



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

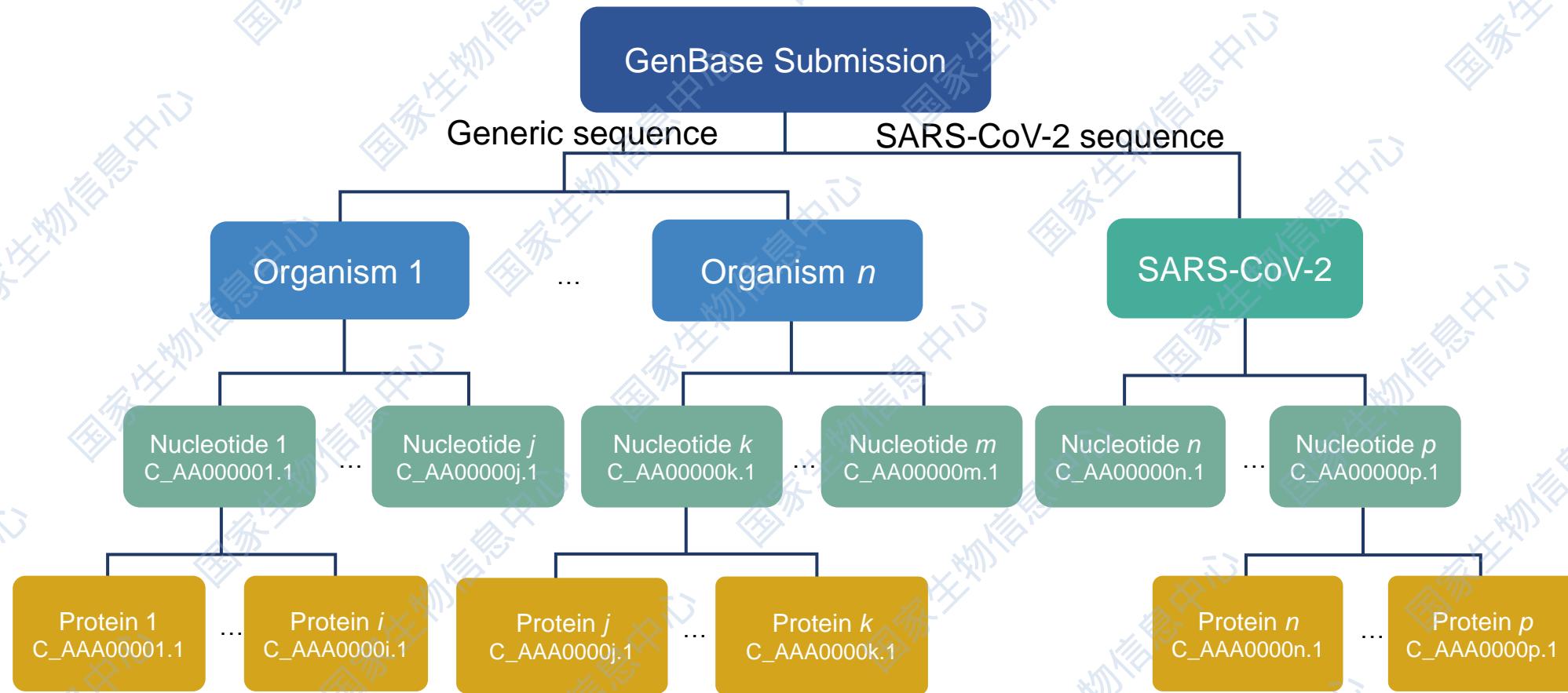
GenBase是一个公共、免费的基因序列数据归档库，收录、存储、管理与共享病毒、细胞器、质粒和基因片段等核酸序列及其注释信息

The screenshot shows the homepage of the GenBase database. At the top, there is a navigation bar with links for Home, Data Submission, Search, Download, Statistics, Standards, and Help Documents. On the right side of the header are buttons for Login, Register, and Language selection. Below the header, a large search bar has "Nucleotide" selected and a placeholder "Type your keywords". A message box on the left states: "GenBase 是一个公开、共享的基因序列归档库。接受用户提交（包括任何生物体的mRNA, DNA片段、非编码RNA或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自INSDC的核酸和蛋白序列数据。" Below the search bar, a status message says "成功! AA001106.1; MH011443.1; GB00003962". The main content area features three circular icons: one for Species (物种) with 592,478 entries, one for Nucleic acids (核酸) with 374,134,720 entries, and one for Proteins (蛋白) with 619,151,637 entries. To the right, a "Recent Updates" section lists four items: "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", and "2024-05-08 数据交换统计信息上线". On the far right, there are sections for "International Data GenBank Integration" (updated 2024-10-21, 890 nucleic acids, 4,032 proteins) and "GenBase Data Release" (released 2024-10-19, 232 nucleic acids, 1,093 proteins), along with a "New Features" link.

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

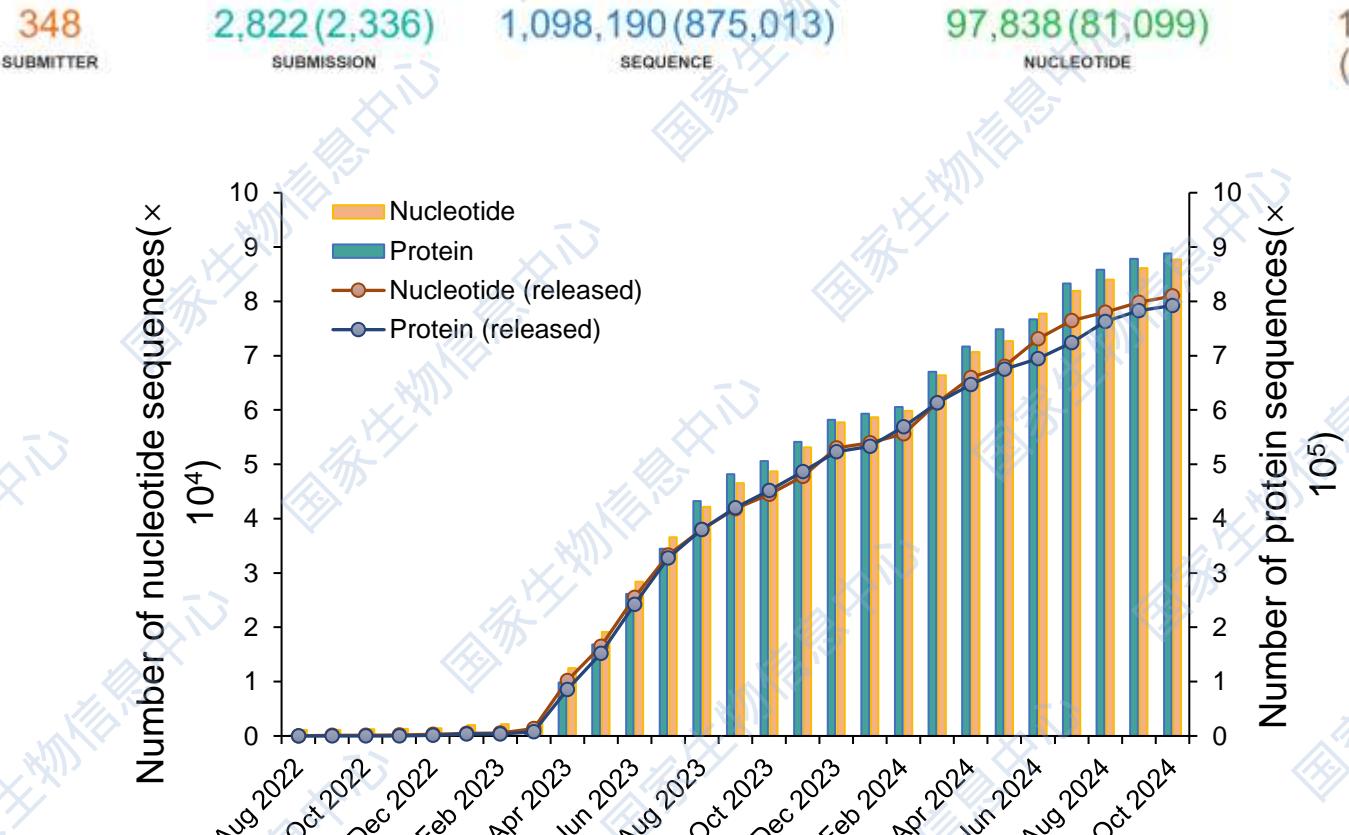
Genomics Proteomics Bioinformatics, 2024

GenBase的数据模型

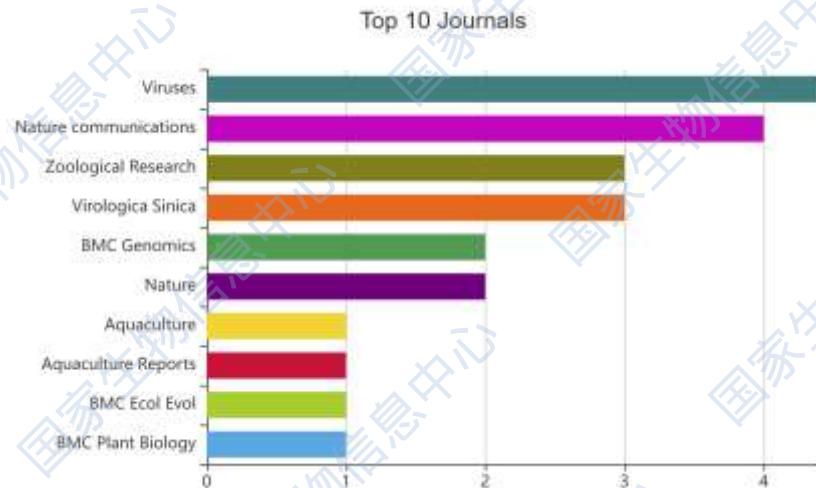


Nucleotide accession: C_ + 2 characters+ 6 numbers + version number

基因序列库GenBase的用户提交数据情况



Genomics Proteomics Bioinformatics (2024)

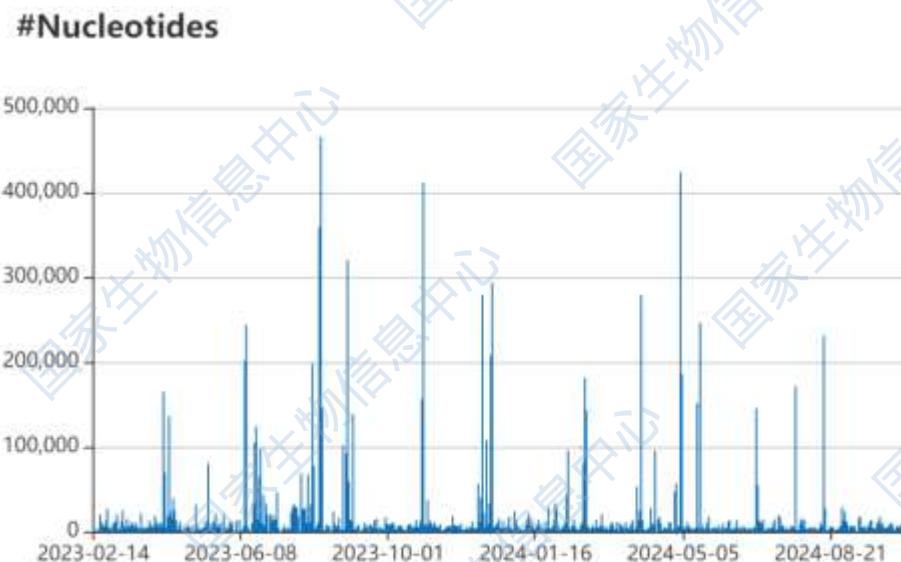


re3data.org

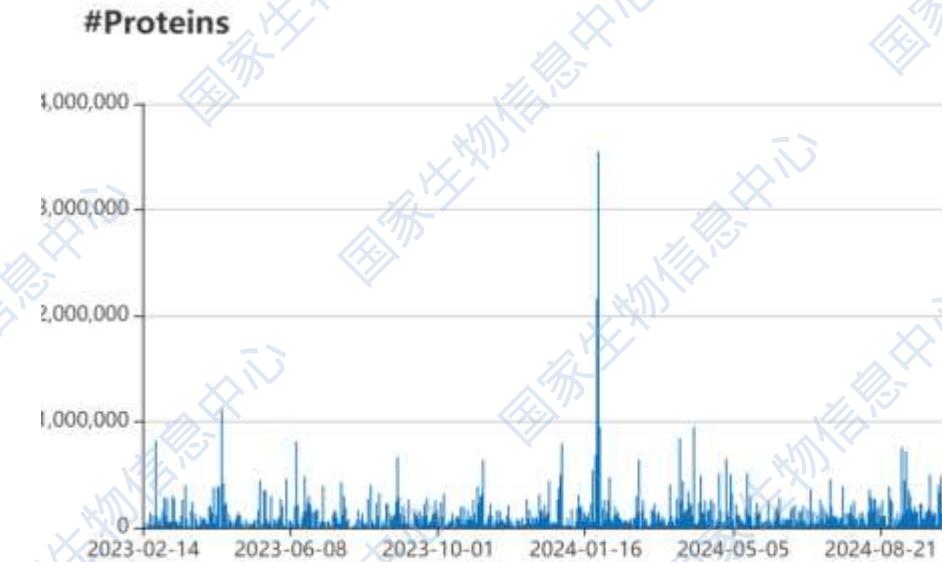
FAIRsharing.org
standards, databases, policies

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	<p>小基因组和基因 • 病毒基因组 • 细胞器基因组 • 质粒基因组 • 所有物种基因片段</p>	<p>• 基因片段序列 • 注释数据（非必需）</p>	<p>可以不涉及 BioProject 和 BioSample</p>
基因组数据库GWH	<p>大基因组 • 真核生物全基因组 • 原核生物全基因组 • 宏基因组组装数据</p>	<p>• 基因组组装序列 • 注释数据（非必需）</p>	<p>涉及 BioProject 和 BioSample</p>

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 shows the GenBase website homepage. At the top, there is a navigation bar with links for Home, Data Submission, Search, Download, Statistics, Standards, Help Documents, Login, Register, and Language selection. The main content area features a search bar with dropdown options for 'Nucleotide' or 'Protein' and a text input field 'Type your keywords'. Below the search bar is a note about the database's purpose: "GenBase 是一个公开、共享的基因序列归档库，接受用户提交（包括任何生物体的 mRNA、DNA 片段、非编码 RNA 或小基因组，如细胞器、病毒、质粒、噬菌体等）核酸序列及其注释数据。同时，已整合来自 INSDC 的核酸和蛋白序列数据。" On the left, there is a section titled "汇交与整合数据" (Data Submission and Integration) showing counts for Species (592,478), Nucleic acids (374,131,350), and Proteins (619,127,880). On the right, there is a "Recent Updates" (Recent Updates) section listing events from July 25, 2024, to May 8, 2024, and a "International Data Integration" (International Data Integration) section for GenBank.

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

Nucleotide Type your keywords

例如 C_AAA001108.1; MH011443.1; GB0003982

检索 高级检索

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

The screenshot shows the NGDC Central Authentication Service login interface. At the top, there are logos for CNCB and NGDC, and navigation links for Databases, Tools, Standards, and Publications. The main form is titled "Enter your Username and Password". It contains three input fields: "Email" (zhaoxuetong@big.ac.cn), "Password" (redacted), and "Check code" (x4ab). Below these are two buttons: "Keep me signed in" (unchecked) and "Login". To the right of the "Login" button is a red box containing the text "用户名" (Username). Further down are "Reset" and "Register" buttons, with "Register" also enclosed in a red box and labeled "密码" (Password). A red box at the bottom left contains the text "点击此处 登陆" (Click here to log in). On the right side of the page, there is a sidebar titled "Central Authentication Service" listing various services: 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), and Scientific Data Archive System (SDAS).

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

准备工作：1.账户准备-注册和登陆

The screenshot shows the 'Welcome to register for an account of NGDC' page. It features two main sections: 'Register information' on the left and 'Account Login Information' on the right. The 'Register information' section contains fields for 'Email*' and 'Password*'. The 'Email*' field includes a note about receiving activation links and avoiding junk email. The 'Password*' field includes a note about password complexity and length. Both fields have validation messages displayed below them.

Welcome to register for an account of NGDC

Register information

Account Login Information

Email *

注意：由于需通过邮件接收激活链接，请确保该邮箱属于自己并能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过滤策略，检查激活邮件是否被邮件系统误认为垃圾邮件，或者被邮件系统拦截！

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.

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

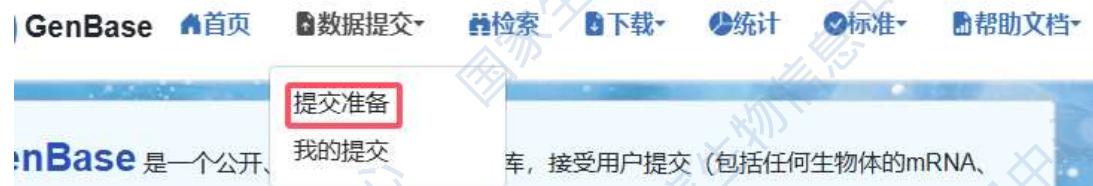
注意：密码必须同时包含大写字母、小写字母、数字、特殊字符，且长度在8~30之间！

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  
aaccgatatacagagagagga  
>Seq2 [organism=Homo Sapiens] Definition Line for Seq2  
atctgaatagagattttt
```

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_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, fill in the first row.							
#!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”、“null”或“unavailable”...							
##Color Code	required	optional (recommended)	optional	##4	##5	##6	
##Column Number	##1	##2	##3				
##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	Age
##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：

```
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								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
33								
34								
35								
36								
37								
myFeature			example	zh-example	feature_inspection_sheet			

3. 请在此处填写具体信息
不要删除表头、行列名称、填写说明等内容

Qualifier hints (提示)

feature  qualifier 

gene  gene 

Example

/gene=ilvE

2. 提示区
提示填写G列内容

准备工作：2. 数据准备-特征注释文件

准备工作：2. 数据准备-特征注释文件

A	B	C	D	E
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 (Optional) citation (Optional) db_xref (Optional) experiment (Optional) function (Optional) gene (Optional) gene_synonym (Optional) inference (Optional) locus_tag (Optional) map (Optional) note (Optional) old_locus_tag (Optional) standard_name (Optional) trans_splicing	text number <database> <identifier> [CATEGORY] text text text text [CATEGORY] [TYPE] (same spec) text (single token) text text text (single token) text boolean	NA
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 (Optional) citation (Optional) db_xref (Optional) experiment (Optional) function (Optional) gene (Optional) gene_synonym (Optional) inference (Optional) locus_tag (Optional) map (Optional) note (Optional) old_locus_tag (Optional) standard_name (Optional) trans_splicing	text number <database> <identifier> [CATEGORY] text text text text [CATEGORY] [TYPE] (same spec) text (single token) text text text text boolean	NA
coding sequence, sequence of nucleotides that corresponds with the sequence of amino acids in a protein (location includes	CDS	(Optional) allele (Optional) artificial_location (Optional) circular_RNA (Optional) citation (Optional) codon_start (Optional) db_xref (Optional) EC_number (Optional) exception (Optional) experiment (Optional) function (Optional) gene (Optional) gene_synonym (Optional) inference (Optional) locus_tag (Optional) map	[artificial_location value] boolean number <1 or 2 or 3> <database> <identifier> text [exception value] [CATEGORY] text text text text text text text text	lesson on new value of 1 or 2 or 3, indicating the offset at which the first complete codon of a coding feature can be found, relative to the first base of that feature. /transl_table defines the genetic code table used if other than the universal genetic code table, genetic code exceptions outside the range of the specified tables is reported in /transl_except qualifier. /protein_id consists of a stable ID portion (from the end of 2018 new

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	10	567	complete	mRNA	regulatory_class	Mat promoter
seq1	789	1320			gene	ubc42
seq1	54	567	complete	CDS	gene	ubc42
seq1	789	1254			product	ubiquitin conjugating enzyme
seq1	10	567	complete	exon	function	cell division control
seq1	568	783	complete	intron	number	1
seq1	789	1320	complete	exon	gene	ubc42
seq1	1310	1317	complete	regulatory	number	1
seq1					product	ubiquitin conjugating enzyme
seq1					function	cell division control
seq1					number	2
seq1					gene	ubc42
seq1					product	ubiquitin conjugating enzyme
seq1					function	cell division control
seq1					number	2
seq1					gene	ubc42
seq1					product	ubiquitin conjugating enzyme
seq1					function	cell division control
seq1					number	2
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

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

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

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

提交流程：步骤1. 提交者信息 (Submitter)

提交者

* 姓名: bx
* 电子邮箱: tangbx@big.ac.cn
* 组织/机构: Beijing Institute of Genomics, Chinese Academy of Sciences
电话:
* 街: 1
邮编: 1

中间名:
* 爵: tang

电子邮件 (次要):
提交者所在组织/机构网址: https://
* 部门: 010
传真:
* 城市: 1
* 省/州: 1
* 国家: Afghanistan

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 群组 元信息 特征 指南或范例

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 bixia	中间名	*姓 tang	后缀	移除
增加更多通讯作者.				

可选择增加通讯作者

文献信息 #1

请提供本次提交相关论文的标题和出版细节(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

根据具体情况如实填写

添加另外的出版信息 **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标题和出版细节(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

文章标题输入框

*文献作者

与通讯作者一致 指定新作者

可选择指定新作者，并且正确添加作者信息

*名	中间名	*姓	后缀	移除
bixia		tang		X

Add 更多参考作者

添加另外的出版信息

保存&下一步

保存并下一步

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

上述序列为：

- 454
- MGISEQ-200RS
- Nanopore

- Helicos
- MGISEQ-2000
- PacBio

- Illumina
- MGI DNBSEQ
- SOLiD

保存&下一步

保存并下一步

根据具体情况，详细填写相关信息

提交流程：步骤4. 核酸序列 (Nucleotide)



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？

审核完立即发布
 拟发布日期

日期格式为 'YYYY-Mon-DD' (例如: 2022-02-07),
发布日期必须是今天起1个月后, 到未来3年前。

发布前阅读免责声明

1. 作者可以指定一个日期。在指定日期前不发布该序列数据。
2. 如果提交者想要更改发布日期, 请联系GenBank工作组: 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, 该序列将在文章发表后立即发布。否则, 在指定日期之后发布。
4. 提交者一旦获取出版信息, 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
5. 数据发布后, 未来可能与NCBI同步。不再联系提交者进行确认。

* 接受 不接受

设置提交标题

BioProject

BioProject:

资助金

资助机构: 资助类别: 资助项目: 资助项目名称: 国家:

更多资助

填写BioProject信息 (可选)

填写资助基金信息

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: **Linear** Circular

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

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

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。

右侧弹出菜单示例:

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

- nucleotide
- mitochondria
- plastid

- Single strand
- Double strand
- Mixed strand

**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)

* 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,.fastal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: <ftp://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 **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，可在下一页补充

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: Homo sapiens

所有序列都使用该物种名称

或者

上传文件: Select file... Browse

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

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)



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

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](#).

* 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

0000062_modifier.xlsx

Remove

验证成功!

保存&下一步

3. 保存并下一步

1. 点击下载Modifiers文件模板

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

2. 点击上传填好的Modifiers文件

提交流程：步骤9. 提交Features文件 (Features)



特征

提交 #0000062

帮助

请注意

- 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](#)
- 请参考[此链接](#)作为TBL格式序列注释的说明
- 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

3.保存并下一步

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

#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)

提交者

出版信息

测序技术

序列

稳定性

集合或批次

类别

元信息

特征

结果预览

结果预览

提交0000062

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GenBank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 genbank@ncbi.nlm.nih.gov。如果您有任何问题，也可以通过此邮箱与我们联系。

审查提交记录

以下存在 1 条 GenBank 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	0530H2379WHT	2483 bp.	RNA	Linear	UHR 14-AUG-2022
DEFINITION	Human picobirnavirus.				
ACCESSION	0530H2379WHT				
VERSION	0530H2379WHT				
KEYWORDS	*				
SOURCE	Human picobirnavirus				
ORGANISM	Human picobirnavirus				
	Riboviria; Orthornavirales; Picuviricota; Duglaspiviriceta;				
	Dornavirales; 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 (14-AUG-2022) National Genome Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, NO.1 Beichen West Road, Chaoyang District, Beijing 100101, China				
CONTENT	#Genome-Assembly-START#				
	Assembly Method :: Trinity-2.11.0				
	Sequencing Technology :: Other				
	#Genome-Assembly-END#				
FEATURES	Location/Qualifiers				
source	1..2483				
	/organism="Human picobirnavirus"				
	/mol_type="genomic RNA"				
	/host=""				
	/country="China; Beijing"				
	/collection_date="2018-06-01 00:00:00"				
CDS	156..701				
	/codon_start=1				
	/product="hypothetical protein"				
	/translation="MIAAICUAAETRERHNITVVERGTRHIVVERIGYGTAAEIR NNKAYEVQIQBALAAETAMIVATEOINMYTAQNLALOFQAANETTSEVGTTTGGAT ESTIQGQLEETGRMULAEPOHNALEOFRMHEFOTLQONIST3AGAYEDVSTOLRQI VEEVVKGIIDHWRKILFGG"				
CDS	701..2814				
	/codon_start=1				
	/product="capid protein"				
	/translation="MEDENRUTIFDRAKYGQEVKRRRBDQIHERYERDICHRYA"				

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：分享申稿人链接

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/ Cancel Modify Grant
0000089	Nzfth1,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/ Cancel Modify Grant
0000086	-	[REDACTED]	2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success Accession	public https://ngdc.cncb.ac.cn/genbase/review/ 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	操作
0000086	-	[REDACTED]	2022-08-25 20:35:03	2022-09-01 17:36:37	2022-08-25	Audit Success	Accession public
0000202	[REDACTED]	[REDACTED]	2022-11-29 11:09:41	2022-11-30 10:35:22	2022-11-29	Audit Success	Accession confidential

点击查看Accession编号

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：

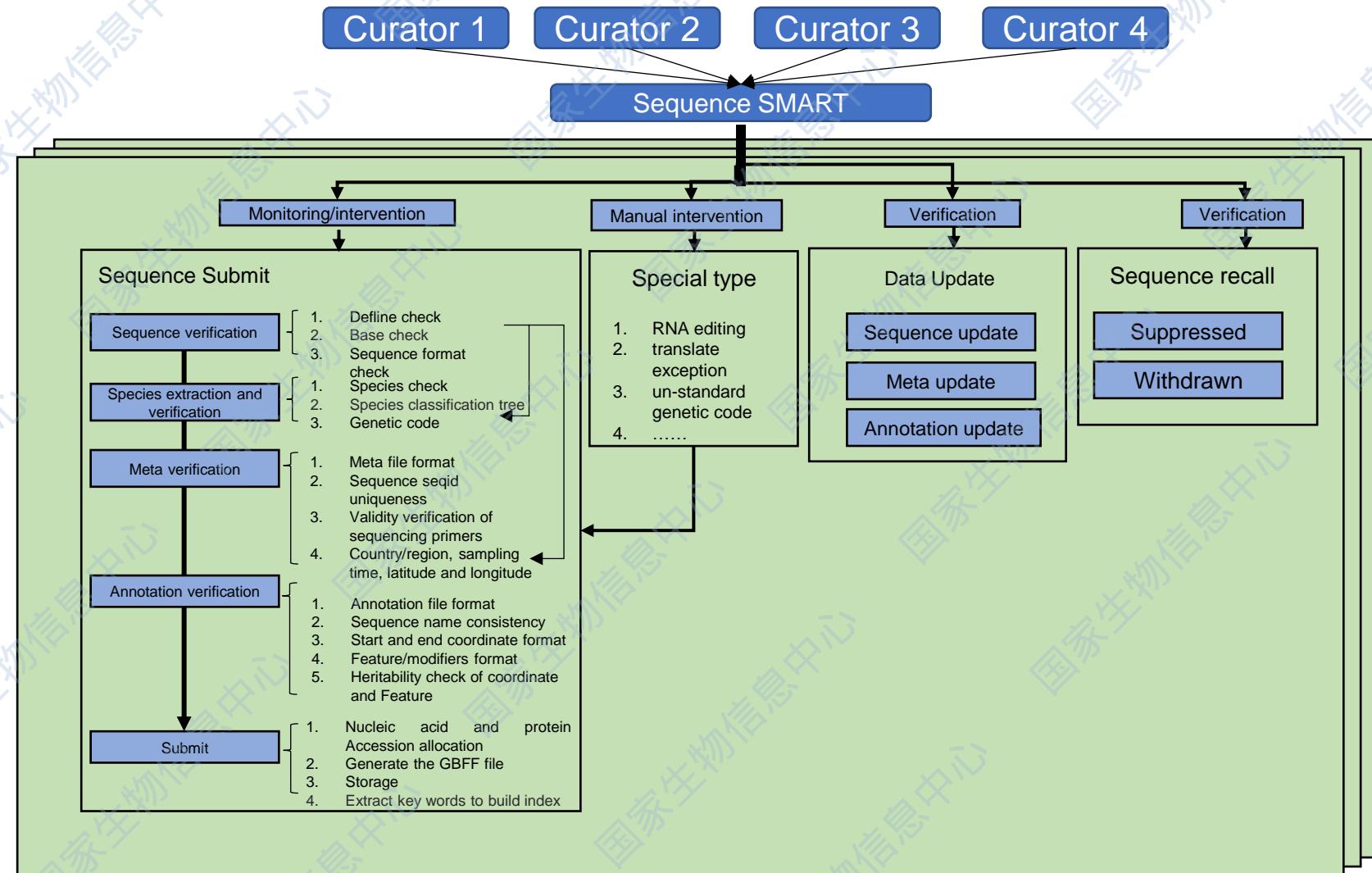
The data reported in this paper have been deposited in the GenBase [1] in National Genomics Data Center [2], Beijing Institute of Genomics, Chinese Academy of Sciences/China National Center for Bioinformation, under accession number C_AA000000 that is publicly accessible at <https://ngdc.cncb.ac.cn/genbase>.

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]
CTTGAGCTAAGTTCCCCAAATCTTAGTGAACCCAGGGTAGCATTGTAAGTGT
GTGTTTGTTCCTTGATGTCTATTACTACTAGGACCATTATATTTCACATACTT
AGCAATAAATAACTCTATAGTTAACTGGTGGTAACGGGTGTATTTTCTTTACT
TTTCCTTCCCCACTTGCTCTGTAGCAACGCTTCTTACCTAACGCTAAAATCCC
TGAAGCTGAGAAAAACTTGGGTGTGCTAGCAAGAGGCTAACTCTACTAGGAC
```

○ 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

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

* 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

提交者 出版信息 测序技术 序列 集合或批次 分类 元信息 特征 结果预览

特征

提交 #0000230 帮助

请注意

- 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
- 请参考此链接作为TBL格式序列注释的说明
- 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

result_err.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。完整的错误信息请查看'result_err.txt'。下面列出了部分错误信息。

ERROR: In line 1, Feature Seq1le is not included in the fasta file

保存&下一步

常见错误类型：Feature table错误示例

The screenshot shows an Excel spreadsheet titled "result_feature - Excel". The ribbon is visible at the top with tabs like File, Start, Insert, etc. The main area displays a table with columns A through G. Rows 1 and 2 contain explanatory text about Feature tables. Row 3 is empty. Row 4 contains Chinese text explaining the table structure. Row 5 starts the data with two rows for "ftchy1" and "ftchy2". The first row has a red box around the "sequence_id" column header. The second row has a red box around the entire row. To the right of the table, there is a sidebar titled "Qualifier hints (提示)" with sections for "feature" and "qualifier", examples, and regulatory class definitions.

	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 feature (column E) is the structure name of this type of sequence, qualifiers (column F) is the attribute of this structure, most attributes are optional for the relevant structure, and some attributes are necessary. After selecting feature and qualifiers, qualifiers hints (column G) will help you prompt the format of the qualifier value.								Qualifier hints (提示)	
2									feature ↗	qualifier ↗
3									regulatory	regulatory_class(Mandatory)
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)								Example	
5									/regulatory_class='promoter'	
6									/regulatory_class='enhancer'	
7	Locations (坐标)				Attributes (属性)					
8	sequence_id	start	end	completeness	feature	qualifier	qualifier value			
9	ftchy1	36695	7518	complete	gene	gene	ft	<= The block of CDS		
10	ftchy2	36695	7518	complete	CDS	product	aaa	<= The block of CDS		
11										
12										
13										
14										

提示此处填写有问题

常见错误类型：Feature table正确示例

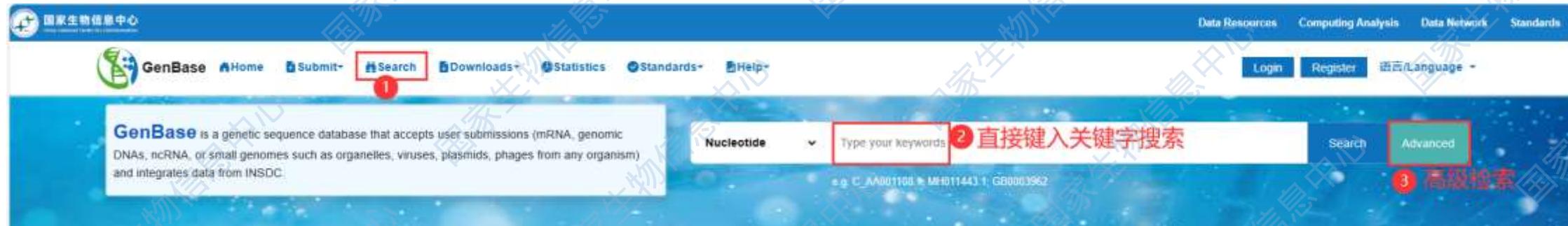
	A	B	C	D	E	F	G	H	I	
1					Feature table xlsx can be regarded as a nucleic acid annotation format with meta information. The user needs to fill in the corresponding annotations in the corresponding columns A ~ G. Where feature (column E) is the structure name of this type of sequence, qualifiers (column 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										
7										
8										
	Locations (坐标)				Attributes (属性)			Qualifier hints (提示)		
	sequence_id	start	end	completeness	feature	qualifier	qualifier value	feature	qualifier	
9	ftchy	36695	7518	complete	gene	gene	ftChyE	regulatory	regulatory_class(Mandatory)	
10	ftchy	36695	7518	complete	CDS	product	nonribosomal peptide synthase	Example		
11								/regulatory_class='promoter' /regulatory_class='enhancer' /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



Advanced Search

Add terms to the query box:

Nucleotide
Protein

Query box

View your search query here

Search Clear

All Fields

- All Fields
- Accession
- Author
- BioProject
- BioSample
- Breed
- Cultivar
- Division
- Data Source
- Grant

Enter a search term

ADD

Add with AND
Add with OR
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

物种

- Viruses (71,553)
- Bacteria (3,675)
- Plants (2,858)
- Animals (1,807)
- Fungi (949)
- Unknown (85)
- Protists (19)

数据来源

- GenBase (80,976)

数据类型

- Nucleotide (80,976)

分子类型

- ss-RNA (63,659)
- DNA (11,203)

Leptodectes potaninii chloroplast clone Tenggeli; complete genome

1. 149,059 bp DNA
Accession: C_AA097739.1
GBFF FASTA

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

或选择所有条目

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

Send to: ▾

Format

Summary

First 1 2 3 4 5

3. 选择格式

Format

Summary

Create File

GBFF

FASTA

Accession List

4. 下载

GBFF

Separated

Merged

Create File

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 

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/>

The screenshot shows the CNCB-NGDC Public Download Server interface. The current directory is 'Home / genbase'. On the left, there is a sidebar with 'Name' and items: 'GenBank', 'RefSeq', and 'daily' (which is highlighted with a red box). On the right, there is a main panel with 'Name' and 'Update Date'. The 'Name' section lists '2022' and '2023' (both highlighted with red boxes), followed by 'process', '20240101.flat.gz', '20240101.fsa.gz' (highlighted with a red box), '20240101.fsa_nt.gz' (highlighted with a red box), '20240101.gnp.gz', '20240102.flat.gz', '20240102.fsa.gz', and '20240102.fsa_nt.gz'. The 'Update Date' section shows three entries: '2024-09-03 10:44:00', '2024-09-03 09:45:00', and '2024-10-19 02:40:00'.

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

GenBase质量控制系统的错误类型和修正方案

在质控过程中，GenBase会输出报错，需要对其进行逐一修改，并重新提交文件到GenBase。下面展示了可能遇到的错误类型和修正方案。[下载PDF版本](#)

SEQ_FEATURE内部错误：EQ_INST_Syntax Problem
错误：序列的蛋白氨基酸与之结合分析不匹配的子句。
建议：
1. 去掉快照分析子句，或者保存为修改过的序列。
2. 后面增加蛋白预测子句，叶预测、结果预测和折叠预测使用不同的进阶密码子表，使用快照的密码子表可能会导致比不应该的终止位的子句出现。如果想要更改之请删除该子句。
3. 后面存在CDS预测的子句，基因预测不正确的CDS生成，不正确的CDS上始符和终止子句，即有多段CDS时需要分段标注出每段的CDS位置。
4. 后面存在codon_Meta变化的子句，此级别的子句请将它放在CDS标记之后或直接放在CDS的最末尾，如果前面有4段的CDS的话，最好用一个parallel的子句。如果前面只有2段的CDS，且后头需要添加一个center_start属性，分别表示将这两段移动一个或两个碱基。对于OFF文件，codon_Meta将其写为1或2表示将这两段移动一个或两个碱基；对于TDF文件，

<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：

```
#Hs001_BsqI LOC03_BsqI C_AA003515.1 29019 1-29019 update_meta  
#Prx1_1D_CDS_BsqI C_AA001556.1 21291 266-13468,13468-21595 update_meta  
CDS_BsqI C_AA001557.1 3822 2146-25197 update_meta  
CDS_BsqI C_AA001558.1 3822 2146-25197 update_meta  
CDS_BsqI C_AA001559.1 658 2162-21647 update_meta  
CDS_BsqI C_AA001560.1 658 2162-21647 update_meta  
CDS_BsqI C_AA001561.1 106 27272-27317 update_meta  
CDS_BsqI C_AA001562.1 106 27272-27317 update_meta  
CDS_BsqI C_AA001563.1 132 27796-27808 update_meta  
CDS_BsqI C_AA001564.1 346 27898-20258 update_meta  
CDS_BsqI C_AA001565.1 3265 -28074-28533 update_meta  
Kitherto_B_gene_BsqI_BsqI 29019 1-29019 update_meta
```

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

② 帮助和支持

提交指南

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

Email: genbase@big.ac.cn

QQ群: 629388189

座机: 86-10-84097298

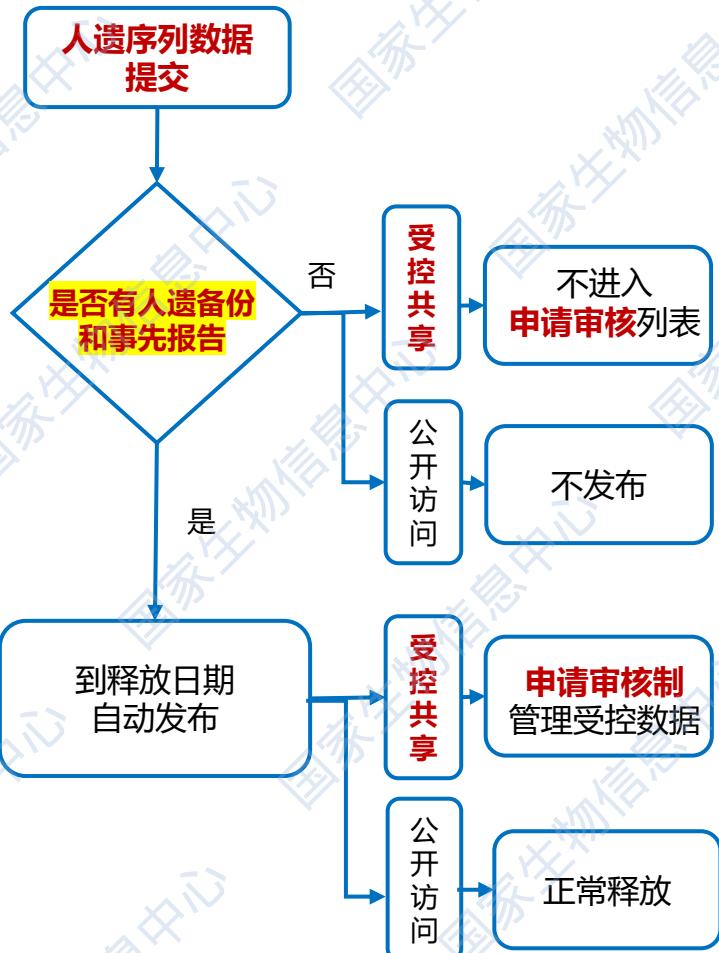
工作时间: 工作日9:00 - 17:00

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

GenBase人类遗传相关序列共享

The screenshot shows the GenBase homepage with a navigation bar including '国家生物信息中心' logo, 'GenBase', '首页', '数据提交', '检索', '下载', '统计', '标准', '帮助文档'. Below is a section titled '提交类型' (Submission Type) with buttons for '常规序列提交', 'SARS-CoV-2快速提交', and '受控序列提交' (highlighted with a red box). The main content area is titled '受控序列提交入口' (Controlled Sequence Submission Interface) with a note: '受控序列提交用于人类遗传资源管理, 只有人的数据才能使用该提交模块进行受控数据提交'. It lists preparation steps and submission requirements.

GenBase采用申请审核制管理受控访问数据



序列提交完成后:

- 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
- 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;
- 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)

目录

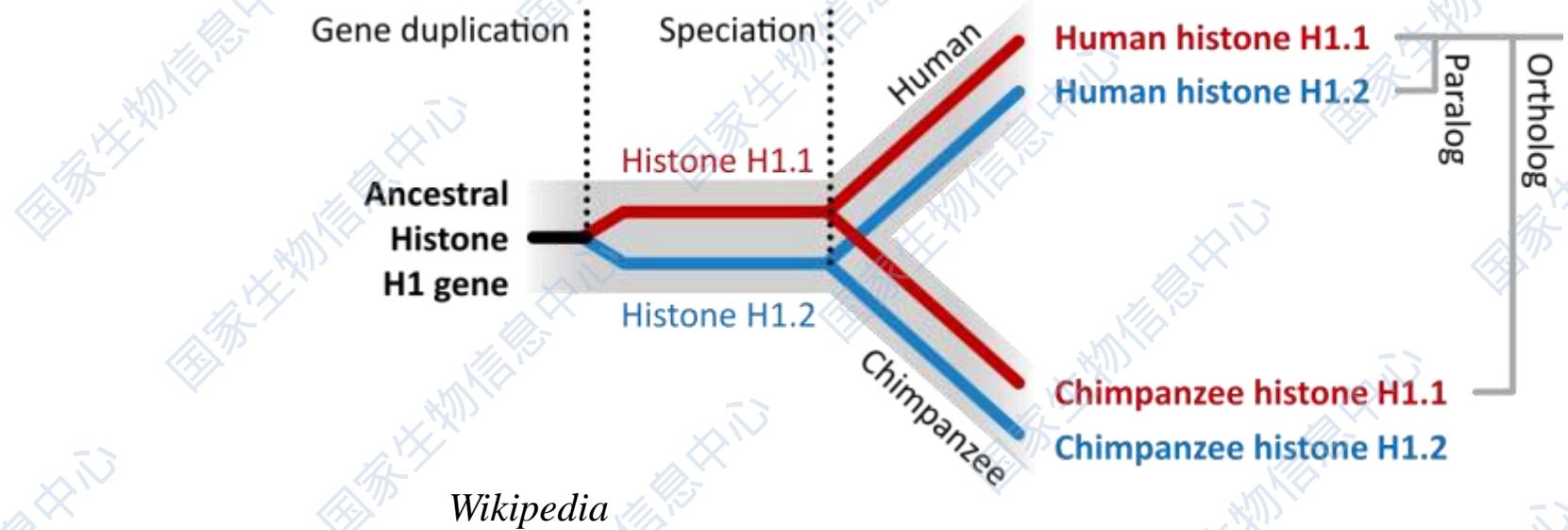
CONTENTS

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



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

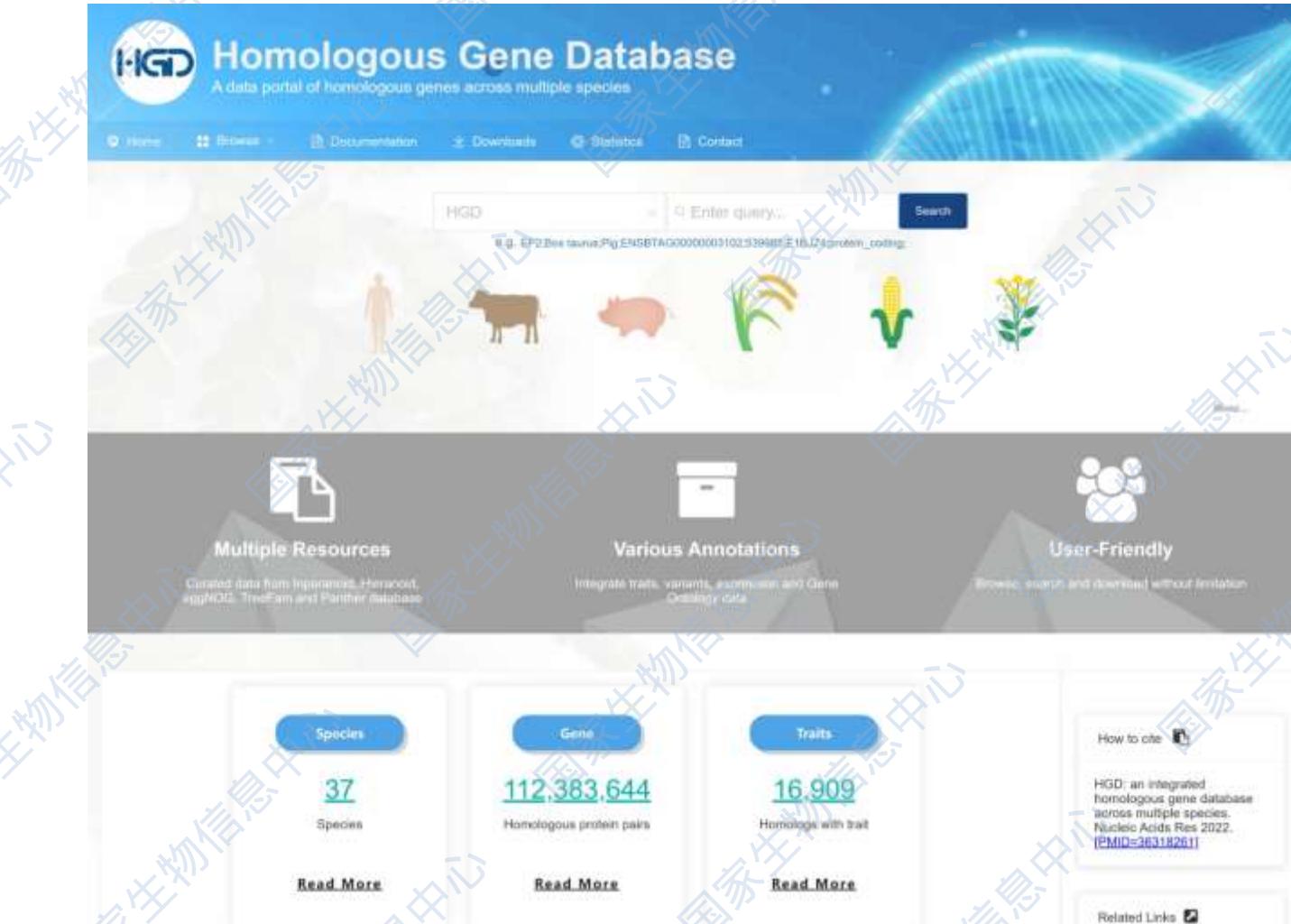
存在的问题：推断方法不同，导致推断结果有差异；

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

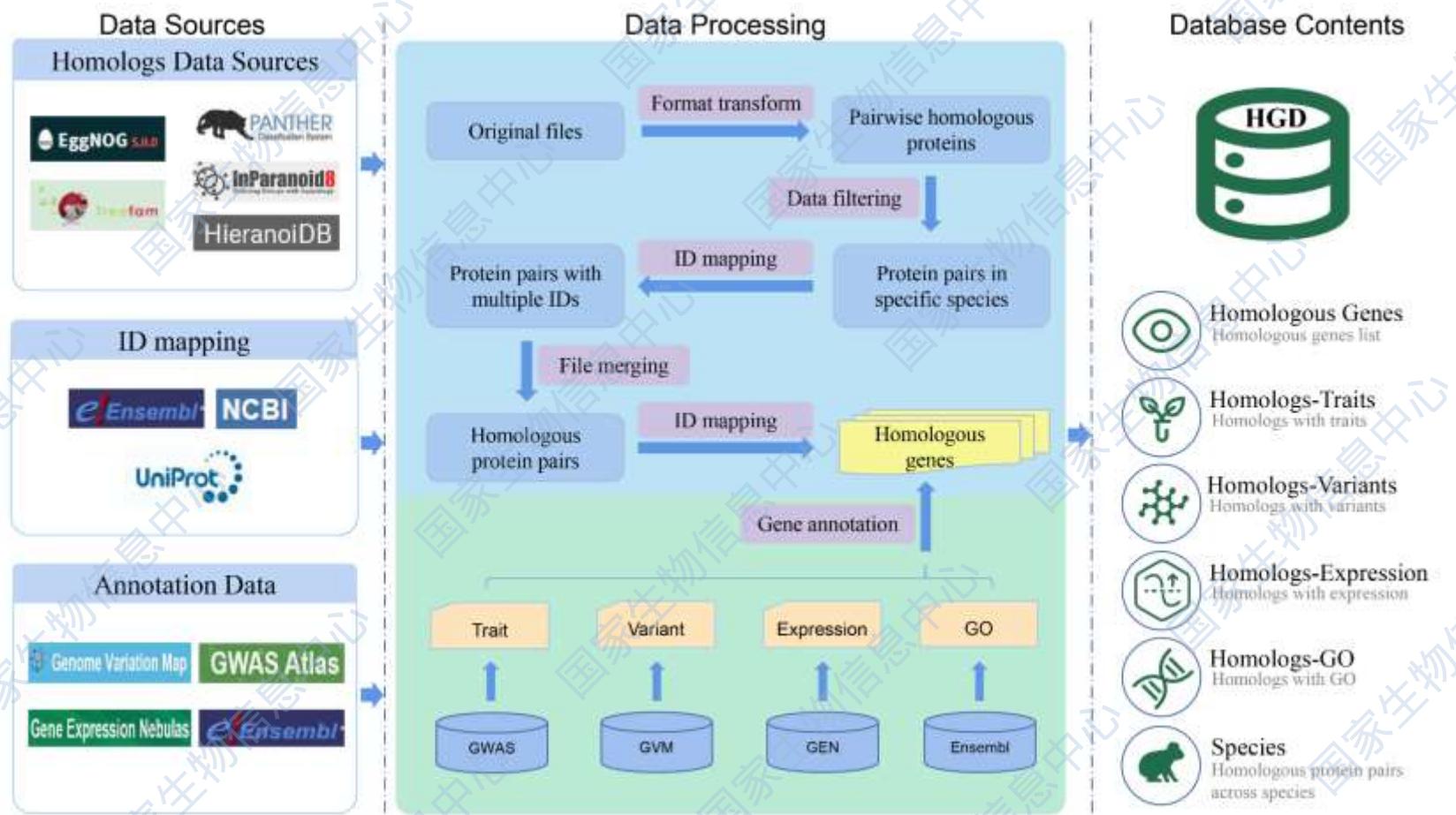


<https://ngdc.cncb.ac.cn/hgd>

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
Filter by Species
Please select
Filter by condition AND
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

Total 1 5/page < 1 > Go to 1

EP2 X

Check All
 EP2 (*Oryza sativa*)

Regulation of panicle erectness, panicle length and grain size

Symbol: EP2

Uniprot: Q0D4N6

Ensembl Protein: Os070616000-01

Synonym:

BioType: protein_coding

Homologous Gene(s):

F6HXQ3_Vitv09g00052_P001 (<i>Vitis vinifera</i>)	A0A1D6IJL5_Zm00001d022143_P006 (<i>Zea mays</i>)	Zm00001d006864 (<i>Zea mays</i>)
Zm00001d022143 (<i>Zea mays</i>)	A0A1D5V1E8_TraesCS2D02G176200 (<i>Triticum aestivum</i>)	A0A3B6AUT7_TraesWEE_scaffold_108679_01G00010C (<i>Triticum aestivum</i>)
A0A3B6C2G9_TraesCAD_scaffold_038815_01G000100	SORBI_3002G374400 (<i>Sorghum bicolor</i>)	SORBI_3002G377000 (<i>Sorghum bicolor</i>)
(<i>Triticum aestivum</i>)	4334871 (<i>Oryza sativa</i>)	Os06g0196500 (<i>Oryza sativa</i>)
4332729 (<i>Oryza sativa</i>)	Q7XI09_Os07I0445600-01 (<i>Oryza sativa</i>)	102600185 (<i>Solanum tuberosum</i>)
OsSTA81 (<i>Oryza sativa</i>)	M1CTF1_PGSC0003DMT400074261 (<i>Solanum tuberosum</i>)	PGSC0003DMG400002567 (<i>Solanum tuberosum</i>)
M1C3M0_PGSC0003DMT400058989 (<i>Solanum tuberosum</i>)	A0A3Q7FXL7 (<i>Solanum lycopersicum</i>)	A0A3Q7JBB5 (<i>Solanum lycopersicum</i>)
	A0A2G2XZ05_PHT62571 (<i>Capsicum annuum</i>)	A0A2G2ZVS5_PHT86078 (<i>Capsicum annuum</i>)
	A0A2C9UYM4_OAY36365 (<i>Manihot esculenta</i>)	100812963 (<i>Glycine max</i>)
	100817241 (<i>Glycine max</i>)	BnaC05g13520D (<i>Brassica napus</i>)
	BnaC05g52130D (<i>Brassica napus</i>)	AT1G17360 (<i>Arabidopsis thaliana</i>)
	AT1G72410 (<i>Arabidopsis thaliana</i>)	F4IDB4_AT1G72410 (<i>Arabidopsis thaliana</i>)
	F4JFS3_AT3G14172 (<i>Arabidopsis thaliana</i>)	A0A2K2B2W6_PNT44118 (<i>Populus trichocarpa</i>)
	A0A3N7EA94 (<i>Populus trichocarpa</i>)	LOC107897450 (<i>Gossypium hirsutum</i>)
	LOC107906061 (<i>Gossypium hirsutum</i>)	LOC107958375 (<i>Gossypium hirsutum</i>)

Trait 1 Variation 6 Expression 15

不同物种中的同源基因

同源基因注释情况

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

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

设置查询条件



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>

Selected Gene
Gene ID: ENSBTAG00000012789
Species Name: Cattle
Uniprot ID: A0A3Q116PF0
Classification: animal
Ensembl Protein ID: ENSBTAP00000067781
Trait Ontology term: growth and meat production trait

Species	Term ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Data Source Count	Data Source
Cattle	PRKD1	ENSBTAG00000012789	PRKD1	A0A3Q116PF0	ENSBTAP00000067781	1	●
Cattle	PRKD1	ENSBTAG00000012789	PRKD1	A0A3Q116PF0	ENSBTAP00000067781	1	●
Cattle	PRKD1	ENSBTNG00000012789	PRKD1	A0A3Q116PF0	ENSBTAP00000067781	1	●

同源来源统计



具体同源来源

GWAS信息

Trait Name	Gene ID	Var ID	Sub Trait Name	Species	Pubmed ID	P-value
growth and meat production trait	ENSBTAG00000012789	EHA4263461	werner-bratzler shear force	Cattle	29163638	0.0000121
growth and meat production trait	ENSBTAG00000012789	EHA4263443	werner-bratzler shear force	Cattle	29163638	0.0000111

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



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

Selected Gene		Uniprot ID: A0A080NPY5		Ensembl Protein ID: ENSPCA00000024153	
Gene ID: ENSPCA00000024430 Species Name: Cat		Classification: avian		Consequence type: transcript:UTR_variant	
Homologous Gene Detail Information					
Species	Taxon ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID
Cat	9603	ENSPCA00000000090-0	TNFSF14	M36208	ENSPCAP00000024462
Cat	9603	ENSPCA00000000449-0	LTK	M36205	ENSPCAP00000024152
Cat	9603	ENSPCA00000000174	TNFSF16	M36200	ENSPCAP00000024116
Cat	9603	ENSPCA00000000032	TNFSF11	M14472	ENSPCAP00000024821
Cat	9603	ENSPCA00000000499	TNF	A0A080NPY5	ENSPCAP00000024153

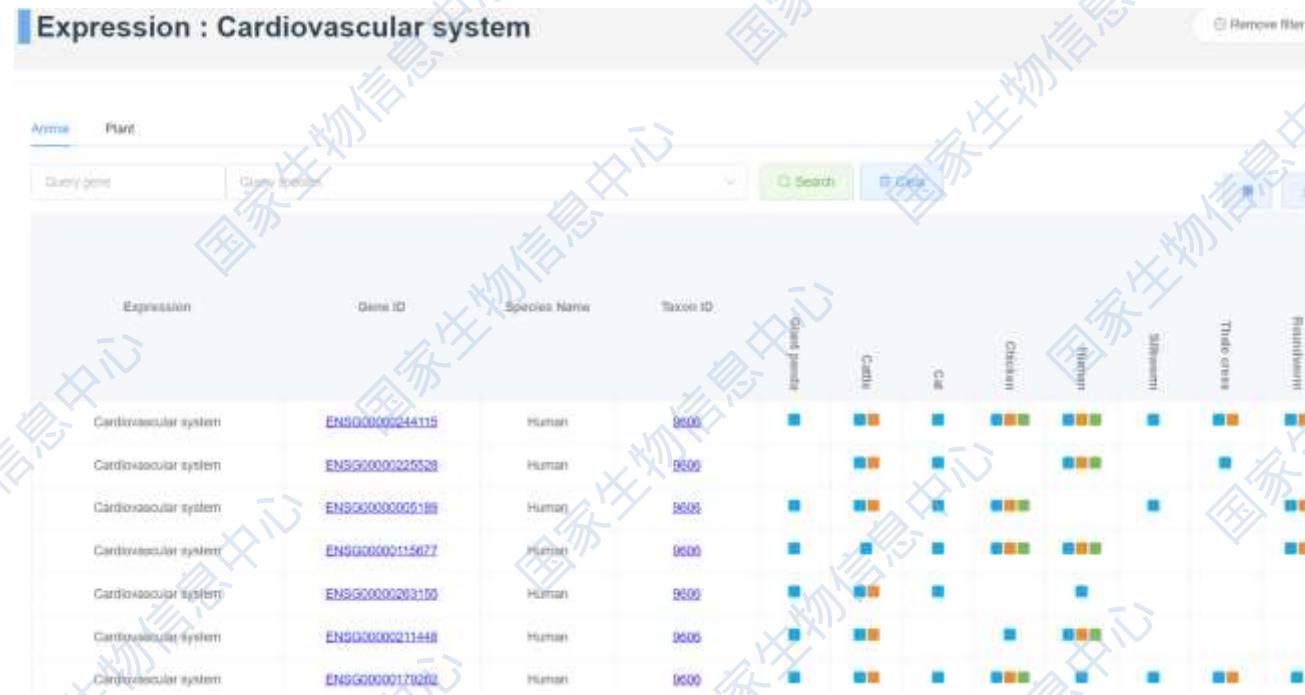
变异注释信息

Variants Detail Information						
Gene ID	Var ID	Position	Alt	MAF	Class	Consequence Type/Effect
ENSPCA000000044						
ENSPCA000000038	kat12959728	B2:325733T	TAC	0.01054	SNP	S_prime_UTR_variant(MODIFIER: upstream_of_ne_variant)MODIFIER;
04499						
ENSPCA000000044						
ENSPCA000000000	kat12959729	B2:325743T	GAA	0.01054	SNP	S_prime_UTR_variant(MODIFIER: upstream_of_ne_variant)MODIFIER;
04498						
ENSPCA000000044						
ENSPCA000000000	kat12959730	B2:325743G	A/G	0.010788	SNP	S_prime_UTR_variant(MODIFIER: upstream_of_ne_variant)MODIFIER;
04498						
ENSPCA000000044						
ENSPCA000000000	kat12959731	B2:325743G	C/T	0.010568	SNP	S_prime_UTR_variant(MODIFIER: upstream_of_ne_variant)MODIFIER;
04498						

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

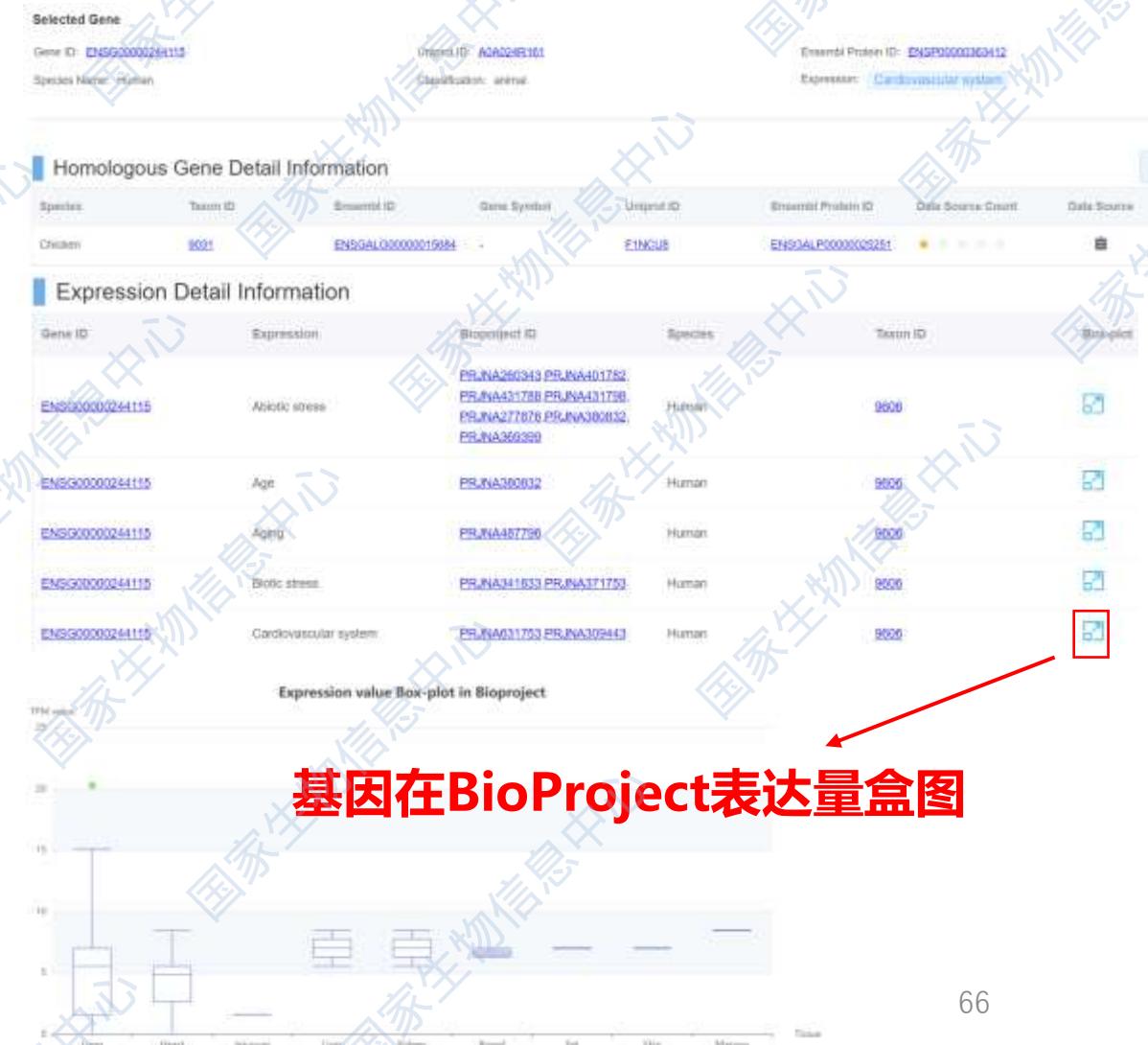


蓝色: 同源基因

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>



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>

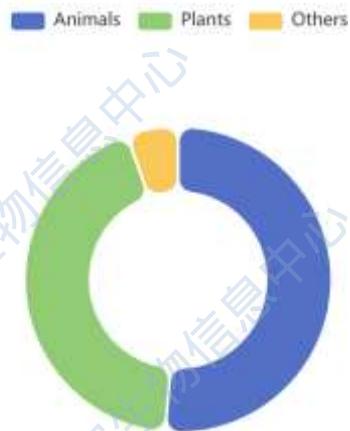
Selected Gene						
Species	Taxon ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Data Source Count
Carrie	9613	ENSGTA00000021251				
Carrie	9613	ENSGTA00000027318				
Carrie	9613	ENSGTA00000029241				
Carrie	9613	ENSGTA00000029611				
Carrie	9613	ENSGTAG00000021483				
Carrie	9613	ENSGTA00000033852				
Homologous Gene Detail Information						
Species	Taxon ID	Ensembl ID	Gene Symbol	Uniprot ID	Ensembl Protein ID	Data Source
Chicken	9031	ENSGAL00000035189	RPL46	A0A106P0W4	ENSGALP00000046293	***
Chicken	9031	ENSGAL00000031327	RPLX1	E1C024	ENSGALP00000036419	***
Chicken	9031	ENSGAL00000037815	ANKK1	E1C023	ENSGALP00000012660	**
Chicken	9031	ENSGAL00000032766	MYL1	E1C024	ENSGALP00000004260	*
GO Detail Information						
Gene ID	GO term	GO ID	SIS-GO ID	Species	Taxon ID	
ENSGAL00000037815	ATP binding	GO:0000024	GO:0000024	Chicken	9031	
ENSGAL00000037815	protein phosphorylation	GO:0006468	GO:0006468	Chicken	9031	
ENSGAL00000037815	intracellular membrane-bounded organelle	GO:0043231	GO:0000039	Chicken	9031	
ENSGAL00000037815	nucleus	GO:0006634	GO:0009234	Chicken	9031	
ENSGAL00000037815	cytoskeleton	GO:0006515	GO:0006515	Chicken	9031	
ENSGAL00000037815	protein kinase activity	GO:0006773	GO:0006772	Chicken	9031	
ENSGAL00000037815	transferrin activity	GO:0016746	GO:0004872	Chicken	9031	

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
<i>Gallus gallus</i>	Chicken	9031	214	13327	15497	13034
<i>Ailuropoda melanoleuca</i>	Giant panda	9646	-	12580	-	13252
<i>Bos taurus</i>	Cattle	9913	238	17738	20721	17672
<i>Canis familiaris</i>	Dog	9615	-	-	1077	973

物种统计分布



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



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: Os07g02016000	Ensembl Gene ID: 434394	RefSeq ID: XP_013647269
Uniprot ID: Q0D4H6	Ensembl Protein ID: Os07g02016000-01	Gene Symbol: -
Gene Symbol: EP2	Gene Type: protein_coding	Latin Name: Oryza sativa
Species Common Name: Rice	Taxon ID: 4530	Chromosome: 7
Gene Start: 25381698	Gene End: 25389932	Gene: Regulator of panicle architecture, panicle length Description: and grain size.

Homologous Gene

Filter by species						
Species	Taxon ID	Gene Symbol	Ensembl ID	Ensembl ID	Uniprot ID	Ensembl Protein ID
Broad sub-esp	4530	-	TrsacCS2032G175200	-	AUA1DSV1E3	Trns-CS2032G175200
Broad sub-esp	4530	-	TransWEE_scattered_10	-	AUA3855AU7	TransWEE_scattered_10

Gene Ontology

同源基因比较

物种筛选过滤

同源基因比较面板



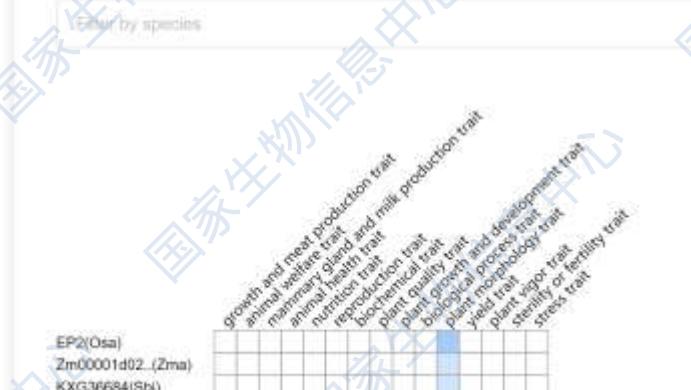
变异

Variants



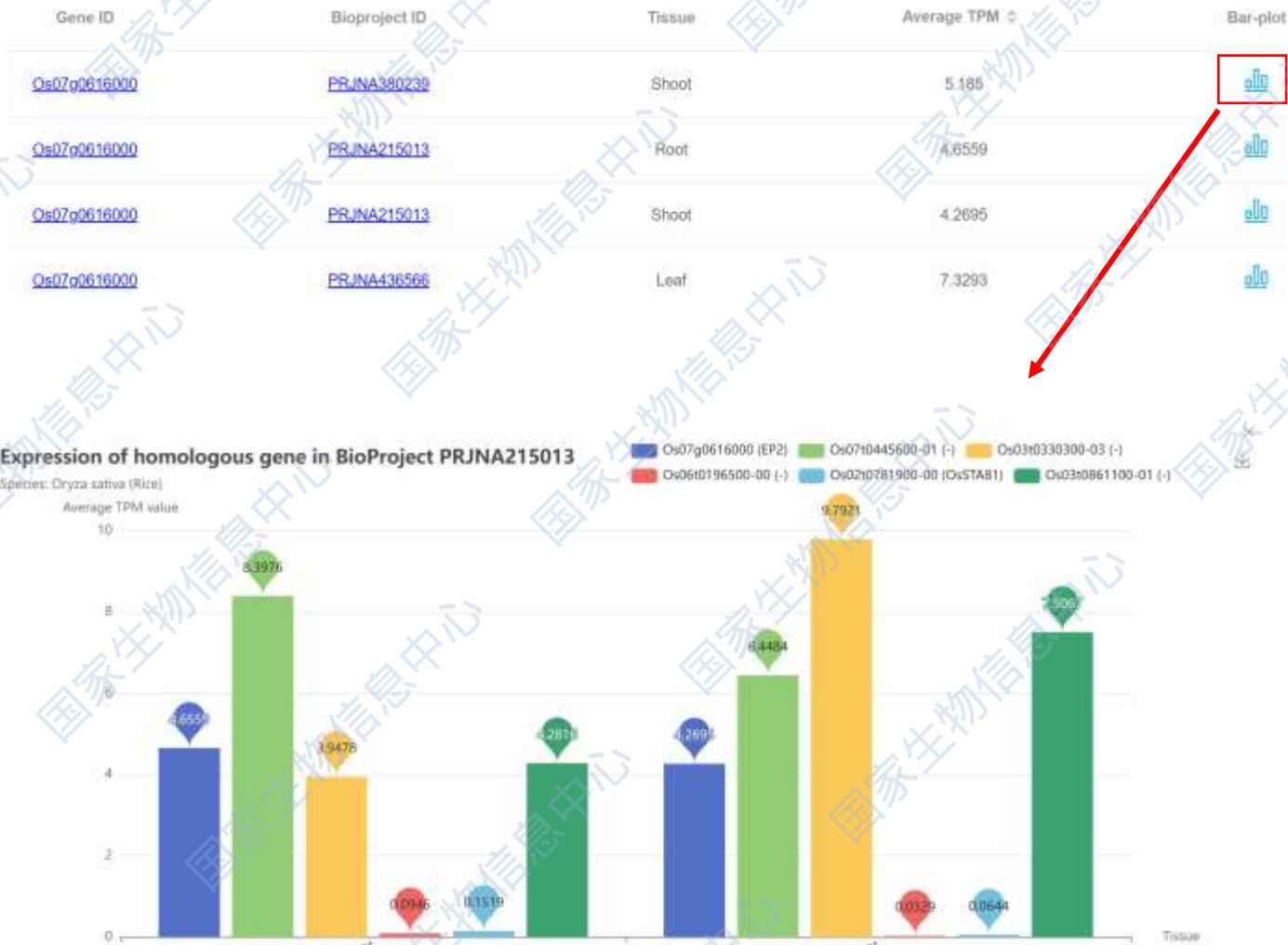
性状

Traits



表达

Expression



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_Chicken_Homolog_protein.txt.gz
Cat_Dog_Homolog_protein.txt.gz
Cat_Rhesus monkey_Homolog_protein.txt.gz
Cat_Potato_Homolog_protein.txt.gz
Cattle_Dog_Homolog_protein.txt.gz
Cattle_Tropical clawed frog_Homolog_protein.txt.gz
Cattle_Rice_Homolog_protein.txt.gz

Cattle_Soybean_Homolog_protein.txt.gz
Cat_Pig_Homolog_protein.txt.gz
Cattle_Rat_Homolog_protein.txt.gz
Cat_Soybean_Homolog_protein.txt.gz
Cattle_Rhesus monkey_Homolog_protein.txt.gz
Cat_Human_Homolog_protein.txt.gz
Cattle_Thale cress_Homolog_protein.txt.gz
Cat_Mouse_Homolog_protein.txt.gz

Cat_Tropical clawed frog_Homolog_protein.txt.gz
Cat_Brewer's yeast_Homolog_protein.txt.gz
Cattle_Brewer's yeast_Homolog_protein.txt.gz
Cat_Giant panda_Homolog_protein.txt.gz
Cattle_E. coli_Homolog_protein.txt.gz
Cattle_Turkey_Homolog_protein.txt.gz
Cat_Grape_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

