国家生物信息中心

China National Center for Bioinformation

基因组序列及变异数据汇交体系介绍

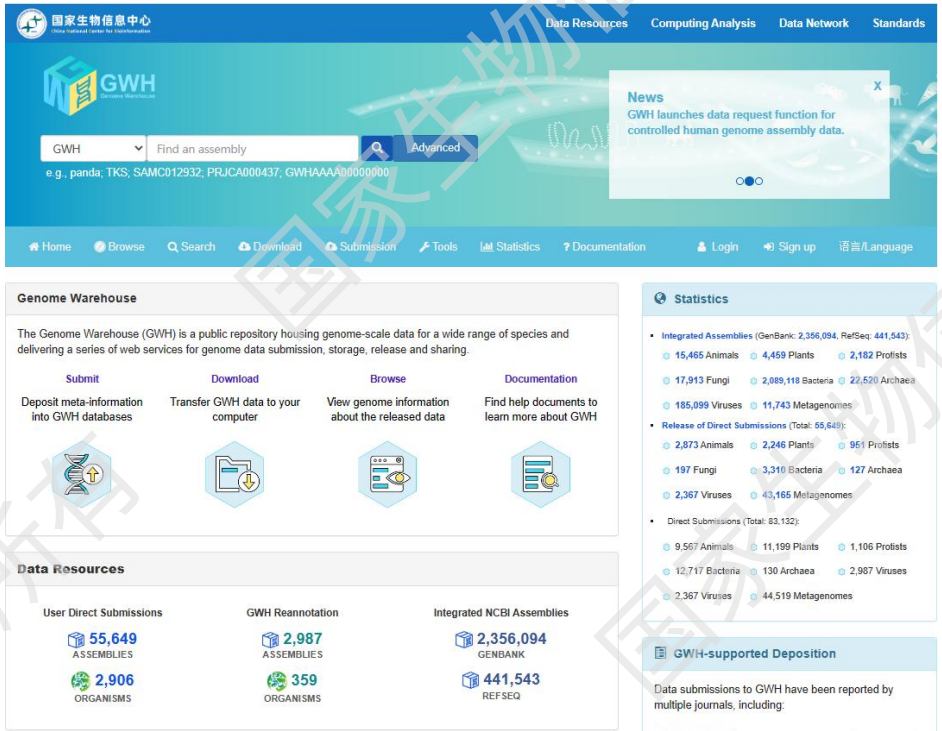
赵学彤 高级工程师

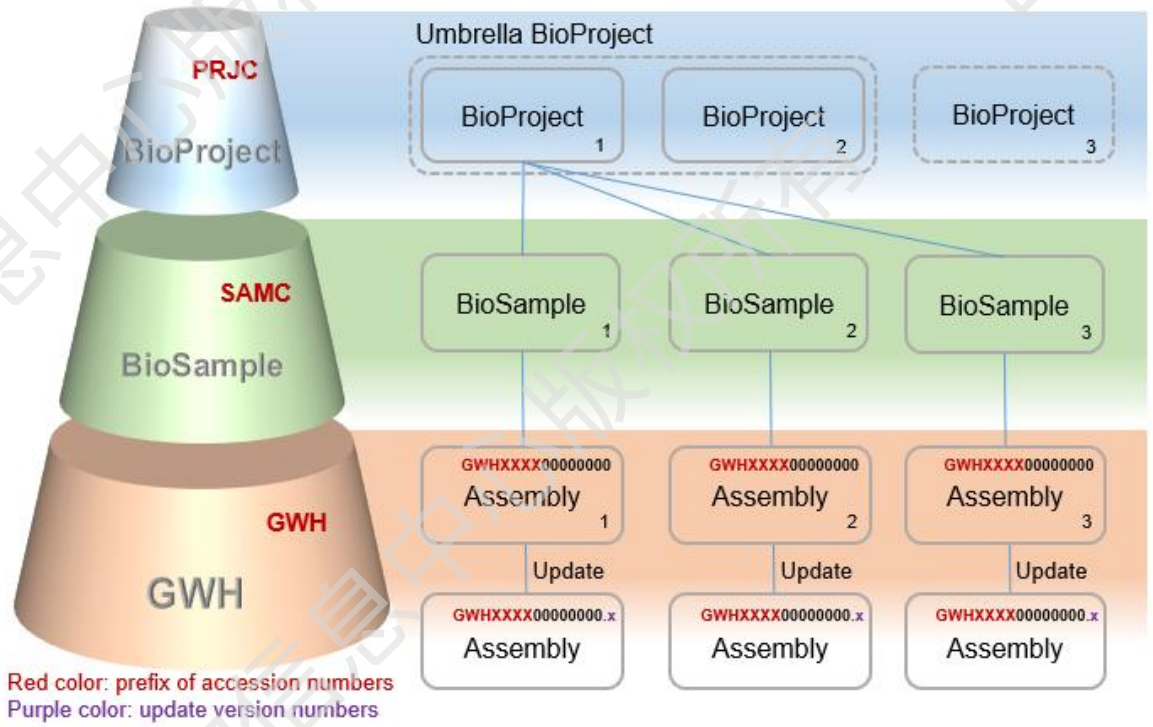
2025-11-20

目录

基因组数据库

围绕基因组大数据汇交与管理，研发基因组数据库（Genome Warehouse，简称GWH），提供基因组组装数据汇交、存储、质控、发布和共享等全链条数据服务





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

参考INSDC对基因组组装数据进行结构化的组织和管理

完善的基因组数据管理平台

Genome Data Standards

1 ID System
2 File Format
- 2.1 Submitted data
  - 2.1.1 Data format
  - 2.1.2 Data format description
- 2.2 Downloaded data

3 Metadata
- 3.1 Submitter
- 3.2 Gneral information
- 3.3 Genome assembly files
- 3.4 Sequence assignment
- 3.5 Reference

4 Data Analysis
- 4.1 Initial validation processing
  - 4.1.1 Genome sequence
  - 4.1.2 Genome annotation
  - 4.1.3 Sequence ordering and orientation information
  - 4.1.4 Sequence assignment
- 4.2 Sequence contamination processing

Data submission

Register account
↓
Register BioProject
↓
Register BioSample
↓
Create a submission
↓
Fill meta-data
↓
Upload files

Quality Control

Filename & md5 code
↓
Genome sequence
- Sequence ID
- Sequence content

↓
Genome annotation
- Gene structure completeness
- Gene structure consistency

↓
Data internal consistency
↓
Sequence contamination (optional)

Archive

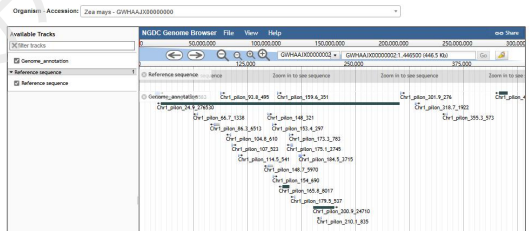
Accession assignment
- Genome assembly
- Genome sequence
- Gene
- Transcript
- Protein

↓
Generation of downloadable files & backup
- Genome sequence
- Genome annotation
- Gene feature
- RNA sequence
- CDS sequence
- Protein sequence

↓
Archival management
- Genome
- Assembly
- Download files
- Genome browser
- BIG Search

Release

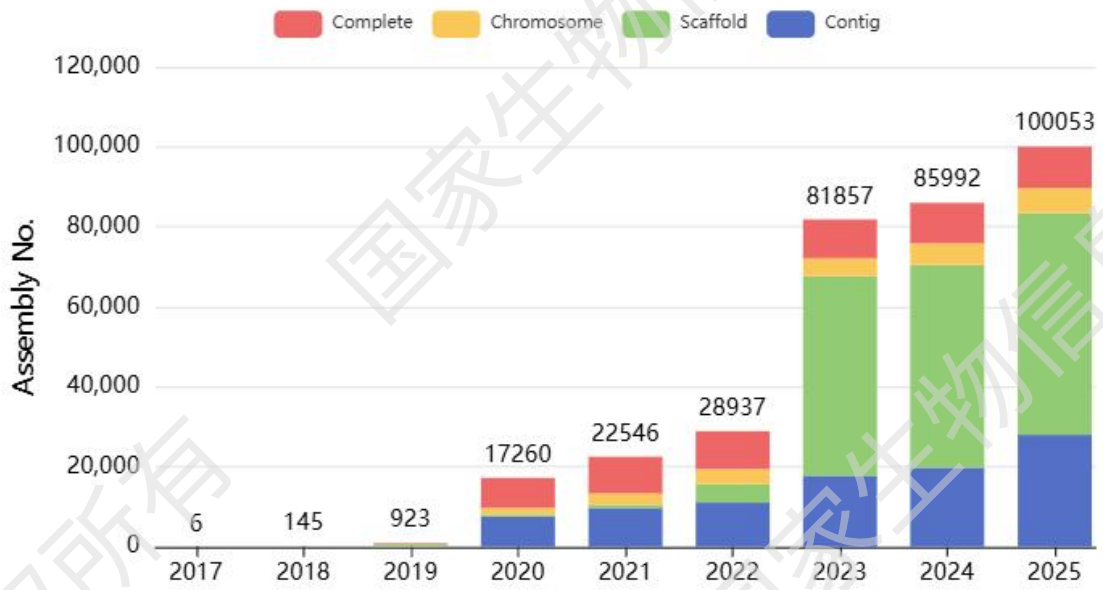
Visualization

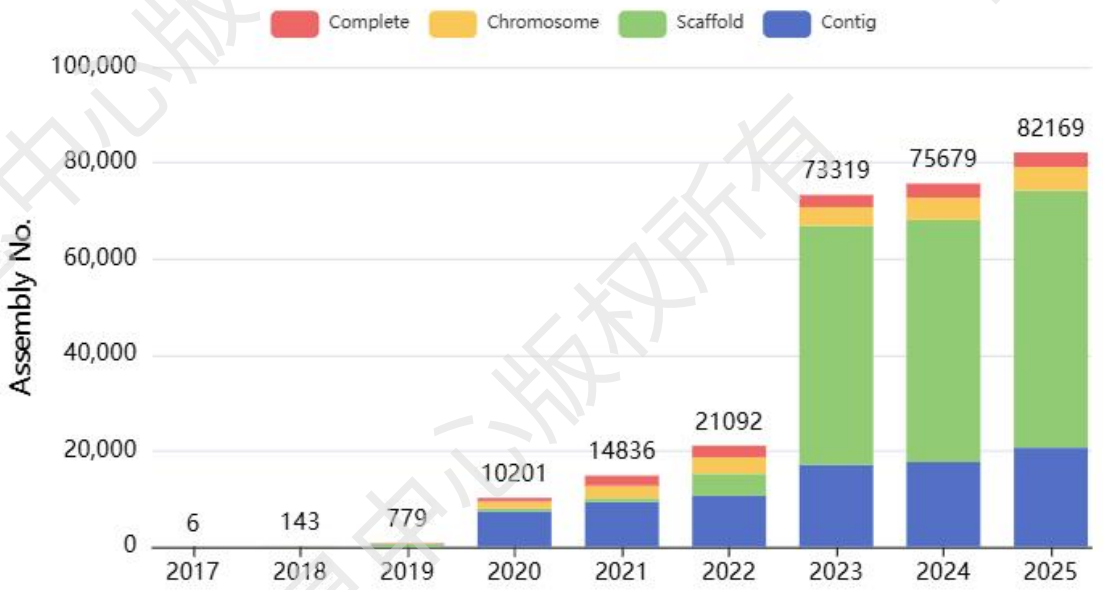


汇交 → 存储 → 质控 → 归档 → 发布 → 共享

实现国内基因组组装数据统一汇交和安全管理

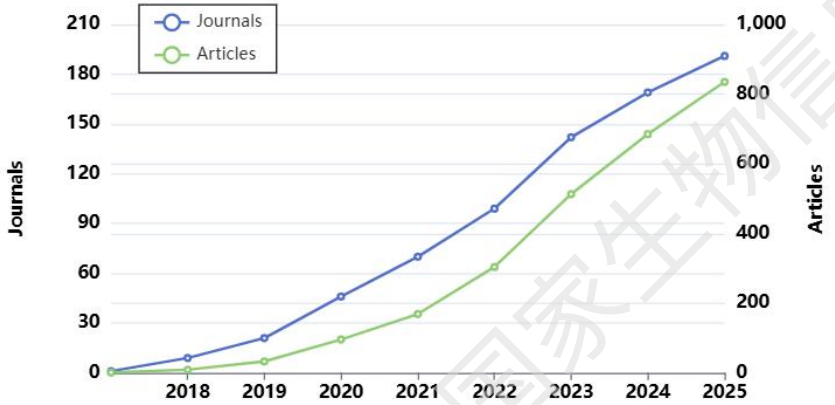
- 2,123 BioProjects
- 58,276 BioSamples
- 100,053 Genome assemblies
- 6,843 Organisms
- 284,030,093 Sequences
- 16,548 Gbases





支撑科技论文发表

期刊数: 191　　文章数：835



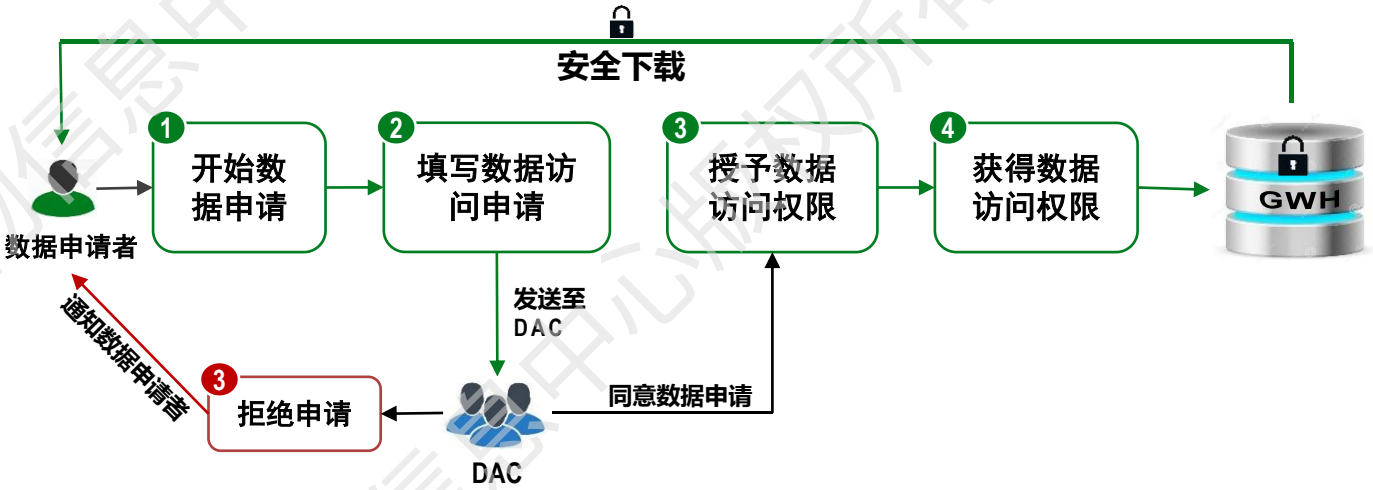
序号	期刊名称	文章标题	发表日期	GWH编号	BioProject编号
1	Plant Diversity	Speciation, endangerment and adaptation in limestone rocky environment of Urophysa (Ranunculaceae)	2025-09	GWHGEYG00000000.1	PRJCA035353
2	Nature Communications	Reticulate allopolyploidy and subsequent dysploidy drive evolution and diversification in the cotton family	2025-08	GWHESEP00000000 GWHESEQ00000000 GWHESER00000000 more	PRJCA019343
3	BMC Infectious Diseases	Whole-genome sequencing for analyzing the transmission characteristics of drug-resistant Mycobacterium tuberculosis in Ganzhou, China	2025-07	GWHFQAW00000000.1 GWHFQAX00000000.1 GWHFQAY00000000.1 more	PRJCA036393
4	Plant Biotechnology Journal	Chromosome-scale haplotype-resolved genome assembly of the autotetraploid alfalfa cultivar Bolivia	2025-07	GWHFIBZ00000000.1 GWHFICL00000000.1	PRJCA033031
5	BMC Bioinformatics	Blastn2dotplots: multiple dot-plot visualizer for genome comparisons	2025-06	GWHBDNP00000000.1	PRJCA005809
6	Molecular Horticulture	Smi-miRmTERF regulates organelle development, retrograde signaling, secondary metabolism and immunity via targeting a subset of SmmTERFs in Salvia miltiorrhiza	2025-06	GWHAOSJ00000000	PRJCA003150
7	iScience	Functional proteomic insights into the lateral oviduct-associated secretions of a bean bug	2025-06	GWHBAZH00000000	PRJCA005094
8	Horticulture Research	Plant resistance inducer AMHA enhances antioxidant capacities to promote cold tolerance by regulating the upgrade of glutathione S-transferase in tea plant	2025-06	GWHAZTZ00000000 GWHBAUV00000000	PRJCA003382
9	BMC Genomics	Comparative transcriptomic and proteomic analyses of hypoxia response in wild and cultivated tomato roots	2025-06	GWHBJTH00000000	PRJCA008297
10	Horticulture Research	A nearly complete haplotype-phased genome assembly of nerve plant (Fittonia albivenis) provides insights into leaf color evolution	2025-06	GWHEUVN00000000.1	PRJCA028152

人类基因组数据资源安全共享

- 两种访问方式
  - 受控访问 (Controlled-access)
  - 公开访问 (Open-access)

  注：共享前通过人遗备份和事先报告
- 受控访问数据，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

✓ 7,736	个人类基因组数据	✓ 58	次下载申请
✓ 1,187	个已公开发布	✓ 22	次获授权
✓ 1,073	个可公开访问	✓ 4,840	次完成下载
✓ 114	个受控访问	✓ 3.2	Tb下载量



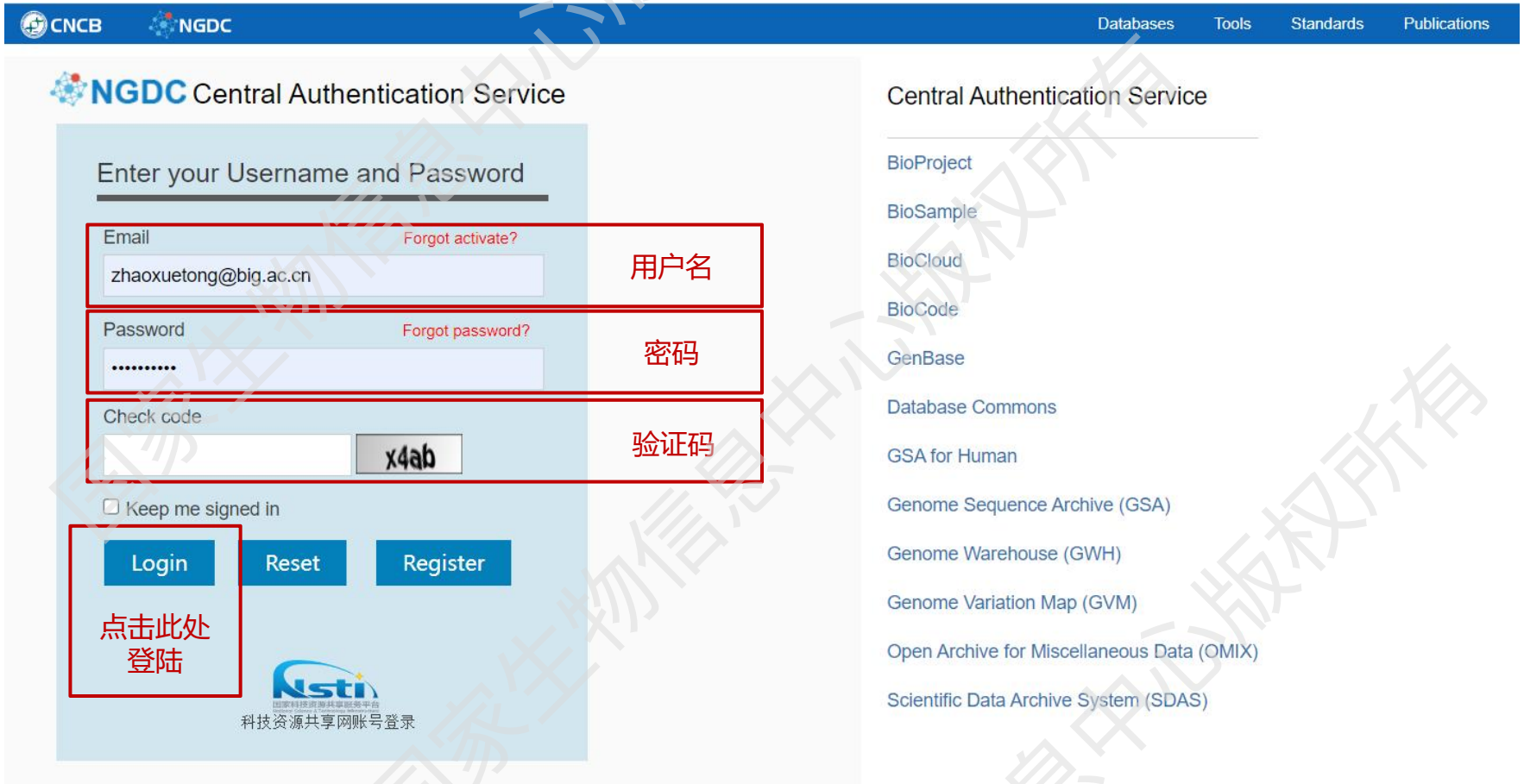
GWH人类基因组受控数据申请审核流程

准备工作：1. 账户准备——注册和登录

- 访问GWH库：https://ngdc.cncb.ac.cn/gwh/，进行注册和登录

SSO单点登录：中心任意一个网站注册和登录一次，就可以被授权访问中心其它数据库

准备工作：1. 账户准备-注册和登陆



- 登陆跳转到如下页面，填写相关信息，点击login登陆
- 未注册用户需要先注册再登陆

准备工作：2. 元信息文件准备

创建新的提交

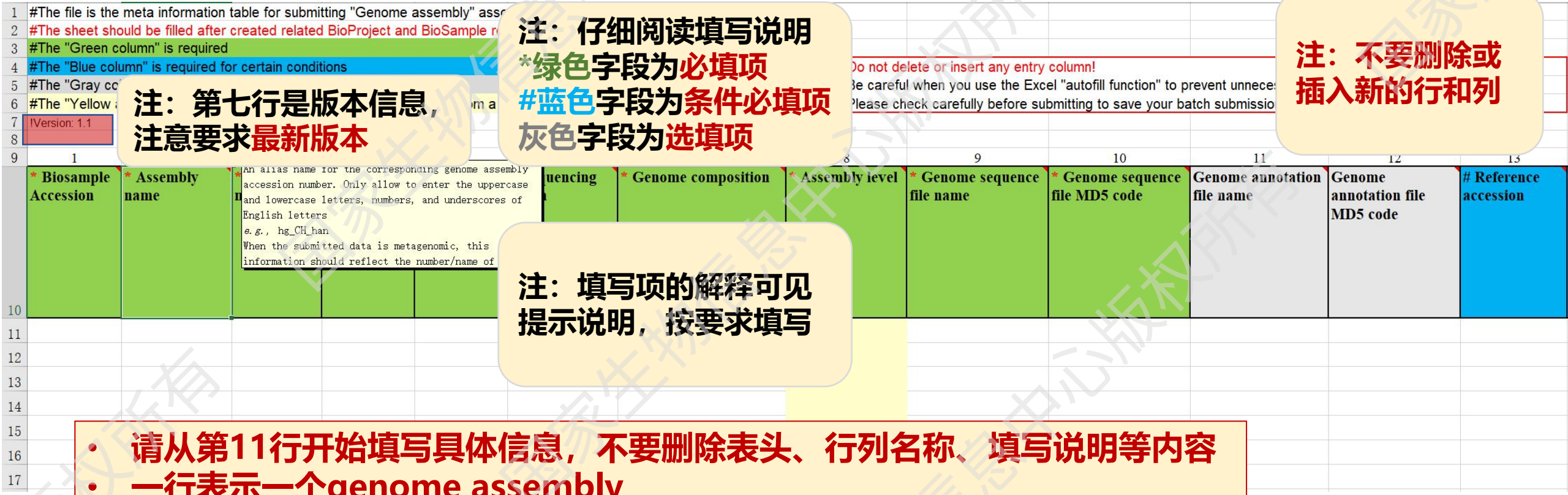
元信息文件下载位置

请注意：GWH接受与基因组相关的数据提交，但我们对提交数据的准确性并不负责，提交者应对数据的准确性负责。

1. 下载GWH批量提交模板文件 GWH_Assembly_en.xlsx / GWH_Assembly_ch.xlsx， 填写并仔细检查后上传
2. 批量提交模板的相关解释和示例，请参见GWH_Assembly_en.xlsx / GWH_Assembly_ch.xlsx。
3. 下载GWH从属信息模板文件 chr.csv / plasmid.csv / organella.csv 及其 说明文档，填写并仔细检查后通过FTP方式上传(可选)。
4. GWH提交相关常见问题解答，请参阅此 pdf。
5. 更多信息，请参阅 帮助文档。

准备工作：2. 元信息文件准备

下载后打开最新版本的GWH-batchsubmission-English.xlsx模板文件



注：第七行是版本信息，注意要求最新版本

注：仔细阅读填写说明
*绿色字段为必填项
#蓝色字段为条件必填项
灰色字段为选填项

注：不要删除或插入新的行和列

注：填写项的解释可见提示说明，按要求填写

- 请从第11行开始填写具体信息，不要删除表头、行列名称、填写说明等内容
- 一行表示一个genome assembly

准备工作：3. 数据文件准备

- 基因组组装序列文件（必需，fasta格式）

```
>seq_1 description line
ATCGATCGATCGATCGATCGATCGATCGATCG
ATCGATCGATCGAT
```

Sequence ID命名要求：

（1）以字母开头

（2）字母、数字、横线-、下划线_、点.、冒号:、星号*和#组成

如上面序列的sequence ID为seq_1

一个基因组只有：

- 1个序列文件
- 1个基因组注释文件，同时含编码、非编码的注释请合并

- 基因组注释（非必需，gff/tbl格式）
  - gff格式（九列）

1. Sequence ID;
2. Annotation information source (null value is '.');
3. Feature type, eg: gene, transcript;
4. Start position of a feature on the sequence (from 1);
5. End position of a feature on the sequence (no more than sequence length);
6. Score (null value is '.');
7. Strand;
8. Frame of a CDS (0, 1, 2);
9. Attributes, additional information about each feature;

For example,

```
##sequence-region ctg123 1 1497228
ctg123 .  gene 1000 4000 .  +  .  ID=gene00001;Name=EDEN
ctg123 .  TF_binding_site  1000 1012 .  +  .  Parent=gene00001
ctg123 .  mRNA 1300 3920 .  +  .  ID=mRNA00001;Parent=gene00001
ctg123 .  exon 1300 1500 .  +  .  ID=exon00001;Parent=mRNA00001
ctg123 .  exon 2000 3950 .  +  .  ID=exon00002;Parent=mRNA00001
ctg123 .  CDS  1300 1500 .  +  0  ID=cds00001;Parent=mRNA00001
ctg123 .  CDS  2000 3901 .  +  0  ID=cds00002;Parent=mRNA00001
ctg123 .  UTR  3902 3920 .  +  0  ID=UTR00001;Parent=mRNA00001
```

准备工作：3. 数据文件准备

- 序列从属和命名说明文件（组装水平为complete或者draft in chromosome时必需，csv格式）

染色体序列从属和命名说明文件

Sequence ID	Chromosome name	Complete	Circular
seq_1	1	true	false

细胞器序列从属和命名说明文件

Sequence ID	Type	Complete	Circular
seq_2	Mitochondrion	true	false

质粒序列从属和命名说明文件

Sequence ID	Plasmid name	Complete	Circular
seq_3	pBR322	true	false

- Sequence ID：必须是来自于基因组序列文件中的Sequence ID，即序列文件中 > 号后面的非空格字符串
- 完整度：“Complete=true” 表示该序列代表确定的染色体、细胞器或质粒。在染色体序列从属说明文件中，“Complete=false”表示与特定的染色体相关，但尚未定位到染色体上的特定位置，同理细胞器和质粒
- 是否环化：“Cirular=true”表示该条序列为可环化（序列首尾可含gap），“Cirular=false”表示该条序列为线性或为环化序列的片段（序列首尾不可含gap）

提交准备页面

详细的提交流程说明

提交准备内容

https://ngdc.cncb.ac.cn/gwh/submit/prepartion

提交流程：1. 创建新的提交

一个批量提交要求关联同一个BioProject，同一篇文章

方法二、通过中心统一汇交入口BIG Sub，登录GWH库的提交系统：https://ngdc.cncb.ac.cn/gsub/

方法一、通过GWH库登录提交系统：https://ngdc.cncb.ac.cn/gwh/submit/submission

提交流程：1. Meta信息在线提交

第一步：提交者信息

注：Submitter的邮箱将用于后续提交过程的信息联系，如有需要请同时填写需抄送的备用邮箱，数据发布后不公开

提交流程：1. Meta信息在线提交

第二步：基本信息

注：标题不对外公开，仅限于用户区分提交列表中的多个提交

注：请注意阅读数据发布正常和免责声明

注：受控访问方式仅限于人类基因组数据资源

提交流程：1. Meta信息在线提交

第三步：序列作者、联系人及文章信息

序列作者：数据发布时仅公开姓名，不公开email
序列联系人：数据发布时公开姓名和email

注：此处为发表该提交的基因组关联的文章信息

注：请确认提供的pubmed ID号对应的详细信息是否准确无误，其它方式需提供详细的文章作者列表

提交流程：1. Meta信息在线提交

第四步：上传组装元信息文件（一次批量提交可提交多个基因组），并在线校验

①选择本地文件并上传
②上传成功后，在线校验

校验失败

报错信息：
①文本文件下载
②在线可视化显示

校验通过

注：

- Error的报错必须修改后重新提交
- Warning的报错则用户根据实际数据情况自行判断

提交流程：1. Meta信息在线提交

第五步：预览提交信息，确认无误后提交

第1步 提交者信息 | 第2步 基本信息 | 第3步 文章信息 | 第4步 Meta | 第5步 预览

信息概览

批量提交号: Batch0023947
标题： test
提交状态： Unfinished Step 4

本提交的数据将于 **2024-05-23 (8** 天后)或引用文章公开发表后公开发布，这取决于哪个日期先到。
请注意：GWH数据的发布将触发关联的BioProject和BioSample一起发布。

FTP提交

您可以使用以下信息将文件传输到GWH的FTP站点:

- **主机地址**: submit.big.ac.cn
- **用户名**: chenml
- **密码**: *与登录GWH/NGDC时的用户密码相同, 例如:123456*
- **路径**: /GWH/Batch0023947

请注意：使用FTP客户端软件（如FileZilla client）登录FTP客户端。
提交文件后，请在您的提交列表中点击"完成上传"按钮 (**https://ngdc.cncb.ac.cn/gwh/submit/submission**)，通知我们审核您的提交。

FTP服务器登录说明

提交者信息

- **提交者信息**: Meili Chen
- **电子邮箱**: chenml@big.ac.cn
- **提交单位**: The CAS Key Laboratory of Genome Sciences and Information,Beijing Institute of Genomics, Chinese Academy of Sciences

基本信息

- **BioProject编号**: PRJCA000494
- **发布日期**: 2024-05-23
- **受控访问**: False

文章信息

- **序列作者**: mei null test
- **发表状态**: published
- **文章标题**: [The laboratory in programs for enteric infection control].
- **文章作者**:
- **通讯作者**: mei null test (test@qq.com)

元信息属性

每个基因组组装特有的元信息属性说明

显示第 1 至 1 项结果，共 1 项

BioSample编号	物种拉丁名	组装名称	组装水平	基因组序列文件名	基因组注释文件名	参考基因组编号	已有的GWH基因组编号	Chromos
SAMC016108	-	wgs_MAG_1	complete	JRJ_duck1.0.fa	N/A	N/A	N/A	

显示 10 项结果

完成提交

提交流程：1. Meta信息在线提交

Meta信息在线提交完成后返回提交列表

创建新的提交

请注意： GWH接受与基因组相关的数据提交，但我们对提交数据的准确性并不负责，提交者应对数据的准确性负责。

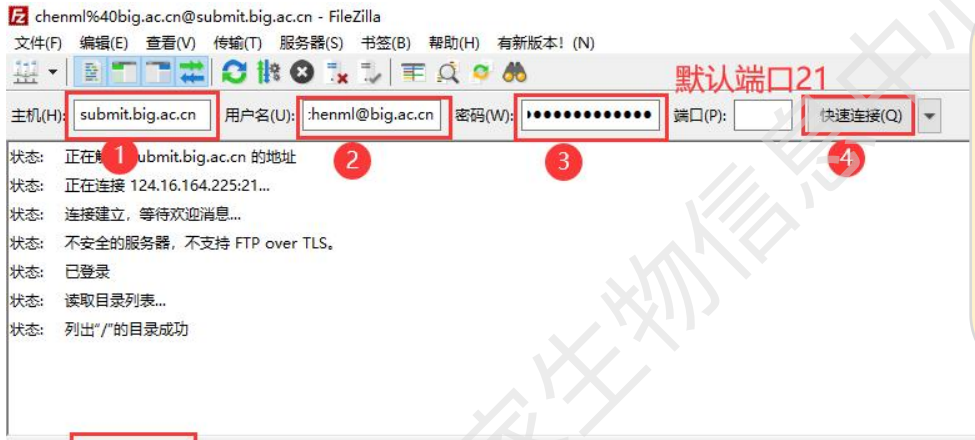
我的提交　已删除的提交

搜索:

7 个提交记录

组装编号	批量提交号	提交标题	BioProject编号	发布日期	创建日期	最后更新日期	提交状态	操作
GWHAZPK01000000 2 个组装编号，显示全部	Batch0018094	SARS-CoV-2_human_CHN_Wuhan_S5_2020	PRJCA004241	2027-07-26	2024-04-23	2024-04-23	SUCCESSFUL	审稿人页面 立即发布 修改发布日期 删除
	Batch0023947	test	PRJCA000494	2024-05-23	2024-04-23	2024-04-23	Waiting for files	完成上传 删除

提交流程：2. 数据文件FTP提交



◘ 登录操作步骤：

①主机地址：submit.big.ac.cn

②用户名：与登录GWH/NGDC时的用户名相同, 例如：chenml@big.ac.cn

③密码：与登录GWH/NGDC时的用户密码相同, 例如：123456

④点快速连接（默认端口21）

◘ 文件FTP上传操作

- 选定本地磁盘路径和文件
- GWH提交指定的FTP路径： /GWH/$Batch_ID，例如：/GWH/Batch0023947
- 本地文件直接拖拽至指定的FTP路径

提交流程：3. 完成提交后触发自动审核

- 元信息和数据文件完成提交后，返回提交列表，点击对应提交的“完成上传”（Finish upload）按钮

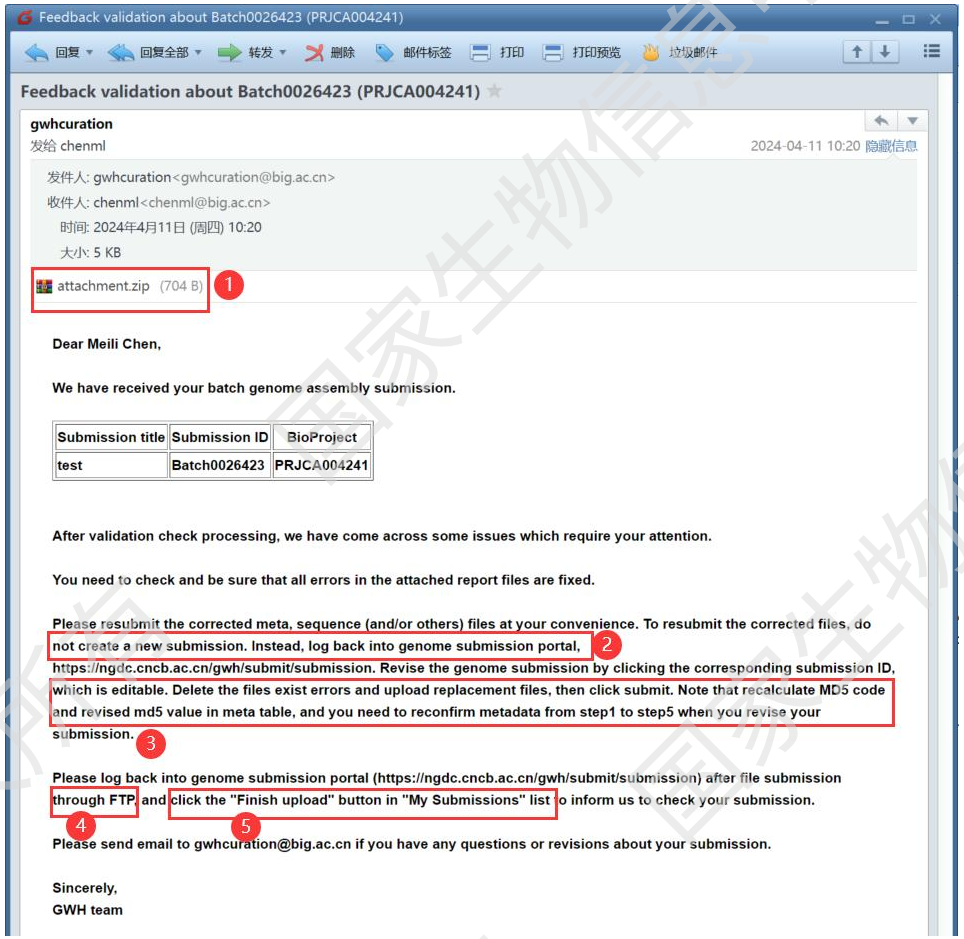
审核内容
- 基因组序列文件内容的合法性和有效性
- 基因组注释文件内容的合法性和有效性
- 基因组序列和注释内容一致性审核
- 基因组序列接头、载体等污染审核

数据提交状态为“Waiting for files”（即等待上传文件）

注：请务必点击该操作按钮来触发系统自动质控和归档
注意查收邮件，系统自动反馈质控报错/通过

提交流程：3. 系统审核及反馈

➢ 审核结果报错



Feedback validation about Batch0026423 (PRJCA004241)

回复 回复全部 转发 删除 邮件标签 打印 打印预览 垃圾邮件

Feedback validation about Batch0026423 (PRJCA004241)

gwhcuration
发给 chenml 2024-04-11 10:20 隐藏信息

发件人: gwhcuration<gwhcuration@big.ac.cn>
收件人: chenml<chenml@big.ac.cn>
时间: 2024年4月11日 (周四) 10:20
大小: 5 KB

attachment.zip (704 B) 1

Dear Meili Chen,

We have received your batch genome assembly submission.

Submission title	Submission ID	BioProject
test	Batch0026423	PRJCA004241

After validation check processing, we have come across some issues which require your attention.

You need to check and be sure that all errors in the attached report files are fixed.

Please resubmit the corrected meta, sequence (and/or others) files at your convenience. To resubmit the corrected files, do not create a new submission. Instead, log back into genome submission portal, 2 https://ngdc.cncb.ac.cn/gwh/submit/submission. Revise the genome submission by clicking the corresponding submission ID, which is editable. Delete the files exist errors and upload replacement files, then click submit. Note that recalculate MD5 code and revised md5 value in meta table, and you need to reconfirm metadata from step1 to step5 when you revise your submission. 3

Please log back into genome submission portal (https://ngdc.cncb.ac.cn/gwh/submit/submission) after file submission through FTP, 4 and click the "Finish upload" button in "My Submissions" list 5 to inform us to check your submission.

Please send email to gwhcuration@big.ac.cn if you have any questions or revisions about your submission.

Sincerely,
GWH team

①查收邮件，根据附件的错误报告修改后提交
②在原来的提交号中修改并提交
③如数据文件有修改，请注意重新计算文件MD5码
④FTP中重新上传修改后的数据文件
⑤返回提交列表，点击“完成上传”（Finish upload）按钮

我的提交 已删除的提交

搜索: 23947

7 个提交记录

组装编号	批量提交号	提交标题	BioProject编号	发布日期	创建日期	最后更新日期	提交状态	操作
	Batch0023947	test	PRJCA000494	2024-05-23	2024-04-23	2024-04-23	ERROR	删除

显示第 1 至 1 项结果，共 1 项 (由 7 项结果过滤)

点击进入修改

数据提交状态为“Error”（即有错误）

提交流程：3. 系统审核及反馈

➢ 审核通过

- 邮件通知审核通过
- 附上正式Accession号
- 发表文章时的引用模板

数据提交状态为 “Successful”（即归档成功）

① 显示分配的Accession号（X表示大写字母，x表示版本号）

✓ 基因组包含多条序列：GWHXXXX00000000.xx，第一版本GWHXXXX00000000.1，其等价于GWHXXXX00000000

✓ 基因组只含一条单序列：GWHXXXXxx000000，第一版本GWHXXXX01000000，如GWHAZPK01000000

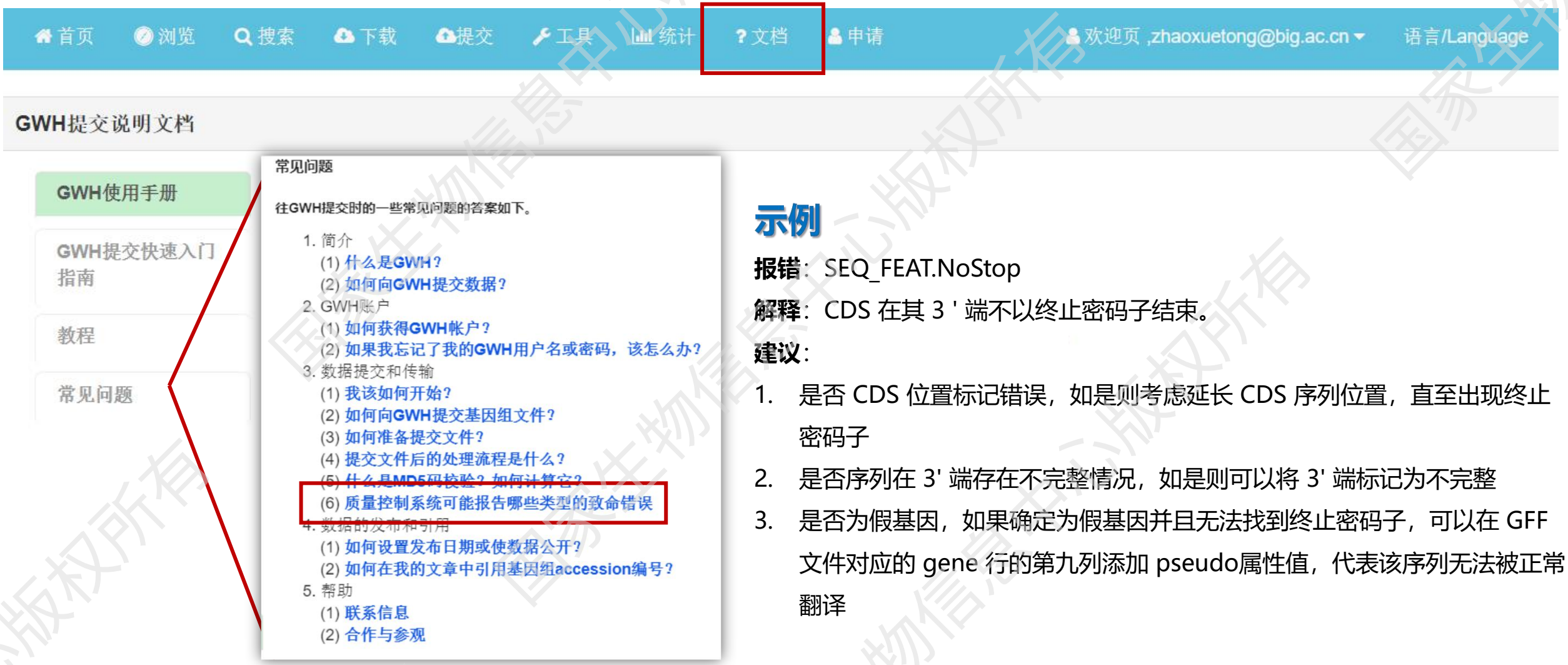
② 显示当前数据的状态为“Successful”（即归档成功）

③ 支持操作：生成审稿人链接、立即发布数据、修改发布日期

报错修正文档1:常见问题解答

1、注册邮箱激活失败
2、GWH接受数据的要求
3、数据FTP形式上传注意事项
4、提交数据注释文件
5、元信息文件中Assembly level填写说明
6、元信息文件中Assignment相关信息填写说明
7、序列文件规范
8、注释文件规范
9、数据释放时间
10、杂志认定
11、获取审稿人链接
12、报错文件（.err和.user）的打开方式
13、部分具体报错信息分析
14、追加注释文件
15、修改数据meta信息
16、文章已发表操作
17、数据update
18、提交者邮箱
19、多个biosample的选择
20、Linkage evidence信息
21、对于提交的MAG数据
22、参考文档

报错修正文档2：常见错误类型及解决方案



示例

报错：SEQ_FEAT.NoStop

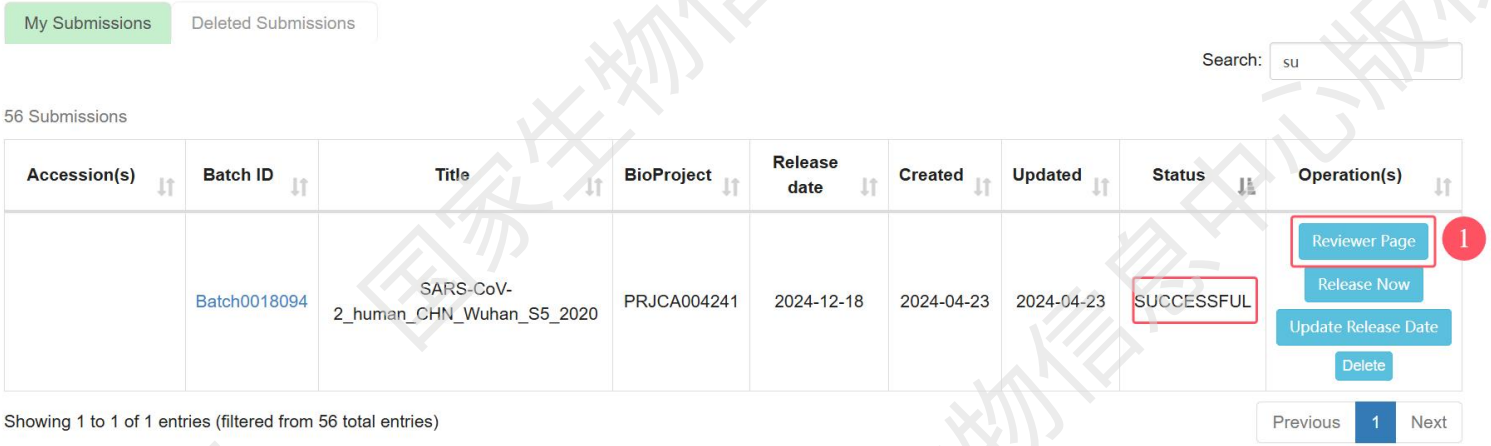
解释：CDS 在其 3 ' 端不以终止密码子结束。

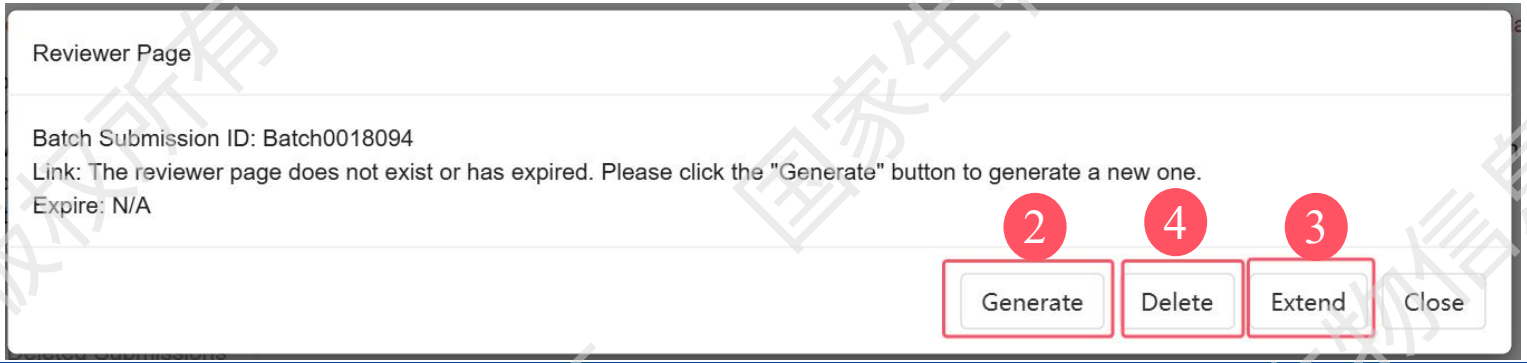
建议：

1. 是否 CDS 位置标记错误，如是则考虑延长 CDS 序列位置，直至出现终止密码子
2. 是否序列在 3' 端存在不完整情况，如是则可以将 3' 端标记为不完整
3. 是否为假基因，如果确定为假基因并且无法找到终止密码子，可以在 GFF 文件对应的 gene 行的第九列添加 pseudo属性值，代表该序列无法被正常翻译

数据发布前审稿人链接自助生成和延期

对于通过质控归档的数据，在发布前（状态为Successful）可自助生成和延期审稿人链接。





① 在提交列表中点击后进入审稿人链接管理

② 首次生成审稿人链接

③ 延期已生成的审稿人链接

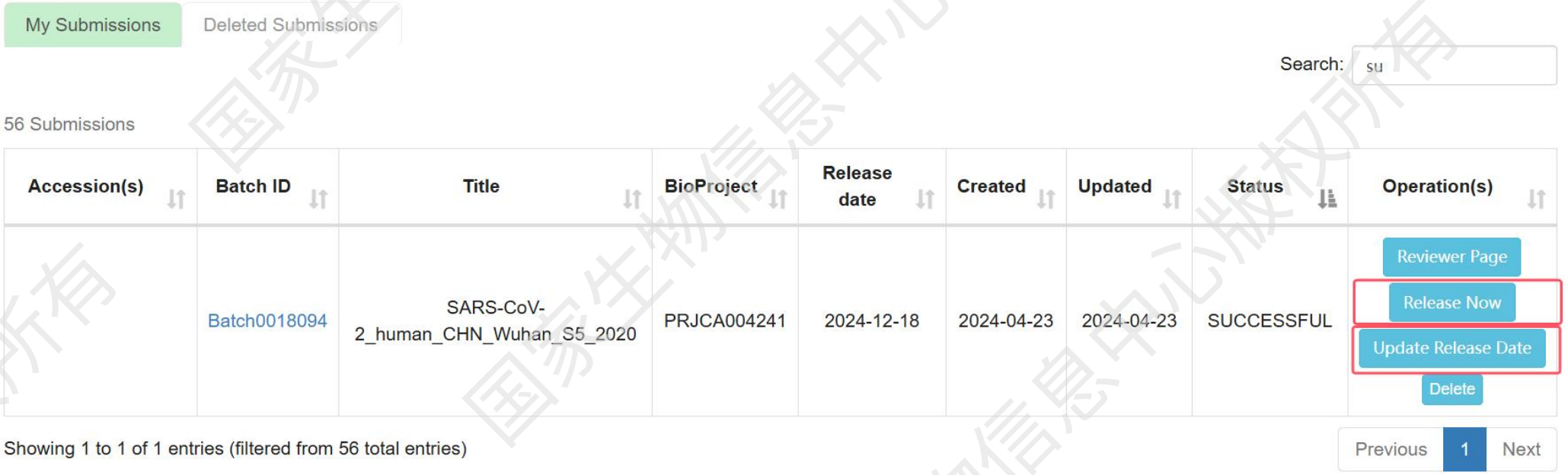
④ 删除已生成的审稿人链接使其失效

注：请注意妥善保管审稿人链接，获取到链接的用户都可以访问数据，尤其是申请将数据文件一起共享的

https://ngdc.cncb.ac.cn/gwh/submit/submission

数据发布自助管理

对于通过质控归档的数据，在发布前（状态为Successful）可自助立即发布和修改发布日期。



目录

基因序列数据库简介

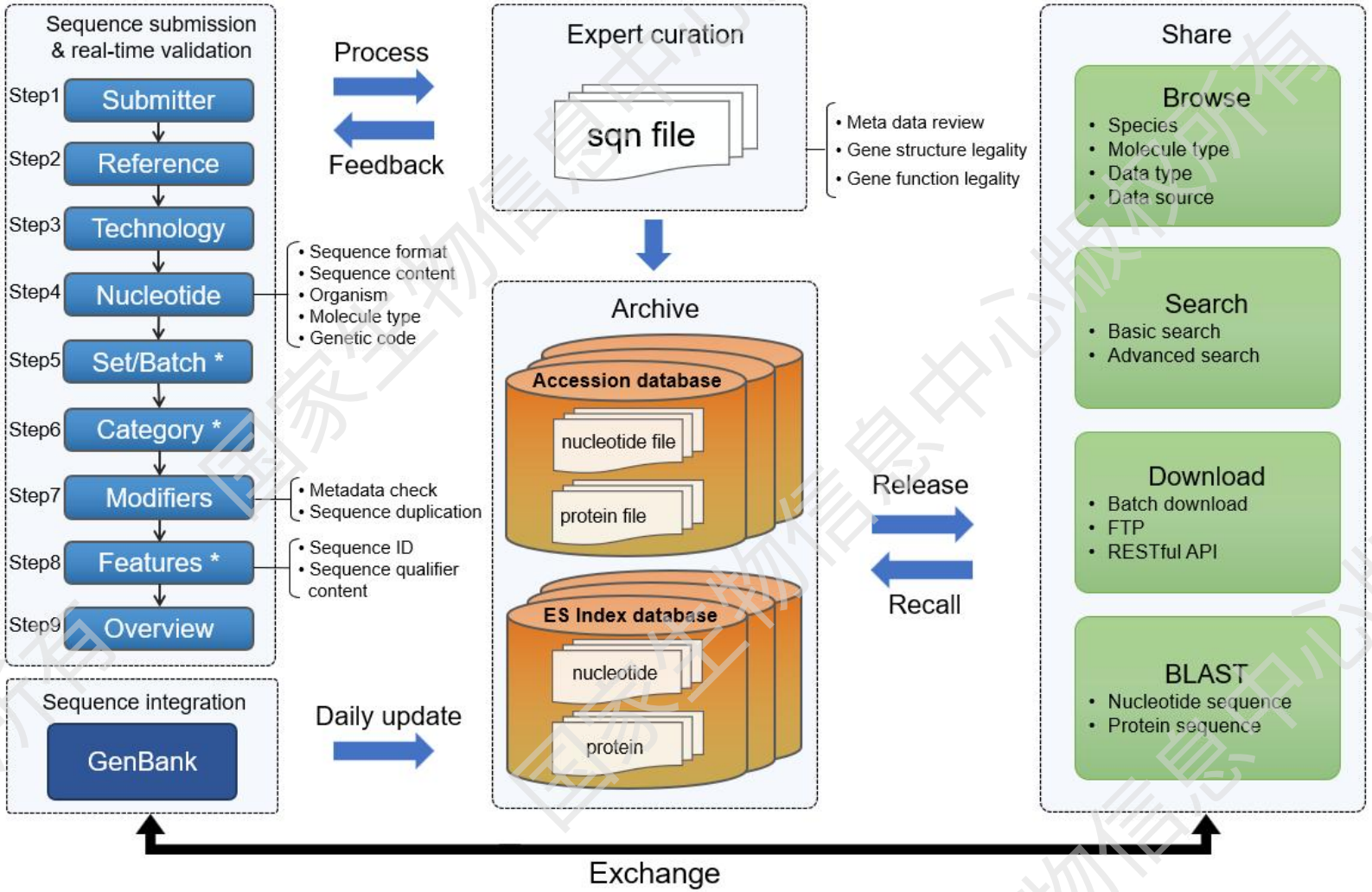
GenBase是一个公共、免费的基因数据归档库，收录、存储、管理与共享**病毒、细胞器、质粒和基因片段等核酸序列及其注释信息**

基因序列数据库首页，https://ngdc.cncb.ac.cn/genbase/

基因组和基因序列汇交的异同

数据库	数据汇交范围	数据类型	数据组织结构和关联
基因组数据库GWH	大基因组 • 真核生物全基因组 • 原核生物全基因组 • 元基因组组装数据	• **基因组**组装序列 • 注释数据（非必需）	**涉及**BioProject和BioSample
基因序列库GenBase	小基因组和基因片段 • 病毒基因组 • 细胞器基因组 • 质粒基因组 • 所有物种基因片段	• **基因片段**序列 • 注释数据（非必需）	**可以不涉及**BioProject和BioSample

规范、智能、用户友好的基因序列数据管理平台



汇交

↓

存储

↓

质控

↓

共享

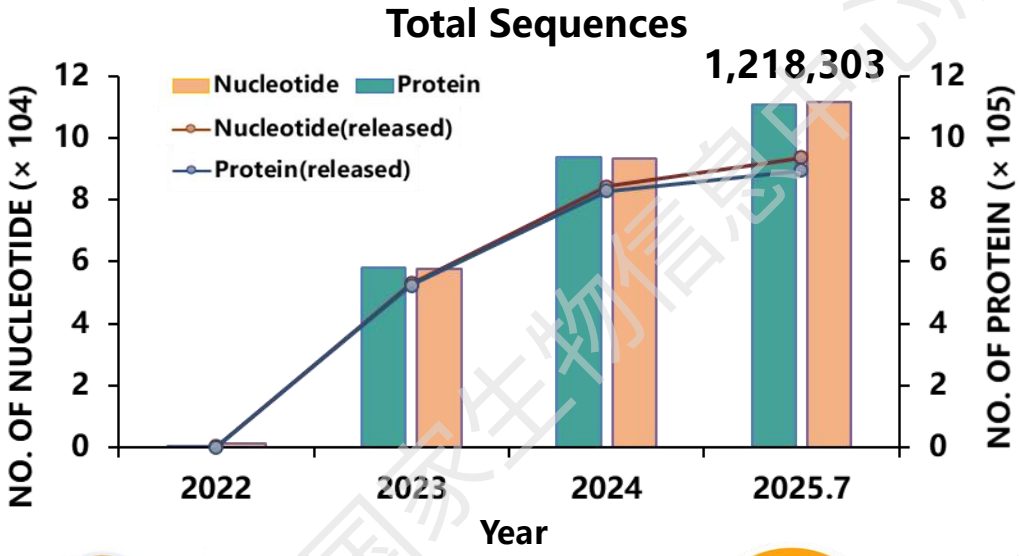
↓

检索

↓

下载

实现基因序列数据统一汇交和安全管理



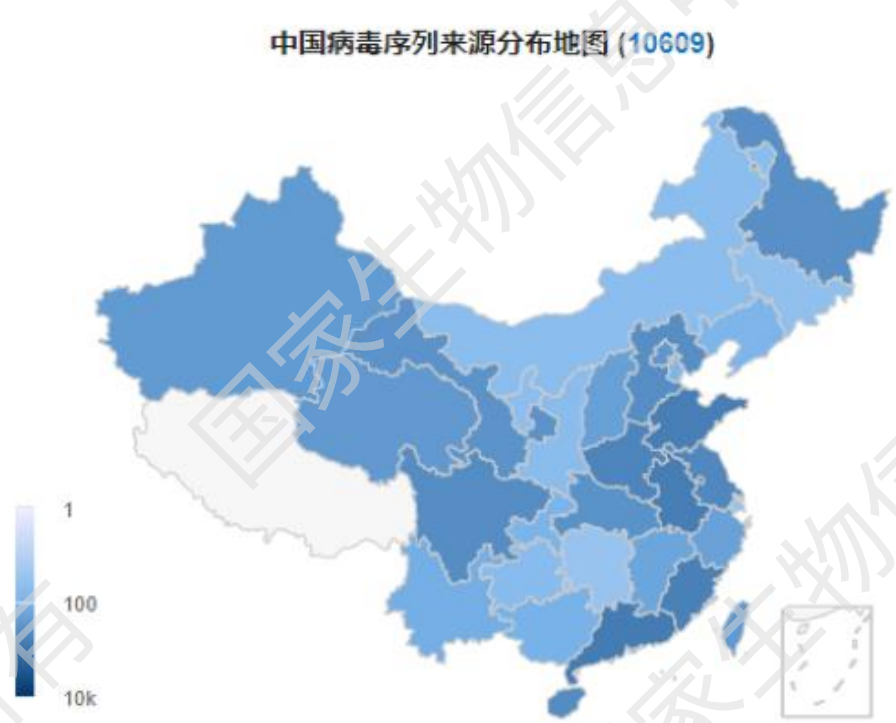
服务成效

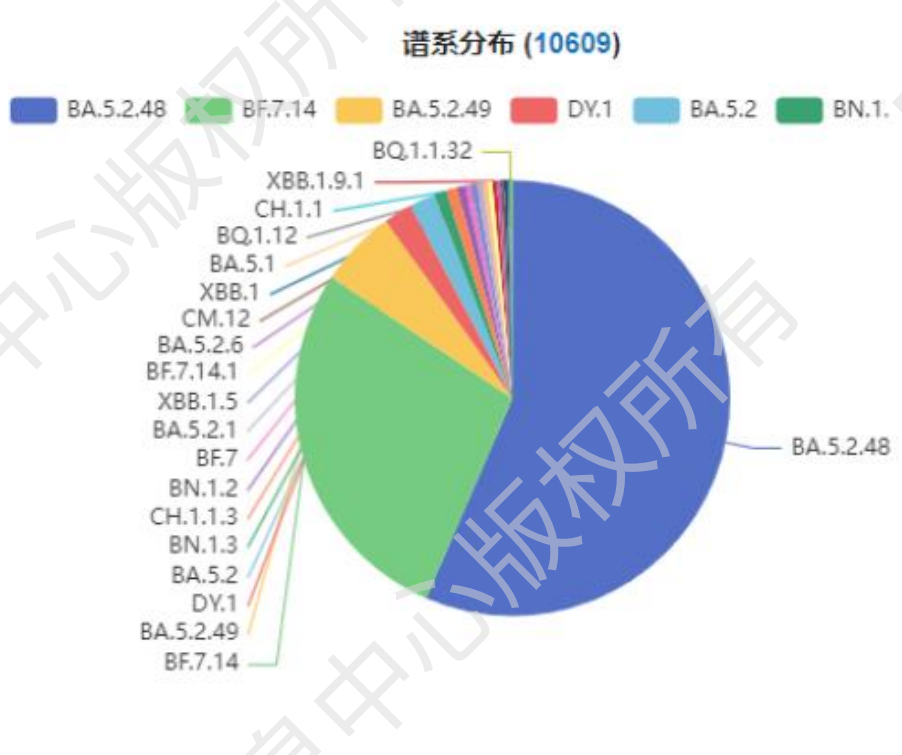
- 291个单位
- 477 名用户
- 3,446 批提交
- 85 论文发表
- 404 个基金
- 105,845 访客

2024年度“中国生物信息学十大进展”

面对国际数据脱钩风险显现的新发展态势，**同步镜像**NCBI GenBank和RefSeq数据库中的基因序列数据资源，并实现**日更新**

受国家疾控局委托
承担新冠病毒基因组数据的汇交和发布





截至2025年11月，共收到86家单位的2,665次数据递交，

共计73,656条新冠病毒基因组序列

准备工作：1 账户准备-登陆

数据库网址：https://ngdc.cncb.ac.cn/genbase/ 进入网站进行登陆

准备工作：2 数据准备

准备工作：2 数据准备

提交类型

常规序列提交 | SARS-CoV-2快速提交 | 受控序列提交

常规序列提交准备

1.概述

请准备以下信息:

1. 基本信息: 您的联系方式，作者，出版物，数据发布日期
2. 提交类型:
    - 原始组装/注释
    - 同一基因座的多个序列集合(如果适用)
    - 分子类型
3. FASTA格式的核酸序列
4. 物种名
5. 元信息, 例如: isolate, strain, collection date, country
6. 特征注释, 例如: CDS (coding region), tRNA, ncRNA, gene

准备工作：2 数据准备

① 常规信息：联系方式，作者信息，出版信息，数据发布日期等

② FASTA格式核苷酸序列

- 利用本地linux运行执行**序列校验**
- GenBaseTools(gbt)：https://ngdc.cncb.ac.cn/genbase/download/template/gbt

③ Source Modifiers文件：包含collection date、country、host等

④ Annotation Features文件：包含gene、CDS、tRNA、ncRNA等

准备工作：2 数据准备-FASTA文件

- FASTA文件，可包含**一条或多条**序列，整合到一个文件中上传
- 请使用FASTA格式，以定义行开始，然后是序列行
- 最简单的定义行需要“>”符号和一个Sequence ID

Sequence ID

For example:
>Seq1 [organism=Homo Sapiens] Definition Line for Seq1
aaccgatatagagagagga
>Seq2 [organism=Homo Sapiens] Definition Line for Seq2
atctgaatagagattattt

注意：定义行仅识别Sequence ID和物种信息，其他信息不识别，其他信息请在元信息和注释信息中体现

Sequence ID命名要求：

（1）以字母开头

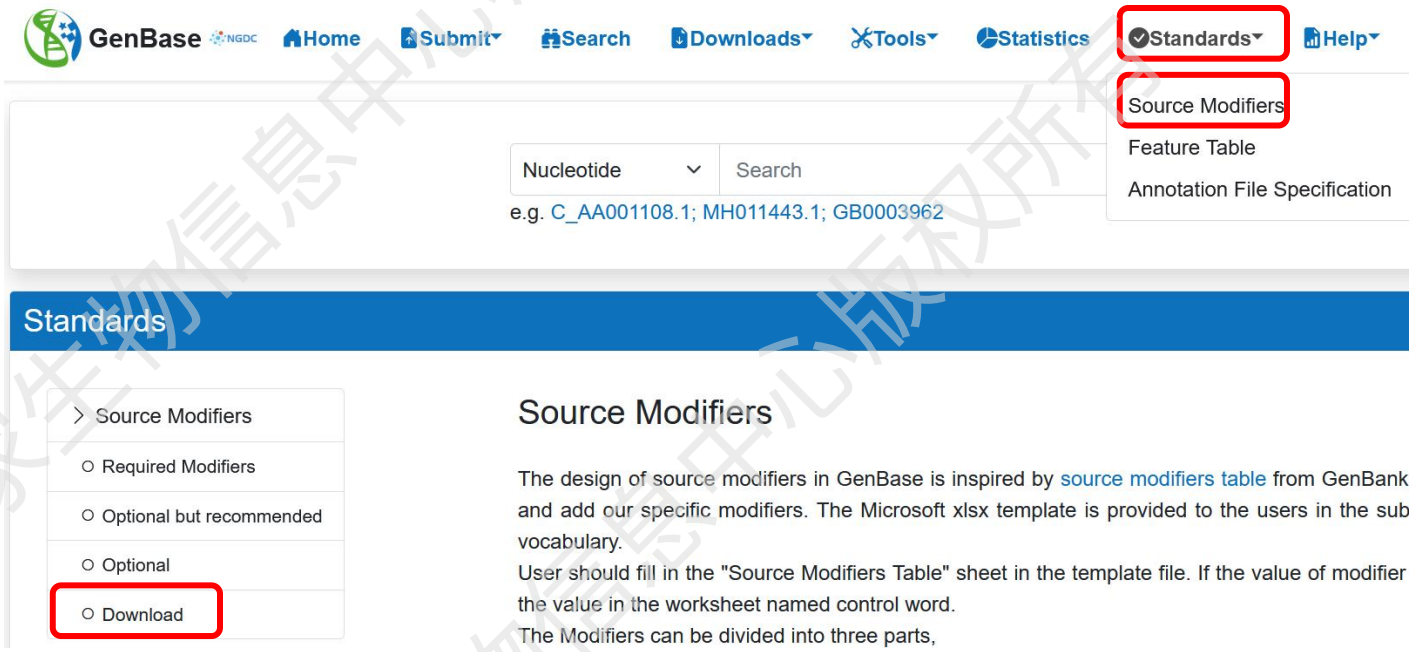
（2）字母、数字、横线-、下划线_、点.、冒号:、星号*和#组成

（3）请短于13个字符

如上面序列的sequence ID为Seq1

准备工作：2 数据准备-Source Modifiers文件

- 模板文件：GenBase_Modifiers.xlsx
  （https://ngdc.cncb.ac.cn/genbase/download/template/Genbase_Modifiers.xlsx）
- GenBase_Modifiers.xlsx文件包含**相关受控词汇表**，用于描述您如何、何时、何处获得样本以及样本的相关信息等



准备工作：2 数据准备-Source Modifiers文件

下载后打开GenBase_Modifiers.xlsx模板：

	A	B	C	D	E	F	G
1	#!Version: 3.3						
2	#!DO NOT MODIFY HEAD LINES AND EXAMPLE LINE! ‖ 不要修改标题行和示例行！						
3	#!The first column contains the Sequence_IDs used to identify each sequence in the nucleotide FASTA file. ‖ 第一列包含用于识别核苷酸 FASTA 文件中每个序列的 Sequence_ID。						
4	#!Specimens are identified in the Source Modifiers Table by the sam						
5	#!The heading for the first column must be exactly Sequence_ID as						
6	#!Each specimen in the set must have a line in the source modifiers f						用于样本。
7	#!Each Sequence_ID may appear only once in the source modifier f						
8	#!Fill in the data in the line beginning with "##number". If it is not enough, you can append lines by yourself. ‖ 填写以“##数字”开头的行中的数据。 如果不够，可以自行追加。						
9	#!Reference: https://ngdc.cncb.ac.cn/genbase/standards ; https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers						
10	#!If you do not know what to fill in the optional modifiers, please leave it blank instead of filling in characters such as "Na", "null", "unavailable"... ‖ 如果您不知道如何填写可选的Modifiers，请将其留空，而不是填写“Na”、“						
11	##Color Code	required	optional (recommended)	optional			
12	##Column Number	##1	##2	##3	##4	##5	##6
13	##Header	Sequence_ID	Collection_date	Country/Region	Clone	Host	Host_sex
14	##Description	Sequence ID, MUST be consistent with the fasta file	Date the specimen was collected. --- It MUST be text format so that we can parse it without any ambiguity.	The country where the sequence's organism was located. May also be an ocean or major sea. Additional region or locality information must be after the country name and separated by a ':'. For example: USA: Riverview Park, Ripkentown, MD. Available values can be found at ControlWords sheet.	Name of clone from which sequence was obtained, typically an alphanumeric ID.	When the sequence submission is from an organism that exists in a symbiotic, parasitic, or other special relationship with some second organism, the 'host' modifier can be used to identify the name of the host species.	Sex of Host
15	##Example(do not delete)	Seq1	2001-01-31 or value in ControlWords sheet	China or value in ControlWords sheet	C.Grant	Homo Sapiens	Male
16	##1						
17	##2						
18	##3						
19	##4						
20	##5						

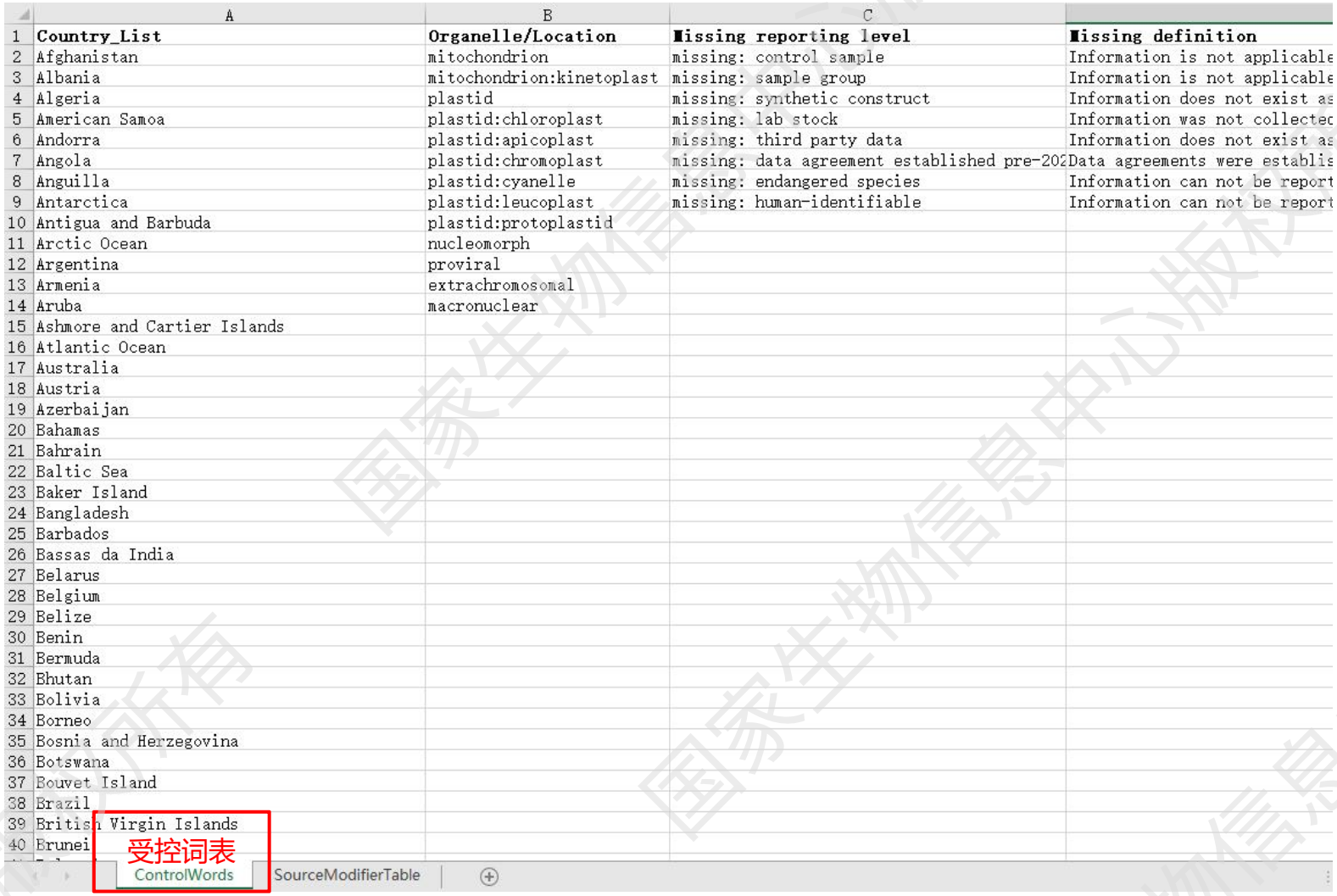
1.仔细阅读填写说明

#蓝色字段为必填项，黄色字段为可选推荐项，绿色字段为可选项

2.请在此处填写具体信息，不要删除表头、行列名称、填写说明等内容

3.请给每个序列提供一个独特的信息，如 isolate, clone, strain, laboratory identifier，至少填写一项，以区分各记录

准备工作：2 数据准备-Source Modifiers文件



	A	B	C	
1	**Country_List**	**Organelle/Location**	**Missing reporting level**	**Missing definition**
2	Afghanistan	mitochondrion	missing: control sample	Information is not applicable
3	Albania	mitochondrion:kinetoplast	missing: sample group	Information is not applicable
4	Algeria	plastid	missing: synthetic construct	Information does not exist as
5	American Samoa	plastid:chloroplast	missing: lab stock	Information was not collected
6	Andorra	plastid:apicoplast	missing: third party data	Information does not exist as
7	Angola	plastid:chromoplast	missing: data agreement established pre-202	Data agreements were establis
8	Anguilla	plastid:cyanelle	missing: endangered species	Information can not be report
9	Antarctica	plastid:leucoplast	missing: human-identifiable	Information can not be report
10	Antigua and Barbuda	plastid:protoplastid		
11	Arctic Ocean	nucleomorph		
12	Argentina	proviral		
13	Armenia	extrachromosomal		
14	Aruba	macronuclear		
15	Ashmore and Cartier Islands			
16	Atlantic Ocean			
17	Australia			
18	Austria			
19	Azerbaijan			
20	Bahamas			
21	Bahrain			
22	Baltic Sea			
23	Baker Island			
24	Bangladesh			
25	Barbados			
26	Bassas da India			
27	Belarus			
28	Belgium			
29	Belize			
30	Benin			
31	Bermuda			
32	Bhutan			
33	Bolivia			
34	Borneo			
35	Bosnia and Herzegovina			
36	Botswana			
37	Bouvet Island			
38	Brazil			
39	British Virgin Islands			
40	Brunei			

受控词表

ControlWords | SourceModifierTable

- Country_List
- Organelle/Location
- Missing

准备工作：2 数据准备-Annotation Features文件

- 对于简单的注释，例如所有序列的功能相同，可以在页面下载**模板文件GenBase_Features.xlsx**填写并上传
  (https://ngdc.cncb.ac.cn/genbase/download/template/Genbase_Features.xlsx)
- 对于复杂的注释，请准备一个**五列制表符分隔的特征表（TBL格式）**或者**九列制表符分隔的注释文件（GFF3格式）**来上传

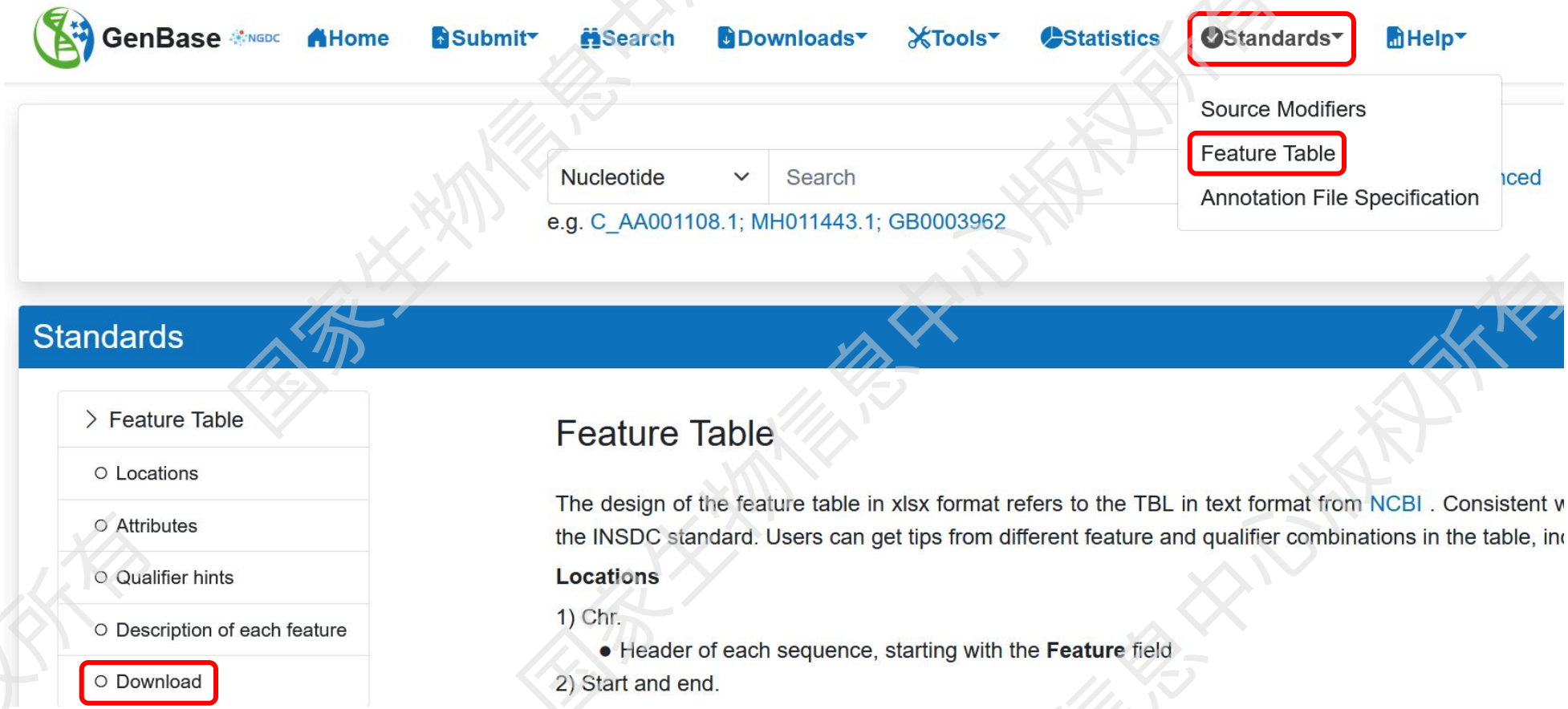
TBL格式：

```
>Feature Sc_16
1    7000    REFERENCE
                     PubMed          8849441
<1   1050    gene
                     gene            ATH1
<1   1009    CDS
                     product         acid trehalase
                     product         Ath1p
                     codon_start     2
<1   1050    mRNA
                     product         acid trehalase
```

GFF3格式：

```
##gff-version 3.1.26
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
```

准备工作：2 数据准备-Annotation Features文件



准备工作：2 数据准备-Annotation Features文件

下载后打开GenBase_Features.xlsx模板：

Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns **A ~ G**. Where **feature** (column E) is the structure name of this type of sequence, **qualifiers** (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, **qualifiers hints** (column G) will help you

1.仔细阅读填写说明，特征内容需填写在A-G列

Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的**A到G列**。其中**feature (column E)** 是该序列中一个对应的结构体，**qualifier (column F)**是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构体和属性，**Qualifier hints**将会获得这些属性的定义、示例或者格式，帮助您正确的填写**qualifier value (column G)**

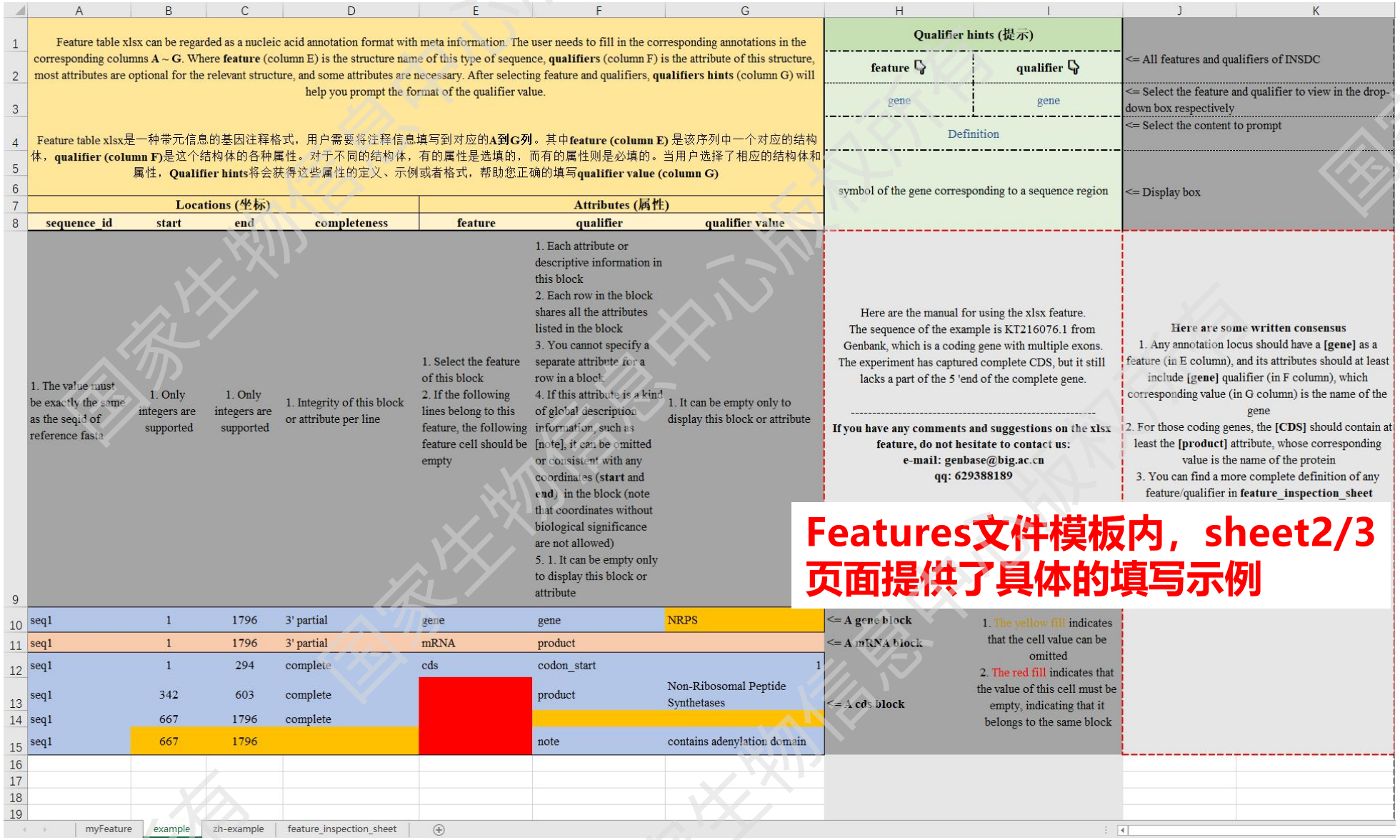
Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value

3. 请在此处填写具体信息
不要删除表头、行列名称、填写说明等内容

Qualifier hints (提示)	
feature	qualifier
gene	gene
Example	
/gene=ilvE	

2. 提示区
根据提示填写G列内容，供参考

准备工作：2 数据准备-Annotation Features文件



Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns **A ~ G**. Where **feature** (column E) is the structure name of this type of sequence, **qualifiers** (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, **qualifiers hints** (column G) will help you prompt the format of the qualifier value.

Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的**A到G列**。其中**feature (column E)** 是该序列中一个对应的结构体，**qualifier (column F)**是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构体和属性，**Qualifier hints**将会获得这些属性的定义、示例或者格式，帮助您正确的填写**qualifier value (column G)**

Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value
1. The value must be exactly the same as the seqid of reference fasta	1. Only integers are supported	1. Only integers are supported	1. Integrity of this block or attribute per line	1. Select the feature of this block 2. If the following lines belong to this feature, the following feature cell should be empty	1. Each attribute or descriptive information in this block 2. Each row in the block shares all the attributes listed in the block 3. You cannot specify a separate attribute for a row in a block 4. If this attribute is a kind of global description information, such as [note], it can be omitted or consistent with any coordinates (**start** and **end**) in the block (note that coordinates without biological significance are not allowed) 5. 1. It can be empty only to display this block or attribute	1. It can be empty only to display this block or attribute
seq1	1	1796	3' partial	gene	gene	NRPS
seq1	1	1796	3' partial	mRNA	product	
seq1	1	294	complete	cds	codon_start	1
seq1	342	603	complete		product	Non-Ribosomal Peptide Synthetases
seq1	667	1796	complete			
seq1	667	1796			note	contains adenylation domain

Qualifier hints (提示)

feature	qualifier
gene	gene
Definition	
symbol of the gene corresponding to a sequence region	

<= All features and qualifiers of INSDC

<= Select the feature and qualifier to view in the drop-down box respectively

<= Select the content to prompt

<= Display box

Here are the manual for using the xlsx feature.
The sequence of the example is KT216076.1 from Genbank, which is a coding gene with multiple exons. The experiment has captured complete CDS, but it still lacks a part of the 5 'end of the complete gene.

If you have any comments and suggestions on the xlsx feature, do not hesitate to contact us:
e-mail: genbase@big.ac.cn
qq: 629388189

Here are some written consensus
1. Any annotation locus should have a **[gene]** as a feature (in E column), and its attributes should at least include **[gene]** qualifier (in F column), which corresponding value (in G column) is the name of the gene
2. For those coding genes, the **[CDS]** should contain at least the **[product]** attribute, whose corresponding value is the name of the protein
3. You can find a more complete definition of any feature/qualifier in **feature_inspection_sheet**

<= A gene block

<= A mRNA block

<= A cds block

1. The yellow fill indicates that the cell value can be omitted
2. The red fill indicates that the value of this cell must be empty, indicating that it belongs to the same block

Features文件模板内，sheet2/3页面提供了具体的填写示例

myFeature | example | zh-example | feature_inspection_sheet

准备工作：2 数据准备-Annotation Features文件

The comparison table lists all the **features** and **qualifiers** that can be filled in. More filling explainations can be obtained at https://www.insdc.org/documents/feature_table.html

definition	feature	qualifiers	type	comment
1) region at the 3' end of a mature transcript (following the stop codon) that is not translated into a protein; 2) region at the 3' end of an RNA virus (following the last stop codon) that is not translated into a protein;	**3'UTR**	(Optional) allele	text	Na
		(Optional) citation	[number]	
		(Optional) db_xref	<database>:<identifier>	
		(Optional) experiment	[CATEGORY:]text	
		(Optional) function	text	
		(Optional) gene	text	
		(Optional) gene_synonym	text	
		(Optional) inference	[CATEGORY:]TYPE[(same spec	
		(Optional) locus_tag	text (single token)	
		(Optional) map	text	
		(Optional) note	text	
		(Optional) old_locus_tag	text (single token)	
		(Optional) standard_name	text	
		(Optional) trans_splicing	boolean	
1) region at the 5' end of a mature transcript (preceding the initiation codon) that is not translated into a protein; 2) region at the 5' end of an RNA virus genome (preceding the first initiation codon) that is not translated into a protein;	**5'UTR**	(Optional) allele	text	NA
		(Optional) citation	[number]	
		(Optional) db_xref	<database>:<identifier>	
		(Optional) experiment	[CATEGORY:]text	
		(Optional) function	text	
		(Optional) gene	text	
		(Optional) gene_synonym	text	
		(Optional) inference	[CATEGORY:]TYPE[(same spec	
		(Optional) locus_tag	text (single token)	
		(Optional) map	text	
		(Optional) note	text	
		(Optional) old_locus_tag	text (single token)	
		(Optional) standard_name	text	
		(Optional) trans_splicing	boolean	
coding sequence; sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes	**CDS**	(Optional) allele	text	/codon_start has valid value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature; /transl_table defines the genetic code table used if other than the universal genetic code table; genetic code exceptions outside the range of the specified tables is reported in /transl_except qualifier; /protein_id consists of a stable ID portion (from the end of 2018 new
		(Optional) artificial_location	[artificial_location_value]	
		(Optional) circular_RNA	boolean	
		(Optional) citation	[number]	
		(Optional) codon_start	<1 or 2 or 3>	
		(Optional) db_xref	<database>:<identifier>	
		(Optional) EC_number	text	
		(Optional) exception	[exception_value]	
		(Optional) experiment	[CATEGORY:]text	
		(Optional) function	text	
		(Optional) gene	text	
		(Optional) gene_synonym	text	
		(Optional) inference	[CATEGORY:]TYPE[(same spec	
		(Optional) locus_tag	text (single token)	
		(Optional) map	text	

myFeature | example | zh-example | feature_inspection_sheet

Feature_inspection_sheet

Features文件模板内，sheet4提供了详细的features、qualifiers、types示例，可供参考

提交流程：创建新的提交

两种方式进入GenBase递交系统：

GenBase主页 ‘数据提交’

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

BIG Sub GenBase提交入口

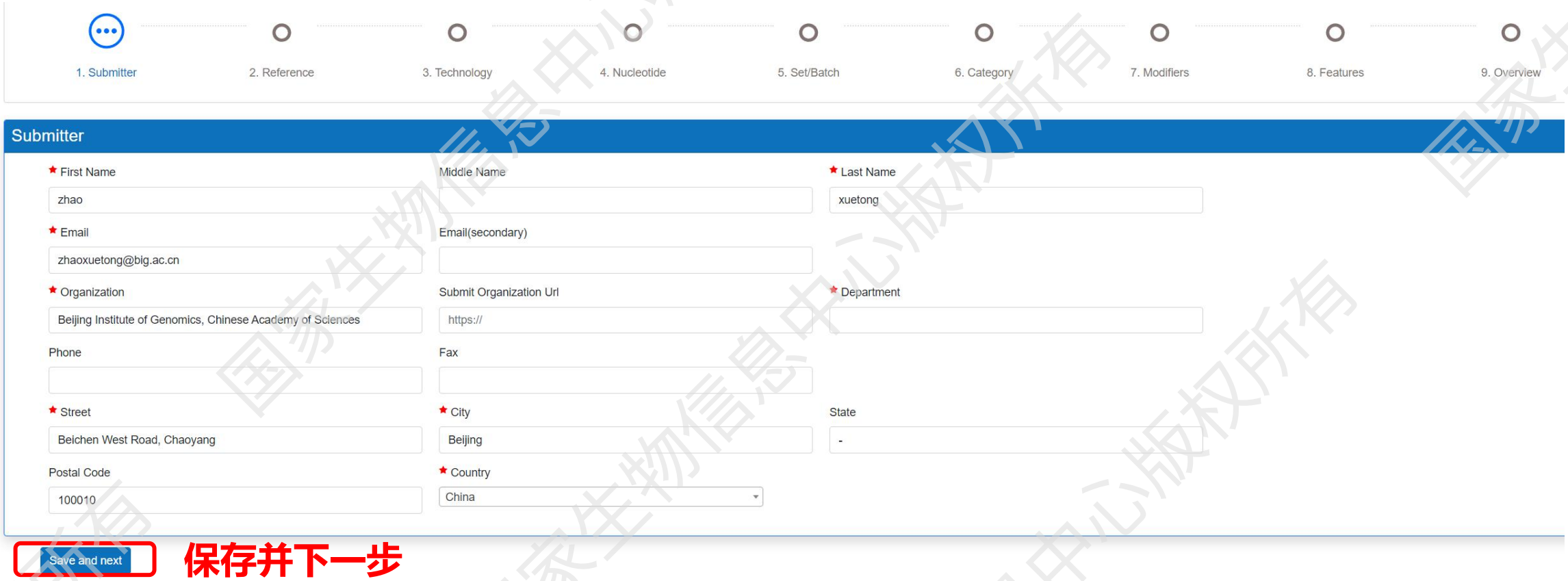
https://ngdc.cncb.ac.cn/gsub/

提交流程：创建新的提交

*受控序列提交：

用于人类遗传资源管理，只有**人的数据**才能使用该提交模块进行受控数据提交（古人不需要）

提交流程： 步骤1 提交者信息 (Submitter)



保存并下一步

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邮件信息**，如部分信息需要调整，也可直接在此处修改

#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**提交者邮箱**

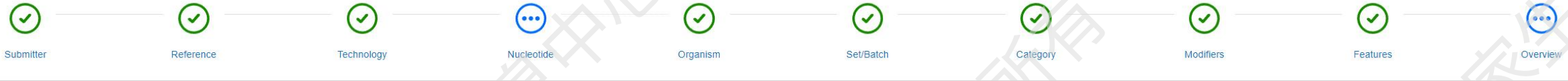
提交流程： 步骤2 参考文献信息(Reference)

保存并下一步

- 可选择增加通讯作者，根据具体情况如实填写
- 如果数据归档后文章发表，可以发邮件到genbase@big.ac.cn邮箱，帮您更新文献信息
- 注意不能写中文!!

提交流程： 步骤3 测序技术(Technology)

提交流程： 步骤4 序列 (Nucleotide)



Nucleotide

Submission #0000099

Submission Release Date

When may we release your sequence record?

- Immediately After Processing
- Release Date:

设置数据释放时间

a.最长不要超过4年

b.只要数据未公开，可随时修改该时间

c.对应文章发表后，数据会自动触发发布

Release date must be after 6 months from today and before 4 years from today.

Help

Release Policies and Disclaimers

1. A date can be set by authors to withhold the release of new submissions for a specified period of time.
2. If submitters want to change release date, please contact GenBase group: genbase@big.ac.cn.
3. If a paper citing the sequence or accession number is published prior to the specified date, the sequence will be released upon publication. Otherwise, GenBase will release sequence data on the specified date.
4. As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address: genbase@big.ac.cn.
5. After the data is released, it may automatically synchronize to NCBI in the future, and submitters will no longer be contacted for confirmation.

仔细阅读数据发布政策和免责声明

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

Submission Short Title

* Short Title:

设置提交标题，尽量有真实含义，此信息可在提交列表中体现，便于检索，不能是中文!!

BioProject

BioProject:

填写BioProject信息（可选）

Grant(s)

*Agency	Program	*Grant ID	Grant Title	Remove
Ministry of Science and Technology	National Key Technologies R&D Pr	123		X

填写资助基金信息，填写真实的，而非aaa/12345

提交流程： 步骤4 序列 (Nucleotide)

根据具体情况如实填写

要保证一批次提交的数据上述情况统一

是否属于**细胞器基因组**、病毒、类病毒、质粒或克隆载体的**完整序列**?

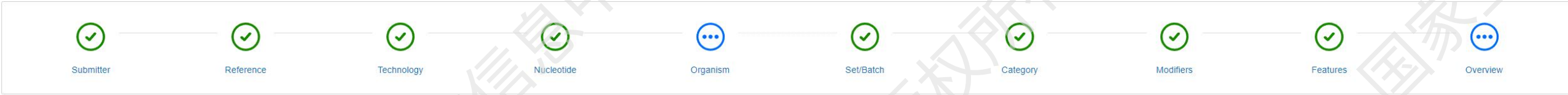
核酸序列填写示例

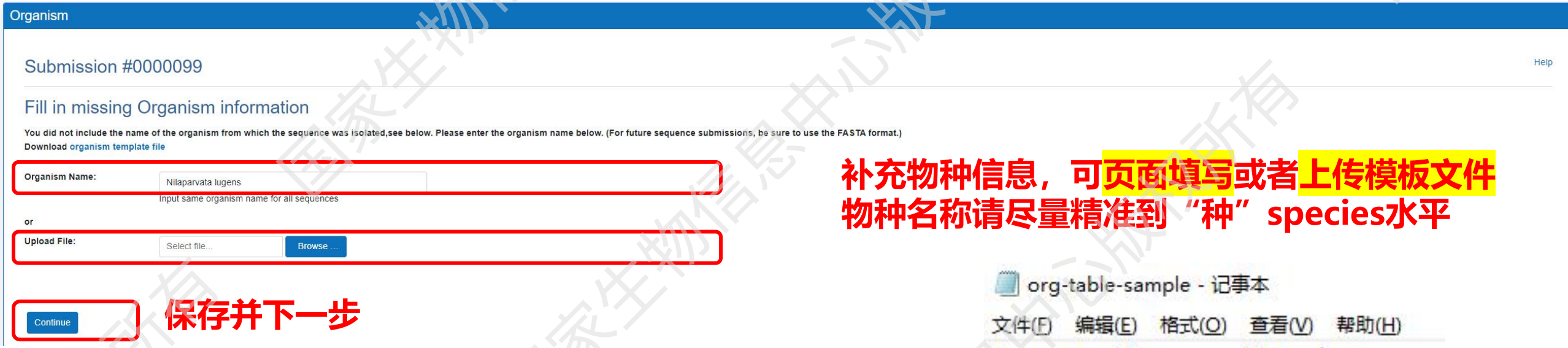
提交流程： 步骤4 序列 (Nucleotide)

保存并下一步

- 提供三种方式上传核酸序列，根据序列数的大小选择最佳方式
- 如果文件大小超过1GB，可通过FTP方式上传数据到GenBase文件夹下
- 如果此处不方便在FASTA定义行填写物种信息，可在下一页面补充

提交流程： 步骤5 物种 (Organism)

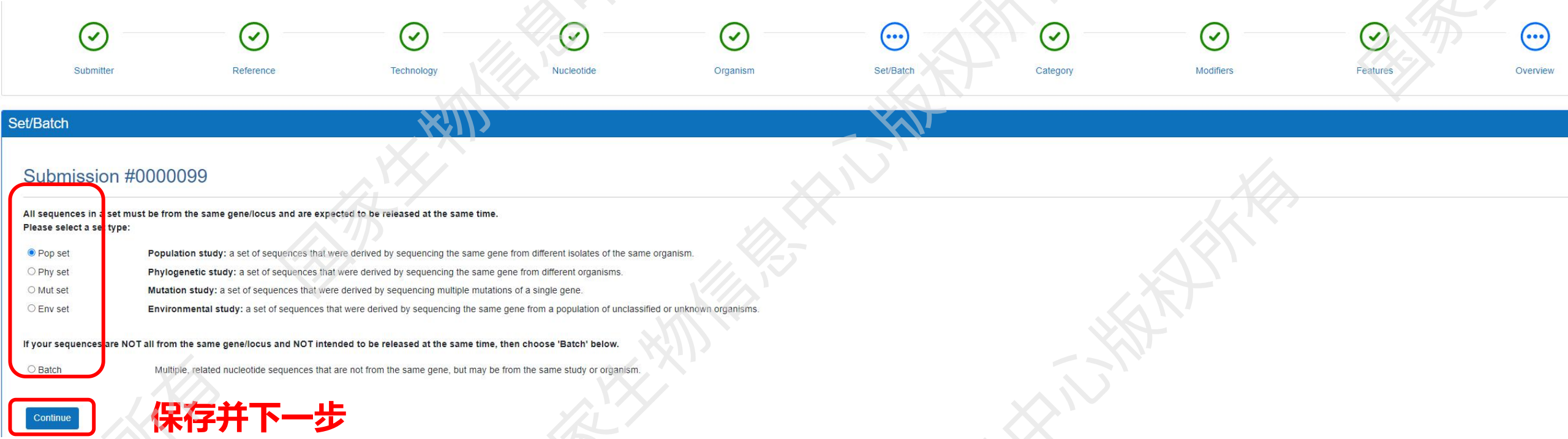




保存并下一步

补充物种信息，可页面填写或者上传模板文件
物种名称请尽量精准到“种”species水平

提交流程： 步骤6 集合/批次 (Set/Batch)



保存并下一步

根据数据具体情况，选择数据是否来自同一集合，并选择一种合适类型

提交流程： 步骤7 类别确认 (Category)

保存并下一步

- 确认您的序列是原始提交或第三方注释
- 第三方注释数据，需要在文章发表后，数据才能公开

提交流程： 步骤8 提交Modifiers文件 (Modifiers)

Submitter — Reference — Technology — Nucleotide — Organism — Set/Batch — Category — Modifiers — Features — Overview

Source Modifiers

Submission #0000099

Help

Note

1. Download Source Modifiers submission template file GenBase_Modifiers.xlsx, then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. GenBase_Modifiers.xlsx.

1.点击下载Modifiers文件模板

* **Are these sequences derived from organelle genomes? (such as Mitochondria or Chloroplast)**
yes no If yes, please fill in the Organelle/Location column in GenBase_Modifiers.xlsx

根据序列是否来自于细胞器基因组，如线粒体或叶绿体,选择yes或no

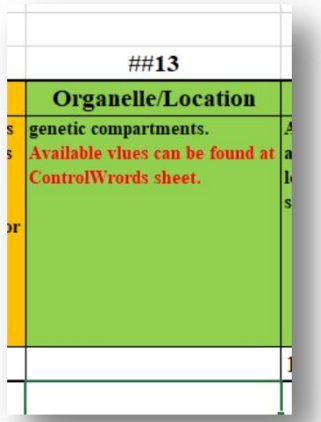
Upload the file in Excel format that includes the attributes for each sequence.

0000099_modifier.xlsx Remove

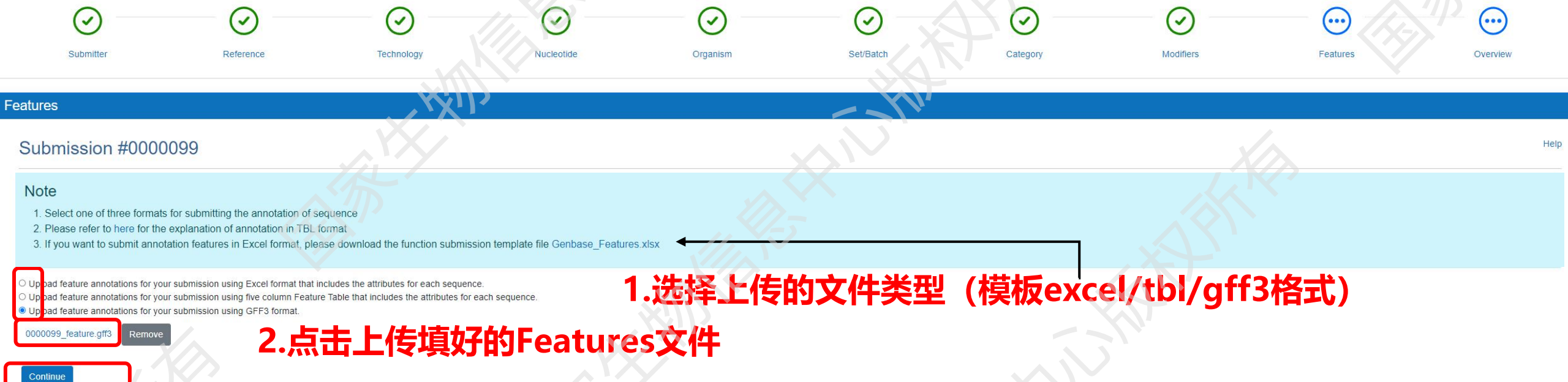
2.点击上传填好的Modifiers文件

Successful verification!

Continue 3.保存并下一步



提交流程： 步骤9 提交Features文件（Features）



1.选择上传的文件类型（模板excel/tbl/gff3格式）

2.点击上传填好的Features文件

3.保存并下一步

#如果此处不上传注释文件，最终归档后的GBF页面会标记：
GenBaase staff is unable to verify sequence and/or annotation provided by the submitter

提交流程： 步骤10 结果预览 (Overview)

1.确认并选填上述信息

2.点击最终提交，等待页面展示

提交流程： 步骤10 结果预览 (Overview)

Review Records of Your Set

Below please find your **1** GenBase submission record(s) for your review.

You can download the complete set as a compressed ZIP file.

Warnings: there are some warnings in your sequence file. We suggest you check the sequence file before you go ahead.Download whole warning file result_err.txt 点击可下载

Severity	Sequence Name	Error Title	Description
●	Luciola_filiformis	SEQ_FEAT.BadRRNAcomponentOverlapTRNA	tRNA-rRNA overlap FEATURE: tRNA: Leu [lcl|Luciola_filiformis:c12463-12402] [lcl|Luciola_filiformis: raw, dna len= 16864]
●	Luciola_filiformis	SEQ_FEAT.BadRRNAcomponentOverlapTRNA	tRNA-rRNA overlap FEATURE: tRNA: Val [lcl|Luciola_filiformis:c13787-13725] [lcl|Luciola_filiformis: raw, dna len= 16864]
●	Luciola_filiformis	SEQ_FEAT.TranslExceptPhase	transl_except qual out of frame. FEATURE: CDS: cytochrome c oxidase subunit III <11> [lcl|Luciola_filiformis:4654-5438] [Luciola_filiformis_6]

查看黄色警告，尽量修改

提交流程： 步骤10 结果预览 (Overview)

3.网页展示数据具体信息

```
LOCUS       G53DN2379VN1              2483 bp    RNA     linear   UNK 16-AUG-2022
DEFINITION  Human picobirnavirus.
ACCESSION   [redacted]
VERSION     G53DN2379VN1
KEYWORDS    .
SOURCE      Human picobirnavirus
  ORGANISM  Human picobirnavirus
            Riboviria; Orthornavirae; Pisuviricota; Duplopiviricetes;
            Durnavirales; Picobirnaviridae; Picobirnavirus.
REFERENCE   1  (bases 1 to 2483)
  AUTHORS   Bao,Y.
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2483)
  AUTHORS   Bao,Y.
  TITLE     Direct Submission
  JOURNAL   Submitted (16-AUG-2022) National Genomics Data Center, Beijing
            Institute of Genomics, Chinese Academy of Sciences, NO.1 Beichen
            West Road, Chaoyang District, Beijing 100101, China
COMMENT     ##Genome-Assembly-Data-START##
            Assembly Method       :: Trinity-2.11.0
            Sequencing Technology :: Other
            ##Genome-Assembly-Data-END##
FEATURES             Location/Qualifiers
     source          1..2483
                     /organism="Human picobirnavirus"
                     /mol_type="genomic RNA"
                     /host="None;gender: None;age: None"
                     /country="China: Beijing"
                     /collection_date="2018-08-01 00:00:00"
     CDS             156..701
                     /codon_start=1
                     /product="hypothetical protein"
                     /translation="MTANQLRYAELRESNRHNTTVEKETGRHNVETERIGYGTLAETSR
                     HNKAYEGIQSALAAETARHNVATEGINWYSAQNLANLQFAQANKTTSEVGETTRHNLAT
                     ESIQSGQLEETGRHNLATEQYNINALGEQFRHNKQTELQGWISTGAGAYKDVSTGLRQN
                     VESVKGIINGVKSILPGG"
     CDS             707..2476
                     /codon_start=1
                     /product="capsid protein"
                     /translation="MKDSRNIKGRNSKYQGKRVNKSRKSELDDQINERYSKDNDWRWYA
```

- 这里的定义行不是最终的
- 最终归档后会生成新的定义行，主要来源于元信息和特征信息，和序列文件定义行无关

等待审核

4.提交成功，等待审核

Submission	Title	Created By	Created Time	Update Time	Release Date	Status	
0000145	-	[redacted]	2022-10-26 22:34:47	2022-10-27 11:31:07	2022-10-27	Unfinished Modifiers	-
0000144	-	[redacted]	2022-10-26 15:25:21	2022-10-26 23:02:34	2022-12-31	Unfinished Modifiers	-
0000142	1	[redacted]	2022-10-24 18:39:43	2022-10-24 18:41:12	2022-10-24	Unfinished Organism	-
0000141	-	[redacted]	2022-10-24 15:34:55	2022-10-25 15:48:33	2026-10-24	Unfinished Features	-
0000140	[redacted]	[redacted]	2022-10-24 13:54:41	2022-10-24 15:42:45	2023-05-01	Unfinished Overview	-
0000139	-	[redacted]	2022-10-21 16:09:10	2022-10-21 16:14:54	2023-12-12	Unfinished Modifiers	-
0000138	Genome-Wide Analysis of the [redacted]	[redacted]	2022-10-21 15:57:39	2022-10-21 16:09:54	2023-10-03	Unfinished Features	-
0000137	-	[redacted]	2022-10-20 22:25:50	2022-10-21 08:59:39	2022-10-20	Unfinished Modifiers	-
0000136	XK10	[redacted]	2022-10-20 16:06:57	2022-10-20 19:42:26	2022-10-20	Unfinished Modifiers	-
0000135	-	[redacted]	2022-10-20 10:28:26	2022-10-21 16:01:55	2022-10-20	Meta Data Finished	-

Show 10 entries

审核结果与反馈

5.审核完毕

Submission	Title	Created By	Created Time	Update Time	Release Date	Status
0000117	-	[redacted]	2022-09-29 13:41:04	2022-10-09 14:02:00	2022-10-09	Unfinished Set/Batch
0000115	Rc_virus	[redacted]	2022-09-28 15:15:20	2022-10-19 15:34:11	2022-09-30	Audit Success Accession
0000114	RcALV-BeLV	[redacted]	2022-09-28 14:47:05	2022-10-19 15:33:55	2022-09-28	Audit Success Accession

无错，点击Accession查看编号

Submission	Title	Created By	Created Time	Update Time	Release Date	Status	
0000065	-	[redacted]	2022-08-19 18:17:50	2022-08-19 20:08:35	2022-08-19	Unfinished Features	-
0000064	-	[redacted]	2022-08-18 20:28:27	2022-08-18 20:28:27	-	Unfinished Reference	-
0000063	-	[redacted]	2022-08-18 08:44:06	2022-11-03 11:30:59	2022-08-18	Audit Fail View	-
0000062	-	[redacted]	2022-08-17 23:40:08	2022-10-19 18:50:07	2022-08-17	Audit Fail View	-

有错，点击View查看待修改内容

数据发布状态

1. 数据已经公开发布

根据Accession编号公开可查询

Submission	Title	Created By	Created Time	Update Time	Release Date	Status	
0000086	-	[redacted]	2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success Accession	public

点击查看Accession编号

0000202	[redacted]	[redacted]	2022-11-29 11:09:41	2022-11-30 10:35:22	2022-11-29	Audit Success Accession	confidential	Review link Modify Grant Modify ReleaseTime

2. 数据处于私密状态

只能自己点击查看

可点击修改发布日期

数据发布

Completed Submission(s)

Submission #0000089

Submitter Organization	Zhejiang University
Released Date	2022-08-26
Sequence author(s)	Xu Haijun,Zhang Jinli
Sequencing Technology	Sanger dideoxy sequencing
Molecule Type	mRNA (cDNA)

2 nucleotide sequence(s) have been submitted

The assigned accession listed as follows, which can be cited in the published paper.

seq1 C_AA001111.1 GBFF Fasta

seq2 C_AA001112.1 GBFF Fasta

点击可查看详细信息

```
LOCUS       C_AA001111              2622 bp    mRNA    linear   INV 31-AUG-2022
DEFINITION  Nilaparvata lugens Zinc finger homeodomain 1 (zfh1) isoform B (Zfh1)
            mRNA, complete cds.
ACCESSION   C_AA001111
VERSION     C_AA001111.1
KEYWORDS    .
SOURCE      Nilaparvata lugens (brown planthopper)
  ORGANISM  Nilaparvata lugens
            Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Protostomia;
            Ecdysozoa; Panarthropoda; Arthropoda; Mandibulata; Pancrustacea;
            Hexapoda; Insecta; Dicondylia; Pterygota; Neoptera; Paraneoptera;
            Hemiptera; Auchenorrhyncha; Fulgoromorpha; Fulgoroidea; Delphacidae;
            Delphacinae; Nilaparvata.
REFERENCE   1  (bases 1 to 2622)
  AUTHORS   Xu,H. and Zhang,J.
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2622)
  AUTHORS   Xu,H. and Zhang,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (31-AUG-2022) institute of insect sciences, Zhejiang
            University, Yuhangtang Road 866, Hangzhou 310058, China
FEATURES             Location/Qualifiers
     source          1..2622
                     /organism="Nilaparvata lugens"
                     /mol_type="mRNA"
                     /db_xref="taxon:108931"
                     /country="China"
     gene            <1..>2622
                     /gene="Zfh1"
     CDS             1..2622
                     /gene="Zfh1"
                     /codon_start=1
                     /product="Zinc finger homeodomain 1 (zfh1) isoform B"
```

已分配的Accession号（可以写到文章中!!!）

同时会发送Accession号到提交者邮件

```
>C_AA001111.1 Nilaparvata lugens Zinc finger homeodomain 1 (zfh1) isoform B (Zfh1) mRNA, complete cds
ATGTATTCATCGCCGACAGCTTCCTGCTACGATAGCTACTCGTACCGGCTGTGGAGTCTGGCAAACGTGT
GGACCACTTATCAATTGCCTTTACCCGGAGTCGTAAGCAGCGGTGGTAAGGAGGAGACGGGTGGCAACCC
CGGGACACAGGTGGCGTCCTCGCCGGACGACTACTTCATCAAATGCCCGCAGTGTGGCGCTGGCTACGCG
GGCTTCCAGGCGCTCAAGGAGTACGTGGAGTTGGCTCATGGAGGCGGTGCGCTCTCCTGTCTGCAGTGCA
CGGCTACCTTCCCCACGAGGGATCTACTCGACAAACATGAGCTGTTGCATTCGCCCAATGCGCAGGTGTC
ATGCAAAGTTTGCAACAAAACATTCGCCAATGTGTACAGACTTCAAAGGCACATGATAAGCCATGATGAA
AGTGCAAACCTTCGCAAATTCAAGTGTCCAGAATGTGAGAAGGCCTTCAAATTCAAGCATCACCTTAAGG
```

数据引用

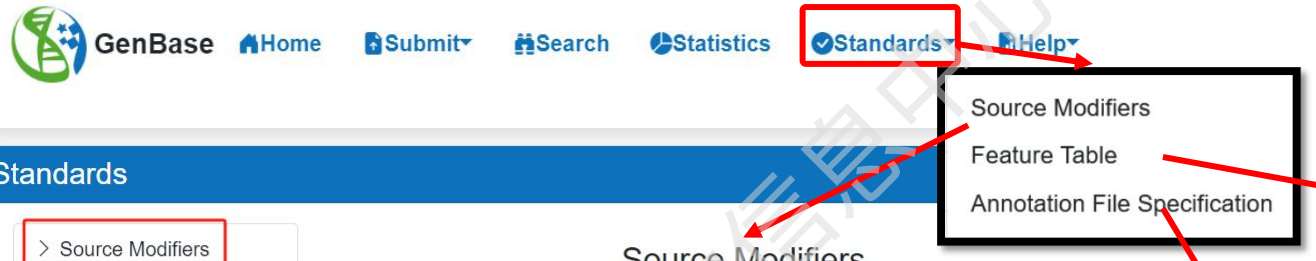
审稿人链接和资助信息

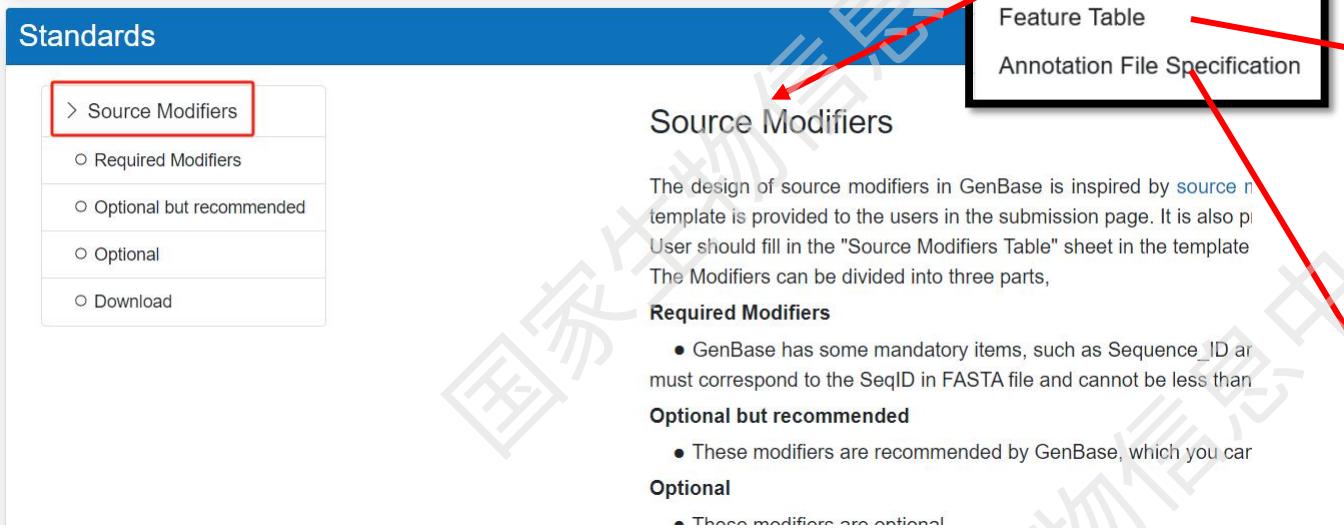
1. 点击Review link按钮
2. 生成Reviewer link

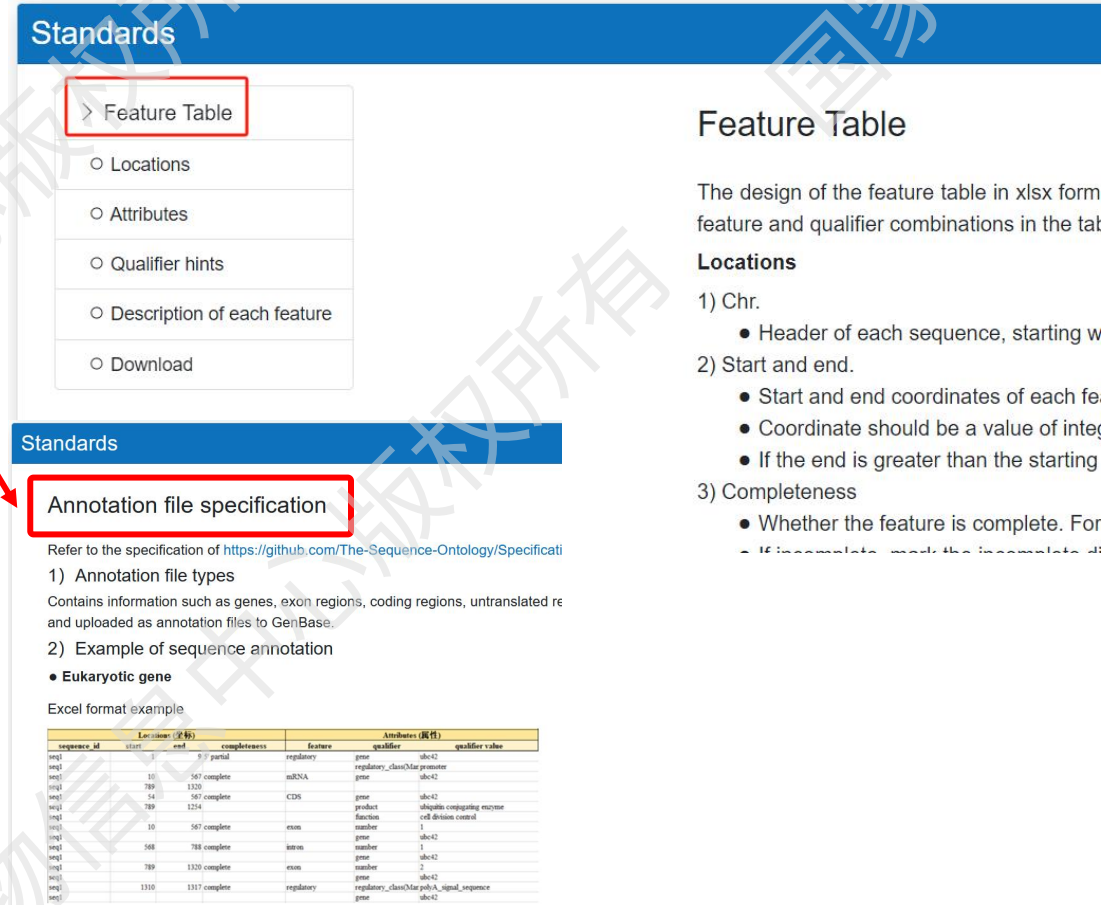
Submission	Title	Created By	Created Time	Update Time	Release Date	Status		User Operation
0000090	LsZfh1	[redacted]	2022-08-26 13:29:11	2022-09-02 15:29:06	2022-08-26	Audit Success	Accession public	https://ngdc.cncb.ac.cn/genbase/review/[redacted] Cancel Modify Grant
0000089	Nlzfh1,isoform B and isoform C	[redacted]	2022-08-26 12:34:44	2022-09-01 17:37:27	2022-08-26	Audit Success	Accession public	https://ngdc.cncb.ac.cn/genbase/review/[redacted] Cancel Modify Grant
0000086	-	[redacted]	2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success	Accession public	https://ngdc.cncb.ac.cn/genbase/review/[redacted] Cancel Modify Grant

3. 点击Modify Grant按钮
4. 可修改资助信息

GenBase标准页面







GenBase说明文档页面

CNCB NGDC Databases Tools Standards Publications

GenBase Home Submit Search Statistics Standards Help Login Register 语言/Language

GenBase Tutorials
GenBase Quality Control Errors
GenBase Sequence Update Tutorials
Data Control

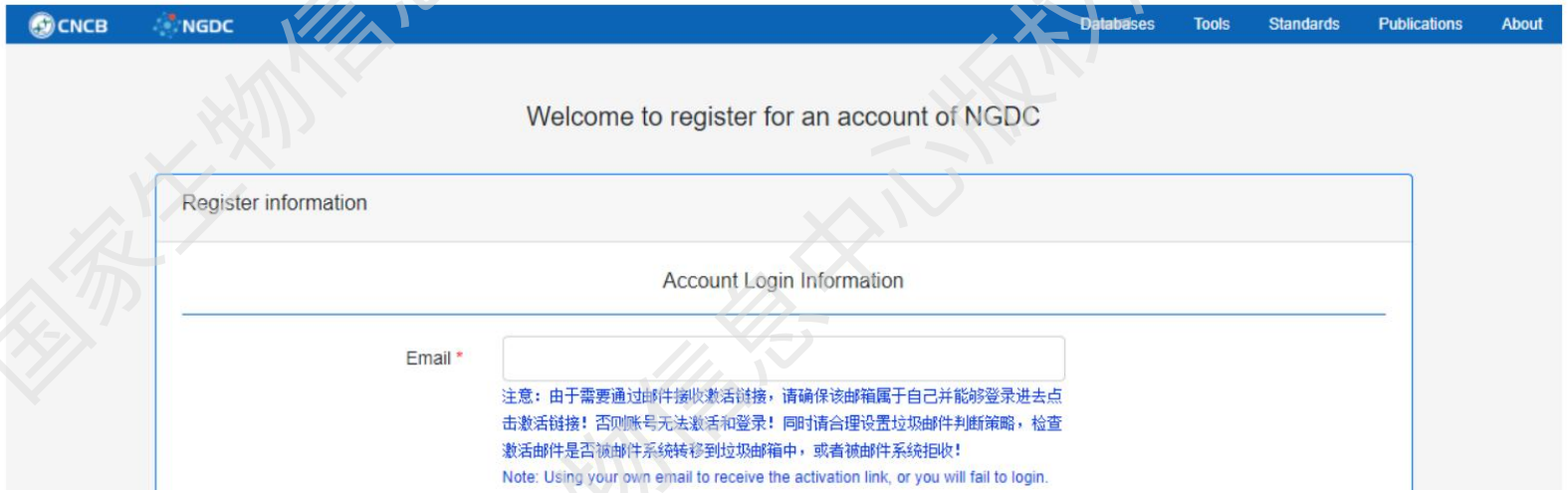
- Overview
- Login to the BIG Sub
- Create a GenBase Submission
  - Enter GenBase submission system
  - Prepare submission files
  - Create new GenBase submission
- Search
- Advanced Search
- Download
- Documentation for Submission
  - Chinese version

Overview

BIG Submission Portal (BIG Sub) offers a number of services through which data can be submitted to the NGDC. You can use this service to submit raw sequence reads, genome assemblies, nucleotide sequences, targeted assembled and annotated sequences and to register projects and samples. The GenBase stores nucleotide and protein sequences with annotations. Click here to see documentation for GenBase submission sections.

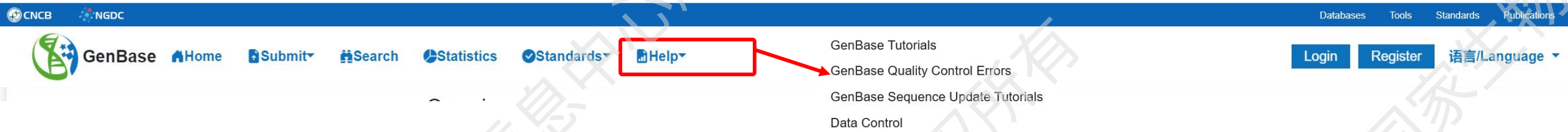
Login to the BIG Sub

Click the login button in BIG Sub, and then enter your user name and password to login. If you do not have an account already, click the Register button to create one. It is recommended to use the laboratory public mailbox for registration.
If you have used an account in the past but no longer see your previous submissions, please contact us at genbase@big.ac.cn for assistance with your account view.
– Recommend the use of Firefox/Google Chrome browser, other browsers may have bugs.
– After the activation of the login system, use our BIG Submission Portal (BIG Sub) and follow steps to finish the submission.



提交说明文档（PDF）

GenBase说明文档页面



Error types and correction schemes of GenBase quality control system

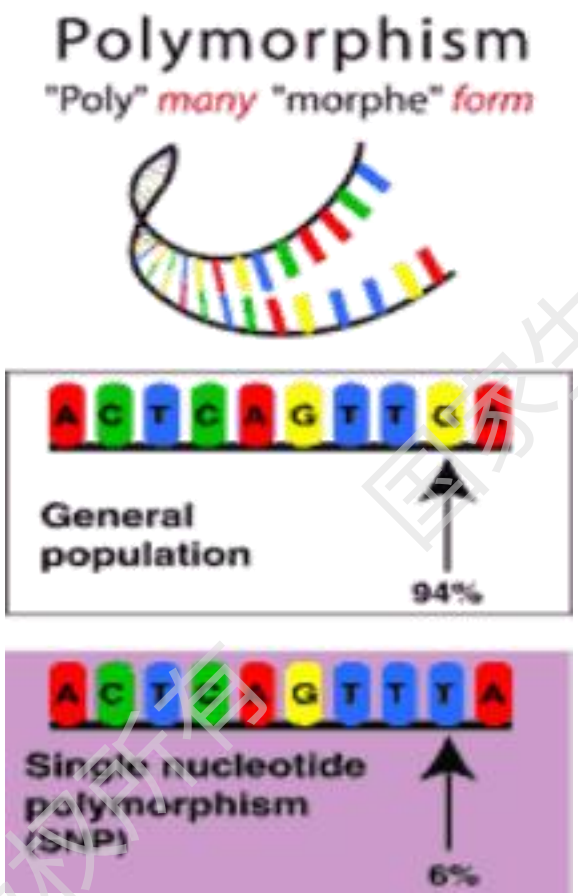
After the GenBase quality control is completed, it is necessary to make modifications based on the error messages in be encountered and the correction methods. Download PDF

SEQ_FEAT.InternalStop or SEQ_INST.StopInProtein

Explanation: The protein translation region of the sequence contains termination codons.
Suggestions:
1. Check if there is any mistake in the nucleic acid sequence and consider correcting it if necessary.
2. Check if the organelle type selection is correct. Chloroplasts, mitochondria, and nuclear genomes use different gen termination codons. If so, consider correcting the organelle type.
3. Check if there are CDS intervals, i.e., the CDS is composed of multiple discontinuous segments, and the terminatic annotate the positions of each individual CDS segment.
4. Check if there is a change in codon_start, which refers to the phase of reading frame shifting or the position of the r CDS is incomplete (partial). If so, specify the incomplete position of the 5' end and add a codon_start qualifier to indic codon_start is 1 or 2 to indicate that the reading frame is shifted by one or two nucleotides; for TBL files, the value of
5. Check if it is a pseudogene. If it is determined to be a pseudogene and there is no termination codon, add pseudo= GFF file to indicate that this sequence cannot be translated normally.
6. Check if it is a special gene coding mechanism, such as the phenomenon of incomplete termination codons often f the CDS sequence and clearly indicate the specific position and corresponding amino acid information.

目录

基因组变异



图片来源：genome news network

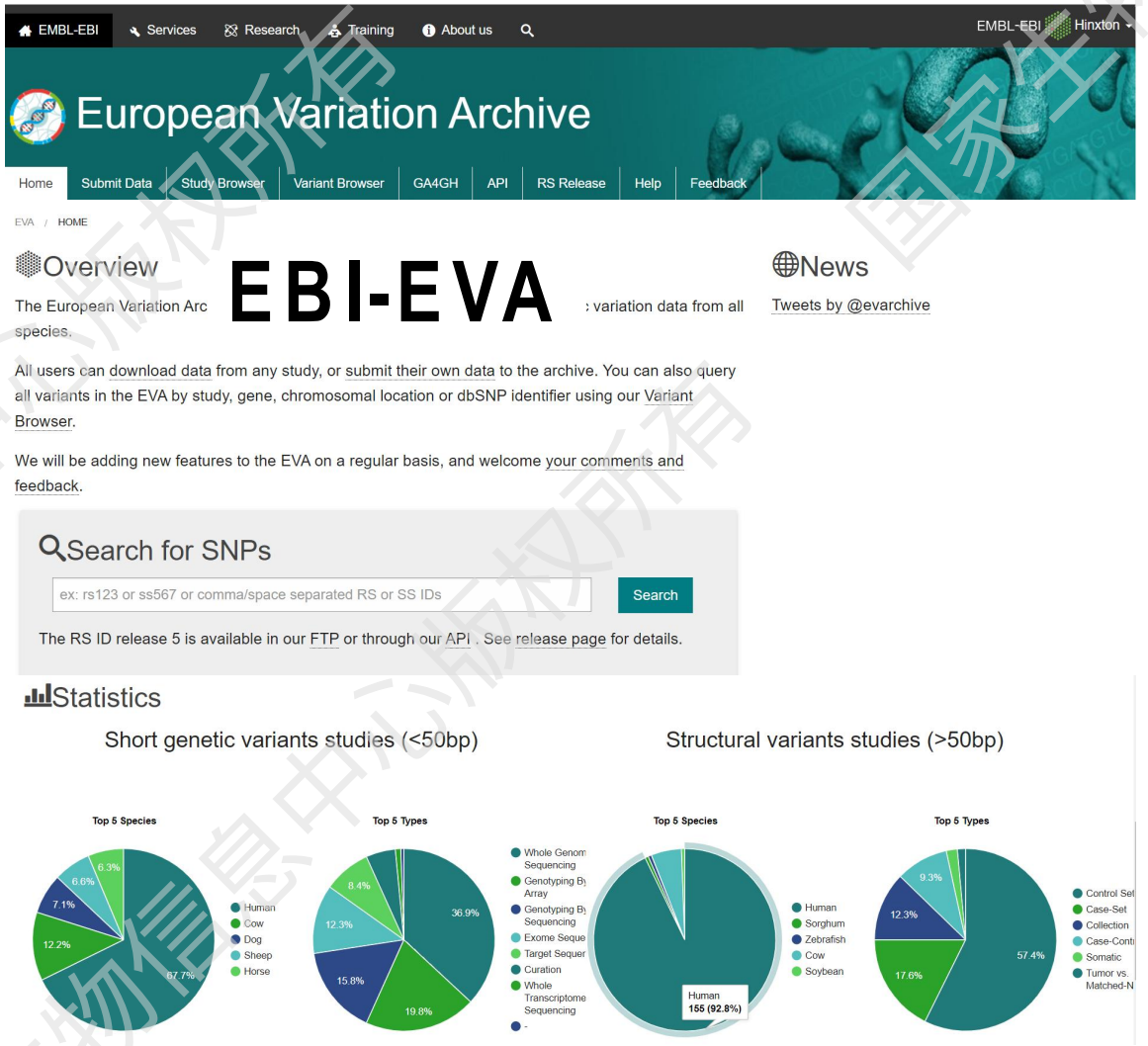
单核苷酸多态性，英文single nucleotide polymorphism，缩写为SNP。SNP主要是指在基因组水平上引起的单个碱基的变异，其在群体中的发生频率不小于1%，包括单碱基的转换、颠换以及单碱基的插入和缺失等。

基因组结构性变异，英文Structure Variations，简称SVs，通常指基因组上大长度的序列变化和位置关系变化。

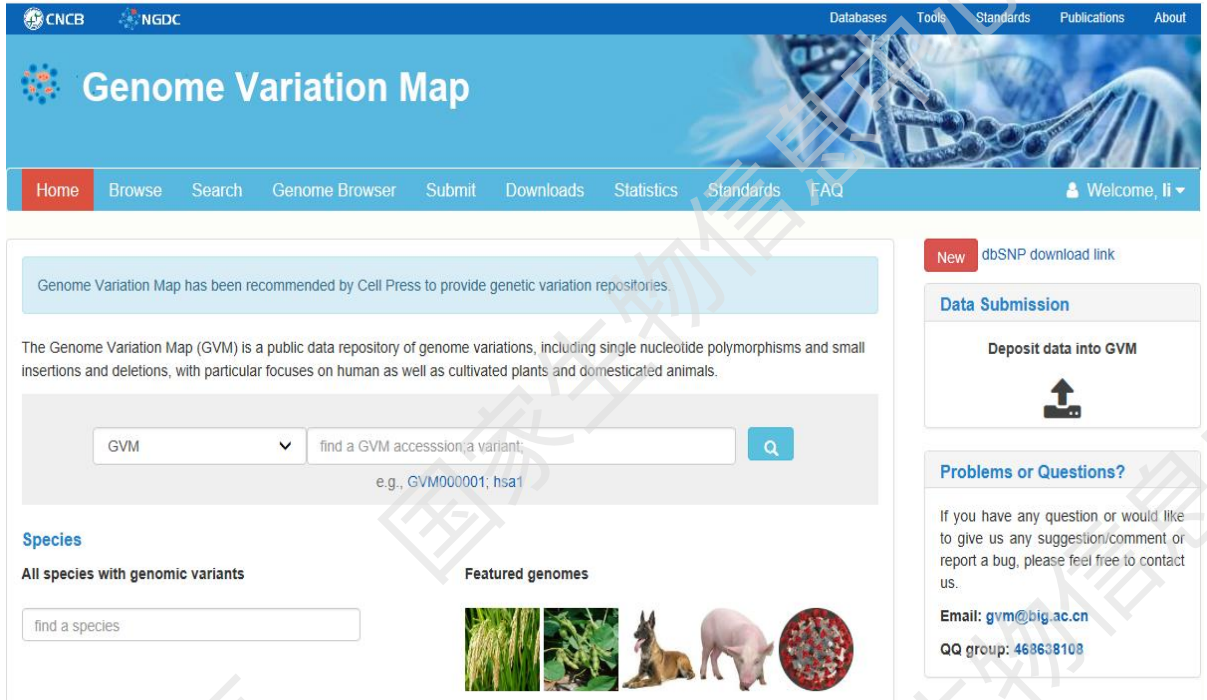
- 物种进化研究
- 表型差异研究
- 人类疾病研究
- 药物研发及精准医疗
- 农业遗传改良育种
- 生物多样性保护

国际变异资源库


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
dbSNP
SNP
Search
Advanced
Help
NCBI-dbSNP/dbVar
Getting Started
dbSNP 20th Anniversary
Overview of dbSNP
About Reference SNP (rs)
Factsheet
Entrez Updates (May 26, 2020)
Submission
How to Submit
Hold Until Published (HUP) Policies
Submission Search
Access Data
Web Search
eUtils API
Variation Services
FTP Download
Tutorials on GitHub
How to Search dbSNP. Additional search terms are here.
All of dbSNP (then use filters on results page)
all[sb]
dbSNP RefSNP ID
Single: 328; Multiple 328,226,200
Gene
Gene symbol PTEN[Gene Name] or gene ID 4023[Gene_ID]
Genomic location of a single position or range on GRCh38. See the announcement and the guide for using GRCh37 coordinates
6[Chromosome] AND (1500000:3000000[Base Position])
Clinical significance
"pathogenic"[Clinical Significance]) OR "likely pathogenic"[Clinical Significance]
Global or study-wide minor allele frequency (GMAF) of a single frequency or range
(Note the required zero padding frequency as shown in example for 0.001 and 0.01)
00000.0010:00000.0100[GLOBAL_MAF]
dbSNP News and Announcements
NCBI Insights
RSS Feed dbSNP News and Announcements(RSS) Feed
Email List
YouTube
NCBI Minute: ALFA Webinar
Accessing Population Allele Frequency
SPDI and Variation Service
Variation Viewer
Variation Databases
dbVar
dbGaP
ClinVar
GTR


EMBL-EBI
Services
Research
Training
About us
EMBL-EBI Hinxton
European Variation Archive
Home
Submit Data
Study Browser
Variant Browser
GA4GH
API
RS Release
Help
Feedback
EVA / HOME
Overview
EBI-EVA
The European Variation Arc : variation data from all species.
All users can download data from any study, or submit their own data to the archive. You can also query all variants in the EVA by study, gene, chromosomal location or dbSNP identifier using our Variant Browser.
We will be adding new features to the EVA on a regular basis, and welcome your comments and feedback.
News
Tweets by @evarchive
Search for SNPs
ex: rs123 or ss567 or comma/space separated RS or SS IDs
Search
The RS ID release 5 is available in our FTP or through our API . See release page for details.
Statistics
Short genetic variants studies (<50bp)
Structural variants studies (>50bp)
Top 5 Species
Top 5 Types
Top 5 Species
Top 5 Types
Human
Cow
Dog
Sheep
Horse
Human
155 (92.8%)
Human
Sorghum
Zebrafish
Cow
Soybean
Control Set
Case-Set
Collection
Somatic
Tumor vs. Matched-N

基因组变异数据库GVM



- 我国最大变异资源库，采用标准分析方法鉴定的物种参考变异数据集及功能注释信息
- 我国首个基因组变异数据汇交、管理与共享平台
- 2017年上线，访客数2.7万，累积下载3500+万次

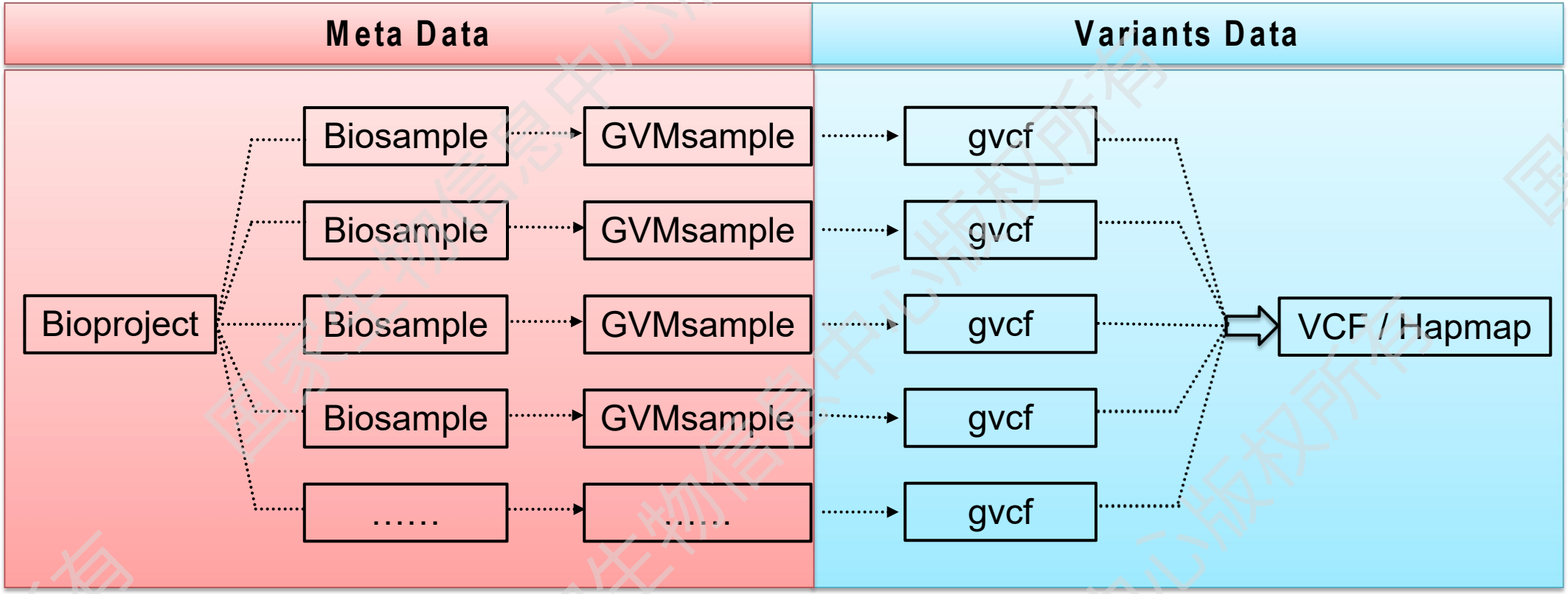
同行评价：

Daniel J. Rigden, *Editor in NAR*: *"A **major new resource** is the Genome Variation Map (58) from the BIG Data Center covering 19 species. **Its arrival is particularly timely** with the announcement that comparable NCBI resources dbSNP and dbVar are to stop accepting non-human."*

Thomas M. Keane, Team leader of EVA: "The EVA, the NCBI-based dbSNP and dbVar databases, and the Chinese Genome Variation Map resource form **a worldwide network for** brokering submissions, assigning permanent study and locus identifiers and exchanging these identifiers to ensure that consistent data is available at all sites."

Nucleic Acids Research (2018，2021) https://ngdc.cncb.ac.cn/gvm/

数据结构



- 项目信息
- 样本采集信息
- 样本变异分析参数等信息

在线填写

- 实体数据文件
- 支持VCF、GVCF、Hapmap等数据格式

FTP上传

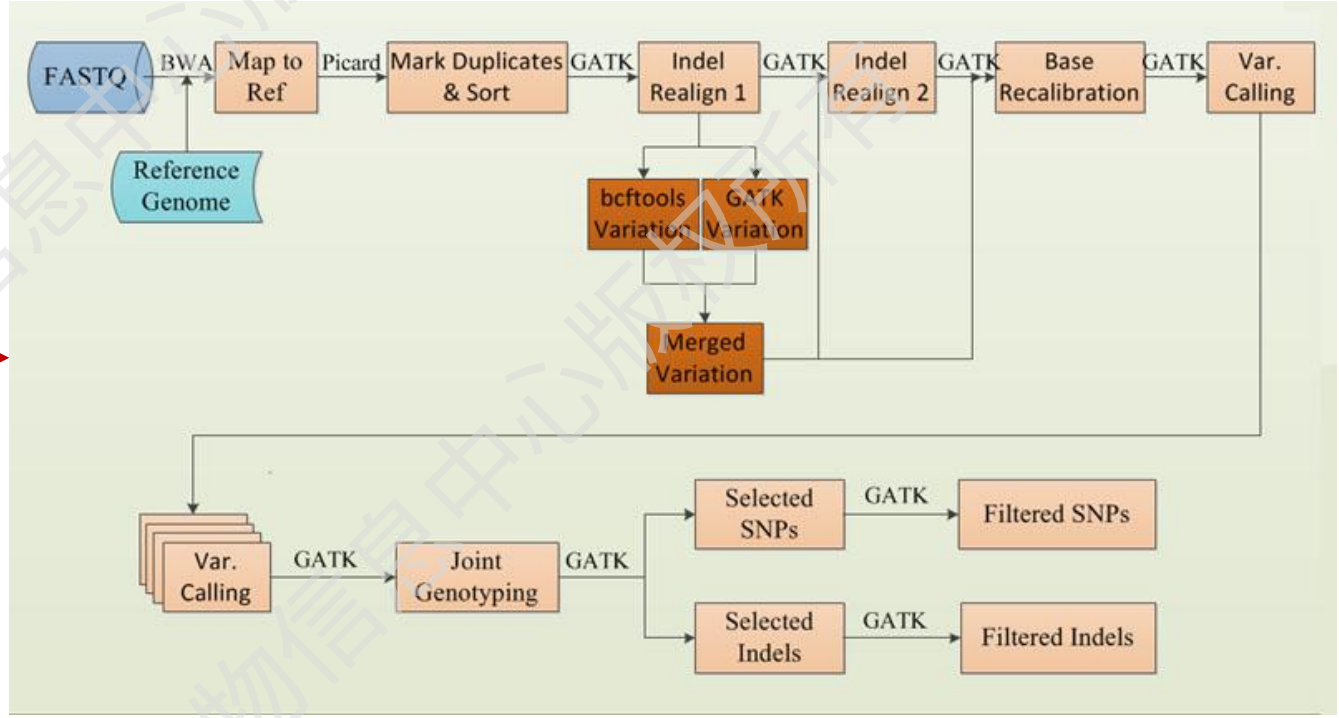
审编数据分析标准

GVM / Standards

Variome Data Standards

Version 2.0 beta, 9 September 2020

The National Genomic Data Center (NGDC), working collaboratively with multiple partner institutions/laboratories, develops the **Variome Data Standards** for variation data representation, analysis, search and exchange. We suggest submitters or users to follow these standards in preparing the metadata data, files and data analysis. We also give a detailed description on nomenclature standards in GVM about variation ID, annotation nomenclature et.al. These standards are our version 2.0 beta; please feel free to contact us (gvm AT big.ac.cn) if you have any suggestions.

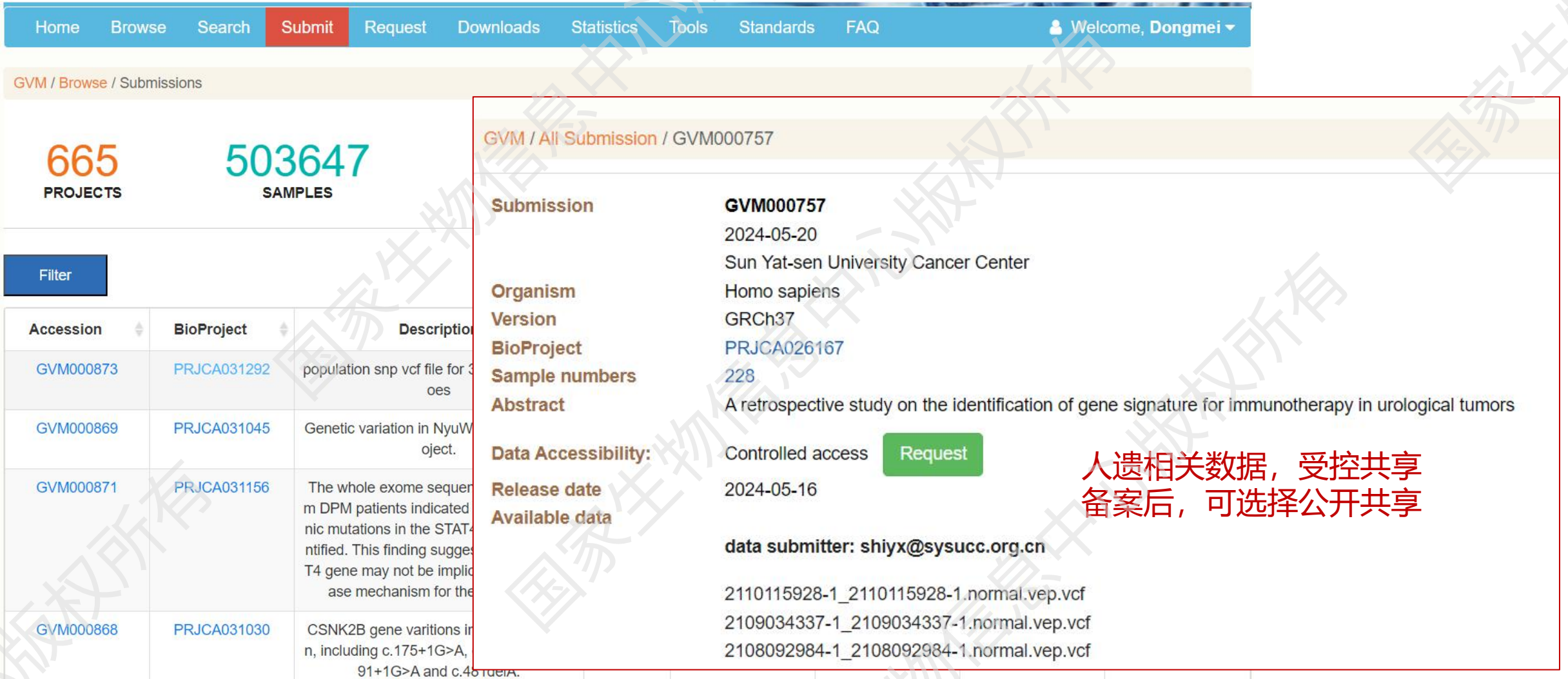


https://ngdc.cncb.ac.cn/gvm/analysis_standards

数据递交与管理

中英文帮助文档，仔细阅读，减少错误率！

数据递交与管理



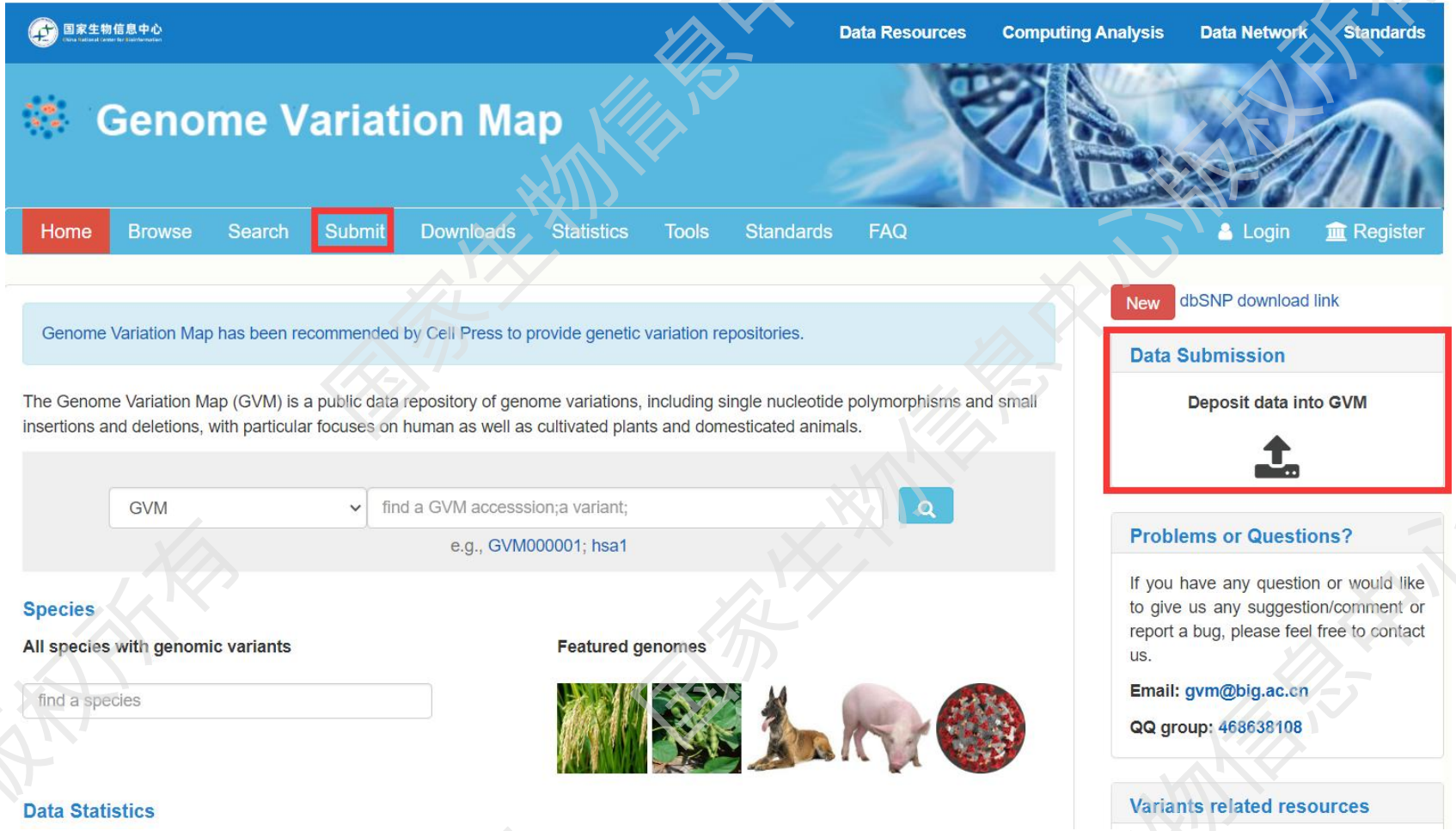
人遗相关数据，受控共享
备案后，可选择公开共享

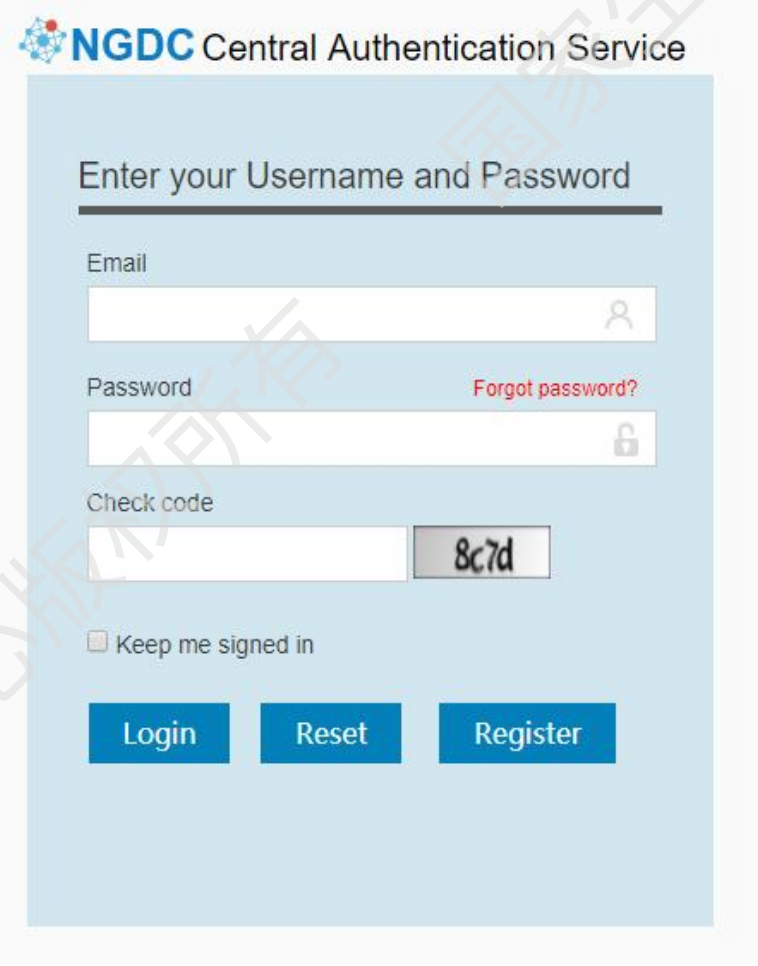
数据提交入口

- 入口URL： https://ngdc.cncb.ac.cn/gsub

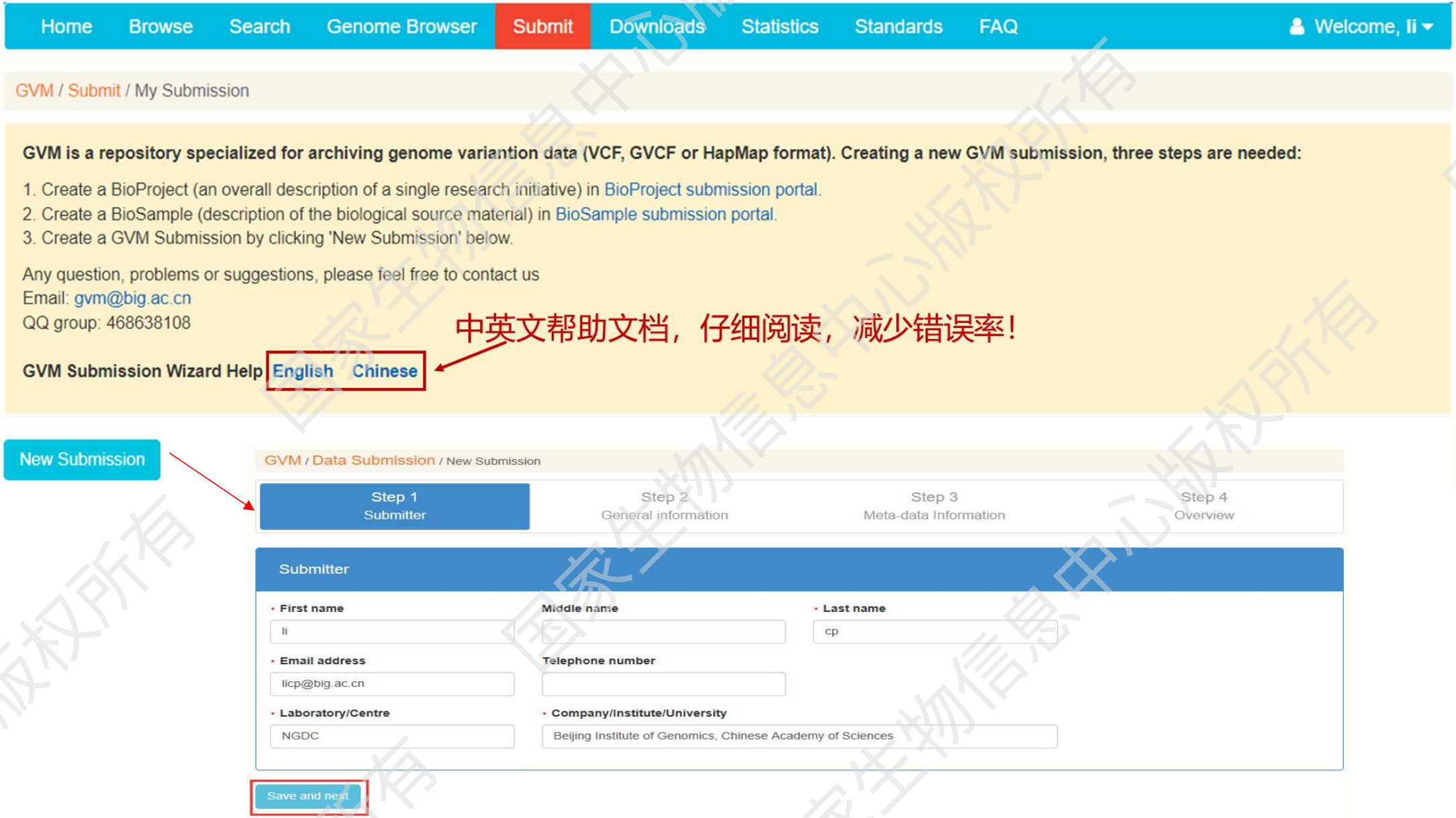
数据提交入口

- 入口URL： https://ngdc.cncb.ac.cn/gvm

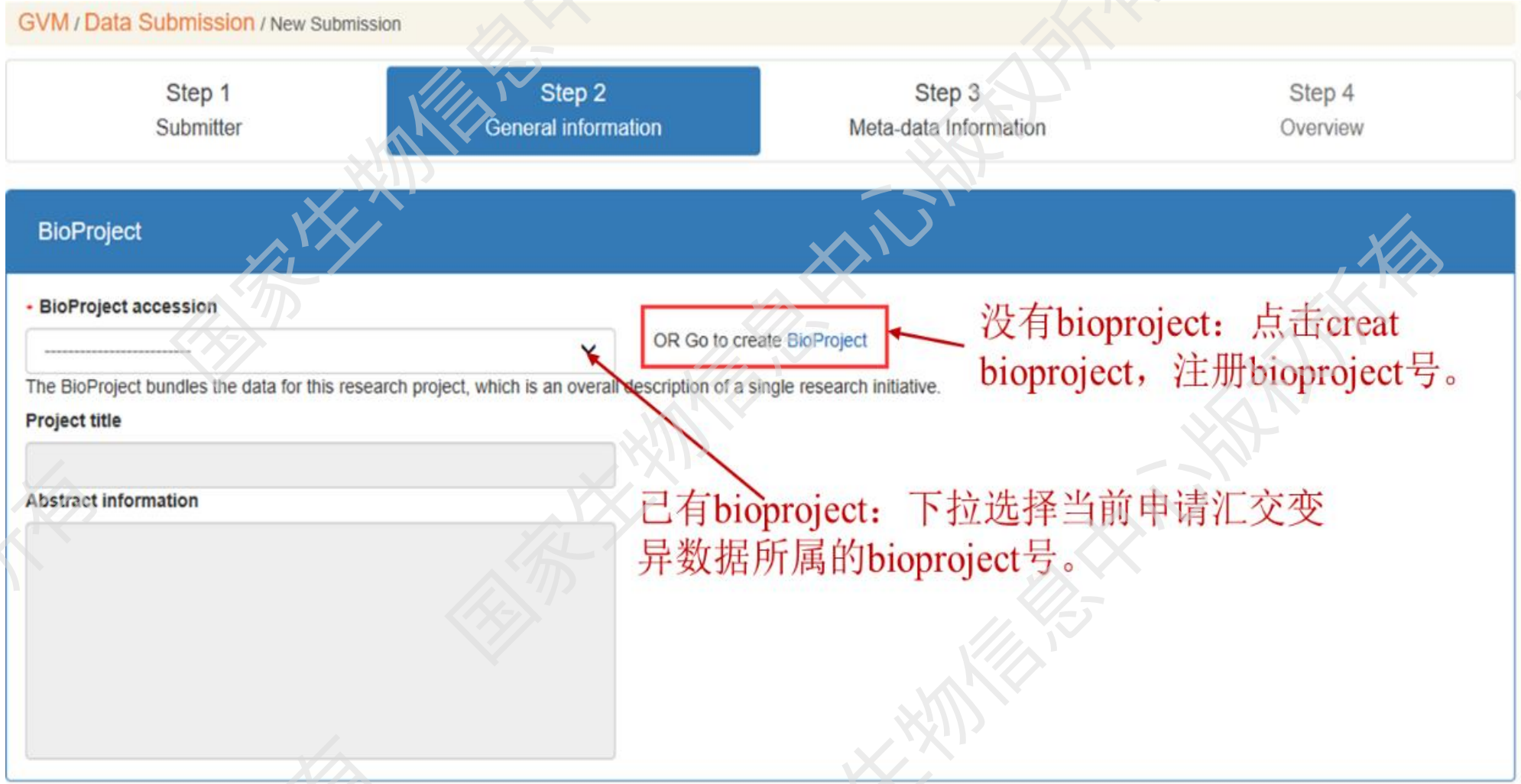




在线填写汇交申请



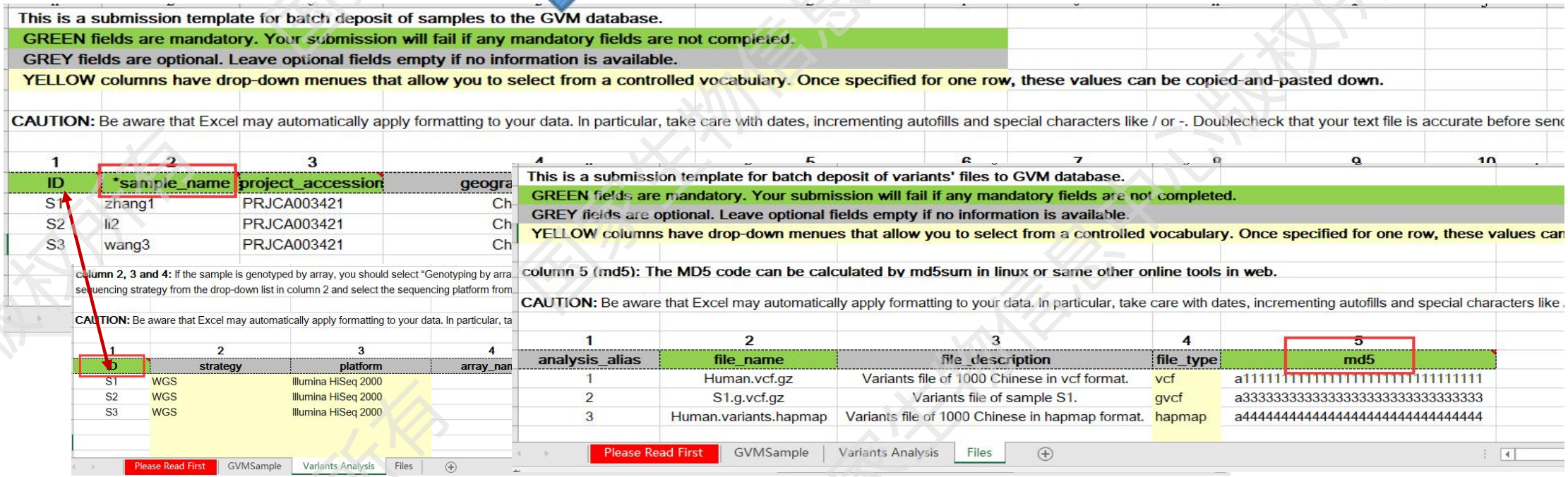
项目信息填写



样本信息填写

样本信息填写


GVM / Data Submission / New Submission
Step 1
Submitter
Step 2
General information
Step 3
Meta-data Information
Step 4
Overview
GVM Meta-data Information
BioProject: PRJCA003421
* Please upload GVM template file
Select file...
Browse ...
Download the GVM Submission Template
下载离线模板

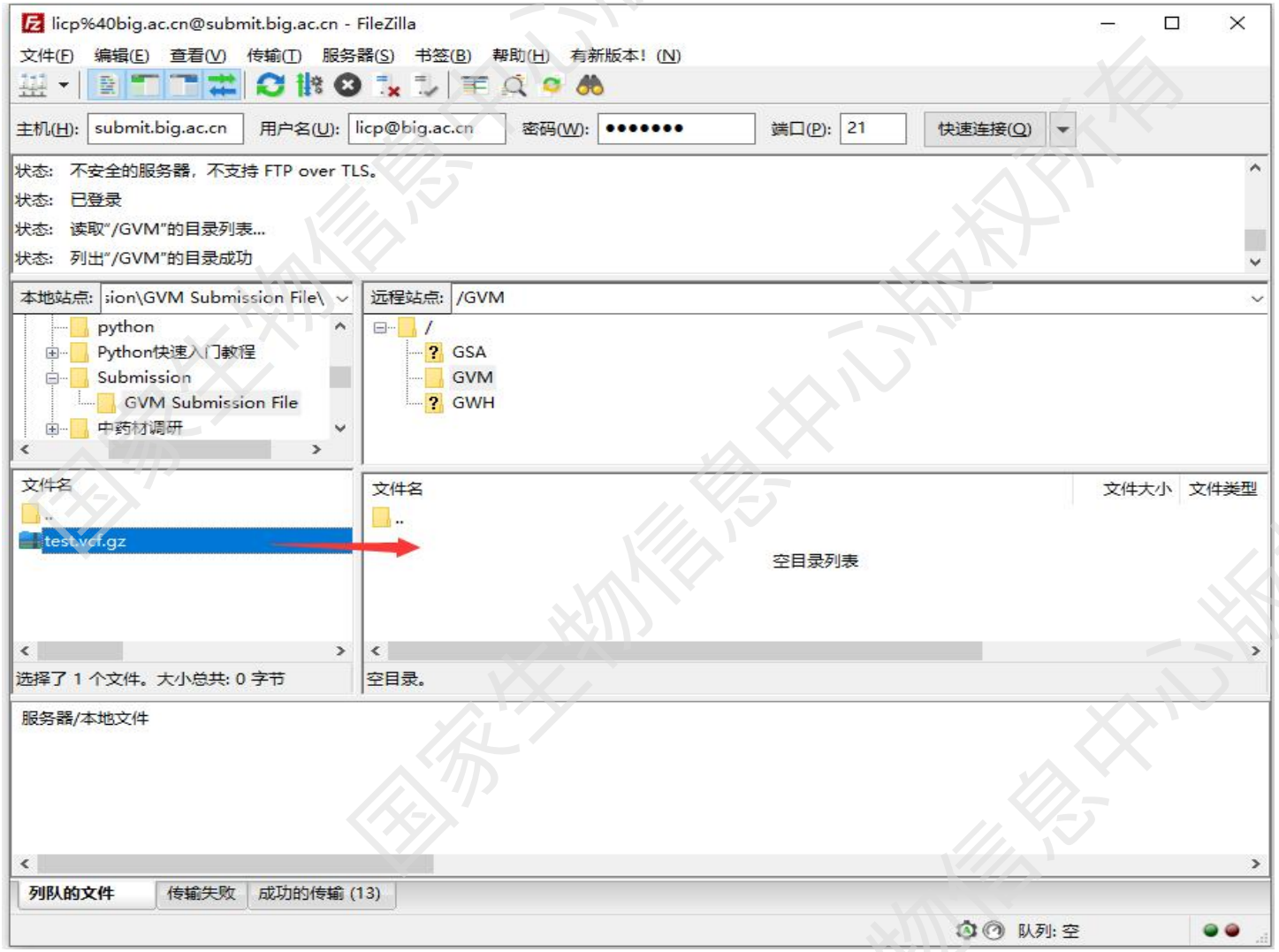

This is a submission template for batch deposit of samples to the GVM database.
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.
GREY fields are optional. Leave optional fields empty if no information is available.
YELLOW columns have drop-down menues that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.
CAUTION: 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 send
ID
*sample_name
project_accession
S1 zhang1 PRJCA003421
S2 li2 PRJCA003421
S3 wang3 PRJCA003421
column 2, 3 and 4: If the sample is genotyped by array, you should select "Genotyping by array
sequencing strategy from the drop-down list in column 2 and select the sequencing platform from
CAUTION: Be aware that Excel may automatically apply formatting to your data. In particular, ta
ID
strategy
platform
S1 WGS Illumina HiSeq 2000
S2 WGS Illumina HiSeq 2000
S3 WGS Illumina HiSeq 2000
Please Read First
GVMSample
Variants Analysis
Files
This is a submission template for batch deposit of variants' files to GVM database.
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
column 5 (md5): The MD5 code can be calculated by md5sum in linux or same other online tools in web.
CAUTION: Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like
analysis_alias
file_name
file_description
file_type
md5
1 Human.vcf.gz Variants file of 1000 Chinese in vcf format. vcf a11111111111111111111111111111111
2 S1.g.vcf.gz Variants file of sample S1. gvcf a33333333333333333333333333333333
3 Human.variants.hapmap Variants file of 1000 Chinese in hapmap format. hapmap a44444444444444444444444444444444
Please Read First
GVMSample
Variants Analysis
Files

样本信息填写

元信息填写

点击链接，下载文件上传工具 FileZilla
https://filezilla-project.org/download.php?type=client

变异数据上传



FileZilla

等待审核

New Submission

Items 1 - 10 of **177** 10 Items per page First Prev 1 of 18 Next Last GOTO

Accession	Submission ID	Project	Sample	Release Date	Create Time	Create By	Update Time	Status		User Operation
Unassigned	sub000254	PRJCA003421	SAMC237135,SAMC237136 more	2020-09-30	2020-09-09 20:06:24	licp@big.ac.cn	2020-09-09 20:47:09	Meta-data Finished	Confidential	

审核未通过，修改完善

Accession	Submission ID	Project	Sample	Release Date	Create Time	Create By	Update Time	Status		User Operation
Unassigned	sub000853	PRJCA013494	SAMC1006538,SAMC10065 more	2022-11-26	2022-11-26 14:53:26	zhiliulin@sina.com	2022-11-26 17:12:10	Checked Failed view	Confidential	delete modify

等待审核

New Submission

Items 1 - 10 of 177 | 10 ▾ Items per page | First | Prev | 1 | of 18 | Next | Last | GOTO

Accession	Submission ID	Project	Sample	Release Date	Create Time	Create By	Update Time	Status		User Operation
Unassigned	sub000254	PRJCA003421	SAMC237135,SAMC237136 more	2020-09-30	2020-09-09 20:06:24	licp@big.ac.cn	2020-09-09 20:47:09	Meta-data Finished	Confidential	

审核通过，获得编号

未公开数据，共享链接

Accession	Submission ID	Project	Sample	Release Date	Create Time	Create By	Update Time	Status		User Operation
GVM000437	sub000836	PRJCA011164	SAMC991091,SAMC991092 more	2025-11-02	2022-11-17 09:57:57	sabr@vip.163.com	2022-11-18 15:51:13	Checked OK	Confidential	share

数据库引用

- **引用示例：**

The variation data reported in this paper have been deposited in the Genome Variation Map (GVM) [1] in National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences and China National Center for Bioinformation, under accession number GVMXXXXXX.

- **References:**

[1] Genome Variation Map: a worldwide collection of genome variations across multiple species. *Nucleic Acids Res* 2021, 49(D1):D1186-D1191. [PMID=33170268].

国家生物信息中心

China National Center for Bioinformation

欢迎访问和提交数据到国家生物信息中心

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

GenBase-QQ交流群

GWH-QQ交流群

GVM-QQ交流群