

组学原始数据归档库（GSA）使用说明

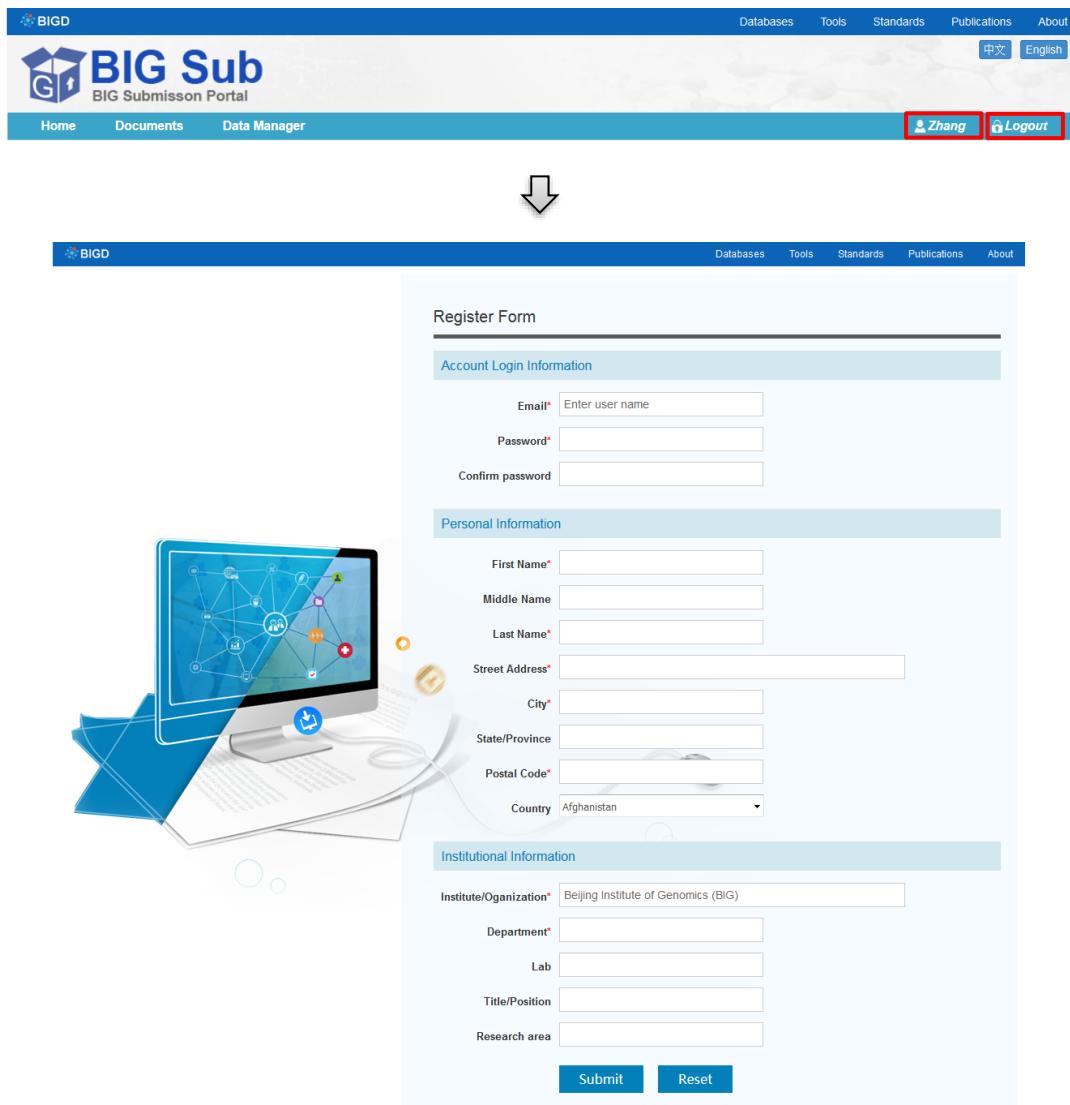
系统简介	2
用户注册	2
GSA 数据集创建	3
GSA 数据集修改、删除和追加	12
GSA 数据集发布	14
GSA 数据集分享链接生成	15
数据文件上传	16
Aspera 命令行上传（推荐）	16
FTP 上传	18
协助上传	25
数据触发机制说明	26
提交状态与操作说明	27

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 遵循 INSDC 数据库系统的数据标准和数据结构，主要汇交实验信息（Experiment Metadata）、测序反应信息（Run Metadata）信息以及归档测序文件数据（Sequence Data file）。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, BIG Sub）完成一站式数据递交。

用户注册

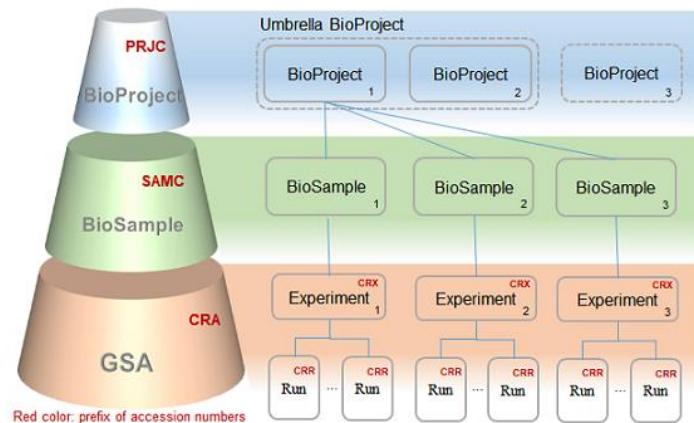
请您进入生物数据递交系统（BIG Submission, BIG Sub, <https://bigd.big.ac.cn/gsub/>）完成账号注册，建议使用实验室公共邮箱进行注册。如果您在账号注册和使用过程中遇到任何问题，请联系 bigd-admin@big.ac.cn。



The screenshot shows the 'Register Form' page of the BIG Sub system. At the top, there is a navigation bar with links for Databases, Tools, Standards, Publications, and About, along with language options for Chinese and English. Below the navigation bar, there are buttons for Home, Documents, and Data Manager, and user account buttons for Zhang and Logout. A large blue downward arrow points from the top navigation area to the registration form below. The registration form itself has three main sections: 'Account Login Information', 'Personal Information', and 'Institutional Information'. The 'Account Login Information' section contains fields for Email*, Password*, and Confirm password. The 'Personal Information' section contains fields for First Name*, Middle Name, Last Name*, Street Address*, City*, State/Province, Postal Code*, and Country (with Afghanistan selected). The 'Institutional Information' section contains fields for Institute/Organization* (Beijing Institute of Genomics (BIG)), Department*, Lab, Title/Position, and Research area. At the bottom of the form are 'Submit' and 'Reset' buttons.

GSA 数据集创建

为保证元数据信息与测序数据文件的一致性与完整性,便于后续数据使用者检索与使用,通过 BIG Sub 统一入口递交 GSA 数据信息时,用户需要为 GSA 数据集的研究任务创建 BioProject, 并为数据集的实验样本创建相应的 BioSample(s)。GSA 各类数据信息间是线性的、一对多的关联关系, 数据结构如下图。



具体提交操作步骤如下:

Screenshot of the BIG Sub submission portal:

- Header:** BIGD Databases Tools Standards Publications About Sign in to BIGD 中文 English
- Section:** BIG Sub BIG Submission Portal Home Documents Data Manager Welcome, Zhang~
- Text:** 请选择您要提交的数据类型
- Text:** BIG Sub是大数据中心生物数据统一汇交入口, 目前集成了BioProject、BioSample、GSA、GWH、GVM、BioCode等数据库的提交接口。
- Buttons:**
 - 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 Sequence Archive (GSA):** 组学原始数据汇交、存储、管理与共享系统
 - Genome Warehouse (GWH):** 面向多物种全基因组数据的归档数据库
 - 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
- Red Box:** A red box highlights the "Genome Sequence Archive (GSA)" button.
- Arrow:** A large downward arrow is positioned below the GSA section.

- **提交者信息 (Submitter)** — 用于收集数据提交者信息，系统会帮您自动填入用户注册时的姓名和电子邮件信息，如部分信息需要调整，可直接修改并通过“**保存并进入下一项 (Save and forward)**”键完成修改。

注： 数据信息审核与文件归档过程中出现任何问题，信息将反馈到您的注册邮箱，而非此处填入的提交者信息邮箱。

- **基本信息 (General)** — 用于收集 GSA 数据集的描述信息，包括发布日期 (Release date)、标题和描述信息 (Description)、项目信息 (BioProject accession)、样本信息 (Sample Information)。

注： 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”，根据系统提示依次完成下文中“元数据信息”和“文件上传”步骤，最终检查无误后完成提交。

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”，依照以下流程完成提交：

BIG Sub / GSA / Submission: subCRA000633

01 提交者信息 02 基本信息 03 元数据信息 04 文件上传 05 概况信息

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
2019-06-25

Release Date 的设置时间，用户可根据项目需求进行设定，但最长不要超过 2 年。

* 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. “发布日期”可以在 GSA 提交系统内进行修改 ([http://bigd.big.ac.cn/gsub/submit/gsa/\[substitute your GSA accession number\]/contents](http://bigd.big.ac.cn/gsub/submit/gsa/[substitute your GSA accession number]/contents))
3. 如果引用这些数据和该 accession 号的文章先于您设定的发布时间而发表，我们将根据文章的发表时间来发布该数据；否则 GSA 将根据您设定的发布日期而发布该数据；
4. 一旦文章发表，数据可以发布，请把已发表文章的全部信息—作者，题目，期刊，刊号，页数，日期信息发送到该邮箱：gsa@big.ac.cn

I accept it. I don't accept it.

发布策略和免责声明

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

如果您已经创建了 BioProject，请选出对应 Accession 号；
如果您还未创建 BioProject，请点击并前往创建 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

此处以未创建 GSA 相关的 BioSample 信息为例

保存并进入下一页

- 样本类型 (Sample Type) — 用于收集有关样本类型信息。

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 概况信息

样本类型

- Pathogen 用于与公共卫生相关的病原体样本
- Clinical or host-associated
- Environmental, food or other
- Microbe 用于能具体写出物种名称的微生物样本，但不包括致病菌或病毒样本，宏基因组数据建议选择的Metagenome/Environmental Sample (GSC MIMS unsupported)或Metagenome/Environmental Sample (GSC MIMS compliant)
- Animal 用于模式生物或动物的多细胞样本或细胞系样本，如大鼠、小鼠、果蝇、线虫、鱼类、两栖类或其他哺乳动物数据
- Human 人类遗传资源相关法律法规提交到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) 目前用于不适合Metagenome/Environmental Sample (GSC MIMS compliant) 的宏基因组生物样本数据
- Metagenome/Environmental Sample (GSC MIMS compliant) 用于宏基因组生物样本数据，下设三个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

以 Plant 为例

保存并进入下一项

注：遵从《中华人民共和国人类遗传资源管理条例》总则规定，如果您确定需要将数据提交到 [GSA-Human 数据库](#)。

• 样本属性 (Attributes) — 用于批量提交样本的属性信息。

- 1) 下载模板文件，如上图中的 [Plant.cn.xlsx \(中文版\)](#), e.g. [Plant.cn.xlsx](#) 为例子文档。更多帮助，请查看 [Help](#);

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 概况信息

植物样本属性信息

* 上传 BioSample 批量提交文件

请选择文件

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

保存并进入下一项

- 2) 编辑模板文件并检查无误后，通过文件选择框进行文件上传；

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 概况信息

植物样本属性信息

* 上传 BioSample 批量提交文件

请选择文件

上传

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

保存并进入下一项

3) 上传完成后, 通过点击“校验”键, 进行批量表格在线审核:

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 概况信息

植物样本属性信息

* 已上傳的 BioSample 批量提交文件

[Plant.xlsx](#) 24KB 删除 校验

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

保存并进入下一项

4) 若文件审核不通过, 请点击“删除”键, 删除已上传的文件并按系统提示信息修改后, 再重新上传批量表格文件, 直至审核通过; 当文件审核通过, 请点击“保存并进入下一项 (Save and forward) ”键, 完成 BioSample 批量提交;

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 概况信息

植物样本属性信息

* 已上傳的 BioSample 批量提交文件

[Plant.cn.xlsx](#) 24KB 删除 校验 ✖ ERROR:

下载错误文件: [error.txt](#)

```
Error in Sample sheet
row 11, column 7: "biomaterial_provider" is empty
row 12, column 7: "biomaterial_provider" is empty
row 13, column 7: "biomaterial_provider" is empty
row 14, column 7: "biomaterial_provider" is empty
```

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

保存并进入下一项



BIG Sub / GSA / Submission: subCRA003595

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 概况信息

植物样本属性信息

* 已上传的BioSample批量提交文件
Plant.xlsx 24KB [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。

[保存并进入下一项](#)



• **元数据信息 (Metadata)** — 用于批量提交 GSA 元数据信息，具体批量提交步骤如下；

- 1) 下载模板文件，如上图中的 [GSA_Template.cn.xlsx \(中文版\)](#)，
e.g. [GSA_Template.cn.xlsx](#) 为例子文档。更多帮助，请查看 [Help](#)；

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 文件上传 07 概况信息

GSA元数据信息

* 请上传GSA批量提交文件
 [上传](#)

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

[保存并进入下一项](#)



- 2) 编辑模板文件并检查无误后，通过文件选择框进行文件上传；

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 文件上传 07 概况信息

GSA元数据信息

* 请上传GSA批量提交文件
 [上传](#)

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

[保存并进入下一项](#)



- 3) 上传完成后，通过点击“校验”键，进行批量表格在线审核；

- 4) 若文件审核不通过，请点击“**删除**”键，删除已上传的文件并系统按提示信息修改后，再重新上传批量表格文件，直至审核通过；当文件审核通过，请点击“**保存并进入下一项（Save and forward）**”键，完成批量样本提交；

- **文件上传 (File Upload)** — 数据文件上传方式选择，分别为 FTP 客户端，Aspera 命令行（推荐）和 Aspera Connect 浏览器插件上传，详见“[数据文件上传](#)”。

- **概况信息(Overview)**—提供对 GSA 数据及其相关信息的整体预览。在正式提交之前，用户可通过点击进度条上的按钮，进入相应页面修改信息。请务必检查无误后再点击“**提交 (Submit)**”完成递交。

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 概况信息

GSA基本信息

提交编号	subCRA003595
标题	Full-length transcriptome sequencing from multiple tissues of annual ryegrass
描述信息	Full-length transcriptome sequencing of annual ryegrass (<i>Lolium multiflorum</i> Lam.)
发布日期	2020-07-29

提交者信息

提交者	Zhang Sisi zhangss@big.ac.cn
单位	Beijing Institute of Genomics, Chinese Academy of Sciences
部门	BIG Data Center
国家/地区	China
地址	Beichen West Road Beijing
邮编	100101

元数据信息

样本类型	Plant
BioSample数据文件	Plant.xlsx
元数据文件	plant_ExR.xlsx

实验名称	测序平台	样本名称	生物名称	实验选择
CK1	PacBio Sequel	CK1	<i>Lolium multiflorum</i>	PCR
Run别名	Run序列文件处理信息			文件类型
CK1	File: m54211_180613_064306.subreads.bam			bam
CK2	PacBio Sequel	CK2	<i>Lolium multiflorum</i>	PCR
Run别名	Run序列文件处理信息			文件类型
CK2	File: m54212_180609_084342.subreads.bam			bam
DK1	PacBio Sequel	DK1	<i>Lolium multiflorum</i>	PCR
Run别名	Run序列文件处理信息			文件类型
DK1	File: m54213_180607_091704.subreads.bam			bam
DK2	PacBio Sequel	DK2	<i>Lolium multiflorum</i>	PCR
Run别名	Run序列文件处理信息			文件类型
DK2	File: m54214_180609_073921.subreads.bam			bam

Total Items: 4 | Items of 1 - 4 | Page size | Page 1/1 | << First | Last >> | Jump To GO

提交



新建 GSA

GSA编号	提交编号	GSA标题	发布日期	提交状态	操作
Unassigned	subCRA003595	Full-length transcriptome sequencing from multiple tissues of annual ryegrass	2020-07-29	Unchecked Confidential	删除

通常状况下，数据信息与文件审核归档约需要 1-2 天（数据量越大相应所需时间越长），归档成功后您会收到一封通知邮件，并可在 GSA 列表中查找的为您分配的 GSA 编

号 (GSA Accession number); 如果归档中数据信息与文件审核归档过程中出现问题，信息将反馈到您的[注册邮箱](#)，因此请您关注邮箱反馈信息。

新建 GSA						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		
CRA002926	subCRA00359	Full-length transcriptome sequencing from multiple tissues of annual ryegrass	2020-07-29	Checked OK Confidential		删除	

注：

- 1) GSA 提交编号 (**Submission ID**)：sub#，如上图中的 subCRA003595。请仅在联系 GSA 工作人员时使用，不要在 [BIG Search](#) 检索信息时或在文章中使用提交编号。
- 2) 请务必在 [BIG Search](#) 检索信息时或在文章中使用 GSA 编号 (GSA Accession Number)：CRA#，如上图中的 CRA002926。

GSA 数据集修改、删除和追加

在 GSA 数据集文件归档完成之前，无论数据信息是否通过审核，用户可通过点击“Submission ID”进入样本总览界面，①更新 GSA 基本信息（Basic Information）中的标题（Title）和发布日期（Release date）；②修改提交者信息（Submitter information）；③使用“追加数据（Add Data）”键，详见“[GSA 数据集创建](#)”；④修改或删除已提交实验（Experiment）和测序反应（Run）基本信息；⑤使用“更新文件（Update File）”键，补充和更新数据文件（推荐 Aspera Connect 浏览器插件上传文件用户使用）。

注： 数据详细提交状态和用户可用操作详见“[提交状态与操作说明](#)”。

The screenshot displays the GSA submission interface with three main sections:

- GSA基本信息**: Shows basic submission details like subCRA000633, Release Date: 2020-12-31, Project: PRJCA000644, and a red box highlights the "查看 BioProject 信息" button. A red box also surrounds the "标题" (Human dataset) and "发布日期" (2020-12-31) input fields, with a red arrow pointing to the "更新" (Update) button labeled "更新 GSA 基本信息中的标题和发布日期".
- 提交者信息**: Displays the submitter's details (Zhang Sisi, zhangss@big.ac.cn, Beijing Institute of Genomics, Chinese Academy of Sciences, etc.) and a red box highlights the "修改" (Modify) button labeled "修改提交者信息".
- 元数据信息**: Shows experimental runs and samples. It includes a red box around the "BioSample 提交编号" (subSAM023445), a red arrow pointing to the "查看 BioSample 信息" button, and red arrows pointing to the "更新文件" (Update File) and "追加数据" (Add Data) buttons. The table lists runs and samples with status indicators like "Unchecked Confidential" or "Checked OK Confidential".

At the bottom, there are navigation links for "Total Items: 6 | Items of 1 - 6 | Page size: 20 | Page 1/1 | << First | Last >> | Jump To: 1 | GO".

在 GSA 数据集文件归档完成之后 (Status 为 check OK; confidential)。用户可通过点击“Submission ID”进入样本总览界面，①更新 GSA 基本信息（Basic Information）中的标题（Title）和发布日期（Release date）；②修改提交者信息（Submitter information）；③使用“追加数据（Add Data）”键，详见“[GSA 数据集创建](#)”；④使用“更新文件（Update File）”键，补充和更新数据文件（推荐 Aspera Connect 浏览器插件上传文件用户使用）。如果您还希望修改或删除已提交实验（Experiment）和测序反应（Run）基本信息，请通过gsa@big.ac.cn 邮箱联系数据库工作组。

注： 数据详细提交状态和用户可用操作详见“[提交状态与操作说明](#)”。

The screenshot shows the BIG Sub system's GSA management interface. At the top, there is a table for viewing GSAs:

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	立即发布 分享

Below this is a large downward arrow pointing to the sample overview section:

BIG Sub / GSA / Submission: subCRA000595

GSA基本信息

GSA 提交信息: subCRA000595 / Release Date: 2020-12-31 / Project: PRJCA000644

状态	Data Archived
标题	1
发布日期	2020-12-31
更新	

更新 GSA 基本信息中的标题和发布日期

修改提交者信息

提交者信息

提交者	Zhang Sisi zhangss@big.ac.cn
单位	Beijing Institute of Genomics, Chinese Academy of Sciences
部门	BIG Data Center
国家/地区	China
地址	Beichen West Road Beijing
邮编	100101

修改提交者信息

元数据信息

样本类型: Human sample
元数据文件: [5.Human.cn.test4.xlsx](#) added on 2019-06-13
[exp_run-Human.cn.test4.xlsx](#) added on 2019-06-17

进入文件上传页面，补充或重新上传文件 批量添加元数据信息

更新文件 追加数据

实验编号	实验名称	物种名称	测序平台	文库布局	样本编号: 样本名称	实验提交状态	操作
CRX020324	scRNA-seq of WJMSC21	Homo sapiens	Illumina HiSeq 2500	FRAGMENT	SAMC018351: scRNA-seq of WJMSC21	Processed Succeed Confidential	
Run编号	Run名称	Run序列文件处理信息					操作
CRR022201	scRNA-seq of WJMSC21	CS300_TGACCA_L002_R1_001.fastq.gz				Status: Processed Succeed	
CRX020325	scRNA-seq of WJMSC22	Homo sapiens	Illumina HiSeq 2500	FRAGMENT	SAMC018352: scRNA-seq of WJMSC22	Processed Succeed Confidential	
Run编号	Run名称	Run序列文件处理信息					操作
CRR022202	scRNA-seq of WJMSC22	CS300_TGACCA_L002_R1_002.fastq.gz				Status: Processed Succeed	
CRX020326	scRNA-seq of WJMSC23	Homo sapiens	Illumina HiSeq 2500	FRAGMENT	SAMC018353: scRNA-seq of WJMSC23	Processed Succeed Confidential	
Run编号	Run名称	Run序列文件处理信息					操作
CRR022203	scRNA-seq of WJMSC23	CS300_TGACCA_L002_R1_003.fastq.gz				Status: Processed Succeed	

Total Items: 6 | Items of 1 - 6 | Page size: 20 | Page 1 / 1 | << First | Last >> | Jump To: 1 | GO

GSA 数据集发布

如用户需要提前发布 GSA 数据集，可点击下图列表中“立即发布（Release Now）”控件，完成提前释放。

Create GSA						GSA Quick Start Guide (US)	GSA Quick Start Guide (CN)
Accession	Submission ID	Title	Release date	Status	Operation		
CRA002926	subCRA003595	Full-length transcriptome sequencing from multiple tissues of annual ryegrass	2020-07-29	Checked OK Confidential	ReleaseNow	Share	

在“释放数据确认框”中点击“Yes”，即可释放 GSA 数据集。请注意当 GSA 数据集发布后，所有与其关联的 BioProject 和 BioSample(s) 将同时发布（具体机制详见[数据触发机制说明](#)）。



注：GSA 数据释放后，需要几个小时归档数据，等数据归档成功后，即可在 [BIG Search](#) 中通过 GSA 序列号（Accession number）搜索到数据集以及相关 BioProject 和 BioSample(s) 信息。

GSA 数据集分享链接生成

1. 用户通过账号登陆 BIG Sub 系统，在 GSA 提交系统列表中，找到 Operation 有个“分享”控件（如图所示）；

GSA Quick Start Guide (US) GSA Quick Start Guide (CN)					
Accession	Submission ID	Title	Release date	Status	Operation
CRA002926	subCRA003595	Full-length transcriptome sequencing from multiple tissues of annual ryegrass	2020-07-29	Checked OK Confidential	ReleaseNow Share

2. 点击“分享”，会生成如下图所示的分享链接，复制该链接并提供给编审，其即可以查看数据；

GSA Quick Start Guide (US) GSA Quick Start Guide (CN)					
Accession	Submission ID	Title	Release date	Status	Operation
CRA002926	subCRA003595	Full-length transcriptome sequencing from multiple tissues of annual ryegrass	2020-07-29	Checked OK Confidential	ReleaseNow Shared URL: http://bigd.big.ac.cn/gsas/T31EvGtU Cancel share

注：此链接为临时链接，用户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。

数据文件上传

在文件上传页面，您有三种选择来上传页面，请选择其中一种方式来上传数据。选中上传方式后，不需要等待文件上传完成，可以直接进入下一项“概况信息(Overview)”查看上传的元数据情况并提交。

Aspera 命令行上传（推荐）

您可以通过 Aspera 命令行，使用以下的命令来上传文件：

[path/to/ascp/] -P33001 -i [path/to/key/file] -QT -I100m -k1 -d [path/to/folder/containing/files]

aspsub@submit.big.ac.cn:uploads/ [user dir]

The screenshot shows a step-by-step file upload process. Step 04, '文件' (Files), is highlighted with a red box. The interface includes a '提示' (Tip) section with instructions about file names and MD5 codes. It also shows a command line input field containing the Aspera command with placeholder values. A note in red text on the right side instructs users to copy the command from step 04 and replace paths before execution.

01 提交者信息 02 基本信息 03 元数据信息 04 文件 05 概览 & 提交

文件上传方式

提示：
上传的文件的文件名和MD5码必须跟您在GSA Metadata表格里填写的一致。
上传的文件名必须是唯一的。
fastq文件必须压缩后上传，目前我们只接收gzip或bzip2格式的压缩文件。

* 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的账号一样。

Aspera Command Line
使用Aspera命令行上传文件。使用Aspera命令行前，需要先安装Aspera Connect插件。如果您没有安装Aspera Connect插件，请先[点击此链接下载安装](#)。

您可以通过Aspera命令行，使用以下的命令来上传文件：
[path/to/ascp/] -P33001 -i [path/to/key/file] -QT -I100m -k1 -d [path/to/folder/containing/files]
aspsub@submit.big.ac.cn:uploads/ @qq.com_2bf1fd3b

注意：每个用户的 [user dir]，即用户目录都不同，命令行中一定要用户目录信息（User Directory）。信息查询方式：点击“提交编号”（Submission ID）进入“概况信息”（Overview）页面，点击“更新文件”（Update file）按钮，进入“04 文件”（04 Files）文件上传页面，即可看到自己的 Aspera 命令行信息

Web browser upload via Aspera Connect plugin
使用Aspera Connect浏览器插件上传文件。如果您没有安装Aspera Connect插件，请先[点击此链接下载安装](#)。

其中：

[path/to/ascp/]: 指 ascp 的执行程序，一般安装了 aspera connect plugin 的操作系统，都有这个执行程序。不同的操作系统，ascp 存在于不同的位置。

Microsoft Windows: C:\Program Files\Aspera\Aspera Connect\bin\ascp.exe

或：C:\users\[username]\AppData\Local\Programs\Aspera\Aspera Connect\bin\ascp.exe

Mac OS X: /Applications/Aspera/Connect.app/Contents/Resources/ascp (admin 用户安装)

或: /Users/[username]/Applications/Aspera/Connect.app/Contents/Resources/ascp (非 admin 用户安装)

Linux: /opt/aspera/bin/ascp or /home/[username]/aspera/connect/bin/ascp

命令行中:

[path/to/key/file] 必须是文件的绝对路径, 如: /home/keys/aspera.openssh

[path/to/folder/containing/files] 应该是包含您要上传的所有文件的本地路径。

[user dir] 是用户目录 (User Directory), 查询方式: 点击“提交编号”(Submission ID) 进入“概况信息”(Overview) 页面, 点击“更新文件”(Update file) 按钮, 进入“04 文件”(04 Files) 文件上传页面, 即可看到自己的 Aspera 命令行信息

请点击“[Get the key file](#)”下载获取此文件。

注:

- 1) 请为您每一个提交创建一个新的子目录(可以用 GSA 的提交编号作为子目录名称)。请注意这个子目录是一个临时文件夹, 当本次提交完成时, 该目录及目录下的文件将被后台处理程序删除。
- 2) 请不要上传复杂的文件夹结构, 也不要上传跟您提交不相关的非序列文件。
- 3) 更新文件: 当元数据信息已经提交后, 无法通过导航条进入文件上传页面。这时, 如果需要重传或者追加数据时, 点击“提交编号”(Submission ID) 进入“概况信息”(Overview) 页面, 点击“更新文件”(Update file) 按钮, 进入“04 文件”(04 Files) 文件上传页面, 重新选择文件上传方式进行文件上传。

GSA 基本信息

GSA 提交信息: subCRA002561 / Release Date: 2020-01-17 / Project: PRJCA000651

状态	5 Runs are checked failed
标题	v1
发布日期	2020-01-17
更新	
联系我们	

提交者信息

提交者	Zhang Sisi zhangs@big.ac.cn
单位	Beijing Institute of Genomics, Chinese Academy of Sciences
部门	BIG Data Center
国家/地区	China
地址	Beichen West Road Beijing
邮编	100101
修改	

元数据信息

样本类型	Virus sample						
BioSample 提交编号	subSAM090716						
样本属性文件	subCRA002561_sample.csv						
元数据文件	subCRA002561_cra.xlsx						
更新文件 添加数据							
实验编号	实验名称	物种名称	测序平台	文库布局	样本编号: 样本名称	实验提交状态	操作
CRX091220	Lactobacillus plantarum L-1	Azorella acaulis	Illumina HiSeq X Ten	PAIRED	SAMC13135: type26	<i>Unchecked</i> <i>Confidential</i>	修改 删除



BIG Sub / GSA / Submission: subCRA002561

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 概况信息

文件上传方式

提示:
上传的文件的文件名和MD5码必须跟您在GSA Metadata表格里填写的一致。
上传的文件名必须是唯一的。
fastq文件必须压缩上传，目前我们只接收gzip或bzip2格式的压缩文件。
△本页面不是本次提交的最终页面。请在本页面选择文件上传方式后，点击页面左下方的“保存并进入下一步”按钮，进入“概况信息”页面，在“概况信息”页面检查、确认上传的元数据无误后，点击“保存”按钮，完成元数据的提交。

* 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的账号一样。
您可以通过客户端软件如[FileZilla](#)或命令行上传文件。
Address: <http://submit.big.ac.cn>
Username: Same as you login in the BIG Sub
Password: Same as you login in the BIG Sub
注意：请进入GSA目录并把文件上传到该目录下。请不要把文件上传到根目录下，这样后台处理程序将扫描不到您的文件。
注意：请使用二进制模式上传文件。如果您使用FTP客户端软件，请参考软件的说明文档来设置传输模式；如果您使用FTP命令行上传文件，请在输入命令之前，先运行binary命令，来设置二进制传输模式。

Aspera Command Line
使用Aspera命令行上传文件。

Web browser upload via Aspera Connect plugin
使用Aspera Connect浏览器插件上传文件。如果您没有安装Aspera Connect插件，请先[点击此链接下载安装](#)。

将入后台上传文件

[保存并进入下一项](#)

FTP 上传

• FTP 客户端方式上传数据

用户需要使用 FTP 客户端软件（比如 [FileZilla Client](#)）登录 FTP 服务器上传数据，

文档中以 FileZilla 作为实例。

1) 第 1 步，下载客户端软件 (<https://filezilla-project.org/>)，下载页面如图 1 所示，点击红色框中的“Download FileZilla Client”，并按照提示安装软件；

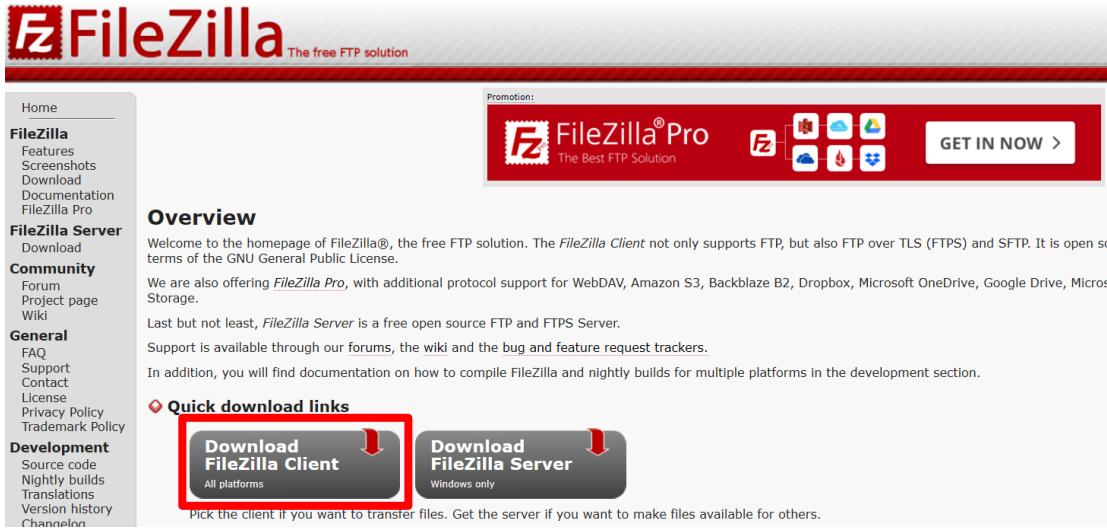


图 1 FileZilla Client 软件下载

2) 第 2 步，打开软件，界面如图 2 所示，填写主机信息为“submit.big.ac.cn”，用户名和密码填写 GSA 数据库的登陆帐号邮箱和密码，然后点击“快速连接”，状态栏显示登陆成功，如果提示错误，请根据提示信息查看错误原因；

3) 第 3 步，登陆成功后，“本地站点”选择需要上传数据的本地数据路径，“远程站点”中，双击 GSA 文件夹，进入 GSA 目录。

4) 第 4 步，在“本地站点”中选择上传的数据文件或者文件夹，点击右键，选择“上传”，或者直接拖拽到“远程站点”，如图 3 所示。

5) 第 5 步，上传的所有数据会进入“队列的文件”，排队上传，上传成功后数据信息会转移到“成功的传输”中，如果上传不成功会转移到“传输失败”，需要重新上传，可以选择“断点续传”。

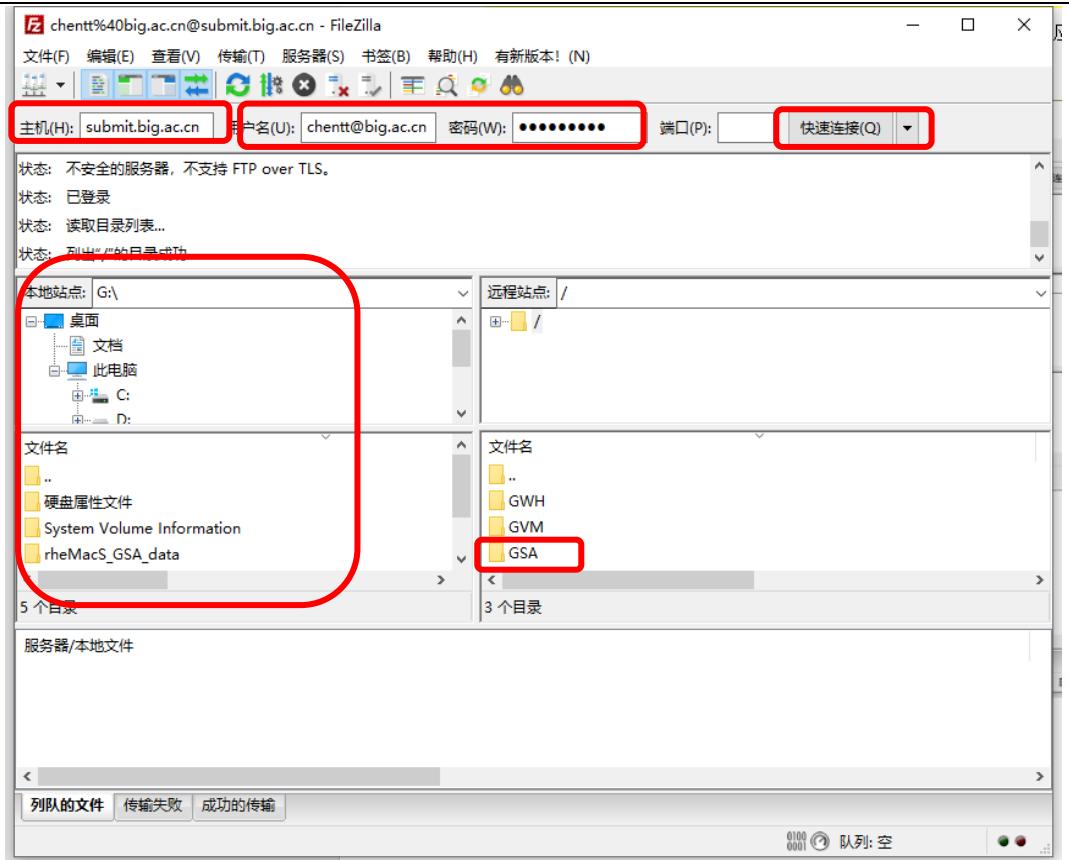


图 2 FileZilla 客户端界面

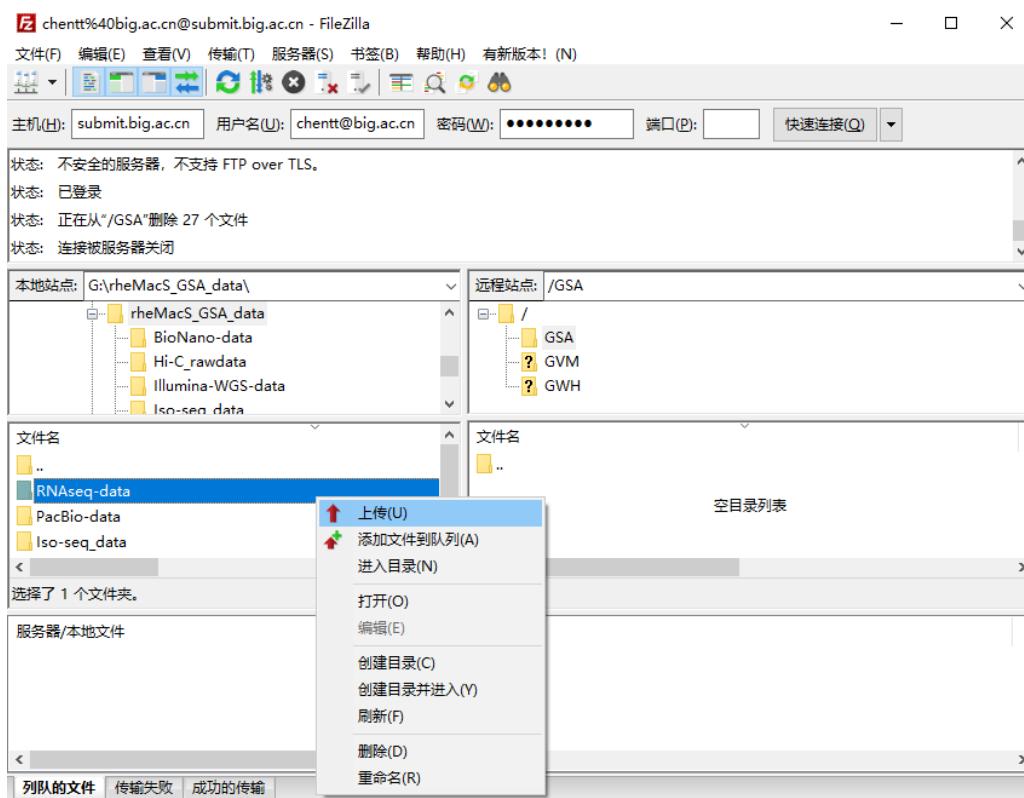


图 3 FileZilla 客户端上传界面

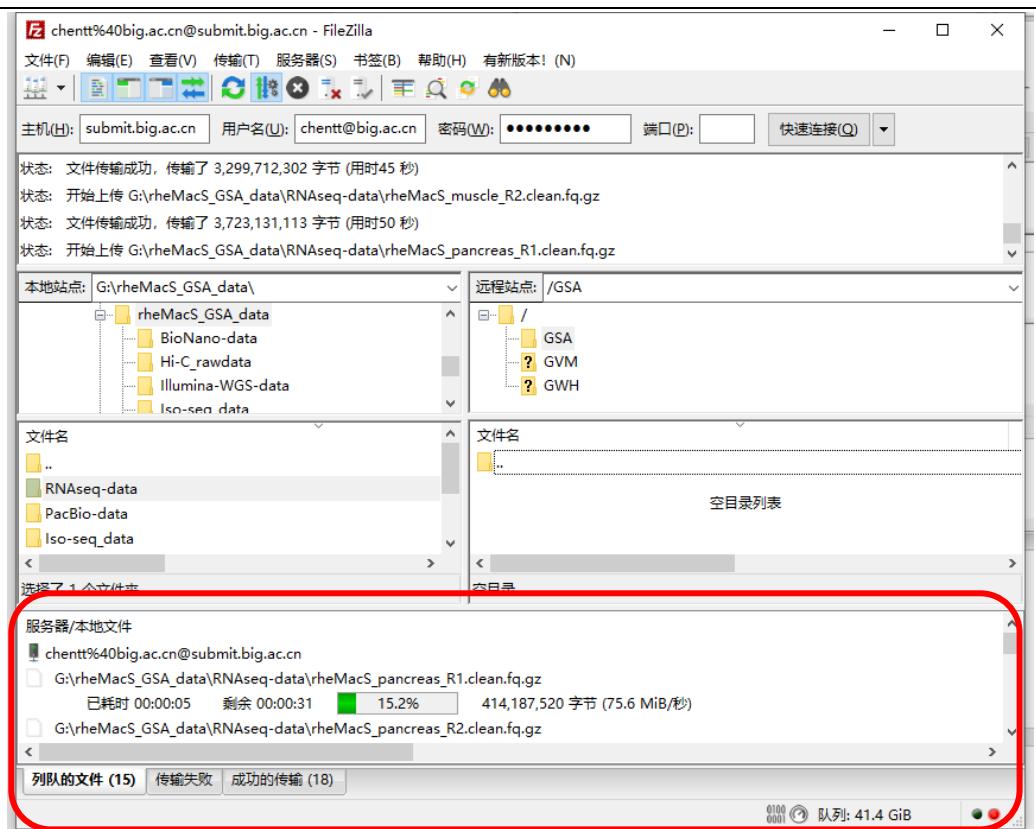


图 4 数据传输状态

• FTP 命令行上传数据

ftp command: 下划线中为需要输入的指令

```

>ftp submit.big.ac.cn
Name: your GSA account
331 User name okay, need password for (your email)
Password: *****
230 User logged in, proceed.
Remote system type is UNIX.
ftp> cd GSA
250 Directory changed to /GSA
ftp> binary
200 Command TYPE okay.
ftp> prompt
Interactive mode off.
ftp> mput *

```

上传成功
交互页面：

```
local: 16294.err remote: 16294.err
227 Entering Passive Mode (192,168,118,121,213,234)
150 File status okay; about to open data connection.
226 Transfer complete.
124 bytes sent in 0.03 seconds (4 Kbytes/s)
local: 16294.finish remote: 16294.finish
227 Entering Passive Mode (192,168,118,121,207,65)
150 File status okay; about to open data connection.
226 Transfer complete.
```

- 可能遇到的问题

问题 1：FTP 登陆时，出现如图 5 所示，状态栏出现 AUTH SSL 的报错信息。

解决方案：如图 6 所示点击菜单栏“文件”中的“站点管理器”，修改“加密”选项为“只使用普通 FTP”或者“”，同时，填写正确的主机地址：submit.big.a.cn，帐号和密码信息。最后点击“连接”即可。



图 5 Filezilla 报错信息

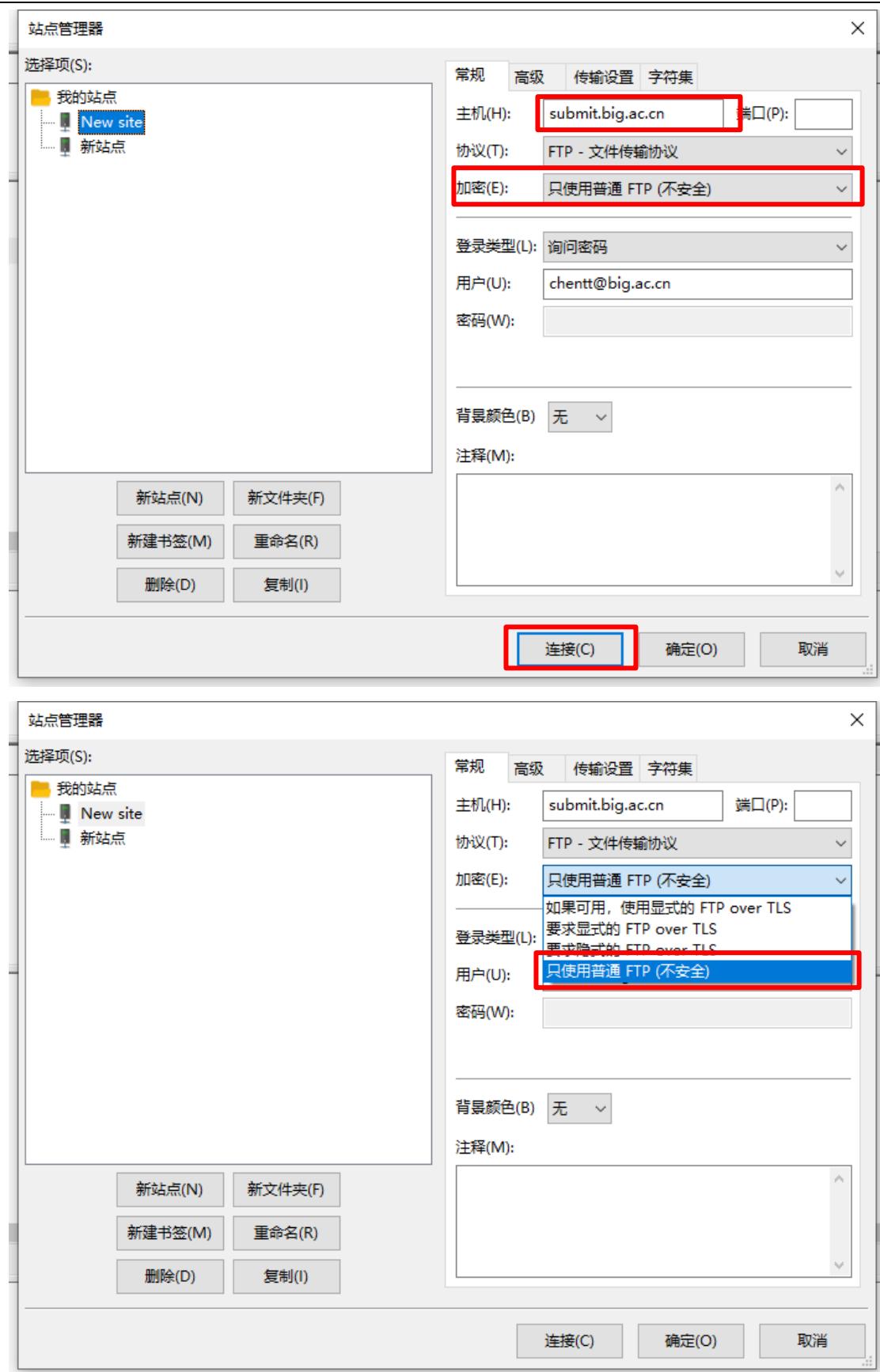


图 6 站点管理设置

问题 2: FTP 登陆时, 出现如图 7 所示, 状态栏出现 MLSD 的报错 (如图 7 所示), 显示“读取目录列表失败”。

解决方案: Filezilla → 编辑->设置中修改传输模式, 改为被动模式 (如图 8 所示)。

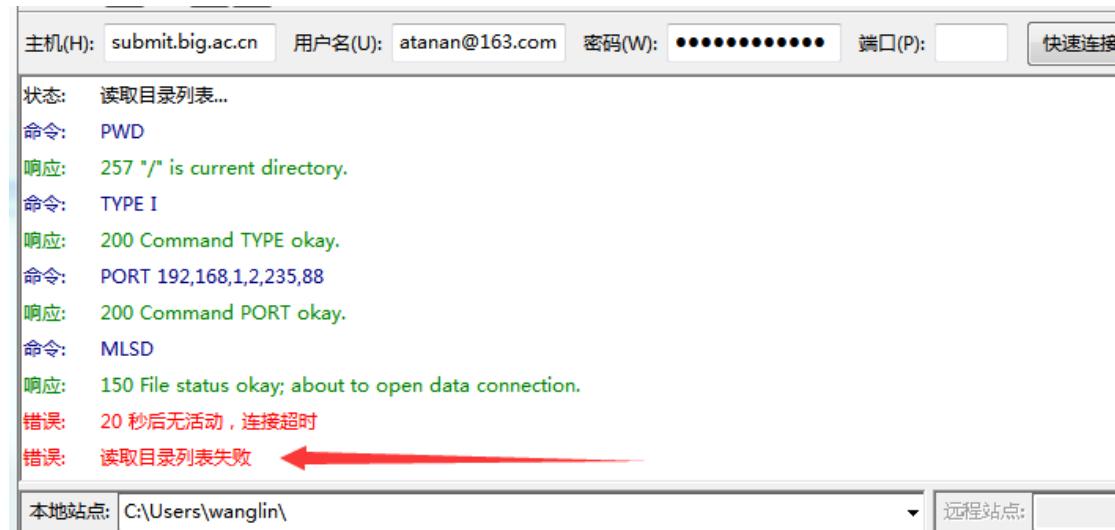
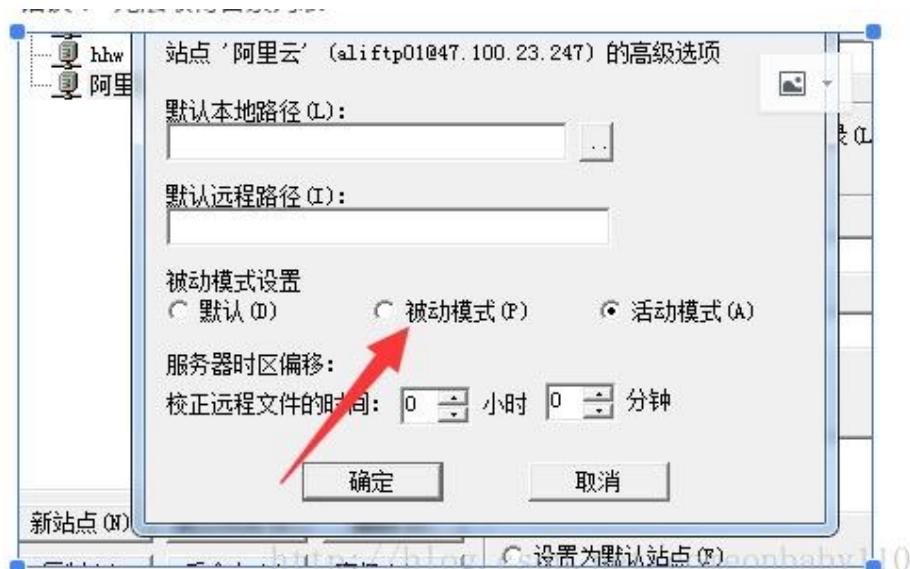


图 7 Filezilla 报错信息



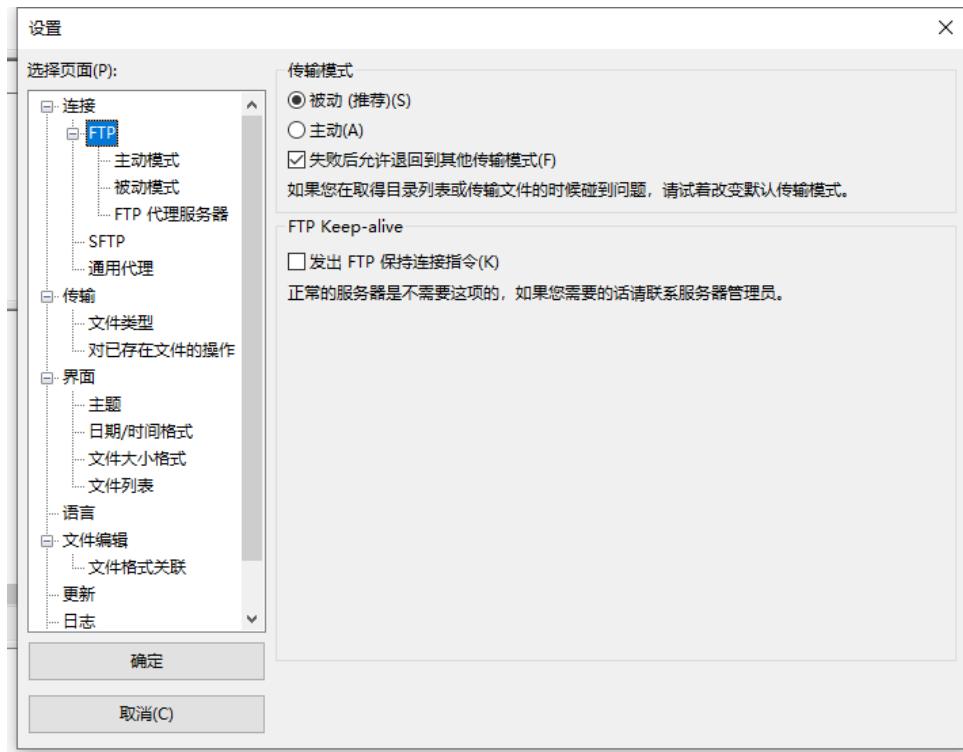


图 8 Filezilla 传输模式修改

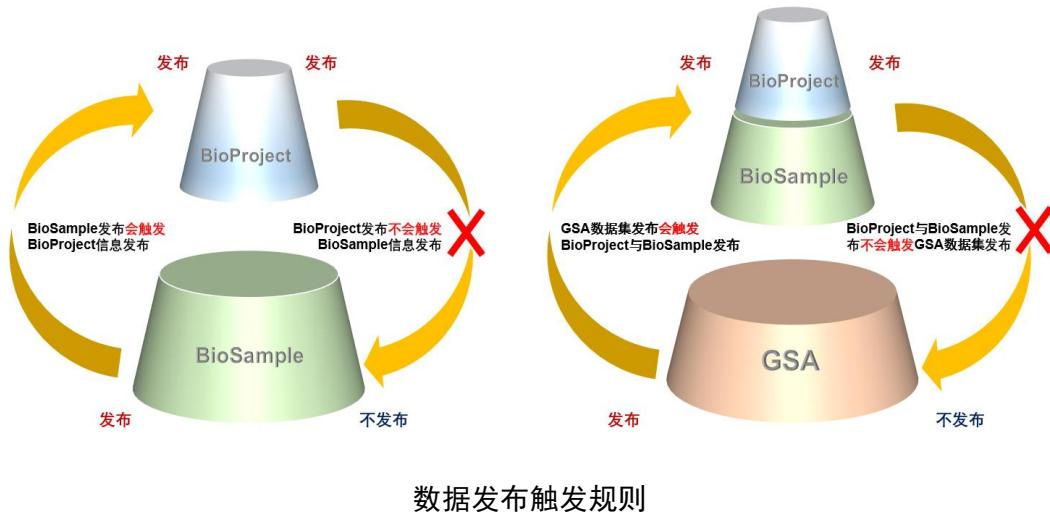
协助上传

GSA 充分考虑到大体量数据递交用户的需求, 为一次性上传数据量大于 **1TB** 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, 填写“**PRJCA[请写上编号]-硬盘填写信息文档**”, 电子版发送至工作组邮箱, 并打印纸质版随数据硬盘寄送到 GSA。

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。



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和下载**。

提交状态与操作说明

GSA 提交状态及可用操作：

新建 GSA				提交状态	操作
Unassigned	subCRA000654	human	2019-08-06	<i>Unfinished at the OverView step Confidential</i>	删除
Unassigned	subCRA000653	Human dataset 1	2020-12-31	<i>Unchecked Confidential</i>	删除
Unassigned	subCRA000652	microbe dataset	2019-08-05	<i>Unfinished at the Attributes step Confidential</i>	删除
CRA000532	subCRA000595	Human dataset test	2020-12-31	<i>Checked OK Confidential</i>	立即发布 分享
Unassigned	subCRA000645	Human dataset 1	2020-12-31	<i>Deleted Confidential</i>	

系统中 GSA 的提交状态共有 10 种，具体情况详见下表：

序号	提交状态	说明	可用操作
1	Unfinished at the General Info step	完成 Submitter 信息提交，进入 general info 步骤	修改 ^[1] 、删除
2	Unfinished at the Sample Type step	完成 General info 信息提交, 如选择“未建立 sample 信息”，进入 Sample type 步骤	修改 ^[1] 、删除
3	Unfinished at the Attributes step	完成 Sample type 信息提交，进入 Attributes 步骤	修改 ^[1] 、删除
4	Unfinished at the Metadata step	完成 Attributes 信息提交，进入 GSA metadata 步骤	修改 ^[1] 、删除
5	Unfinished at the File Upload step	完成 GSA metadata 信息提交，进入文件上传步骤	修改 ^[1] 、删除
6	Unfinished at the Overview step	完成所有信息填写，进入 overview 步骤	修改 ^[1] 、删除
7	Unchecked	完成所有信息填写并已提交	修改 ^[1] 、删除
8	Checked failed	数据文件处理失败	修改 ^[1] 、删除
9	Checked OK	数据处理完成，归档成功。用户提交页面显示 Accession 编号	立即发布、生成分享链接
10	Deleted	已删除	

^[1]: 用户可通过点击 “Submission ID” 进入样本总览界面修改 GSA 元数据信息，详细修改原则详见 “[GSA 数据集修改、删除和追加](#)”。

Experiment 提交状态及可用操作：

实验编号	实验名称	物种名称	测序平台	文库布局	样本编号: 样本名称	实验提交状态	操作
CRX022949	scRNA-seq of WJMSC1	Homo sapiens	Illumina HiSeq 2500	FRAGMENT	SAMC021054: scRNA-seq of WJMSC1	Unchecked Confidential	<button>修改</button> <button>删除</button>
Run编号	Run名称	Run序列文件处理信息					
CRR022922	scRNA-seq of WJMSC1	CS300_TGACCA_L002_R1_001.fastq.gz				Status: Unchecked	<button>修改</button> <button>删除</button>

系统中 Experiment 的提交状态共有 4 种，具体情况详见下表：

序号	提交状态	说明	可用操作
1	Unchecked	Experiment 信息已提交，等待后台审核	修改 ^[1] 、删除
2	Checked OK	Experiment 信息审核通过	修改 ^[1] 、删除(如果其关联的 Run 还没有进入后台处理流程)
3	Checked failed	Experiment 信息审核未通过	修改 ^[1] 、删除
4	Deleted	已删除	无

^[1]: 用户可通过点击“Submission ID”进入样本总览界面修改 GSA 元数据信息，详细修改原则详见“[GSA 数据集修改、删除和追加](#)”

Run 提交状态及可用操作：

实验编号	实验名称	物种名称	测序平台	文库布局	样本编号: 样本名称	实验提交状态	操作
CRX022949	scRNA-seq of WJMSC1	Homo sapiens	Illumina HiSeq 2500	FRAGMENT	SAMC021054: scRNA-seq of WJMSC1	Unchecked Confidential	<button>修改</button> <button>删除</button>
Run编号	Run名称	Run序列文件处理信息					
CRR022922	scRNA-seq of WJMSC1	CS300_TGACCA_L002_R1_001.fastq.gz				Status: Unchecked	<button>修改</button> <button>删除</button>

系统中 Run 的提交状态共有 8 种，具体情况详见下表：

序号	提交状态	说明	可用操作
1	Unchecked	Run metadata 信息已提交，等待后台审核	修改 ^[1] 、删除
2	Checked OK	Run metadata 信息审核通过，等待数据文件上传	修改 ^[1] 、删除
3	Checked failed	Run metadata 信息审核未通过	修改 ^[1] 、删除
4	Uploaded Succeed	数据文件匹配完毕	
5	Processing	数据文件正在处理	
6	Processed succeed	数据文件处理完成	
7	Processed error	数据文件处理错误	修改 ^[1] 、删除；
8	Deleted	已删除	

^[1]: 用户可通过点击“Submission ID”进入样本总览界面修改 GSA 元数据信息，详细修改原则详见“[GSA 数据集修改、删除和追加](#)”。