



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

# CNCB Genomic Data Submission and Sharing System

**Yanqing Wang**

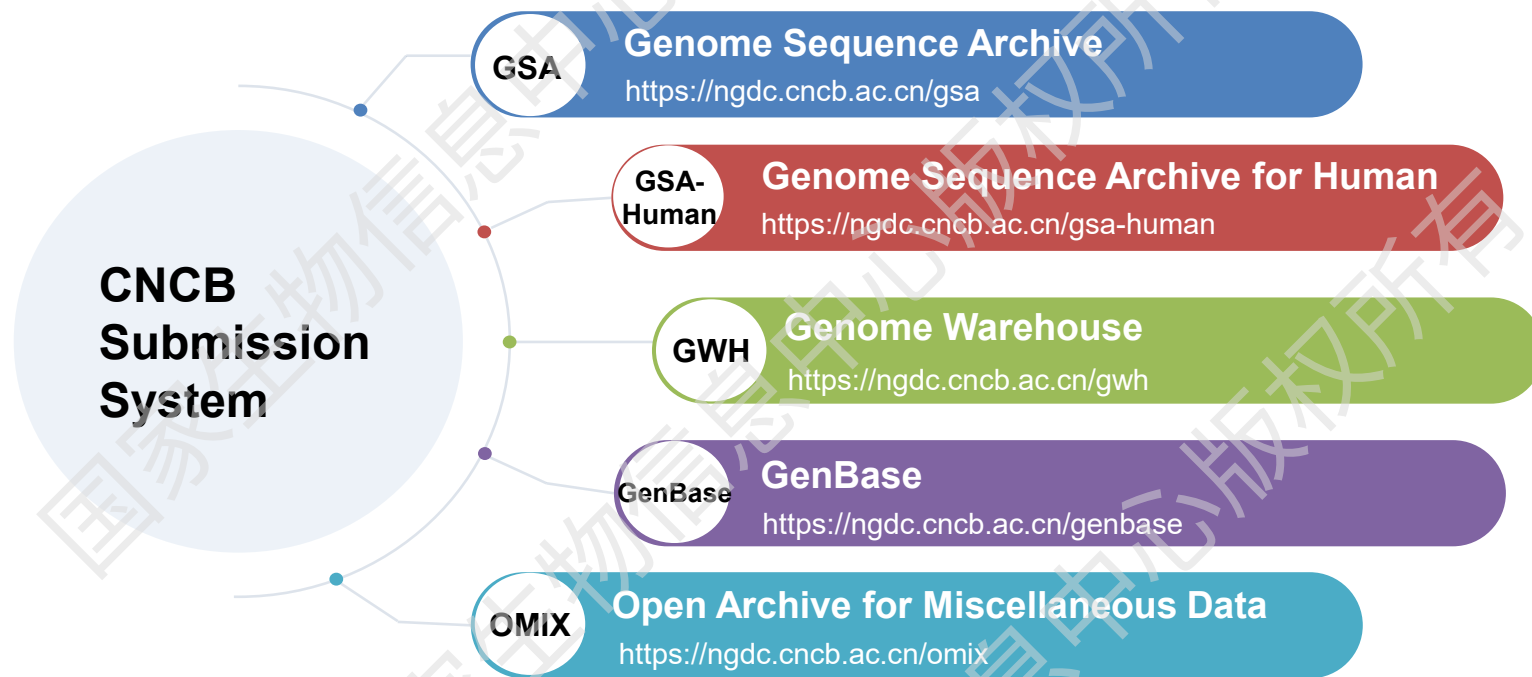
National Genomics Data Center  
China National Center for Bioinformation

October 13, 2025

# Contents

- 1 CNCB Data Submission System Overview**
- 2 GSA Data Submission & Sharing**
- 3 GSA-Human Data Submission & Sharing**
- 4 GWH Data Submission & Sharing**
- 5 GenBase Data Submission & Sharing**
- 6 OMIX Data Submission & Sharing**

# Overview of the CNCB Data Submission System



A set of omics data repositories, providing data submission, storage, and sharing services

# GSA: Genome Sequence Archive



Genomics Proteomics Bioinformatics (2017, 2021, 2025)

GSA is a public repository for the archiving, retrieving, and sharing of non-human raw sequence reads.

equivalent

SRA  
Sequence  
Read Archive  
@NCBI

DRA  
DDBJ Read  
Archive  
@DDBJ

ENA  
European  
Nucleotide  
Archive  
@EBI

International Nucleotide Sequence Database  
Collaboration (INSDC)



GSA was selected as one of the **Global Core Biodata Resources (GCBR)** by the Global Biodata Coalition (GBC)

# GSA-Human: Genome Sequence Archive for Human



Genomics Proteomics Bioinformatics (2021, 2025) Yi Chuan (2021)

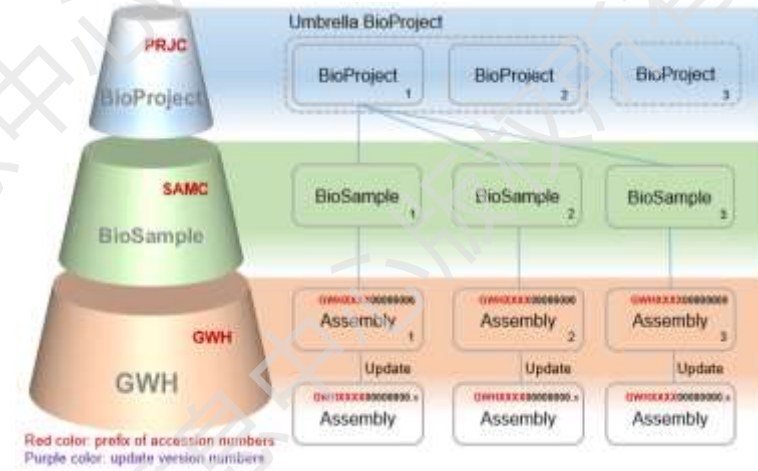
- **Focuses on Human related omics sequence data**
  - Authorized as the national human genetic resource data management platform
  - **Features two data access mechanisms**
    - Controlled access
    - Open access
  - The “Request and Provel” model for the controlled access data
    - Data Access Committee, DAC
    - Review the data application, grant data access permissions to data requesters

# GWH: Genome Warehouse

GWH is a public repository housing genome assembly data for a wide range of species, delivering a series of web services for genome data submission, storage, release and sharing.



Homepage of GWH, <https://ngdc.cncb.ac.cn/gwh/>

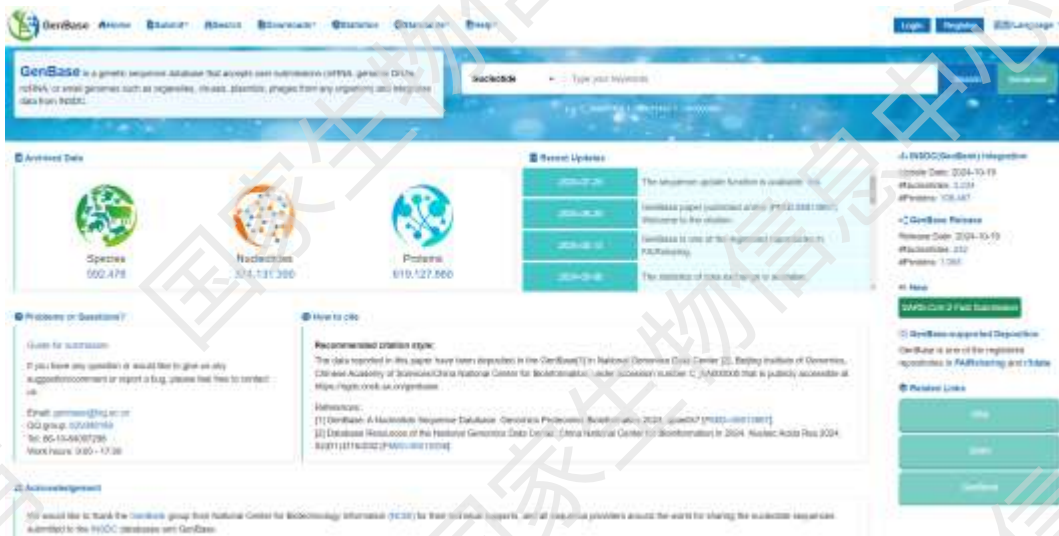


Structured organization and management of genome assembly data like INSDC



# GenBase

**GenBase is an open-access data repository designed for nucleotide sequence archiving, searching, and sharing. It supports the submission of mRNA, genomic DNAs, ncRNA, or small genomes**



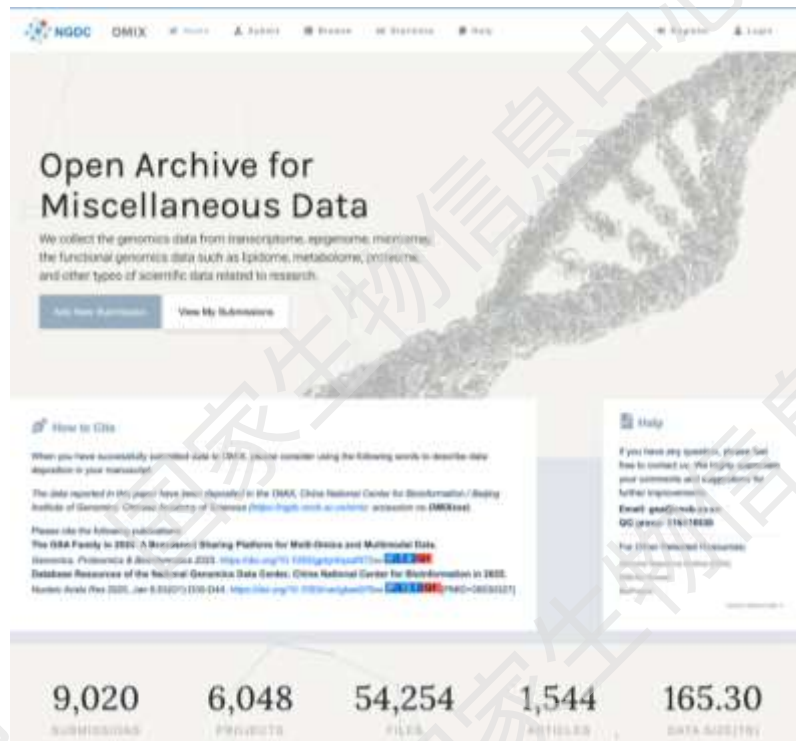
### # Submitted Nucleotide (Cumulative)



# Genomics Proteomics Bioinformatics (2024)

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

# OMIX: Open Archive for Miscellaneous Data



Genomics Proteomics Bioinformatics (2021, 2025)

## Generalist repository for multimodal biological data

- Supports **10** main categories, **32** subcategories
  - Supports **2** file upload methods:
    - HTTP online upload
    - FTP pre-upload
  - Supports **2** data access policies:
    - Open access
    - Controlled access (for human data only)
- Certified by **FAIRSharing** and **re3data.org**
  - Acknowledged by over **420** Journals, supporting **1544** articles



# Data Flows

## Data submissions

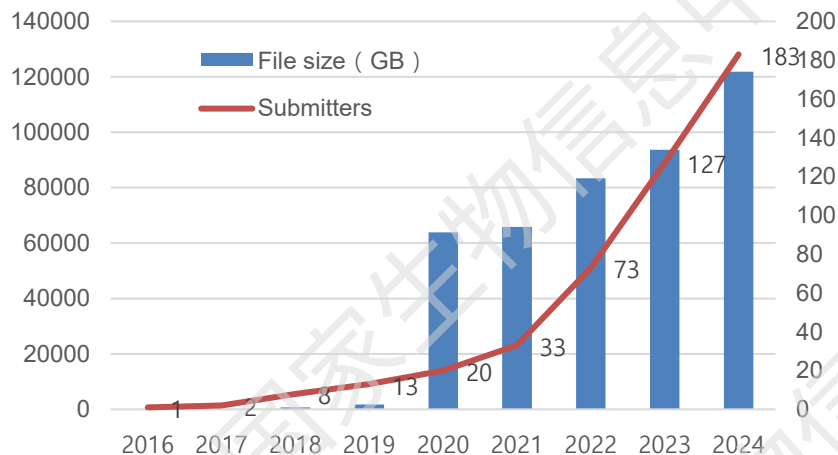
- **86.6 PB** Raw reads
- **98,108** Genome assemblies
- **1,264,944** Sequences
- **18,000+** Submitters
- **780+** Journals
- **5500+** Supported articles

## Visits & downloads



> **25 million** Page views    > **2.2 million** Unique IPs  
> **200** Countries/regions    > **19PB** Data downloads

# International Submissions



| Country/Region     | Submitters | Datasets |
|--------------------|------------|----------|
| United States      | 48         | 64       |
| Germany            | 14         | 16       |
| Republic of Korea  | 10         | 24       |
| Singapore          | 9          | 10       |
| France             | 8          | 16       |
| Pakistan           | 8          | 54       |
| Australia          | 7          | 37       |
| Canada             | 7          | 9        |
| Italy              | 7          | 7        |
| Japan              | 7          | 13       |
| Russian Federation | 7          | 85       |
| United Kingdom     | 6          | 8        |
| India              | 4          | 4        |
| Malaysia           | 4          | 4        |
| Denmark            | 3          | 3        |
| Korea, Republic of | 3          | 3        |
| Sweden             | 3          | 3        |
| Thailand           | 3          | 5        |
| Brazil             | 2          | 21       |
| Egypt              | 2          | 2        |

| Country/Region             | Submitters | Datasets |
|----------------------------|------------|----------|
| Israel                     | 2          | 2        |
| Netherlands                | 1          | 1        |
| Switzerland                | 1          | 2        |
| Argentina                  | 1          | 1        |
| Austria                    | 1          | 1        |
| Belarus                    | 1          | 1        |
| Belgium                    | 1          | 1        |
| Chile                      | 1          | 1        |
| Holland                    | 1          | 1        |
| Indonesia                  | 1          | 1        |
| Iran (Islamic Republic of) | 1          | 1        |
| Iraq                       | 1          | 1        |
| Ireland                    | 1          | 1        |
| Kazakhstan                 | 1          | 1        |
| Monaco                     | 1          | 1        |
| Philippines                | 1          | 1        |
| Qatar                      | 1          | 1        |
| Spain                      | 1          | 1        |
| Sri Lanka                  | 1          | 1        |
| Zimbabwe                   | 1          | 1        |

**183** international submitters from **40** countries/regions, **410** datasets submitted

As of Dec 2024

Research data policy

Mandated data types

For the following data types submission to a community-accepted, public repository is mandatory. Personnel (authors, ORIDs) and institutions (employers) assigned to the data by the repository must be appropriately cited and referenced in the published article.

Data types

|                                                          |                                                                                                                                                                                             |
|----------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| DNA sequence data*                                       | Any IPDCC member repository / Genome Sequence Archive (GSA)                                                                                                                                 |
| RNA sequence data*                                       |                                                                                                                                                                                             |
| Genomic assembly data*                                   |                                                                                                                                                                                             |
| Protein sequence data                                    |                                                                                                                                                                                             |
| Proteomics data                                          |                                                                                                                                                                                             |
| Genetic variation data                                   |                                                                                                                                                                                             |
| Sensitive human genomic data requiring controlled access | <ul style="list-style-type: none"> <li>The European Genomic Archive (EGA)</li> <li>European Genomic Phenotype Archive (EGPA)</li> <li>Genome Sequence Archive for Humans (GSA-H)</li> </ul> |

RELATED STANDARDS (3)

RELATED DATABASES (92)

Genome Sequence Archive

Elsevier's Data Policy recommends Genome Sequence Archive

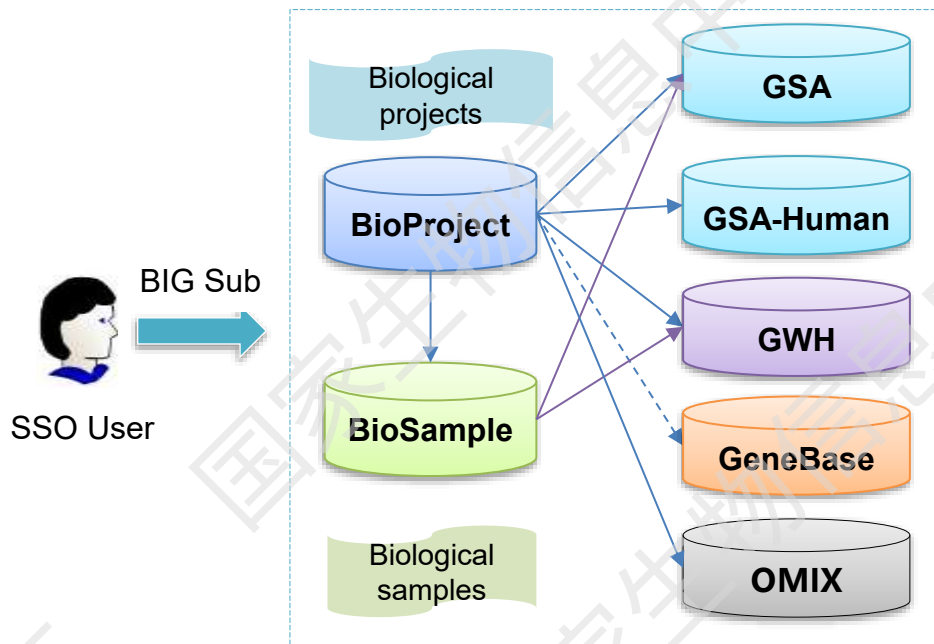
<https://www.elsevier.com/researcher/author/tools-and-resources/research-data/data-base-linking#repositories>

# INSDC Data Mirroring & Integrating

- Automated pipelines are developed to download data from INSDC, and integrate them to the local databases
- Data search, browse, visualization and download functions are provided

| CNCB database | INSDC database | Data type                 | Integrated data                                        |
|---------------|----------------|---------------------------|--------------------------------------------------------|
| GSA           | SRA            | Raw sequence reads        | 36,513,811 experiment<br>38,712,098 Run<br>24 PB files |
| GWH           | GenBank        | Genome assembly           | 3,235,203                                              |
|               | RefSeq         | Reference genome assembly | 569,560                                                |
| GenBase       | GenBank        | Nucleotide & Protein      | Nucleotides: 278,877,556<br>Proteins: 380,993,958      |

# Connectivity of the Depositories



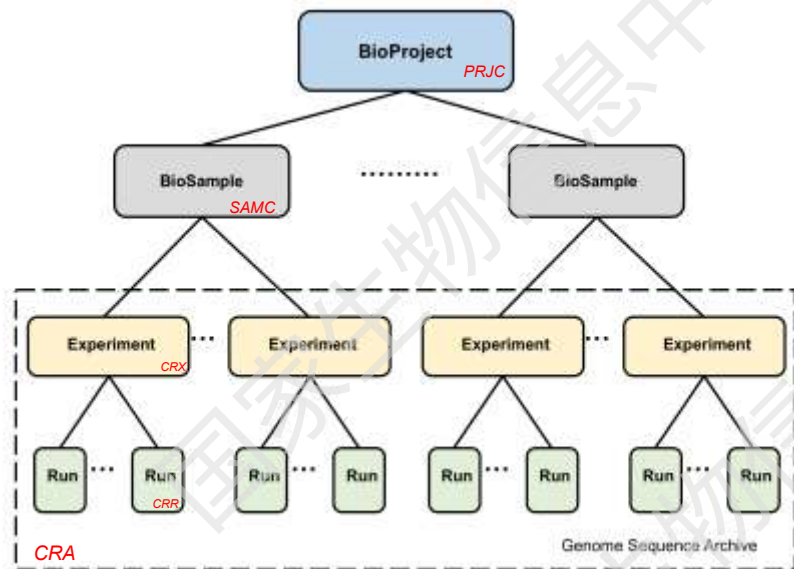
<https://ngdc.cncb.ac.cn/gsub>

# Contents

- 1 CNCB Data Submission System Overview
- 2 GSA Data Submission & Sharing**
- 3 GSA-Human Data Submission & Sharing
- 4 GWH Data Submission & Sharing
- 5 GenBase Data Submission & Sharing
- 6 OMIX Data Submission & Sharing



# Data Model and Supported Sequencing Platforms



Data model and data standard: **In accordance with the INSDC standard**

| Platform          | File format                          |
|-------------------|--------------------------------------|
| ILLUMINA          | FASTQ, BAM                           |
| PACBIO_SMRT       | HDF5, BAM, FASTQ                     |
| Oxford Nanopore   | FASTQ, BAM, Fast5                    |
| COMPLETE_GENOMICS | Complete_Genomics_native, FASTQ, BAM |
| BGISEQ            | FASTQ, BAM                           |
| BIONANO           | BNX                                  |
| ABI_SOLiD         | FASTA/QUAL, FASTQ, BAM               |
| ...               | ...                                  |

# Accepted Sample Types

## ☐ Pathogen

Used for pathogen samples that are relevant to public health. Required attributes include those considered useful for the rapid analysis and trace back of pathogens.

### ☐ Clinical or host-associated

### ☐ Environmental, food or other

## ☐ Microbe

Used for bacteria or other microbes when it is not appropriate or advantageous to use the Pathogen or Virus packages.

## ☐ Animal

Use for multicellular samples or cell lines derived from common laboratory model organisms, or large mammals including zoo and farm animals.

### ☐ Model organism or animal

Used for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, *Drosophila*, worm, fish, frog, or large mammals including zoo and farm animals.

### ☐ Animal for TCM syndrome

Use for model animal samples in Traditional Chinese Medicine syndrome researches.

## ☒ Human

Human genetic related raw sequence reads should be submitted to the [GSA for Human](#) database.

## ☐ Plant

Used for any plant sample or parts.

## ☐ Virus

Used for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen, Clinical or host-associated pathogen package.

## ☐ Metagenome/Environmental Sample (GSC MIMS unsupported)

Use for metagenome/environmental samples when it is not appropriate or advantageous to use the GSC (Genome Standards Consortium) MIMS (Minimum Information about a MetaGenome Sequence) standards. Organism must be a metagenome, where lineage starts with unclassified sequences and scientific name ends with metagenome.

## Metagenome/Environmental Sample (GSC MIMS compliant)

Describe and standardize sample methods, defined by the GSC (Genome Standards Consortium) MIMS (Minimum Information about a MetaGenome Sequence) standards for metagenome/environmental samples. Organism must be a metagenome, where lineage starts with unclassified sequences and scientific name ends with metagenome.

### ☐ human-gut

### ☐ soil

### ☐ water

<https://www.genesc.org/pages/standards-intro.html>

For samples related to pathogens (such as COVID-19), this type is preferred.

For samples related to human, please submit to GSA-Human.

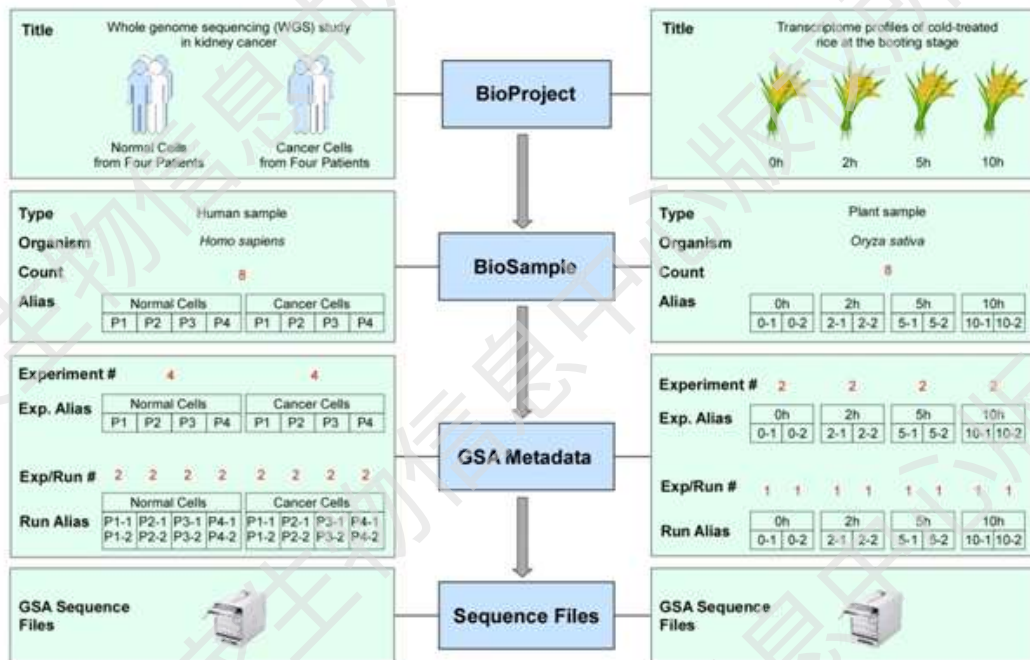
Use for metagenome/environmental samples when it is not appropriate or advantageous to use the GSC (Genome Standards Consortium) MIMS (Minimum Information about a MetaGenome Sequence) standards.

# Data Submission Portal : BIG-Sub

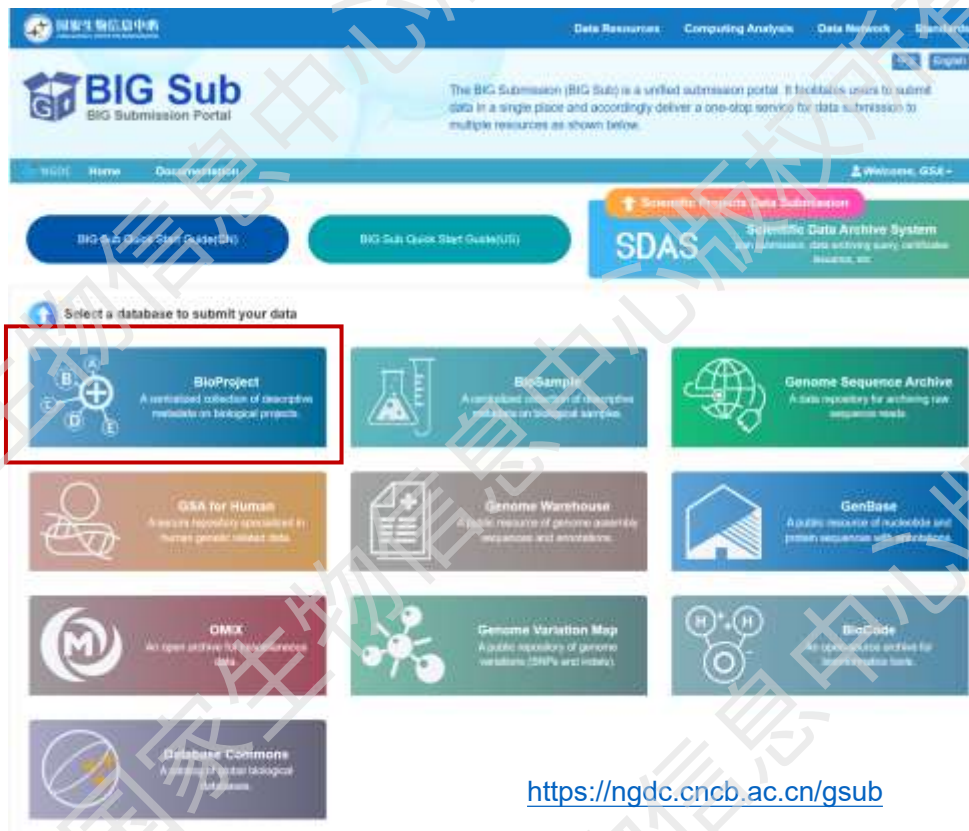


<https://ngdc.cncb.ac.cn/gsub>

# Data Submission Process



# Create a New BioProject Submission



# Step 1. Submitter Information

NCDC Home Documentation Welcome, GSA

BIG Sub BioProject New BioProject

01 Submitter 02 Genomes 03 Project Type 04 Publications 05 Overview & Submit

## Submitter Information

|                                   |                      |                |
|-----------------------------------|----------------------|----------------|
| * First name                      | Middle name          | * Last name    |
| GSA                               | middle name          | BIGD           |
| * Email                           | Email (secondary)    |                |
| genet2@big.ac.cn                  | secondary email      |                |
| * Organization                    | Organization website | * Department   |
| Beijing Institute of Genomics, CH | http://www.big.ac.cn | BIGD           |
| Phone                             | Fax                  |                |
|                                   |                      |                |
| * Street                          | * City               | State/Province |
| No.1 Beichen West Road, Chaoyi    | Beijing              |                |
| * Postal code                     | * Country/Region     |                |
| 100101                            | China                |                |
| <a href="#">Save and forward</a>  |                      |                |

## Step 2. General Information

Sub ID SubProject Submitters subPRJ0000077

Submitter General Project Title Publication Version Subject

General Information

Release Date

☐ After the approval is passed, release immediately following creation  
☒ Release on specified date

2027-09-30  
YYYY-MM-DD

Specify project

Branch center

Project

Project title Relevance

Description

Grants

Agency Program Grant ID Grant Title

External link

Link description URL

Related projects

Accession Description

Accessions in other database

Accession Database name

Save and forward



# Step 3. Project Type

BIG Sub / BioProject Submission: subPRO000277

01 Submitter 02 General 03 Project Type 04 Publication 05 Overview & Submit

### Project type

**\* Project data type**

- ☐ Whole genome sequencing
- ☐ Clono ants
- ☐ Epigenomics
- ☐ Exome
- ☐ Map
- ☐ Metagenome
- ☐ Phenotype or Genotype
- ☐ Random survey
- ☐ Targeted Locus (Loci)
- ☐ Transcriptome or Gene expression
- ☐ Variation
- ☐ Genome sequencing and assembly
- ☐ Raw sequence reads
- ☐ Genome sequencing
- ☐ Assembly
- ☐ Metagenomic assembly
- ☐ Targeted loci cultured
- ☐ Targeted loci environmental
- ☐ Single cell sequencing
- ☐ Clinical information
- ☐ Biomarkers
- ☐ Proteome
- ☐ Metabonomics
- ☐ Lipidomics
- ☐ Other

**\* Sample scope**

Monoisolate

**Sample scope choices**

**Monoisolate:** a single animal, cultured cell-line, inbred population (or possibly a heterogeneous population when a single genome assembly is generated from the pooled sample; not preferred).

**Multisolate:** multiple individuals, a population (representative of a species).

**Multi-species:** sample represents multiple species.

**Environment:** the species content of the sample is not known.

**Synthetic:** the sample is synthetically created by a machine.

**Single cell:** the sample is single cell sequencing.

**Other:** specify the sample scope that was used.

Save and forward

# Step 4. Publications

BIG Sub / BioProject / Submission: subPRO089277

01 Submitter 02 General 03 Project Type 04 Publication 05 Overview & Submit

Publication

PubMed ID OR DOI

+ Add another publication

Save and forward

# Step 5. Overview & Submit

BIG Sub / BioProject / Submission: subPRO069277

01 Submitter 02 General 03 Project Type 04 Publication 05 Overview & Submit

Overview

Submitter Information

|                |                                                            |
|----------------|------------------------------------------------------------|
| Submitter      | GSA BIGD                                                   |
| Organization   | <a href="mailto:gsatest2@big.ac.cn">gsatest2@big.ac.cn</a> |
| Department     | BIGD                                                       |
| Country/Region | China                                                      |
| Address        | No.1 Beichen West Road,Chaoyang District Beijing           |
| Postal code    | 100101                                                     |

General Information

|               |                    |
|---------------|--------------------|
| Project title | test2              |
| Relevance     | Agricultural       |
| Description   | test2              |
| Release Date  | 2027-09-30         |
| Grants        | Agency             |
|               | No funding support |

Project type

|                   |                         |
|-------------------|-------------------------|
| Project data type | Whole genome sequencing |
| Sample scope      | Multisolate             |

Submit

# Submission Completed

BIG Sub / BioProject

Create BioProject

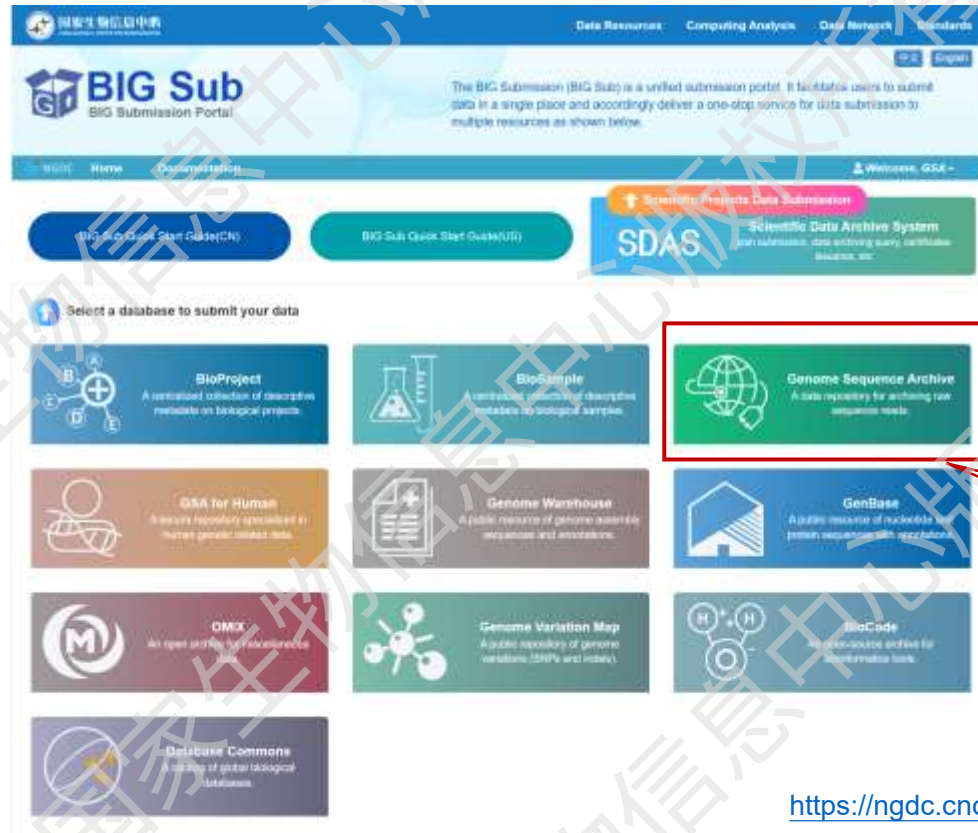
BioProject accession numbers

Remember this number. It will be used when submitting the GSA later

[BioProject Quick Start Guide\(US\)](#) [BioProject Quick Start Guide\(CN\)](#)

| Accession   | Submission ID | Project title | Release Date | Status     |              | Operation                                   |
|-------------|---------------|---------------|--------------|------------|--------------|---------------------------------------------|
| PRJCA047199 | subPRO069277  | test2         | 2027-09-30   | finished   | confidential | <a href="#">Edit</a> <a href="#">Delete</a> |
| PRJCA039910 | subPRO058697  | test_20250508 | 2027-05-08   | finished   | confidential | <a href="#">Edit</a> <a href="#">Delete</a> |
| Unassigned  | subPRO006710  | xx            | 2026-11-28   | unfinished | confidential | <a href="#">Edit</a> <a href="#">Delete</a> |
| PRJCA004553 | subPRO006709  | Disease study | 2021-02-25   | finished   | confidential | <a href="#">Edit</a> <a href="#">Delete</a> |
| PRJCA001016 | subPRO001507  | rice project  | 2020-02-03   |            |              | <a href="#">Edit</a>                        |

# Create a New GSA Submission



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

# Step 1. Submitter

BiG Sub GSA New GSA

01 Submitter Information 02 General 03 Metadata 04 Files 05 Overview & Submit

Submitter Information

|                                    |                      |                |
|------------------------------------|----------------------|----------------|
| * First name                       | Middle name          | * Last name    |
| GSA                                | middle name          | BIGD           |
| * Email                            | Email (secondary)    |                |
| gsa@big.ac.cn                      | secondary email      |                |
| * Organization                     | Organization website | * Department   |
| Beijing Institute of Genomics, Chi | http://www.big.ac.cn | BIGD           |
| Phone                              | Fax                  |                |
|                                    |                      |                |
| * Street                           | * City               | State/Province |
| No. 1 Beichen West Road, Chaoyi    | Beijing              |                |
| * Postal code                      | * Country/Region     |                |
| 100101                             | China                |                |

Save and forward

Save and forward

# Step2. General Information

01  
Submitter

02  
General

03  
Metadata

04  
File Upload

05  
Overview

General information

Release date

☐ After the approval is passed, release immediately following curation (recommended)

☒ Release on specified date

2019-08-06

(yyyy-mm-dd)

\* Release Policies and Disclaimers

1. A date can be set by authors to withhold the release of new submissions for a specified period of time.

2. The release date can be changed through the GSA submission portal ([http://bigd.big.ac.cn/gsub/submit/gsa/\[substitute your GSA accession number\]/contents](http://bigd.big.ac.cn/gsub/submit/gsa/[substitute your GSA accession number]/contents))

3. If a paper citing the sequence or accession number is published prior to the specified date, the sequence will be released upon publication. Otherwise, GSA will release sequence data on the specified date.

4. As soon as they are available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following address: [gsa@big.ac.cn](mailto:gsa@big.ac.cn)

☒ I accept it.

☐ I don't accept it.

The release date is the date your data can be publicly accessed. If your submitted information need to be protected for some time, you can set this date to a future date within two years.



# Step 2. General Information

**Title And Description**

\* title

\* description

**Project Information**

\* Please select Bioproject number

OR Go to create [BioProject](#)

**Sample Information**

☒ No GSA related BioSample information was created

☐ GSA related BioSample information has been created

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

[Save and forward](#)

Select an existing BioProject number  
or go to create a BioProject first

Select whether the related  
bioSamples has been created

# Step 3. Choose a Sample Type

BIG Sub / GSA / Submission: sub01488208

01 Submission 02 General 03 **Sample Type** 04 Attribution 05 Metadata 06 Files 07 Review & Submit

### Sample Type

☐ **Pathogen**  
Used for pathogens because they are relevant to public health. Required attributes include those associated with the rapid isolation and treatment of pathogens.

☐ **Clinical or host-associated**

☐ **Environmental, food or other**

☐ **Microbe**  
Used for bacteria or other microbes when it is not appropriate or advantageous to use the Pathogen or Virus packages.

☐ **Animal**  
Used for publication samples or cell lines derived from common laboratory model organisms, or large animals including porcine and non-human primates.

☐ **Model organism or animal**  
Used for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, Drosophila, worm, fish, frog, or large mammals, including pig and farm animals.

☐ **Animal for TCM syndrome**  
Used for model animal samples in Traditional Chinese Medicine syndrome research.

☐ **Plant**  
Used for any plant sample or cell line.

☐ **Virus**  
Used for all virus samples not directly associated with disease. New pathogens should be identified using the Pathogen (Clinical or host-associated) pathogen package.

☐ **Metagenome/Environmental Sample (GSC IMBA unsupported)**  
Use for metagenome/environmental samples when it is not appropriate or advantageous to use the GSC Genome Standards Compliant MBI. Minimum information about a Metagenome (Minimum Information about a Metagenome) must be a metagenome, where sample data will be unclassified sequences and scientific name will be metagenome.

☐ **Metagenome/Environmental Sample (GSC IMBA compliant)**  
Despite the publication sample metadata, defined by the GSC Genome Standards Compliant MBI (Minimum information about a Metagenome) standards, for metagenome/environmental samples. Organisms need to be metagenome, where sample data will be unclassified sequences and scientific name will be metagenome.

☐ **Human gut**

☐ **soil**

☐ **water**

[Go and forward](#)

# Step 4. Sample Attributes

BIG Sub : GSA Submission: subCRA050200

01 Submitter 02 General 03 Sample Type 04 Attributes 05 Metadata 06 File 07 Overview & Submit

## Attributes of clinical or host-associated pathogen

- Upload BioSample batch submission file using Excel format that includes the attributes for each of Samples

Please select file

Click here to download the BioSample submission template file, fill it out and upload it

- Download BioSample batch submission template [Pathogen\\_cl.us.xlsx](#) then fill in and double-check it before uploading.

For column explanations and examples, please see the [e.g.Pathogen\\_cl.us.xlsx](#).

This is an example file. Note: Please do not edit or upload this file

For more information, please see the [Help](#).



This is a submission template for batch deposit of "Clinical or host-associated pathogen" to the BioSample database.

GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.

GREY fields are optional. Leave optional fields empty if no information is available.

YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied and pasted down.

Organism (column 5) and host (column 11): Please fill in the name of the scientific species in the Taxon category of NCBI.

### CAUTION: DO NOT delete or insert column!

Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text

| 1           | 2                  | 3               | 4            | 5        | 6       | 7      | 8            | 9               | 10                  |
|-------------|--------------------|-----------------|--------------|----------|---------|--------|--------------|-----------------|---------------------|
| sample_name | public_description | project_voc_str | sample_title | organism | isolate | strain | collected_by | collection_date | geographic_location |

※Note: The Sample name needs to be consistent with the sample name in the experiment sheet of the GSA metadata file

| 10                  | 11   | 12          | 13               | 14                 | 15                 | 16       | 17       | 18            |
|---------------------|------|-------------|------------------|--------------------|--------------------|----------|----------|---------------|
| geographic_location | host | host_strain | isolation_source | latitude_longitude | culture_collection | genotype | host_age | host_age_unit |
|                     |      |             |                  |                    |                    |          |          | year          |

pathogen\_cl.cn.xlsx

# Step 4. Sample Attributes

Attributes of clinical or host-associated pathogen

• Uploaded BioSample batch submission file using Excel format that includes the attributes for each of Samples

Pathogen\_cl.us.xlsx 34KB

Please click the "Check" button to verify the submitted batch information online. If the file checked OK, please click the "Save and forward" to complete your submission. If not, please delete the file first, then modify and reload until it passed the examination.

Attributes of clinical or host-associated pathogen

• Uploaded BioSample batch submission file using Excel format that includes the attributes for each of Samples

Pathogen\_cl.us.xlsx 34KB   **✖ ERROR:**

Please download the error file: [error.txt](#)

Error in Sample sheet  
row 11, column 5: "organism" ("human") Human data please submit to <https://ngdc.cncb.ac.cn/gsa-human>

Attributes of clinical or host-associated pathogen

• Uploaded BioSample batch submission file using Excel format that includes the attributes for each of Samples

Pathogen\_cl.us.xlsx 34KB   ☒ Checked OK

Please click the "Check" button to verify the submitted batch information online. If the file checked OK, please click the "Save and forward" to complete your submission. If not, please delete the file first, then modify and reload until it passed the examination.

Click here to download the error report file

The error messages displayed online

Checked error, delete the uploaded file, fix it as prompted, re-upload and re-check it

# Step 5. GSA Metadata Information

BIG Sub: GSA Submission: subCRA000000

01 Submission 02 Samples 03 Sample Type 04 Attributes 05 Metadata 06 Files 07 Download & Submit

### GSA Metadata Information

Upload the GSA batch submission file using Excel format.

Please select file

Download GSA batch submission template [GSA\\_Template.us.xlsx](#) then fill in and double-check it before uploading.

For column explanations and examples, please see the [e.g. GSA\\_Template.us.xlsx](#)

For more information, please see the [Help](#)

Click here to download the GSA metadata template file, fill it out and upload it

This is an example file. **Note: Please do not edit or upload this file**

This sheet is used for Experiment information and needs to be created after creating BioProject and BioSample.

GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.

GREY fields are optional. Leave optional fields empty if no information is available.

BLUE fields require for paired-end data only.

YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.

When the Platform (column 5) is Helicos Heliscope, the Planned number of cycles (column 17) is required.

**CAUTION: DO NOT delete or insert column!**

Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is acco

| 1  | 2                | 3                    | 4              | 5                   | 6        | 7                                          |
|----|------------------|----------------------|----------------|---------------------|----------|--------------------------------------------|
| ID | Experiment title | BioProject accession | BioSample name | BioSample accession | Platform | Library Construction / Experimental Design |
| E1 |                  |                      |                |                     |          |                                            |

Experiment Run Please Read First!

※Note: The Sample name needs to be consistent with the sample name in the Biosample metadata file

This sheet is used for Run information and needs to be created after Experiment sheet

GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.

GREY fields are optional. Leave optional fields empty if no information is available.

BLUE fields require for paired-end data only.

PURPLE When your Run data file type (column 8) select bam. If you want to submit your reference file to our FTP Site, you need to fill in the Reference file name and

YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-paste

**CAUTION: DO NOT delete or insert column!**

Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Double

| 1  | 2         | 3                    | 4             | 5                  | 6           | 7              |
|----|-----------|----------------------|---------------|--------------------|-------------|----------------|
| ID | Run title | BioProject accession | Experiment ID | Run data file type | File name 1 | MD5 checksum 1 |
| R1 |           |                      |               |                    |             |                |
| R2 |           |                      |               |                    |             |                |

Experiment Run Please Read First!

GSA\_Template.cn.xlsx

# Step 5. GSA Metadata Information

BiG Sub / GSA / Submission: subCRA050200

01 Submitter 02 General 03 Sample Type 04 Attributes 05 Metadata 06 Files 07 Overview & Submit

### GSA Metadata Information

• Uploaded GSA batch submission file.

GSA\_Template.us.xlsx 41KB   ☒ Checked OK.

1 Please click the "Check" button to verify the submitted batch information online. If the file checked OK, please click the "Save and forward" to complete your submission. If not, please delete the file first, then modify and reload until it passed the examination.

# Step 6. Upload Sequence Files



How do you want to provide files for submission?

☒ FTP

Use FTP client to upload files. FTP Username and Password are same as the GSA Sub account.

You may upload your files to our FTP via desktop FTP client software (such as the [FileZilla Client](#)) or using the FTP command.

Address: <http://submit2.bgi.ac.cn>

Username: [submit2@bgi.ac.cn](mailto:submit2@bgi.ac.cn)

Password: Same as you login the GSA Sub.

Notice:

1. After you connected to the FTP server, please navigate to the GSA directory, create a folder named subGSA000000, and upload your files to subGSA000000. DO NOT upload files to the root directory.

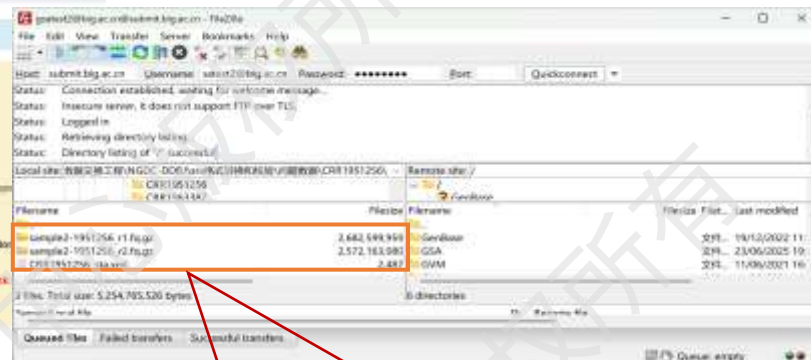
2. Please do not upload files irrelevant to the submissions.

3. Please upload files using the binary mode. If you are using an FTP client software, please refer to the software's documentation to set the binary mode. If you are using the FTP command line, run the "binary" command before running the "upload" command to set the binary mode.

☐ Aspera Command Line

Use Aspera Command Line to upload files. You need to install the Aspera Connect plugin before uploading your files. If you don't have Aspera Connect plugin installed in your computer, please [Download Aspera Connect](#) and install it first.

Save and forward



The file name and MD5 code need to be consistent with those filled in the Run sheet

It is recommended to upload files using a FTP client software.

Note: Please upload in binary mode



# Step 7. Overview & Submit

NGS Sub - GSA - Submission: testORA000000

01 Submitter 02 Submit 03 Submit Data 04 Submission 05 Metadata 06 Files 07 Overview & Submit

**Basic Information**

|               |               |
|---------------|---------------|
| Submission ID | testORA000000 |
| Site          | test20250000  |
| Description   | test20250000  |
| Release date  | 2027-09-30    |

**Submitter Information**

|                |                                                            |
|----------------|------------------------------------------------------------|
| Submitter      | GSA (GSC)                                                  |
| Organization   | gustest2@gsc.ac.cn                                         |
| Organization   | Beijing Institute of Genomics, Chinese Academy of Sciences |
| Department     | IGAD                                                       |
| Country/Region | China                                                      |
| Address        | No. 1 Beichen West Road, Chaoyang District, Beijing        |
| Postal code    | 100101                                                     |

**Metadata**

|                             |                                 |
|-----------------------------|---------------------------------|
| Sample type                 | Chloroplast-associated pathogen |
| Disc sample attributes file | Pathogen_classification         |
| Metadata file               | GSA_Template.xlsx               |

| Experiment Title | Platform                  | Sample Name | Taxon Name                                      | Selection |
|------------------|---------------------------|-------------|-------------------------------------------------|-----------|
| test1            | illumina hbbio 2000       | sample1     | Severe acute respiratory syndrome coronavirus 2 | ProxA     |
| Run Alias        | Run Data File information |             |                                                 | File Type |
| test1            | Pathogen assembly         |             |                                                 | Raw       |

Total Items: 1 Item of 1 Item Page size: 15 Page 1/1 << First Last >> Jump To: 1 30

**Submit**

After checking, click "Submit" to complete the submission

# Metadata Submission Completed

Create GSA

GSA Quick Start Guide (US) GSA Quick Start Guide (CN)

Waiting for check

| Accession  | Submission ID | Title         | Create date | Release date | Status                          | Operation           |
|------------|---------------|---------------|-------------|--------------|---------------------------------|---------------------|
| Unassigned | subCRA050200  | test120250930 | 2025-09-30  | 2027-09-30   | Unchecked<br>Confidential       | Delete              |
| Unassigned | subCRA041627  | test_20250508 | 2025-05-08  | 2027-05-08   | Checking detail<br>Confidential | Delete              |
| CRA011238  | subCRA027044  | test_1        | 2024-06-03  | 2027-05-08   | Checked OK<br>Confidential      | ReleaseNow<br>Share |

The GSA number assigned to a successful submission, which **can be cited in an article**

## Data Review Process

### Phase 1 : Manual Review

Metadata double check

Verify associated relationships

Verify completeness of metadata

Metadata review:  
1-2 working days.

### Phase 2 : Data Association and Quality Control

Data file association

Data quality control

Data quality analysis

The file processing  
cycle is related to the  
**size and number** of files.

Please ensure that the relevant files have been uploaded. Wait until the data are verified and archived. If the data validation fails, please make the necessary corrections and resubmit the metadata or re-upload the files.

# Data Retrieval and Access

**GSA** Genome Sequence Archive

Search:

NGDC Home Submit Browse Search Services Support Login Register 语言/language

**Center**

- INSDC (7261700)
- NGDC (251857)

**Downloadable**

- Non-downloadable (1076548)
- Downloadable (6437011)

**File type**

- ab1 (20)
- sff (382)
- SRA (6155018)
- undefined (1074655)
- oxfordnanopore\_native (5)
- pacbio\_sequel\_native (19)
- fastq (245679)
- fasta (502)
- bam (5186)
- Bam (1187)

**Platform**

- ULTIMA (397)
- unspecified (74872)
- ILLUMINA (11471)

**Search result:**

Total items: 7513537 | Items: 1 of 20 | Page: 1 / 32678 |    |

|                          |                       |                                                                                              |                        |             |
|--------------------------|-----------------------|----------------------------------------------------------------------------------------------|------------------------|-------------|
| <input type="checkbox"/> | <b>Title:</b>         | Illumina NovaSeq 6000 paired end sequencing: Land use transition effects on soil microbiomes |                        |             |
| 1.                       | <b>Accession:</b>     | ERX11458488                                                                                  | <b>Source name:</b>    | METAGENOMIC |
|                          | <b>Platform name:</b> | ILLUMINA                                                                                     | <b>Selection name:</b> | RANDOM      |
|                          | <b>Strategy name:</b> | Genome                                                                                       |                        |             |
| <input type="checkbox"/> | <b>Title:</b>         | Illumina NovaSeq 6000 paired end sequencing: Land use transition effects on soil microbiomes |                        |             |
| 2.                       | <b>Accession:</b>     | ERX11458487                                                                                  | <b>Source name:</b>    | METAGENOMIC |
|                          | <b>Platform name:</b> | ILLUMINA                                                                                     | <b>Selection name:</b> | RANDOM      |
|                          | <b>Strategy name:</b> | Genome                                                                                       |                        |             |
| <input type="checkbox"/> | <b>Title:</b>         | Illumina NovaSeq 6000 paired end sequencing: Land use transition effects on soil microbiomes |                        |             |
| 3.                       | <b>Accession:</b>     | ERX11458519                                                                                  | <b>Source name:</b>    | METAGENOMIC |
|                          | <b>Platform name:</b> | ILLUMINA                                                                                     | <b>Selection name:</b> | RANDOM      |
|                          | <b>Strategy name:</b> | Genome                                                                                       |                        |             |
| <input type="checkbox"/> | <b>Title:</b>         | Illumina NovaSeq 6000 paired end sequencing: Land use transition effects on soil microbiomes |                        |             |
| 4.                       | <b>Accession:</b>     | ERX11458498                                                                                  | <b>Source name:</b>    | METAGENOMIC |
|                          | <b>Platform name:</b> | ILLUMINA                                                                                     | <b>Selection name:</b> | RANDOM      |
|                          | <b>Strategy name:</b> | Genome                                                                                       |                        |             |
| <input type="checkbox"/> | <b>Title:</b>         | Illumina NovaSeq 6000 paired end sequencing: Land use transition effects on soil microbiomes |                        |             |
| 5.                       | <b>Accession:</b>     | ERX11458508                                                                                  | <b>Source name:</b>    | METAGENOMIC |
|                          | <b>Platform name:</b> | ILLUMINA                                                                                     | <b>Selection name:</b> |             |

**Send to :**

**Send to :**

**Choose Destination**

- ☒ File

Download 5804385 items.

**Format**

- ☒ Accession List
- ☐ RunInfo
- ☐ DataSet Accession List

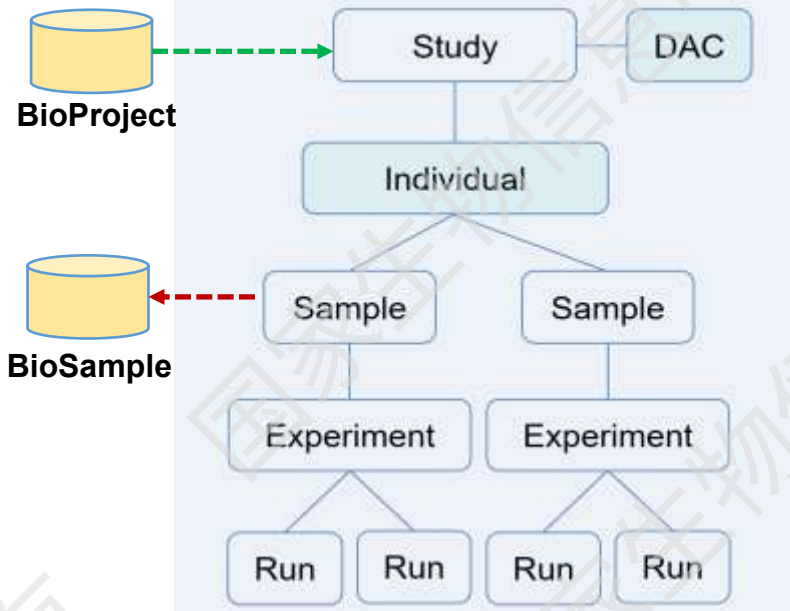
# Contents

- 1 CNCB Data Submission System Overview
- 2 GSA Data Submission & Sharing
- 3 GSA-Human Data Submission & Sharing**
- 4 GWH Data Submission & Sharing
- 5 GenBase Data Submission & Sharing
- 6 OMIX Data Submission & Sharing



# Data Model and Supported Data Types

## Data Organization Model



To organize metadata and data files centered on "Individual".

| Study type                  | Applicable metadata template |
|-----------------------------|------------------------------|
| Disease study               | Solid tumor                  |
|                             | Hematologic tumor            |
|                             | Autoimmune diseases          |
|                             | Other diseases               |
| Cohort study                |                              |
| Cell line related study     |                              |
| Clinical pathogen           |                              |
| Human associated metagenome |                              |

| Sequencing platform | File format                          |
|---------------------|--------------------------------------|
| ILLUMINA            | FASTQ, BAM                           |
| PACBIO_SMRT         | HDF5, BAM, FASTQ                     |
| Oxford Nanopore     | FASTQ, BAM, Fast5                    |
| COMPLETE_GENOMICS   | Complete_Genomics_native, FASTQ, BAM |
| BGISEQ              | FASTQ, BAM                           |
| BIONANO             | BNX                                  |
| ABI_SOLID           | FASTA/QUAL, FASTQ, BAM               |
| ...                 | ...                                  |

For more file format information: [https://ngdc.cncb.ac.cn/gsa/support/standardsGsa#file\\_1](https://ngdc.cncb.ac.cn/gsa/support/standardsGsa#file_1)

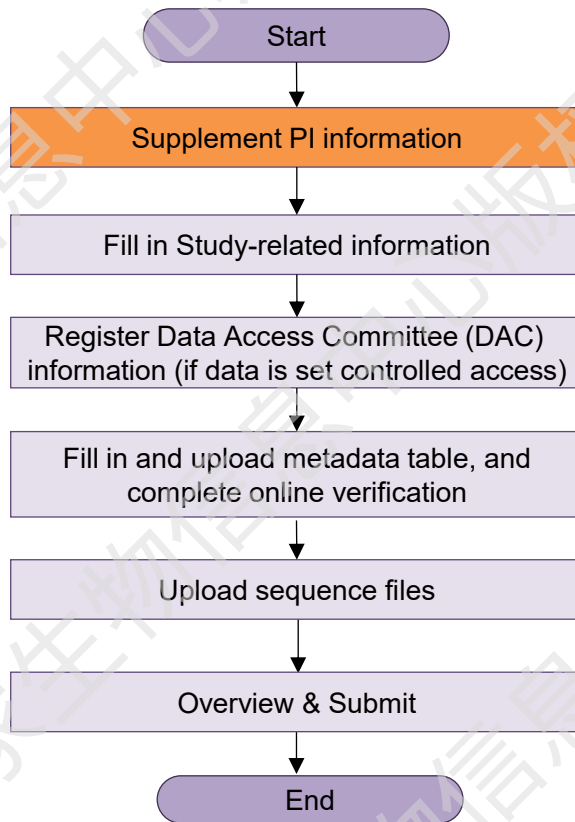


# Pre-Submission Notes

- **A PI account is required to make the submission – the data responsible person**
  - Principle Investigator (PI), Full or Associate Professor of a research institute
  - Full or Associate Professor of an university
  - Chief or Deputy Chief Physician of a hospital
  - Department Manager of a company
- The metadata to be submitted must be desensitized (no subject privacy information)
- Submitters should set the data access method upon submission
  - **Controlled access (default)**
  - Open access
- For the controlled access data, information of the Data Access Committee (DAC) and its members are required, and a DAC contact person should be designated



# Data Submission Process





# Create a New GSA-Human Submission



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

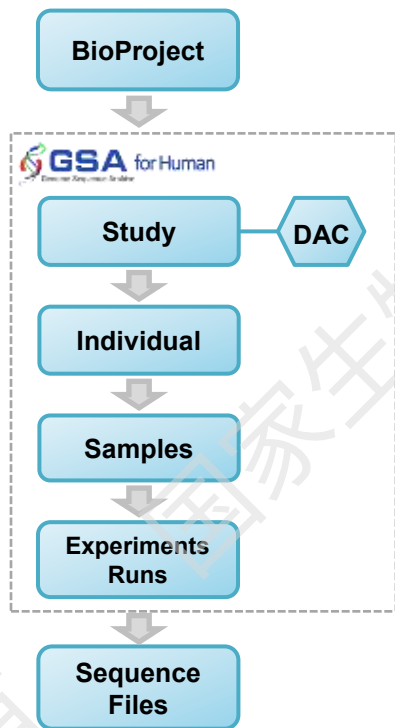
The screenshot shows the GSA for Human submission form. Red boxes highlight the "Professional Title" field, the "Institution" field, and the "Institutional Homepage" field. The form includes sections for "Complete Personal Information" and "Personal Information".

After logging in, please complete the following personal information:

- Professional title, personal homepage/corresponding publication URL, and institutional homepage information.
- The email on the personal homepage should preferably be consistent with the account email.
- **Submissions from unqualified accounts will not be approved**



# Step 1. Submitter



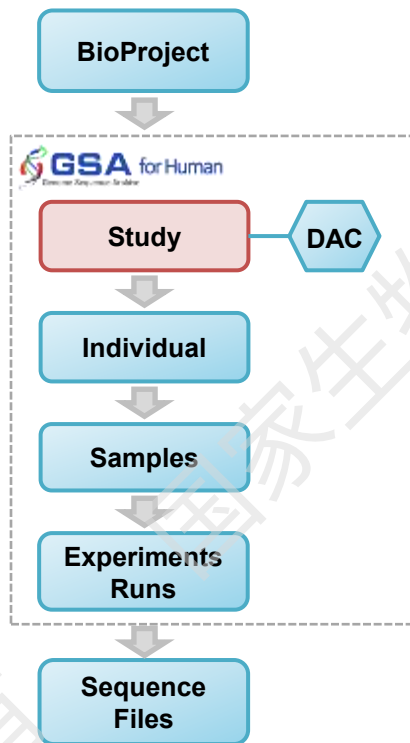
The screenshot shows the 'New Submission' page for GSA for Human. The 'Submitter Information' section includes a checkbox for 'Do you want to assign another person as Contact Person?' and a checkbox for 'I want to provide my contact information'. The 'Alternative Contact Information' section contains fields for First Name, Middle Name, Last Name, Email, Institute/Organization, Department, Street, City, State/Province, Postal Code, and Country/Region. The 'Privacy for the Access of Human Genetic Resource Data in NCDB' section contains a text area with the following text: 'The GSA for Human Database of NCDB is designed for the collection, management and access of human genetic resource information. There are two modes of data access in GSA for Human, namely Open Access and Controlled Access. The Open Access data can be downloaded and used directly, while the Controlled Access data can only be imported after applying to the original data submitter and getting authorized. 1. The data submitter has the right to restrict the specific usage of the submitted data regarding data and related materials of China's human genetic resource management and protection with ethics. 2. The data submitter can set the time of dataset access and revision, even if the related research result has been published or the intellectual property has been explained. The dataset revision conditions should be approved with consensus of the data submission committee (DAC) of the data submitter. 3. The data submitter and the DAC have the right to Agree or Disagree for the applicants of the data access and usage. 4. Responsibilities of data submitter 1. The data submitter is the person responsible for the dataset submitted, which should be the group leader. Before submitting to GSA for Human, the following conditions must be met: 1. There must not be any other data above.'

**It is recommended to add contact person information. If not added, feedback on dataset issues will only be sent to the registered account.** Since data submission and issue handling may not be done by the PI personally, adding a contact person ensures both parties receive notifications.

**Before proceeding to the next step, please make sure to read and accept our data management rules.**



# Step 2. Study Information



The screenshot shows the 'Study Information' section of the GSA for Human submission form. The form is divided into several sections, each with a red box highlighting the relevant fields:

- Release Date:** A dropdown menu with options: 'Release immediately following creation' and 'Release on specified date'. The 'Release on specified date' option is selected, and a date picker shows '2024-06-27'.
- Basic Information:** Fields for 'Study title', 'Description', 'Study type' (with radio buttons for 'Genome study', 'Cancer study', 'Cell line related study', 'Clinical pathway', 'Human development', 'Anatomy or function', 'Other study'), 'Disease name', 'Is your study a longitudinal study?', and 'Yes'/'No' buttons.
- BioProject Information:** A field for 'Please select Bioproject number' with a dropdown menu showing 'PNCAG001182' and a button 'Or Do it myself (BioProject)'.
- Data Accessibility:** A section for 'Data release policy' with radio buttons for 'Open access' and 'Controlled access'. The 'Open access' option is selected. Below this, there are checkboxes for 'Controlled access' and 'Non-profitable use', and a button 'I have read and accepted the DATA ACCESS AGREEMENT'.

**Release Date:** Any date **within 2 years** can be selected.

**Basic Information:** Choose the study type and its relevant information according to the research.

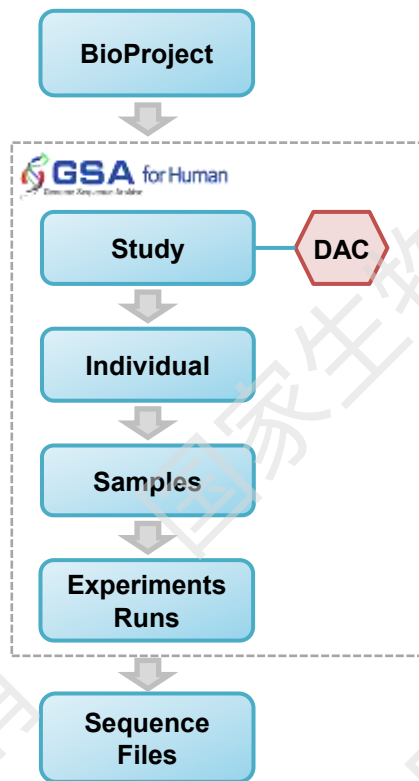
**BioProject Information:** Select an existing BioProject number, or jump to create a new BioProject.

**Data Access Methods**

- Select data access method (Controlled-access / Open-access) and usage restrictions.

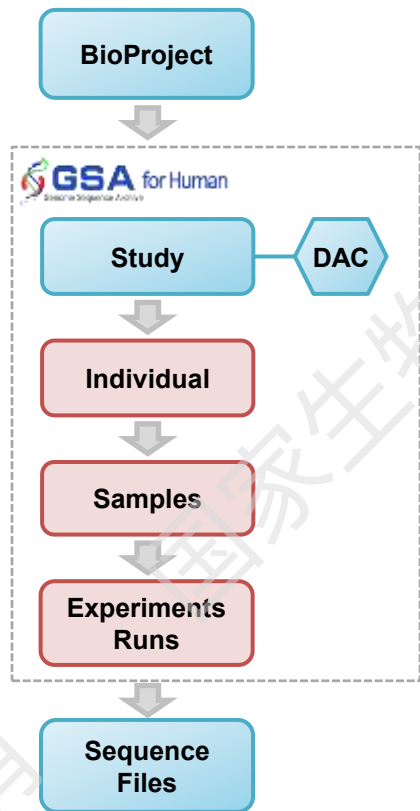


## Step 3. Data Access Committee (DAC) Information

[illegible]

If "Controlled-access" is selected as the data access method, a DAC must be established and a DAC contact person designated to review and authorize data access applications after data release.

# Step 4. Metadata Information



The screenshot shows the 'GSA for Human' submission page. The submission ID is 'subHRA000650'. The 'Metadata' tab is selected. Under the 'General Package' section, there is an option to 'Download Excel Template General 1.0 spreadsheet'. A red box highlights this option, and a red arrow points from it to the 'Study Type' section of the form below.

## Study Type of research information

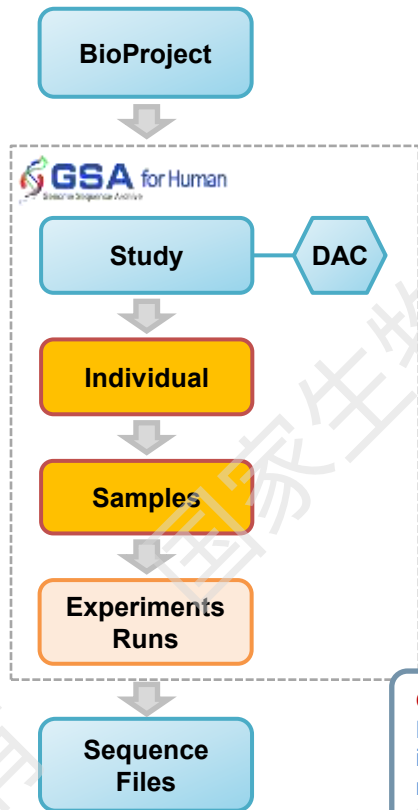
The form shows the 'Study Type' section with radio buttons for 'Disease study', 'Cohort study', 'Cell line related study', 'Clinical pathogen', 'Human metagenome', 'Ancients or Fossils', and 'Other study'. Below this is a 'Disease Name' field with a note: 'The name of disease should be chosen from Human Disease Ontology'. At the bottom, there is a question 'Is your study a longitudinal study?' with 'No' and 'Yes' radio buttons.

| Study Type                  | Applicable Metadata Template         |
|-----------------------------|--------------------------------------|
| Disease study               | General_1.0.xlsx                     |
| Cohort study                | General_1.0.xlsx                     |
| Cell line related study     | Cellline_1.0.xlsx                    |
| Clinical pathogen           | Clinical_pathogen_1.0.xlsx           |
| Human associated metagenome | Human_associated_metagenome_1.0.xlsx |

The metadata template is automatically matched based on the Study Type, to adapt to different research needs.



**Taking General\_1.0.xlsx as an example: Mainly used for Disease studies and Cohort studies.**

[illegible]

## Individual Sheet

## 1:N

## Common Issue 1: Corresponding Relationships

**one individual → multiple samples.** Be sure to pay attention to the affiliation of each sample. This is the part where problems are most likely to occur, since metadata is organized around "Individual", **any mistake will affect all metadata.**

[illegible]

## Sample Sheet

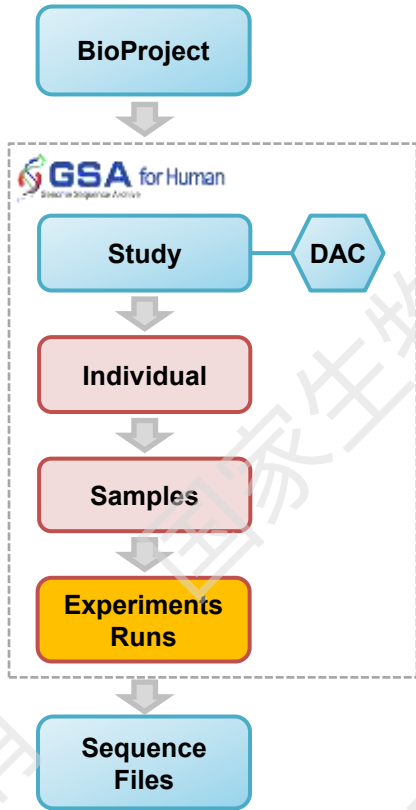
## Common Issue 2: Sensitive Information

Names of individuals and samples must not include full or abbreviated Pinyin of subjects' real names.

**Note:** Ensure metadata is **complete and accurate**. Comprehensive information provides sufficient background and context, which is critical for **administrator review, reviewer assessment, and future data reuse**. If existing columns cannot describe information in detail, add custom fields.

## Step 4. Metadata Information

**Taking General\_1.0.xlsx as an example: Mainly used for Disease studies and Cohort studies.**

[illegible]

## Experiment Sheet

### Common Issue 1 : Corresponding Relationships

If multiple sequencing strategies are used for one sample: 1 Sample  $\rightarrow$  N Experiments  
Relationship between Experiment and Run: 1 Experiment  $\rightarrow$  N Runs

This sheet is used for Run information and needs to be created after Experiment sheet.

GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed.

BLUE fields require for paired-end data only

PURPLE When your Run data file type (column 4) select bam, If you want to submit your reference file to our FTP Site, you need to fill in the Reference file name and MD5 for reference file (columns 9-10). If your run data file type is not bam, you do not need to fill in the Reference file name and MD5 for reference file (columns 9-10).

YELLOW columns have drop-down menus that allow you to select from a controlled vocabulary. Once specified for one row, these values can be copied-and-pasted down.

**CAUTION: DO NOT delete or insert column!**

Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before sending to us.

| 1  | 2         | 3             | 4                  | 5           | 6              | 7           | 8              | 9                   | 10                     | 11                         | 12                     |
|----|-----------|---------------|--------------------|-------------|----------------|-------------|----------------|---------------------|------------------------|----------------------------|------------------------|
| ID | Run title | Experiment ID | Run data file type | File name 1 | MD5 checksum 1 | File name 2 | MD5 checksum 2 | Reference file name | MD5 for reference file | Assembly Name or Accession | Assembly Accession URL |
| R1 |           |               |                    |             |                |             |                |                     |                        |                            |                        |

**Run Sheet**

## Run Sheet

## Common Issue 2 : Sensitive Information

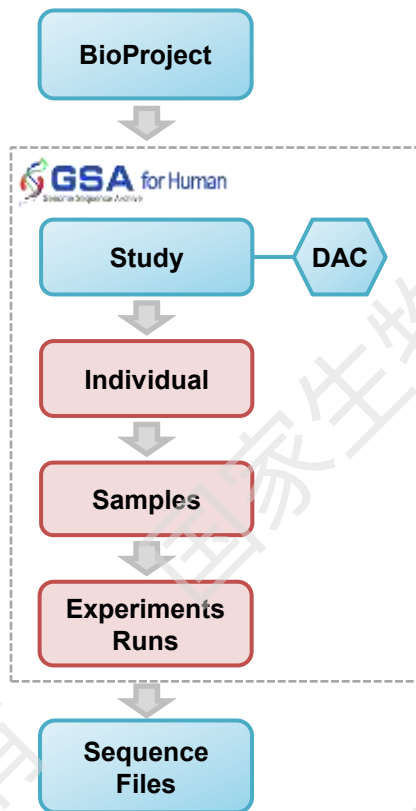
Titles of Experiments and Runs must not include full or abbreviated Pinyin of patients' real names.  
File names must not contain such sensitive information.

### Common Issue 3 : File Name Consistency

**File names in the metadata must be exactly the same as uploaded files.** Inconsistency will prevent the system from locating files during subsequent quality control, affecting file processing and archiving efficiency.



# Step 4. Metadata Information



The screenshot shows the 'GSA for Human' submission interface. The submission ID is 'subHRA000560'. The 'Metadata' tab is selected. The 'General Package' section shows the 'Upload the metadata file using Excel format' option. The 'Upload' button is highlighted with a red box. The 'Check' button is also highlighted with a red box. The 'Save and Forward' button is highlighted with a red box.

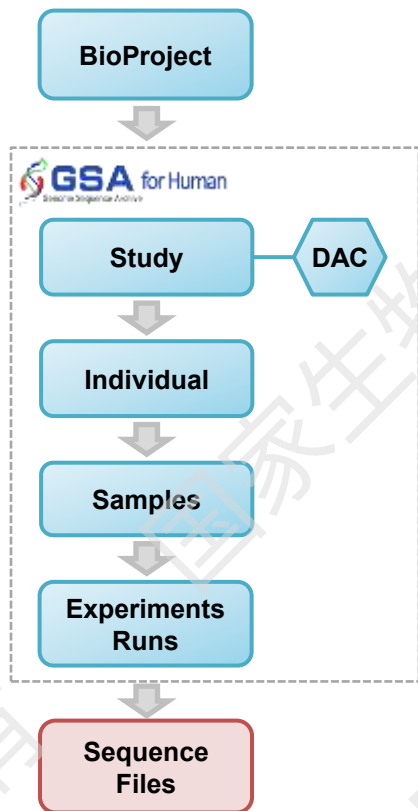
**Upload the completed metadata table.**

**Click "Check" to verify the format.**

**After successful verification, click "Save and forward" to proceed to the next step.**

- If the metadata **review fails**, modify the information according to system prompts and re-upload for verification.

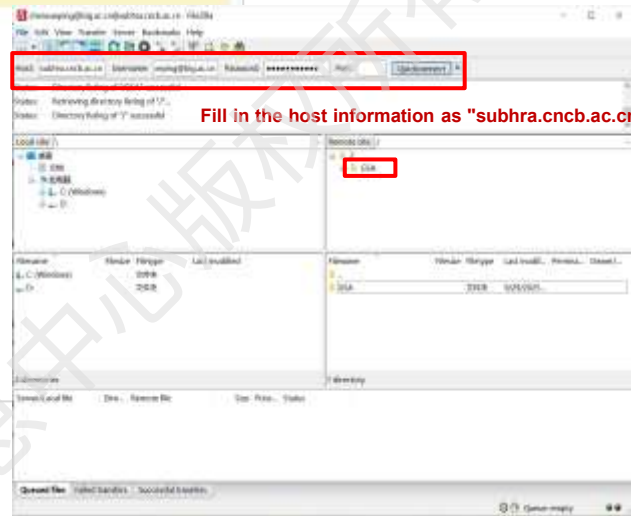
# Step 5. Sequence File Upload



The screenshot shows the 'File upload mode' section of the GSA submission interface. It includes a 'Notice' box with instructions: 'Each file must be added in the GSA metadata file you uploaded. Unique file names should be used for all files. Files must be compressed using gzip or bzip2. Take as the last step of the submission. Please click the 'Save and forward' button on the bottom of this page and step into the 'Overview' page, where you can overview and check the metadata you submitted and click the 'Submit' button to finish your submission.'

Below the notice, a red box highlights the 'FTP' option under 'How do you want to provide files for submission?'. The text for FTP states: 'Use FTP only to upload files. FTP username and password are same as the GSA Sub account. You only upload your files to our FTP site using an FTP client software (such as the [FileZilla Client](#)) or using the FTP command. Address: ftp.cnbc.ac.cn Username: Same as you register the GSA Sub Password: Same as you login the GSA Sub. After you connected to the FTP server, please navigate to the GSA directory and upload your files here. DO NOT upload files to the root directory. NOTE: Please use the binary mode to transfer files. If you are using the FTP client software to log in to the FTP server, please the software's help option to set the transfer mode. If you are using the command, type the binary command before the input command.'

Below the FTP section, there is a 'Save and forward' button.



# Step 6. Overview & Submit

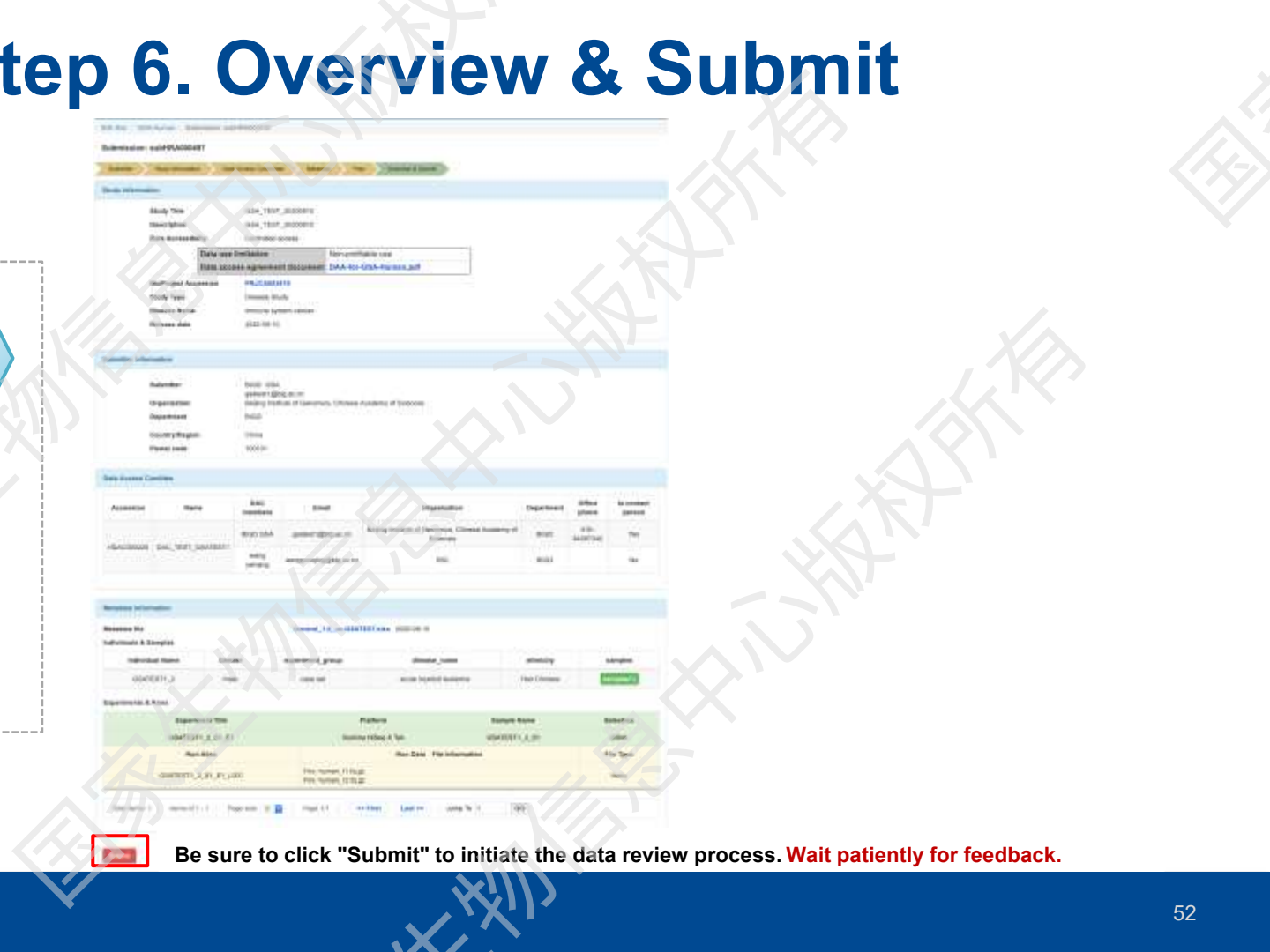
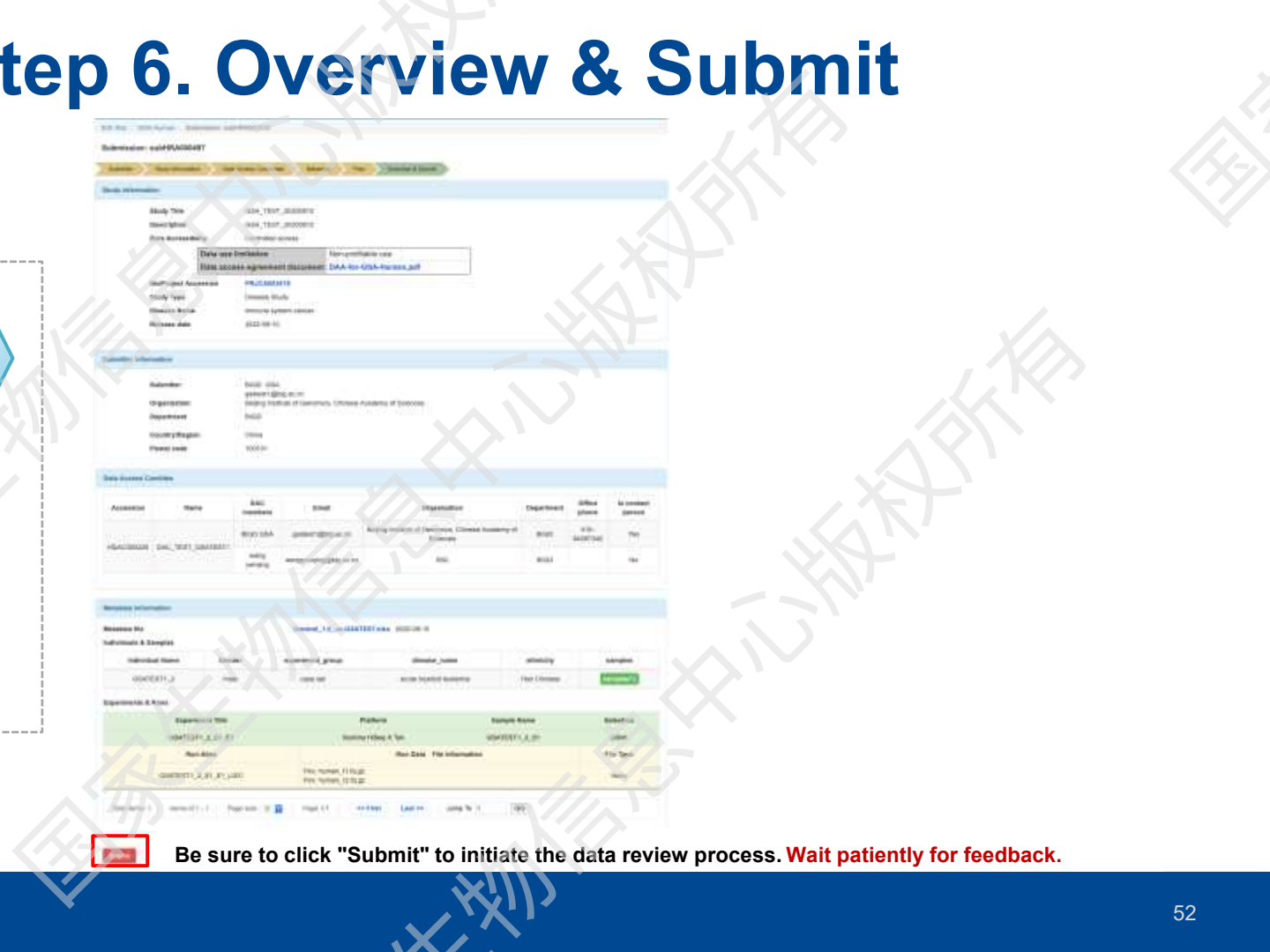
The screenshot displays the 'Overview & Submit' page for a GISAID submission. The submission ID is GISAID\_1000000000. The page is divided into several sections:

- Study Information:** Includes fields for Study Title, Description, Data use limitations (Non-proprietary use), and a link to the Data use limitations document (DAA-for-GISAID-users.pdf).
- Organization Information:** Includes fields for Organization, Department, Country/Region, and Phone number.
- Data Access Control:** A table showing access control for the submission. It includes columns for Accession, Name, Email, Organization, Department, and Is member. The table shows two rows of data.
- Sequence Information:** Includes fields for Submission ID, Submission Name, and Submission Date.
- Biospecimen & Assay:** A table showing biospecimen and assay information. It includes columns for Submission ID, Platform, Sample Name, and Submission Date. The table shows two rows of data.

A red box highlights the 'Submit' button at the bottom left of the page.

Be sure to click "Submit" to initiate the data review process. **Wait patiently for feedback.**

52



# Step 6. Overview & Submit

The screenshot displays the 'Overview & Submit' page for a GISAID submission. The submission ID is GISAID\_1000000000. The page is divided into several sections:

- Study Information:** Includes fields for Study Title, Description, Data use limitations (Non-proprietary use), and a link to the Data use limitations document (DAA-for-GISAID-users.pdf).
- Organization Information:** Includes fields for Organization, Department, Country/Region, and Phone number.
- Data Access Controls:** A table showing access controls for the submission, including columns for Accession, Name, Email, Organization, Department, and Is member.
- Sequence Information:** Includes fields for Submission ID, Submission Name, and Submission Date.
- Biospecimen & Assay:** A table showing biospecimen and assay information, including columns for Submission ID, Platform, Sample Name, and Submission ID.

A red box highlights the 'Submit' button at the bottom left of the page.

Be sure to click "Submit" to initiate the data review process. **Wait patiently for feedback.**

52

# Metadata Submission Completed

## Data Review Process

### Phase 1 : Manual Review

Authenticate user identity

Check sensitive information

Verify associated relationships

Verify completeness of metadata

Metadata review:  
1-2 working days.

### Phase 2 : Data Association and Quality Control

Data file association

Data quality control

Data quality analysis

The file processing  
cycle is related to the  
**size and number** of  
files.

| Accession        | Title  | Submission ID | Release Date | Package   | Status                         | Access(Request) | Operation                                                |
|------------------|--------|---------------|--------------|-----------|--------------------------------|-----------------|----------------------------------------------------------|
| Unassigned       | aaaaaa | subHRA000350  | 2021-09-30   | General   | Checking data/<br>Confidential | Controlled(0)   | <div>Delete</div> <div>Update</div>                      |
| <b>HRA000165</b> | test   | subHRA000201  | 2024-03-22   | Cell Line | Checked OK<br>Confidential     | Open            | <div>ReleaseNow</div> <div>Store</div> <div>Update</div> |

The assigned  
accession number

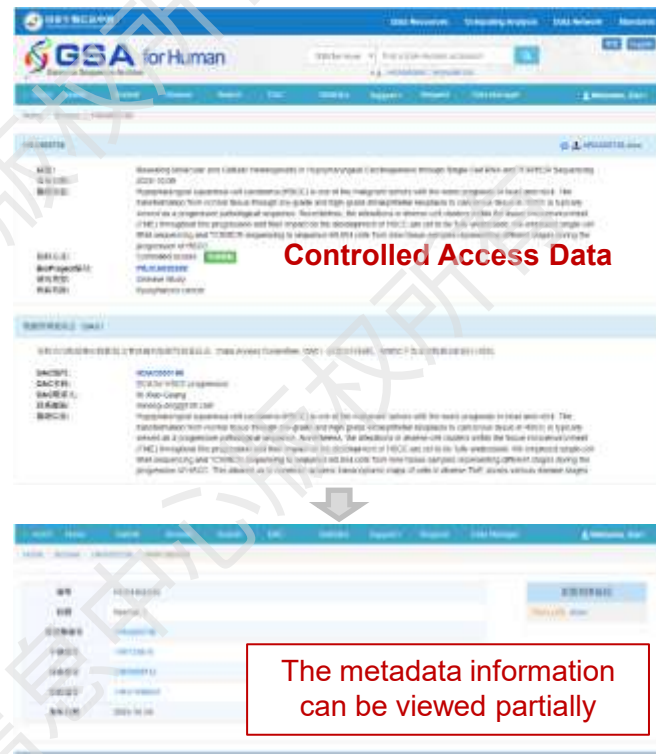
- Please ensure that the relevant files have been uploaded. Wait until the data are verified and archived. If the data validation fails, please make the necessary corrections and resubmit the metadata or re-upload the files.
- Accession Allocation: After data archiving, **a unique accession number will be assigned.**



# GSA-Human Data Access Permissions

| Access Type            | Data Status                                              | Access Permission Categories | Metadata View | Metadata Download | Sequence File Download |
|------------------------|----------------------------------------------------------|------------------------------|---------------|-------------------|------------------------|
| Open Access Data       | • Meets open sharing criteria*                           | Accessible                   | Full          | Full              | Allowed                |
|                        | • Does not meet open sharing criteria                    | Inaccessible                 | Partial       | Partial           | Not Allowed            |
| Controlled Access Data | • Meets open sharing criteria*                           | Accessible                   | Full          | Full              | Allowed                |
|                        | • Data application submitted                             |                              |               |                   |                        |
|                        | • DAC approval obtained                                  | Accessible                   | Partial       | Partial           | Not Allowed            |
|                        | • No data application submitted or DAC approval rejected |                              |               |                   |                        |
|                        | • Does not meet open sharing criteria                    | Inaccessible                 | Partial       | Partial           | Not Allowed            |

\*Note: "Meets open sharing criteria" refers to data that has been registered in the Human Genetic Resource Management Platform, or is foreign data, or is data of clinical pathogens, ancient humans/fossils, or commercial cell lines.



**Controlled Access Data**

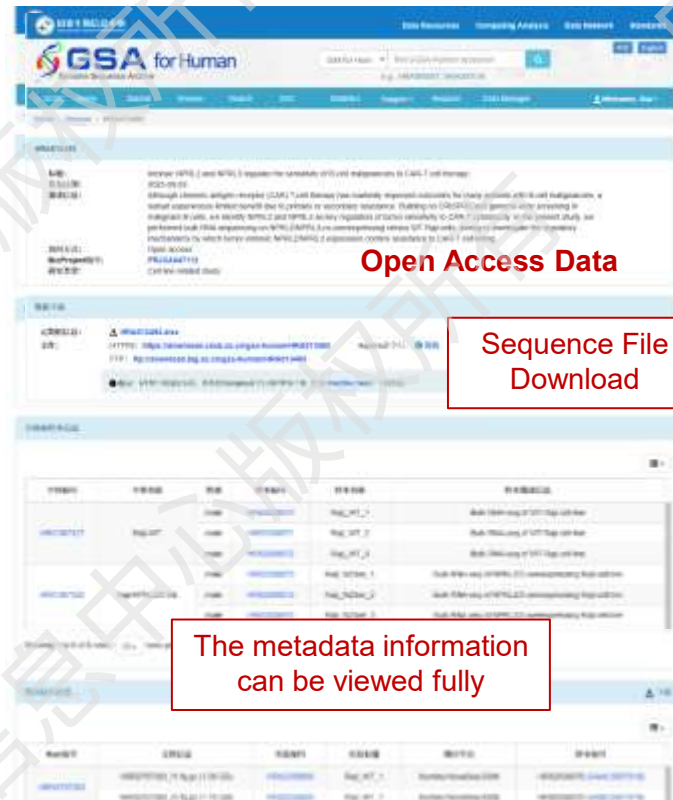
The screenshot shows the GSA for Human website interface. The top navigation bar includes links for 'GSA for Human', 'GSA for Human', 'GSA for Human', and 'GSA for Human'. The main content area is titled 'Controlled Access Data' in red. Below this, there is a detailed explanation of the controlled access process. A red box at the bottom states: 'The metadata information can be viewed partially'.



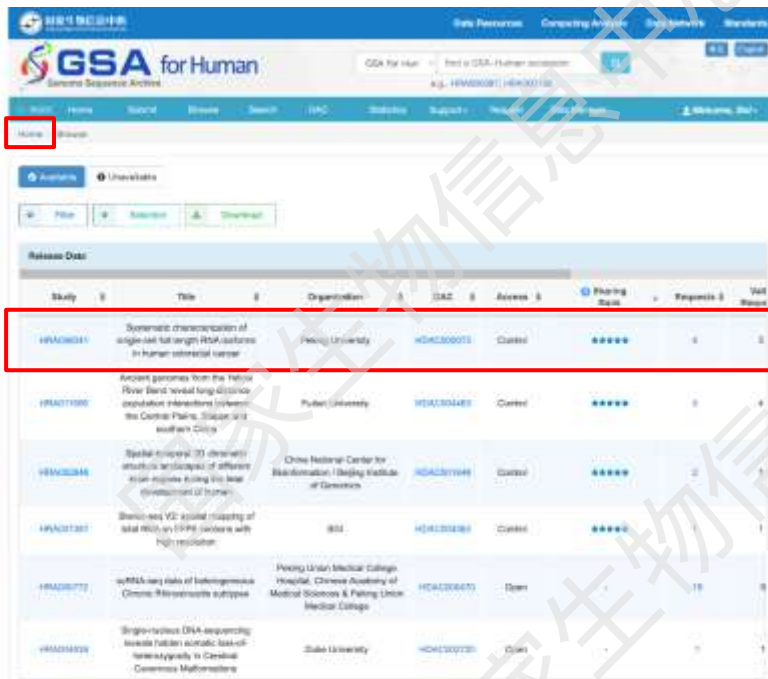
# GSA-Human Data Access Permissions

| Access Type            | Data Status                                                                                | Access Permission Categories | Metadata View | Metadata Download | Sequence File Download |
|------------------------|--------------------------------------------------------------------------------------------|------------------------------|---------------|-------------------|------------------------|
| Open Access Data       | • Meets open sharing criteria*                                                             | Accessible                   | Full          | Full              | Allowed                |
|                        | • Does not meet open sharing criteria                                                      | Inaccessible                 | Partial       | Partial           | Not Allowed            |
| Controlled Access Data | • Meets open sharing criteria*<br>• Data application submitted<br>• DAC approval obtained  | Accessible                   | Full          | Full              | Allowed                |
|                        | • Meets open sharing criteria*<br>• No data application submitted or DAC approval rejected | Accessible                   | Partial       | Partial           | Not Allowed            |
|                        | • Does not meet open sharing criteria                                                      | Inaccessible                 | Partial       | Partial           | Not Allowed            |

\*Note: "Meets open sharing criteria" refers to data that has been registered in the Human Genetic Resource Management Platform, or is foreign data, or is data of clinical pathogens, ancient humans/fossils, or commercial cell lines.

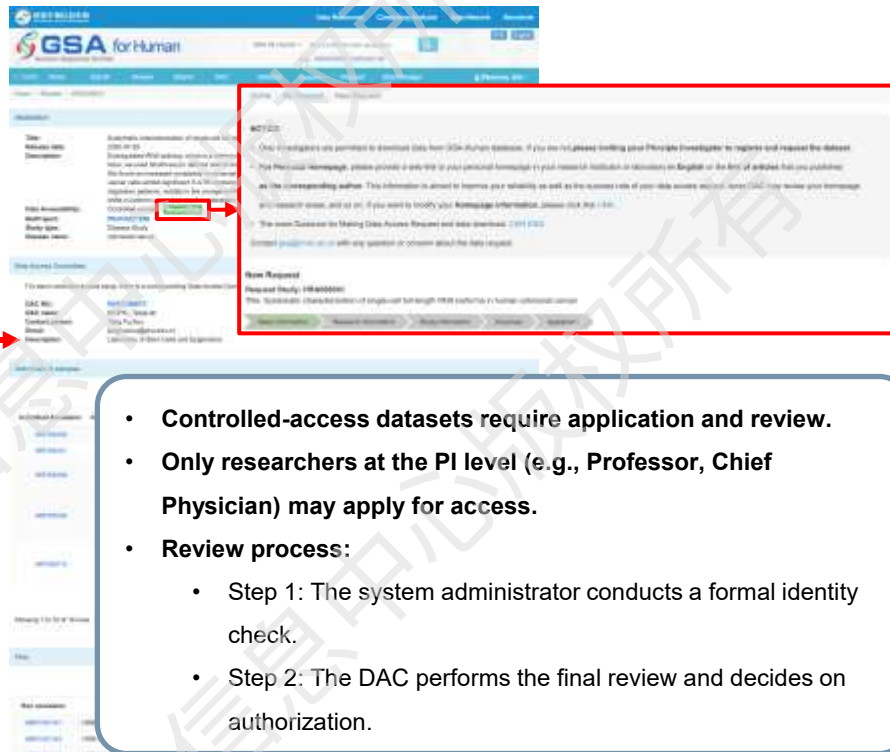


# GSA-Human Data Browse and Access



| Study     | Title                                                                                                                                               | Organization                                                                                              | GAZ       | Access  | Phasing Data | Requests | Wait |
|-----------|-----------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|-----------|---------|--------------|----------|------|
| HPAC0001  | Systematic characterization of origin and full-length HLA sequences in human ancestral human                                                        | Peking University                                                                                         | HGAC00001 | Control | *****        | 0        | 0    |
| HPAC01000 | Ancient genomes from the Neolithic Yellow River reveal long-distance population interactions between the Central Plains, Sogdian and southern China | Peking University                                                                                         | HPAC00040 | Control | *****        | 0        | 0    |
| HPAC00044 | Special temporal (3) chromosome studies and analysis of different human genomes during the last development of human                                | China National Center for Biotechnology / Beijing Institute of Genomics                                   | AGAC01000 | Control | *****        | 2        | 0    |
| HPAC01001 | Genomic V2: whole-genome mapping of total RNA-seq (RNA-seq) with high resolution                                                                    | IBL                                                                                                       | HPAC00040 | Control | *****        | 0        | 0    |
| HPAC00077 | Whole-genome DNA sequencing of human whole-genome (WGS) and whole-genome (WGS) in Chinese                                                           | Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College | HPAC00040 | Open    | *****        | 1        | 10   |
| HPAC00009 | Single-nucleotide DNA sequencing of human whole-genome (WGS) and whole-genome (WGS) in Chinese                                                      | State University                                                                                          | HPAC00077 | Open    | *****        | 1        | 10   |

<https://ngdc.cncb.ac.cn/gsa-human/browse/>



**Request**

One researcher is permitted to download data from GSA-Human database. If you are not a registered user, please register first and request the dataset.

**Request**

One researcher is permitted to download data from GSA-Human database. If you are not a registered user, please register first and request the dataset.

**Request**

One researcher is permitted to download data from GSA-Human database. If you are not a registered user, please register first and request the dataset.

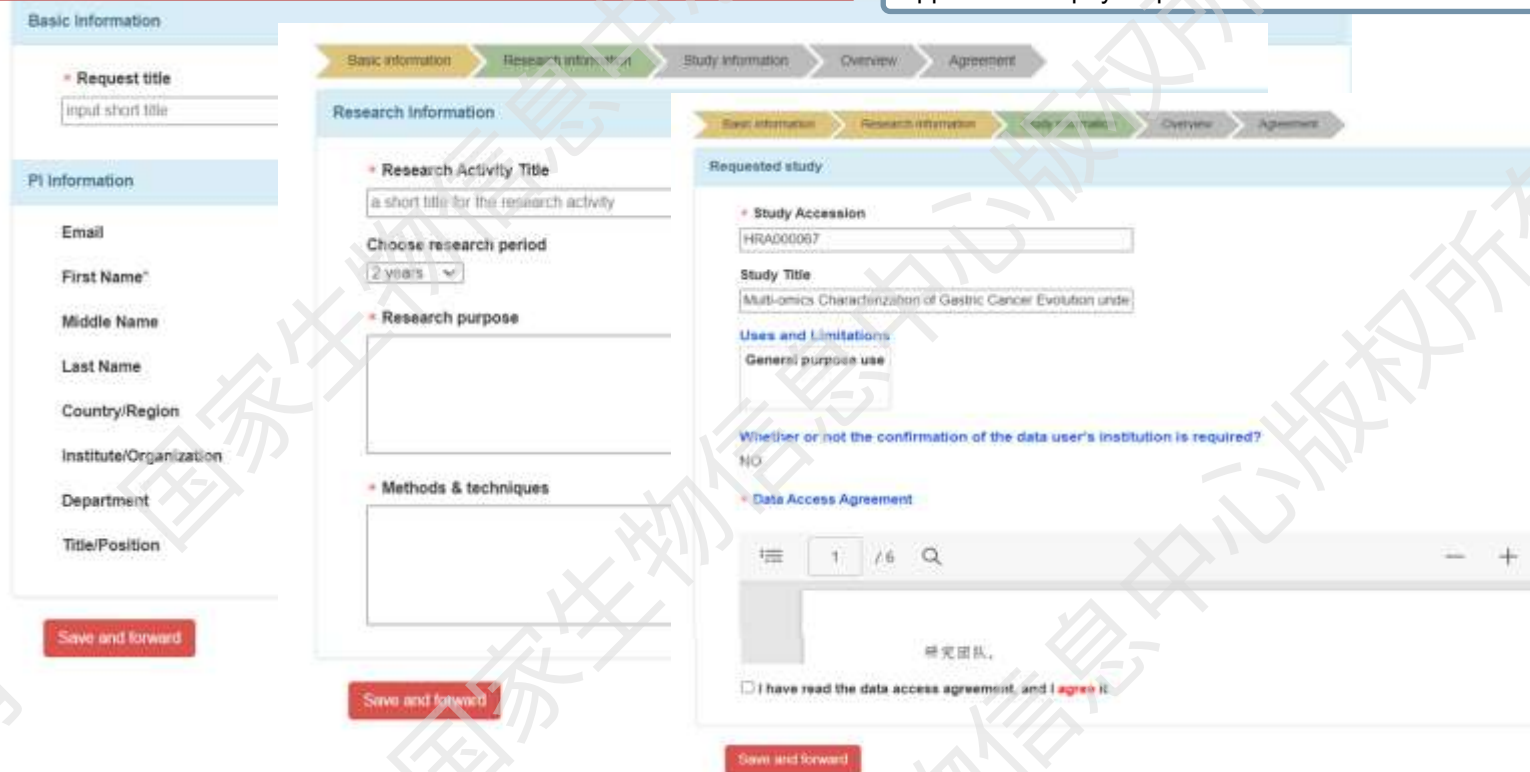
- Controlled-access datasets require application and review.
- Only researchers at the PI level (e.g., Professor, Chief Physician) may apply for access.
- Review process:
  - Step 1: The system administrator conducts a formal identity check.
  - Step 2: The DAC performs the final review and decides on authorization.





# Controlled-Access Data Application

A wizard that helps you to make a data application step by step



**Basic information**

Request title  
input short title

**PI Information**

Email  
First Name\*  
Middle Name  
Last Name  
Country/Region  
Institute/Organization  
Department  
Title/Position

Save and forward

**Research information**

Research Activity Title  
a short title for the research activity

Choose research period  
2 years

Research purpose

Methods & techniques

Save and forward

**Study information**

Requested study

Study Accession  
HRA000067

Study Title  
Multi-omics Characterization of Gastric Cancer Evolution under

Uses and Limitations  
General purpose use

Whether or not the confirmation of the data user's institution is required?  
NO

Data Access Agreement

1 / 6

研究团队,

☐ I have read the data access agreement, and I agree it

Save and forward



# Controlled-Access Data Application

Basic Information

Request Title

Personal Information

Email

First Name

Middle Name

Last Name

Country/Region

Institute/Organization

Department

Title/Position

Personal Homepage / Corresponding Fulltext URL

Institute / Organization Homepage

Research Information

Research Activity Title

Research purpose

Methods & techniques

## Step4 Overview

It summarizes all previously entered information—**Basic Information, Personal Information, Research Information, and Study Selection**—displays notices and contact details, provides the **Data Access Agreement** for review and consent, and lets you proceed via **Save and forward**.

Study Selection

Study Selection Information

Study Title

Users and Limitations

Whether or not the verification of the data user's institution is required?

Data Access Agreement

Save and forward



# Controlled-Access Data Application

Basic information Research information Study information Overview Agreement

Agreement

☐ I have read the following GSA-Human data policy, and I agree it.

**1 Rights of data user**

- 1) The data user can make applications for the data needed through the GSA-Human system.
- 2) The data user can download the data in certain time after obtaining the approval of accessing the dataset.
- 3) The data user has the intellectual property of the subsequent research result based on the dataset.
- 4) The data user can download the approved data by tools like ftp or Aspera, which is designated by GSA-Human.

**2 Responsibilities of data user**

- 1) The data user should use the downloaded dataset in the promised research scope of application. The dataset is for research purpose, and can only be used for the research group and its research collaborators. Dataset distribution to other persons cannot be used to identify individuals.
- 2) The data user should verify the quality, content and scientific validity of the downloaded dataset by his/her own.
- 3) If the data user publish a paper using the downloaded dataset from GSA-Human, the data accession number must be included with the note that the data can be acquired from GSA-Human.

**Choose signature type**

☐ Online signature

☐ Offline signature

☒ Online signature

- ☐ Upload a signature picture (recommend size: 120px \* 50px)
- ☐ Generate signature by using mouse
- ☐ Create a signature automatically

Save and forward

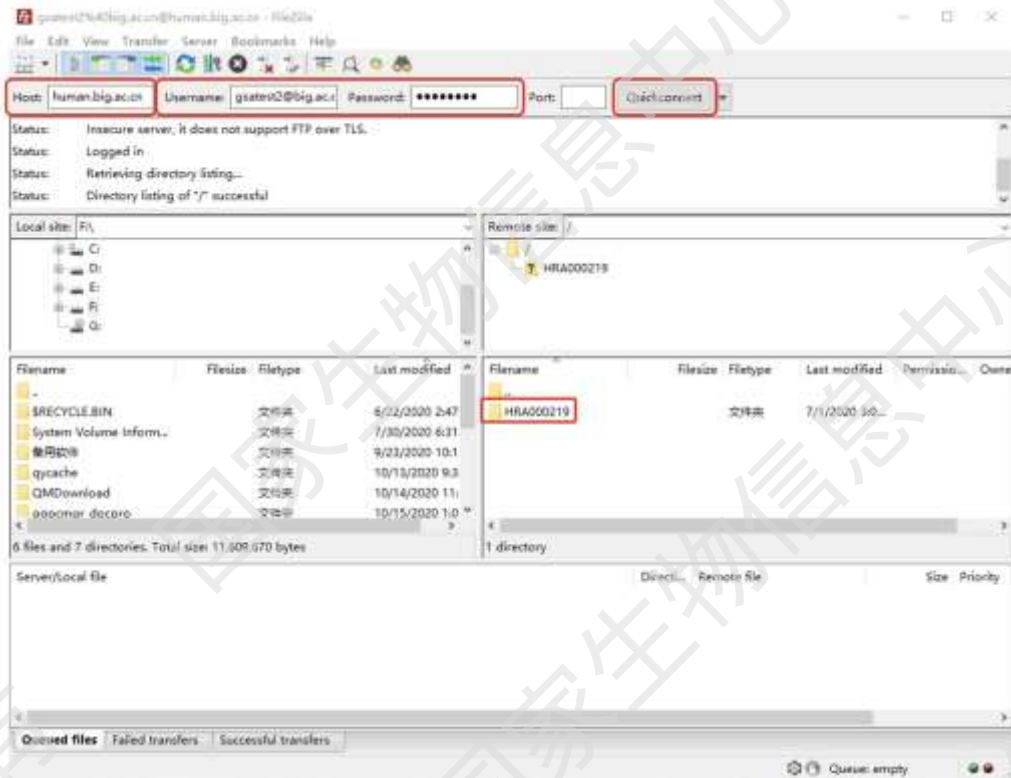
## Step5 Agreement

- **Check to agree:** Read the GSA-Human data policy and tick "I agree."
- **Choose a signature method (required):**
  - **Online signature:**
    - Upload a signature image (recommended), or
    - Draw a signature with the mouse, or
    - Generate a signature automatically;
  - **Offline signature:** Sign offline.

**NOTICE: If the data submitter additionally requires the user's institution to sign the DAA, the agreement must be completed offline with signatures and an official institutional stamp.**



# Download the Approved Data



## Quick Guide (FileZilla Download & Transfer)

- Download & install FileZilla Client from the official site



- Open FileZilla: fill in the port and click “Quickconnect”
  - Set Host: [human.big.ac.cn](http://human.big.ac.cn);
  - User/Password: the PI account that made the application;
- In the local pane, choose the destination folder for your downloads
- In the remote pane, open the your authorized dataset folder(s) (e.g., HRA000219) and drag files/folders to the local pane to download



# Contents

- 1 CNCB Data Submission System Overview
- 2 GSA Data Submission & Sharing
- 3 GSA-Human Data Submission & Sharing
- 4 GWH Data Submission & Sharing**
- 5 GenBase Data Submission & Sharing
- 6 OMIX Data Submission & Sharing

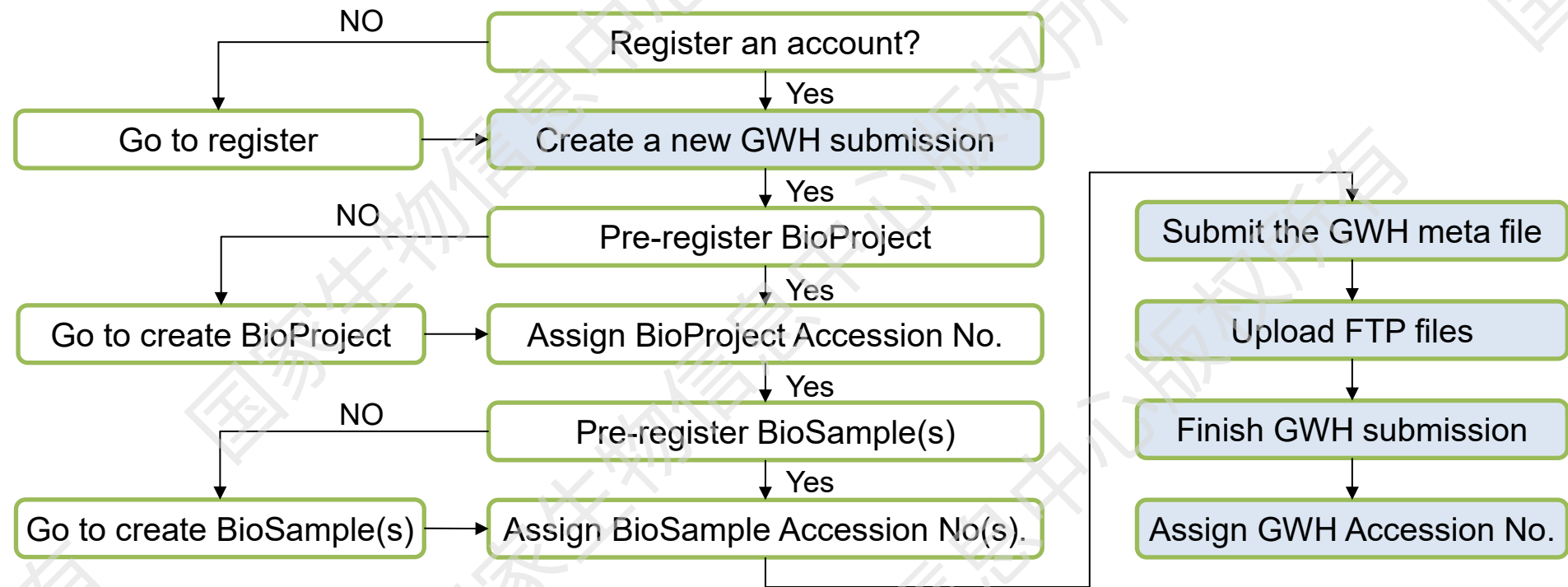


# What Types of Data does GWH Accept?

- GWH accepts **full genomes of eukaryote, prokaryote** (including/excluding organelle(s), plasmid(s), etc.) and **metagenomes** data
- You can submit all genome assembly data associated with the **same BioProject and publication** simultaneously in one batch submission

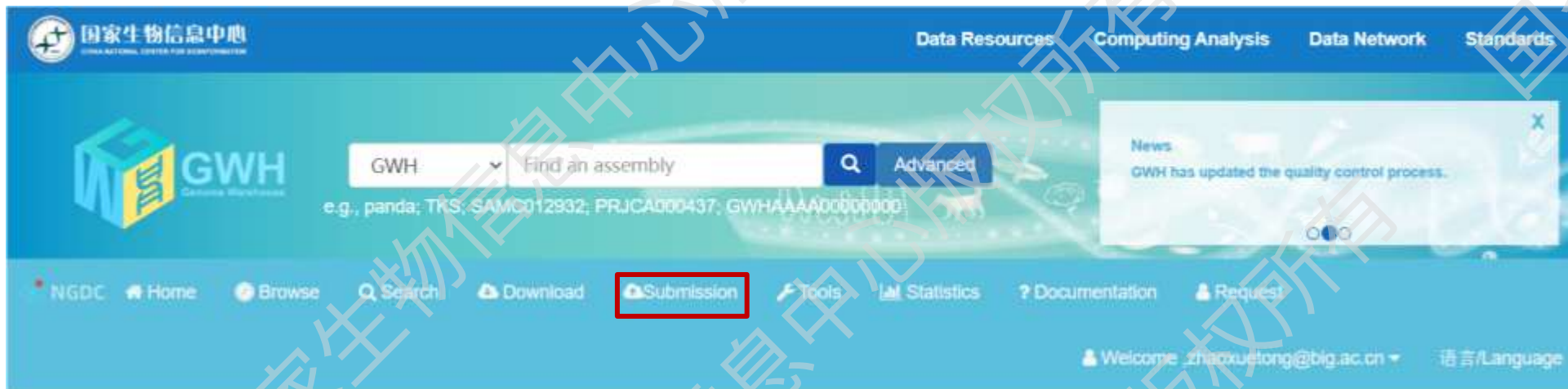


# GWH Submission Pipeline





# Preparation 1: Metadata Files



**You can download the metadata file in here:**

**Note:** GWH accepts genome associated data submission, but we are not responsible for the accuracy of the submitted data. The submitter should be responsible for the data accuracy.

1. Download GWH batch submission template file [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#), then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#).
3. Download the GWH assignment information template file [chr.csv](#) / [plasmid.csv](#) / [organelle.csv](#) and documentation, fill in and double-check it before upload it through FTP (optional).
4. FAQ for GWH submission, please see the [pdf](#).
5. For more information, please see the [Help](#).



The latest version of the GWH-batchsubmission-English.xlsx template file

**Note: The explanations for the filled-in items can be found in the instructions.**

**Note: Do not delete or insert new rows or columns.**

# Preparation 2: Data Files

## ➤ Genome assembly file (Required, fasta format)

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

Sequence ID:

- (1) Start with a letter
- (2) Letters, numbers, - \_ . : \* #

A genome only has:

- One genome assembly file
- One genome annotation file

## ➤ Genome annotation file (Not required, gff/tbl format)

### ■ gff format

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

For example,

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



# Preparation 3: Assignment Information File

- Assignment Information File (When the assembly level is complete or draft in chromosome, chromosome assignment file is required)

chromosome assignment file

| Sequence ID | Chromosome name | Complete | Circular |
|-------------|-----------------|----------|----------|
| seq_1       | 1               | true     | false    |

organelle assignment file

| Sequence ID | Type          | Complete | Circular |
|-------------|---------------|----------|----------|
| seq_2       | Mitochondrion | true     | false    |

plasmid assignment file

| Sequence ID | Plasmid name | Complete | Circular |
|-------------|--------------|----------|----------|
| seq_3       | pBR322       | true     | false    |

- The Sequence ID in the assignment file must come from the Sequence ID in the genome sequence file. For example, if the genome sequence file contains: ">ctg1 chromosome1", then the sequence ID is "ctg1".
- In the chromosome assignment file, "Complete=true" indicates that the sequence represents a certain chromosome. "Complete=false" means that is associated with a certain chromosome but has not been localized to a specific position on the chromosome.



# Online Documentation about the Preparations

Genome data submission preparation

GenBank accepts whole genome assembly data (including/excluding organelle(s), plasmids(s), etc.). Individual organelle/plasmid genome, viral genome, and gene fragment sequence data are submitted to the [GenBank](#) database of the NCBI. You can submit all genome assembly data associated with the same BioProject and publication simultaneously in one batch submission.

● Submission process description (Hide)

- (1) Sign up an account and Login.
- (2) Submit meta data:
  - Create a new submission
  - Fill in step1-3
  - Upload "Batch Meta excel File" in step4 and validate it until it passes the online quality control.
  - Go to the overview page in step5 and click the "Finish" button to complete the upload.
- (3) Upload files through [ftp](#).
- (4) Returns to the submission list <https://ncbi.nlm.nih.gov/submit/submission>, click the "FTP finish upload" button of the corresponding Batch ID. After that, the system will trigger automatic quality control system to audit it. Please do not click repeatedly to avoid exception.
- (5) Wait for the data audit, and the audit result will be informed by email.
  - If the submit files pass quality control system, the submission status will turn into "Successful". And GenBank will assign a unique genome assembly accession No. for each submitted genome assembly in a batch submission.
  - On the contrary, the submission status will turn into "Error". And GenBank will send error report to submitter through email. Submitter needs to check and be sure that all errors in the attached report files are fixed, and resubmits the corrected files at your convenience.

● Prepare the following information:

1. BioProject and BioSample accession number
2. Batch Meta File
3. Genome Sequence File
4. Genome Annotation File (optional)
5. Assignment Information File (required for certain condition)
6. Uploading Files through FTP
7. Metagenome data submission
8. Haplotype data submission

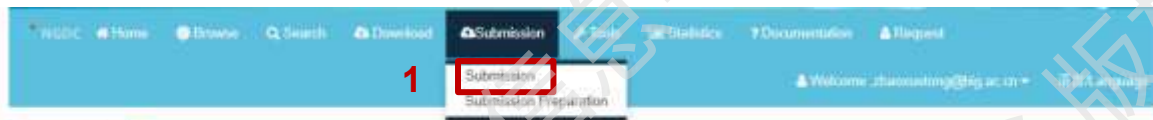
Detailed description of the submission process →

Prepare the following information: →



# Create a New Submission

A batch submission is required to be associated with the same BioProject and the same article



New submission

Note: GWH accepts genome associated data submission, but we are not responsible for the accuracy of the submitted data. The submitter should be responsible for the data accuracy.

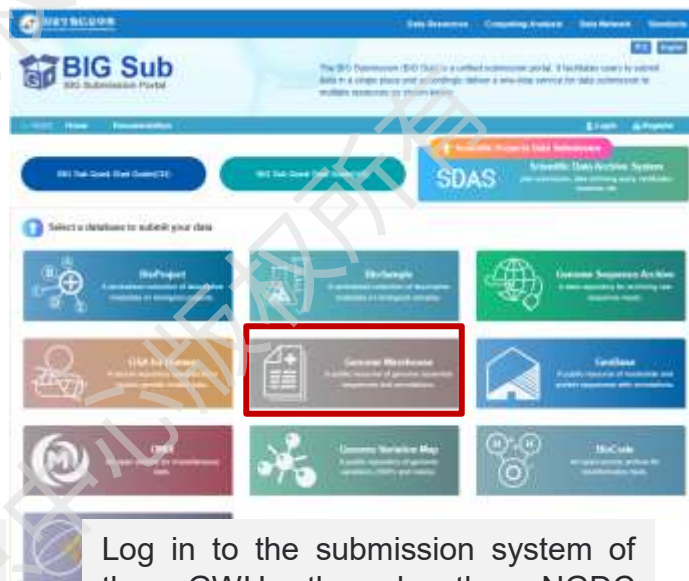
1. Download GWH batch submission template file *GWH\_Assembly\_en.xlsx* / *GWH\_Assembly\_ch.xlsx*, then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. *GWH\_Assembly\_en.xlsx* / *GWH\_Assembly\_ch.xlsx*.
3. Download the GWH assignment information template file *chr.csv* / *plasmid.csv* / *organism.csv* and documentation, fill in and double-check it before upload it through FTP (optional).
4. FAQ for GWH submission, please see the [pdf](#).
5. For more information, please see the [Help](#).

2 My Submissions Deleted Submissions

8 Submissions

| Accession(s) | Batch ID | Title | BioProject | Release date | Created    | Updated    | Status | Operation(s)             |
|--------------|----------|-------|------------|--------------|------------|------------|--------|--------------------------|
| 202400010001 | 000      | 000   | 0000000000 | 2024-07-24   | 2024-07-24 | 2024-07-24 | 00000  | <a href="#">Batch ID</a> |

Log in to the submission system through the GWH  
<https://ngdc.cncb.ac.cn/gwh/submit/submission>



Log in to the submission system of  
the GWH through the NGDC  
submission  
<https://ngdc.cncb.ac.cn/gsub/>





# Step 1. Submitter

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

**Submitter**

|                                              |                                                                                            |                      |
|----------------------------------------------|--------------------------------------------------------------------------------------------|----------------------|
| First name<br>zhao                           | Middle name                                                                                | Last name<br>xuetong |
| Email<br>zhaoxuetong@big.ac.cn               | Secondary Email (The recipients of feedback email such as students)<br>13027751205@163.com |                      |
| Organization<br>Academy of Mathematics an... | Organization URL<br>https://                                                               | Department<br>NGDC   |
| Phone                                        | Fax                                                                                        |                      |
| Street<br>Beichen West Road, Chaoyang        | City<br>Beijing                                                                            | State/Province       |
| Postal code<br>100010                        | Country/Region<br>China                                                                    |                      |

Save and next

Note: The email address of the Submitter will be used for information communication during the subsequent submission process. If necessary, please also fill in the Secondary email address to be copied.





## Step 2. General Info

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

### General Info

**Submission Title:**

### Release Date

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

**Note:** Release of BioProject and BioSample is also triggered by the released date.

### Release Policies and Disclaimers

1. A date can be set by authors to withhold the release of new submissions.
2. The release date can be changed through the genome portal (<https://www.genome.cn/>).
3. If a paper citing the sequence or accession number is published prior to the release date on the specified date.
4. As soon as it is available, please send the full publication data—all authors, title, journal, volume, pages and date—to the following address: [publications@cnj.ac.cn](mailto:publications@cnj.ac.cn).

☒ I accept it ☐ I don't accept it

Note: The title is only used by users to distinguish multiple submissions in the submission list, and will never be made public

Note: Please pay attention to the Release Policies and Disclaimers

### BioProject

**BioProject accession:**  [Click here to create BioProject](#)  
The BioProject bundles the data for this research project, which is an overall description of a single research initiative.

**Note:** If you have not created related BioSample, go to create [BioSample](#) (fill it in the meta-information box of the sample).

### Data Accessibility

**Data Accessibility:**  
☒ Open-access data (default) ☐ Controlled-access data

**Note:** controlled access only work for human genome data resources, and the release of human genome data, regardless of the data access method, requires completion of backup in HGDP and registration number in HRC.

[View and read](#)

Note: Controlled access is used for human genome data resources only



# Step 3. Reference

Step 1: Submission | Step 2: General info | **Step 3: Reference** | Step 4: Note | Step 5: Overview

**Sequence authors**

| First name | Middle name | Last name | Email               |
|------------|-------------|-----------|---------------------|
| zhao       |             | xuefeng   | zhaoxuefeng1@big... |

**Contact persons**

| First name | Middle name | Last name | Email               |
|------------|-------------|-----------|---------------------|
| zhao       |             | xuefeng   | zhaoxuefeng2@big... |

- **Sequence author:** Only the name is disclosed at the time of data release, not the email address
- **Contact person:** Publicly disclose the name and email at the time of data release

**Reference**

Publication status: ☐ Unpublished

Publication Source: ☒ PubMed ID: 12345

Reference details for:  
Title: A new granulation method for compressed tablets [proceedings]  
Author: Rubinstein MH  
Journal: The Journal of pharmacy and pharmacology

Is this PubMed info? ☐ Yes ☐ No

☐ LIFE with DOI  
☐ Publication Detail Information

Save and next

Note: Here is the information of the article that published the genomic accession of this submission

Note: Please confirm whether the detailed information provided corresponding to the PubMed ID is accurate.

## Step 4. Meta

Upload the meta-information file (multiple genomes can be submitted in one batch) and verify it online

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

### Batch Meta Submission

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

20240528-1.xlsx

upload validate

Note:

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

Save and Next

fa

X

①Select a local file and upload it

②After successful upload, verify online

fail

pass

[illegible]

Error message:

- ① Download the file
- ② View online

- **Error message:** must be modified and resubmitted
- **Warning message:** can choose to make modified and resubmitted

# Step 5. Overview



## Overview

Batch submission: Batch0355846

Title: xxx

Status: Overview

This WGS submission will be released on 2025-04-07 (-175 days from now) or publication, depending on which is first.

Note: Release of BioProject and BioSample is also triggered by the release of WGS data.

### FTP submission

You can transfer your data files to the GWH FTP site using the following credentials:

- Host address: submit.big.ac.cn
- User name: zhaoxutong@big.ac.cn
- Password: the same to the user password that you log in GWHBIOID, eg: 123456
- Path: /GWH/Batch0355846

Note: Use FTP client software (such as FileZilla Client) to log in FTP.

After files submission, please click the "FTP File(s) upload" button in your submission list (<https://ngdc.cncb.ac.cn/gwh/submits/submission>) to inform us to check your submission.

FTP account and the file upload location

### Submitter

- Submitter: zhaoxutong
- Email: zhaoxutong@big.ac.cn
- Submitting organization: National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation

### General info

- BioProject accession: PRJCA38205
- Release date: 2025-04-07
- Is controlled access: false

### Reference

- Sequence authors: zhaoxutong
- Publication status: unpublished
- Reference title: xxx
- Reference authors: zhaoxutong
- Contact authors: zhaoxutong (zhaoxutong@big.ac.cn)

### Attribution

Showing 1 to 1 of 1 entries

Search:

| BioSample accession | Organism | Assembly name | Assembly level | Genome sequence file name | Genome annotation file name | Reference access |
|---------------------|----------|---------------|----------------|---------------------------|-----------------------------|------------------|
| SAMCUB001           |          | eggs_G1ha     | chromosome     | genome.final.genome.fa    | N/A                         | N/A              |

Show:

A description of the unique meta-information attributes of each genome assembly



# Step 6. Return to the Submission List

After the Meta information is submitted online, return to the submission list

My Submissions

Deleted Submissions

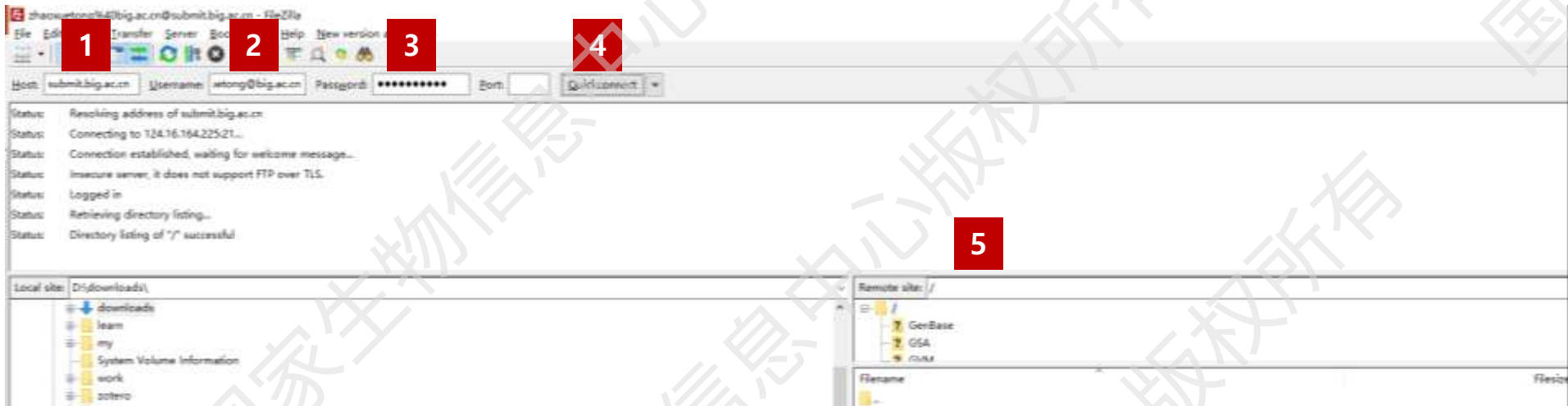
Search:

8 Submissions

| Accession(s) ↑↓ | Batch ID ↑↓                  | Title ↑↓ | BioProject ↑↓ | Release date ↑↓ | Created ↑↓ | Updated ↑↓ | Status ↑↓         | Operation(s) ↑↓                                             |
|-----------------|------------------------------|----------|---------------|-----------------|------------|------------|-------------------|-------------------------------------------------------------|
|                 | <a href="#">Batch0031050</a> | abc      | PRJCA001233   | 2024-07-31      | 2024-04-23 | 2024-04-23 | ERROR             | <a href="#">Download Error Zip</a>                          |
|                 | <a href="#">Batch0088496</a> | ddd      | PRJCA008260   | 2024-05-31      | 2024-05-24 | 2024-05-24 | Waiting for files | <a href="#">FTP Finish Upload</a><br><a href="#">Delete</a> |
|                 | <a href="#">Batch0088499</a> | cc       | PRJCA008260   | 2024-05-31      | 2024-05-24 | 2024-05-24 | Waiting for files | <a href="#">FTP Finish Upload</a><br><a href="#">Delete</a> |



# Step 7. Upload the Data Files via FTP



Upload your data files to the GWH FTP site:

1. Host address: submit.big.ac.cn
2. User name: the same as the username used when logging in to GWH, eg: chenml@big.ac.cn
3. Password: the same to the user password that you log in GWH, eg:123456
4. Click "Quick Connection" (port 21)
5. Path: **/GWH/Batch00xxxxx** , eg:/GWH/Batch0090866



# Step 8. Trigger Automatic Check

- Return to the submission list
- Click the "Finish upload" button corresponding to the submission

New submission

**Note:** GWH accepts genome associated data submission, but we are not responsible for the accuracy of the submitted data. The submitter should be responsible for the data accuracy.

1. Download GWH batch submission template file [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#), then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#).
3. Download the GWH assignment information template file [chr.csv](#) / [plasmid.csv](#) / [organella.csv](#) and documentation, fill in and double-check it before upload it through FTP (optional).
4. FAQ for GWH submission, please see the [pdf](#).
5. For more information, please see the [Help](#).

My Submissions

Deleted Submissions

Search:

8 Submissions

| Accession(s) | Batch ID                     | Title | BioProject  | Release date | Created    | Updated    | Status            | Operation(s)                                                |
|--------------|------------------------------|-------|-------------|--------------|------------|------------|-------------------|-------------------------------------------------------------|
|              | <a href="#">Batch0031050</a> | abc   | PRJCA001233 | 2024-07-31   | 2024-04-23 | 2024-04-23 | ERROR             | <a href="#">Download Error Zip</a>                          |
|              | <a href="#">Batch0088496</a> | ddd   | PRJCA008260 | 2024-05-31   | 2024-05-24 | 2024-05-24 | Waiting for files | <a href="#">FTP Finish Upload</a><br><a href="#">Delete</a> |
|              | <a href="#">Batch0088499</a> | cc    | PRJCA008260 | 2024-05-31   | 2024-05-24 | 2024-05-24 | Waiting for files | <a href="#">FTP Finish Upload</a><br><a href="#">Delete</a> |

**Note:** Please click the operation button to trigger the system's automatic quality control and archiving.

Remember to check your email for the system's automatic feedback on quality control errors or passes.





# Step 8. Data Validation and Feedback

➤ **If failed:** please check the attached file that reports the errors



- ① Please resubmit the metadata/data files after modifying the errors report in the attachment
- ② Modify/resubmit under the original submission number
- ③ Recalculate the MD5 code of the data file if it was modified
- ④ Re-upload the modified data file in FTP
- ⑤ Return to the submission list and click the "Finish upload" button

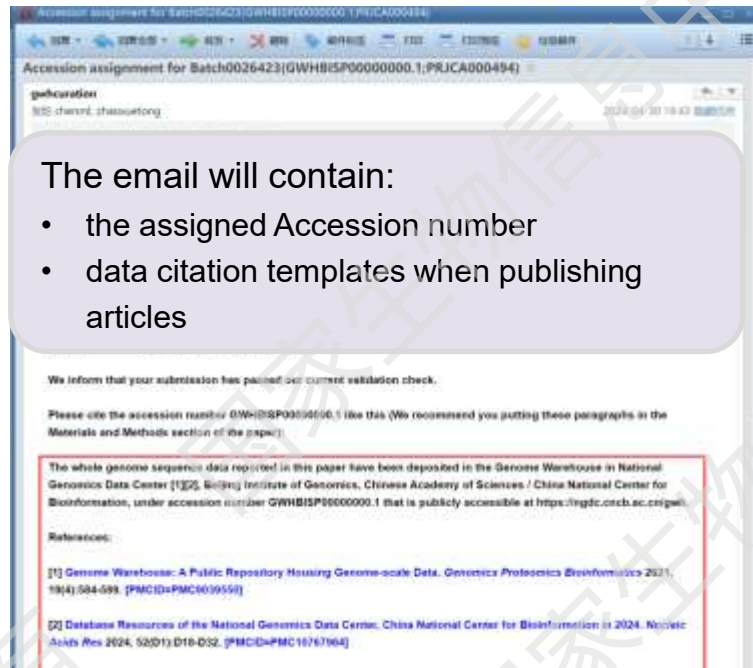
The screenshot shows the "My Submissions" table in the submission portal. The table has columns: Accession(s), Batch ID, Title, BioProject, Release date, Created, Updated, Status, and Operation(s). The first row shows a submission with Accession(s) SRR111111, Batch ID Batch001000, Title abc, BioProject PRJCA001233, Release date 2024-07-31, Created 2024-04-23, Updated 2024-04-23, and Status ERROR. The Status column has a red box around the word "ERROR". The Operation(s) column has a button labeled "Download Error File".

| Accession(s) | Batch ID    | Title | BioProject  | Release date | Created    | Updated    | Status | Operation(s)        |
|--------------|-------------|-------|-------------|--------------|------------|------------|--------|---------------------|
| SRR111111    | Batch001000 | abc   | PRJCA001233 | 2024-07-31   | 2024-04-23 | 2024-04-23 | ERROR  | Download Error File |



# Step 8. Data Validation and Feedback

➤ **If passed:** obtain the accession number



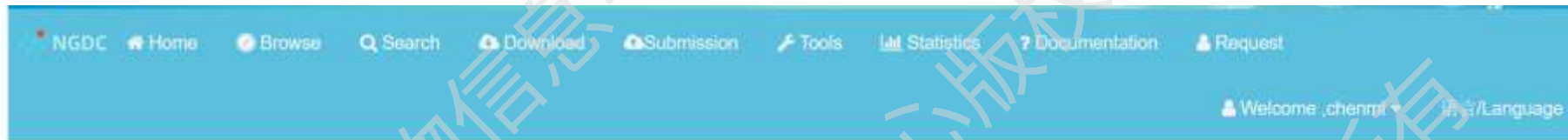
83 Submissions

| Accession(s)      | Batch ID     | Title                              | BioProject   | Release date | Created    | Updated    | Status     | Op                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                            |
|-------------------|--------------|------------------------------------|--------------|--------------|------------|------------|------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| GWHBIS000000000.1 | Batch0018064 | SARS-CoV-2 human_CHN_Wuhan_SS_2020 | PRJCA0004241 | 2024-12-18   | 2024-04-23 | 2024-04-23 | SUCCESSFUL | <div>3</div> <div>1</div> <div>2</div> <div>3</div> <div>4</div> <div>5</div> <div>6</div> <div>7</div> <div>8</div> <div>9</div> <div>10</div> <div>11</div> <div>12</div> <div>13</div> <div>14</div> <div>15</div> <div>16</div> <div>17</div> <div>18</div> <div>19</div> <div>20</div> <div>21</div> <div>22</div> <div>23</div> <div>24</div> <div>25</div> <div>26</div> <div>27</div> <div>28</div> <div>29</div> <div>30</div> <div>31</div> <div>32</div> <div>33</div> <div>34</div> <div>35</div> <div>36</div> <div>37</div> <div>38</div> <div>39</div> <div>40</div> <div>41</div> <div>42</div> <div>43</div> <div>44</div> <div>45</div> <div>46</div> <div>47</div> <div>48</div> <div>49</div> <div>50</div> <div>51</div> <div>52</div> <div>53</div> <div>54</div> <div>55</div> <div>56</div> <div>57</div> <div>58</div> <div>59</div> <div>60</div> <div>61</div> <div>62</div> <div>63</div> <div>64</div> <div>65</div> <div>66</div> <div>67</div> <div>68</div> <div>69</div> <div>70</div> <div>71</div> <div>72</div> <div>73</div> <div>74</div> <div>75</div> <div>76</div> <div>77</div> <div>78</div> <div>79</div> <div>80</div> <div>81</div> <div>82</div> <div>83</div> |

- ① Display the **assigned Accession number**
- ② Display the current data status as “**Successful**”
- ③ Supported operations: **Generate reviewer links, Release the data immediately, Modify the publication date** and so on



# How to Fix the Errors?



New submission

**Note:** GWH accepts genome associated data submission, but we are not responsible for the accuracy of the submitted data. The submitter should be responsible for the data accuracy.

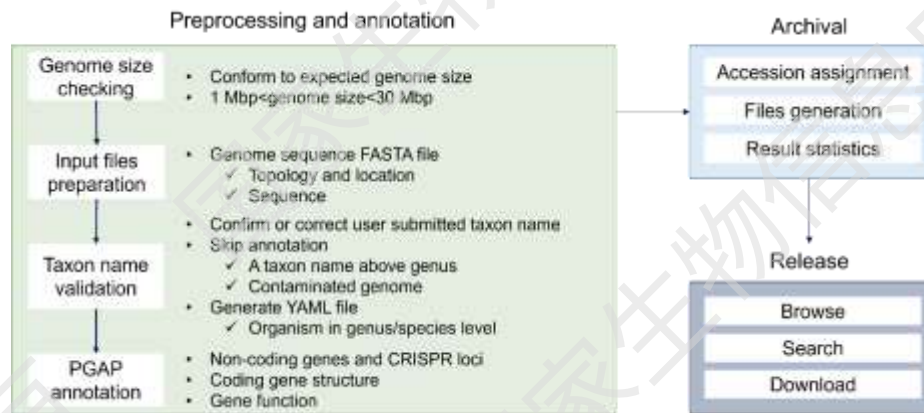
1. Download GWH batch submission template file [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#), then fill in and double-check it before uploading.
2. For column explanations and examples, please see the e.g. [GWH\\_Assembly\\_en.xlsx](#) / [GWH\\_Assembly\\_ch.xlsx](#).
3. Download the GWH assignment information template file [chr.csv](#) / [plasmid.csv](#) / [organella.csv](#) and [documentation](#), fill in and double-check it before upload it through FTP (optional).
4. FAQ for GWH submission, please see the [pdf](#).
5. For more information, please see the [Help](#).

**Please click here to see the manuals**



# Prokaryotic Genome Reannotation

Genome annotation is the basic of functional genomics research. GWH establishes a prokaryotic genome reannotation resource to support scientific research. It can effectively enhance the utilization value of submitted data.



The process of reannotation and sharing of prokaryotic genomes in GWH

## Features

- Unified and standardized** pipeline: 1,108 organisms, 9,646 genome assemblies
- Improved annotated percentage: 1.5% -> **88.4%**
- Various** genomic features: protein-coding genes, tRNA, rRNA, sRNA, CRISPR elements, pseudogenes
- Completeness:** gene structure, protein products, gene name and gene function, etc.

- Browse with multi-condition filters



<https://ngdc.cncb.ac.cn/qwh/search/advanced>

# Access the Prokaryotic Genome Reannotation Resources

- Browse page

Scientific name:  Common name:  Group:  Species:  Accession:  Genomic Annotation:  See Reannotation:

Genome representation:  Assembly Level:  Genome size (Mb):  GC content (%):  Release date:  Source release date:

Showing 1 to 10 of 3,701 entries

| Genome representation | Assembly Level | Genome size (Mb) | GC content (%) | Release date | Source release date | Download Reannotation                                                                                   |
|-----------------------|----------------|------------------|----------------|--------------|---------------------|---------------------------------------------------------------------------------------------------------|
| some                  | Scaffold       | 2.7325           | 33.12          | 2023-07-20   | 2023-07-20          | <a href="#">DNA</a> <a href="#">GFF</a> <a href="#">RNA</a> <a href="#">CDS</a> <a href="#">Protein</a> |
| some                  | Scaffold       | 1.68M            | 64.45          | 2023-07-20   | 2023-07-20          | <a href="#">DNA</a> <a href="#">GFF</a> <a href="#">RNA</a> <a href="#">CDS</a> <a href="#">Protein</a> |
| some                  | Scaffold       | 5.8320           | 64.19          | 2023-07-20   | 2023-07-20          | <a href="#">DNA</a> <a href="#">GFF</a> <a href="#">RNA</a> <a href="#">CDS</a> <a href="#">Protein</a> |
| some                  | Scaffold       | 4.1188           | 58.81          | 2019-10-13   | 2019-10-13          | <a href="#">DNA</a> <a href="#">GFF</a> <a href="#">RNA</a> <a href="#">CDS</a> <a href="#">Protein</a> |

<https://ngdc.cncb.ac.cn/gwh/browse/assembly>

- Detailed information of an Assembly

Scientific Name: *Acetobacterium sticklandii*  
Common Name: *Clostridium sticklandii*  
Bioproject: [PRJCA018239](#)  
Biosample: [SAMC2923546](#)  
Accession No.: GWHDPN000000000  
GWH reannotation accession: GWHRA\_DPCN000000000.1  
GSA raw reads: -  
GVM variations of the species: -  
Submitter Organization: Marine life science, Ocean University of China  
Contact: Han Cui, cuihan@oou.edu.cn  
Sequence author(s): Han Cui  
Released Date: 2023-07-21  
Assembly Level: Draft genome in scaffold level  
Genome Representation: Full Genome  
Assembly method: maxbin 2.2.6  
metabat 0.32.5  
concoct 1.1  
Sequencing & coverage: Illumina 10  
Nanopore 30

Download: [DNA](#) [Reannotation DNA](#) [Reannotation GFF](#) [Reannotation RNA](#) [Reannotation Protein](#)

Publication(s): -

<https://ngdc.cncb.ac.cn/gwh/Assembly/64710/show>



# Contents

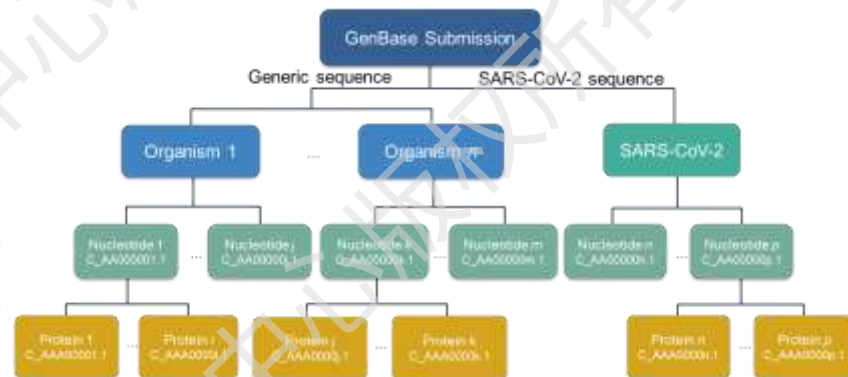
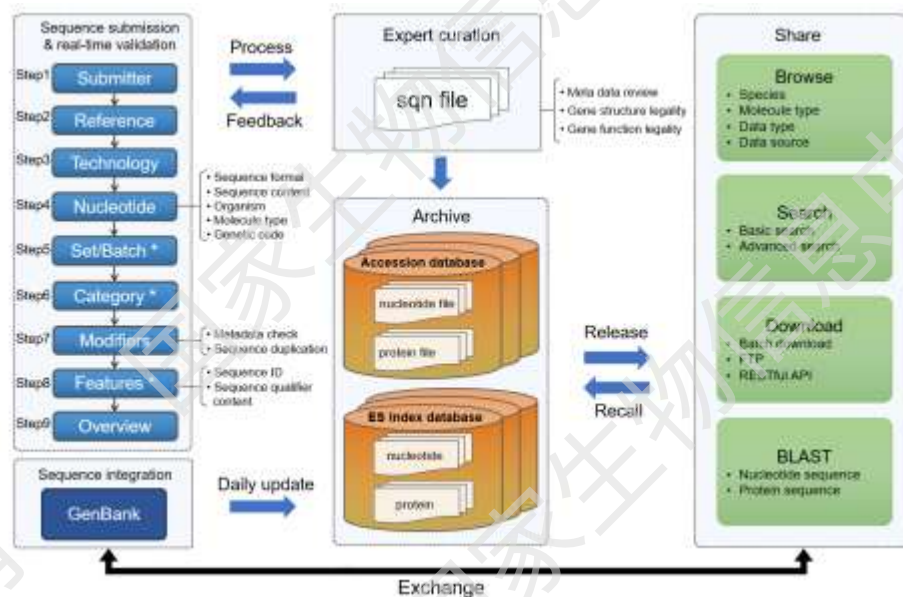
- 1 CNCB Data Submission System Overview
- 2 GSA Data Submission & Sharing
- 3 GSA-Human Data Submission & Sharing
- 4 GWH Data Submission & Sharing
- 5 GenBase Data Submission & Sharing**
- 6 OMIX Data Submission & Sharing





# GenBase Submission System

A bilingual submission system with online real-time validation, alongside expert curation to ensure the quality of archived data



INSDC Compatible Data Model

# Preparation 1-Fasta File

- FASTA files can contain one or more sequences and can be consolidated into a single file for upload
- Please use the FASTA format, starting with a definition line followed by sequence lines
- The simplest definition line requires a ">" symbol and a Sequence ID

**Sequence ID**

For example:

```
>Seq1 [organism=Homo Sapiens] Definition Line for Seq1  
aaccgatatagagagagga  
>Seq2 [organism=Homo Sapiens] Definition Line for Seq2  
atctgaatagagattattt
```

Note: The definition line only recognizes the Sequence ID and organism information

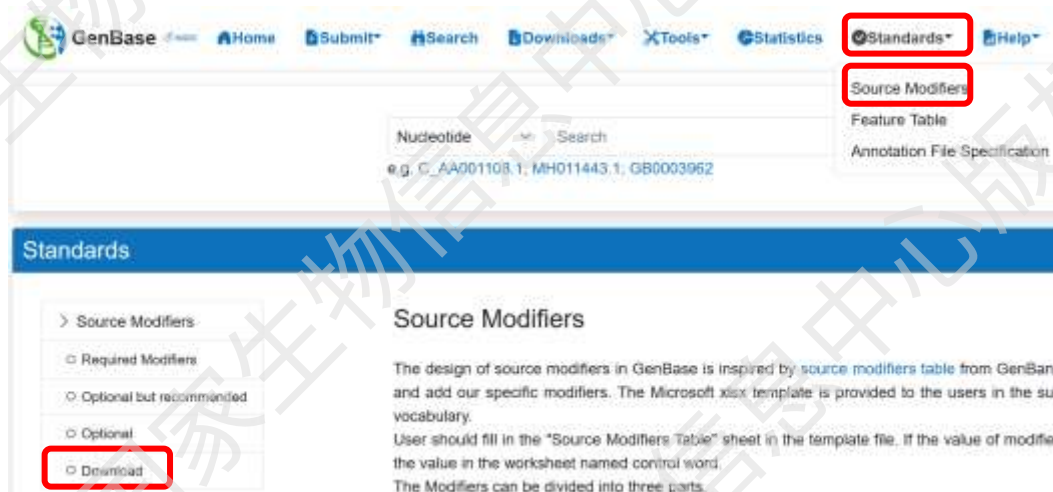
Sequence ID:

- (1) Start with a letter
- (2) letter number - \_ : \* #
- (3) shorter than 13 characters



# Preparation 2-Source Modifier File

- Template: 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)
- The document contains relevant controlled vocabularies to describe how, when, where you obtained the samples and the relevant information about the samples, etc



# Preparation 2-Source Modifier File

|    | A                                                                                                                                                                                                                                                                                                  | B                                                                                                         | C                                                                                                                                                                                                                                                                                                 | D                                                                          | E                                                                                                                                                                                                                          | F            | G        |
|----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|----------|
| 1  | #!Version: 3.3                                                                                                                                                                                                                                                                                     |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 2  | #!DO NOT MODIFY HEAD LINES AND EXAMPLE LINE! 不要修改标题行和示例行!                                                                                                                                                                                                                                          |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 3  | #!The first column contains the Sequence_IDs used to identify each sequence in the nucleotide FASTA file. 第一列包含用于识别核苷酸 FASTA 文件中每个序列的 Sequence_ID.                                                                                                                                                 |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 4  | #!Specimens are identified in the Source Modifiers Table by the same Sequence_ID used in the FASTA file. Specimens 在 Source Modifiers Table 中由 FASTA 文件中使用的相同 Sequence_ID 标识。                                                                                                                      |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 5  | #!The heading for the first column must be exactly Sequence_ID as shown in the below. 第一列的标题必须正好是 Sequence_ID, 如下所示。                                                                                                                                                                               |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 6  | #!Each specimen in the set must have a line in the source modifiers file, even if there are no modifiers to apply to the specimen. 集合中的每个样本在 source modifiers 文件中都必须有一行, 即使没有 source modifiers 应用于样本。                                                                                              |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 7  | #!Each Sequence_ID may appear only once in the source modifier file. 每个 Sequence_ID 在 source modifiers 文件中只能出现一次。                                                                                                                                                                                  |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 8  | #!Fill in the data in the line beginning with "##number". If it is not enough, you can append lines by yourself. 填写以"##数字"开头的行中的数据。如果不够, 可以自行追加。                                                                                                                                                   |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 9  | #!Reference: <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> |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 10 | #!If you do not know what to fill in the optional modifiers, please leave it blank instead of filling in characters such as "Na", "null", "unavailable"... 如果您不知道如何填写可选的Modifiers, 请将其留空, 而不是填写"Na", "null", "unavailable"...                                                                      |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 11 | ##Color Code                                                                                                                                                                                                                                                                                       | required                                                                                                  | optional (recommended)                                                                                                                                                                                                                                                                            | optional                                                                   |                                                                                                                                                                                                                            |              |          |
| 12 | ##Column Number                                                                                                                                                                                                                                                                                    | ##1                                                                                                       | ##2                                                                                                                                                                                                                                                                                               | ##3                                                                        | ##4                                                                                                                                                                                                                        | ##5          | ##6      |
| 13 | ##Header                                                                                                                                                                                                                                                                                           | Sequence_ID                                                                                               | Collection_date                                                                                                                                                                                                                                                                                   | Country/Region                                                             | Clade                                                                                                                                                                                                                      | Host         | Host sex |
|    | Sequence ID. MUST be consistent with the fasta file                                                                                                                                                                                                                                                | Date the specimen was collected.<br>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, Ripkintown, MD. Available values can be found at ControlWords sheet. | Name of clade from which sequence was obtained, typically an accession ID. | When the sequence submission is from an organism that exists in a symbiotic, parasitic, or other special relationship with some record organism, the 'host' modifier can be used to identify the name of the host species. | Sex of Host  |          |
| 14 | ##Description                                                                                                                                                                                                                                                                                      |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 15 | ##Example(do not delete)                                                                                                                                                                                                                                                                           | Seq1                                                                                                      | 2001-01-31 or value in ControlWords sheet                                                                                                                                                                                                                                                         | China or value in ControlWords sheet                                       | C.Grant                                                                                                                                                                                                                    | Homo Sapiens | Male     |
| 16 | ##1                                                                                                                                                                                                                                                                                                |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 17 | ##2                                                                                                                                                                                                                                                                                                |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 18 | ##3                                                                                                                                                                                                                                                                                                |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 19 | ##4                                                                                                                                                                                                                                                                                                |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |
| 20 | ##5                                                                                                                                                                                                                                                                                                |                                                                                                           |                                                                                                                                                                                                                                                                                                   |                                                                            |                                                                                                                                                                                                                            |              |          |



# Preparation 2-Feature File

- Template: GenBase\_Features.xlsx

[https://ngdc.cncb.ac.cn/genbase/download/template/Genbase\\_Features.xlsx](https://ngdc.cncb.ac.cn/genbase/download/template/Genbase_Features.xlsx)

- You can upload annotation files in gff, tbl or excel formats

## TBL:

```
>Feature 5c_10
1 7000 REFERENCE
      PubMed      8849441
      gene        ATH1
      gene
<1 1050 CDS
      product     acid trehalase
      product     Ath1p
      codon_start 2
<1 1050 mRNA
      product     acid trehalase
```

## GFF3:

```
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
```



# Preparation 2-Feature File

The screenshot shows the GenBase website interface. At the top, there is a navigation bar with links: GenBase, Home, Submit, Search, Downloads, Tools, Statistics, Standards (highlighted with a red box), and Help. Below the navigation bar, there is a search area with a dropdown menu set to 'Nucleotide' and a search button. Below the search area, there is a search bar containing the text 'e.g. C\_AA001108.1; MH011443.1; GB0003962'. On the right side of the search area, there is a dropdown menu with options: Source Modifiers, Feature Table (highlighted with a red box), and Annotation File Specification. Below the search area, there is a blue header bar labeled 'Standards'. On the left side of the 'Standards' section, there is a sidebar with a list of options: Feature Table (expanded), Locations, Attributes, Qualifier hints, Description of each feature, and Download (highlighted with a red box). The main content area of the 'Standards' section is titled 'Feature Table' and contains text explaining the design of the feature table in xlsx format, referring to the TBL in text format from NCBI. It also lists 'Locations' and provides instructions for the header of each sequence and the start and end of the sequence.

GenBase

Home Submit Search Downloads Tools Statistics Standards Help

Nucleotide Search

e.g. C\_AA001108.1; MH011443.1; GB0003962

Source Modifiers

Feature Table

Annotation File Specification

Standards

> Feature Table

Locations

Attributes

Qualifier hints

Description of each feature

Download

## Feature Table

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

### Locations

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



# Preparation 2-Feature File

Annotation format with meta information. The user needs to fill in the corresponding annotations in the structure name of this type of sequence, **qualifiers (column F)** is the attribute of this structure, most of them should be filled in columns **A to G**

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的名称，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后，示例或者格式，帮助您正确的填写**qualifier value (column G)**

| Attributes (属性) |         |           |                 |
|-----------------|---------|-----------|-----------------|
| feature         | feature | qualifier | qualifier value |
|                 |         |           |                 |
|                 |         |           |                 |
|                 |         |           |                 |
|                 |         |           |                 |
|                 |         |           |                 |

**Specific information here**

**Qualifier hints (提示)**

| feature    | qualifier |
|------------|-----------|
| gene       | gene      |
| Example    |           |
| /gene=ilvE |           |

**2.Hints area, according to the prompt to fill in G column content, for your reference**

91

**1. The feature information should be filled in columns A to G**

### 3. Please fill in the specific information here

feature qualifier 

2000

gene

### Example

```
/gene=ihvE
```

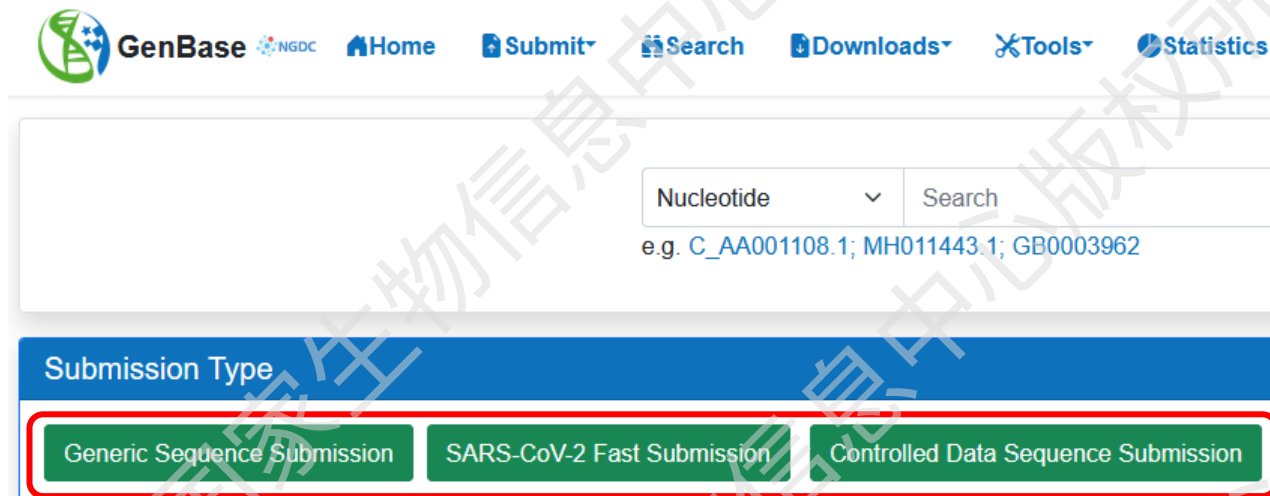
## 2.Hints area, according to the prompt to fill in G column content, for your reference



**CNCB-NGDC**



# Create a New Submission



The screenshot shows the GenBase NGDC website interface. At the top, there is a navigation bar with the GenBase logo and links for Home, Submit, Search, Downloads, Tools, and Statistics. Below the navigation bar is a search area with a dropdown menu set to 'Nucleotide' and a 'Search' button. Below the search area is a 'Submission Type' section with three buttons: 'Generic Sequence Submission', 'SARS-CoV-2 Fast Submission', and 'Controlled Data Sequence Submission'. The 'Controlled Data Sequence Submission' button is highlighted with a red border. Below the submission type section, there is a text box with the example sequence 'e.g. C\_AA001108.1; MH011443.1; GB0003962'.

GenBase NGDC

Home Submit Search Downloads Tools Statistics

Nucleotide Search

e.g. C\_AA001108.1; MH011443.1; GB0003962

Submission Type

Generic Sequence Submission SARS-CoV-2 Fast Submission Controlled Data Sequence Submission

Click here to start submitting

## Controlled sequence submission:

Only human data can be used for controlled data submission through this submission module

# Step 1. Submitter

Progress bar: 1. Submitter, 2. Reference, 3. Technology, 4. Nucleotide, 5. SetBatch, 6. Category, 7. Modifiers, 8. Features, 9. Overview

**Submitter**

\* First Name: zhao

\* Email: zhaoxuelong@ig.ac.cn

\* Organization: Beijing Institute of Genomics, Chinese Academy of Sciences

Phone:

\* Street: Beichen West Road, Chaoyang

Postal Code: 100010

Middle Name:

\* Last Name: xuelong

Email (secondary):

Submit Organization Unit:

\* Department:

Fax:

\* City: Beijing

State:

\* Country: China

**Save and next**

If any problems occur during the data information review and file archiving process, the information will be fed back to the submitter's email



## Step 2. Reference

Submission #0000099

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

| * First Name                          | Middle (Initials)    | * Last Name                       | Suffix               | Remove                                |
|---------------------------------------|----------------------|-----------------------------------|----------------------|---------------------------------------|
| <input type="text" value="swelling"/> | <input type="text"/> | <input type="text" value="OTAC"/> | <input type="text"/> | <input type="button" value="Remove"/> |

More Corresponding Author(s)

Reference Information #1

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

\* Publication Status

☒ Unpublished ☐ In-Press ☐ Published

Reference Title

\* Reference Authors

☒ Same As Corresponding Author(s) ☐ Specify New Authors

If the data is archived and the article is published, you can send an email to [genbase@big.ac.cn](mailto:genbase@big.ac.cn) and we will update the paper informations for you



# Step 3. Technology

Submission #0000099

## Sequencing Technology

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

What methods were used to obtain these sequences?

|                                                               |                                 |                                  |
|---------------------------------------------------------------|---------------------------------|----------------------------------|
| <input checked="" type="checkbox"/> Sanger dideoxy sequencing | <input type="checkbox"/> 454    | <input type="checkbox"/> Helicos |
| <input type="checkbox"/> Illumina                             | <input type="checkbox"/> PacBio | <input type="checkbox"/> SOLiD   |
| <input type="checkbox"/> Other                                |                                 |                                  |

Are these sequences:

Assembled sequences

**Continue**

Assembled sequences

Unassembled sequence reads

Assembled sequences

Here to choose Assembled sequences or need to submit to the GSA rather than GenBase

# Step 4. Nucleotide

Submission #0000099

Submission Release Date

When may we release your sequence record?

☒ Immediately After Processing

☐ Release Date:

Date format is "YYYY-MM-DD" (example: 2022-02-07)  
Release date must be after 6 months from today and before 4 years from today.

Release Policies and Disclaimers

1. A date can be set by authors to withhold the release of new submissions for a specified period of time.  
2. If submitters want to change release date, please contact GenBank group: [genbank@ncbi.nlm.nih.gov](mailto:genbank@ncbi.nlm.nih.gov)  
3. If a paper citing the sequence or accession number is published prior to the specified date, the sequence will be released upon publication.  
4. As soon as it is available, please send the full publication data--all authors, title, journal, volume, pages and date--to the following:  
5. After the data is released, it may automatically synchronize to NCBI in the future, and submitters will no longer be contacted for comments.

☒ I accept it ☐ I don't accept it

Submission Short Title

Short Title:

Set the submission title and try to make it have a true meaning

BioProject

BioProject:

Grant(s)

Agency:  Program:  Grant ID:  Grant Title:  Remarks:

Ministry of Science and Technology National Key Technologies R&D Program 123

BioProject ( optional )

Funding information



# Step 4. Nucleotide

## Sequence(s) and Definition Line(s)

\* Molecule Type:

genomic DNA

Topology

Linear

Strand:

To which compartment does the one you submitted belong?

mitochondria

Are you submitting the complete sequence of an organelle genome, virus, viral segment, viroid, plasmid, or cloning vector?

☒ Yes ☐ No

genomic DNA

mRNA (cDNA)

genomic RNA

precursor RNA

tRNA

rRNA

cRNA

transcribed RNA

Other genetic: RNA

Other genetic: DNA

Linear

Circular

Single strand

Double strand

Mixed strand

nucleotide

mitochondria

plastid

It is necessary to ensure that the data submitted in one batch is consistent with the above situations

### Nucleotide Sequence Format

Sequence data format:

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

For example:

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

aaccgatatagagagagga

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

atctgaatagagattttt

Example of nucleic acid sequence

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





## Step 4. Nucleotide

**Nucleotide Sequence(s)**

☐ Paste sequence data (when sequence number < 40)

```
>Alligator_VPSS2-RNASE  
CTTGAGCTAAGTTTCCCAATACTTAGTGAACCCAGGGTAGCATTGTAACGTGT  
TGTGTTTGTTCCTTTGCATGTCTATTACTAGBACCATTATATATTCACATAC  
TTAGCAATAAAATAACTTGATAGTTAACTGGTGGTAACTGGGTGATTTTTCTTT  
ACTTTTCCTCCCCTCGCTCTGTAGCAACGCTCTTTTACCTAAGCTAAAAATC  
GTCGTAAGCTGAAAAGCTTCGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT
```

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

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

\* File Name

\* MD5

Note:

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

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

User: Same as you login the GenBase

Password: Same as you login the GenBase

Three methods are provided for uploading nucleic acid sequences

# Step 5. 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. Please enter the organism name below. (For future sequence submissions, be sure to use the FASTA format.)

Download organism template file

Organism Name:  Placeholder: Species  
Input same organism name for all sequences

or

Upload File:  Select File

Please fill the organism name as precisely as possible to the "species" level



# Step 6. Set/Batch

The screenshot shows a progress bar at the top with 10 steps: Submitter, Reference, Technology, Nucleotide, Organism, Set/Batch (highlighted with a blue circle), Category, Modifiers, Features, and Overview. Below the progress bar, the 'Set/Batch' section is titled 'Submission #0000099'. It contains a red-bordered box with the following text: 'All sequences in a set must be from the same gene locus and are expected to be released at the same time. Please select a set type:'. Below this, there are four radio button options: 'Pop set' (selected), 'Phy set', 'Mut set', and 'Env set'. To the right of these options, there are four definitions: 'Population study: a set of sequences that were derived by sequencing the same gene from different isolates of the same organism.', 'Phylogenetic study: a set of sequences that were derived by sequencing the same gene from different organisms.', 'Mutation study: a set of sequences that were derived by sequencing multiple mutations of a single gene.', and 'Environmental study: a set of sequences that were derived by sequencing the same gene from a population of sequestered or unknown organisms.'. Below these definitions, there is another red-bordered box with the text: 'If your sequences are NOT all from the same gene locus and NOT intended to be released at the same time, then choose 'Batch' below.' followed by a 'Batch' radio button option. At the bottom left, there is a 'Continue' button.

Submission #0000099

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

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

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

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

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

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

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

☐ Batch

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

Continue

Whether the data comes from the same batch or set

# Step 7. Category

Category

Submission #0000099

Indicate whether your sequence is an original submission. If your sequence is a third-party annotation submission, please contact [genbase@big.ac.cn](mailto:genbase@big.ac.cn).

☒ Original Directly sequenced by submitter

☐ Third Party Annotation Derived from other primary sequence data.

[Continue](#)

- Is the sequence the original submission or a third-party annotation
- Third-party annotated data can only be made public after the article is published

# Step 8. Modifiers

Submission #0000099

Note

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

\* Are these sequences derived from organelle genomes? (such as Mitochondria or Chloroplast)

☒ yes ☐ no If yes, please fill in the Organelle/Location column in GenBase\_Modifiers.xlsx.

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

0000099\_modifier.xlsx Remove

Successful verification!

Continue

Modifiers file template

Choose yes or no based on whether the sequence comes from the organelle genome, such as mitochondria or chloroplasts

| id   | Organelle/Location                                                                  |
|------|-------------------------------------------------------------------------------------|
| g#13 | genetic compartments.<br>Available viruses can be found at<br>GenBank/Genome sheet. |

# Step 9. Features

Submission #0000099

Note

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

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

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

☐ Upload feature annotations for your submission using GFF3 format.

0000099\_features.gff3 Remove

Continue

Select the type of file

Upload features file

If no annotation file is uploaded here, the final archived page will be marked:

**GenBase staff is unable to verify sequence and/or annotation provided by the submitter**

# Step 10. Overview

Submission #0000021

1. Additional Email Addresses?  
Correspondence regarding this submission will be sent to the following email address:  
  
Separate multiple email addresses with commas.

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

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

4. Update:  
You may update or review your submission at any time by sending new or corrected information in an email to [genbank@ncbi.nlm.nih.gov](mailto:genbank@ncbi.nlm.nih.gov). You may also contact us at this address with any questions.

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

[Finish Submission](#)

Confirm above information

Click " Finish Submission " and wait for the page to display





# Step 10. Overview

## Review Records of Your Set

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

You can download the [complete set](#) as a compressed ZIP file.

Warnings: there are some warnings in your sequence file. We suggest you check the sequence file before you go ahead. Download whole warning file [result\\_err.txt](#) Click to download

| Severity | Sequence Name      | Error Title                          | Description                                                                                                                                    |
|----------|--------------------|--------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| ●        | Luciola_filiformis | SEQ_FEAT.BadRRNAcomponentOverlapTRNA | tRNA-rRNA overlap FEATURE: tRNA: Leu [lcl Luciola_filiformis:c12463-12402] [lcl Luciola_filiformis: raw, dna len= 16864]                       |
| ●        | Luciola_filiformis | SEQ_FEAT.BadRRNAcomponentOverlapTRNA | tRNA-rRNA overlap FEATURE: tRNA: Val [lcl Luciola_filiformis:c13787-13725] [lcl Luciola_filiformis: raw, dna len= 16864]                       |
| ●        | Luciola_filiformis | SEQ_FEAT.TranslExceptPhase           | transl_except qual out of frame. FEATURE: CDS: cytochrome c oxidase subunit III <11> [lcl Luciola_filiformis:4654-5438] [Luciola_filiformis 6] |

Check for the yellow warnings and try to make changes (optional)



# Awaiting review

| 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 | 2025-10-24   | 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 22:26:50 | 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 10:26:26 | 2022-10-21 16:01:55 | 2022-10-20   | Meta Data Finished   | - |

Submission successful, awaiting review



# Data Release

## Completed Submission(s)

Submission #0000089

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

2 nucleotide sequence(s) have been submitted

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

seq1 C\_AA001111.1 GBFF Fasta

seq2 C AA001112.1 GBFF Fasta

|            |                                                                                                                                                                                                                                         |         |      |        |                 |
|------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------|------|--------|-----------------|
| LOCUS      | C_AAC01113                                                                                                                                                                                                                              | 3452 bp | mRNA | linear | JV9-81-626-2022 |
| DEFINITION | <i>Milnespera lugens</i> Zinc finger homeodomain 1 (zfhl) variant B (Zfhl)                                                                                                                                                              |         |      |        |                 |
| FEATURES   | cDNA, non-template strand.                                                                                                                                                                                                              |         |      |        |                 |
| SEQUENCE   | C_AAC01113                                                                                                                                                                                                                              |         |      |        |                 |
| VERSION    | C_AAC01113.1                                                                                                                                                                                                                            |         |      |        |                 |
| KEYWORDS   | -                                                                                                                                                                                                                                       |         |      |        |                 |
| SOURCE     | <i>Milnespera lugens</i> (Horse plant hopper)                                                                                                                                                                                           |         |      |        |                 |
| ORGANISM   | <i>Milnespera lugens</i>                                                                                                                                                                                                                |         |      |        |                 |
| REFERENCE  | Eukaryotes: Opiliones; Mesostigmata: Eumetastigmata; Milnesperidae: Synonymis<br>Kutymysini; Pentatomorpha: Arthropoda: Hemiptera: Coreidae: Pentatominae:<br>Mesopoda: Trochila: Elasmophylla: Pteroptera: Heteroptera: Pentatomorpha: |         |      |        |                 |
| REFERENCE  | 1 (bases 1 to 3452)<br>Xu, S. and Zhang, J.<br><i>OpenLife</i>                                                                                                                                                                          |         |      |        |                 |
| REFERENCE  | 2 (bases 1 to 3452)<br>Xu, S. and Zhang, J.<br><i>BioRx</i>                                                                                                                                                                             |         |      |        |                 |
| TITLE      | Zinc finger homeodomain                                                                                                                                                                                                                 |         |      |        |                 |
| JOURNAL    | Submitted (J1-JUN-2022) Institute of Insect Sciences, Zhejiang University, Yuhangtang Road 866, Hangzhou 311000, China                                                                                                                  |         |      |        |                 |
| FEATURES   | Location/Qualifiers                                                                                                                                                                                                                     |         |      |        |                 |
| exon       | 1..3452<br>/organism="Milnespera lugens"<br>/mol_type="mRNA"<br>/db_xref="taxon:108911"<br>/country="China"                                                                                                                             |         |      |        |                 |
| gene       | 1..3452<br>/gene="Zfhl"                                                                                                                                                                                                                 |         |      |        |                 |
| CDS        | 1..3452<br>/gene="Zfhl"<br>/codon_start=1<br>/product="Zinc finger homeodomain 1 (zfhl) variant B"                                                                                                                                      |         |      |        |                 |

>C\_AA001113.1 Hisapervatus Zinc finger homeodomain 1 (zfhl1) isoform 2 (zfhl1) mRNA, complete cds

## Accession numbers

**It can be written in the article**



# Data Search

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 INSDC.

Nucleotide Type your keywords Search Advanced

Advanced Search

Add terms to the query box

Nucleotide Protein

Query box

View your search query here

Nucleotide All Fields

Enter a search term

ADD

Add with AND

Add with OR

Add with NOT

- Click on "Search", enter the keywords, then click "Search", and the search can be completed
- Or you can conduct the search through the advanced search page



# Data Search

## Search Result

Search

Species

Bacteria (326)

Fungi (12)

Animals (1)

Plants (1)

Data types

Nucleotide (340)

Molecule types

DNA (340)

Sequence length (bp)

Revision date

Summary

10 per page

Sort by Default order

Send to:

Found 1 to 10 of 340 items

Select All

Total 34 pages

<< First 1 2 3 4 5 6 7 8 9

☐ [Helicobasidium mompa isolate ZWY-4 large subunit ribosomal RNA gene, partial sequence](#)

1. 549 bp DNA  
Accession: ON055754.1

☐ [Helicobasidium mompa isolate ZWY-1 large subunit ribosomal RNA gene, partial sequence](#)

2. 563 bp DNA  
Accession: ON055750.1

☐ [Falsiroseomonas bella strain CQN31 16S ribosomal RNA gene, partial sequence](#)

3. 1,489 bp DNA  
Accession: MT573348.1

CNCB-NGDC

110

# Data Download

**You can select all of them here**

Found 1 to 10 of 3,588 items

☐ Select All

**1. Select the items that interest you**

- ☒ 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. Click Send to**

**3. Choose an appropriate format**

**4. Click to create a file**

Send to: ▼

Format

Summary

Create File

Format

- Summary
- Summary
- GBFF
- FASTA
- Accession List
- CI List



# Contents

- 1 CNCB Data Submission System Overview
- 2 GSA Data Submission & Sharing
- 3 GSA-Human Data Submission & Sharing
- 4 GWH Data Submission & Sharing
- 5 GenBase Data Submission & Sharing
- 6 OMIX Data Submission & Sharing**





# Accepted Data Types

| Data Type           | Sub Type                                                                                                                                                                                                              |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b>Genomics</b>     | Methylation Profiling by NGS<br>Expression Profiling by NGS<br>Non-coding RNA Profiling by NGS<br>Genome Binding/Occupancy Profiling by NGS<br>Chromatin Accessibility Profiling by NGS<br>Other Type of Genomic Data |
| <b>Microarray</b>   | Methylation profiling by Array<br>Expression Profiling by Array<br>Non-coding RNA Profiling by Array<br>Genome binding/occupancy profiling by Array<br>Other type of Microarray Data                                  |
| <b>Proteomics</b>   | Protein 3D Structure Data<br>Proteomic Data by Mass Spectrometry<br>Other Type of Proteomic Data                                                                                                                      |
| <b>Metabolomics</b> | Metabolome Data by Mass Spectrometry<br>Lipidome Data by Mass Spectrometry<br>Other Type of Metabolome Data                                                                                                           |

| Data Type                   | Sub Type                                                                                                                 |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------|
| <b>Clinical Information</b> | Demographic Data<br>Clinical Research Data<br>Other Type of Clinical Information                                         |
| <b>Biomarker</b>            | Genetic Biomarkers<br>Karyotype Biomarkers<br>Protein Biomarkers<br>Chemical Biomarkers<br>Condition-specific Biomarkers |
| <b>Imaging</b>              | Magnetic Resonance Imaging<br>Behavioral Video Data<br>Other Type of Image Data                                          |
| <b>Neurobiological Data</b> | /                                                                                                                        |
| <b>Flow cytometry</b>       | /                                                                                                                        |
| <b>Spectra Data</b>         | Raman Spectra Data<br>Other type of Spectra Data                                                                         |

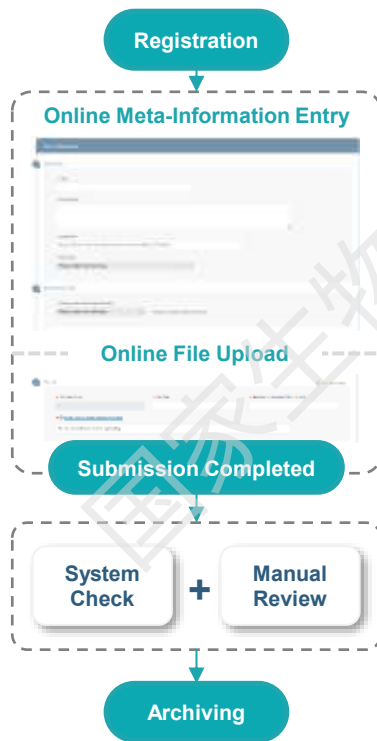


# Recommended Data Formats

| Formats                        | Preferred                                                                                                                                                                       | Non-Preferred                                 |
|--------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------|
| Compressed Files               | .gz, .zip, .tar<br><i>For single files <math>\geq 50</math> GB, please use .zip compression.</i>                                                                                | .rar                                          |
| Post-NGS Analysis Data         | .txt, .tsv, .csv, .mtx, .bed                                                                                                                                                    | Binary files such as .bedgraph, .bigwig, etc. |
| Microarray Data                | Raw Data: .cel, .idat, .txt, .csv, .tsv, .chp, etc.                                                                                                                             | -                                             |
| Metabolomics MS Data           | Raw Instrument Data: .D, .BAF, .FID, .YEP, .WIFF, etc.<br>Analysis Data: .mzML, .mzXML, .xml, .txt, etc.                                                                        | -                                             |
| Proteomics MS Data             | Raw Data: .D, .BAF, .FID, .YEP, .WIFF, etc.<br>Peak Data: .mzML, .mzXML, .mgf, etc.<br>Results Data: .mzIdentML, .mzTab; Database Search Results: .xml, .pepxml, .protxml, .txt | -                                             |
| Image Data                     | JPEG (.jpg, .jpeg); TIFF (.tif, .tiff); MRC                                                                                                                                     | -                                             |
| Video Data                     | .MP4, .avi                                                                                                                                                                      | -                                             |
| Spreadsheet / Statistical Data | .xls, .csv, .txt                                                                                                                                                                | -                                             |



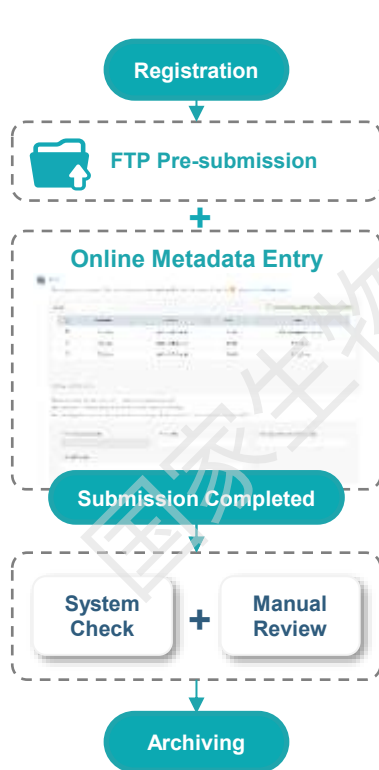
# Data Submission: Online Upload



The screenshot shows a multi-step online submission form. The steps visible are: 1. Data Info (with fields for Title, Description, Organism, and Data Type), 2. BioProject Info (with a dropdown for BioProject selection), 3. Release Date (with radio buttons for 'Release today after submitting' and 'Release on specified date'), and 4. Files (with a section for file uploads). The form includes various instructional text and validation messages.

- **Data Info:** Enter basic metadata, include Title, Description, Organism, and Data Type.
- **BioProject Info:** Select the associated **BioProject** access or jump to create a new BioProject.
- **Release Date:** Set when to make the data public — immediately after submission or on a specified date.
- **Files:** Upload data files online

# Data Submission: FTP Upload



The screenshots show the FTP Upload interface. The top screenshot displays the 'Files' section with a table of pre-uploaded files. The bottom screenshot shows the 'Metadata' section where users can select pre-uploaded files and complete the metadata.

**Files Table:**

| File Name  | Last Modified       | Size    | Path       |
|------------|---------------------|---------|------------|
| FTS_01.jpg | 2023-09-14 12:00:04 | 1.34 KB | FTS_01.jpg |
| FTS_01.jpg | 2023-09-14 12:00:04 | 1.34 KB | FTS_01.jpg |
| FTS_02.jpg | 2023-09-14 12:00:04 | 1.34 KB | FTS_02.jpg |

**Metadata Section:**

select pre-uploaded files

complete the metadata

Unlike browser-based uploads, FTP lets users select those pre-uploaded files directly in the interface, then add a file title and sample count for each, completing the pairing of metadata with the data.

# Data Browse and Access

OMIX

[Home](#)
[About](#)
[Get in touch](#)
[Contact](#)

View all released data of OMIX

Released Data - 1416 records

Items

12

Columns

| OMIX ID | Project    | Title                                                                                                             | Organization                         | Access Type | Release Time | Operation                                                                                                                                                           |
|---------|------------|-------------------------------------------------------------------------------------------------------------------|--------------------------------------|-------------|--------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| OMIX026 | PJLJ000038 | Protein mass spectrometry in the peritoneal lymphatic system                                                      | Shanghai Institute of Materia Medica | Open        | 2018.12.23   |   |
| OMIX020 | PJLJ000048 | Drug reuse potential                                                                                              | China Food and Drug Administration   | Open        | 2017-05-08   |   |
| OMIX021 | PJLJ000039 | Proteomic profiling of drug targets after target activation                                                       | China Food and Drug Administration   | Open        | 2017-08-20   |   |
| OMIX022 | PJLJ000070 | HLI                                                                                                               | Protein synthesis                    | Controlled  | 2022-04-01   |   |
| OMIX024 | PJLJ000218 | Topography of the brain                                                                                           | Protein synthesis                    | Controlled  | 2022-08-17   |   |
| OMIX025 | PJLJ000035 | Key enzyme production status mark the production of pancreatic endocrine diseases in mice and humans              | Protein synthesis                    | Open        | 2017-09-07   |   |
| OMIX021 | PJLJ000040 | A type of normal protein reference sequence                                                                       | Protein synthesis                    | Open        | 2017-08-19   |   |
| OMIX026 | PJLJ000052 | Maximum synergies with thermodynamic to cause responses molecularly targeting HIF-1 in DMS                        | Multi-records                        | Open        | 2017-02-25   |   |
| OMIX040 | PJLJ000470 | Infra-early expressed gene and alternative splicing affected by RNA                                               | RNA                                  | Open        | 2019-05-17   |   |
| OMIX041 | PJLJ000472 | Proteomic Analysis of Heat Shocking in Proteomic Analysis of Heat Shocking in Proteomic Analysis of Heat Shocking | Enzyme/Protein Synthesis             | Open        | 2019-11-17   |   |

Showing 401 to 414 of 450 records

Previous

1

2

3

4

5

6

7

8

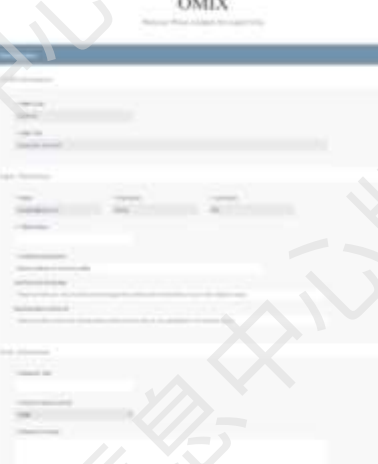
9

10

Next

| DOI ID  | Project     | Title                                                             | Co-ordinator         | Access Type | Release Date | Operation                                                                                                                                                                                             |
|---------|-------------|-------------------------------------------------------------------|----------------------|-------------|--------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 0004118 | PLACOCK2008 | Protein-protein interactions of the pathogenesis Angerium complex | Researcher contact   | Open        | 2009-05-24   |                                 |
| 0004119 | PLACOCK2010 | Drug target prediction                                            | Open access hardware | Open        | 2011-05-05   |   HTTP Download<br>FTP Download |

The Open-access datasets can be fully viewed and downloaded by anyone.



Controlled-access datasets require application and review



# Contact Us

GSA  
GSA-Human  
OMIX



gsa@big.ac.cn

GWH



gwhcuration@big.ac.cn  
gwh@big.ac.cn

GenBase



genbase@big.ac.cn





**THANKS**