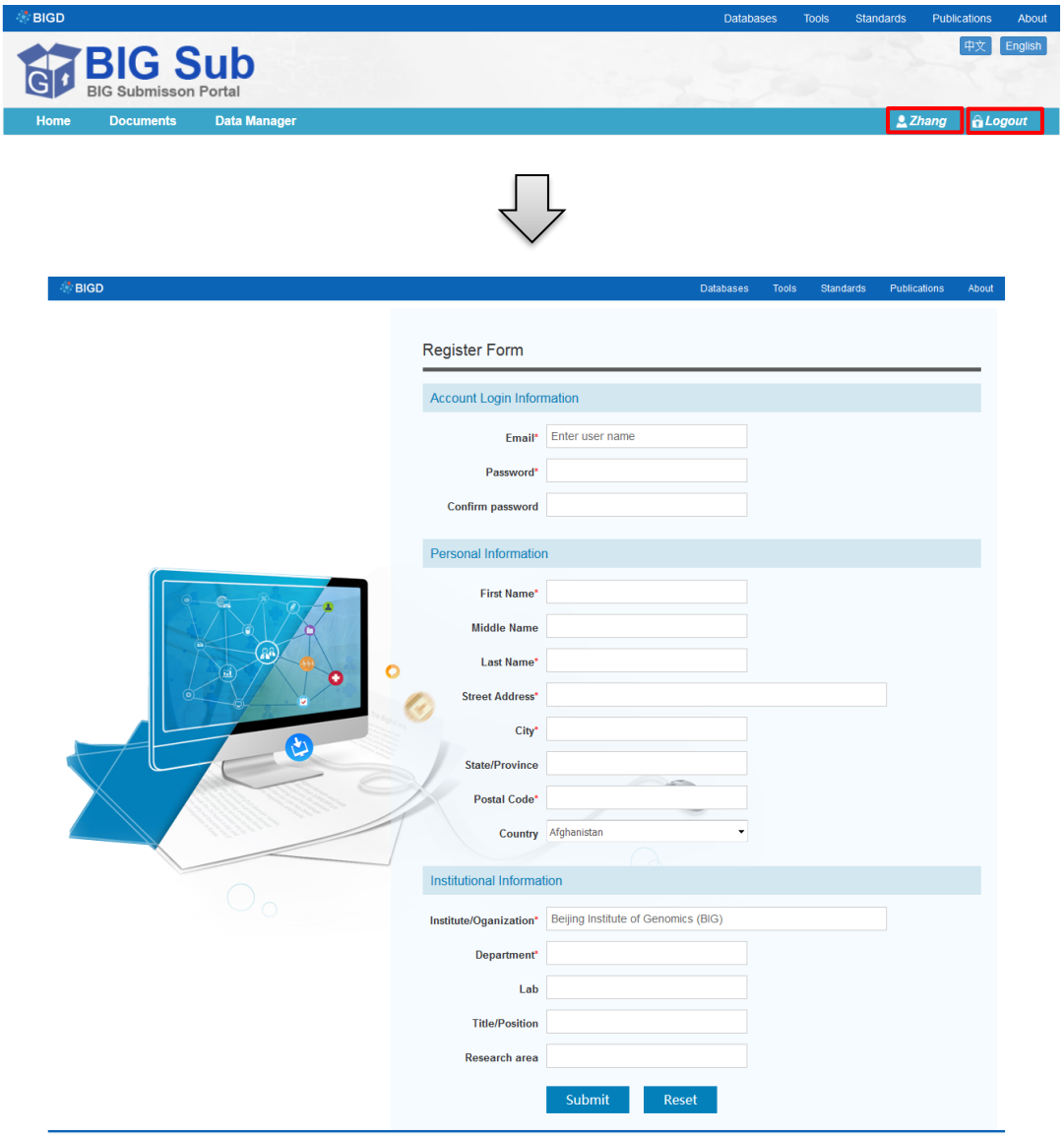
数据触发机制说明.....19

系统简介

组学原始数据归档库（Genome Sequence Archive, GSA）是**组学原始数据汇交、存储、管理与共享系统**。GSA 用户可通过大数据中心生物数据统一汇交入口——生物数据递交系统（BIG Submission）完成一站式数据递交。

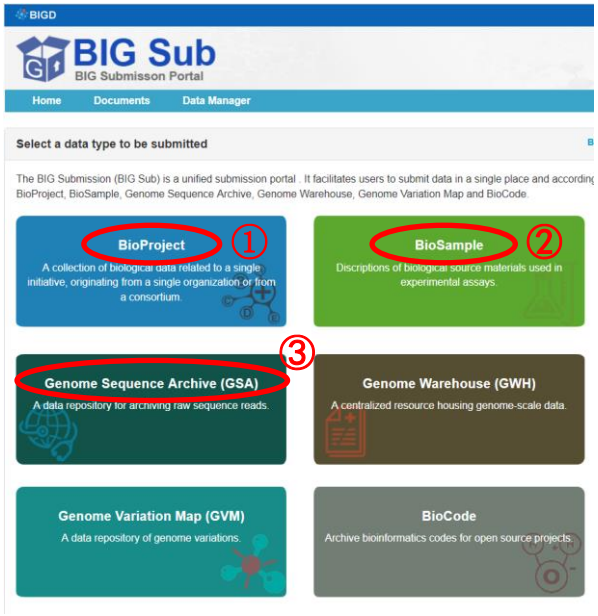
用户注册

请您进入生物数据递交系统（BIG Submission，BIG Sub，http://bigd.big.ac.cn/gsub/）完成账号注册，建议**使用实验室公共邮箱进行注册**。如果您在账号注册和使用过程中遇到任何问题，请联系 bigd-admin@big.ac.cn。



GSA 数据集信息录入

账户注册完成后，您可遵循以下原则进行数据信息录入，再通过 FTP 完成测序文件上传：



① 所有用户都需先在 **BioProject** 中创建 **Project**；

② 当数据样本（Sample）数量＜10 个，建议直接进入 BioSample 数据库，通过“在线模式”创建相应数量 Sample；

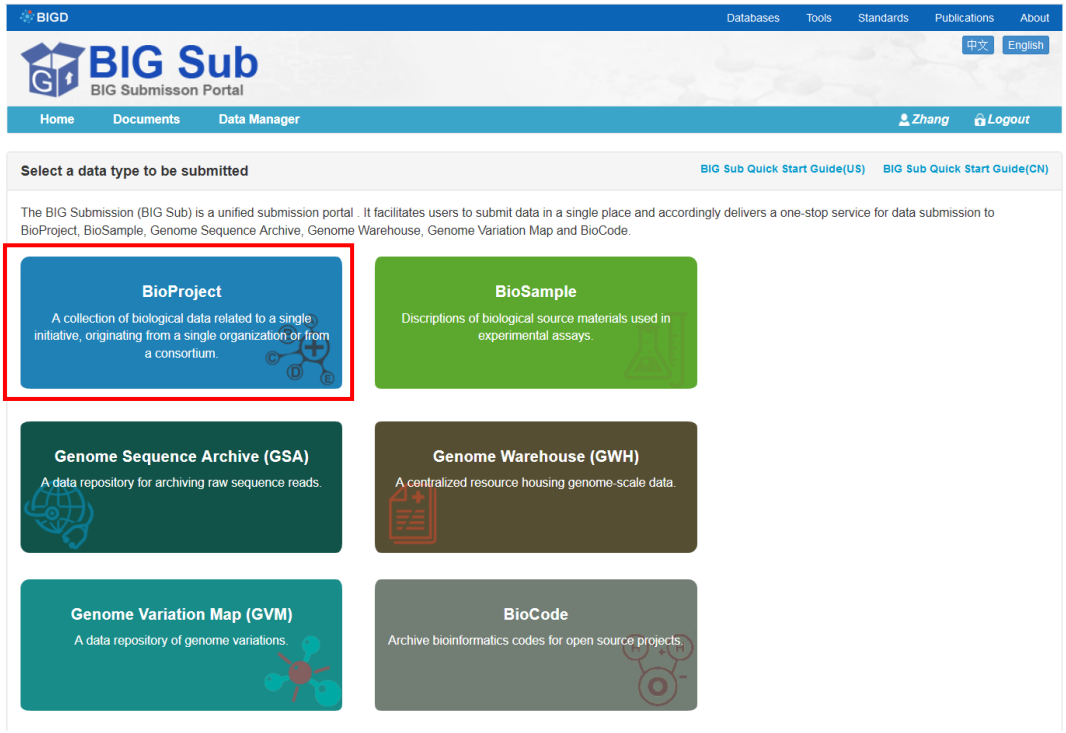
当数据样本（Sample）数量≥10 个，并希望采用“离线模式”进行元数据信息批量导入的用户，详见“数据信息批量上传说明”；

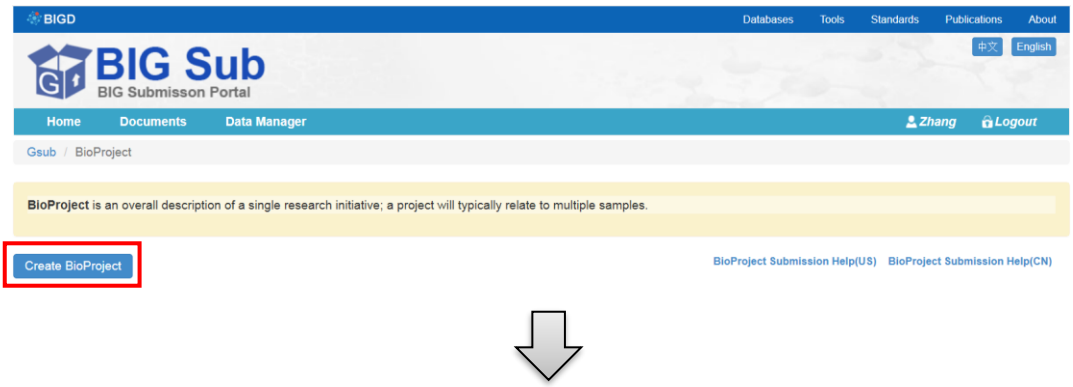
③ 选择“在线模式”的用户，请在 GSA 数据库中完成 GSA 数据集创建，并完善相应 Experiment 和 Run 信息录入

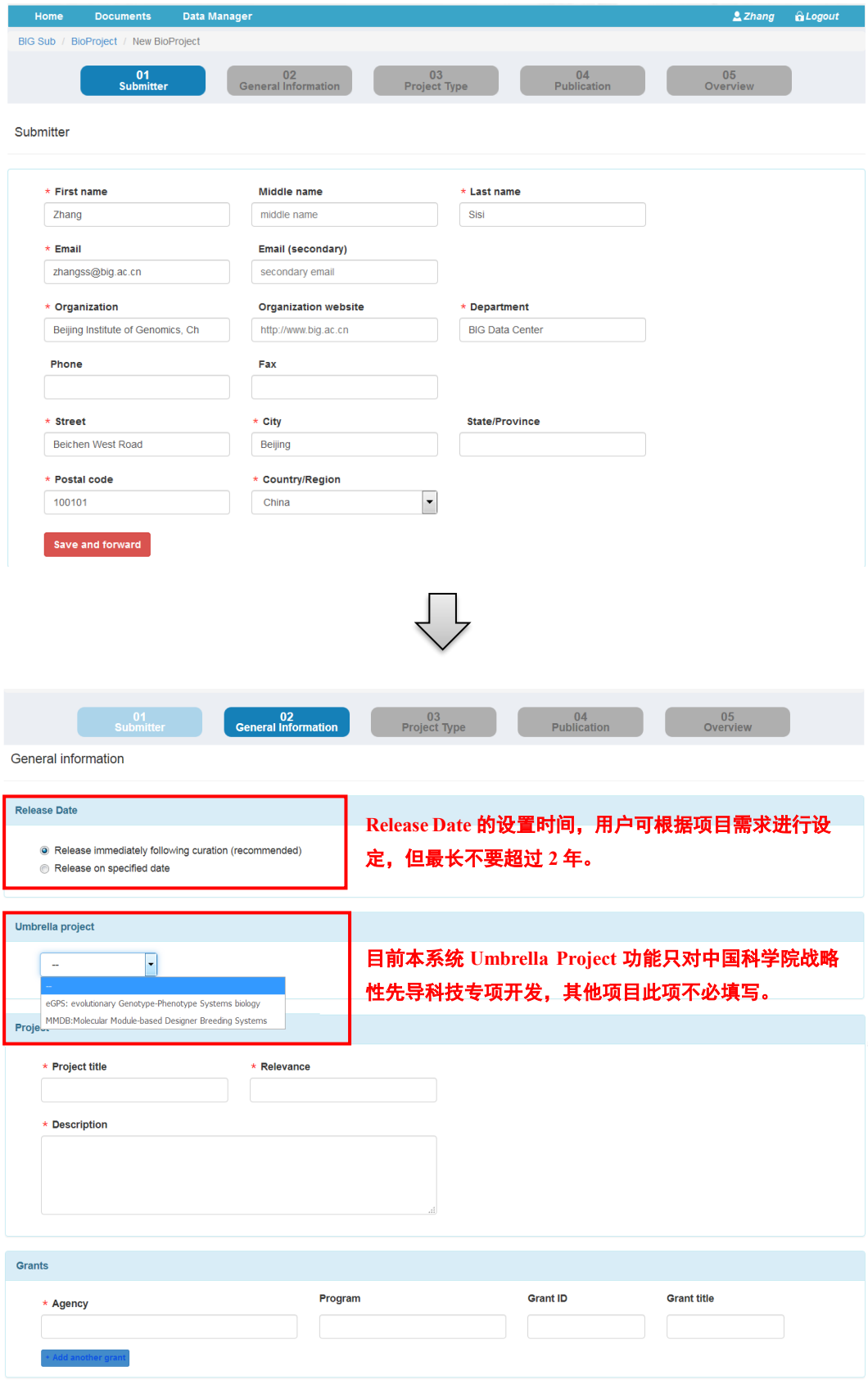
注：GSA 数据集的整体 Accession Number 为 CRAxxxxxx，从属于同一个 BioProject 且使用同一批 BioSample 的数据，**只需要创建一个 GSA 数据集**。

第一步：创建 BioProject

如果之前没有创建过 BioProject，依照以下步骤，进入 BioProject 数据库创建 BioProject 并完成相关信息填写：


BIGD
Databases
Tools
Standards
Publications
About
中文
English
BIG Sub
BIG Submisson Portal
Home
Documents
Data Manager
Zhang
Logout
Select a data type to be submitted
BIG Sub Quick Start Guide(US)
BIG Sub Quick Start Guide(CN)
The BIG Submission (BIG Sub) is a unified submission portal . It facilitates users to submit data in a single place and accordingly delivers a one-stop service for data submission to BioProject, BioSample, Genome Sequence Archive, Genome Warehouse, Genome Variation Map and BioCode.
BioProject
A collection of biological data related to a single initiative, originating from a single organization or from a consortium.
BioSample
Discriptions of biological source materials used in experimental assays.
Genome Sequence Archive (GSA)
A data repository for archiving raw sequence reads.
Genome Warehouse (GWH)
A centralized resource housing genome-scale data.
Genome Variation Map (GVM)
A data repository of genome variations.
BioCode
Archive bioinformatics codes for open source projects.

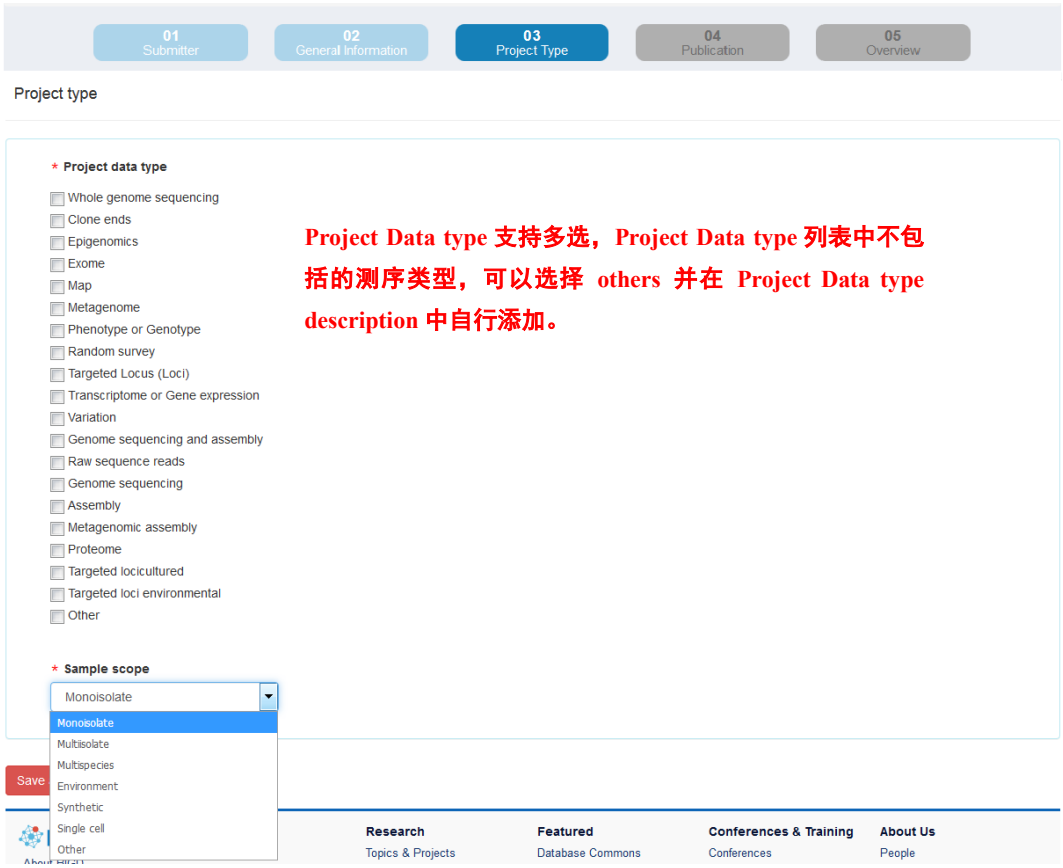

BIGD
Databases
Tools
Standards
Publications
About
中文
English
BIG Sub
BIG Submisson Portal
Home
Documents
Data Manager
Zhang
Logout
Gsub / BioProject
BioProject is an overall description of a single research initiative; a project will typically relate to multiple samples.
Create BioProject
BioProject Submission Help(US)
BioProject Submission Help(CN)



注：

eGPS：evolutionary Genotype-Phenotype biology 为先导 B 动物复杂性状的进化解析与调控项；

MMDB：Molecular Module-based Designer Breeding System 为先导 A 分子模块设计育种创新体系项目。



Monoisolate：同一物种且同一基因型样本，主要用于模式动物/植物或细胞系等。

Multiisolate：同一物种的多基因型的个体样本

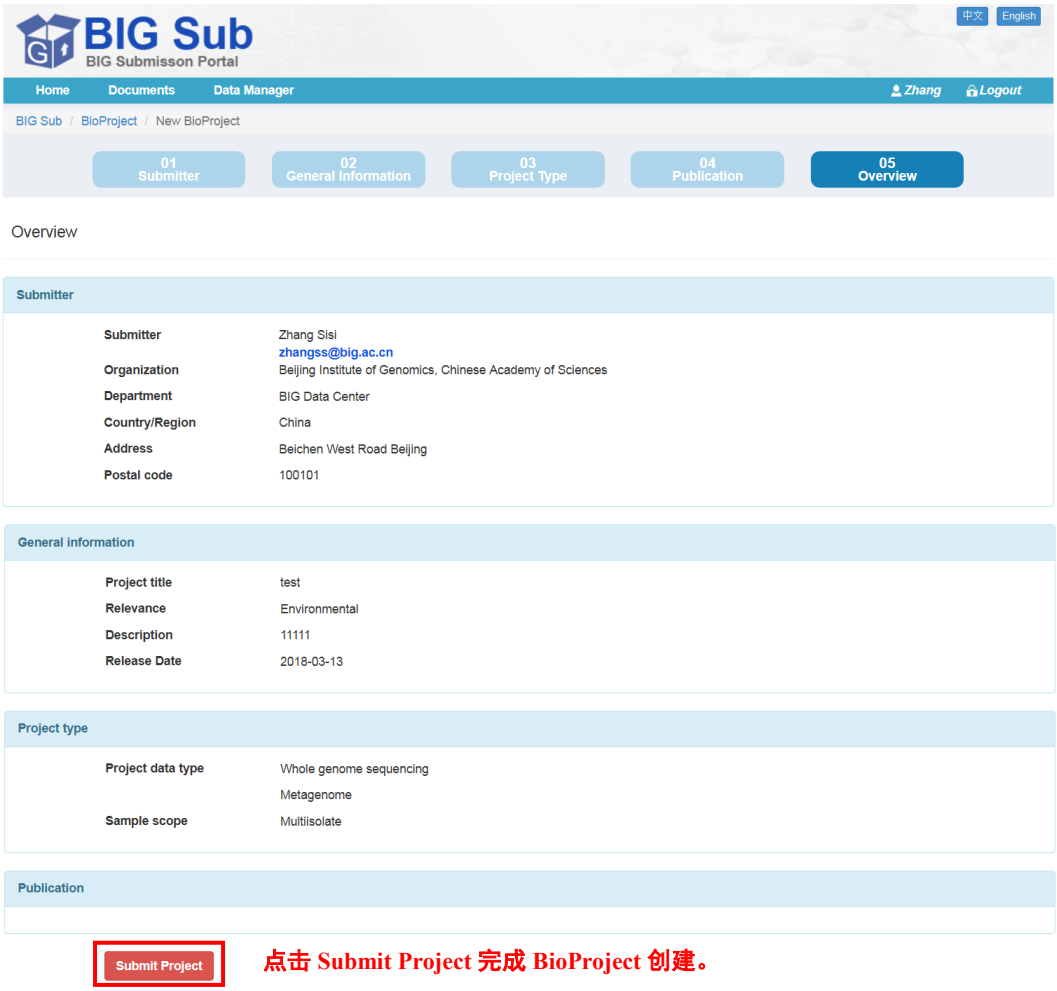
Multi-species：多物种样本集

Environment：环境样本集，样本的种类含量尚不清楚，多用于宏基因组研究

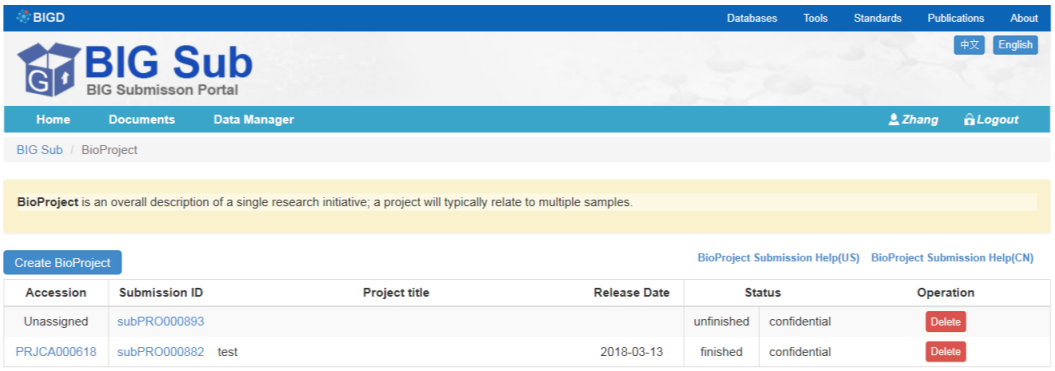
Synthetic：在实验室里制造/合成的样本

Single cell：单个细胞测序得到的序列信息

Other：其他样本/无法归入以上类别的特殊样本



点击 Submit Project 完成 BioProject 创建。

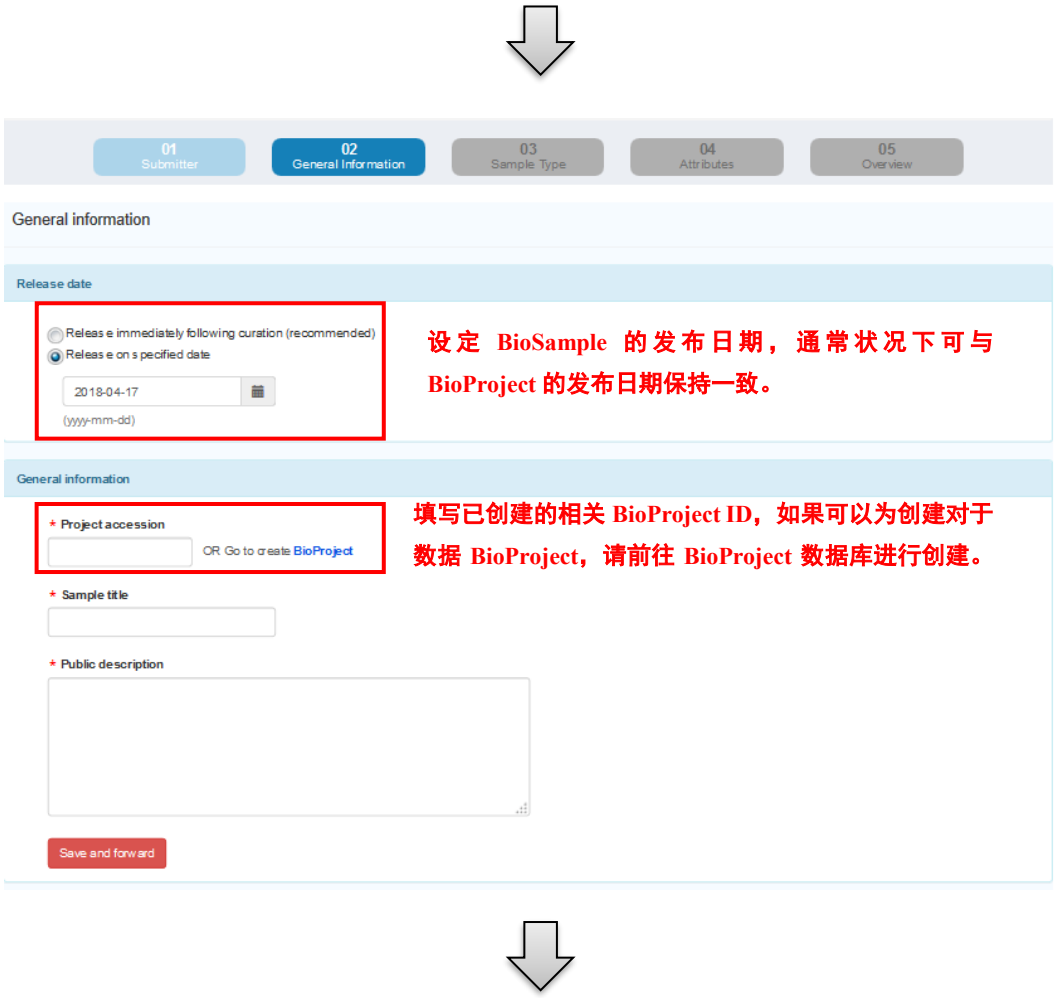


注：用户完成 BioProject 创建后，Status 状态为 Finish; Cofidential，系统将自动为用户生成 Accession number。在管理员审核通过之前，用户可在 BioProject 界面随时修改和删除所创建的 BioProject。 管理员审核通过后，Status 状态会变为 Checked OK; Cofidential，到达用户设定的发布日期后，会变成 Checked OK；Public。

第二步：创建 BioSample

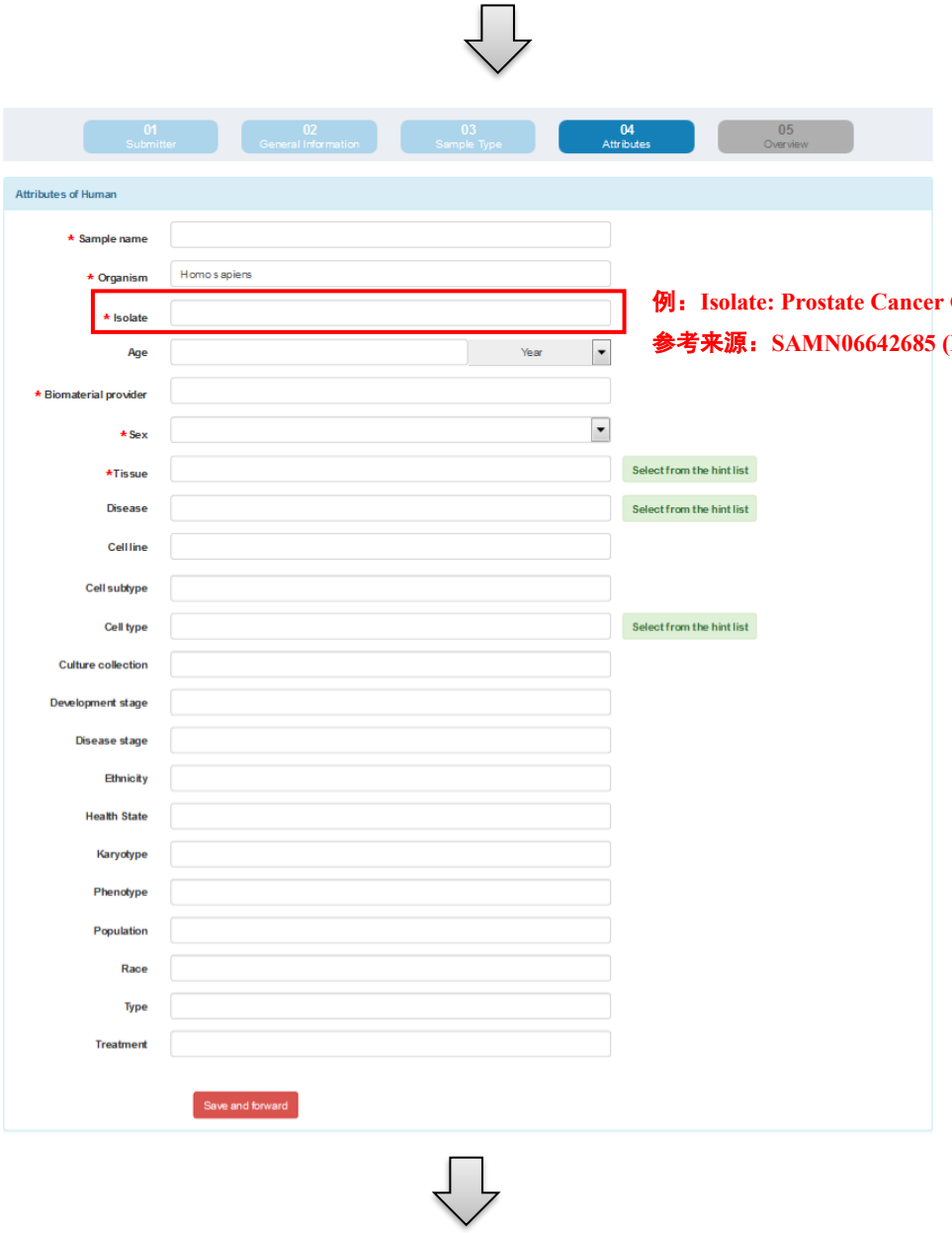
如果数据样本（Sample）数量＜10 个，请依照以下步骤，进入 BioSample 数据库创建相应数量的 BioSample 并完成相关信息填写：

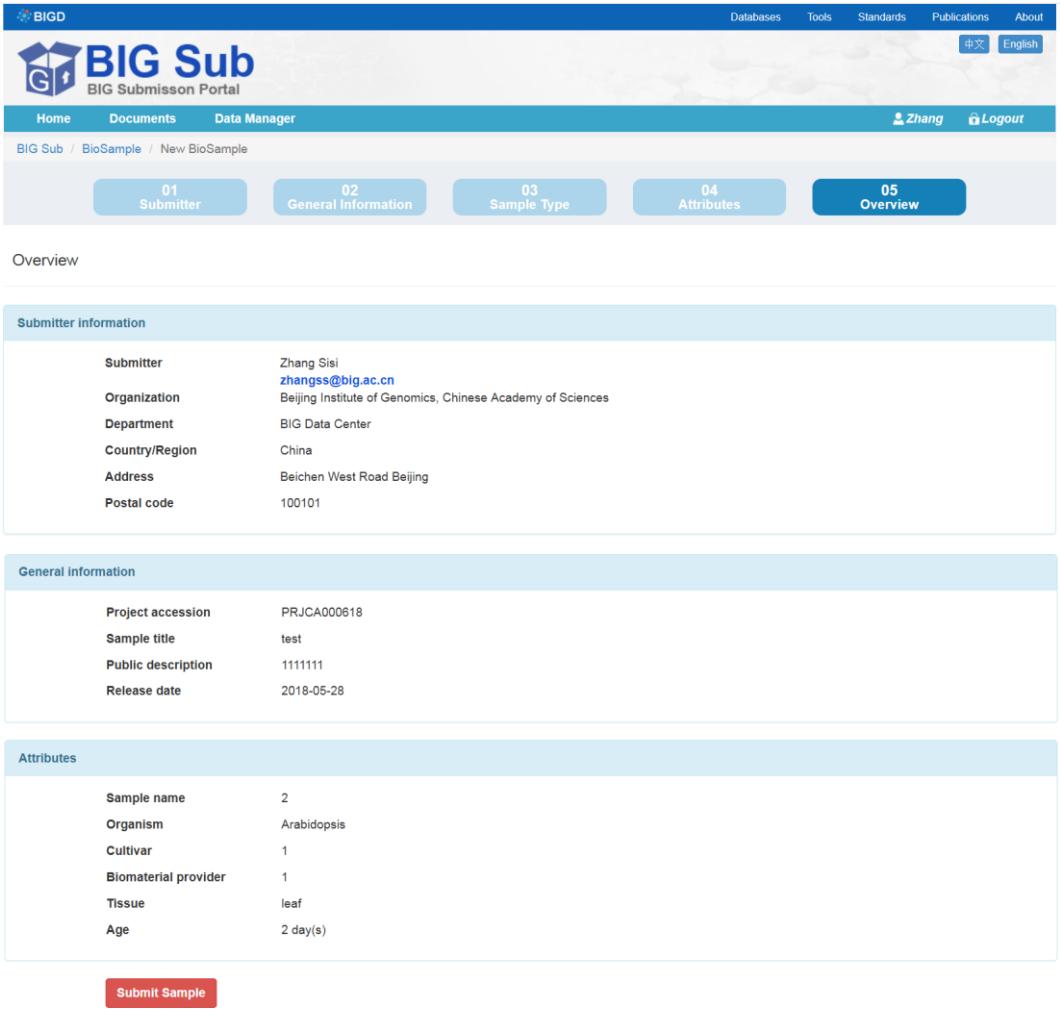
当数据样本（Sample）数量≥10 个，并希望采用“离线模式”进行元数据信息批量导入的用户，详见“数据信息批量上传说明”

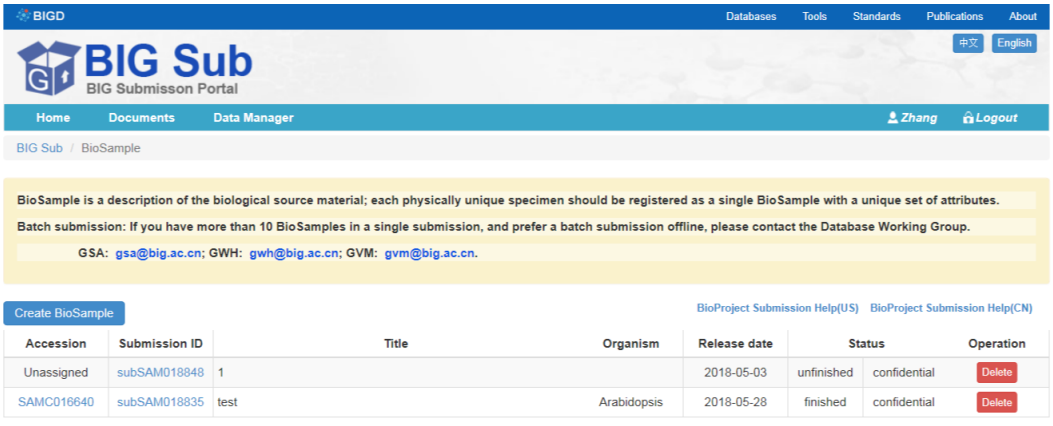


- Clinical or host-associated pathogen：临床或宿主相关病原体样本数据。
- Environmental, food or other pathogen：环境，食品等方面的微生物数据。
- Microbe：来自细菌或其他单细胞微生物，但不包括致病菌或病毒数据。
- Model organism or animal sample：模式生物或动物的多细胞样本或细胞系，如大鼠、小鼠、果蝇、线虫、鱼类、两栖类或其他哺乳动物数据。

- Human Sample：该表格仅用于没有隐私问题的人体样本或细胞系样本。 如果需要提交人类受控数据，请联系 gsa@big.ac.cn 将数据提交到 GSA for human 数据库。
- Plant sample：植物样本或植物细胞系数据。
- Virus Sample：所有与疾病无关的病毒数据，病原体应归类为 Clinical or host-associated pathogen。
- Attributes of metagenome or environmental：目前用于不适于 Environmental/Metagenome sample 的宏基因组生物样本数据。
- Environmental/Metagenome sample (GSC MIMS)：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、土壤和泥土的数据，即 human-gut，soil，water。



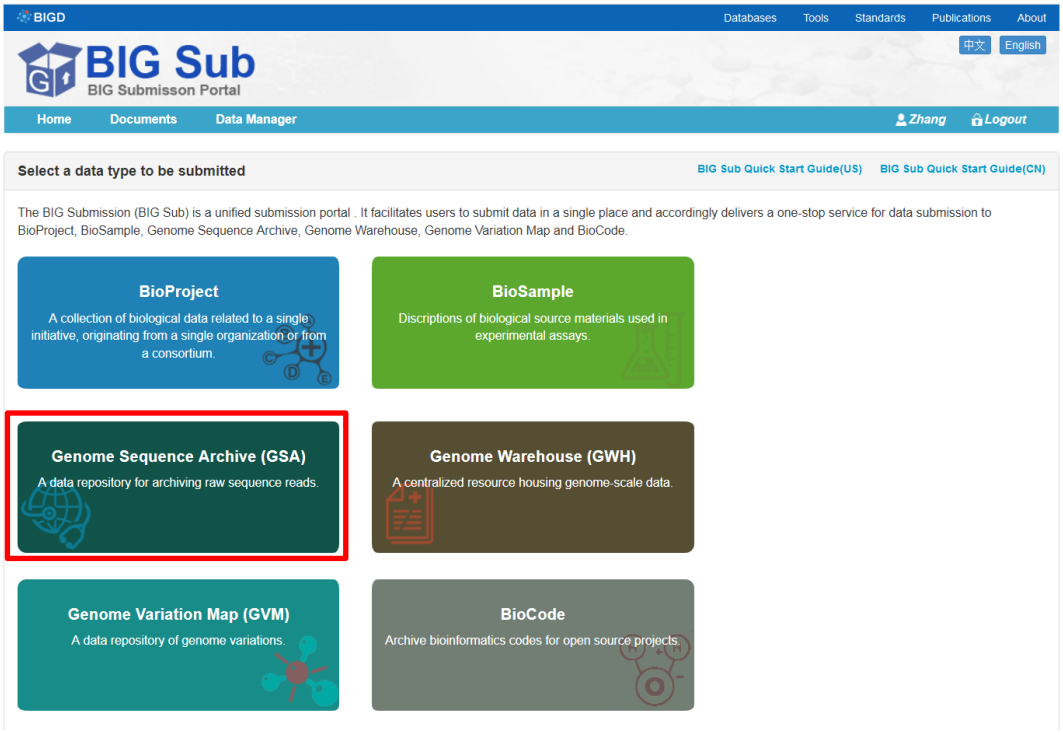


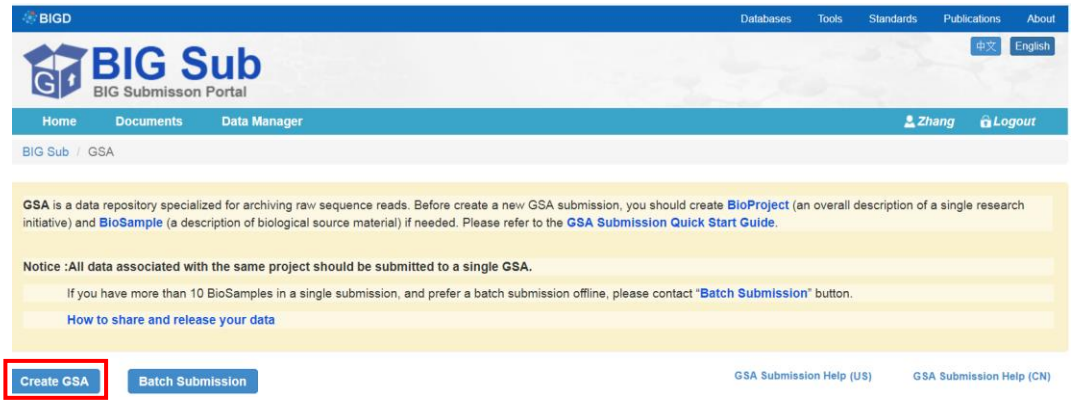


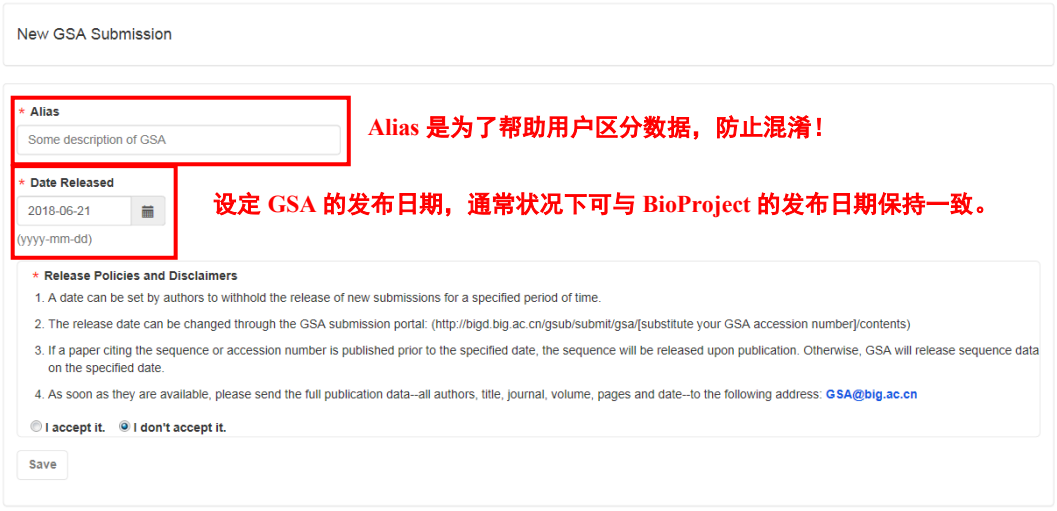
注： 用户完成 BioSample 创建后，Status 状态为 Finish; Cofidential，系统将自动为用户生成 Accession number。在管理员审核通过之前，用户可在 BioSample 界面随时修改和删除所创建的 BioSample。 管理员审核通过后，Status 状态会变为 Checked OK; Cofidential，到达用户设定的发布日期后，会变成 Checked OK；Public。

第三步：创建 GSA

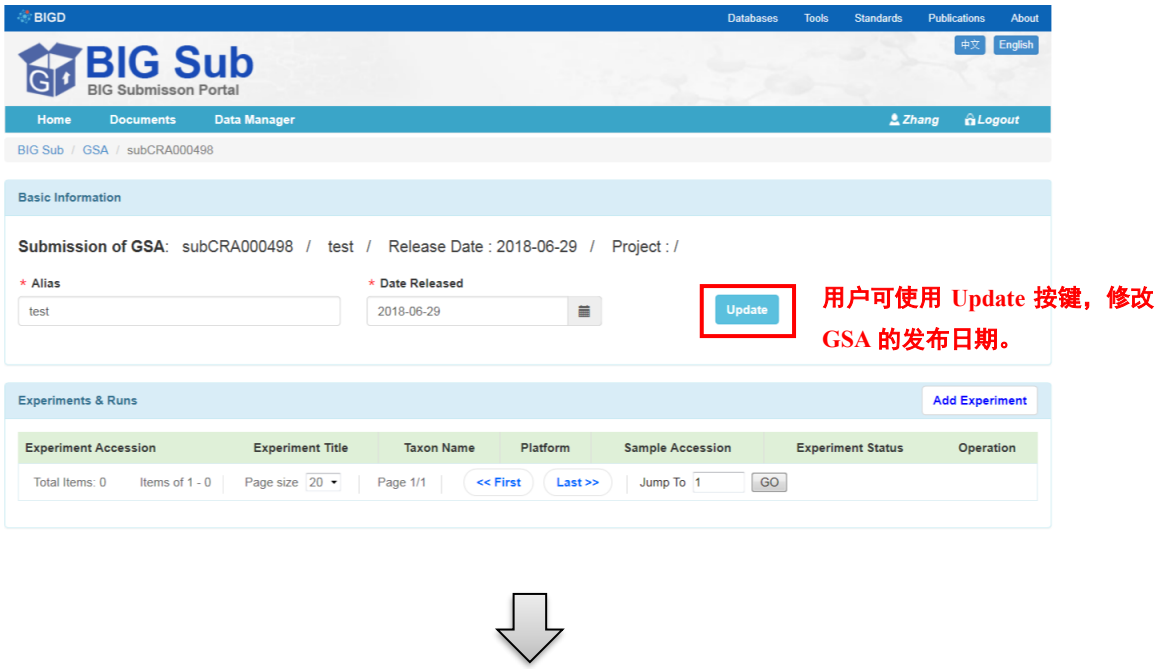
依照以下步骤，进入 GSA 数据库创建 GSA 并完成 Experiment 和 Run 的创建和相关信息填写。注意同一批次提交的数据，且同属一个 BioProject，**一般建议只创建一个 GSA**。

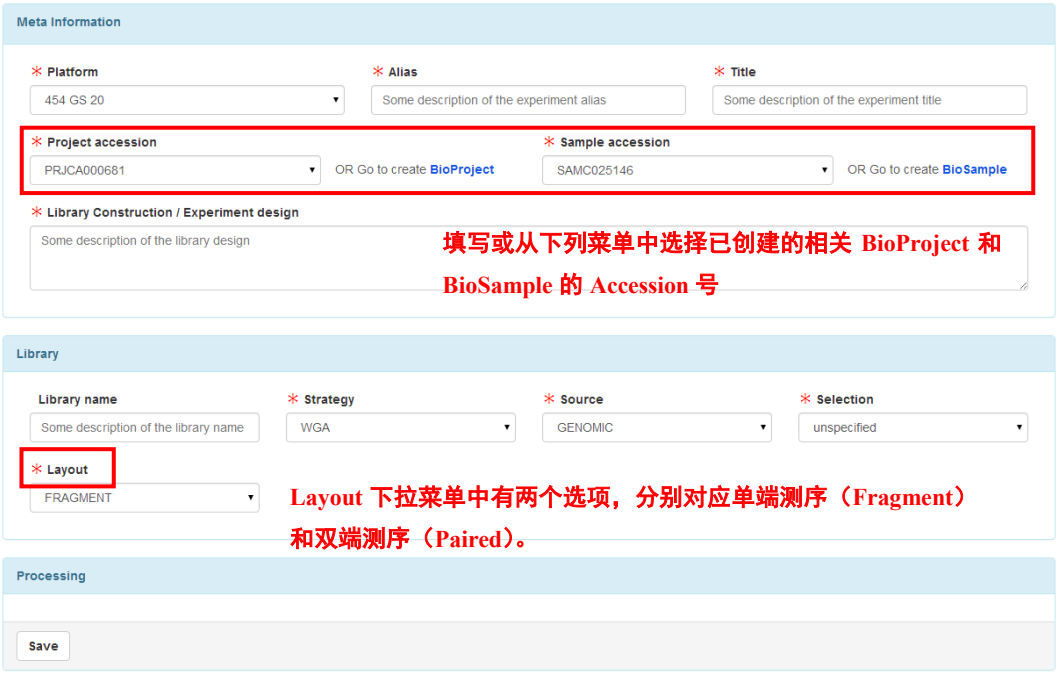




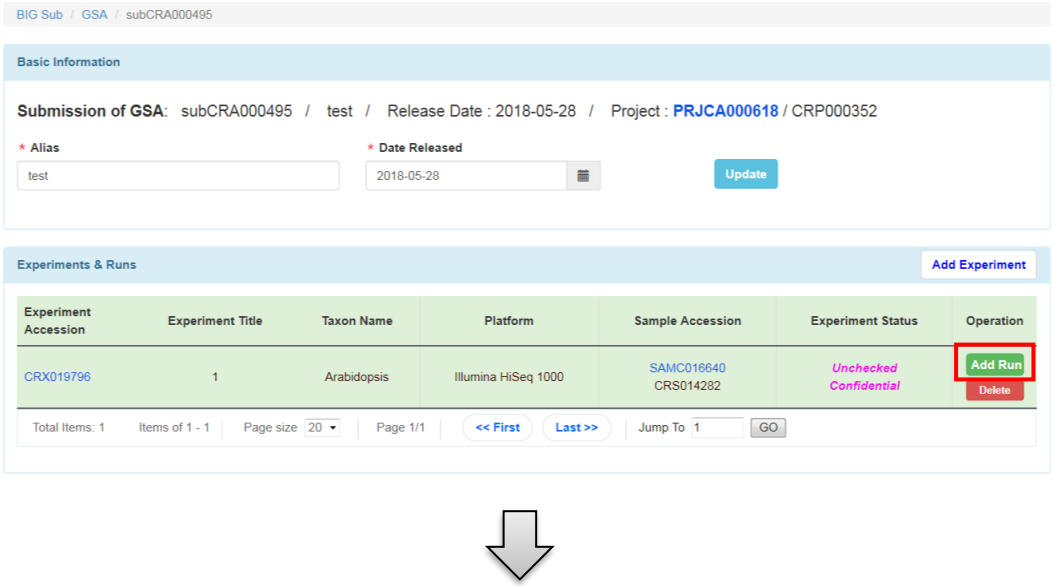


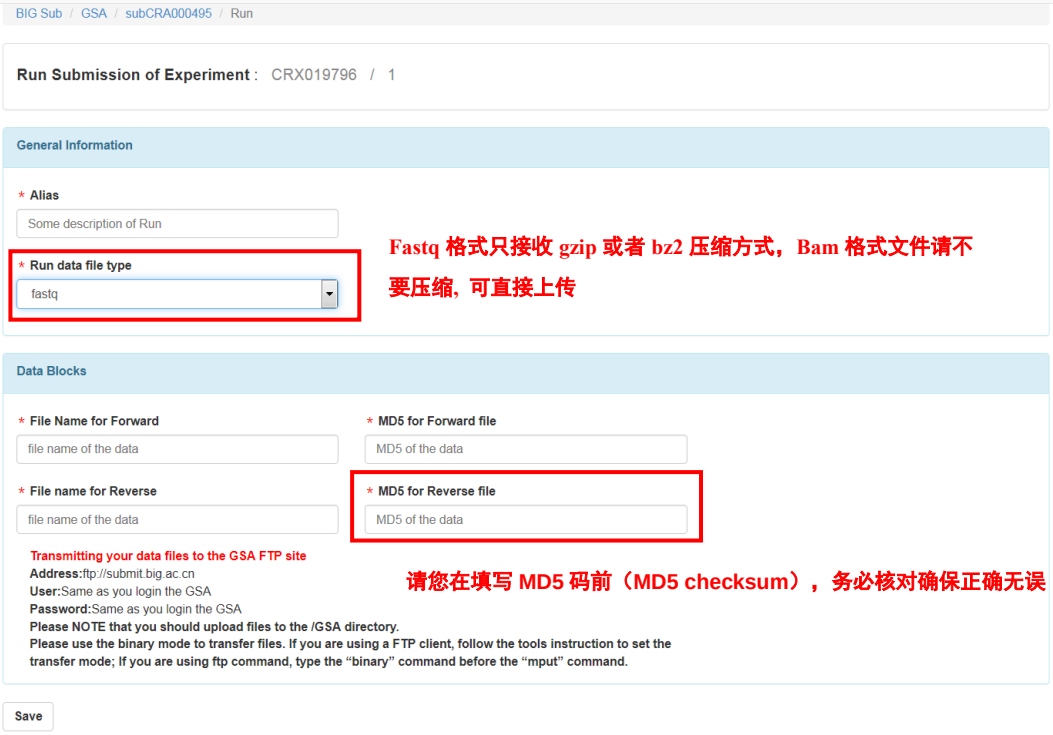
- 使用 Add Experiment 创建新 Experiment，填写完善相关信息；





- 使用 Add Run 创建新 Run：填写相关信息，推荐上传格式为 Fastq 或 Bam，如其他测序格式数据，建议先转换为以上两种格式再进行上传。





注：

1) MD5 码主要是用来校验递交的数据在网络传输过程中是否损坏或丢包，它是由数字和英文字母组成的长度为 32 的定长字符串。

   - Linux 用户请使用 $ md5sum 命令计算；
   - Mac 用户请使用$ md5 命令计算；
   - Windows 用户请使用第三方工具进行计算，例如 winmd5free。

2) 当用户完成 Run 创建后，所对应的 GSA 将只可修改不可删除，如有特殊需求必须删除 GSA 的用户，请联系 GSA 工作组。

3) 用户完成 GSA 创建后，GSA、Experiment 和 Run 的 Status 状态为 Unchecked; Confidential，用户将原始数据通过 FTP 上传，经过质量审核通过并归档成功后，系统才会分配 GSA 的 Accession number。在数据审核归档期间，用户如果需要修改或删除数据信息，请联系 gsa@big.ac.cn。

数据文件上传

FTP 上传

请使用 FTP 客户端软件（比如 FileZilla Client）登录 FTP 服务器，用户账号与 BIG sub 账号一致。

FTP 服务器地址:

ftp://submit.big.ac.cn

注意：用户登录自己的 FTP 路径后，先 cd 到 /GSA 目录下再上传文件。请采用二进制模式上传，如果是用 FTP 软件上传，请参考软件说明进行设置；如果是用 FTP 命令上传，请在 put 命令前，先运行 binary 命令。

此外，数据上传完毕后，GSA 后台系统需要进行相应的审核，请耐心等待并密切关注系统和邮箱的情况反馈。

协助上传

GSA 充分考虑到大体量数据递交用户的需求，开启了硬盘寄送和协助上传的绿色通道。请您联系 GSA 工作组，邮箱: gsa@big.ac.cn；QQ 工作群: 548170081，填写批量提交表格并将把数据硬盘寄送到 GSA。通信地址：北京市朝阳区北辰西路 1 号院 104 号楼；联系电话：+86（01）84097340。

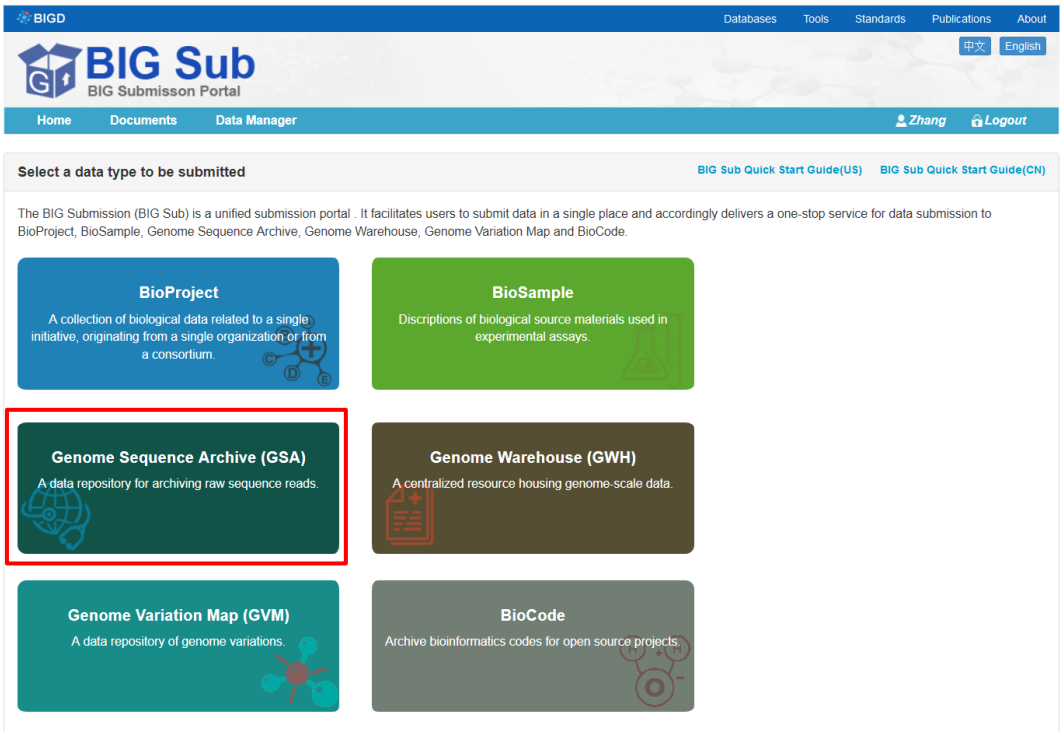
同网大型机协助拷取

如果数据存储在基因组所大型机上，可以联系 GSA 工作人员协助拷取。

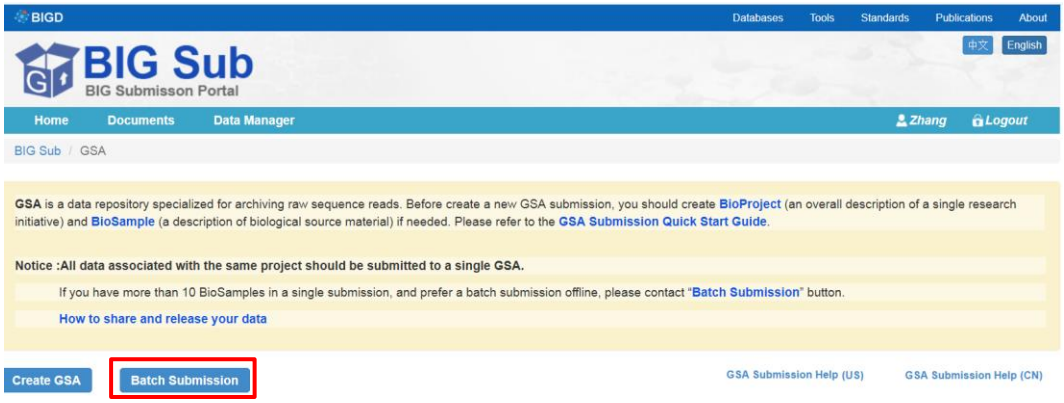
数据信息批量上传说明

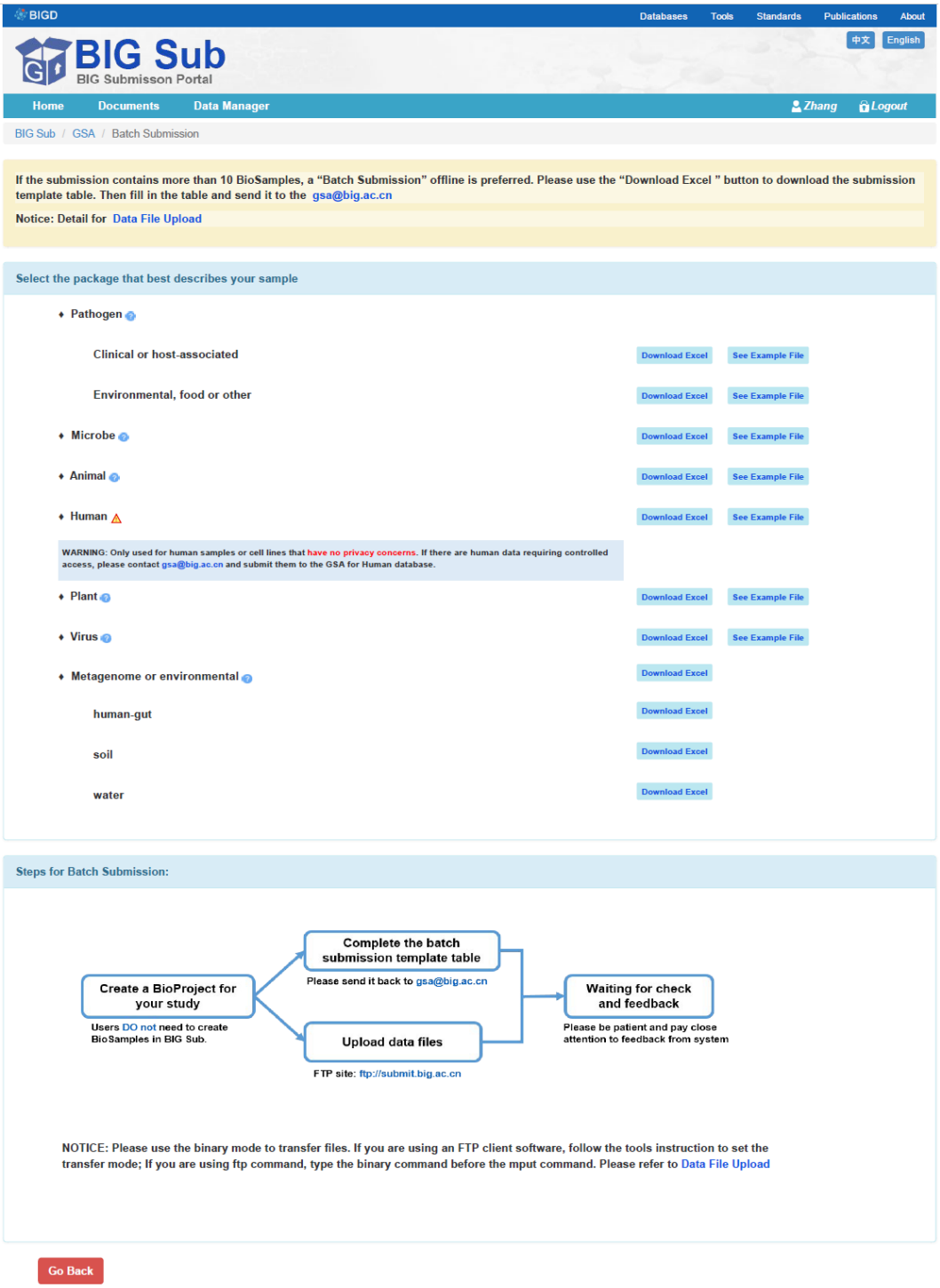
当数据样本（Sample）数量≥10 个，在完成 BioProject 创建后，建议您使用“离线模式”进行元数据信息批量录入，具体步骤如下：

1) 通过 BIG Sub 数据统一汇交入口，进入 GSA 数据库：



2) 点击“**Batch Submission**”进入“批量上传表格”下载页面，请根据提示信息下载对应的表格模板和例子，每个“批量提交表格”内包括四张表，分别为 Sample，GSA，Experiment 和 Run，填好后请发送至 gsa@big.ac.cn。

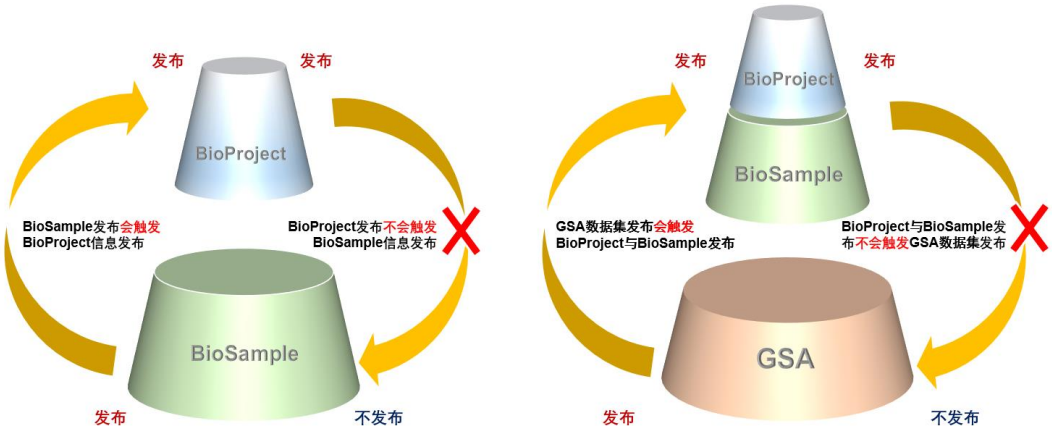



BIGD
Databases
Tools
Standards
Publications
About
中文
English
BIG Sub
BIG Submisson Portal
Home
Documents
Data Manager
Zhang
Logout
BIG Sub / GSA / Batch Submission
If the submission contains more than 10 BioSamples, a "Batch Submission" offline is preferred. Please use the "Download Excel " button to download the submission template table. Then fill in the table and send it to the gsa@big.ac.cn
Notice: Detail for Data File Upload
Select the package that best describes your sample
Pathogen
Clinical or host-associated
Environmental, food or other
Microbe
Animal
Human
WARNING: Only used for human samples or cell lines that have no privacy concerns. If there are human data requiring controlled access, please contact gsa@big.ac.cn and submit them to the GSA for Human database.
Plant
Virus
Metagenome or environmental
human-gut
soil
water
Download Excel
See Example File
Steps for Batch Submission:
Create a BioProject for your study
Users DO not need to create BioSamples in BIG Sub.
Complete the batch submission template table
Please send it back to gsa@big.ac.cn
Upload data files
FTP site: ftp://submit.big.ac.cn
Waiting for check and feedback
Please be patient and pay close attention to feedback from system
NOTICE: Please use the binary mode to transfer files. If you are using an FTP client software, follow the tools instruction to set the transfer mode; If you are using ftp command, type the binary command before the mput command. Please refer to Data File Upload
Go Back

数据触发机制说明

数据发布时，相关的 BioProject、BioSample 与 GSA 数据集遵循以下触发机制（如下图所示）：

1. BioProject 发布不会触发相关联 BioSample 信息与 GSA 数据集释放；
2. GSA 数据集发布，会触发相关联 BioProject 和 BioSample 信息释放。



数据发布触发规则

因此，请慎重填写 BioProject、BioSample 与 GSA “发布时间”，一旦发布就代表数据或信息可供其他用户公开检索或下载。