



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

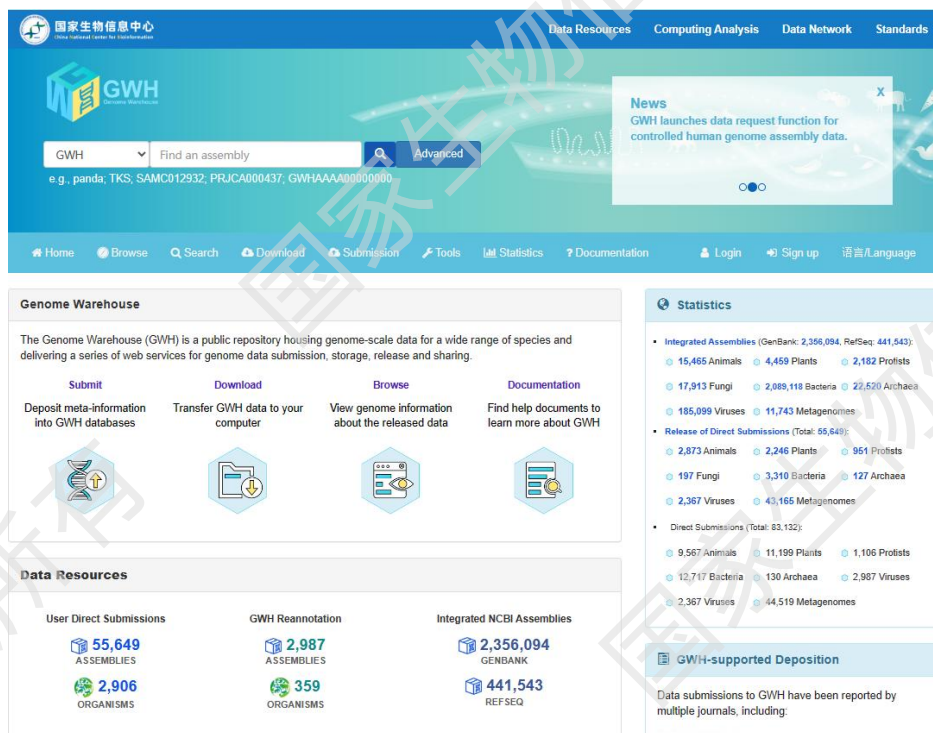
2025-11-20

目录

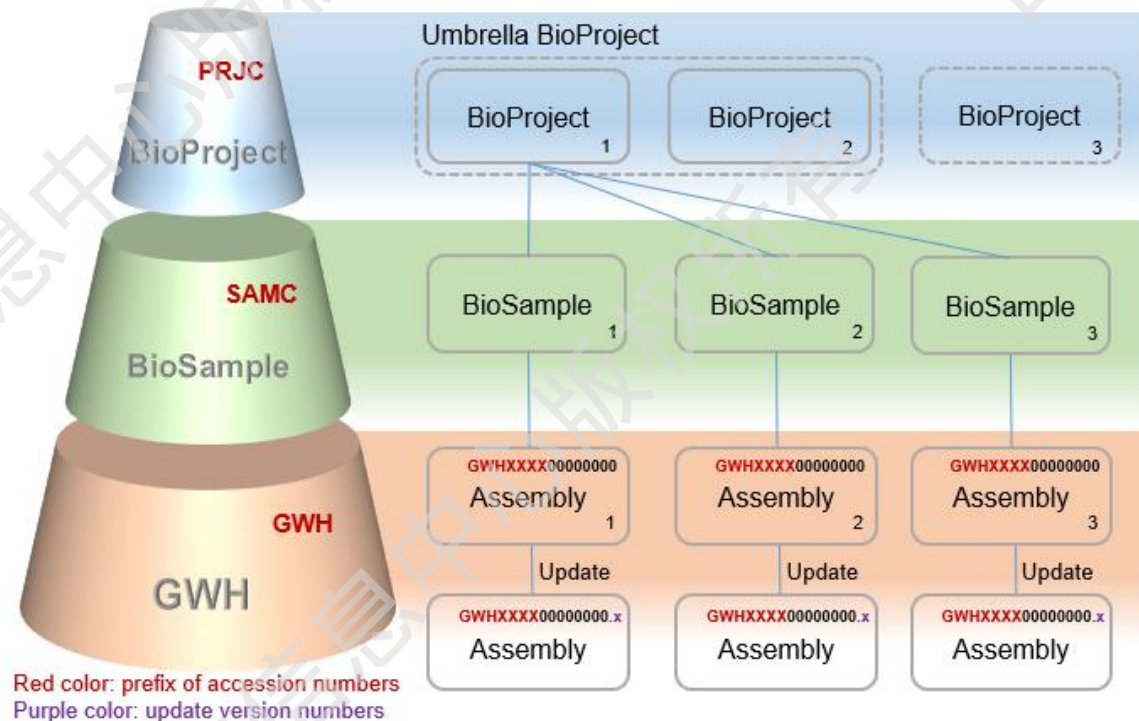
1. 基因组数据库GWH
2. 基因序列数据库GenBase
3. 基因组变异数据库GVM

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/>



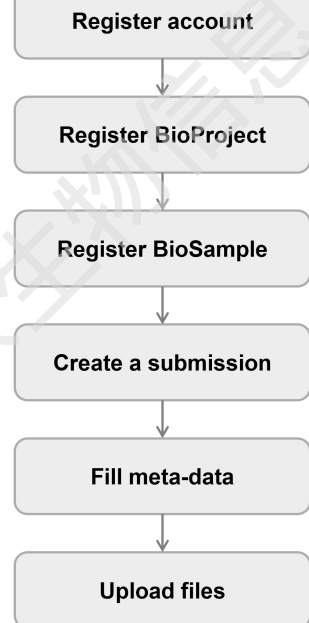
参考INSDC对基因组组装数据进行结构化的组织和管理

完善的基因组数据管理平台

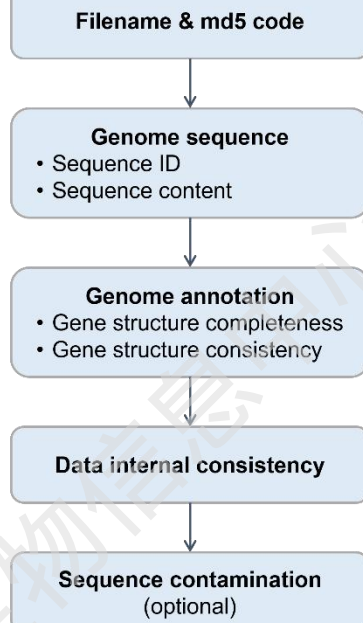
Genome Data Standards

- 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 General 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

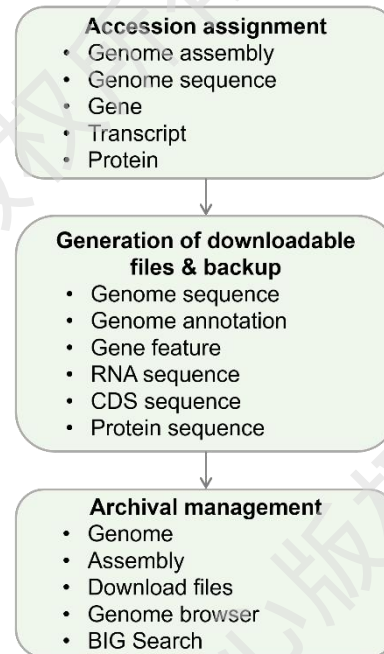
Data submission



Quality Control



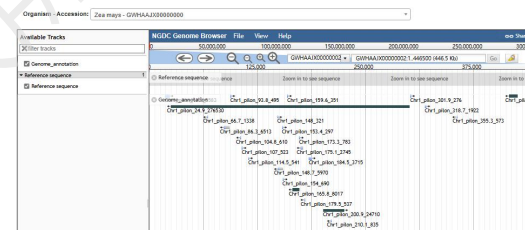
Archive



Release



Visualization



汇交

存储

质控

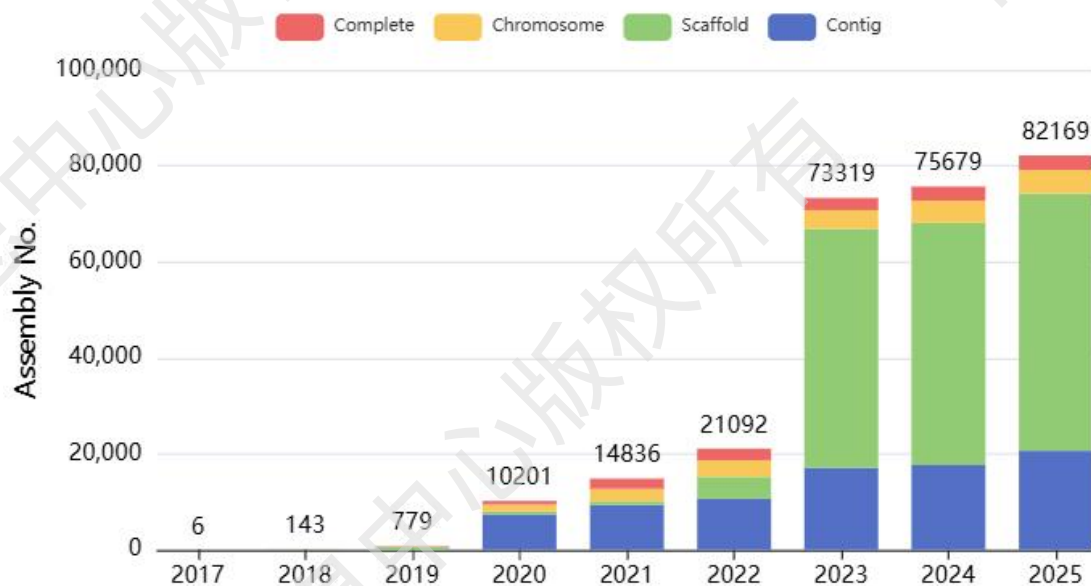
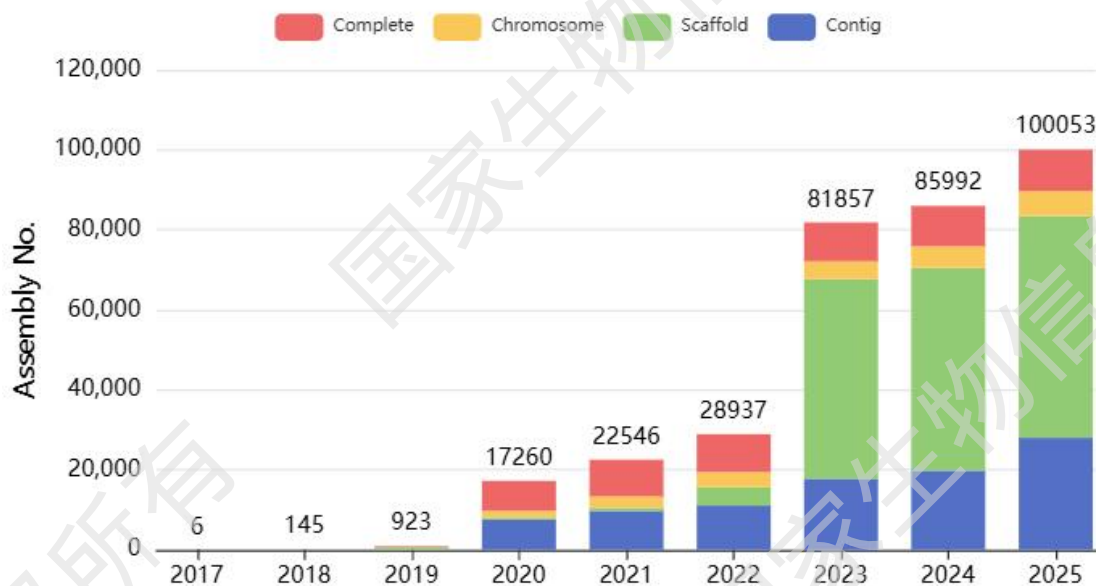
归档

发布

共享

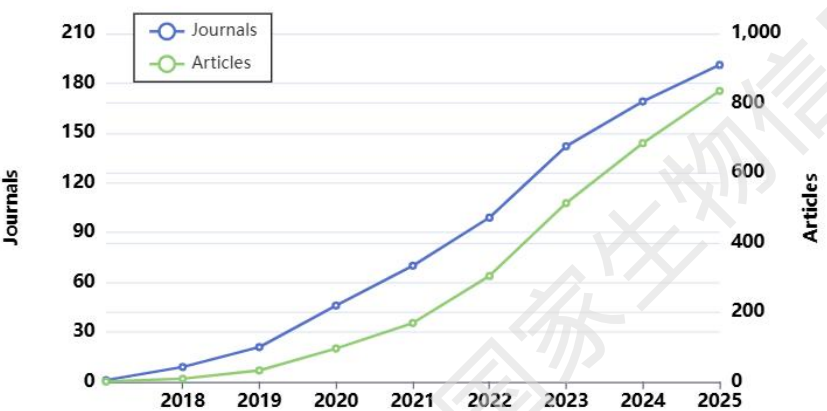
实现国内基因组组装数据统一汇交和安全管理

- 2,123 BioProjects
- 100,053 Genome assemblies
- 284,030,093 Sequences
- 58,276 BioSamples
- 6,843 Organisms
- 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	GWHSEEP000000000 GWHSEEQ000000000 GWHESER000000000 more	PRJCA019343
3	BMC Infectious Diseases	Whole-genome sequencing for analyzing the transmission characteristics of drug-resistant Mycobacterium tuberculosis in Ganzhou, China	2025-07	GWHFQAW000000000.1 GWHFQAX000000000.1 GWHFQAY000000000.1 more	PRJCA036393
4	Plant Biotechnology Journal	Chromosome-scale haplotype-resolved genome assembly of the autotetraploid alfalfa cultivar Bolivia	2025-07	GWHFIBZ000000000.1 GWHFICL000000000.1	PRJCA033031
5	BMC Bioinformatics	Blastn2dotplots: multiple dot-plot visualizer for genome comparisons	2025-06	GWHBDNP000000000.1	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	GWHAOSJ000000000	PRJCA003150
7	iScience	Functional proteomic insights into the lateral oviduct-associated secretions of a bean bug	2025-06	GWHBAZH000000000	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	GWHAZTZ000000000 GWHBAUV000000000	PRJCA003382
9	BMC Genomics	Comparative transcriptomic and proteomic analyses of hypoxia response in wild and cultivated tomato roots	2025-06	GWHBJTH000000000	PRJCA008297
10	Horticulture Research	A nearly complete haplotype-phased genome assembly of nerve plant (Fittonia albivenis) provides insights into leaf color evolution	2025-06	GWHEUVN000000000.1	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

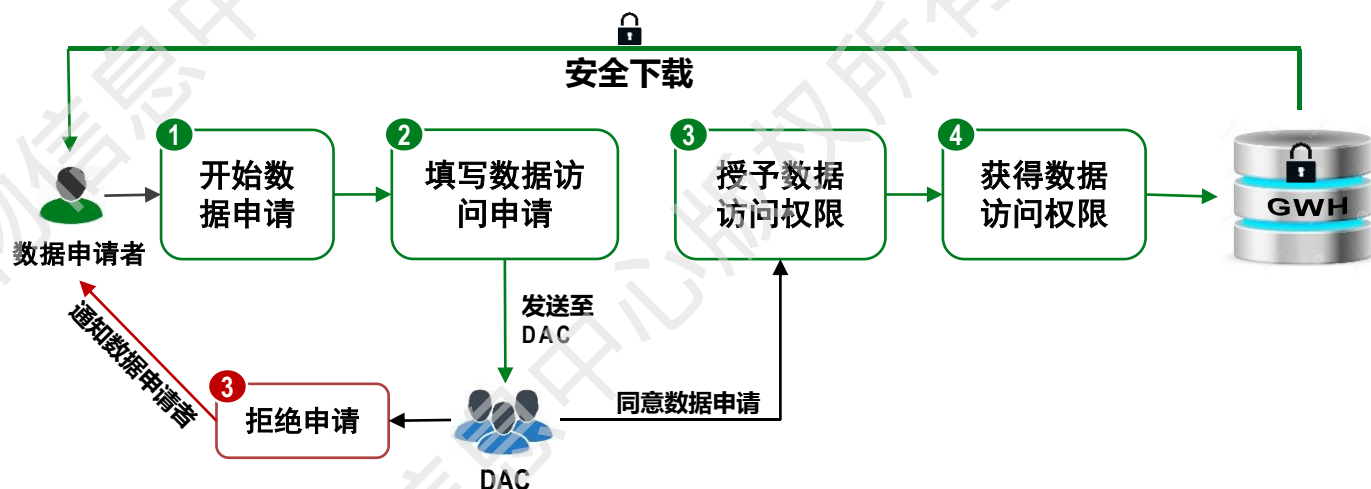
- 受控访问数据，采用申请审核制

- 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

CNCB NGDC Databases Tools Standards Publications

NGDC Central Authentication Service

Enter your Username and Password

Email [Forgot activate?](#) **用户名**
zhaoxuetong@big.ac.cn

Password [Forgot password?](#) **密码**
.....

Check code **验证码**
x4ab

☐ Keep me signed in

Login **Reset** **Register**

点击此处 登陆

Nsti
国家科技资源共享服务平台
科技资源共享网账号登录

Central Authentication Service

- BioProject
- BioSample
- BioCloud
- BioCode
- GenBase
- Database Commons
- GSA for Human
- Genome Sequence Archive (GSA)
- Genome Warehouse (GWH)
- Genome Variation Map (GVM)
- Open Archive for Miscellaneous Data (OMIX)
- Scientific Data Archive System (SDAS)

- 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

1	#The file is the meta information table for submitting "Genome assembly" ass												
2	#The sheet should be filled after created related BioProject and BioSample r												
3	#The "Green column" is required												
4	#The "Blue column" is required for certain conditions												
5	#The "Gray co												
6	#The "Yellow												
7	Version: 1.1												
8													
9	1												
10	* Biosample Accession	* Assembly name	An alias name for the corresponding genome assembly accession number. Only allow to enter the uppercase and lowercase letters, numbers, and underscores of English letters e.g., hg_CH_han When the submitted data is metagenomic, this information should reflect the number/name of	Sequencing	* Genome composition	* Assembly level	* Genome sequence file name	* Genome sequence file MD5 code	Genome annotation file name	Genome annotation file MD5 code	# Reference accession		
11													
12													
13													
14													
15													
16													
17													

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：第七行是版本信息，注意要求**最新版本**

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：“Circular=true”表示该条序列为可环化（序列首尾可含gap），“Circular=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

基因组数据提交准备

GWH接受全基因组组装数据（可包括/不包括细胞器基因组/质粒基因组、病毒基因组和基因片段序列数据，请提交到NGDC的 **GenBase** 数据库。您可以在一次批量提交中同时提交与同一个BioProject和文章相关的所有基因组组装数据。

● 提交流程说明（收起）

(1) 提交元数据:

- 创建新提交;
- 填写步骤1-3;
- 在步骤4中上传“批量元信息文件”并验证它，直到它通过质量控制;
- 转到步骤5中的预览页面，单击finish按钮完成上传。

(2) 通过ftp上传文件。

(3) 返回提交列表，点击对应Batch ID的“FTP完成上传”按钮。之后，系统将触发自动质控系统对其进行审核。请不要重复点击，以免出现异常。

(4) 等待系统质控，结果会以邮件形式通知。

- 如果提交的文件通过质控，则提交状态为“Successful”。GWH将对该批提交中的每个基因组组装分配一个唯一的基因组组装Accession编号。
- 反之，提交状态变为“Error”。GWH通过邮件将错误报告发送给提交者。提交者需要检查并确保所附错误报告文件中的所有错误都已修复，并在您方便时重新提交更正后的文件。

● 请提前做好以下准备工作:

1. **BioProject**、**BioSample**编号
2. 批量元信息文件
3. 基因组序列文件
4. 基因组注释文件（可选）
5. 序列从属信息文件（特定情况下必需）
6. **FTP**上传文件
7. 元基因组数据提交
8. 单倍型数据提交

详细的提交
流程说明

提交准备内容

<https://ngdc.cncb.ac.cn/gwh/submit/preparation>

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



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，请参见此 pdf。
5. 更多信息，请参见 帮助文档。

我的提交 已删除的提交

9 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
	Batch0031050	xxx	PRJCA001233	2024-07-31	2024-04-23	2024-04-23	ERROR	删除



方法二、通过中心统一汇交入口BIG Sub，
登录 G W H 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

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

提交者

* 名字	中间名	* 姓氏
Meili		Chen
* 电子邮箱	电子邮箱 (备用)	
chenml@big.ac.cn	303329406@qq.com	
* 单位	单位网址	部门
Beijing Institute of Genomics, C	https://	The CAS Key Laboratory of Gei
电话	传真	
* 街道	* 城市	省/州
No.1 Beichen West Road, Chac	Beijing	
* 邮政编码	* 国家/地区	
100010	China	

保存并下一步

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

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

• 基本信息

提交标题: ⓘ

test

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

• 发布日期

☐ 审核后立即发布（推荐）

☒ 指定日期发布

2024-05-23

示例: 发布日期 (yyyy-MM-dd)

注：请仔细阅读数据发布正常和免责声明

请注意：WGS相关数据的发布将触发BioProject和BioSample的发布。

发布政策和免责声明

1. 作者可以设定一个日期，在一段指定的时间内不发布该提交的数据。

2. 发布日期可以通过基因组提交入口进行更改（<https://ngdc.cncb.ac.cn/gwh/submit/submission>）。

3. 如果一篇引用序列或Accession号的论文已发表且先于用户指定的发布日期，文章发表后序列将立即发布。否则，GWH将在用户设定的指定日期发布序列数据。

4. 一旦获得，请将文章相关数据——所有作者、标题、期刊、卷号、页码和发表时间等，发送到邮箱：gwhcuration@big.ac.cn。

☒ 我接受。

☐ 我不接受。

• BioProject和BioSample

• BioProject accession编号

PRJCA000494: wcr

或者点击创建 BioProject

BioProject关联了该研究项目的数据，这是对一项研究计划的整体描述。

请注意：如果您还没有创建相关的BioSample，点击创建 BioSample

• 数据公开后访问方式设置

• 数据访问方式

☒ 公开访问（默认）

☐ 受控访问

请注意：仅适用于人类基因组数据资源

保存并下一步

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

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

序列作者

名字

中间名

姓氏

电子邮箱

增加

删除

mei

test

test@qq.com

序列联系人

名字

中间名

姓氏

电子邮箱

增加

删除

mei

test

test@qq.com

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

文章信息

文章发布状态

未发表

已接收但未正式刊印

已发表

文章信息来源

PubMed ID号

123456

PubMed ID的文章详细信息:123456
标题: [The laboratory in programs for enteric infection control].
作者: Grados OB
期刊名称: Boletin de la Oficina Sanitaria Panamericana. Pan American Sanitary Bureau

这是出版信息吗

是

否

文章DOI链接

文章详细信息

保存并下一步

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

CNCB-NGDC

18

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

Step 1 Submitter Step 2 General Info Step 3 Reference Step 4 Meta Step 5 Overview

Batch Meta Submission

Upload genome assembly batch submission file using Excel format that includes the attributes for each genome assembly.

20240508-1.xlsx 1 Browse ... 2 validate  

Note :

1. Download GWH batch submission template file [GWH_Assembly_Template.xlsx](#)
2. For column explanations and examples, please see the e.g. [GWH_Assembly_Template.xlsx](#)
3. For more information, please see the [Help](#).

Save and next

- ①选择本地文件并上传
②上传成功后，在线校验

校验失败

校验通过

Batch Meta Submission

Upload genome assembly batch submission file using Excel format that includes the attributes for each genome assembly.

20240508-1.xlsx   validate Checked OK. 3 校验通过

Batch Meta Submission

Upload genome assembly batch submission file using Excel format that includes the attributes for each genome assembly.

20231129_1.xlsx validate   **ERROR**

Uploaded file - 20231129_1.xlsx successfully.

error.txt 1

Row	Column	Column Name	Column Value	Message
1	NA	NA	NA	The number of columns filled in row 11 exceeds the number of columns that should be filled in the template, please delete the extra columns.
2	Assembly name			Need to start with English letters, only uppercase and lowercase English letters, digits.
6	Sequencing depth			
6	Sequencing depth			
10	Genome sequence file MD5 code			
12	Genome annotation file MD5 code			
14	Existing genome accession			
11	Chromosome assignment			
15	Chromosome assignment			

报错信息：

- ①文本文件下载
②在线可视化显示

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。

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

FTP服务器登录说明

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 页结果

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

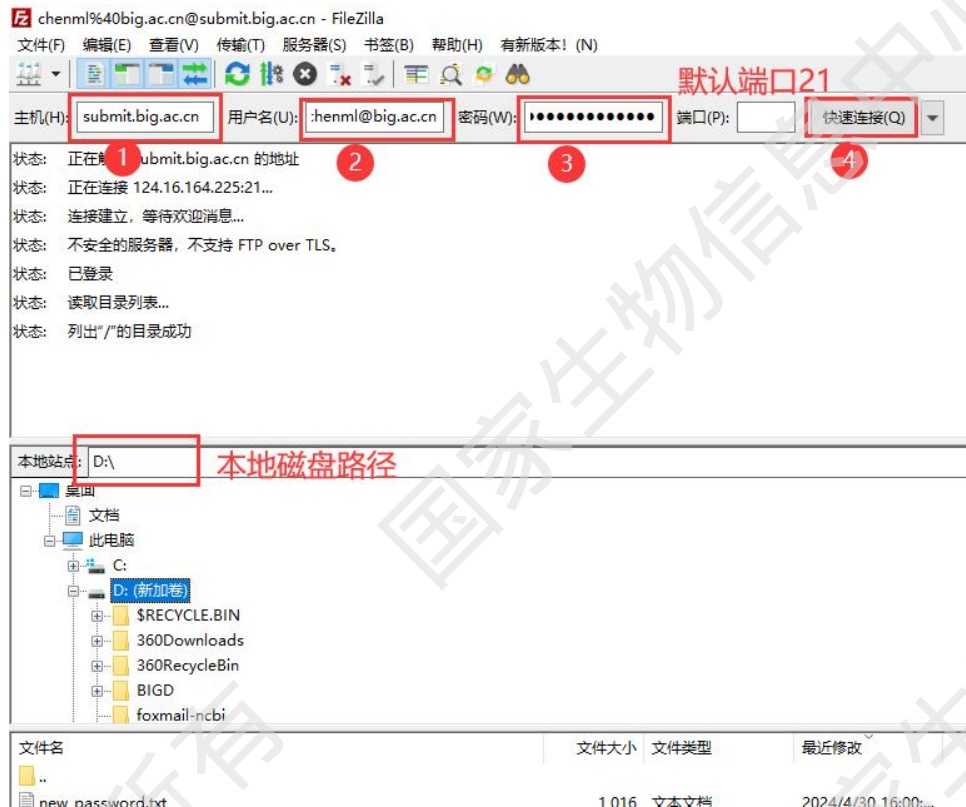
已删除的提交

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 and 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

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

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	审核人页面 立即发布
	Batch0023947	test	PRJCA000494	2024-05-23	2024-04-23	2024-04-23	Waiting for files	完成上传 删除

审核内容

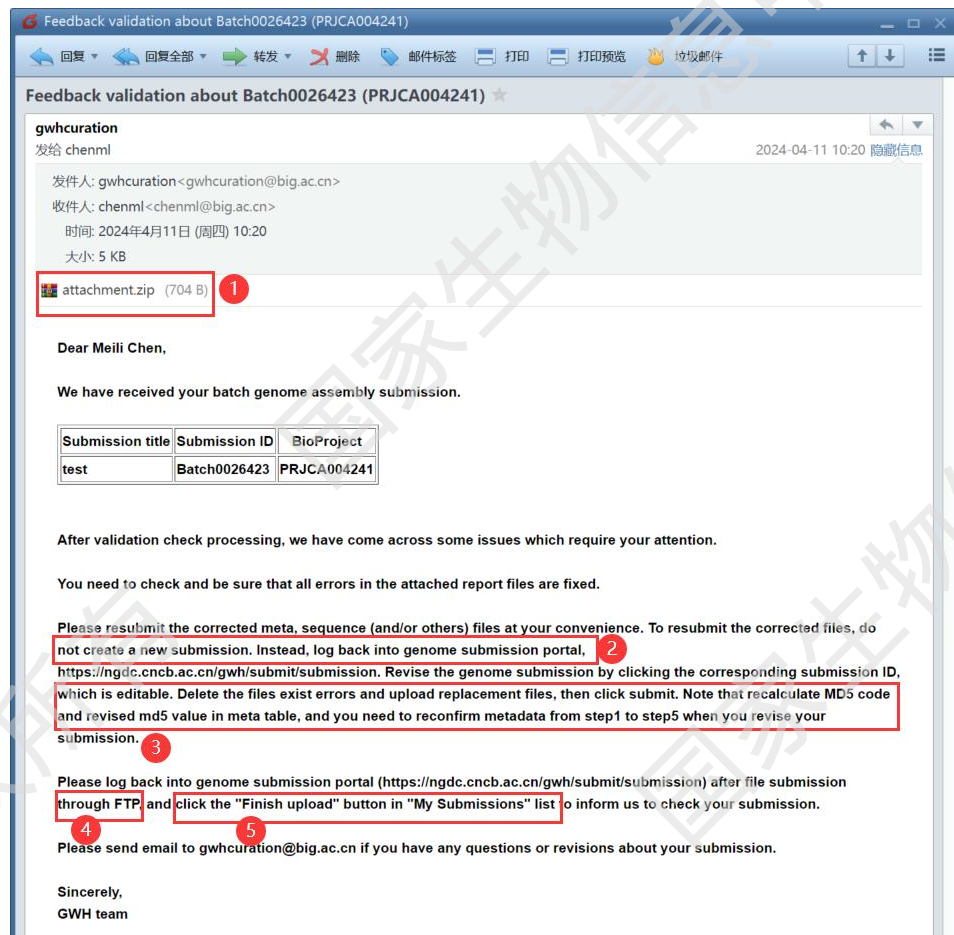
- 基因组序列文件内容的合法性和有效性
- 基因组注释文件内容的合法性和有效性
- 基因组序列和注释内容一致性审核
- 基因组序列接头、载体等污染审核

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



- 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	点击进入修改

显示第 1 至 1 项结果, 共 1 项 (由 7 项结果过滤)

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**

We inform that your submission has passed our current validation check.

Please cite the accession number GWHBISP00000000.1 like this (We recommend you putting these paragraphs in the Materials and Methods section of the paper):

The whole genome sequence data reported in this paper have been deposited in the Genome Warehouse in National Genomics Data Center [1][2]. Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformatics, under accession number GWHBISP00000000.1 that is publicly accessible at <https://ngdc.cncb.ac.cn/gwh>.

References:

[1] Genome Warehouse: A Public Repository Housing Genome-scale Data. *Genomics Proteomics Bioinformatics* 2021, 19(4):584-589. [PMCID=PMC9039550]

[2] Database Resources of the National Genomics Data Center, China National Center for Bioinformatics in 2024. *Nucleic Acids Res* 2024, 52(D1):D18-D32. [PMCID=PMC10767964]

我的提交 已删除的提交

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
2个组装编号, 隐藏 GWHAZPK01000000 GWHAZPM01000000	Batch0018094	human_CHN_Wuhan_S5_2020	PRJCA004241	2027-07-26	2024-04-23	2024-04-23	SUCCESSFUL	审稿人页签 立即发布 修改发布日期 删除

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：GWHXXXXX00000000.xx，第一版本GWHXXXXX00000000.1，其等价于GWHXXXXX00000000
 - ✓ 基因组只含**一条**单序列：GWHXXXXXxx000000，第一版本GWHXXXXX01000000，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



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](#) / [organelle.csv](#) 及其说明文档，填写并仔细检查后通过FTP方式。
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](#)。

- 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

首页 浏览 搜索 下载 提交 工具 统计 ? 文档 申请 欢迎页 ,zhaoxuetong@big.ac.cn 语言/Language

GWH提交说明文档

GWH使用手册

GWH提交快速入门指南

教程

常见问题

常见问题

往GWH提交时的一些常见问题的答案如下。

1. 简介
 - (1) 什么是GWH?
 - (2) 如何向GWH提交数据?
2. GWH账户
 - (1) 如何获得GWH帐户?
 - (2) 如果我忘记了我的GWH用户名或密码, 该怎么办?
3. 数据提交和传输
 - (1) 我该如何开始?
 - (2) 如何向GWH提交基因组文件?
 - (3) 如何准备提交文件?
 - (4) 提交文件后的处理流程是什么?
 - (5) 什么是MD5码校验? 如何计算它?
 - (6) 质量控制系统可能报告哪些类型的致命错误
4. 数据的发布和引用
 - (1) 如何设置发布日期或使数据公开?
 - (2) 如何在我的文章中引用基因组accession编号?
5. 帮助
 - (1) 联系信息
 - (2) 合作与参观

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。

My Submissions Deleted Submissions

Search: su

56 Submissions

Accession(s)	Batch ID	Title	BioProject	Release date	Created	Updated	Status	Operation(s)
	Batch0018094	SARS-CoV-2_human_CHN_Wuhan_S5_2020	PRJCA004241	2024-12-18	2024-04-23	2024-04-23	SUCCESSFUL	Reviewer Page 1 Release Now Update Release Date Delete

Showing 1 to 1 of 1 entries (filtered from 56 total entries)

Previous 1 Next

Reviewer Page

Batch Submission ID: Batch0018094

Link: The reviewer page does not exist or has expired. Please click the "Generate" button to generate a new one.

Expire: N/A

2

Generate

4

Delete

3

Extend

Close

- 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**。

My Submissions

Deleted Submissions

Search:

56 Submissions

Accession(s) ↑↓	Batch ID ↑↓	Title ↑↓	BioProject ↑↓	Release date ↑↓	Created ↑↓	Updated ↑↓	Status ↓	Operation(s) ↑↓
	Batch0018094	SARS-CoV-2_human_CHN_Wuhan_S5_2020	PRJCA004241	2024-12-18	2024-04-23	2024-04-23	SUCCESSFUL	Reviewer Page Release Now Update Release Date Delete

Showing 1 to 1 of 1 entries (filtered from 56 total entries)

Previous

1

Next

目录

1. 基因组数据库GWH
2. 基因序列数据库GenBase
3. 基因组变异数据库GVM

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

The screenshot displays the GenBase website interface. At the top, there is a navigation bar with links for 'Data Resources', 'Computing Analysis', 'Data Network', and 'Standards'. Below this, a secondary navigation bar includes 'GenBase', '首页', '数据提交', '检索', '下载', '统计', '标准', and '帮助文档'. A search bar is prominently featured with a dropdown menu set to 'Nucleotide' and a placeholder text 'Type your keywords'. To the right of the search bar are buttons for '登录', '注册', and '语言/Language'. Below the search bar, a text box describes GenBase as a public, shared gene sequence archive. The main content area is divided into two sections: '汇总与整合数据' (Summary and Integrated Data) and '最近更新' (Recent Updates). The '汇总与整合数据' section shows three categories: '物种' (Species) with 592,355 entries, '核酸' (Nucleic Acid) with 372,555,634 entries, and '蛋白' (Protein) with 598,805,212 entries. The '最近更新' section lists four recent updates with dates and descriptions. On the right side, there is a section for '国际数据GenBank整合' (International Data GenBank Integration) showing update dates and counts for nucleic acids and proteins.

国家生物信息中心
China National Center for Bioinformation

GenBase 首页 数据提交 检索 下载 统计 标准 帮助文档 登录 注册 语言/Language

GenBase 是一个公开、共享的基因序列归档库，接受用户提交（包括任何生物体的mRNA、DNA片段、非编码RNA或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自INSDC的核酸和蛋白序列数据。

Nucleotide Type your keywords 检索 高级检索

例如 C_AA001108.1, MH011443.1, GB0003962

汇总与整合数据

物种 592,355

核酸 372,555,634

蛋白 598,805,212

最近更新

2024-05-08	数据交换统计信息上线
2024-04-17	GB0001891交换到NCBI GenBank
2024-04-16	GB0002669交换到NCBI GenBank
2024-04-01	GB0001146, GB0003021交换到NCBI GenBank

国际数据GenBank整合

更新日期: 2024-05-13

核酸数: 3,149

蛋白数: 49,794

GenBase数据释放

释放日期: 2024-05-13

核酸数: 111

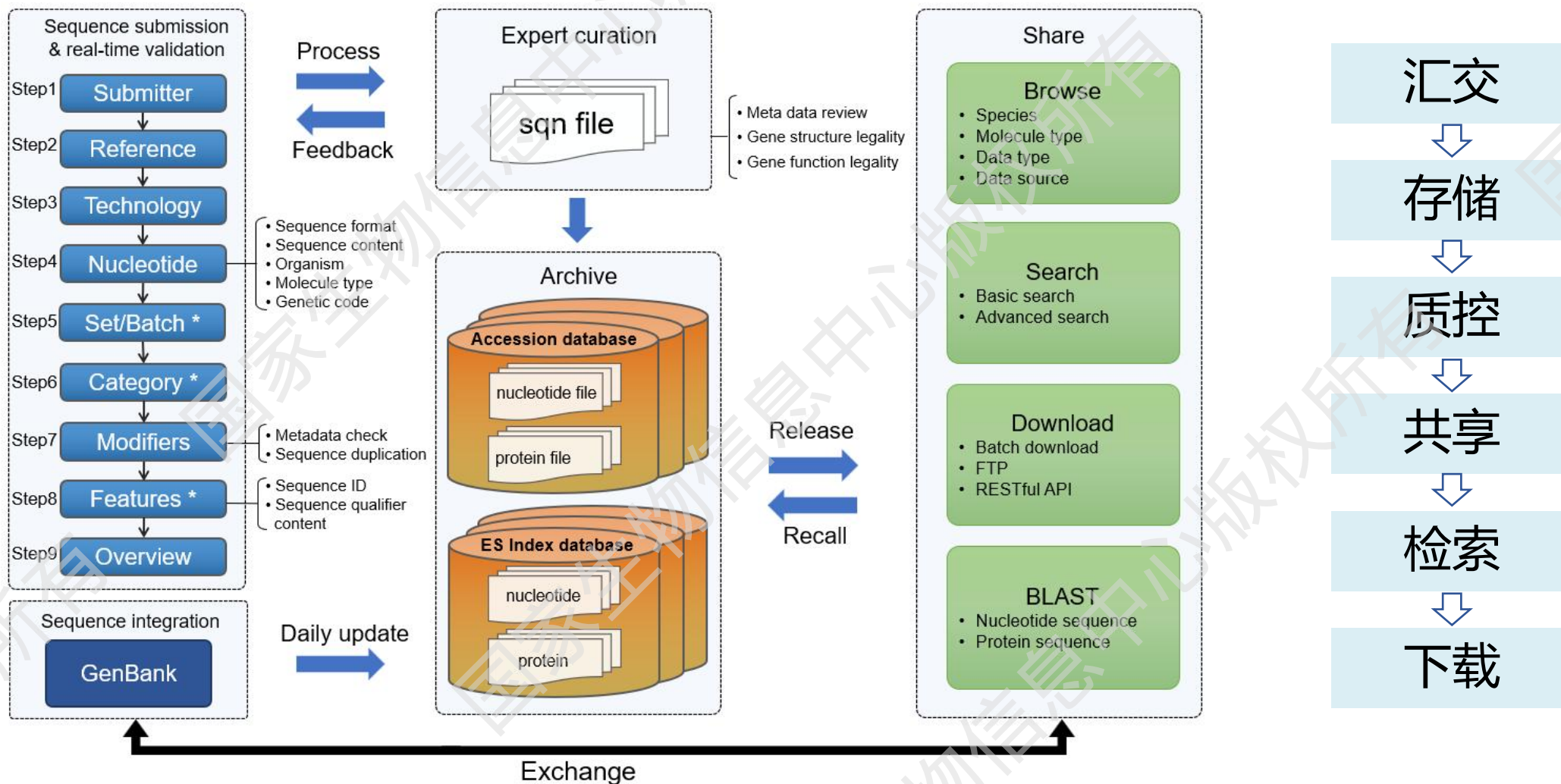
蛋白数: 1,331

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

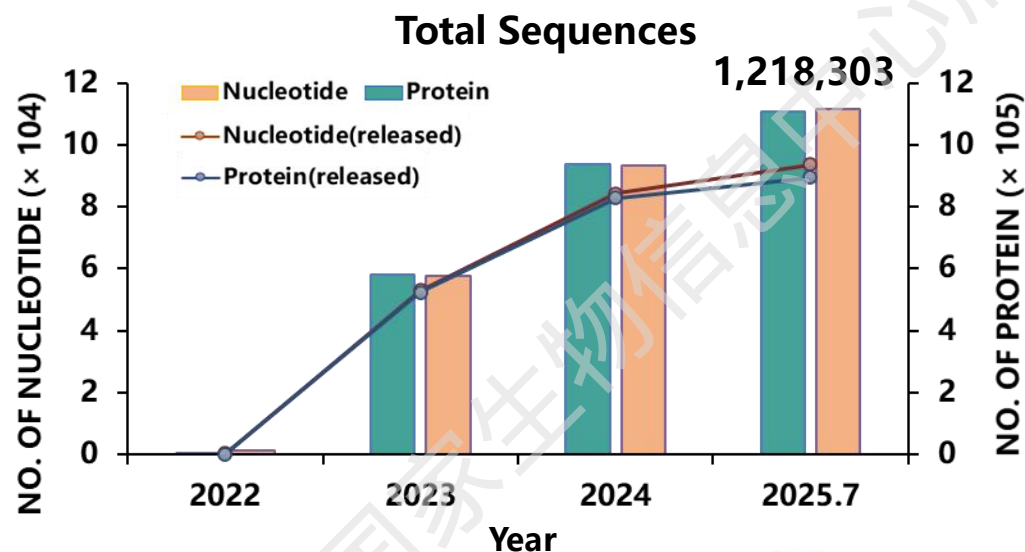
基因组和基因序列汇交的异同

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



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”



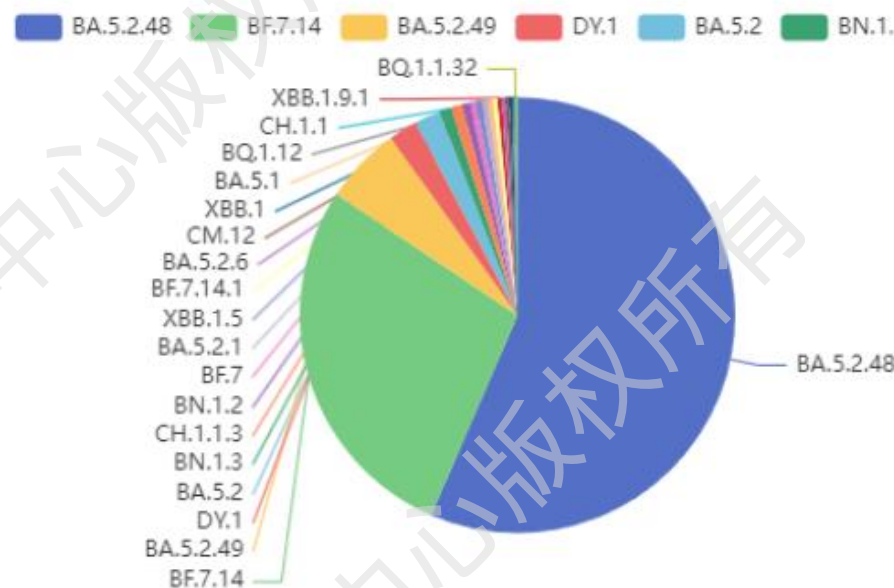
面对**国际数据脱钩**风险显现的新发展态势，**同步镜像**NCBI GenBank和RefSeq数据库中的基因序列数据资源，并实现**日更新**

受国家疾控中心委托 承担新冠病毒基因组数据的汇交和发布

中国病毒序列来源分布地图 (10609)



谱系分布 (10609)



**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

 **GenBase** 

[首页](#) [数据提交](#) [检索](#) [下载](#) [工具](#) [统计](#) [标准](#) [帮助文档](#)

[登录](#) [注册](#) [语言/Language](#)

GenBase 是一个公开、共享的基因序列归档库，接受用户提交（包括任何生物体的 mRNA、DNA 片段、非编码 RNA 或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自 INSDC 的核酸和蛋白序列数据。

Nucleotide

Type your keywords

检索

高级检索

例如 C_AA001108.1; MH011443.1; GB0003962

汇交与整合数据



物种
592,641



核酸
380,069,817



蛋白
664,750,416

最近更新

2024-12-09	病毒注释工具上线，欢迎使用。
2024-07-25	序列更新功能上线，帮助。
2024-06-25	GenBase 文章在线发表 (PMID:38913867)。欢迎引用。
2024-05-13	GenBase 通过 FAIRsharing 认证

国际数据 GenBank 整合

更新日期: 2025-08-07
核酸数: 6,129
蛋白数: 73,303

GenBase 数据释放

释放日期: 2025-08-05
核酸数: 10
蛋白数: 14

新加功能

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

 CNCB-NGDC

36

准备工作: 2 数据准备



提交准备

我的提交

GenBase 是一个公开、共享的数据库，接受用户提交（包括任何生物体的 mRNA、DNA 片段、非编码 RNA 或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自 INSDC 的核酸和蛋白序列数据。

Nucleotide

Type your keywords

例如 C_AA001108.1; MH011443.1; GB0003962

汇交与整合数据



物种

592,641



核酸

380,069,817



蛋白

664,750,416

最近更新

2024-12-09	病毒注释工具上线，欢迎使用。
2024-07-25	序列更新功能上线，帮助。
2024-06-25	GenBase 文章在线发表 (PMID:38913867)。欢迎引用。
2024-05-13	GenBase 通过 FAIRsharing 认证

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

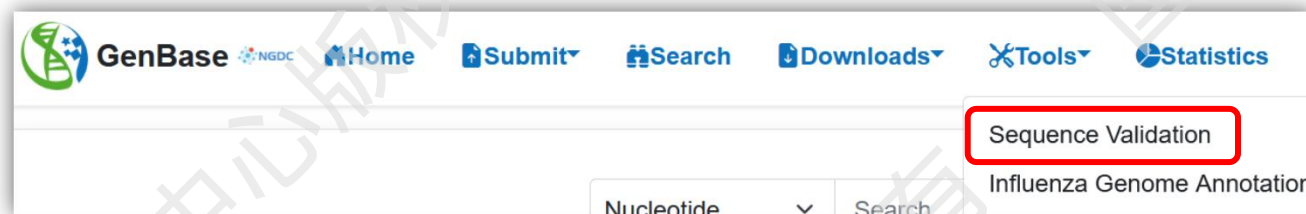
准备工作：2 数据准备

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

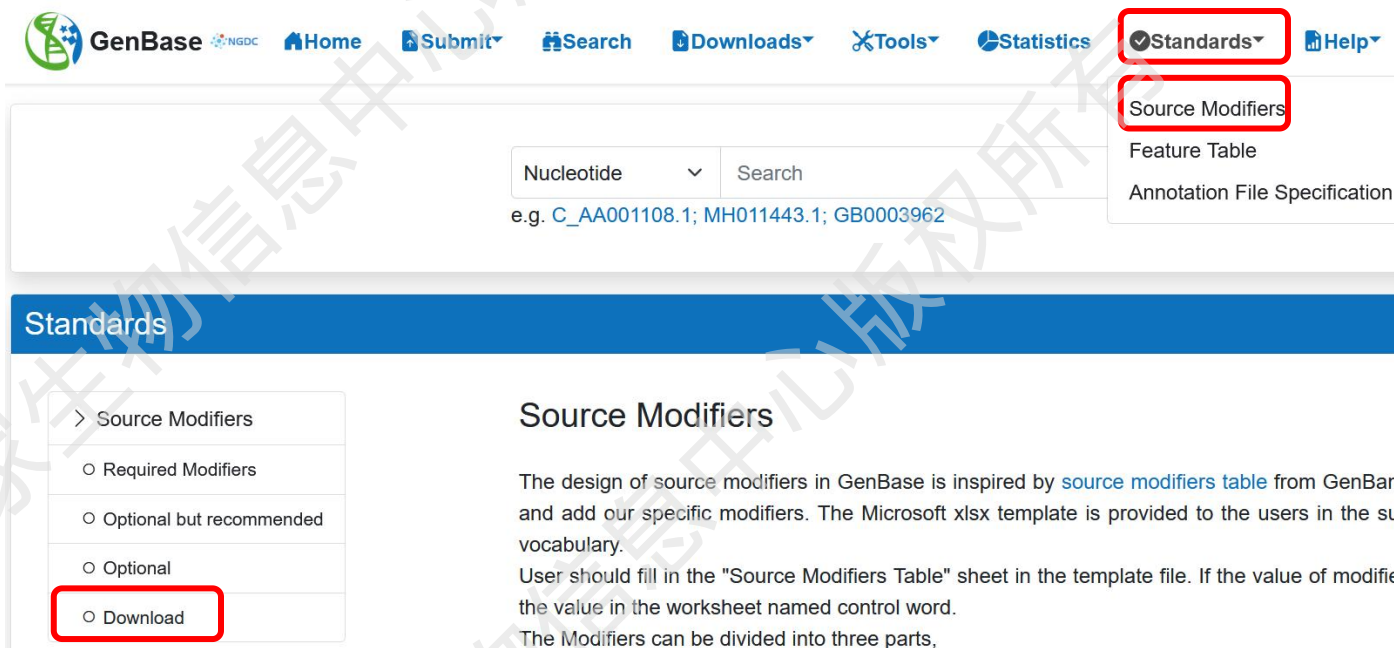
Sequence ID命名要求：

- (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 same Sequence_ID used in the FASTA file. 在 GenBase Modifiers table 中由 FASTA 文件中使用的相同 Sequence_ID 标识						
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						
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
		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
14	##Description						
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-2020	Data agreements were established
8	Anguilla	plastid:cyanelle	missing: endangered species	Information can not be reported
9	Antarctica	plastid:leucoplast	missing: human-identifiable	Information can not be reported
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:

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

准备工作：2 数据准备-Annotation Features文件



The screenshot shows the GenBase website interface. At the top, the navigation bar includes links for Home, Submit, Search, Downloads, Tools, Statistics, Standards, and Help. The 'Standards' link is highlighted with a red box. Below the navigation bar, there is a search bar with a dropdown menu set to 'Nucleotide' and a search button. Below the search bar, there is a text input field with the example text 'e.g. C_AA001108.1; MH011443.1; GB0003962'. To the right of the search bar, a dropdown menu is open, showing options for 'Source Modifiers', 'Feature Table' (highlighted with a red box), and 'Annotation File Specification'. Below the search bar, there is a blue header bar labeled 'Standards'. Under the 'Standards' header, there is a sidebar with a list of options: 'Feature Table', 'Locations', 'Attributes', 'Qualifier hints', 'Description of each feature', and 'Download' (highlighted with a red box). The main content area is titled 'Feature Table' and contains text explaining the design of the feature table in xlsx format, referring to the TBL in text format from NCBI. It also includes a section for 'Locations' with a list of items: '1) Chr.' and '2) Start and end.'.

GenBase NGDC Home Submit Search Downloads Tools Statistics Standards Help

Nucleotide Search
e.g. C_AA001108.1; MH011443.1; GB0003962

Source Modifiers
Feature Table
Annotation File Specification

Standards

- > Feature Table
 - Locations
 - Attributes
 - Qualifier hints
 - Description of each feature
 - Download**

Feature Table

The design of the feature table in xlsx format refers to the TBL in text format from [NCBI](#). Consistent with the INSDC standard. Users can get tips from different feature and qualifier combinations in the table, including:

Locations

- 1) Chr.
 - Header of each sequence, starting with the **Feature** field
- 2) Start and end.

准备工作：2 数据准备-Annotation Features文件

下载后打开GenBase_Features.xlsx模板：

</

准备工作：2 数据准备-Annotation Features文件

	A	B	C	D	E	F	G	H	I	J	K
1	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.							Qualifier hints (提示)			
2								feature	qualifier	<= All features and qualifiers of INSDC	
3								gene	gene	<= Select the feature and qualifier to view in the drop-down box respectively	
4	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)							Definition		<= Select the content to prompt	
5											
6	Locations (坐标)				Attributes (属性)			symbol of the gene corresponding to a sequence region		<= Display box	
7	sequence id	start	end	completeness	feature	qualifier	qualifier value				
8	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. 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 2. If the following lines belong to this feature, the following feature cell should be empty	1. It can be empty only to display this block or attribute	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	
9										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	
10	seq1	1	1796	3' partial	gene	gene	NRPS	<= A gene block		1. The yellow fill indicates that the cell value can be omitted	
11	seq1	1	1796	3' partial	mRNA	product		<= A mRNA block		2. The red fill indicates that the value of this cell must be empty, indicating that it belongs to the same block	
12	seq1	1	294	complete	cds	codon_start	1	<= A cds block			
13	seq1	342	603	complete		product	Non-Ribosomal Peptide Synthetases				
14	seq1	667	1796	complete							
15	seq1	667	1796			note	contains adenylation domain				
16											
17											
18											
19											
	myFeature	example	zh-example	feature_inspection_sheet							

准备工作：2 数据准备-Annotation Features文件

	A	B	C	D	E
1	The comparison table lists all the features and qualifiers that can be filled in. More filling explanations can be obtained at https://www.insdc.org/documents/feature_table.html				
2					
3					
4	definition	feature	qualifiers	type	comment
5	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
6			(Optional) citation	[number]	
7			(Optional) db_xref	<database>:<identifier>	
8			(Optional) experiment	[CATEGORY]:text	
9			(Optional) function	text	
10			(Optional) gene	text	
11			(Optional) gene_synonym	text	
12			(Optional) inference	[CATEGORY]:[TYPE] (same spec	
13			(Optional) locus_tag	text (single token)	
14			(Optional) map	text	
15			(Optional) note	text	
16			(Optional) old_locus_tag	text (single token)	
17			(Optional) standard_name	text	
18			(Optional) trans_splicing	boolean	
19	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
20			(Optional) citation	[number]	
21			(Optional) db_xref	<database>:<identifier>	
22			(Optional) experiment	[CATEGORY]:text	
23			(Optional) function	text	
24			(Optional) gene	text	
25			(Optional) gene_synonym	text	
26			(Optional) inference	[CATEGORY]:[TYPE] (same spec	
27			(Optional) locus_tag	text (single token)	
28			(Optional) map	text	
29			(Optional) note	text	
30			(Optional) old_locus_tag	text (single token)	
31			(Optional) standard_name	text	
32			(Optional) trans_splicing	boolean	
33	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
34			(Optional) artificial_location	[artificial_location_value]	
35			(Optional) circular_RNA	boolean	
36			(Optional) citation	[number]	
37			(Optional) codon_start	<1 or 2 or 3>	
38			(Optional) db_xref	<database>:<identifier>	
39			(Optional) EC_number	text	
40			(Optional) exception	[exception_value]	
41			(Optional) experiment	[CATEGORY]:text	
42			(Optional) function	text	
43			(Optional) gene	text	
44			(Optional) gene_synonym	text	
45			(Optional) inference	[CATEGORY]:[TYPE] (same spec	
46			(Optional) locus_tag	text (single token)	
47			(Optional) map	text	

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

Feature_inspection_sheet

提交流程：创建新的提交

两种方式进入GenBase递交系统：



GenBase NGDC 首页 数据提交 检索 下载 工具 统计 标准 帮助文档 登录 注册 语言/Language

GenBase 是一个公开、共享的基因序列归档库，接受用户提交（包括任何生物体的mRNA、DNA片段、非编码RNA或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自INSDC的核酸和蛋白序列数据。

例如 C_AA001108.1; MH011443.1; GB0003962

检索 高级检索

汇总与整合数据

- 物种: 592,641
- 核酸: 380,032,432
- 蛋白: 663,761,427

最近更新

- 2024-12-09 病毒注释工具上线，欢迎使用。
- 2024-07-25 序列更新功能上线，帮助。
- 2024-06-25 GenBase文章在线发表 (PMID:38913867)。欢迎引用。
- 2024-05-13 GenBase通过FAIRsharing认证

国际数据GenBank整合

更新日期: 2025-08-01
核酸数: 6,294
蛋白数: 339,302

GenBase数据释放

释放日期: 2025-08-01
核酸数: 3
蛋白数: 0



国家生物信息中心 Data Resources Computing Analysis Data Network Standards

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

NGDC Home Documentation Login Register

BIG Sub 使用说明 BIG Sub Quick Start Guide SDAS 科技计划项目科学数据汇交 科学数据汇交服务系统 提供汇交计划提交、实时数据跟踪查询、汇交证明出具等服务

请根据您的数据类型选择对应的数据库提交

- BioProject 收集与共享生物项目信息的数据库
- BioSample 收集与共享实验组为生物样本信息数据库
- Genome Sequence Archive 电子学数据汇交、存储、管理与共享系统
- GSA for Human 人类遗传资源组学数据数据库
- Genome Warehouse 多物种基因组序列和注释数据数据库
- GenBase 核酸和蛋白质序列及注释数据数据库
- OMIX 收集和共享表达谱、甲基化谱、蛋白质组学、代谢组学等数据的数据库
- Genome Variation Map 基因组序列变异数据汇交、管理与检索的数据库
- BioCode 整合并开发生物信息软件工具的数据库
- Database Commons 生物数据收集、分类、检索系统

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



GenBase

首页

数据提交

检索

统计

标准

帮助文档

Nucleotide

Search

检索

高级检索

例如 C_AA001108.1; MH011443.1

提交类型

常规序列提交

SARS-CoV-2快速提交

受控序列提交

点击此处开始递交

*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）



CNCB-NGDC

提交流程： 步骤1 提交者信息 (Submitter)

The screenshot shows a multi-step submission process. At the top, a progress bar contains nine steps: 1. Submitter (active), 2. Reference, 3. Technology, 4. Nucleotide, 5. Set/Batch, 6. Category, 7. Modifiers, 8. Features, and 9. Overview. Below the progress bar is the 'Submitter' form. The form is divided into three columns. The first column contains fields for First Name (zhao), Email (zhaoxuetong@big.ac.cn), Organization (Beijing Institute of Genomics, Chinese Academy of Sciences), Phone, Street (Beichen West Road, Chaoyang), and Postal Code (100010). The second column contains fields for Middle Name, Email(secondary), Submit Organization Url (https://), Fax, City (Beijing), and Country (China). The third column contains fields for Last Name (xuetong), Department, and State (-). At the bottom left of the form is a 'Save and next' button, which is highlighted with a red box and the text '保存并下一步'.

Submitter

★ First Name: zhao

Middle Name

★ Last Name: xuetong

★ Email: zhaoxuetong@big.ac.cn

Email(secondary)

★ Organization: Beijing Institute of Genomics, Chinese Academy of Sciences

Submit Organization Url: https://

★ Department

Phone

Fax

★ Street: Beichen West Road, Chaoyang

★ City: Beijing

State: -

Postal Code: 100010

★ Country: China

Save and next 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)



Submitter



Reference



Technology



Nucleotide



Organism



Set/Batch



Category



Modifiers



Features



Overview

Reference

Submission #0000099

Corresponding Author(s) (Note: The information of professors or Associate Professors is recommended)

* First Name

xuetong

Middle Initial(s)

* Last Name

zhao

Suffix

Remove

Add More Corresponding Author(s).

Reference Information #1

Please provide the title and relevant publication details (volume, issue, etc.) of a paper that discusses this submission.

* Publication Status

☒ Unpublished ☐ In-Press ☐ Published

Reference Title

* Reference Authors

☒ Same As Corresponding Author(s)

☐ Specify New Authors

Add Another 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)

Submission #0000099

Sequencing Technology

This information is required if you are submitting over 500 sequences or if your sequences were generated using next-generation sequencing technology.

What methods were used to obtain these sequences?

☒ Sanger dideoxy sequencing
☐ Illumina
☐ Other

☐ IonTorrent

☐ 454
☐ PacBio

☐ Helicos
☐ SOLID

Are these sequence(s):

Assembled sequences

Continue

保存并下一步

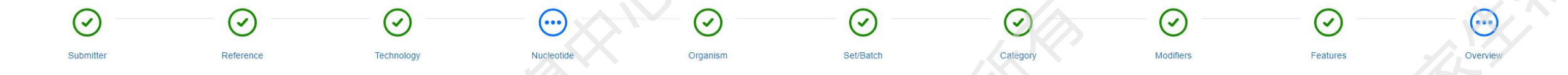
根据具体情况填写

上述序列为:

- Assembled sequences
- Unassembled sequence reads
- Assembled sequences

**这里要选Assembled sequences
否则需要提交到GSA而非GenBase**

提交流程： 步骤4 序列 (Nucleotide)



Nucleotide

设置数据释放时间

- 最长不要超过4年
- 只要数据未公开，可随时修改该时间
- 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.

Submission #0000099

Help

Submission Release Date

When may we release your sequence record?

☒ Immediately After Processing

☐ Release Date:

Release Policies and Disclaimers

仔细阅读数据发布政策和免责声明

- A date can be set by authors to withhold the release of new submissions for a specified period of time.
- If submitters want to change release date, please contact GenBase group: genbase@big.ac.cn.
- 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.
- 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.
- After the data is released, it may automatically synchronize to NCBI in the future, and submitters will no longer be contacted for confirmation.

☒ 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)

Sequence(s) and Definition Line(s)

*** Molecule Type:** Select the type of molecule to be submitted.

Topology:

Strand:

To which compartment does the one you submitted belong? please select one compartment type. If multiple genetic compartment types, you should submit them separately.

Are you submitting the complete sequence of an organelle genome, virus, viral segment, viroid, plasmid, or cloning vector? ☒ Yes ☐ No

Sequence Type Selections:

- genomic DNA
- mRNA (cDNA)
- genomic RNA
- precursor RNA
- tRNA
- rRNA
- cRNA
- transcribed RNA
- Other genetic: RNA
- Other genetic: DNA

Topology Selections:

- Linear
- Circular

Strand Selections:

- Single strand
- Double strand
- Mixed strand

Compartment Selections:

- nucleotide
- mitochondria
- plastid

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

Nucleotide Sequence Format

Sequence data format:

FASTA sequences (not an alignment, most common data format, [FASTA help](#))

For example:

>Seq1 [organism=Homo Sapiens] Definition Line for Seq1

aaccgatatagagagga

>Seq2 [organism=Homo Sapiens] Definition Line for Seq2

atctgaatagattatt

[Definition Lines](#) which are used to describe each sequence, should be included with your sequence data.

核酸序列填写示例

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？

提交流程： 步骤4 序列 (Nucleotide)

★ Nucleotide Sequence(s)

☐ Paste sequence data (when sequence number < 40)

```
>Alligator_VPS52-RNASE
CTTGAGCTAAGTTTCCCCAAATACTTAGTGAACCCAGGGTAGCATTGTAAGTGT
TGTGTTTGTTCCTTTGCATGTCTATTACTACTAGGACCATATATATTTACATAC
TTAGCAATAAATACTTCTATAGTTAACTGGTGGTAAGTGGGTGTATTTTCCTTT
ACTTTTCCTTCCCACTTGCTCTGTAGCAACGCTTCTTTACCTAAGCTAAAAATC
CCTCTCAACCTCAGAAAAAAGTTCCGCTCTCTCTCAGCAACGCTAAGCTAAGT
```

☐ Upload local file (.zip is supported. The maximum allowed uploaded file size is 1GB)

Select file... Browse ...

☒ Upload by FTP (when file size > 1GB)

★ File Name

★ MD5

Note

Please transmitting your data files to the GenBase FTP site (download FileZilla client). We recommend you transer the file via FTP first, then fill the file name and md5 here and click continue.

Address: ftp://submit.big.ac.cn

User: Same as you login the GenBase

Password: Same as you login the GenBase

Continue

- 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

保存并下一步

提交流程： 步骤5 物种 (Organism)



Organism

Submission #0000099

Help

Fill in missing Organism information

You did not include the name of the organism from which the sequence was isolated, see below. Please enter the organism name below. (For future sequence submissions, be sure to use the FASTA format.)

[Download organism template file](#)

Organism Name:

Nilaparvata lugens

Input same organism name for all sequences

or

Upload File:

Select file...

Browse ...

Continue

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

org-table-sample - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

Sequence_ID	Organism
Seq1	Homo sapiens

提交流程： 步骤6 集合/批次 (Set/ Batch)



Set/ Batch

Submission #0000099

All sequences in a set must be from the same gene/locus and are expected to be released at the same time.
Please select a set type:

- ☒ Pop set
- ☐ Phy set
- ☐ Mut set
- ☐ Env set

Population study: a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.

Phylogenetic study: a set of sequences that were derived by sequencing the same gene from different organisms.

Mutation study: a set of sequences that were derived by sequencing multiple mutations of a single gene.

Environmental study: a set of sequences that were derived by sequencing the same gene from a population of unclassified or unknown organisms.

If your sequences are NOT all from the same gene/locus and NOT intended to be released at the same time, then choose 'Batch' below.

- ☐ Batch
- Multiple, related nucleotide sequences that are not from the same gene, but may be from the same study or organism.

Continue

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)



Category

Submission #0000099

Indicate whether your sequence is an original submission. If your sequence is a third-party annotation submission, please contact genbase@big.ac.cn.

☒ Original Directly sequenced by submitter.

☐ Third Party Annotation Derived from other primary sequence data.

[Continue](#)

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)

Submission progress bar showing steps: Submitter, Reference, Technology, Nucleotide, Organism, Set/Batch, Category, Modifiers (active), Features, Overview.

Source Modifiers

Submission #0000099

Note

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

*** 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

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

0000099_modifier.xlsx Remove

Successful verification!

Continue

1.点击下载Modifiers文件模板

2.点击上传填好的Modifiers文件

3.保存并下一步

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

##13
Organelle/Location
genetic compartments.
Available vlues can be found at
ControlWrords sheet.

提交流程： 步骤9 提交Features文件 (Features)

Submission #0000099

Note

1. Select one of three formats for submitting the annotation of sequence
2. Please refer to here for the explanation of annotation in TBL format
3. If you want to submit annotation features in Excel format, please download the function submission template file Genbase_Features.xlsx

☐ Upload feature annotations for your submission using Excel format that includes the attributes for each sequence.

☐ Upload feature annotations for your submission using five column Feature Table that includes the attributes for each sequence.

☒ Upload feature annotations for your submission using GFF3 format.

0000099_feature.gff3 Remove

Continue

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)

Overview

Submission #0000021 [Help](#)

1. Additional Email Addresses?
Correspondence regarding this submission will be sent to the following email address:

Separate multiple email addresses with commas.

2. Resubmission?
If you were asked by GenBase staff to resubmit your sequence data, check here: ☐

3. Additional Information
If you have additional or corrected source modifier or feature files, or other plain text description for your sequence data submission, check here: ☐

4. Updates
You may update or revise your submissions at any time by sending new or corrected information in an email to genbase@big.ac.cn. You may also contact us at this address with any questions.

Review Records of Your Set
Below please find your 12 GenBase submission record(s) for your review.
You can download the complete set as a compressed ZIP file.
Warn: there are some warnings in your sequence file, you can continue finish, but we suggest you need to check the sequence before your finish.

[Finish Submission](#)

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	GS3DN2379VN1	2483 bp	RNA	linear	UNK 16-AUG-2022
DEFINITION	Human picobirnavirus.				
VERSION	GS3DN2379VN1				
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				
	ESIQSGGLEETGRHNLATEQYNINALGEQFRHNKQTELQGWISTGAGAYKDVSTGLRQN				
	VESVKGIINGVKSILPGG"				
CDS	707..2476				
	/codon_start=1				
	/product="capsid protein"				
	/translation="MKDSRNIGRNSKYQGKRVNKSRSKSELDQINERYSKDNDWRWYA				

- 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	-		2022-10-26 22:34:47	2022-10-27 11:31:07	2022-10-27	Unfinished Modifiers	-
0000144	-		2022-10-26 15:25:21	2022-10-26 23:02:34	2022-12-31	Unfinished Modifiers	-
0000142	1		2022-10-24 18:39:43	2022-10-24 18:41:12	2022-10-24	Unfinished Organism	-
0000141	-		2022-10-24 15:34:55	2022-10-25 15:48:33	2026-10-24	Unfinished Features	-
0000140			2022-10-24 13:54:41	2022-10-24 15:42:45	2023-05-01	Unfinished Overview	-
0000139	-		2022-10-21 16:09:10	2022-10-21 16:14:54	2023-12-12	Unfinished Modifiers	-
0000138	Genome-Wide Analysis of the		2022-10-21 15:57:39	2022-10-21 16:09:54	2023-10-03	Unfinished Features	-
0000137	-		2022-10-20 22:25:50	2022-10-21 08:59:39	2022-10-20	Unfinished Modifiers	-
0000136	XK10		2022-10-20 16:06:57	2022-10-20 19:42:26	2022-10-20	Unfinished Modifiers	-
0000135	-		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	-		2022-09-29 13:41:04	2022-10-09 14:02:00	2022-10-09	Unfinished Set/Batch
0000115	Rc_virus		2022-09-28 15:15:20	2022-10-19 15:34:11	2022-09-30	Audit Success Accession
0000114	RcALV-BelV		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	-		2022-08-19 18:17:50	2022-08-19 20:08:35	2022-08-19	Unfinished Features -
0000064	-		2022-08-18 20:28:27	2022-08-18 20:28:27	-	Unfinished Reference -
0000063	-		2022-08-18 08:44:06	2022-11-03 11:30:59	2022-08-18	Audit Fail View -
0000062	-		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	Accession	
0000086	-		2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success		public
0000202			2022-11-29 11:09:41	2022-11-30 10:35:22	2022-11-29	Audit Success		confidential

点击查看Accession编号

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

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

```
LOCUS       C_AA001111                2622 bp    mRNA    linear   INV 31-AUG-2022
DEFINITION  Nilaparvata lugens Zinc finger homeodomain 1 (zfhl) isoform B (Zfhl)
            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="Zfhl"
     CDS               1..2622
                     /gene="Zfhl"
                     /codon_start=1
                     /product="Zinc finger homeodomain 1 (zfhl) isoform B"
```

```
>C_AA001111.1 Nilaparvata lugens Zinc finger homeodomain 1 (zfhl) isoform B (Zfhl) mRNA, complete cds
ATGTATTTCATCGCCGACAGCTTCCTGCTACGATAGCTACTCGTACCGGCTGTGGAGTCTGGCAACGCTGT
GGACCACTTATCAATTGCGCTTTACCCGAGTGTGTAAGCAGCGGTGTAAGGAGGAGACGGGTGGCAACCC
CGGGACACAGGTGGCGTCTCGCCGCGGACGACTTTCATCAAAATGCCCGCAGTGTGGCGCTGGCTACGCG
GGCTTCCAGGCGCTCAAGGAGTACGTGGAGTTGGCTCATGGAGGCGGTGCGCTCTCCTGTCTGCAGTGCA
CGGCTACCTTCCCCACGAGGGATCTACTCGACAAACATGAGCTGTTCATTGCCCCAATGCGCAGGTGTG
ATGCAAAAGTTTGCAACAAACATTGCGCAATGTGTACAGACTTCAAAGGCACATGATAAGCCATGATGAA
AGTGCAAAAGCTTGGCAATTCAGTGTCCAGAATGTGAGAGGCTTCAAATTCAGCATCACCTTAAGG
```

数据引用



GenBase 是一个公开、共享的基因序列归档库，接受用户提交（包括任何生物体的mRNA、DNA片段、非编码RNA或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自INSDC的核酸和蛋白序列数据。

Nucleotide

Type your keywords

例如 C_AA001108.1, MH011443.1, GB0003962

汇交与整合数据



物种
592,641



核酸
380,069,817



蛋白
664,750,416

最近更新

2024-12-09

病毒注释工具上线，欢迎使用。

2024-07-25

序列更新功能上线，帮助。

2024-06-25

GenBase文章在线发表 (PMID:38913867)。欢迎引用。

2024-05-13

GenBase通过FAIRsharing认证

帮助和支持

提交指南

如果您在数据提交过程中遇到任何问题，或想向我们提出任何建议/意见或报告系统错误，请随时联系我们。

Email: genbase@big.ac.cn

QQ群: 629388189

座机: 86-10-84097298

工作时间: 工作日9:00 - 17:00

如何引用?

当您成功提交数据到GenBase并通过审核后，请在您要发表的论文中添加如下语句：

The data reported in this paper have been deposited in the GenBase [1] in National Genomics Data Center [2], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number C_AA000000 that is publicly accessible at <https://ngdc.cncb.ac.cn/genbase>.

References:

[1] GenBase: A Nucleotide Sequence Database. Genomics Proteomics Bioinformatics 2024, 22(3):qzae047 [PMID=38913867].

[2] Database Resources of the National Genomics Data Center, China National Center for Bioinformation in 2024. Nucleic Acids Res 2024, 52(D1):D18-D32 [PMID=38018256].

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		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/ Cancel Modify Grant
0000089	Nlzf1,isoform B and isoform C		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/ Cancel Modify Grant
0000086	-		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/ Cancel Modify Grant

3. 点击Modify Grant按钮

4. 可修改资助信息

GenBase标准页面



GenBase

Home

Submit

Search

Statistics

Standards

Help

Login

Register

语言/Language

Standards

> Source Modifiers

- Required Modifiers
- Optional but recommended
- Optional
- Download

Source Modifiers

The design of source modifiers in GenBase is inspired by [source n](#) template is provided to the users in the submission page. It is also p User should fill in the "Source Modifiers Table" sheet in the template The Modifiers can be divided into three parts,

Required Modifiers

- GenBase has some mandatory items, such as Sequence_ID ar must correspond to the SeqID in FASTA file and cannot be less than

Optional but recommended

- These modifiers are recommended by GenBase, which you car

Optional

- These modifiers are optional

Source Modifiers

Feature Table

Annotation File Specification

Standards

> Feature Table

- Locations
- Attributes
- Qualifier hints
- Description of each feature
- Download

Feature Table

The design of the feature table in xlsx form feature and qualifier combinations in the ta

Locations

- 1) Chr.
 - Header of each sequence, starting w
- 2) Start and end.
 - Start and end coordinates of each fe
 - Coordinate should be a value of integ
 - If the end is greater than the starting
- 3) Completeness
 - Whether the feature is complete. For
 - If incomplete, mark the incompleteness

Standards

Annotation file specification

Refer to the specification of <https://github.com/The-Sequence-Ontology/Specification>

1) Annotation file types

Contains information such as genes, exon regions, coding regions, untranslated re and uploaded as annotation files to GenBase.

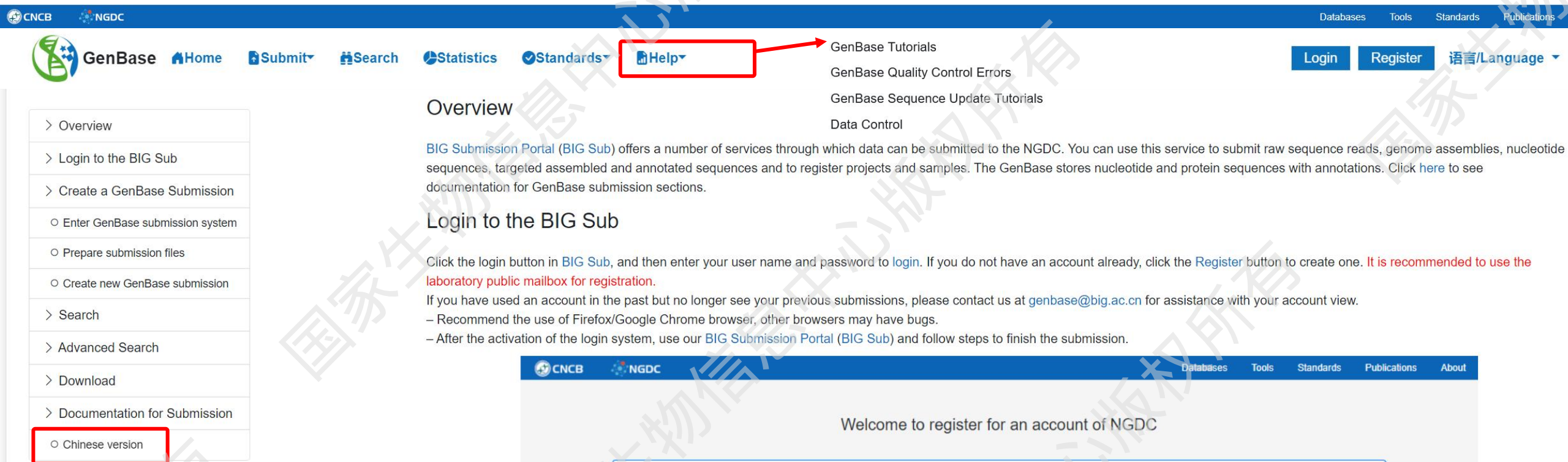
2) Example of sequence annotation

• Eukaryotic gene

Excel format example

Locations (位置)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value
seq1	1	500	500/500 complete	regulatory	gene	abc1
seq1	1	500	500/500 complete	regulatory	regulatory_class	promoter
seq1	780	1320	1320/1320 complete	mRNA	gene	abc2
seq1	14	567	567/567 complete	CDS	gene	abc2
seq1	789	1234	1234/1234	product	sequence	coding sequence
seq1	10	567	567/567 complete	function	cell division control	
seq1	568	788	788/788 complete	exon	number	1
seq1	568	788	788/788 complete	intron	number	1
seq1	789	1320	1320/1320 complete	exon	number	2
seq1	1310	1317	1317/1317 complete	regulatory	regulatory_class	signal_sequence
seq1				gene	abc1	

GenBase说明文档页面



The screenshot shows the GenBase website interface. The top navigation bar includes links for Databases, Tools, Standards, and Publications. The main navigation menu on the left includes Overview, Login to the BIG Sub, Create a GenBase Submission, Search, Advanced Search, Download, and Documentation for Submission. The 'Help' dropdown menu is highlighted with a red box and an arrow pointing to the 'GenBase Tutorials' link. The 'Overview' section is visible, providing information about the BIG Submission Portal (BIG Sub) and the login process.

GenBase Home Submit Search Statistics Standards Help

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

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

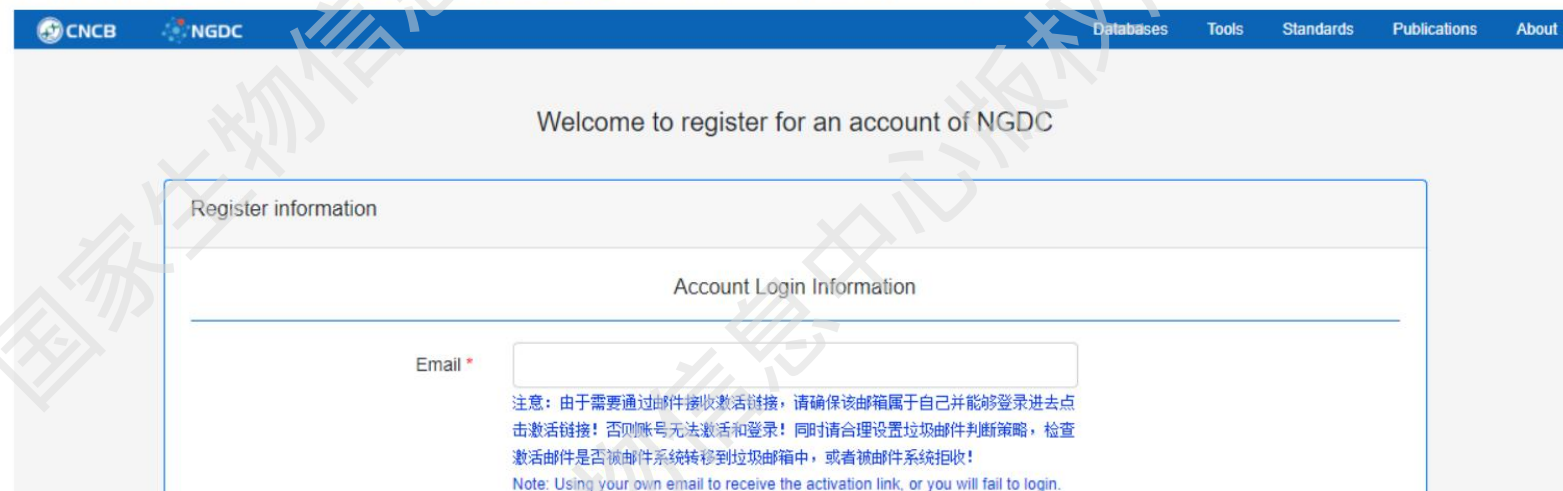
Click the login button in [BIG Sub](#), and then enter your user name and password to [login](#). If you do not have an account already, click the [Register](#) button to create one. **It is recommended to use the laboratory public mailbox for registration.**

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.

Documentation for Submission

- Chinese version



The screenshot shows the NGDC registration page. The header includes links for Databases, Tools, Standards, Publications, and About. The main heading is 'Welcome to register for an account of NGDC'. The registration form is divided into two sections: 'Register information' and 'Account Login Information'. The 'Account Login Information' section includes a form for Email and a note about using a personal email for activation.

Register information

Account Login Information

Email *

注意：由于需要通过邮件接收激活链接，请确保该邮箱属于自己并能够登录进去点击激活链接！否则账号无法激活和登录！同时请合理设置垃圾邮件判断策略，检查激活邮件是否被邮件系统转移到垃圾邮箱中，或者被邮件系统拒收！

Note: Using your own email to receive the activation link, or you will fail to login.

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 terminatio 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 fo the CDS sequence and clearly indicate the specific position and corresponding amino acid information.

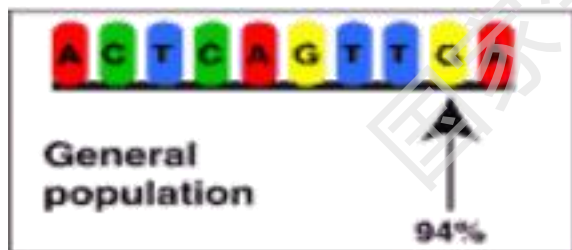
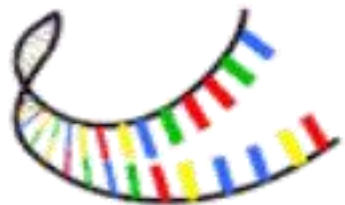
目录

1. 基因组数据库GWH
2. 基因序列数据库GenBase
3. 基因组变异数据库GVM

基因组变异

Polymorphism

"Poly" *many* "morphe" *form*



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

图片来源: genome news network

国际变异资源库

An official website of the United States government [Here's how you know](#)

NIH National Library of Medicine
National Center for Biotechnology Information

Log in

dbSNP [Advanced](#) [Help](#)

NCBI-dbSNP/dbVar

ns and deletions along with
ing information for both

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](#)
- [sUtils 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[ab]
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

EBI-EVA

News : variation data from all Tweets by @evarchive

Overview
The European Variation Archive (EVA) is a public archive of human 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.

Search for SNPs

ex: rs123 or ss567 or comma/space separated RS or SS IDs

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)

Top 5 Species

Species	Percentage
Human	67.7%
Cow	12.2%
Dog	7.1%
Sheep	6.6%
Horse	6.3%

Top 5 Types

Type	Percentage
Whole Genome Sequencing	36.9%
Exome Sequencing	19.8%
Target Sequencing	15.8%
Genotyping By Array	12.3%
Genotyping By Sequencing	8.4%

Structural variants studies (>50bp)

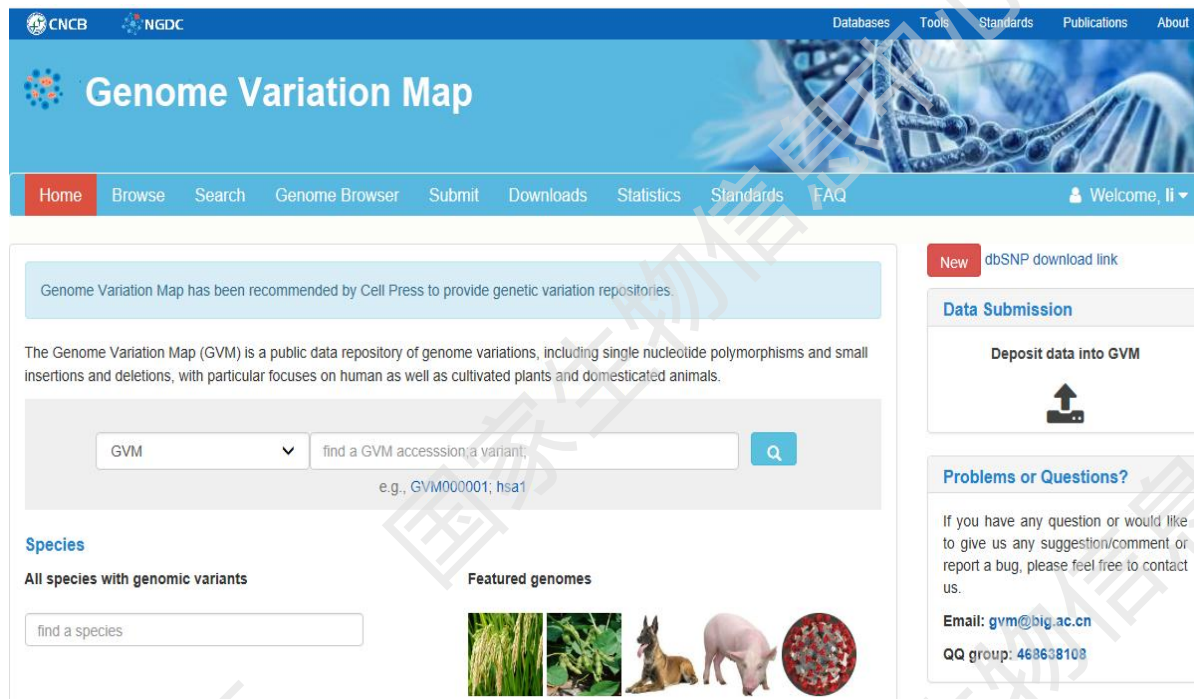
Top 5 Species

Species	Percentage
Human	92.8%
Sorghum	17.6%
Zebrafish	12.3%
Cow	9.3%
Soybean	6.3%

Top 5 Types

Type	Percentage
Control Set	57.4%
Case-Set	17.6%
Case-Cont	12.3%
Somatic	9.3%
Tumor vs. Matched-N	6.3%

基因组变异数据库GVM



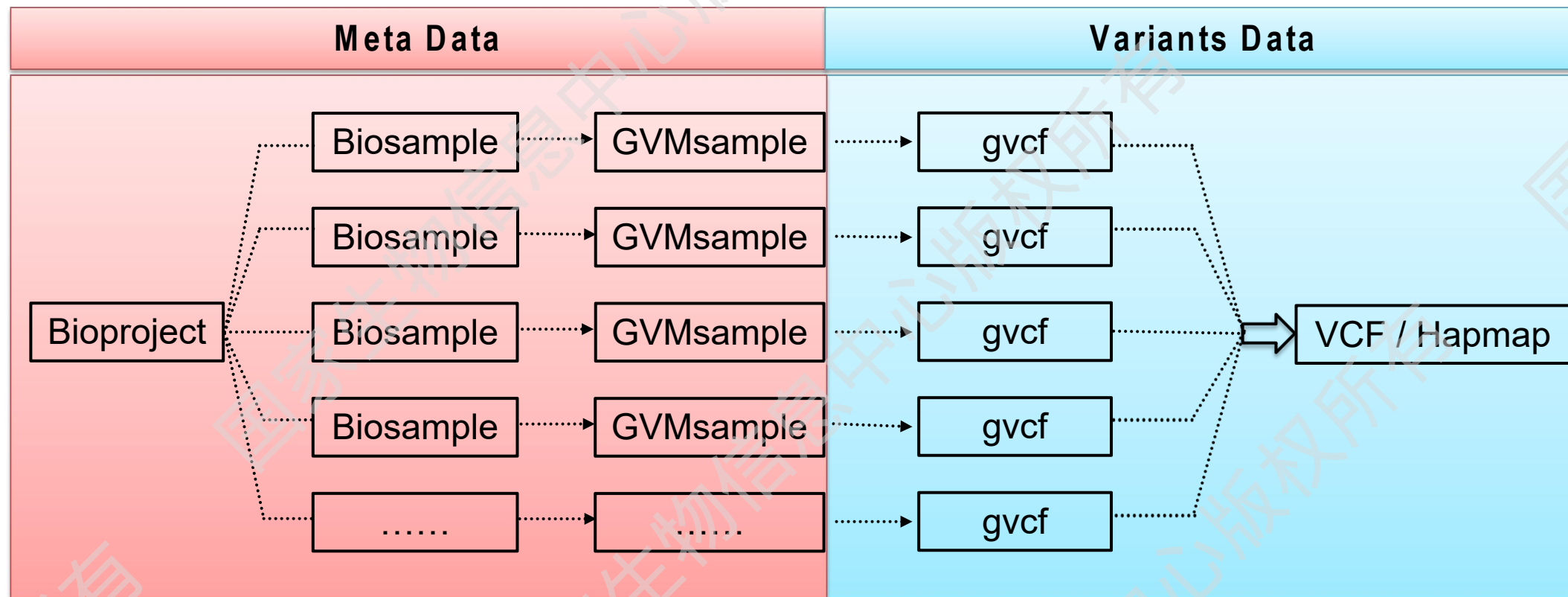
- 我国最大变异资源库，采用标准分析方法鉴定的物种**参考变异数据集**及功能注释信息
- 我国**首个**基因组变异数据汇交、管理与共享平台
- 2017年上线，访客数2.7万，累积下载**3500+万次**

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.”

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@big.ac.cn) if you have any suggestions.

1. Metadata

1.1 Submitter details

1.2 Project info

1.3 Sample details

1.4 Analysis methods

1.5 File names

2. File formats

2.1 VCF

2.2 HapMap

3. Data analysis standards

3.1 SNP & small I/D

3.1.1 Sequencing data analysis

3.1.2 Array data analysis

3.2 Large I/D (coming soon)

3.3 CNV (unavailable)

3.4 SV(unavailable)

4. Nomenclature standards

4.1 GVM Nomenclature

4.2 Variation ID Nomenclature

4.2.1 SNP & small I/D

4.2.2 Large I/D (NA)

4.2.3 CNV (NA)

4.2.4 SV (NA)

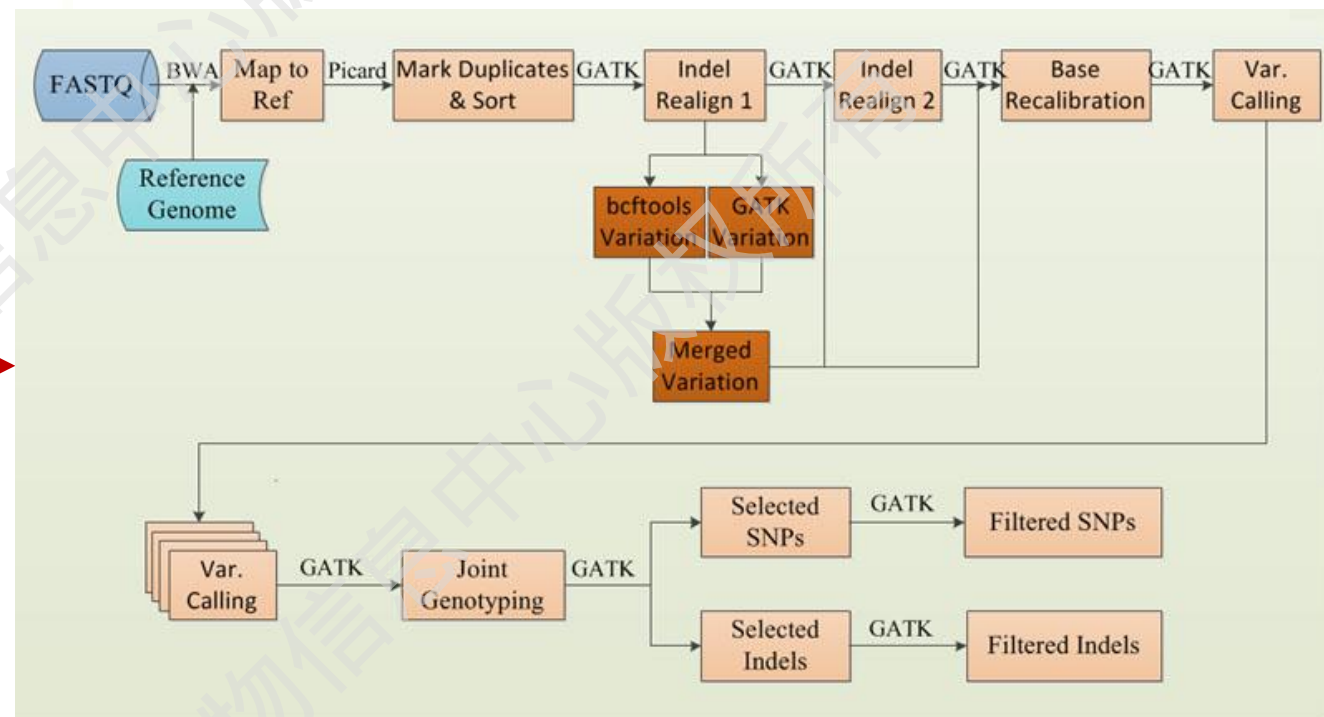
4.3 Shorten name of the organisms

4.4 Variation annotation

4.5 Pfam Annotate

4.6 Nucleotides and amino acid nomenclature

5. Tutorials



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

数据递交与管理

Home Browse Search Genome Browser **Submit** Downloads Statistics Standards FAQ

Welcome, li ▾

GVM / **Submit** / My Submission

GVM is a repository specialized for archiving genome variation data (VCF, GVCF or HapMap format). Creating a new GVM submission, three steps are needed:

1. Create a BioProject (an overall description of a single research initiative) in [BioProject submission portal](#).
2. Create a BioSample (description of the biological source material) in [BioSample submission portal](#).
3. Create a GVM Submission by clicking 'New Submission' below.

Any question, problems or suggestions, please feel free to contact us

Email: gvm@big.ac.cn

QQ group: 468638108

GVM Submission Wizard Help **English** Chinese

中英文帮助文档，仔细阅读，减少错误率！

New Submission

数据递交与管理

665
PROJECTS

503647
SAMPLES

Filter

Accession	BioProject	Description
GVM000873	PRJCA031292	population snp vcf file for 3 oes
GVM000869	PRJCA031045	Genetic variation in NyuW oject.
GVM000871	PRJCA031156	The whole exome sequen m DPM patients indicated nic mutations in the STAT4 ntified. This finding sugges T4 gene may not be implic ase mechanism for the
GVM000868	PRJCA031030	CSNK2B gene varitions in n, including c.175+1G>A, 91+1G>A and c.46 tuerA.

GVM / All Submission / GVM000757

Submission

GVM000757

2024-05-20

Sun Yat-sen University Cancer Center

Organism

Homo sapiens

Version

GRCh37

BioProject

[PRJCA026167](#)

Sample numbers

228

Abstract

A retrospective study on the identification of gene signature for immunotherapy in urological tumors

Data Accessibility:

Controlled access

Request

Release date

2024-05-16

Available data

data submitter: shiyx@sysucc.org.cn

2110115928-1_2110115928-1.normal.vcf

2109034337-1_2109034337-1.normal.vcf

2108092984-1_2108092984-1.normal.vcf

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>

国家生物信息中心
China National Center for Bioinformation

Data Resources Computing Analysis Data Network Standards

中文 English

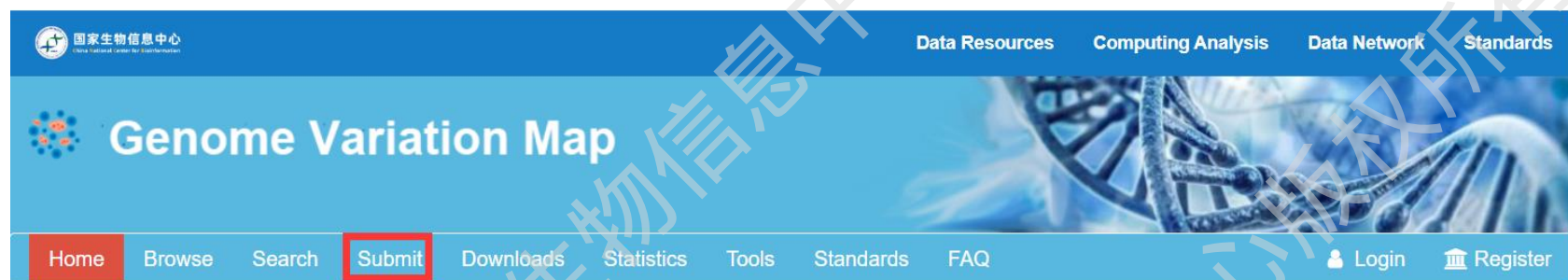
BIG Sub
BIG Submission Portal

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

- 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**
组学原始数据汇交、存储、管理与共享系统
- GSA for Human**
人类遗传资源组学原始数据库
- Genome Warehouse**
多物种基因组组装序列和注释数据归档库
- GenBase**
核酸和蛋白序列及注释数据归档库
- OMIX**
收集和共享表达谱、甲基化谱、蛋白质组学、代谢组学等数据的资源库
- Genome Variation Map**
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

数据提交入口

◆ 入口URL: <https://ngdc.cncb.ac.cn/gvm>



Genome Variation Map has been recommended by Cell Press to provide genetic variation repositories.

The Genome Variation Map (GVM) is a public data repository of genome variations, including single nucleotide polymorphisms and small insertions and deletions, with particular focuses on human as well as cultivated plants and domesticated animals.

GVM



find a GVM accesssion;a variant;



e.g., GVM000001; hsa1

Species

All species with genomic variants

find a species

Featured genomes



Data Statistics

New dbSNP download link

Data Submission

Deposit data into GVM



Problems or Questions?

If you have any question or would like to give us any suggestion/comment or report a bug, please feel free to contact us.

Email: gvm@big.ac.cn

QQ group: 468638108

Variants related resources

NGDC Central Authentication Service

Enter your Username and Password

Email

Password

[Forgot password?](#)

Check code

☐ Keep me signed in

Login

Reset

Register

在线填写汇交申请

GVM / Submit / My Submission

GVM is a repository specialized for archiving genome variation data (VCF, GVCF or HapMap format). Creating a new GVM submission, three steps are needed:

1. Create a BioProject (an overall description of a single research initiative) in [BioProject submission portal](#).
2. Create a BioSample (description of the biological source material) in [BioSample submission portal](#).
3. Create a GVM Submission by clicking 'New Submission' below.

Any question, problems or suggestions, please feel free to contact us

Email: gvm@big.ac.cn

QQ group: 468638108

GVM Submission Wizard Help

[English](#) [Chinese](#)

中英文帮助文档，仔细阅读，减少错误率！

New Submission

GVM / Data Submission / New Submission

Step 1
Submitter

Step 2
General information

Step 3
Meta-data Information

Step 4
Overview

Submitter

• First name

li

Middle name

• Last name

cp

• Email address

licp@big.ac.cn

Telephone number

• Laboratory/Centre

NGDC

• Company/Institute/University

Beijing Institute of Genomics, Chinese Academy of Sciences

Save and next

项目信息填写

GVM / Data Submission / New Submission

Step 1
Submitter

Step 2
General information

Step 3
Meta-data information

Step 4
Overview

BioProject

• BioProject accession

OR Go to create [BioProject](#)

没有bioproject: 点击creat
bioproject, 注册bioproject号。

The BioProject bundles the data for this research project, which is an overall description of a single research initiative.

Project title

Abstract information

已有bioproject: 下拉选择当前申请汇交变
异数据所属的bioproject号。

样本信息填写

★ BioSample

- ☒ No GVM related BioSample information was created
- ☐ GVM related BioSample information has been created

没有BioSample

已有BioSample

If you have not created GVM related Biological Sample(s), please select the 'No GVM related BioSample information was created' and create BioSample(s) in this wizard.
If you have already created GVM related Biological Sample(s), please select the 'GVM related BioSample information has been created'.

GVM

★ Release Date for GVM

- ☒ Release immediately following curation (recommended)
- ☐ Release on specified date

选择数据释放日期，立即释放 / 择期释放

★ Abstract for the variation data

变异信息描述

★ Short title for the variation data (less than 150 characters)

简短标题

Step 4
Overview

Template 下载离线模板

Column 2, 3 and 4: If the sample is genotyped by array, you should select "Genotyping by array" as the sequencing strategy from the drop-down list in column 2 and select the sequencing platform from the drop-down list in column 3.

CAUTION: Be aware that Excel may automatically apply formatting to your data. In particular, the following table shows that the first column of the data table is automatically formatted as text.

1	2	3	4
D	strategy	platform	array_name
S1	WGS	Illumina HiSeq 2000	
S2	WGS	Illumina HiSeq 2000	
S3	WGS	Illumina HiSeq 2000	

The screenshot shows a red box around the cell containing 'D' in the first column of the data table, indicating that the first column is automatically formatted as text. A red arrow points to the cell containing 'D'.

样本信息填写

• Please upload GVM template file

Select file...

选择已填好的离线模板

Browse ...

Download the GVM Submission [Template](#)

• Please upload GVM template file

上传已填好的离线模板

GVM Submission Template V3.0.xlsx

Remove

Upload

Browse ...

Download the GVM Submission [Template](#)

• Please upload GVM template file

GVM_Submission_Template_V3.0.xlsx

Check

Remove

校验元信息格式及内容

• Please upload GVM template file

GVM_Submission_Template_V3.0.xlsx

Check

Remove

Validate Success !

Save and next

下一步

元信息填写

You now at: [Submit](#) > [Submission](#)

Step 1
Submitter

Step 2
General information

Step 3
Meta-data Information

Step 4
Overview

Overview

Please transmitting your data files to the GVM FTP site (download FileZilla client)

Address: <ftp://submit.big.ac.cn>
User: Same as you login the GVM
Password: Same as you login the GVM

Finish

点击链接，下载文件上传工具 FileZilla

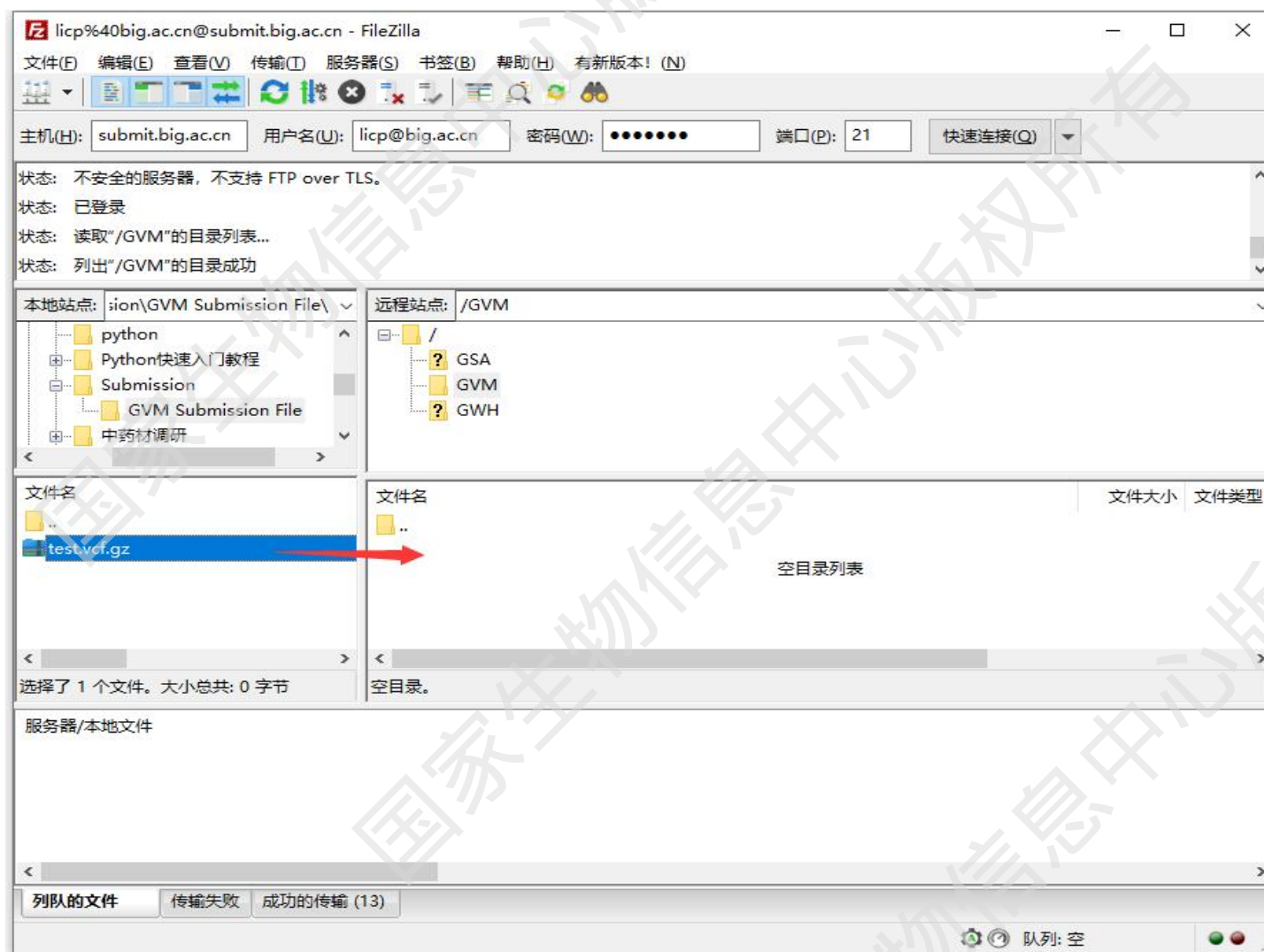
<https://filezilla-project.org/download.php?type=client>

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	Admin Operation
Unassigned	sub000254	PRJCA003421	SAMC237135, SAMC237136 more	2020-09-30	2020-09-09 20:06:24	llcp@big.ac.cn	2020-09-09 20:47:09	Meta-data Finished	Confidential	audit

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