



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

一、基因序列和基因组数据汇交的异同

二、基因序列数据汇交和共享

三、同源基因数据库HGD

# 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

The screenshot displays the GenBase website interface. At the top, there is a navigation bar with links for 'GenBase', '首页' (Home), '数据提交' (Data Submission), '检索' (Search), '下载' (Download), '统计' (Statistics), '标准' (Standards), and '帮助文档' (Help Documents). On the right side of the navigation bar are links for '登录' (Login), '注册' (Registration), and '语言/Language'.

Below the navigation bar, a large blue banner contains a search bar with a dropdown menu set to 'Nucleotide' and a text input field labeled 'Type your keywords'. To the right of the search bar are buttons for '检索' (Search) and '高级检索' (Advanced Search). Below the search bar, a text box provides information about GenBase: '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的核酸和蛋白质序列数据。'

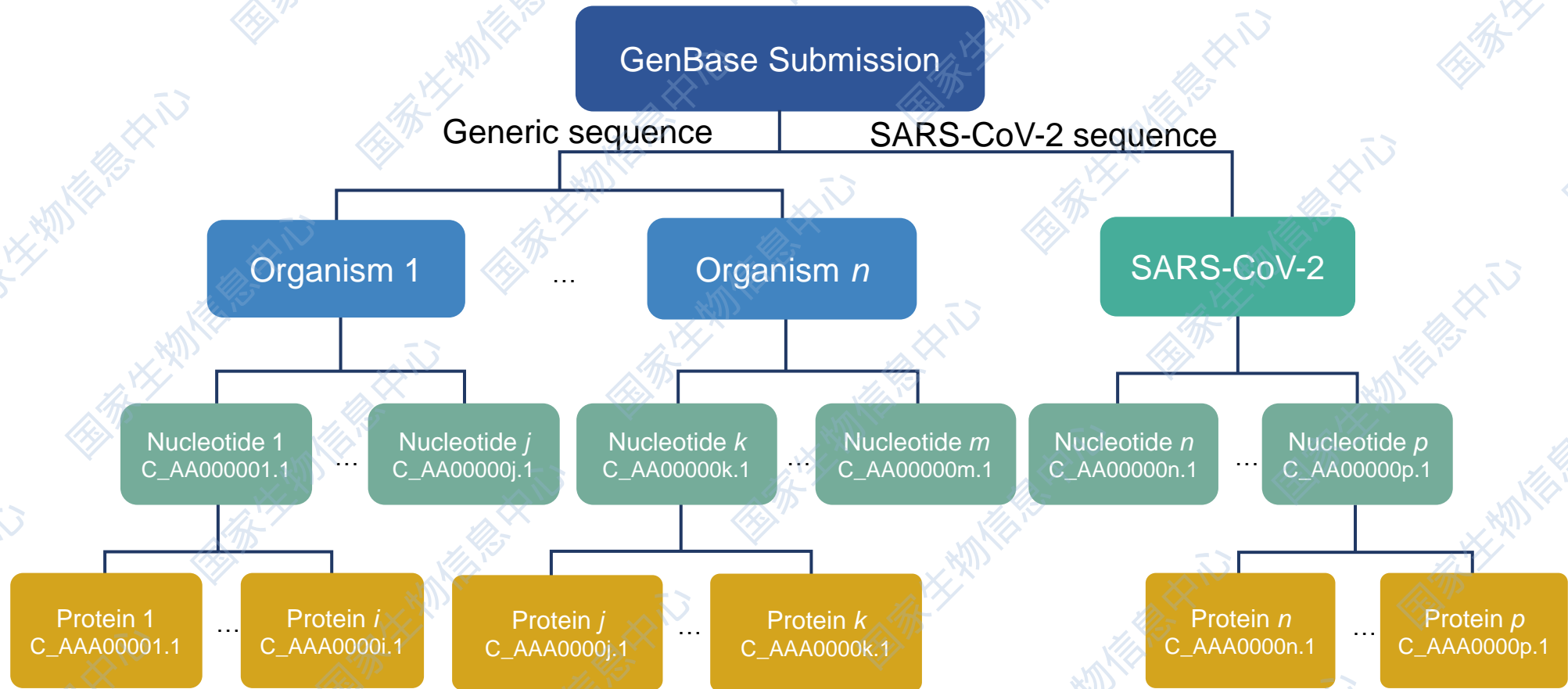
Below the banner, there are three main sections:

- 汇交与整合数据** (Data Submission and Integration): This section features three circular icons representing different data types: '物种' (Species) with 592,478 entries, '核酸' (Nucleic Acid) with 374,134,720 entries, and '蛋白' (Protein) with 619,151,637 entries.
- 最近更新** (Recent Updates): This section lists four updates with dates and descriptions:
  - 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
  - 2024-05-08: 数据交换统计信息上线
- 国际数据GenBank整合** (International Data GenBank Integration): This section provides information about the integration with GenBank, including the update date (2024-10-21), the number of nucleic acid sequences (890), the number of proteins (4,032), the release date (2024-10-19), the number of nucleic acid sequences (232), and the number of proteins (1,093).

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

*Genomics Proteomics Bioinformatics, 2024*

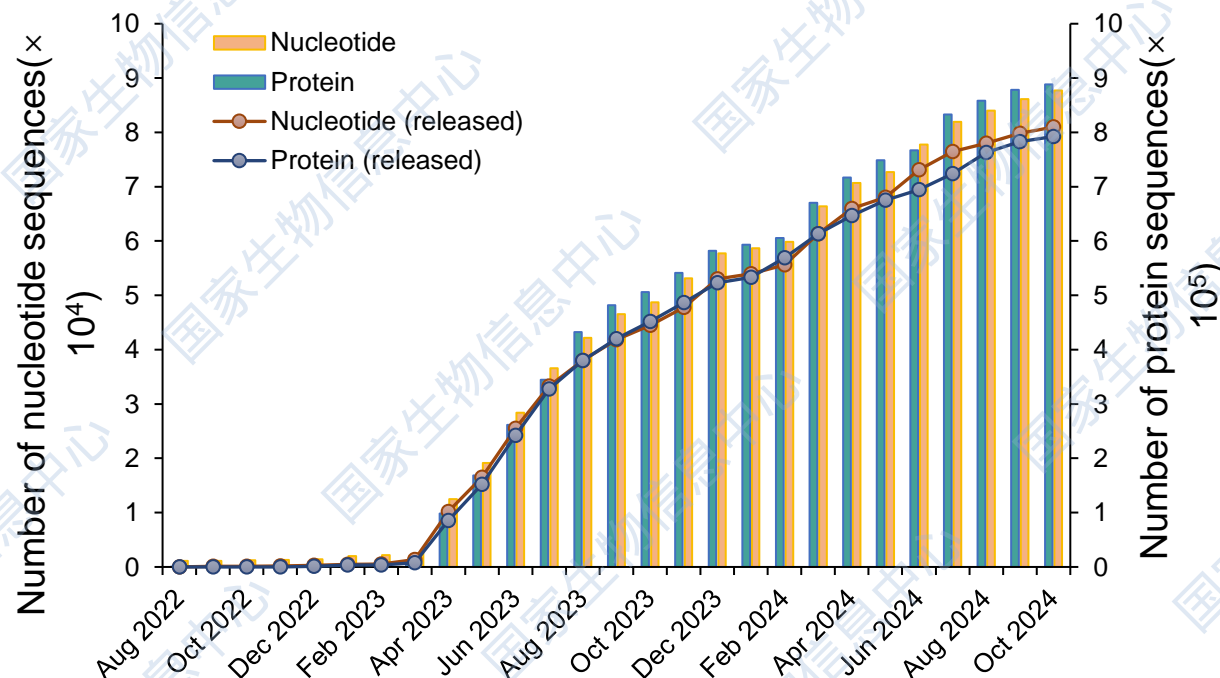
# GenBase的数据模型



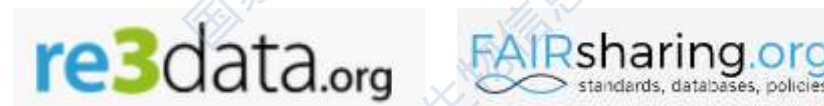
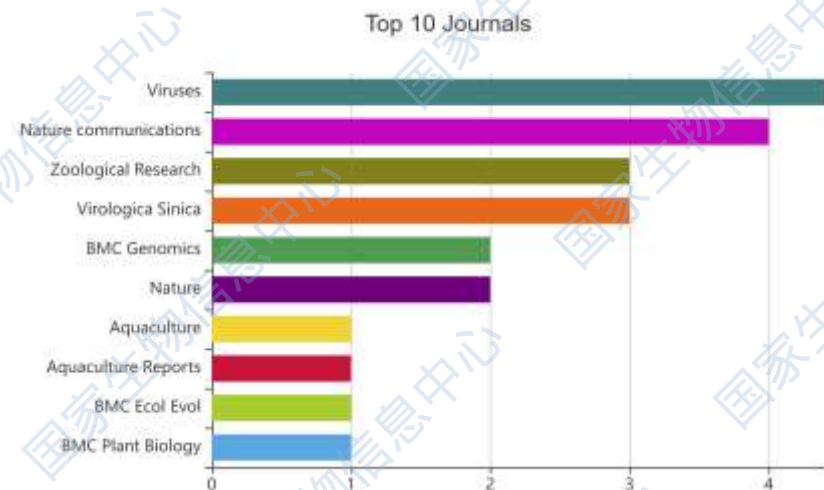
**Nucleotide accession: C\_ + 2 characters + 6 numbers + version number**

# 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



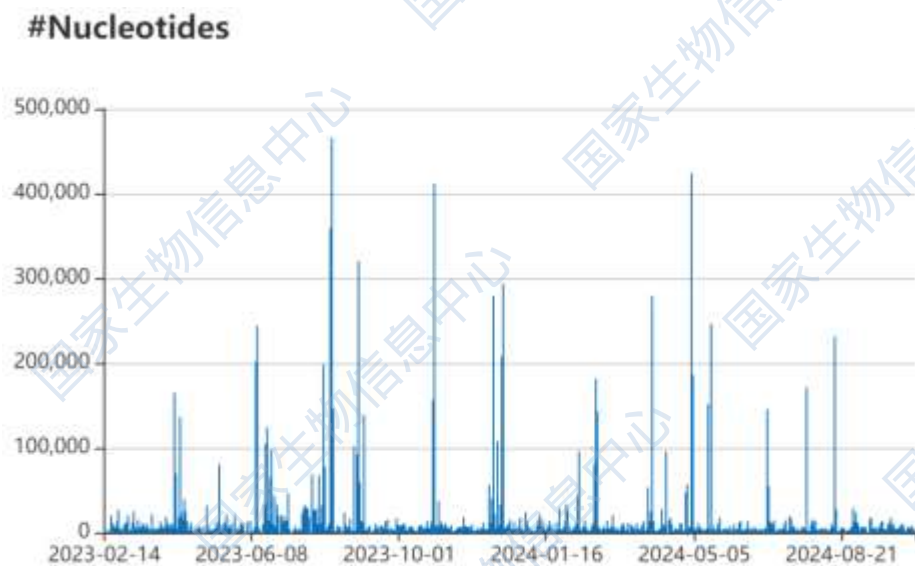
Genomics Proteomics Bioinformatics (2024)



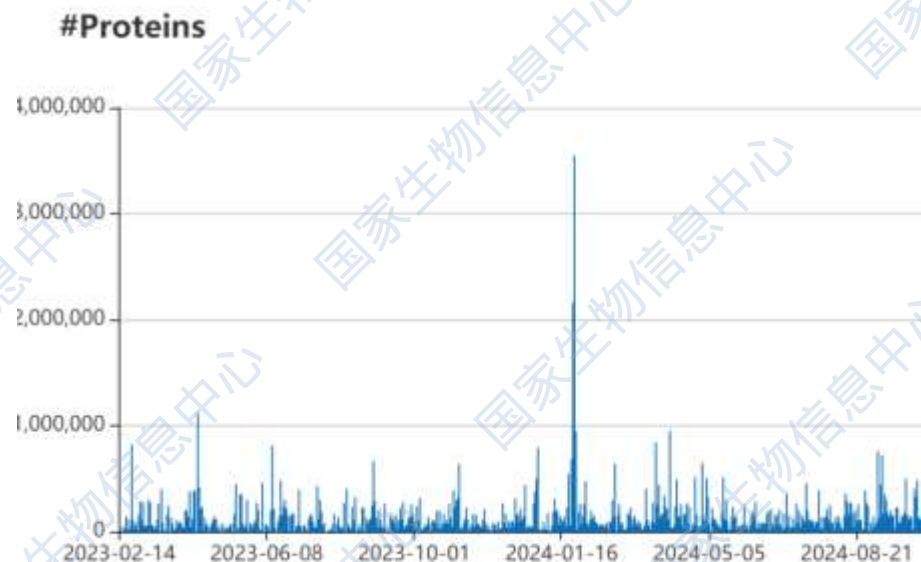
Since March 2023, GenBase has supported data submissions for 67 articles and 285 funded projects.



# 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 <ul style="list-style-type: none"><li>• 病毒基因组</li><li>• 细胞器基因组</li><li>• 质粒基因组</li><li>• 所有物种基因片段</li></ul>	<ul style="list-style-type: none"><li>• <b>基因片段</b>序列</li><li>• 注释数据（非必需）</li></ul>	<b>可以不涉及</b> BioProject 和 BioSample
基因组数据库GWH	大基因组 <ul style="list-style-type: none"><li>• 真核生物全基因组</li><li>• 原核生物全基因组</li><li>• 宏基因组组装数据</li></ul>	<ul style="list-style-type: none"><li>• <b>基因组</b>组装序列</li><li>• 注释数据（非必需）</li></ul>	<b>涉及</b> 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 GWH	重注释	RefSeq	Ensembl	/
基因组序列库 GenBase	用户递交	GenBank：全基因组 WGS	ENA： Assembled/annotated sequence	DDBJ
	用户递交	GenBank：传统 GenBank		



# 目录

## CONTENTS

一、基因序列和基因组数据汇交的异同

二、基因序列数据汇交和共享

三、同源基因数据库HGD

# 基因序列数据汇交共享——GenBase

1. 提交前准备

2. 提交流程

3. 数据发布和共享

4. 常见错误类型

# 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



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

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

# 准备工作：1.账户准备-注册和登陆

The screenshot displays the GenBase website interface. At the top, there is a navigation bar with links for GenBase, Home, Data Submission, Search, Download, Statistics, Standards, and Help Documents. On the right, there are buttons for Login (highlighted with a red box), Register, and Language. Below the navigation bar, a large blue banner contains a search bar with a dropdown menu set to 'Nucleotide' and a text input field for keywords. To the right of the search bar are buttons for 'Search' and 'Advanced Search'. Below the search bar, there is a section titled 'GenBase' with a description of the database and a list of recent updates. To the left of the search bar, there is a section titled 'GenBase' with a description of the database. Below the search bar, there is a section titled '汇交与整合数据' (Data Submission and Integration) with three icons representing Species, Nucleotide, and Protein, each with a corresponding count. To the right of the search bar, there is a section titled '最近更新' (Recent Updates) with a list of updates. On the far right, there is a section titled '国际数据GenBank整合' (International Data GenBank Integration) with a list of statistics.

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的核酸和蛋白质序列数据。

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: 592,478
- 核酸: 374,131,350
- 蛋白: 619,127,880

最近更新

- 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
- 2024-05-08 数据交换统计信息上线

国际数据GenBank整合

更新日期: 2024-10-19  
核酸数: 3,234  
蛋白数: 108,467

GenBase数据释放

释放日期: 2024-10-19  
核酸数: 232  
蛋白数: 1,093

新加功能

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

Enter your Username and Password

Email [Forgot activate?](#)

zhaoxuetong@big.ac.cn

用户名

Password [Forgot password?](#)

密码

Check code

x4ab

验证码

☐ Keep me signed in

Login Reset Register

点击此处  
登陆

注册

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)

Nstn  
科技资源共享网账号登录

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

Welcome to register for an account of NGDC

Register information

Account Login Information

Email \*

注意：由于需要通过邮件接收激活链接，请确保该邮箱属于白名单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拦截策略，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. Detection junk email avoid transferred by anti-virus email system.

Password \*

must contain A-Z, a-z, 0-9 and special character, length 8-30  
Password can not be empty

Confirm Password \*

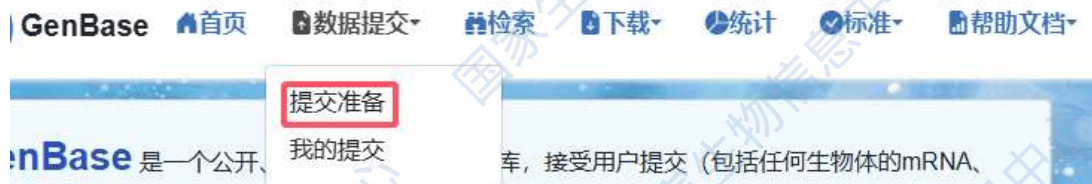
must contain A-Z, a-z, 0-9 and special character, length 8-30  
注意：密码必须同时包含大写字母、小写字母、数字、特殊字符，且长度在8-30位之间！  
Note: password must contain one uppercase letter, one lowercase letter, one number, one special character, and should be 8-30 characters in length.

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**

For example:

```
>Seq1 [organism=Homo Sapiens] Definition Line for Seq1  
aaccgatatagagagga  
>Seq2 [organism=Homo Sapiens] Definition Line for Seq2  
atctgaatagagatttt
```

**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](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_ID						
#!Each specimen in the set must have a line in the source modifiers table						
#!Each Sequence_ID may appear only once in the source modifiers table						
#!Fill in the data in the line beginning with "##number". If it is 1						
#!Reference: <a href="https://ngdc.cncb.ac.cn/genbase/standards">https://ngdc.cncb.ac.cn/genbase/standards</a> ; <a href="https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers">https://www.ncbi.nlm.nih.gov/WebSub/html/help/genbank-source-table.html#modifiers</a>						
#!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						
##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
	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
##Description						
##Example(do not delete)	Seq1	2001-01-31 or value in ControlWords sheet	China or value in ControlWords sheet	C.Grant	Homo Sapiens	Male
##1						
##2						
##3						
##4						
##5						
##6						
##7						

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

ControlWords

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](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 Athlp
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-特征注释文件

## 下载后打开GenBase Features.xlsx模板:

	A	B	C	D	E	F	G	H	I
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 <b>feature</b> (column E) is the structure name of this type of sequence, <b>qualifiers</b> (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, <b>qualifiers hints</b> (column G) will help you							<b>Qualifier hints (提示)</b>	
2								<b>feature</b>	<b>qualifier</b>
3	<b>1.仔细阅读填写说明，特征内容需填写在A-G列</b>							gene	gene
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 <b>A到G列</b> 。其中 <b>feature (column E)</b> 是该序列中一个对应的结构体， <b>qualifier (column F)</b> 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， <b>Qualifier hints</b> 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 <b>qualifier value (column G)</b>							Example	
5								/gene=ilvE	
6									
7	<b>Locations (坐标)</b>				<b>Attributes (属性)</b>				
8	sequence_id	start	end	completeness	feature	qualifier	qualifier value		
9	<b>3. 请在此处填写具体信息 不要删除表头、行列名称、填写说明等内容</b>								
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									

**2. 提示区**  
**提示填写G列内容**



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

A	B	C	D	E	F	G	H	I	J	K
<p>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 <b>feature</b> (column E) is the structure name of this type of sequence, <b>qualifiers</b> (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, <b>qualifiers hints</b> (column G) will help you prompt the format of the qualifier value.</p> <p>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<b>feature</b> (column E) 是该序列中一个对应的结构体, <b>qualifier</b> (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, <b>Qualifier hints</b>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<b>qualifier value</b> (column G)</p>							<b>Qualifier hints (提示)</b>			
							feature	qualifier	<= All features and qualifiers of INSDC	
							gene	gene	<= Select the feature and qualifier to view in the drop-down box respectively <= Select the content to prompt	
							Definition			
							symbol of the gene corresponding to a sequence region		<= Display box	

# 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 <a href="https://www.insdc.org/documents/feature_table.html">https://www.insdc.org/documents/feature_table.html</a>				
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	
		(Optional) note	text	
		(Optional) old_locus_tag	text (single token)	
		(Optional) standard_name	text	
		(Optional) trans_splicing	boolean	

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

# 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/](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		10	567 complete	mRNA	regulatory_class(Mar promoter)	ubc42
seq1	789	1320			gene	ubc42
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)	ubc42
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
```



GenBase

首页

数据提交

检索

下载

统计

标准

帮助文档

元信息

特征注释

注释文件规范

例如 C\_AA00110

0039

## 8种类型

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

<https://ngdc.cnbc.ac.cn/genbase/filespec>



# 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

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

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

## 新的提交

**GenBase** 接受来自任何生物体mRNA、基因组DNA、ncRNA或小基因组，如细胞器、病毒、质粒、噬菌体。

- 组装完成的真核和原核基因组(可能包括也可能不包括细胞器或质粒)应提交到 [GWH](#)。
- 由下一代测序技术生成的未组装(原始)序列应提交到 [GSA](#)。

已知晓，继续提交

# 提交流程： 步骤1. 提交者信息 (Submitter)

提交者

姓: tang

中间名:

名: box

电子邮箱 (次要):

电子邮箱: tangbx@big.ac.cn

提交者所在组织机构网址: https://

组织机构: Beijing Institute of Genomics, Chinese Academy of Sciences

部门: 010

电话:

传真:

街: 1

城市: 1

省/州: 1

国家: Afghanistan

邮政编码: 1

保存并下一步

**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)

提交者 出版信息 测序技术 序列 物种 集合或批次 类别 元信息 特征 结果预览

出版信息

提交 #0000062

通讯作者 (请注意: 推荐教授或副教授的信息)

名	中间名	姓	后缀	移除
bixia		tang		X

增加 更多通讯作者

文献信息 #1

请提供本次提交相关论文的题目和出版细节(卷号, 期号, 页号等)。

出版状态

☒ 未公开出版 ☐ In-Press ☐ 已出版

文章标题

文献作者

☒ 与通讯作者一致 ☐ 指定新作者

增加另外的出版信息

增加参考文献

# 提交流程： 步骤2. 参考文献信息(Reference)

文献信息 #1

请提供本次提交相关论文的题目和出版细节(卷号, 期号, 页号等)。

\* 出版状态

☒ 未公开出版

☐ In-Press

☐ 已出版

文章标题

\* 文献作者

☐ 与通讯作者一致

☒ 指定新作者

\* 名

bixia

中间名

\* 姓

tang

后缀

移除

X

Add

更多参考作者

可选择指定新作者，并且正确添加作者信息

添加另外的出版信息

保存&下一步

保存并下一步

 CNCB-NGDC

28

# 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？

☒ Sanger dideoxy sequencing

☐ MGI

☐ IonTorrent

☐ Other

☐ 454

☐ MGISEQ-200RS

☐ Nanopore

☐ Helicos

☐ MGISEQ-2000

☐ PacBio

☐ Illumina

☐ MGI DNBSEQ

☐ SOLiD

上述序列为：

Assembled sequences



保存&下一步

保存并下一步

根据实际情况，详细填写相关信息

# 提交流程：步骤4. 核酸序列 (Nucleotide)

提交者

出版信息

测序技术

序列

物种

集合或批次

类别

元信息

特征

结果预览

核酸序列

提交 #0000062

数据发布日期

什么时候可以发布你的序列?

☒ 审核完成即发布

☐ 拟发布日期

日期格式为 'YYYY-MM-DD' (例如: 2022-02-07),  
发布日期必须从今天起1个月后, 到未来3年内。

发布政策和免责声明

1. 作者可以指定一个日期, 在指定日期前暂不发布该序列数据。

2. 如果提交者想要更改发布日期, 请联系GenBase工作组: [genbase@big.ac.cn](mailto:genbase@big.ac.cn)

3. 如果在指定日期之前有引用该序列或编号的文章发表, 则该序列将在文章发表后立即发布, 否则,

4. 提交者一旦获取出版信息, 烦请将完整的出版物数据, 即所有作者、标题、期刊、卷数、页数和日期发送到邮箱: [genbase@big.ac.cn](mailto:genbase@big.ac.cn)

5. 数据发布后, 未来可能会自动与NCBI同步, 不再联系提交者进行确认。

☒ 接受 ☐ 不接受

提交任务小标题

BioProject

BioProject:

资助金

\* 资助机构

资助类别

\* 资助项目编号

资助项目名称

资助

Ministry of Science and Technology of the

National Key Technologies R&D Program

NA

更多资助

设置数据释放时间  
但最长不要超过4年

仔细阅读数据发布政策和免责声明

设置提交标题

填写BioProject信息 (可选)

填写资助基金信息

30

# 提交流程： 步骤4. 核酸序列 (Nucleotide)

**核酸序列和定义行**

\* 分子类型:  选择测序的分子类型。

拓扑结构:

Strand信息:  如果您提交的是病毒序列，则必须填此信息。如果不是，请留空。

您提交的序列所对应的遗传区属于:  请注意，如果您的序列对应多种遗传区类型，则应该分开提交。

您是否提交了细胞器基因组、病毒、病毒片段、类病毒、质粒或克隆载体的完整序列? ☒ Yes ☐ No

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

Single strand  
Double strand  
Mixed strand

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

**核酸序列格式**

数据格式:

**FASTA序列** (最常见的数据格式, seqid的长度不超过23字符。 [帮助](#))

例如:

```
>Seq1 [organism=Homo Sapiens] Definition Line for Seq1  
aaccgatataagagagga  
>Seq2 [organism=Homo Sapiens] Definition Line for Seq2  
atctgaatagattatt
```

定义行 (Definition line) 用于描述每条序列，因此必须包含在每条提交序列中。

核酸序列填写示例



# 提交流程： 步骤4. 核酸序列 (Nucleotide)

\* Nucleotide Sequence(s)

☐ 粘贴序列 (当序列数目小于40)

☐ 上传本地文件 (支持 fa, fsa, fas, fst, fna, fasta 以及 gz 格式。允许上传的最大文件为 1GB)

0000062\_seq.fsa Remove

☐ FTP 上传 (文件 > 1GB)

\* 文件名

\* MD5

请注意  
请将您的数据文件传输到 GenBase FTP 站点(下载 FileZilla 客户端)。我们建议您先通过 FTP 传输文件，然后在这里填写文件名和 md5 码，然后单击继续。  
地址: <http://submit.big.ac.cn>  
用户名: 与登录 GenBase 一致  
密码: 与登录 GenBase 一致

保存&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

保存并下一步



# 提交流程： 步骤5. 物种 (Organism)



物种

提交 #0000062 [Help](#)

请填写缺失的物种信息

您提交的序列中未包含物种信息。请在下面输入物种名称。(对于后续的序列提交，请务必在FASTA文件中包含物种信息。)  
[下载 物种模板文件](#)

物种名称:

所有序列都使用该物种名称

或者

上传文件:

**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**

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)



## 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
- ☐ Phy set
- ☐ Mut set
- ☐ Env 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。

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。

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。

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

**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&下一步

**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)



## 特征

提交 #0000062 帮助

请注意

- 1. 请选择提交序列注释的以下三种格式之一，可参考注释文件规范
- 2. 请参考 [此链接](#) 作为TBL格式序列注释的说明
- 3. 如果需要以Excel格式提交序列注释特征，请下载模板文件 [GenBase\\_Features.xlsx](#)

上传您提交的序列注释:

- ☒ Excel格式
- ☐ TBL格式
- ☐ GFF3 格式

0000062\_feature.xlsx Remove

验证成功!

保存&下一步

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)



提交0000062

帮助

## 1. 其他邮箱地址?

关于本次提交的邮件将发送到以下地址。多个电子邮件地址之间用逗号分隔。

多个电子邮件地址之间用逗号分隔。

## 2. 重新提交?

如果GenBase工作人员要求您重新提交序列数据 ☐

## 3. 其他信息

如果您有额外的或更正的元信息或特征文件，或其他纯文本描述的序列数据提交 ☐

## 4. 更新

您可以随时更新或修改您的提交，请通过电子邮件发送新的或更新的信息到 [genbase@big.ac.cn](mailto:genbase@big.ac.cn)。如果您有任何问题，也可以通过此邮箱与我们联系。

1.确认并选填上述信息

审查提交记录

以下存在 1 条GenBase提交序列供您审查。

完成提交

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       053DN2379VH1               2483 bp    RNA    linear    DUX 16-AGO-2022
DEFINITION  Human picobirnavirus.
ACCESSION   053DN2379VH1
VERSION     053DN2379VH1
KEYWORDS    .
SOURCE      Human picobirnavirus
  ORGANISM  Human picobirnavirus
            Riboviria; Orthornavirales; Pleovirales; Duplodnavsivirales;
            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-AGO-2022) National Genomic Data Center, Beijing
            Institute of Genomics, Chinese Academy of Sciences, NO.1 Weichen
            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="MTNPQLATAELESHRHITVEHETGRHIVETERIOGTLAGTSR
                     HSKAYEIOIALAARTANNDVATKOINHTSAQFLAHQFAQAHETTSVSEETTRGLAT
                     ESTQSGLEETGRHMLATEQVHTHALGQFRHFRQTELQWISTGAGAYKOVSTGLAQH
                     VESVKGIHQVKEELGG"
     CDS               707..2474
                     /codon_start=1
                     /product="capsid protein"
                     /translation="MKDSRHITGRGRKYGKAVIKRRRELDQQIHRYKDCNHWYA
```

# 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	-		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



# 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	-		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	<b>Audit Success</b> <a href="#">Accession</a>
0000114	RcALV-BelV		2022-09-28 14:47:05	2022-10-19 15:33:55	2022-09-28	<b>Audit Success</b> <a href="#">Accession</a>

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	<b>Audit Fail</b> <a href="#">View</a> -
0000062	-		2022-08-17 23:40:08	2022-10-19 18:50:07	2022-08-17	<b>Audit Fail</b> <a href="#">View</a> -

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		2022-08-26 13:29:11	2022-09-02 15:29:06	2022-08-26	Audit Success	Accession	public	<a href="https://ngdc.cnbi.ac.cn/genbase/review/">https://ngdc.cnbi.ac.cn/genbase/review/</a> 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	<a href="https://ngdc.cnbi.ac.cn/genbase/review/">https://ngdc.cnbi.ac.cn/genbase/review/</a> Cancel Modify Grant
0000086	-		2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success	Accession	public	<a href="https://ngdc.cnbi.ac.cn/genbase/review/">https://ngdc.cnbi.ac.cn/genbase/review/</a> Cancel 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	-		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			2022-11-29 11:09:41	2022-11-30 10:35:22	2022-11-29	Audit Success Accession confidential
						Review link Modify Grant Modify ReleaseTime

## 2. 数据处于私密状态

只能自己点击查看

可点击在线修改发布日期

# 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：

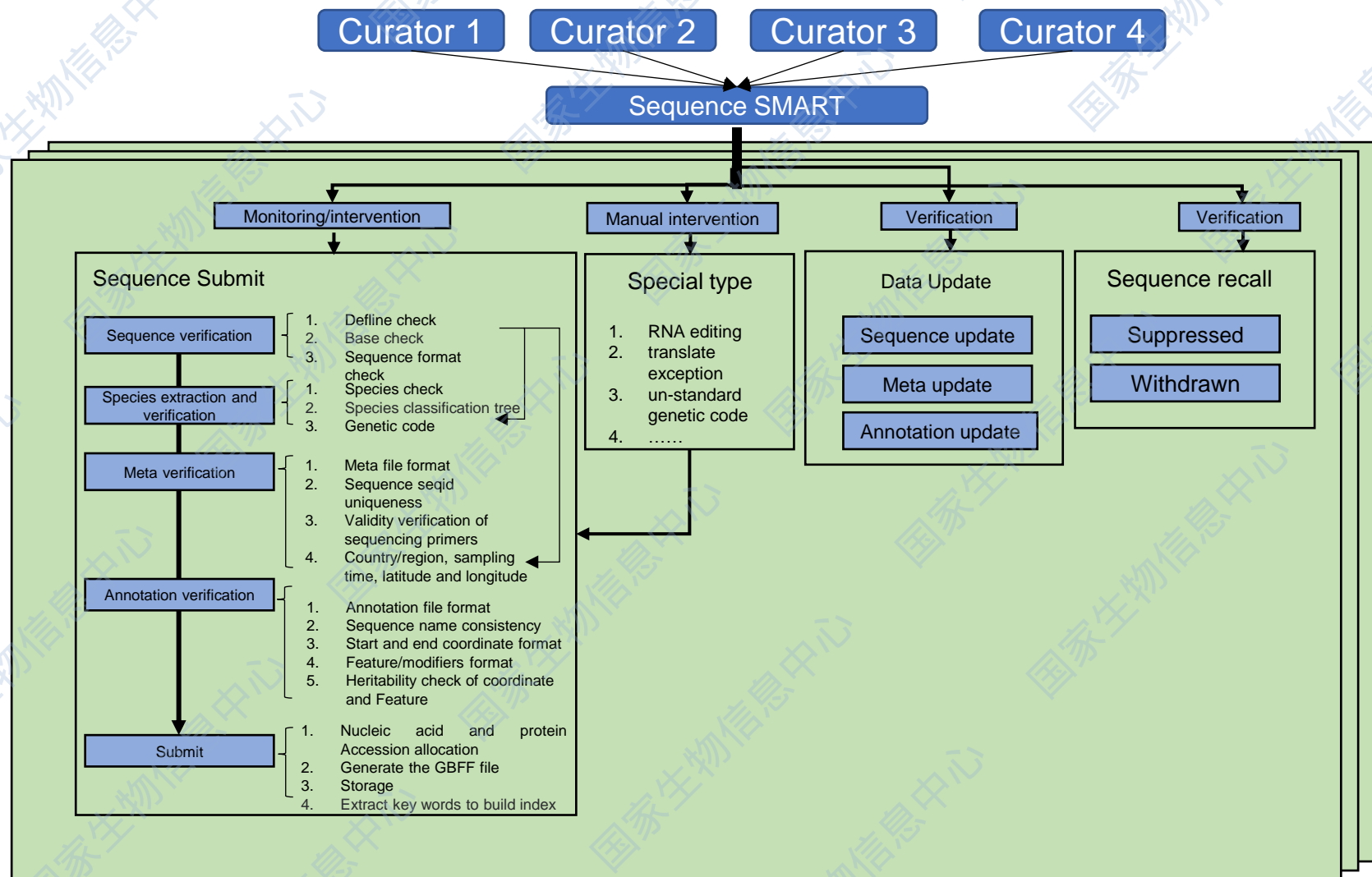
The data reported in this paper have been deposited in the GenBase [1] in National Genomics Data Center [2], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number **C\_AA000000** that is publicly accessible at <https://ngdc.cncb.ac.cn/genbase>.

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

★ Nucleotide Sequence(s)

☒ Paste sequence data (when sequence number < 40)

>Alligator\_VPS52-RNASE [Homo Sapiens]

CTTGAGCTAAGTTTCCCCAAATACTTAGTGAACCCAGGGTAGCATTGTAACGTGTT  
GTGTTTGTTCCTTTGCATGTCTATTACTACTAGGACCATTATATATTTTACATACTT  
AGCAATAAATACTTCTATAGTTAAACTGGTGGTAACTGGGTGTATTTTCTTTACT  
TTTCCTTCCCCACTTGCTCTGTAGCAACGCTTCTTTTACCTAAGCTAAAAATCCC  
TGAAGCTGAGAAAAAATTTGGGGTCTGCTCAGCAAGAGGCTAACTCTACTAGGAC

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

Select file...

Browse ...

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

Continue

Organism

Submission #0000099

Fill in missing Organism information

You did not include the name of the organism from which the sequence was isolated, see below.  
[Download organism template file](#)

>Alligator\_VPS52-RNASE does not have an organism

Organism Name:

Input same organism name for all sequences

or

Upload File:

Select file...

Browse ...

Continue

# 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

报错提示文件, 可点击下载查看

报错描述

result\_error.txt

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

提交者

出版信息

测序技术

序列

集合或批次

类别

元信息

特征

结果预览

特征

提交 #0000230

帮助

请注意

1. 请选择提交序列注释的以下三种格式之一，可参考注释文件规范

2. 请参考 [此链接](#) 作为TBL格式序列注释的说明

3. 如果需要以Excel格式提交序列注释特征，请下载模板文件 [GenBase\\_Features.xlsx](#)

上传您提交的序列注释:

☒ Excel格式

☐ TBL格式

☐ GFF3 格式

0000230\_feature.xlsx

Remove

result\_feature.xlsx

报错提示文件，可点击下载查看

result\_err.txt

报错描述

验证失败！您提交的文件中存在错误。完整的错误信息请查看'result\_err.txt'。下面列出了部分错误信息。

ERROR: In line , Feature Seq1fe is not included in the fasta file

保存&下一步

# 常见错误类型： Feature table错误示例

result\_feature - Excel

文件 开始 插入 绘图 页面布局 公式 数据 审阅 视图 帮助

等线 11 A A 自动换行 常规 条件格式 套用 表格格式 检查单元格 差 好 适中 计算 解释性文本 警告文本 链接单元格 输出

G28

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.

2

3

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)

5

6

Locations (坐标)				Attributes (属性)		
sequence_id	start	end	completeness	feature	qualifier	qualifier value
hchv1	86695	7518	complete	gene	gene	ft
hchv2	86695	7518	complete	CDS	product	aaa

Qualifier hints (提示)

feature	qualifier
regulatory	regulatory_class(Mandatory)
Example	
/regulatory_class='promoter'	
/regulatory_class='enhancer'	
/regulatory_class='ribosome_binding_site'	

<= The block of CDS

<= The block of CDS

提示此处填写有问题



# 常见错误类型： Feature table正确示例

Genbase\_Features (1) - Excel

搜索

文件

开始

插入

绘图

页面布局

公式

数据

审阅

视图

帮助

剪贴板

剪贴

复制

格式刷

等线

11

A

A

B

I

U

A

文

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



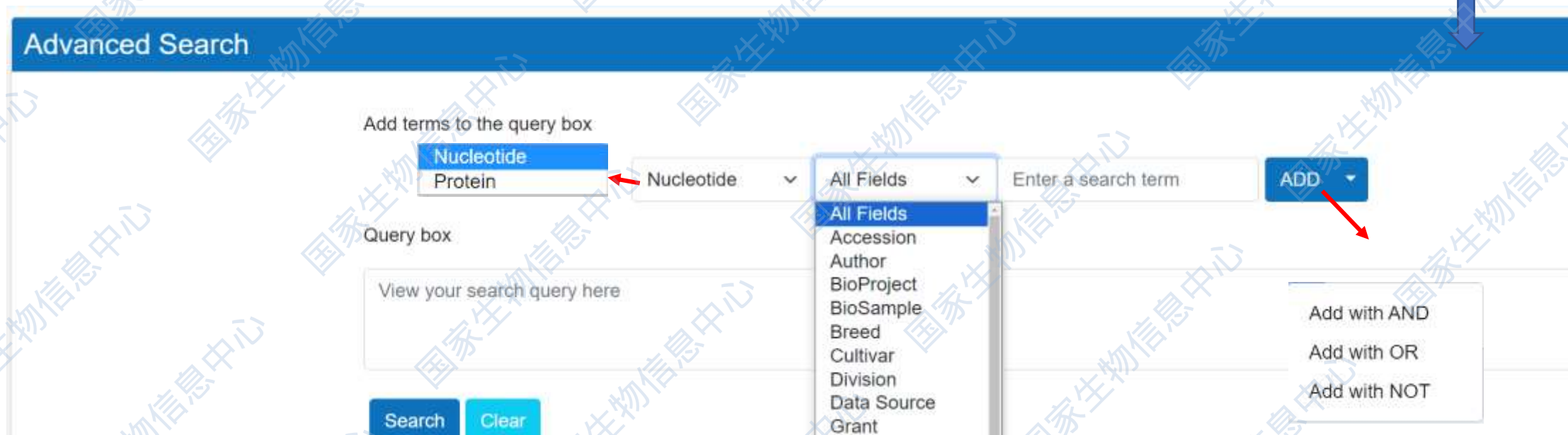
# GenBase数据发布与共享：搜索



The image shows the GenBase homepage with several annotations for the search process:

- 1**: A red box highlights the "Search" link in the top navigation bar.
- 2**: A red box highlights the "Type your keywords" input field, with the text "直接键入关键字搜索" (Directly enter keywords to search) next to it.
- 3**: A red box highlights the "Advanced" button, with the text "高级搜索" (Advanced search) next to it.

The homepage also includes a description of GenBase: "GenBase is a genetic sequence database that accepts user submissions (mRNA, genomic DNAs, ncRNA, or small genomes such as organelles, viruses, plasmids, phages from any organism) and integrates data from INSOC".



The image shows the "Advanced Search" interface with the following components:

- Add terms to the query box**: A section with a dropdown menu showing "Nucleotide" and "Protein".
- Query box**: A text area labeled "View your search query here".
- Search and Clear buttons**: Located at the bottom of the query box.
- Field Selection**: A dropdown menu showing "All Fields" and a list of fields: Accession, Author, BioProject, BioSample, Breed, Cultivar, Division, Data Source, Grant.
- Enter a search term**: A text input field.
- ADD button**: A blue button with a dropdown arrow.
- Logic Options**: A box showing "Add with AND", "Add with OR", and "Add with NOT".

# GenBase数据发布与共享：搜索结果

## 搜索结果

### 卡片格式

Summary 10/页 默认排序 下载文件

☒ 卡片或表格

1 到 10 条, 共 80,976 条

☐ 全选 共 8,098 页 << 首页 1 2 3 4 5 6 7 8 9

☐ Lespedeza potaninii chloroplast clone Tenggeii, complete genome

1. 149,059 bp DNA  
Accession: C\_AA097739.1  
GBFF FASTA

☐ Arenavirus isolate Hubei Rodents arena-like virus/2021 2 RNA directed RNA polymerase L (RdRp) gene, partial cds; and Z protein (Z) gene, complete cds

2. 7,241 bp DNA  
Accession: C\_AA087509.1  
GBFF FASTA

### 表格格式

☒ 卡片或表格

1 到 10 条, 共 80,976 条

☐ 全选 共 8,098 页 << 首页 1 2 3 4 5 6 7 8 9

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
<input type="checkbox"/> C_AA097739.1	Lespedeza potaninii	China			149,059	DNA	GenBase	2021-01-30
<input type="checkbox"/> C_AA087509.1	Arenavirus	China: Hubei	Hubei Rodents arena-like virus/2021 2	Apodemus agrarius	7,241	DNA	GenBase	2021-06-30
<input type="checkbox"/> C_AA087510.1	Chuviridae	China: Hubei	Hubei Rodents chu-like virus/2021 1	Rattus norvegicus	4,424	DNA	GenBase	2021-06-30
<input type="checkbox"/> 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

Summary ▾ 10 per page ▾ Sort by Default order ▾

Found 1 to 10 of 3,588 items

☐ Select All

**1. 选择感兴趣的条目**

- ☒ WO 2021206054-A/15: Genome modification method and genome modification Kit  
1. 20 bp RNA  
Accession: PA342009.1
- ☒ WO 2021206054-A/16: Genome modification method and genome modification Kit  
2. 20 bp RNA  
Accession: PA342010.1
- ☒ Homo sapiens isolate TWH-2683-0-1 truncated mutant tumor protein p53 (TP53) mRNA, complete cds  
3. 1,182 bp mRNA  
Accession: OL856015.1

**2. 点击**

**3. 选择格式**

**4. 下载**

Send to: ▾

Format

Summary

Create File

Format

GBFF

Separated

Merged

Create File

Format

Summary

Summary

GBFF

FASTA

Accession List

# GenBase数据发布与共享：搜索结果下载

Home / Search

GBFF ▾

Homo sapiens isolate SARS-CoV-2/human/CHN/0411\_4/2020 ORF1ab

GenBase: C\_AA000001.1

[FASTA](#)

LOCUS	C_AA000001	29819 bp	DNA	linear	PRI 26-APR-2022
DEFINITION	Homo sapiens isolate SARS-CoV-2/human/CHN/0411_4/2020 ORF1ab polyprotein gene, complete cds.				
ACCESSION	C_AA000001				
VERSION	C_AA000001.1				
KEYWORDS	.				
SOURCE	Homo sapiens (human)				
ORGANISM	Homo sapiens Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo.				

1. 点击

2. 选择格式

3. 下载

Send to: ▾

Format

GBFF

Create File

Format

GBFF

GBFF

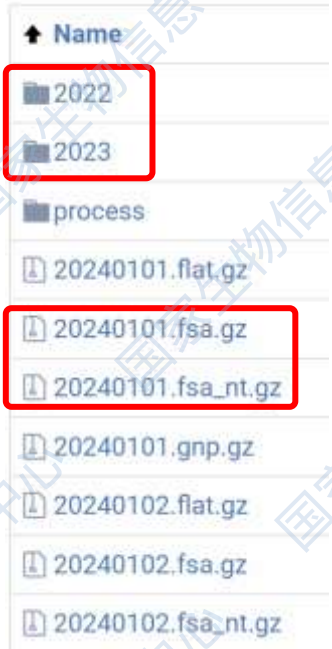
FASTA



# 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](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

#### GenBase质量控制系统的错误类型和修正方案

在提交数据时，GenBase会输出质量控制报告，您可以根据报告进行修正。并重新提交数据到GenBase。下面总结了可能遇到的错误类型和修正方案。 [下载PDF版本。](#)

SEQ\_FEATURE Internal Stop Codon, Internal Stop Codon

说明：序列中如果出现内部终止密码子，

建议：

1. 是否核酸序列中存在，如有请考虑是否修正核酸序列。
2. 是否核酸序列中存在，叶绿体、线粒体和核基因组使用不同的遗传密码子，使用核基因组密码子表可能会存在不匹配的终止密码子的问题，如是请考虑提交线粒体基因组。
3. 是否存在CDS不连续的情况，如果存在不连续的CDS情况，不连续的位置上标注存在终止密码子，如有多条CDS则需要分别标注出每条CDS的位置。
4. 是否存在partial的情况，如果存在partial的情况，标注下一个核苷酸位于前一个CDS的终止密码子后的位置，这里使用partial的情况，如果标注的是partial的情况，标注下一个核苷酸位于前一个CDS的终止密码子后的位置，这里使用partial的情况，标注下一个核苷酸位于前一个CDS的终止密码子后的位置，这里使用partial的情况。

<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

情况1：更新source modifier信息

以result\_assign为结尾的文件修改示例1：

```
#Header 1
LOCUS      Seq1      C_AA001108.1      29619      1-29619      update_meta
#Features 10
CDS       Seq1      C_AA01555.1      21281      344-13468,13468-21538      update_meta
CDS       Seq1      C_AA01557.1      3822      21543-25108      update_meta
CDS       Seq1      C_AA01558.1      820      25193-24220      update_meta
CDS       Seq1      C_AA01559.1      228      26245-26472      update_meta
CDS       Seq1      C_AA01560.1      469      26523-27191      update_meta
CDS       Seq1      C_AA01561.1      104      27202-27387      update_meta
CDS       Seq1      C_AA01562.1      146      27394-27719      update_meta
CDS       Seq1      C_AA01563.1      132      27756-27887      update_meta
CDS       Seq1      C_AA01564.1      146      27894-28030      update_meta
CDS       Seq1      C_AA01565.1      1265      28034-28513      update_meta
#Other: 0
gene      Seq1      Water      29619      1-29619      update_meta
```

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

## 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](mailto:genbase@big.ac.cn)

QQ群: 629388189

座机: 86-10-84097298

工作时间: 工作日9:00 - 17:00

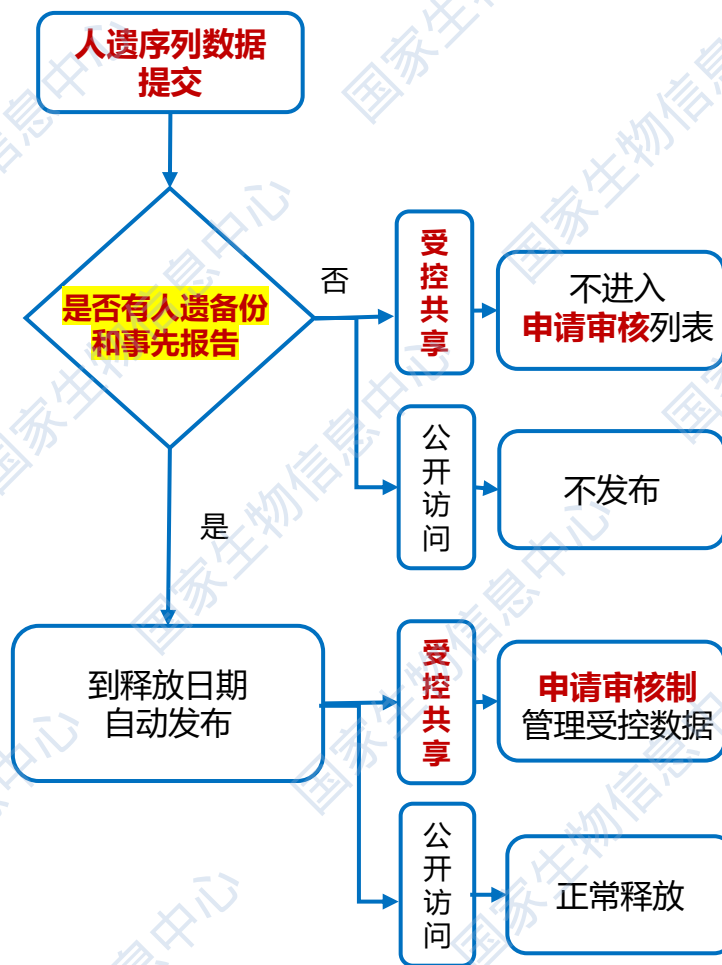
<https://ngdc.cncb.ac.cn/genbase>

# GenBase人类遗传相关序列共享



## 受控序列提交入口

GenBase采用申请审核制管理受控访问数据



## 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](mailto: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)

# 目录

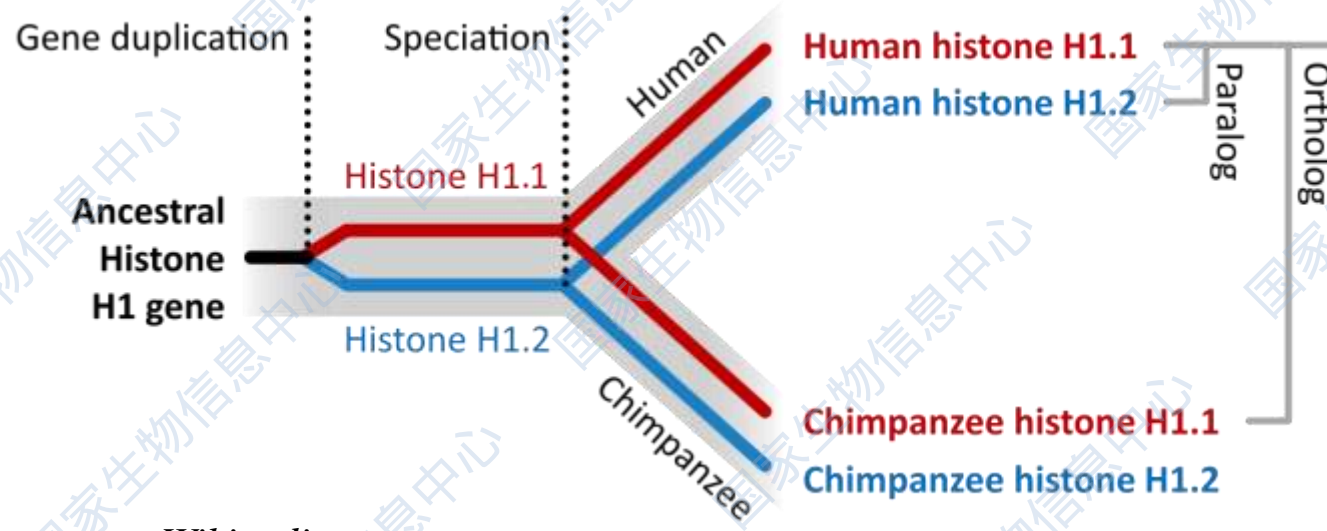
## CONTENTS

一、基因序列和基因组数据汇交的异同

二、基因序列数据汇交和共享

三、同源基因数据库HGD

# 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

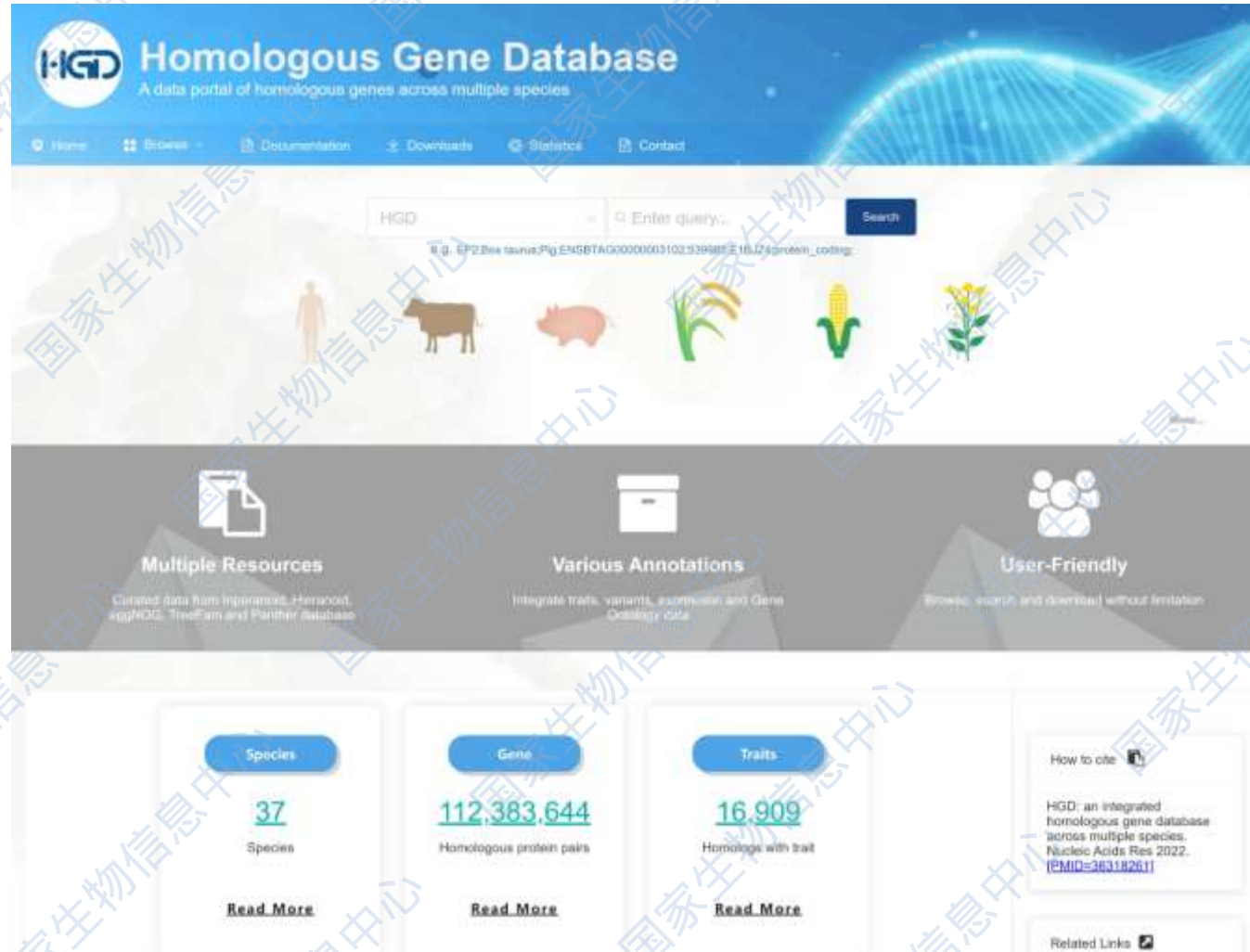
- 5个同源库

- 多组学注释信息

- 性状、变异、表达

- 基因功能注释信息

- Gene Ontology

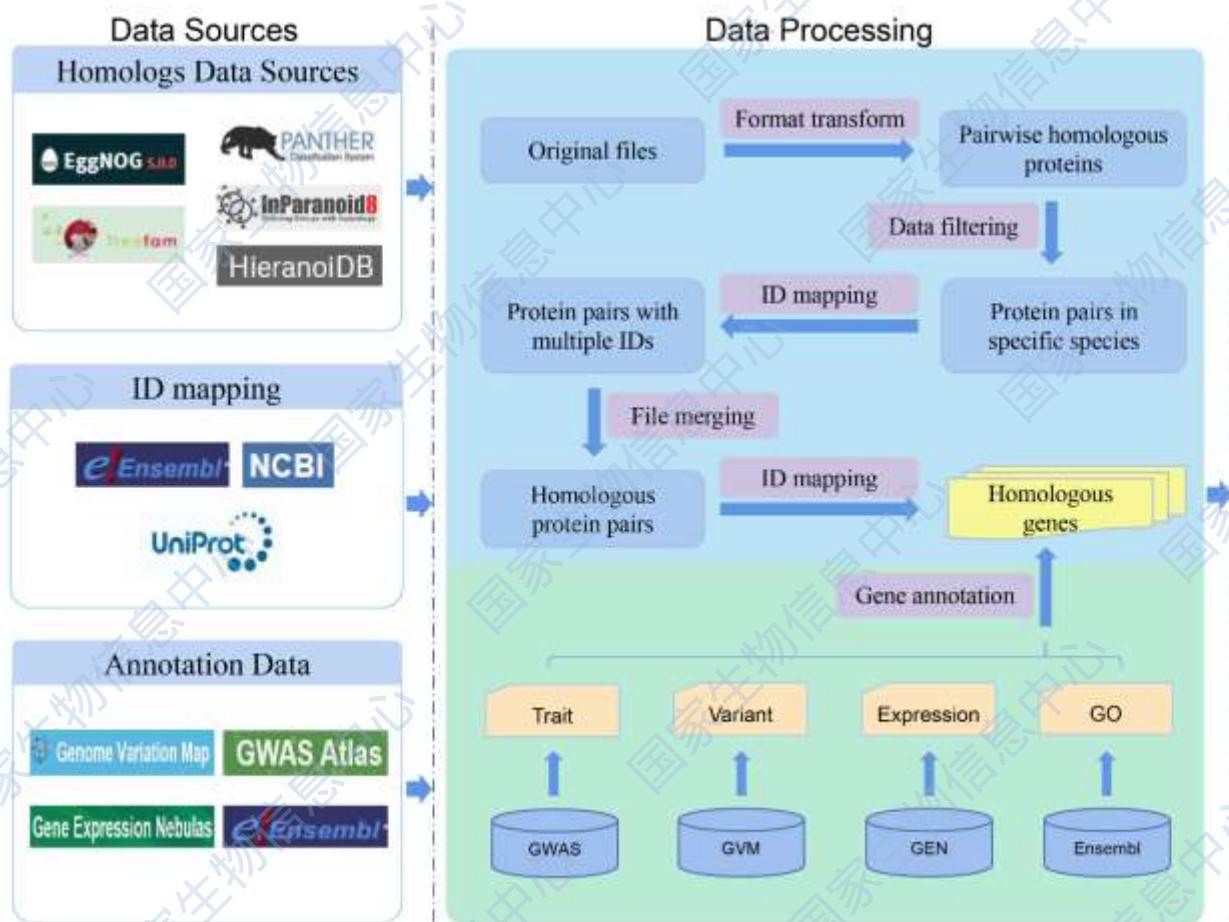


<https://ngdc.cncb.ac.cn/hgd>

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

Clear

Total 1 5/page < 1 > Go to 1

EP2 x

☐ Check All

☐ EP2 (*Oryza sativa*)

Regulation of panicle erectness, panicle length and grain size

Symobl: EP2

Uniprot: Q0D4N6

Ensembl Protein: Os07t0616000-01

Synonym:

BioType: protein coding

Homologous Gene(s):

F6HXQ3\_Vitvi09g00052\_P001 (*Vitis vinifera*)

Zm00001d022143 (*Zea mays*)

A0A3B6C2G9\_TraesCAD\_scaffold\_038815\_01G000100 (*Triticum aestivum*)

4332729 (*Oryza sativa*)

OsSTA81 (*Oryza sativa*)

M1C3M0\_PGSC0003DMT400058959 (*Solanum tuberosum*)

A0A3Q7FXL7 (*Solanum lycopersicum*)

A0A2G2XZ05\_PHT62571 (*Capsicum annuum*)

A0A2C9UYM4\_OAY36365 (*Manihot esculenta*)

100817241 (*Glycine max*)

BnaC05g52130D (*Brassica napus*)

AT1G72410 (*Arabidopsis thaliana*)

F4JFS3\_AT3G14172 (*Arabidopsis thaliana*)

A0A3N7EA94 (*Populus trichocarpa*)

LOC107906061 (*Gossypium hirsutum*)

A0A1D6IUL5\_Zm00001d022143\_P006 (*Zea mays*)

A0A1D5V1E8\_TraesCS2D02G176200 (*Triticum aestivum*)

SORBI\_3002G374400 (*Sorghum bicolor*)

4334871 (*Oryza sativa*)

Q7XI09\_Os07t0445600-01 (*Oryza sativa*)

M1CTF1\_PGSC0003DMT400074261 (*Solanum tuberosum*)

A0A3Q7GYT5 (*Solanum lycopersicum*)

A0A2G2ZBB8\_PHT79244 (*Capsicum annuum*)

A0A2C9W422\_OAY53252 (*Manihot esculenta*)

BnaA06g11670D (*Brassica napus*)

S007999 (*Arabidopsis thaliana*)

AT3G14172 (*Arabidopsis thaliana*)

Q9LQI8\_AT1G17360 (*Arabidopsis thaliana*)

101213033 (*Cucumis sativus*)

LOC107913848 (*Gossypium hirsutum*)

Zm00001d006864 (*Zea mays*)

A0A3B6AUT7\_TraesWEE\_scaffold\_108679\_01G000100 (*Triticum aestivum*)

SORBI\_3002G377000 (*Sorghum bicolor*)

Os06g0196500 (*Oryza sativa*)

102600185 (*Solanum tuberosum*)

PGSC0003DMG400002567 (*Solanum tuberosum*)

A0A3Q7JBB5 (*Solanum lycopersicum*)

A0A2G2ZV55\_PHT86078 (*Capsicum annuum*)

100812963 (*Glycine max*)

BnaC05g13520D (*Brassica napus*)

AT1G17360 (*Arabidopsis thaliana*)

F4IDB4\_AT1G72410 (*Arabidopsis thaliana*)

A0A2K2B2W6\_PNT44118 (*Populus trichocarpa*)

LOC107897450 (*Gossypium hirsutum*)

LOC107958375 (*Gossypium hirsutum*)

Filter by Species

Please select

Filter by condition AND OR

Filter by Trait

Select your favourite(s)

Animal

☐ growth and meat prod...

☐ animal welfare trait

☐ mammary gland and ...

☐ animal health trait

☐ nutrition trait

☐ reproduction trait

Plant

Filter by GO

Select your favourite(s)...

☐ nucleotide binding

☐ nucleic acid binding

☐ DNA binding

☐ DNA-binding transcription fa...

☐ RNA binding

☐ catalytic activity

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



Plant Trait Ontology  
biological process trait  
plant growth and development trait  
plant morphology trait  
plant quality trait  
plant stage trait  
study or facility trait  
stress trait  
yield trait  
Animal Trait Ontology for Livestock  
animal health trait  
animal welfare trait  
growth and meat production trait  
reproductive plant and wild production trait  
nutrition trait  
reproductive trait

Selected Trait : growth and meat production trait | II  
Definition: Any measurable or observable characteristics related to animal growth and meat production quantity

Animal

Plant

Query genes:

设置查询条件

Trait Name	Gene ID	Species Name	Taxon ID	Ensembl ID	Uniprot ID	Ensembl Protein ID	GWAS ID	P-value	GWAS Trait
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force
growth and meat production trait	ENSBTAG00000012789	Cattle	9913	ENSBTAG00000012789	A0A0Q1N8E9	ENSBTAP000000067781	20163038	0.0000121	warmer-bratzer shear force

Selected Gene

Gene ID: ENSBTAG00000012789

Uniprot ID: A0A0Q1N8E9

Ensembl Protein ID: ENSBTAP000000067781

Species Name: Cattle

Classification: animal

Trait Ontology term: growth and meat production trait

Homologous Gene Detail Information

Species	Taxon ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Data Source Count	Data Source
Cattle	9913	ENSBTAG00000012789	PRKD1	A0A0Q1N8E9	ENSBTAP000000067781	1	Ensembl
Cattle	9913	ENSBTAG00000012789	PRKD1	A0A0Q1N8E9	ENSBTAP000000067781	1	Ensembl
Cattle	9913	ENSBTAG00000012789	PRKD1	A0A0Q1N8E9	ENSBTAP000000067781	1	Ensembl

同源来源统计

Data Source Count	Data Source
1	Ensembl
1	Ensembl
1	Ensembl

具体同源来源

GWAS信息

GWAS Detail Information

Trait Name	Gene ID	Var ID	Sub Trait Name	Species	Pubmed ID	P-value
growth and meat production trait	ENSBTAG00000012789	1044283481	warmer-bratzer shear force	Cattle	20163038	0.0000121
growth and meat production trait	ENSBTAG00000012789	1044283444	warmer-bratzer shear force	Cattle	20163038	0.0000121

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

Consequence Type : frameshift\_variant

Animal Plant

Query gene: Query species: Search Clear

Consequence Type	Gene ID	Species Name	Accession ID	Query position	Gene	Protein	RefSeq	Ensembl	UniProt
frameshift_variant	ENSFCA00000001556	Cat	9685						
frameshift_variant	ENSFCA00000004499	Cat	9689						
frameshift_variant	ENSFCA00000000279	Cat	9685						
frameshift_variant	ENSFCA00000007382	Cat	9685						
frameshift_variant	ENSFCA00000004206	Cat	9689						

同源基因信息

Selected Gene

Gene ID: ENSFCAG00000004499 Uniprot ID: A6A0AMPY5 Ensembl Protein ID: ENSFCAP00000004158

Species Name: Cat Classification: animal Consequence type: frameshift\_variant

Homologous Gene Detail Information

Species	Accession ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Data Source Count	Data Source
Cat	9685	ENSFCA00000000092	TNFSF14	M0W0C8	ENSFCA000000004602	1	
Cat	9689	ENSFCA00000000498	LTA	M0W2A5	ENSFCA00000004158	1	
Cat	9685	ENSFCA00000000074	TNFSF15	M0W0U0	ENSFCA000000000318	1	
Cat	9685	ENSFCA00000000092	TNFSF14	M0W4Y2	ENSFCA000000004601	1	
Cat	9685	ENSFCA00000000499	TNF	A6A0AMPY3	ENSFCA000000004159	1	

Variants Detail Information

Gene ID	Uniprot ID	Position	Allele	MAF	Class	Consequence Type/Effect	Gene Symbol
ENSFCA00000000044 96 ENSFCAG000000004499	Uni299728	B2:32573377	T/C	0.31054	SNP	5_prime_UTR_variant(MODIFIER), upstream_gene_variant(MODIFIER)	LTA,TNF
ENSFCA00000000044 96 ENSFCAG000000004499	Uni299729	B2:32573378	G/A	0.11054	SNP	5_prime_UTR_variant(MODIFIER), upstream_gene_variant(MODIFIER)	LTA,TNF
ENSFCA00000000044 96 ENSFCAG000000004499	Uni299730	B2:32573438	A/G	0.05788	SNP	5_prime_UTR_variant(MODIFIER), upstream_gene_variant(MODIFIER)	LTA,TNF
ENSFCA00000000044 96 ENSFCAG000000004499	Uni299731	B2:32573637	C/T	0.11054	SNP	5_prime_UTR_variant(MODIFIER), upstream_gene_variant(MODIFIER)	LTA,TNF

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

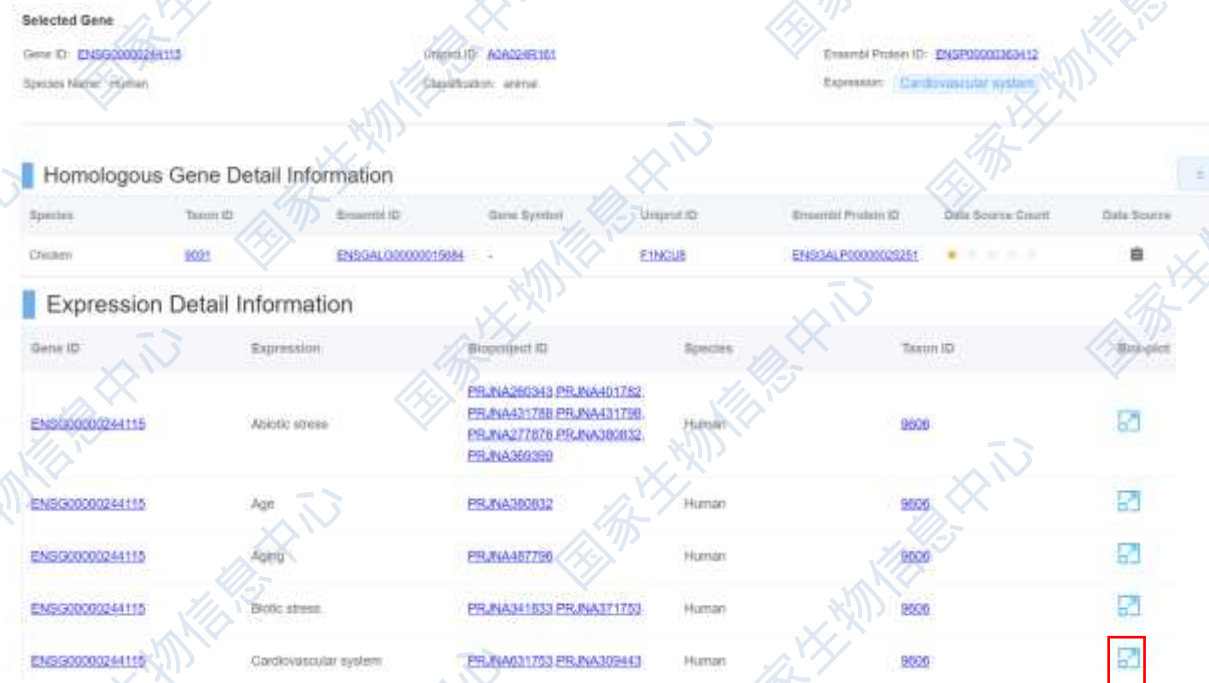


蓝色：同源基因

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

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



基因在BioProject表达量盒图



# 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

biological\_process

- negative regulation of transcription by RNA polymerase
- carbohydrate metabolic process
- transcription, DNA-templated
- regulation of transcription, DNA-templated
- regulation of transcription by RNA polymerase II
- translation

Gene Ontology Annotation : protein phosphorylation | ID: GO:0006468

Definition: The process of introducing a phosphate group on to a protein. [GOC:nh]

Animal Plant

Query genes: Query species:

Gene Ontology

Gene Ontology	GO ID	Gene ID	Species Name	Taxon ID	Chicken	Guinea pig	Canine	Human	Mouse	Rat	Macaca	Hu	Shrew	Other
protein phosphorylation	GO:0006468	ENSGALG000000021251	Cattle	9913										
protein phosphorylation	GO:0006468	ENSGALG000000021252	Cattle	9913										
protein phosphorylation	GO:0006468	ENSGALG000000021241	Cattle	9913										
protein phosphorylation	GO:0006468	ENSGALG000000021251	Cattle	9913										
protein phosphorylation	GO:0006468	ENSGALG000000021493	Cattle	9913										
protein phosphorylation	GO:0006468	ENSGALG000000021493	Cattle	9913										

Selected Gene

Gene ID: ENSGALG000000021251  
Species Name: Cattle  
Uniprot ID: A6CLAV  
Classification: animal  
Ensembl Protein ID: ENSGALP000000021251  
Gene ontology: protein phosphorylation

Homologous Gene Detail Information

Species	Taxon ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Defn Source Count	Data Source
Chicken	9031	ENSGALG000000042189	RPL6A	AAA1C6PWH	ENSGALP000000042189	4	
Chicken	9031	ENSGALG000000021252	RPL6I	E1C036	ENSGALP000000021252	4	
Chicken	9031	ENSGALG000000021251	ANKK1	E1B2F3	ENSGALP000000021251	4	
Chicken	9031	ENSGALG000000021250	MLKL	E1C0P2	ENSGALP000000021250	4	

GO Detail Information

Gene ID	GO term	GO ID	Sub-GO ID	Species	Taxon ID
ENSGALG000000021251	ATP binding	GO:0005524	GO:0005524	Chicken	9031
ENSGALG000000021251	protein phosphorylation	GO:0006468	GO:0006468	Chicken	9031
ENSGALG000000021251	intracellular membrane-bounded organelle	GO:0043231	GO:0000034	Chicken	9031
ENSGALG000000021251	nucleus	GO:0005634	GO:0005634	Chicken	9031
ENSGALG000000021251	protein binding	GO:0005515	GO:0005515	Chicken	9031
ENSGALG000000021251	protein kinase activity	GO:0004673	GO:0004673	Chicken	9031
ENSGALG000000021251	transferase activity	GO:0016740	GO:0004673	Chicken	9031

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

# 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	<a href="#">9031</a>	<a href="#">214</a>	<a href="#">13327</a>	<a href="#">15497</a>	<a href="#">13034</a>
Ailuropoda melan oleuca	Giant panda	<a href="#">9646</a>	-	<a href="#">12580</a>	-	<a href="#">13252</a>
Bos taurus	Cattle	<a href="#">9913</a>	<a href="#">238</a>	<a href="#">17738</a>	<a href="#">20721</a>	<a href="#">17872</a>
Canis familiaris	Dog	<a href="#">9615</a>	-	-	<a href="#">1077</a>	<a href="#">971</a>

## 物种统计分布

Animals Plants Others



## 同源基因检索列表



## 跨物种同源基因统计情况



## 同源基因详细信息

## Gene Basic Information

Ensembl Gene ID: <a href="#">Os07t0010000</a>	Enrez Gene ID: <a href="#">4343954</a>	RefSeq ID: <a href="#">XP_013647289</a>
Uniprot ID: <a href="#">Q0D4N6</a>	Ensembl Protein ID: <a href="#">Os07t0010000-01</a>	Gene Synonym: -
Gene Symbol: <a href="#">EP2</a>	Gene Type: <a href="#">protein_coding</a>	Latin Name: <a href="#">Oryza sativa</a>
Species Common Name: <a href="#">Rice</a>	Taxon ID: <a href="#">4530</a>	Chromosome: <a href="#">7</a>
Gene Start: <a href="#">25381698</a>	Gene End: <a href="#">25389532</a>	Gene Description: <a href="#">Regulation of pericarp electrophoresis; pericarp length and grain size</a>

**Homologous Gene**

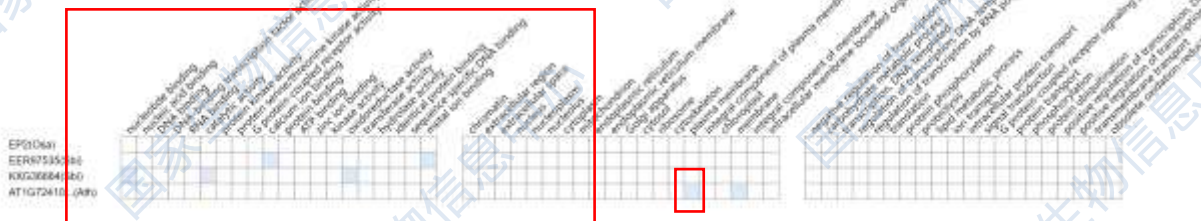
Species	Taxon ID	Gene Symbol	Ensembl ID	Ensembl ID	UniProt ID	Ensembl Protein ID	Data Source Count	Data Source
Broad soft off	4566		<a href="#">TrnaCS2032G17620</a>		<a href="#">A0A1DSV1E8</a>	<a href="#">TrnaCS2032G17620</a>	1	<a href="#">UniProt</a>
Broad soft off	4566		<a href="#">TrnaSVEE_scotfold_10 8078_01G000100</a>		<a href="#">A0A365A0T7</a>	<a href="#">TrnaSVEE_scotfold_10 8078_01G000100</a>	1	<a href="#">UniProt</a>

### Gene Ontology

## 同源基因比较

## 物种筛选过滤

## 同源基因比较面板



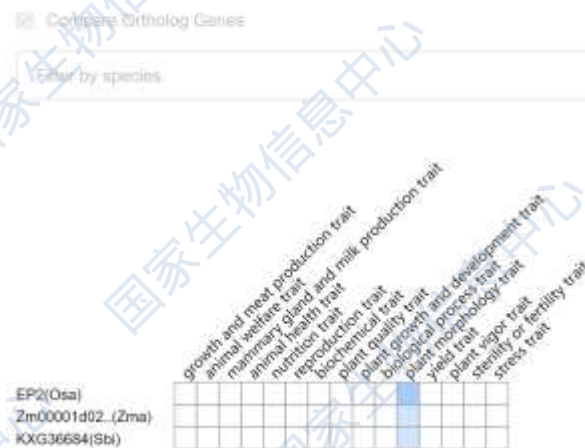
## ■ Variants

# 变异



## Traits

## 性状

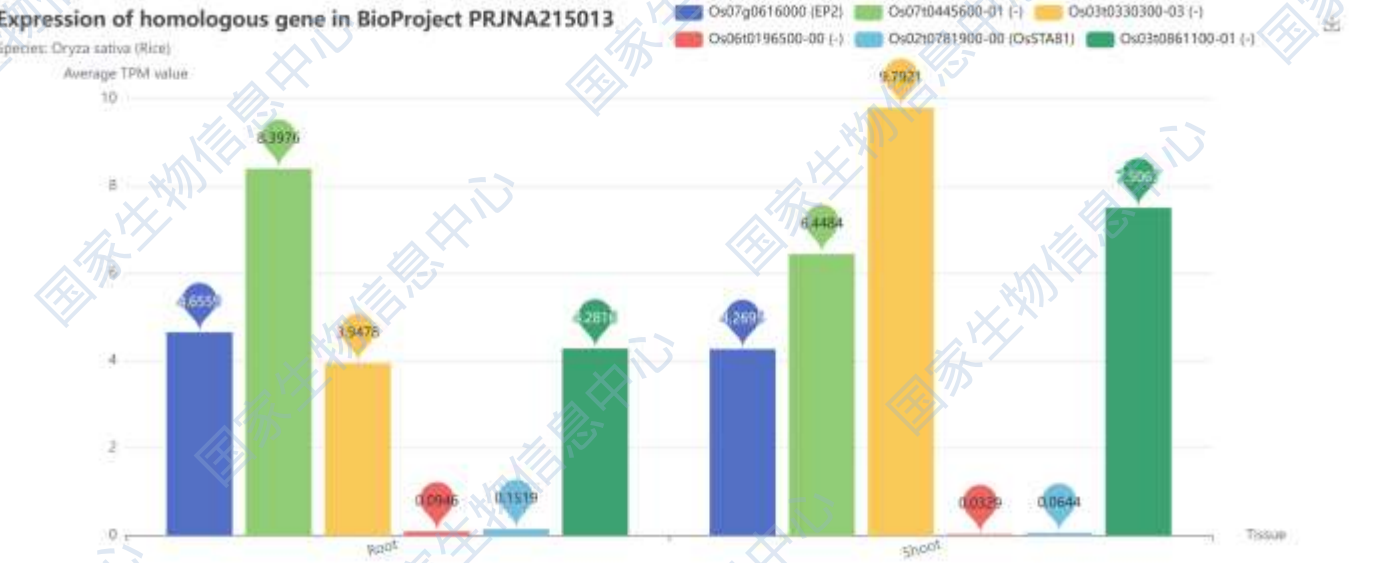


表达

Expression



Gene ID	Bioproject ID	Tissue	Average TPM	Bar-plot
<a href="#">Os07g0616000</a>	<a href="#">PRJNA380230</a>	Shoot	5.185	
<a href="#">Os07g0616000</a>	<a href="#">PRJNA215013</a>	Root	4.6559	
<a href="#">Os07g0616000</a>	<a href="#">PRJNA215013</a>	Shoot	4.2695	
<a href="#">Os07g0616000</a>	<a href="#">PRJNA436566</a>	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](mailto:genbase@big.ac.cn)

**QQ群:** 629388189

