中国科学院北京基因组研究所（国家生物信息中心）
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES / CHINA NATIONAL CENTER FOR BIOINFORMATION

# 基因序列数据汇交共享及同源基因数据库

唐碧霞 高级工程师

2024-10-26

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

# 目录
CONTENTS

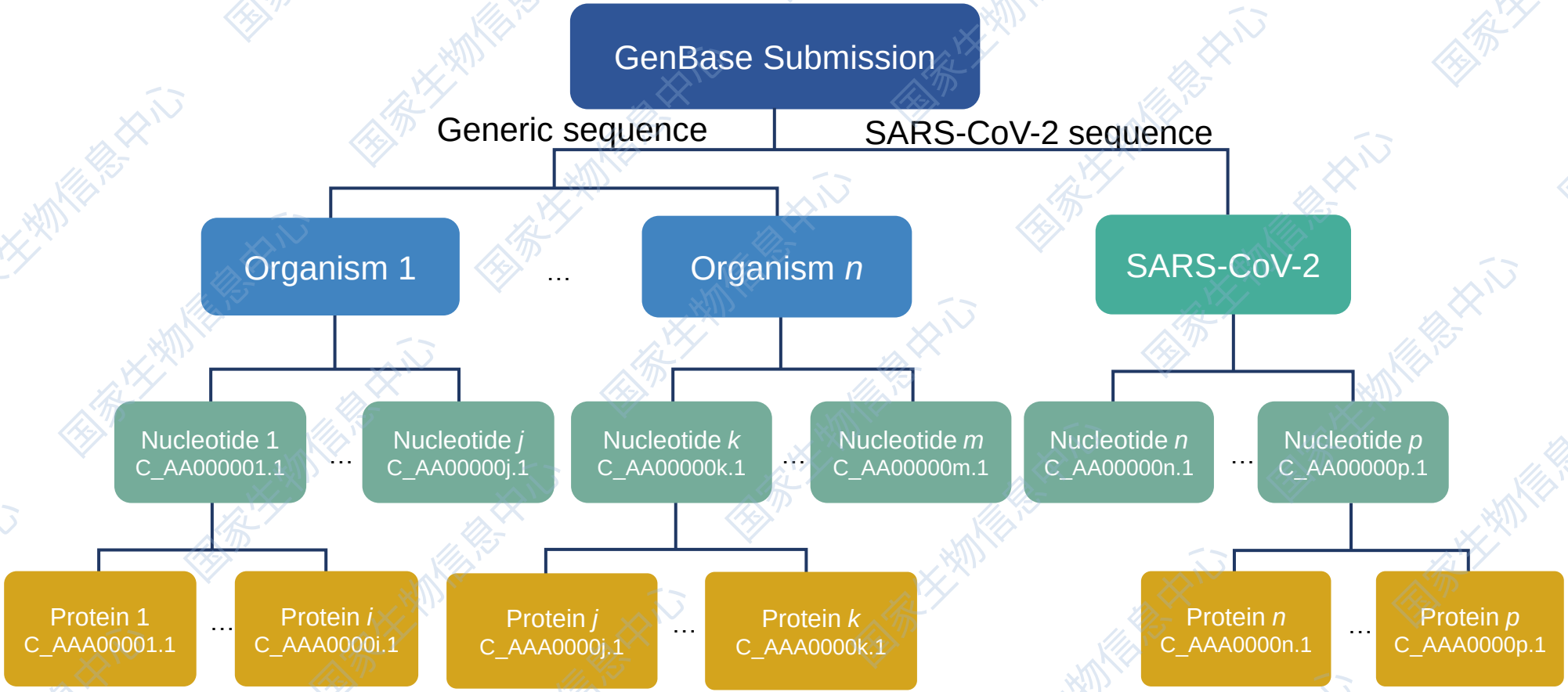
# 基因序列数据库GenBase简介

GenBase是一个公共、免费的基因序列数据归档库，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**


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

*Genomics Proteomics Bioinformatics, 2024*

# GenBase的数据模型




**Nucleotide accession: C_ + 2 characters+ 6 numbers + version number**

# 基因序列库GenBase的用户提交数据情况

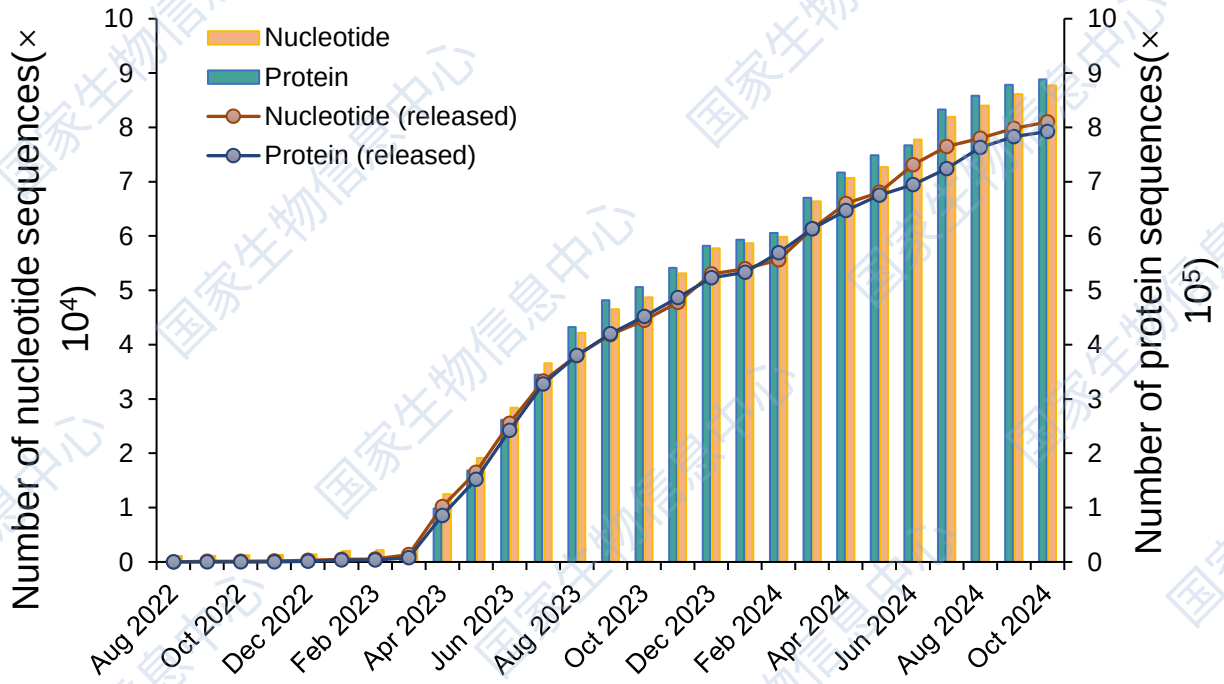
348
SUBMITTER

2,822 (2,336)
SUBMISSION

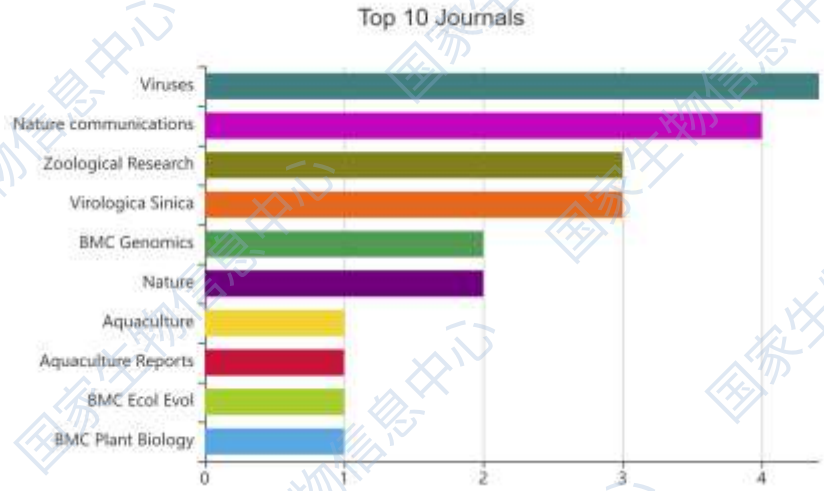
1,098,190 (875,013)
SEQUENCE

97,838 (81,099)
NUCLEOTIDE

1,000,352 (793,914)
PROTEIN




*Genomics Proteomics Bioinformatics*（2024）

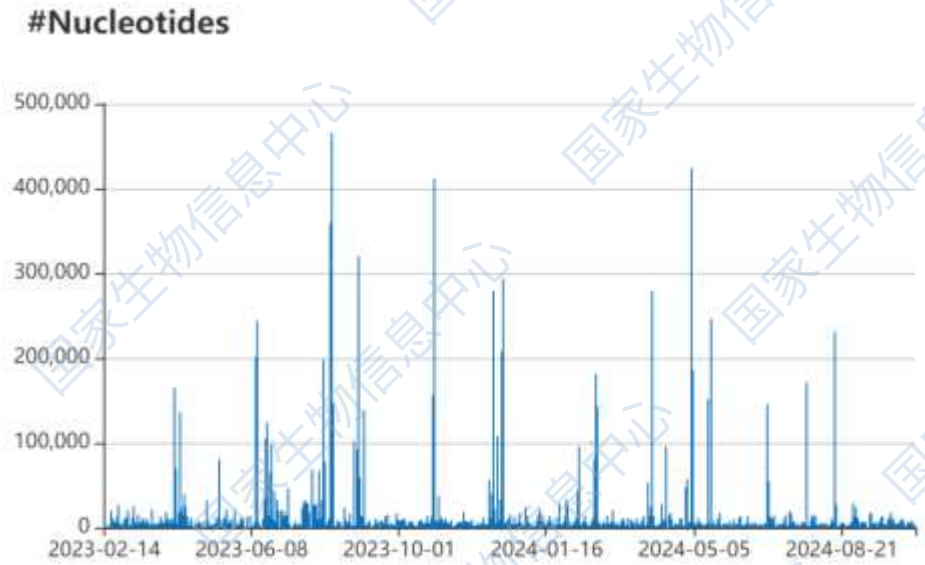


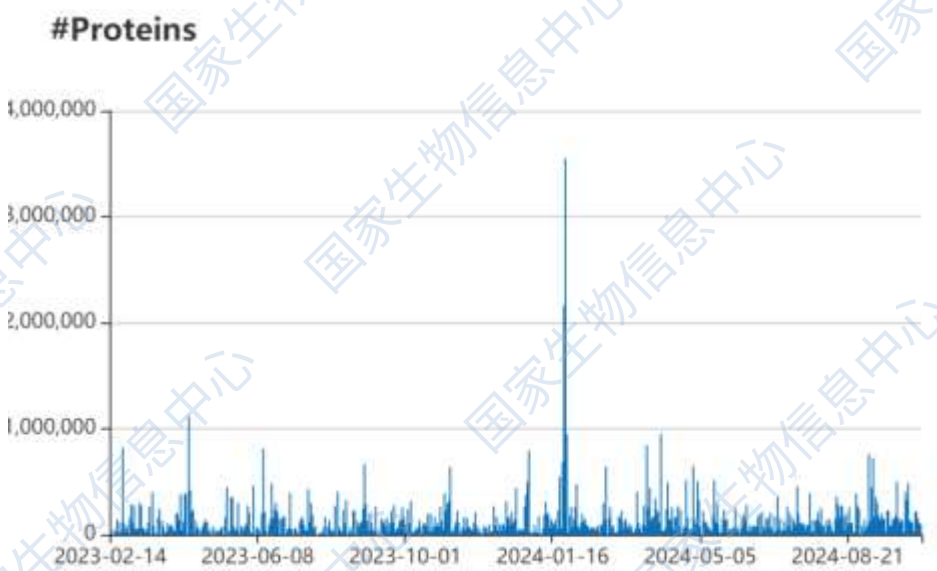

re3data.org

FAIRsharing.org
standards, databases, policies

**Since March 2023,** GenBase has supported data submissions for 67 articles and 285 funded projects.

# NCBI GenBank数据每日镜像




The total number of INSDC nucleotides sequences is **272,165,418**

The total number of INSDC proteins sequences is **329,065,402**

# 基因序列和基因组汇交的异同

| 数据库 | 收录范围 | 数据类型 | 数据关联 |
| --- | --- | --- | --- |
| 基因序列库GenBase | 小基因组和基因<br>• 病毒基因组<br>• 细胞器基因组<br>• 质粒基因组<br>• 所有物种基因片段 | • **基因片段**序列<br>• 注释数据（非必需） | **可以不涉及** BioProject 和 BioSample |
| 基因组数据库GWH | 大基因组<br>• 真核生物全基因组<br>• 原核生物全基因组<br>• 宏基因组组装数据 | • **基因组**组装序列<br>• 注释数据（非必需） | **涉及**BioProject和BioSample |

# 国际核酸数据库联盟INSDC对标库

| 国家基因组科学数据中心 | 数据资源 | 美国NCBI | 欧洲EBI | 日本DDBJ |
| --- | --- | --- | --- | --- |
| 基因组数据库<br>GWH | 重注释 | RefSeq | Ensembl | / |
| 基因序列库<br>GenBase | 用户递交 | GenBank：全基因组<br>WGS | ENA:<br>Assembled/annotated sequence | DDBJ |
| | 用户递交 | GenBank：传统<br>GenBank | | |

# 目录
CONTENTS

# 基因序列数据汇交共享——GenBase

# GenBase数据提交模块

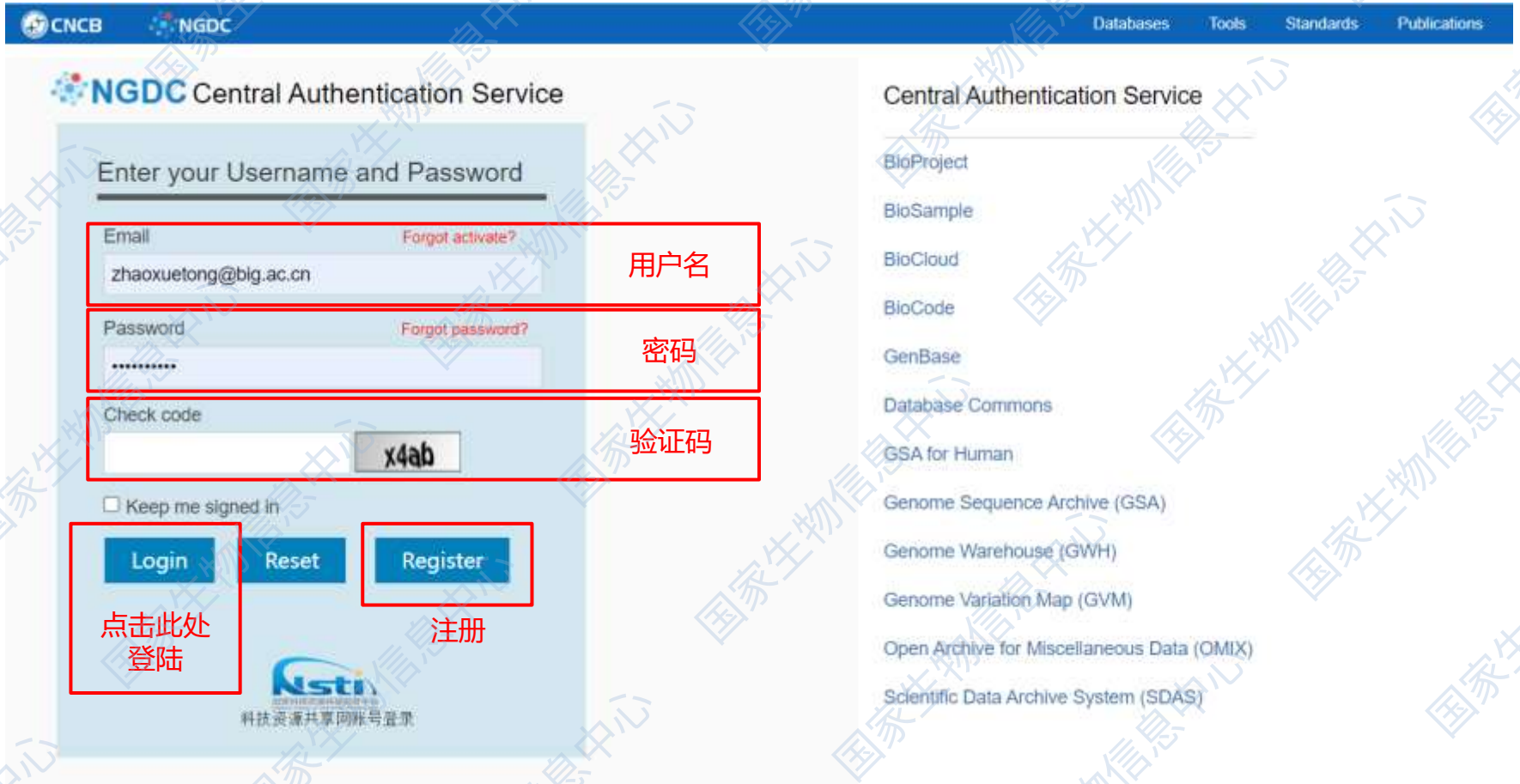
- 三类数据提交模块
  - 常规序列提交
  - 新冠序列提交
    - 自动注释
  - 受控序列提交
    - 人类遗传核酸序列


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

# 准备工作：1.账户准备-注册和登陆

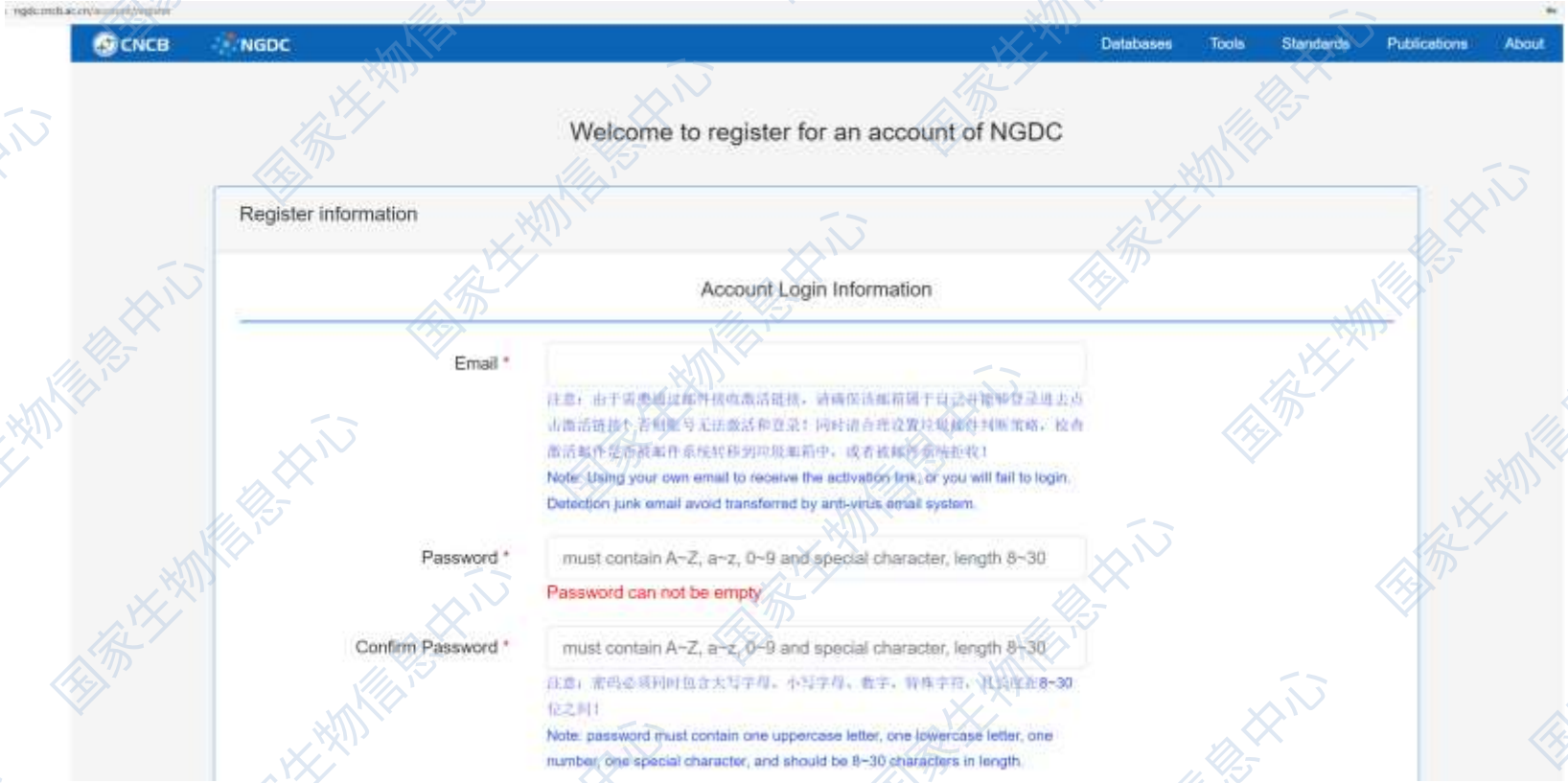

数据库网址：https://ngdc.cncb.ac.cn/genbase/ 进入网站进行登陆

# 准备工作：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：1.账户准备-注册和登陆




注册网址：https://ngdc.cncb.ac.cn/account/register

如果您在账号注册和使用过程中遇到任何问题，请联系bigd-admin@big.ac.cn

# 准备工作：2. 常规序列提交准备


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

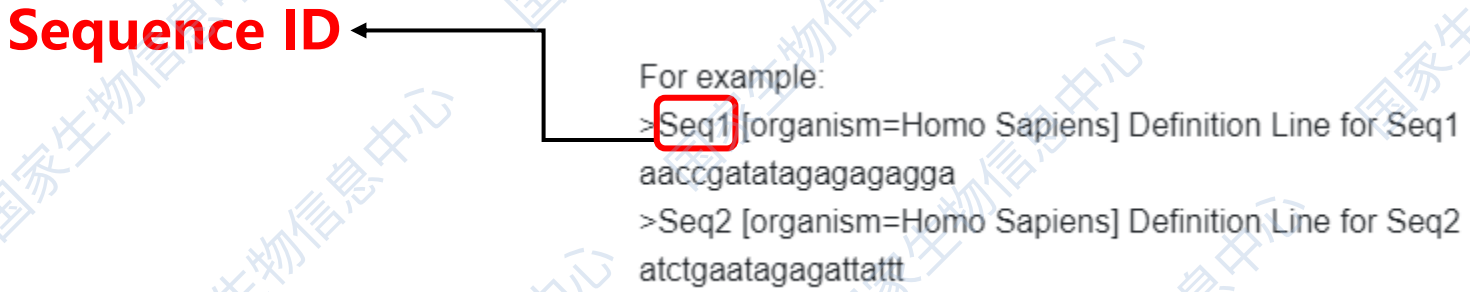
③ 元信息文件：包含collection date、country、host等

④ 特征注释文件：包含gene、CDS、tRNA、ncRNA等

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

# 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
- 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seq1

# 准备工作：2. 数据准备-元数据文件

- 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-元数据文件

## 下载后打开GenBase_Modifiers.xlsx模板：

#!Version: 3.3

#!DO NOT MODIFY HEAD LINES AND EXAMPLE LINE! ‖ 不要修改标题行和示例行！

#!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。

#!Specimens are identified in the Source Modifiers Table by the same Sequence_ID used in the FASTA file. ‖ Specimens 在 Source Modifiers Table 中由 FASTA 文件中使用的相同 Sequence_ID 标识。

#!The heading for the first column must be exactly Sequence_I

#!Each specimen in the set must have a line in the source modi

#!Each Sequence_ID may appear only once in the source mod

#!Fill in the data in the line beginning with "##number". If it is n

#!Reference: https://ngdc.cncb.ac.cn/genbase/standards ; https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers

#!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”、“n

**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**

于样本。

| ##Color Code | required | optional (recommended) | optional | | | | |
|---|---|---|---|---|---|---|---|
| ##Column Number | ##1 | ##2 | ##3 | ##4 | ##5 | ##6 | |
| ##Header | Sequence_ID | Collection_date | Country/Region | Clone | Host | Host_sex | |
| ##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 | Ag |
| ##Example(do not delete) | Seq1 | 2001-01-31 or value in ControlWords sheet | China or value in ControlWords sheet | C.Grant | Homo Sapiens | Male | 43 |
| ##1 | | | | | | | |
| ##2 | | | | | | | |
| ##3 | | | | | | | |
| ##4 | | | | | | | |
| ##5 | | | | | | | |
| ##6 | | | | | | | |
| ##7 | | | | | | | |

**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**

ControlWords　SourceModifierTable

控制词汇信息

# 准备工作：2. 数据准备-特征注释文件

- 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：**

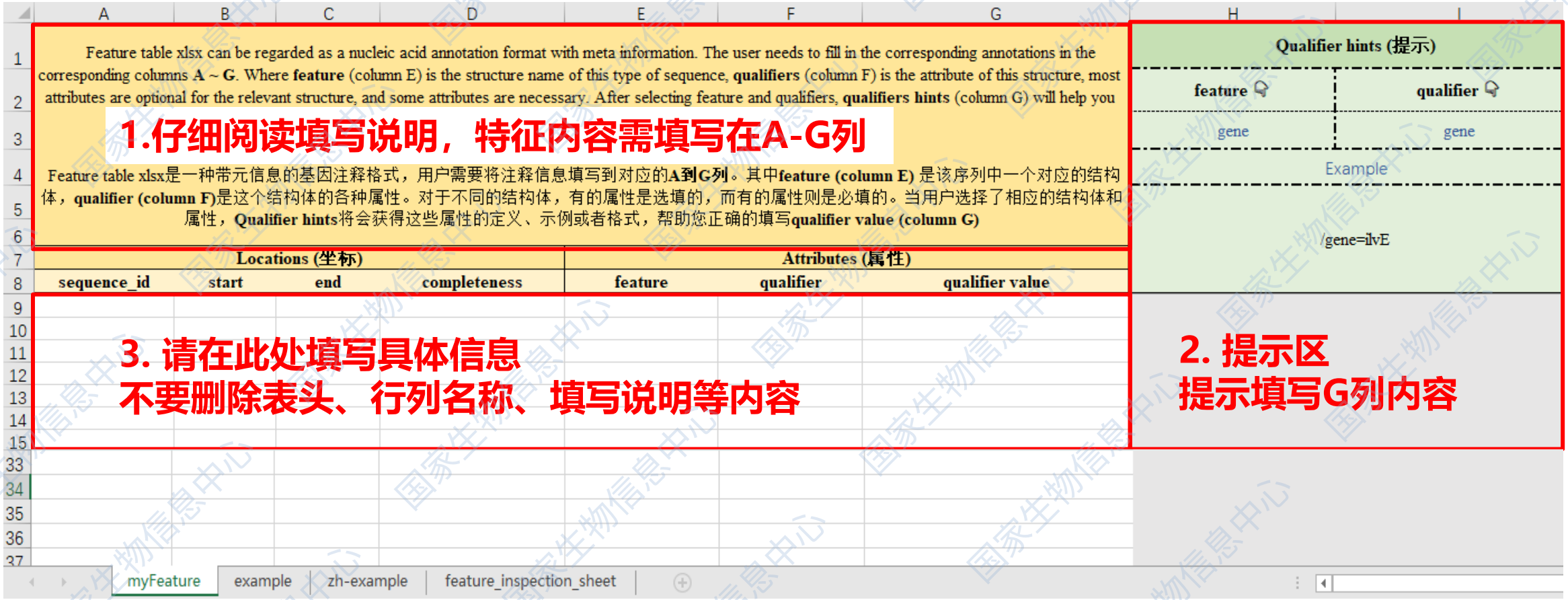
```
>Feature Sc_16
1       7000    REFERENCE
                        PubMed          8849441
<1      1050    gene
                        gene            ATH1
<1      1009    CDS
                        product         acid trehalase
                        product         Ath1p
                    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

**gff3格式：**

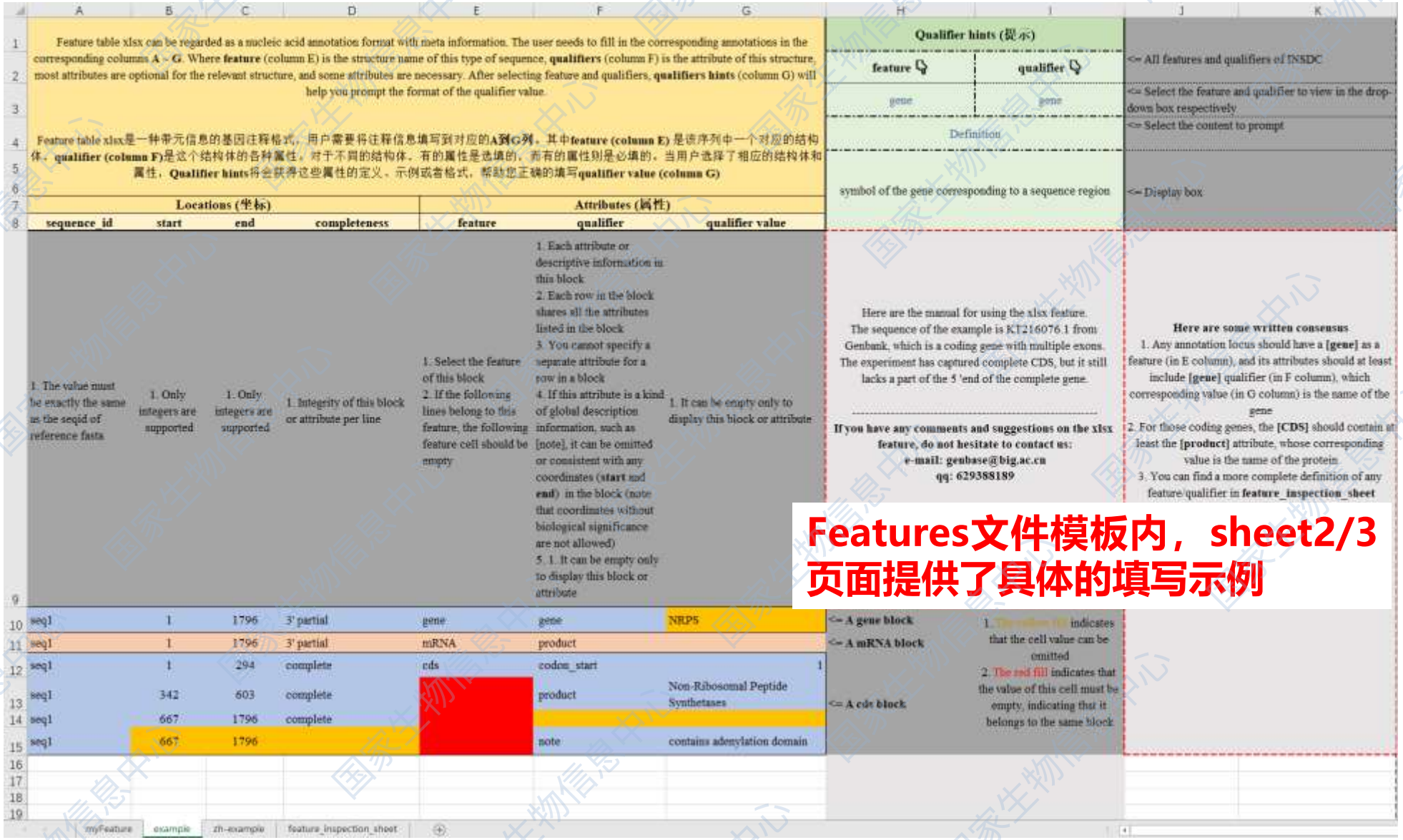
```
##gff-version 3.1.26
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . 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-特征注释文件

**下载后打开GenBase_Features.xlsx模板：**



# 准备工作：2. 数据准备-特征注释文件



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<br>2. If the following lines belong to this feature, the following feature cell should be empty | 1. Each attribute or descriptive information is this block<br>2. Each row in the block shares all the attributes listed in the block<br>3. You cannot specify a separate attribute for a row in a block<br>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)<br>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 | NRP5 |
| 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**

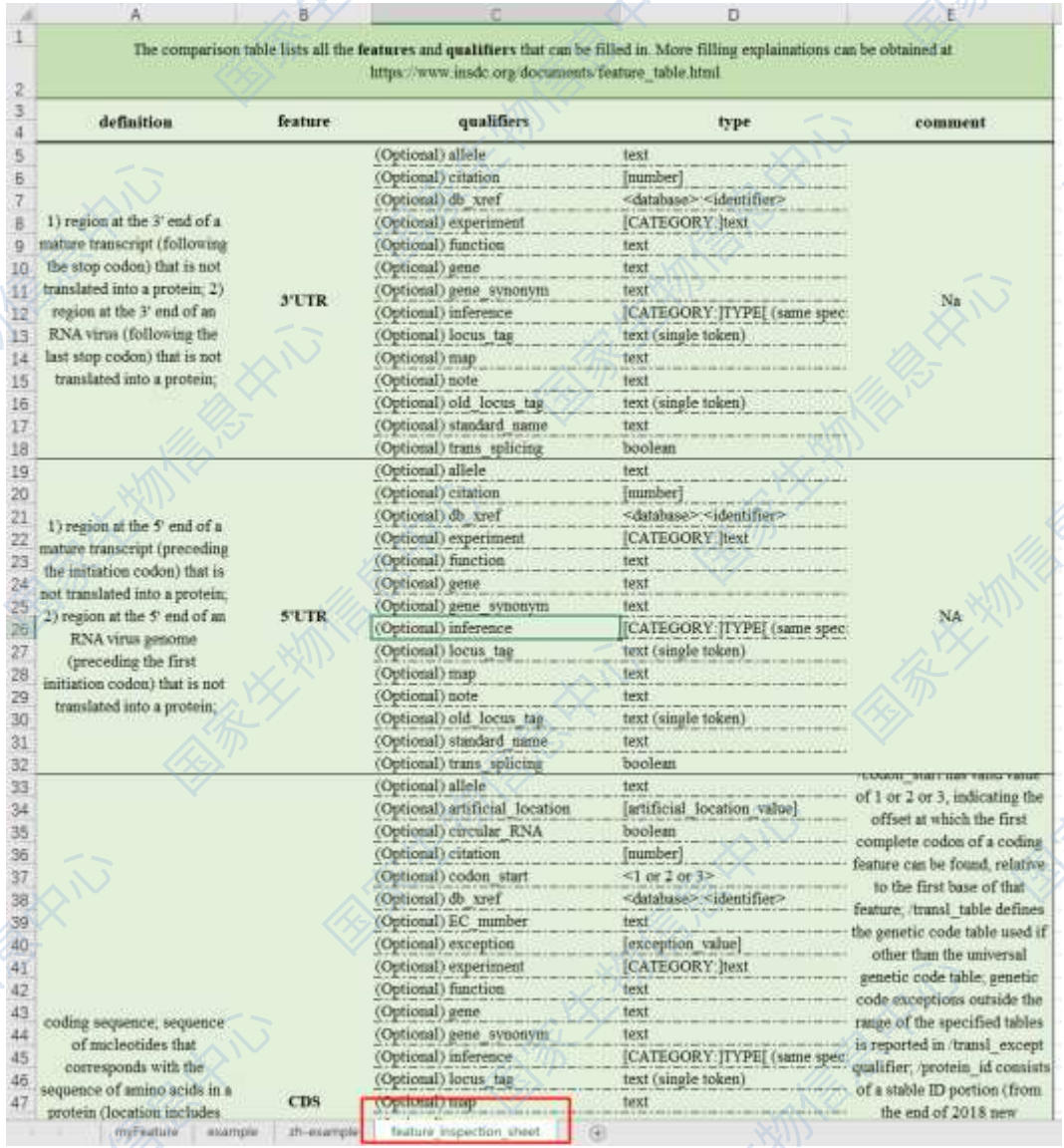
<= A gene block
<= A mRNA block
<= A cds block

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

myFeature | example | zh-example | feature_inspection_sheet

**Features文件模板内，sheet2/3页面提供了具体的填写示例**

# 准备工作：2. 数据准备-特征注释文件



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

| 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 | 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**

# 准备工作：2. 数据准备-特征注释文件

## 注释文件规范

参考https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md和https://www.ncbi.nlm.nih.gov/genbank/feature_table/.

1）注释文件

包含基因、外显子区、编码区、非翻译区、转座子、重复区等信息，上述特征信息都需要放进注释文件excel、GFF或者TBL文件中，作为注释文件上传至GenBase。

2）序列注释示例

- **Eukaryotic gene**

Excel格式示例

| Locations (坐标) | | | | Attributes (属性) | | |
|---|---|---|---|---|---|---|
| sequence_id | start | end | completeness | feature | qualifier | qualifier value |
| seq1 | 1 | 9 | 5' partial | regulatory | gene | ubc42 |
| seq1 | | | | | regulatory_class(Mar | promoter |
| seq1 | 10 | 567 | complete | mRNA | gene | ubc42 |
| seq1 | 789 | 1320 | | | | |
| seq1 | 54 | 567 | complete | CDS | gene | ubc42 |
| seq1 | 789 | 1254 | | | product | ubiquitin conjugating enzyme |
| seq1 | | | | | function | cell division control |
| seq1 | 10 | 567 | complete | exon | number | 1 |
| seq1 | | | | | gene | ubc42 |
| seq1 | 568 | 788 | complete | intron | number | 1 |
| seq1 | | | | | gene | ubc42 |
| seq1 | 789 | 1320 | complete | exon | number | 2 |
| seq1 | | | | | gene | ubc42 |
| seq1 | 1310 | 1317 | complete | regulatory | regulatory_class(Mar | polyA_signal_sequence |
| seq1 | | | | | gene | ubc42 |

TBL格式示例

```
>Feature seq1

<1  9   regulatory
            gene    ubc42
            regulatory_class    promoter
10  567 mRNA
789 1320
            gene    ubc42

54  567 CDS
```


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

**8种类型**

- **Eukaryotic gene**
- **Bacterial operon**
- **Artificial cloning vector (circular)**
- **Plasmid**
- **Repeat element**
- **Immunoglobulin heavy chain**
- **T-cell receptor**
- **Transfer RNA**

# 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


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

# 提交流程： 步骤1. 提交者信息 (Submitter)

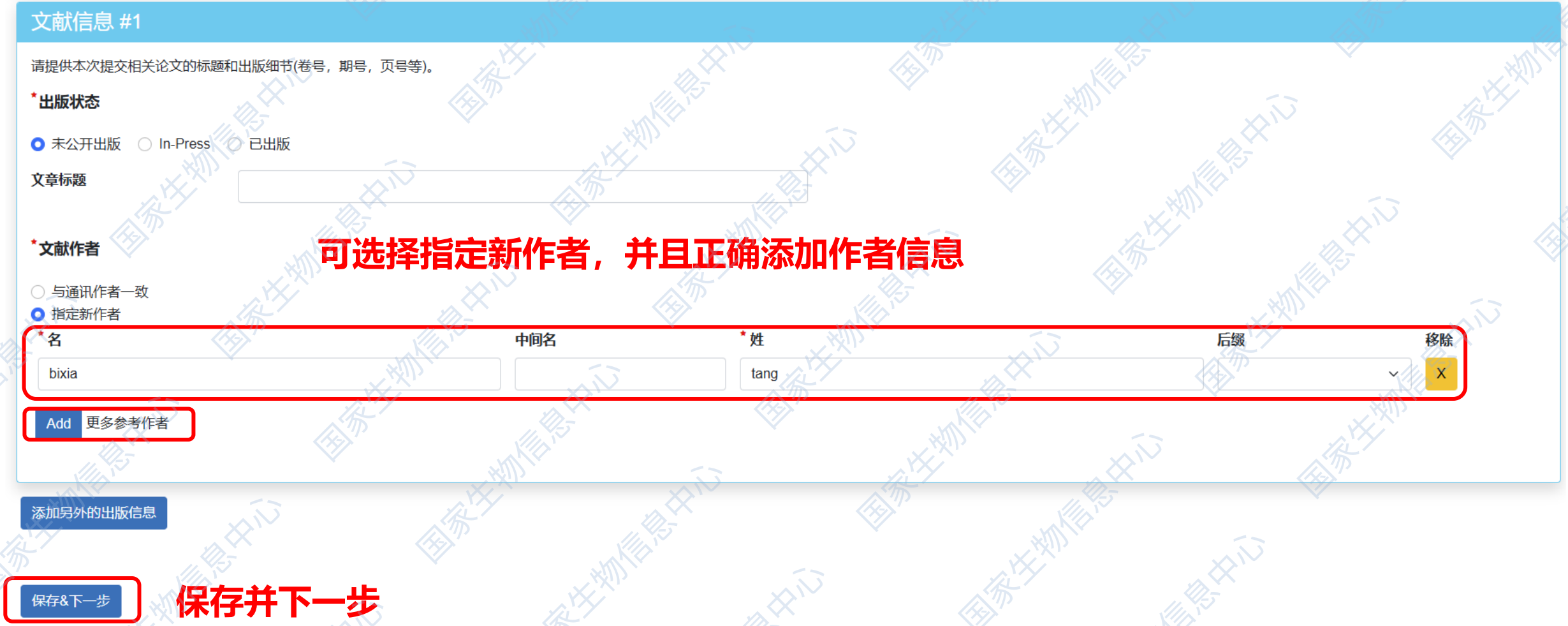

**保存并下一步**

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，信息将反馈到您的**注册邮箱**，而非此处填入的提交者信息邮箱

# 提交流程： 步骤2. 参考文献信息(Reference)

# 提交流程： 步骤2. 参考文献信息(Reference)



# 提交流程： 步骤3. 技术信息 (Technology)


技术信息

提交 #0000062

## 测序技术

**如果您提交的序列超过 500条 或者您的序列是使用 下一代测序技术 生成的，则需要提供此信息。**

用什么技术获得这些序列?

- [x] Sanger dideoxy sequencing
- [ ] 454
- [ ] Helicos
- [ ] Illumina
- [ ] MGI
- [ ] MGISEQ-200RS
- [ ] MGISEQ-2000
- [ ] MGI DNBSEQ
- [ ] IonTorrent
- [ ] Nanopore
- [ ] PacBio
- [ ] SOLiD
- [ ] Other

上述序列为:

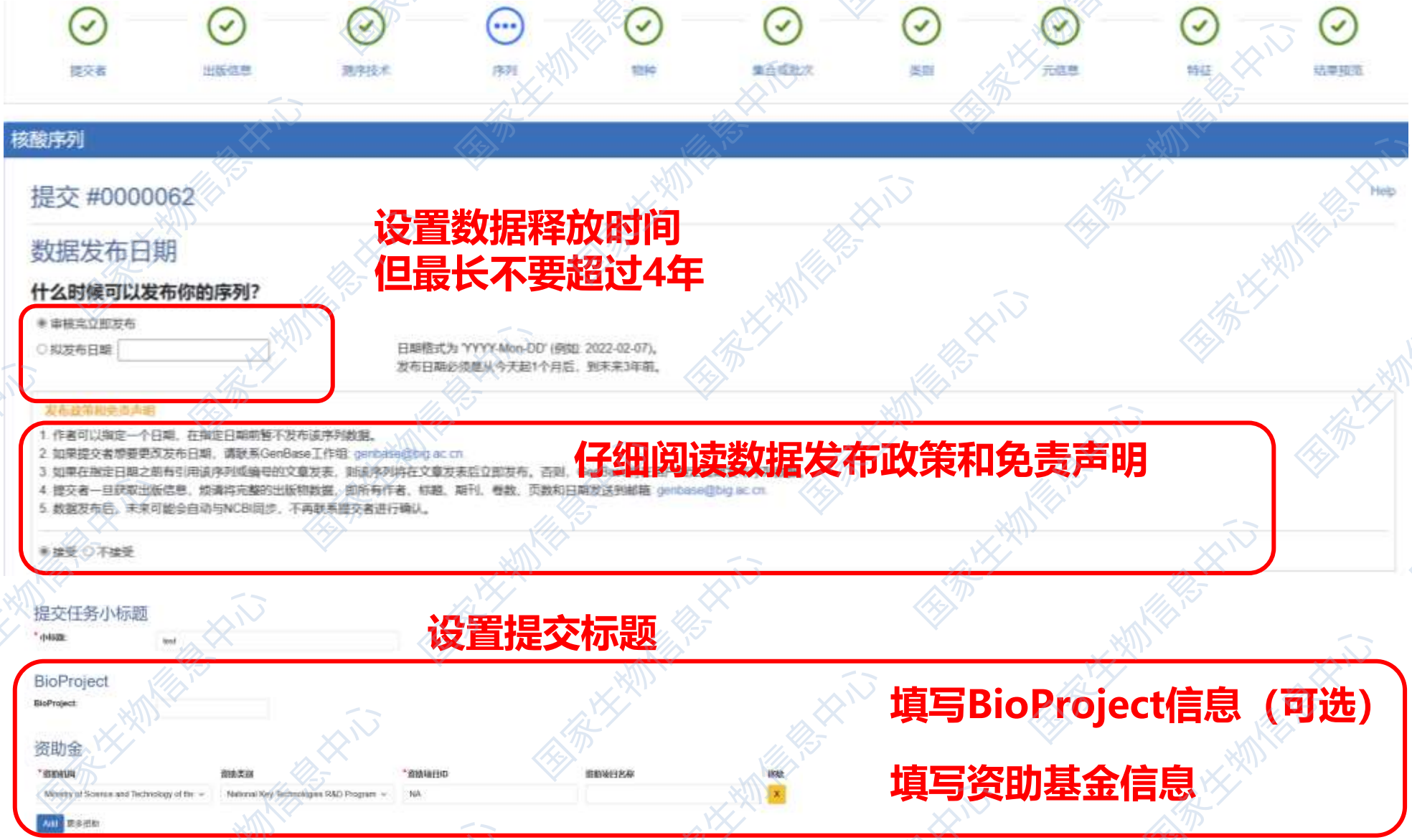
Assembled sequences

保存&下一步

**根据具体情况，详细填写相关信息**

**保存并下一步**

# 提交流程： 步骤4. 核酸序列 (Nucleotide)



# 提交流程： 步骤4. 核酸序列 (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**

**核酸序列填写示例**

# 提交流程： 步骤4. 核酸序列 (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**

# 提交流程： 步骤6. 集合/批次 (Set/Batch)


## 集合或批次

提交 #0000062

**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**

帮助

**一个集合中的所有序列必须来自相同的基因/locus，并且预计将在同一时间释放。**
**请选择集合（set）研究类型：**

- ◉ Pop set — **群体研究：** 通过对同一生物的不同分离株的同一基因进行测序而获得的一组序列。
- ○ Phy set — **系统发育研究：** 通过对不同生物的同一基因进行测序而获得的一组序列。
- ○ Mut set — **突变研究：** 通过对单个基因的多个突变进行测序而获得的一组序列。
- ○ Env set — **环境研究：** 通过对未分类或未知生物的同一基因进行测序而获得的一组序列。

**如果你的序列不是全部来自同一个基因/locus，也不打算同时释放，请选择下面的"Batch"。**

- ○ Batch — 多个相关的核苷酸序列，不来自同一基因，但可能来自同一研究或同一生物。

保存&下一步

**保存并下一步**

# 提交流程：步骤7. 类别确认 (Category)


## 类别

提交 #0000062

帮助

表明你的序列是否是原始提交。如果您的序列是第三方注释，请联系genbase@big.ac.cn。

◉ **原始提交** 由提交者直接测序

○ **第三方注释** 非自产数据的注释

保存&下一步

**确认您的序列是原始提交或第三方注释**

**注意：选择使用第三方注释的数据，在文章被接收后，提交的序列才会释放**

**保存并下一步**

# 提交流程： 步骤8. 提交Modifiers文件 (Modifiers)


元信息

提交 #0000062

帮助

请注意

1. 下载元信息模板文件 GenBase_Modifiers.xlsx, 在上传前填写并仔细检查
2. 有关列的解释和示例，请参见示例 GenBase_Modifiers.xlsx.

1.点击下载Modifiers文件模板

* 这些序列来自于细胞器基因组吗？ (如线粒体或叶绿体)

○ 是 ◉ 否 如果选择是, 请在GenBase_Modifiers.xlsx中填写Organelle/Location列

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

使用Excel格式上传元信息文件

0000062_modifier.xlsx Remove

2.点击上传填好的Modifiers文件

验证成功!

保存&下一步

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，comment注释会标记为：GenBase 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)

## 3.网页展示数据具体信息

```
LOCUS       OL3DN2379VN1            2483 bp    RNA    linear   UNK 16-AUG-2022
DEFINITION  Human picobirnavirus.
ACCESSION   OL3DN2379VN1
VERSION     OL3DN2379VN1
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="MIAPQLRYAELRESNRHNTTVEKETGRHHVETERIGVGTLAETSR
                     HGKAYEGIQSALAAETAAGVATGIHWYSAQWLAHLQFAQANETTSEVSETTRKHLAT
                     ESIQSGQLEETGRSLATEQYHINALGEQFRHHKQTELQGWISTGAGAYKDVSTGLRQV
                     VESVKGIIHGVKSILPGG"
     CDS             707..2474
                     /codon_start=1
                     /product="capsid protein"
                     /translation="MKDSNIIRGHSKYQGKAVSKRKEELDDQIHERYREDWDWMYA
```

# 提交流程： 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

# 提交流程：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 |

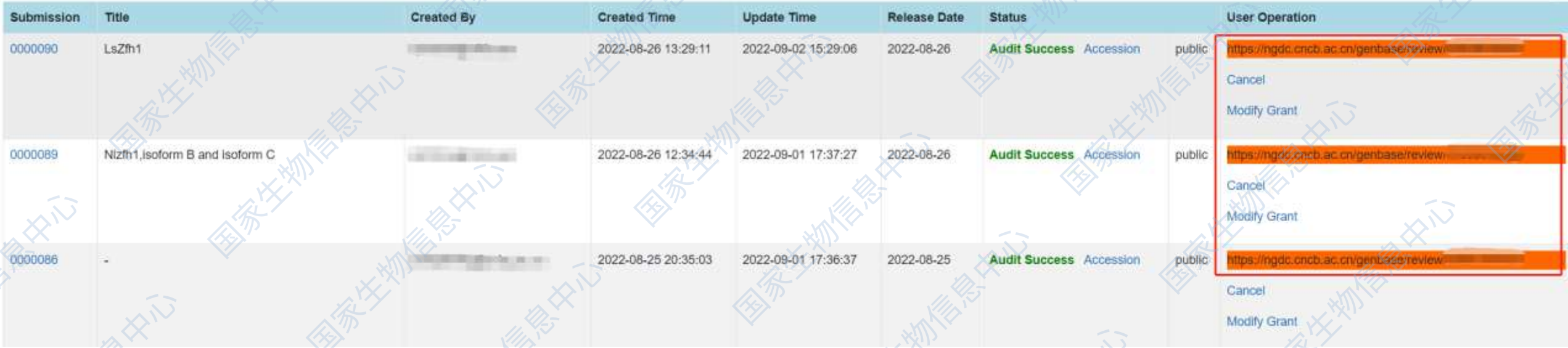
无错，点击Accession查看编号

| 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

# 提交流程：分享审稿人链接

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]<br>Cancel<br>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]<br>Cancel<br>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]<br>Cancel<br>Modify Grant |

3. 点击Modify Grant按钮

可修改资助信息

# 提交流程：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<br>Modify Grant<br>Modify ReleaseTime |

2. 数据处于私密状态

   只能自己点击查看

   可点击在线修改发布日期

# 提交流程：基因序列库归档编号

## 提交 #0000250

| | |
|---|---|
| **提交组织** | Beijing Changping Laboratory |
| **释放日期** | 2023-01-03 |
| **序列作者** | Cao Yunlong,Song Weiliang,Fu Haoyi,Ma Wentai,Liu Shujun,Yang Sijie,Li Mingkun |
| **测序技术** | Illumina |
| **分子类型** | genomic RNA |

**21** 提交的核酸序列

序列编号如下，可以在文章中引用。

| | | | |
|---|---|---|---|
| **CPL-Dec-50-2022** | **C_AA001438.1** | GBFF | Fasta |
| **CPL-Dec-51-2022** | **C_AA001439.1** | GBFF | Fasta |
| **CPL-Dec-52-2022** | **C_AA001440.1** | GBFF | Fasta |
| **CPL-Dec-53-2022** | **C_AA001441.1** | GBFF | Fasta |

## 文章引用

**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.

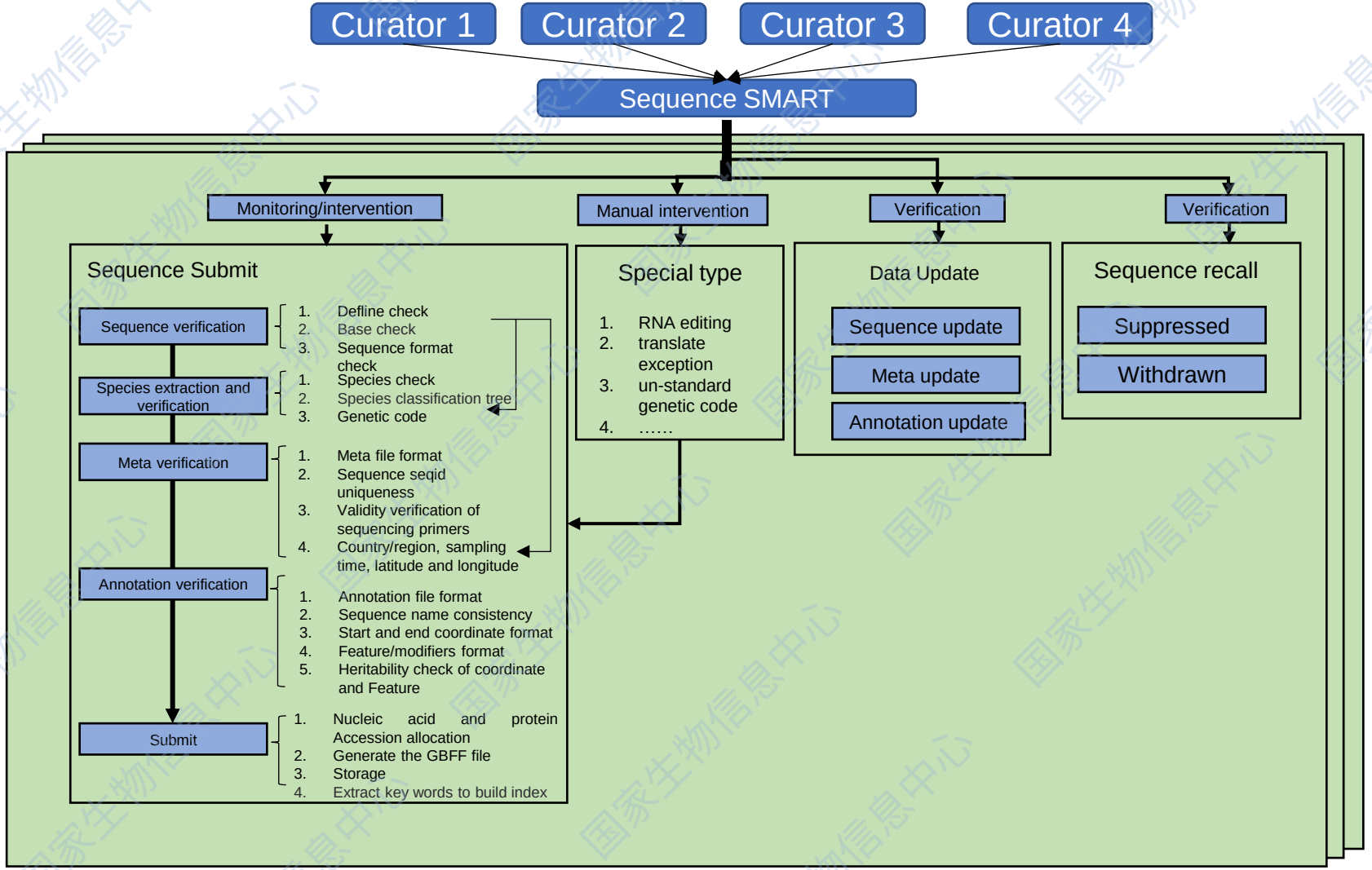
References:

[1] GenBase: A Nucleotide Sequence Database. Genomics Proteomics Bioinformatics 2024, 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].

# GenBase数据质量控制

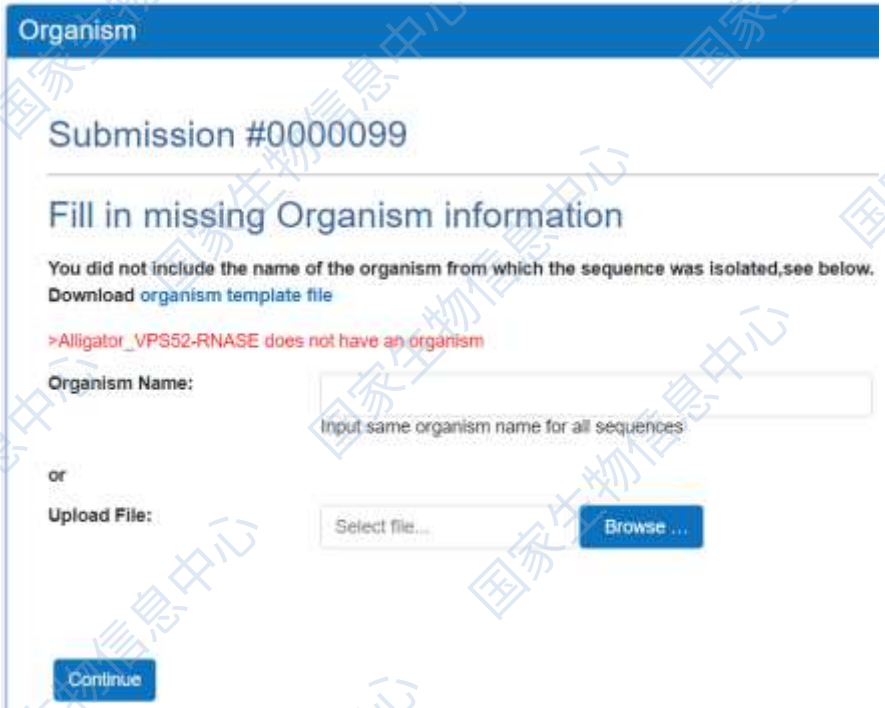
- 必填项，核酸序列，物种，元数据，特征信息在线实时校验
- Table2Asn对所填写的必要内容，尤其是注释信息的合法性，进行再校验
- 审核人员对物种，分子类型，遗传区，拓扑结构，注释合法性等进行再审核，保证序列提交质量



# 常见错误类型：物种名错误示例

**正确写法 [organism=Homo Sapiens]**

**否则会跳转到organism页面重新填物种名**

# 常见错误类型： Source modifiers错误示例

元信息

提交 #0000230

请注意

1. 下载元信息模板文件 GenBase_Modifiers.xlsx, 在上传前填写并仔细检查
2. 有关列的解释和示例，请参见示例 GenBase_Modifiers.xlsx

* 这些序列来自于细胞器基因组吗？ （如线粒体或叶绿体）

○ 是 ◉ 否 如果选择是, 请在GenBase_Modifiers.xlsx中填写Organelle/Location列

使用Excel格式上传元信息文件

0000230_modifier.xlsx Remove

result_error.txt

报错提示文件，可点击下载查看

报错描述

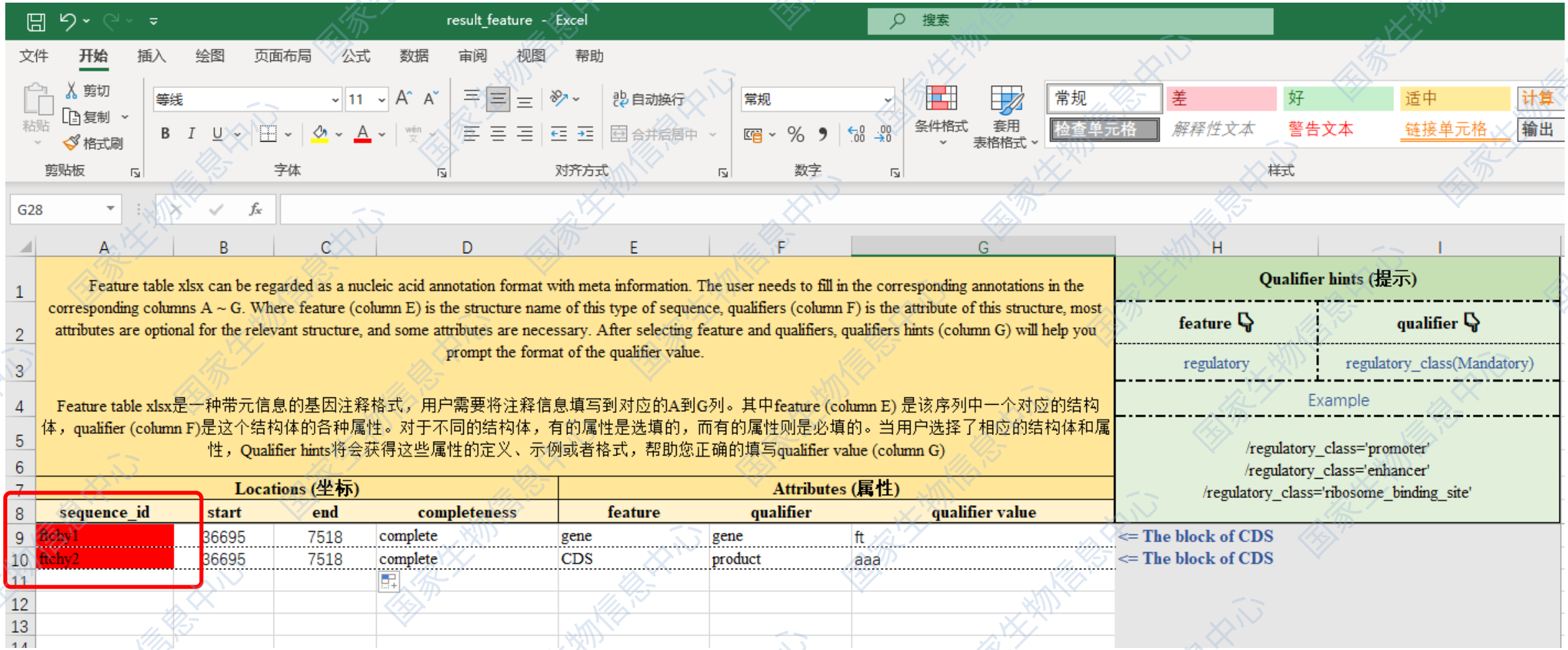
验证失败! 您提交的文件中存在错误，完整的错误信息显示在'result_error.txt'文件中，以下列出了具体前20行的错误消息。

| Error Type | Level | Modifier Name | Sequence Index | Message |
|---|---|---|---|---|
| Fatal | error | - | - | Columns ['chromosome', 'clone_lib', 'db_xref', 'map', 'mating_type', 'metagenome_source', 'type_material'] are not found in excel |
| Fatal | error | - | - | The version of Source Modifiers Excel is not correct, please use the latest version |

保存&下一步

# 常见错误类型： Feature table错误示例

# 常见错误类型： Feature table错误示例




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.

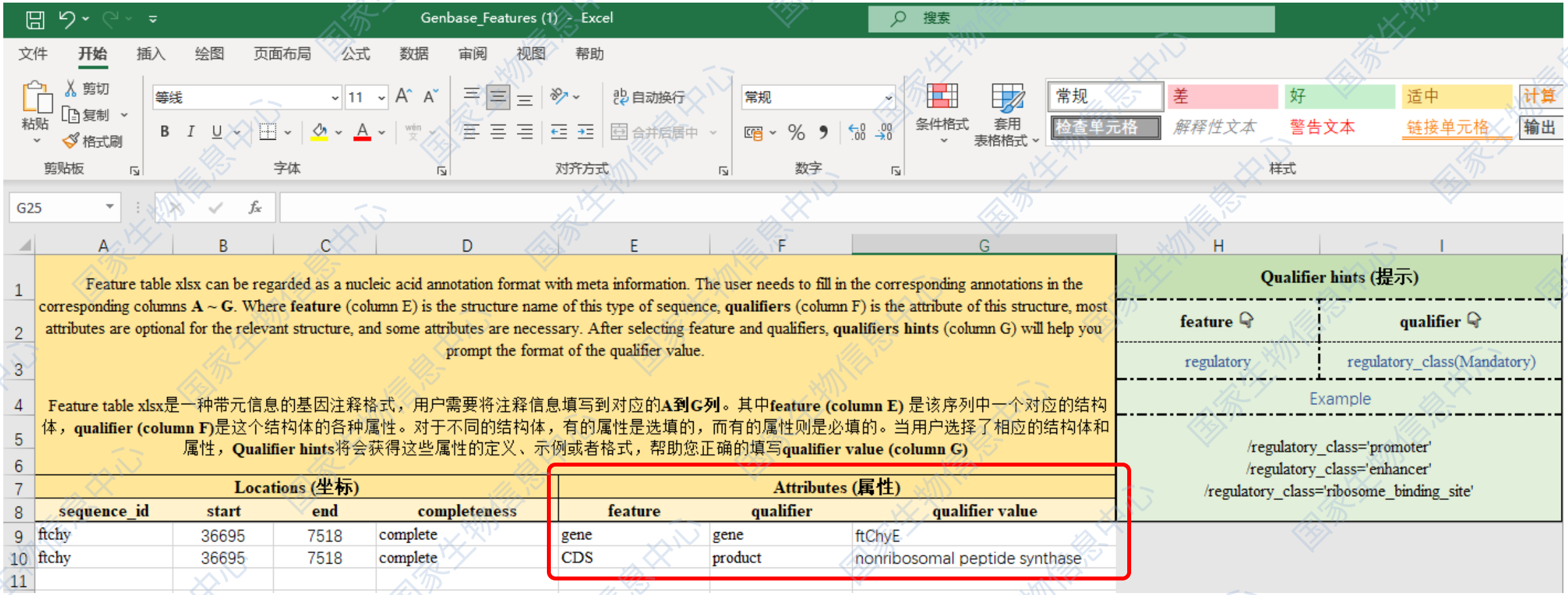
Feature table xlsx是一种带元信息的基因注释格式，用户需要将注释信息填写到对应的A到G列。其中feature (column E) 是该序列中一个对应的结构体，qualifier (column F)是这个结构体的各种属性。对于不同的结构体，有的属性是选填的，而有的属性则是必填的。当用户选择了相应的结构体和属性，Qualifier hints将会获得这些属性的定义、示例或者格式，帮助您正确的填写qualifier value (column G)

| Locations (坐标) | | | | Attributes (属性) | | | |
|---|---|---|---|---|---|---|---|
| sequence_id | start | end | completeness | feature | qualifier | qualifier value | |
| ftchy1 | 36695 | 7518 | complete | gene | gene | ft | <= The block of CDS |
| ftchy2 | 36695 | 7518 | complete | CDS | product | aaa | <= The block of CDS |

| Qualifier hints (提示) | |
|---|---|
| feature | qualifier |
| regulatory | regulatory_class(Mandatory) |
| Example | |
| /regulatory_class='promoter'<br>/regulatory_class='enhancer'<br>/regulatory_class='ribosome_binding_site' | |

**提示此处填写有问题**

# 常见错误类型： Feature table正确示例




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 |
| ftchy | 36695 | 7518 | complete | gene | gene | ftChyE |
| ftchy | 36695 | 7518 | complete | CDS | product | nonribosomal peptide synthase |

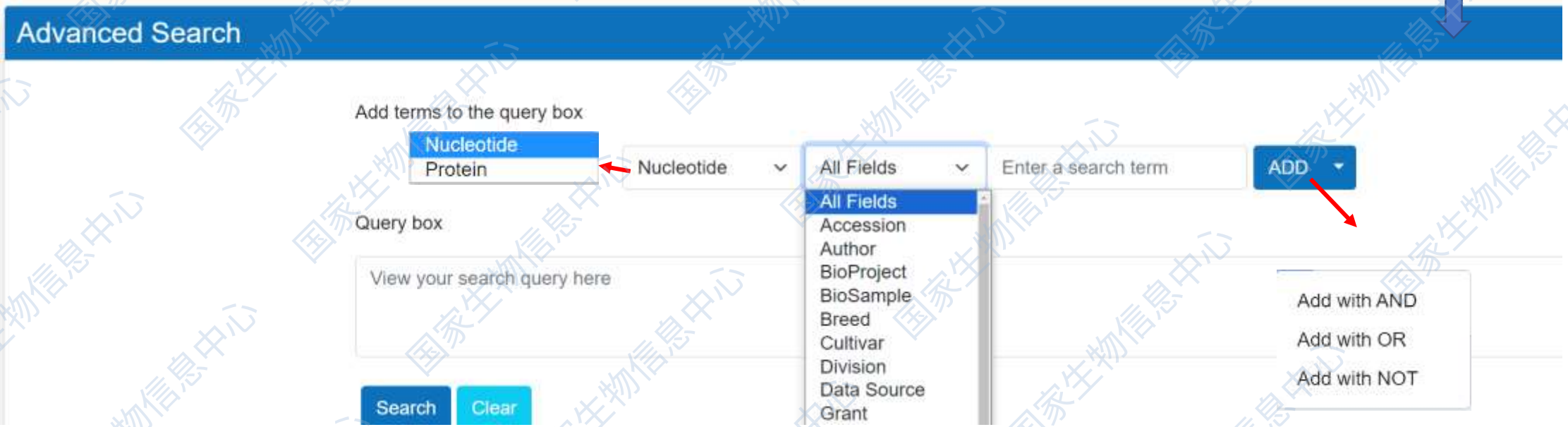
| Qualifier hints (提示) | |
|---|---|
| feature | qualifier |
| regulatory | regulatory_class(Mandatory) |
| Example | |
| /regulatory_class='promoter'<br>/regulatory_class='enhancer'<br>/regulatory_class='ribosome_binding_site' | |

**如果feature是gene，qualifier是gene， qualifier value应该是gene name(缩写)**

**如果feature是CDS， qualifier是product，qualifier value应该是product name(全称)**

**并且gene feature应该在CDS feature之前**

# GenBase数据发布与共享：搜索

# GenBase数据发布与共享：搜索结果

## 搜索结果

### 卡片格式


### 表格格式


| Accession | 物种 | 采样地点 | Isolation Source | 宿主 | 长度 | 分子类型 | 数据来源 | 采样日期 |
|---|---|---|---|---|---|---|---|---|
| C_AA097739.1 | Lespedeza potaninii | China | | | 149,059 | DNA | GenBase | 2021-01-30 |
| C_AA087509.1 | Arenavirus | China: Hubei | Hubei Rodents arena-like virus/2021 2 | Apodemus agrarius | 7,241 | DNA | GenBase | 2021-06-30 |
| C_AA087510.1 | Chuviridae | China: Hubei | Hubei Rodents chu-like virus/2021 1 | Rattus norvegicus | 4,424 | DNA | GenBase | 2021-06-30 |
| C_AA087514.1 | Orthohantavirus | China: Hubei | Hubei Rodents orthohantavirus/2021 3 | Apodemus agrarius | 1,688 | DNA | GenBase | 2021-06-30 |

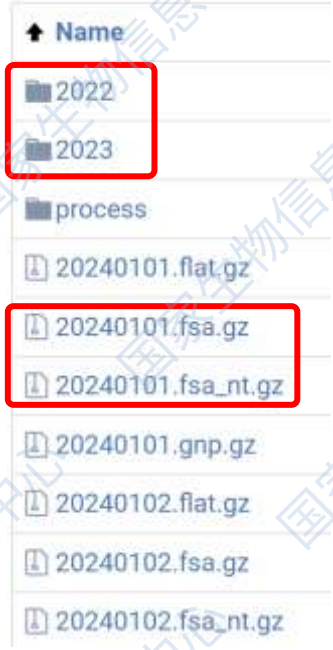
# GenBase数据发布与共享：搜索结果下载

# GenBase数据发布与共享：搜索结果下载

# GenBase数据发布与共享：数据下载其他方式

## 1. FTP下载

https://download2.cncb.ac.cn/genbase/


核酸序列

## 2. Rest接口

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

下载给定Accession编号的序列

例如：https://ngdc.cncb.ac.cn/genbase/api/file/fasta?acc=C_AA004835.1

| 名称 | 类型 | 描述 | 示例 |
| --- | --- | --- | --- |
| acc | string | Accession | C_AA004835.1 |

响应： byte[]

# GenBase数据提交帮助

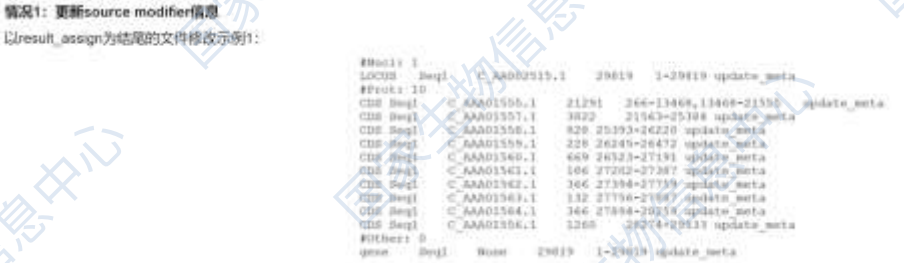

## 1. 质量控制帮助文档

## 52种常见的质量控制问题以及修改建议


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


## 2. 序列更新帮助文档

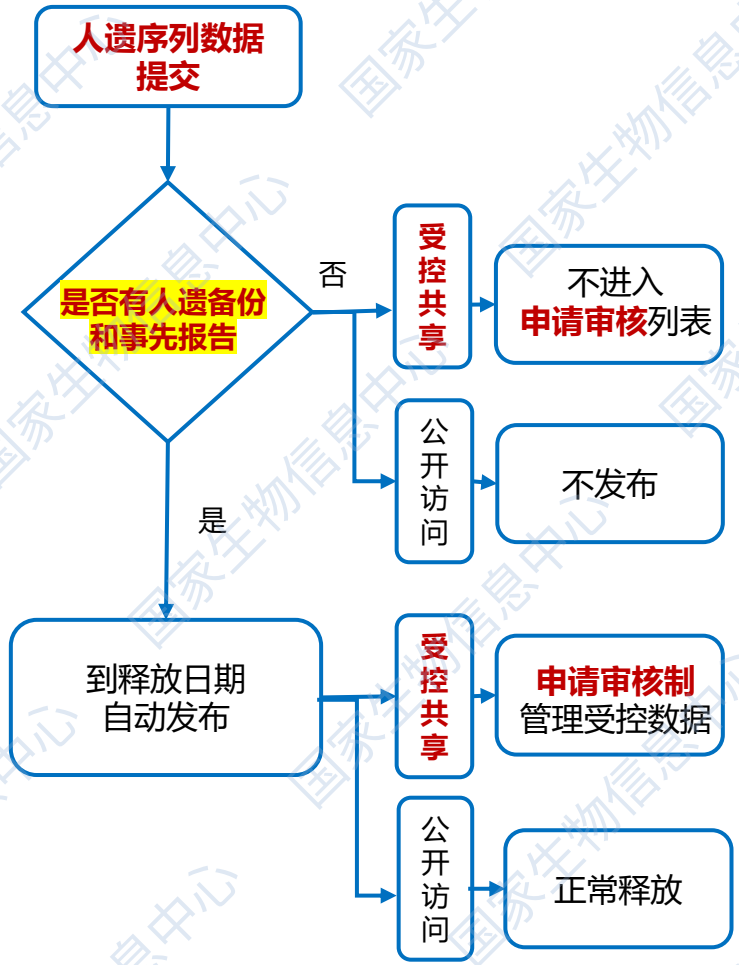
## 3种核酸序列，5种蛋白序列修改场景示例




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


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

# GenBase人类遗传相关序列共享


**序列提交完成后：**

1. 在人类遗传资源信息管理备份平台(https:/ngdc.cncb.ac.cn/hgrip/login)通过GenBase编号完成数据备份，获得备份号

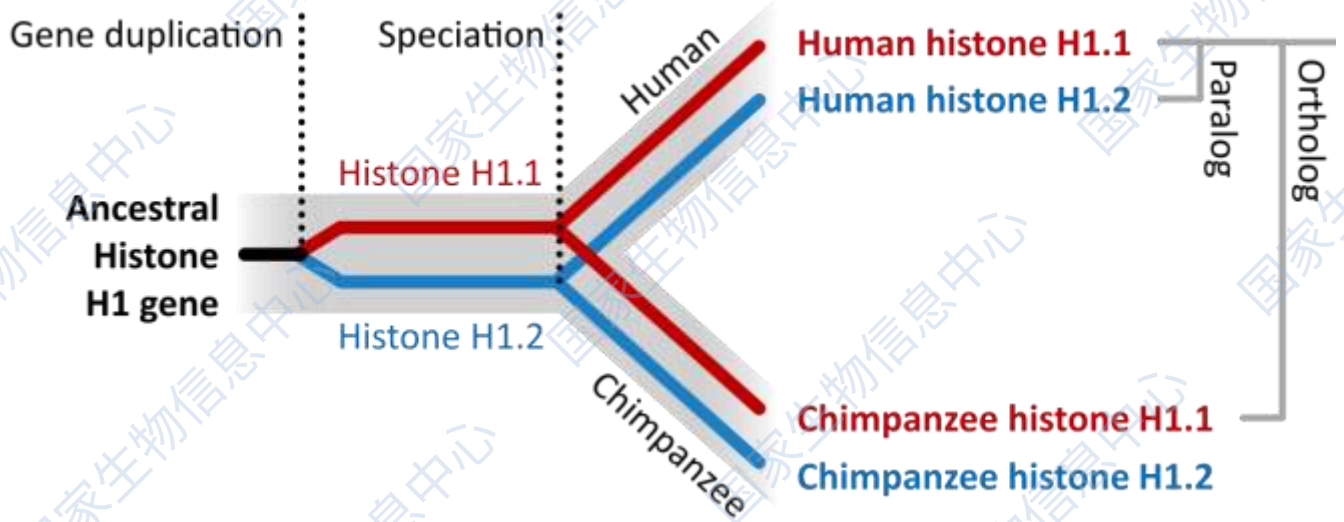
2. 在人类遗传资源服务管理系统(https://apply.hgrg.net/login)通过备份号进行事先报告，获得事先报告编号;

3.将备份编号和事先报告编号通过邮件返回给GenBase工作邮箱genbase@big.ac.cn。请同时在邮件中注明数据是否公开释放，以及计划释放的日期(释放日期可通过用户账号自行修改)

GenBase采用申请审核制管理受控访问数据

# 目录
CONTENTS

# 同源基因数据库HGD-背景




*Wikipedia*

直系同源基因由于其基因结构和功能的相似性以及进化中的保守性具有重要的研究意义

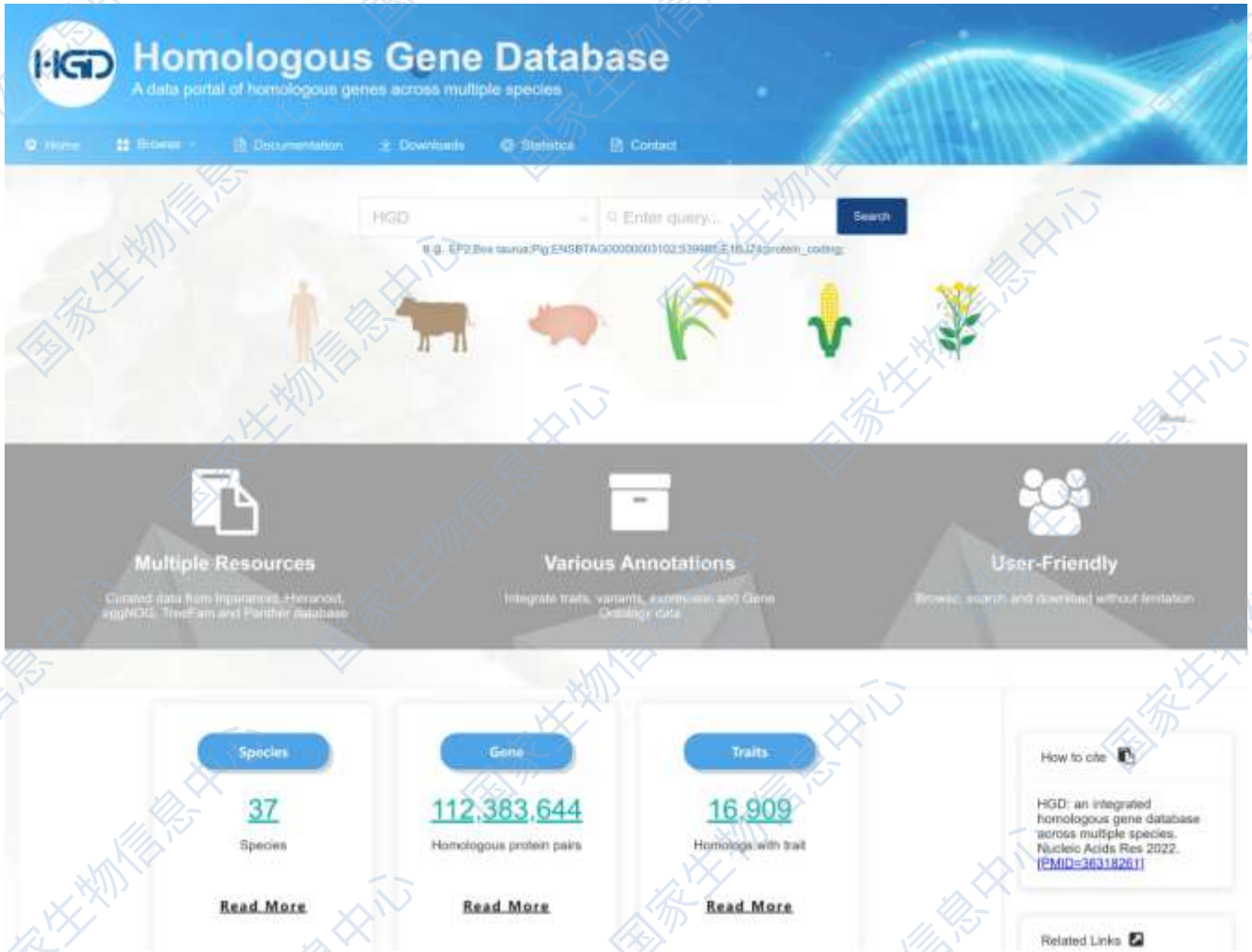
# 研究背景

| 资源名称 | 鉴定方法 | 优缺点 |
| --- | --- | --- |
| OMA (orthologous matrix) | 序列比对 | 支持上传数据与检索，基因树与同源蛋白，未关联表型信息 |
| Inparanoid | blast序列比对 | 提供基因注释和比较分析，未关联表型信息，可视化不好 |
| eggNOG (evolutionary genealogy of genes: non-supervised orthologous groups) | 进化关系推断 | 支持检索与注释，整合了功能注释但没有表型信息，可视化有所不足 |
| OrthoDB (the hierarchical catalog of orthologs) | 序列比对 | 可视化不好，未关联表型信息 |
| Gramene (comparative genomics and pathway resources for plant research) | 进化关系推断 | 仅包含主流作物和植物，没有动物信息，可视化不好 |
| Treefam (phylogenetic trees information with homology predictions ) | 进化关系推断 | 关注动物基因的基因进化发育树关系，基因家族的进化历史，可视化不足，也未关联表型信息。 |
| Triticeae-GeneTribe database | 序列比对 | 小麦专有数据库，专注于小麦的某种，且交互性不好 |

存在的问题：推断方法不同，导致推断结果有差异；
使用的基因标识符不同，很难在不同同源库之间对照；
同源基因缺少组学信息；

# 同源基因数据库HGD

- 数据整合
  - 同源基因信息
    - 5个同源库
  - 多组学注释信息
    - 性状、变异、表达
  - 基因功能注释信息
    - Gene Ontology

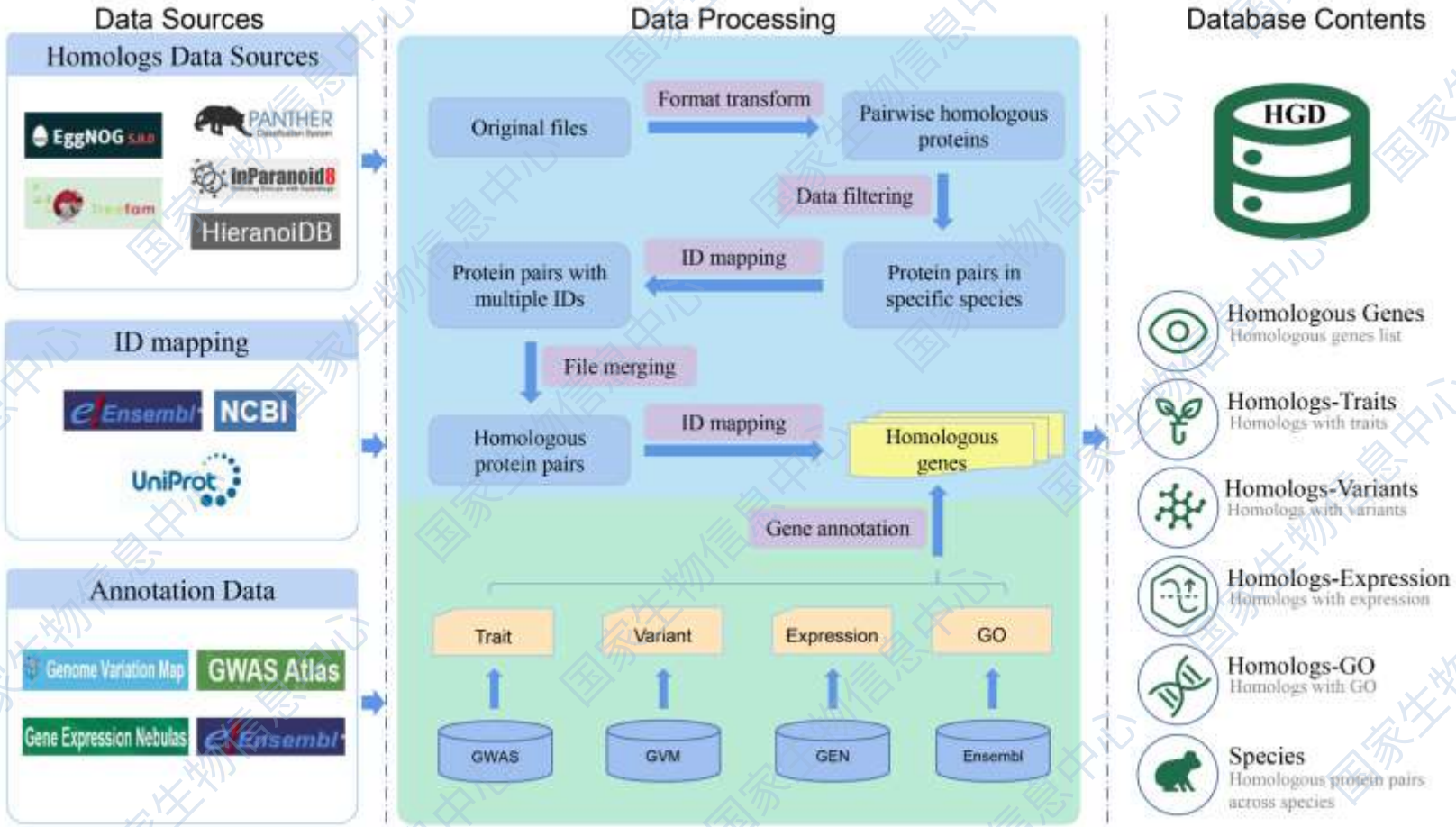



https://ngdc.cncb.ac.cn/hgd

*NAR, 2023*

# 同源基因数据库HGD

数据整合思路

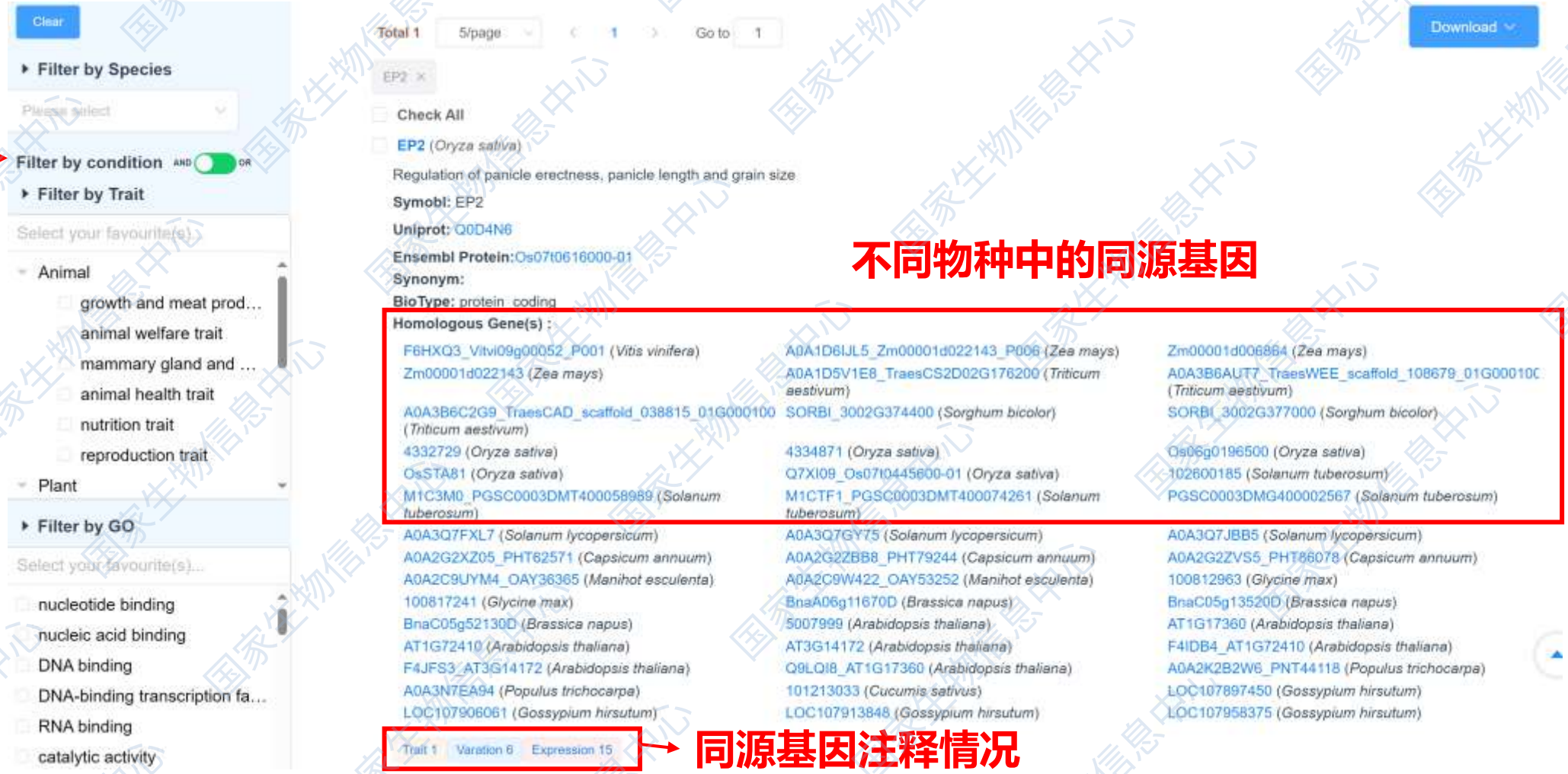



## 数据整合结果

- 物种数：37（动物19，植物16）
- 同源蛋白数：112,383,644
- Trait注释同源基因：16,909
- 变异注释同源基因：276,670
- 表达注释同源基因：398,573
- GO注释同源基因：536,852

# 同源基因浏览、检索

检索过滤条件




不同物种中的同源基因

同源基因注释情况

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

# 跨物种的同源基因trait注释

GWAS Atlas, 3个动物，6个植物，16 909 同源基因

选择性状本体信息



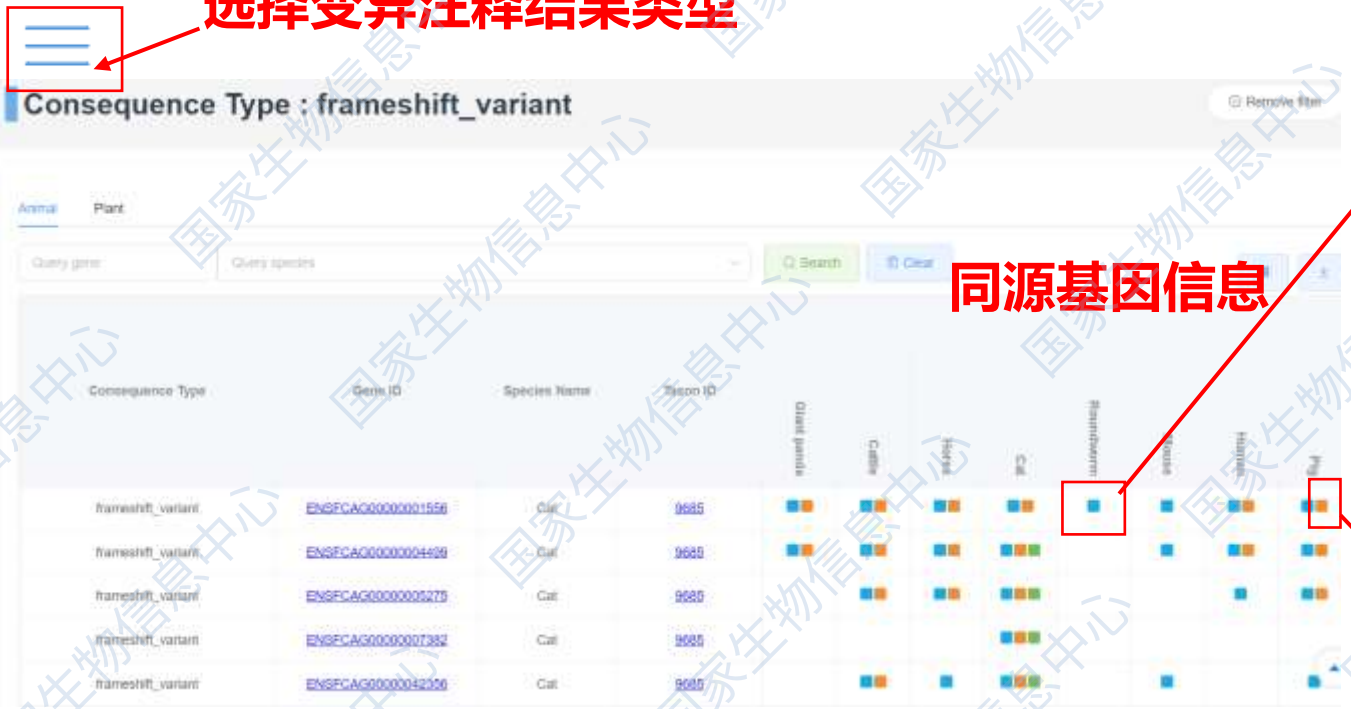
蓝色：同源基因　橙色：同源基因，注释不同trait　绿色：同源基因，注释相同的trait

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

# 跨物种的同源基因变异信息

GVM　7个动物, 9个植物, 276 670 同源基因

选择变异注释结果类型




同源基因信息


变异注释信息



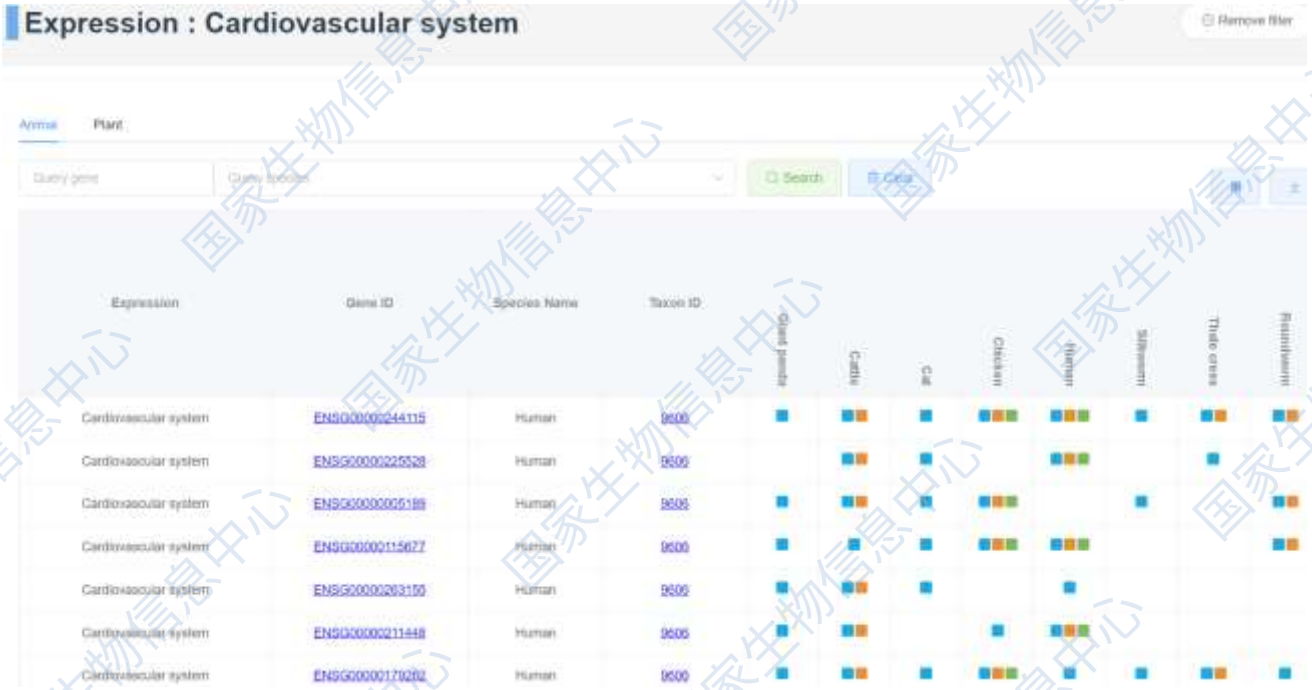

蓝色：同源基因
橙色：同源基因，注释不同变异consequence type
绿色：同源基因，注释相同的consequence type

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

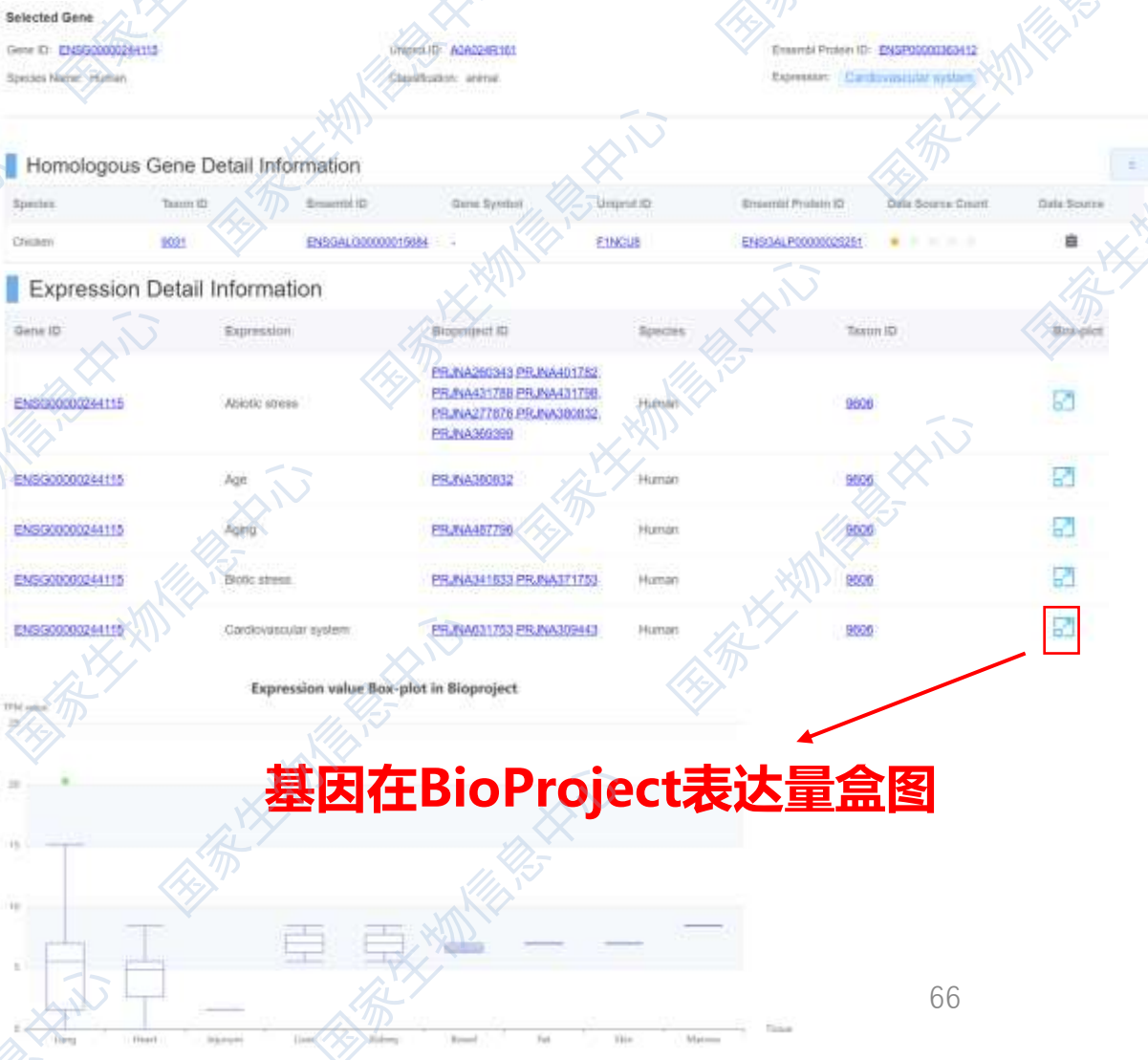
# 跨物种同源基因表达注释

GEN　12个动物, 9个植物, 1个微生物，398,573同源基因

选择表达上下文




蓝色：同源基因
橙色：同源基因，注释不同表达Context
绿色：同源基因，注释相同表达Context



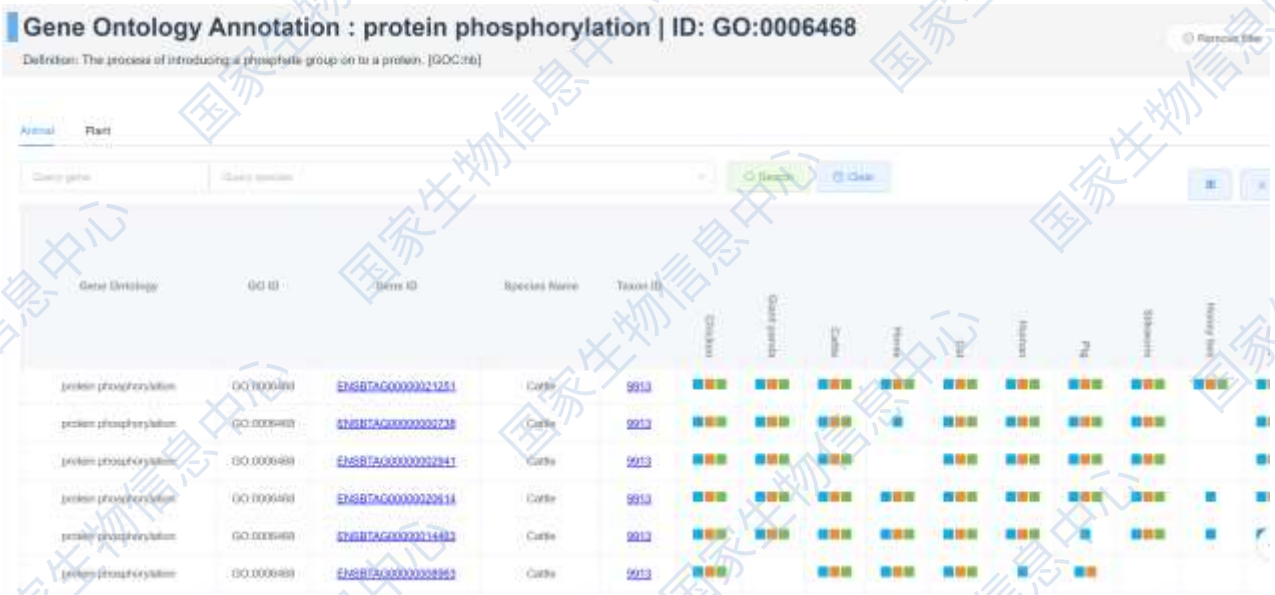

基因在BioProject表达量盒图

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

# 跨物种的同源基因GO注释

Ensembl　19个动物, 14个植物, 2个微生物，536,852个同源基因

选择GO功能



蓝色：同源基因
橙色：同源基因，注释不同GO
绿色：同源基因，注释相同GO

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

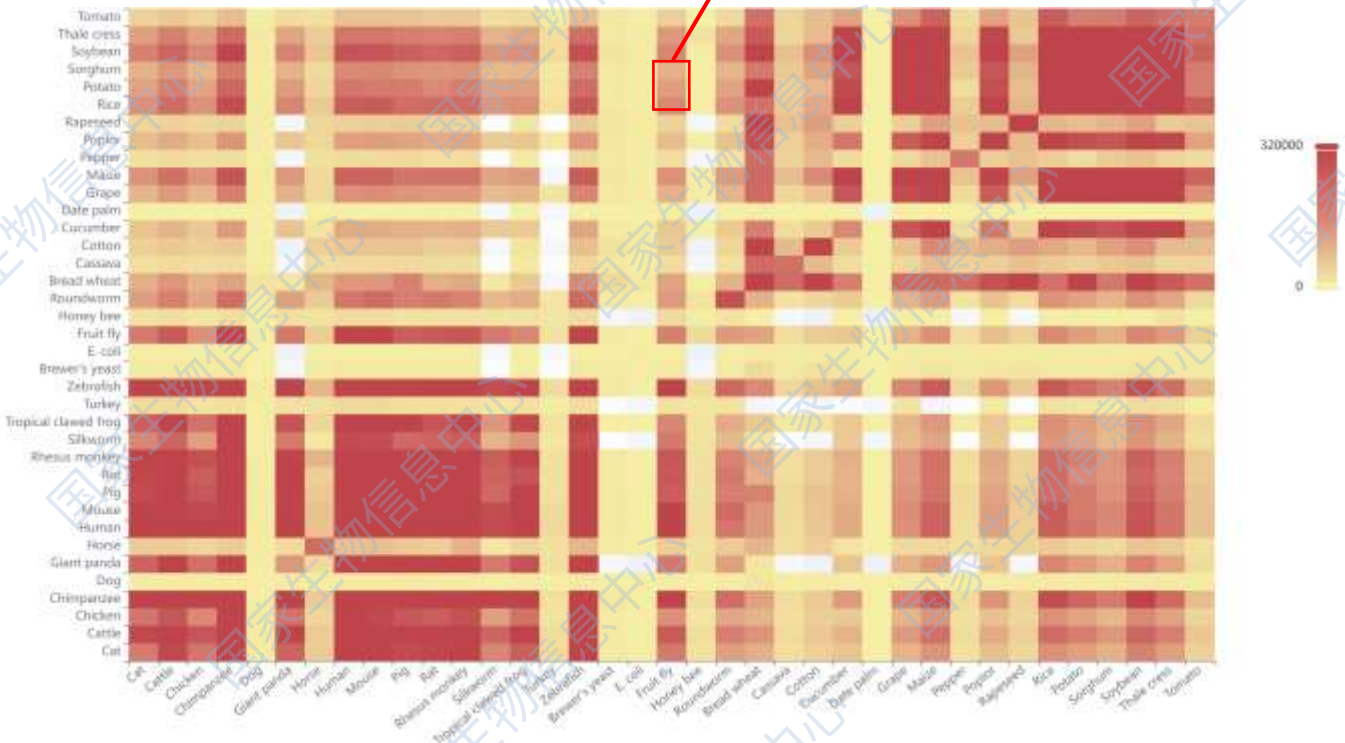
# 物种信息列表

## 37个物种在不同模块中的同源基因统计信息

| Organism | Common Name | NCBI Taxon ID | #Trait | #Variant | #Expression | #GO |
|---|---|---|---|---|---|---|
| *Gallus gallus* | Chicken | 9031 | 214 | 13327 | 15497 | 13034 |
| *Ailuropoda melanoleuca* | Giant panda | 9646 | - | 12580 | - | 13252 |
| *Bos taurus* | Cattle | 9913 | 236 | 17738 | 20721 | 17872 |
| *Canis familiaris* | Dog | 9615 | - | - | 1077 | 973 |

## 物种统计分布


同源基因检索列表




跨物种同源基因统计情况

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

# 同源基因详细信息

## Gene Basic Information

Ensembl Gene ID: Os07g0616000 | Entrez Gene ID: 4343904 | RefSeq ID: XP_015647289
Uniprot ID: Q0D4N6 | Ensembl Protein ID: Os07t0616000-01 | Gene Synonym: -
Gene Symbol: EP2 | Gene Type: protein_coding | Latin Name: Oryza sativa
Species Common Name: Rice | Taxon ID: 4530 | Chromosom: 7
Gene Start: 25381698 | Gene End: 25389632 | Gene Description: Regulation of panicle erectness, panicle length and grain size

## Homologous Gene

| Species | Taxon ID | Gene Symbol | Ensembl ID | Entrez ID | Uniprot ID | Ensembl Protein ID | Data Source Count | Data Source |
|---|---|---|---|---|---|---|---|---|
| Broad wheat | 4565 | - | TraesCS2D02G176200 | - | A0A1D5V1E8 | TraesCS2D02G176200 | | |
| Broad wheat | 4565 | - | TraesWEE_scaffold_106679_01G000100 | - | A0A3B5AUT7 | TraesWEE_scaffold_106679_01G000100 | | |

## Gene Ontology

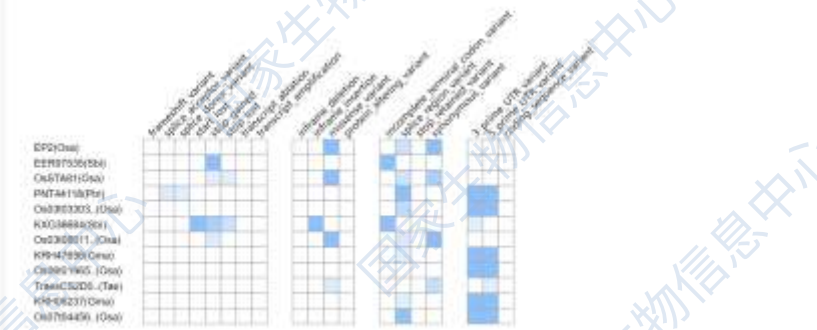
同源基因比较

物种筛选过滤

同源基因比较面板

## Variants

变异

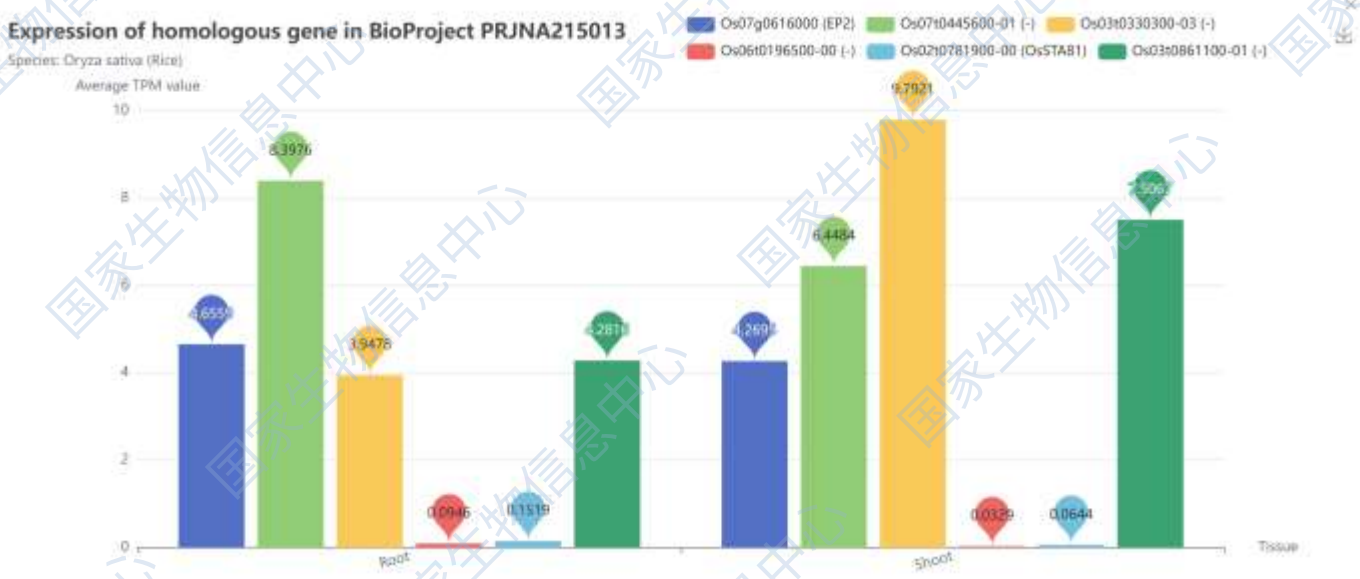


## Traits

性状

# 表达


| Gene ID | Bioproject ID | Tissue | Average TPM | Bar-plot |
|---|---|---|---|---|
| Os07g0616000 | PRJNA380239 | Shoot | 5.185 | |
| Os07g0616000 | PRJNA215013 | Root | 4.6559 | |
| Os07g0616000 | PRJNA215013 | Shoot | 4.2695 | |
| Os07g0616000 | PRJNA436566 | Leaf | 7.3293 | |




同源基因的表达信息

# 同源基因数据下载

- Homologous Protein
- Homologous Gene
- Trait Files
- Variation
- Gene Ontology
- Expression

**Cat**

| | | |
|---|---|---|
| Cattle_Cattle_Homolog_protein.txt.gz | Cattle_Soybean_Homolog_protein.txt.gz | Cat_Tropical clawed frog_Homolog_protein.txt.gz |
| Cattle_Chicken_Homolog_protein.txt.gz | Cat_Pig_Homolog_protein.txt.gz | Cat_Brewer's yeast_Homolog_protein.txt.gz |
| Cat_Dog_Homolog_protein.txt.gz | Cattle_Rat_Homolog_protein.txt.gz | Cattle_Brewer's yeast_Homolog_protein.txt.gz |
| Cat_Rhesus monkey_Homolog_protein.txt.gz | Cat_Soybean_Homolog_protein.txt.gz | Cat_Giant panda_Homolog_protein.txt.gz |
| Cat_Potato_Homolog_protein.txt.gz | Cattle_Rhesus monkey_Homolog_protein.txt.gz | Cattle_E. coli_Homolog_protein.txt.gz |
| Cattle_Dog_Homolog_protein.txt.gz | Cat_Human_Homolog_protein.txt.gz | Cattle_Turkey_Homolog_protein.txt.gz |
| Cattle_Tropical clawed frog_Homolog_protein.txt.gz | Cattle_Thale cress_Homolog_protein.txt.gz | Cat_Grape_Homolog_protein.txt.gz |
| Cattle_Rice_Homolog_protein.txt.gz | Cat_Mouse_Homolog_protein.txt.gz | Cattle_Date palm_Homolog_protein.txt.gz |

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

# Thanks!

## 欢迎访问和提交数据到国家基因组科学数据中心

基因序列库GenBase

**Email:** genbase@big.ac.cn

**QQ群:** 629388189