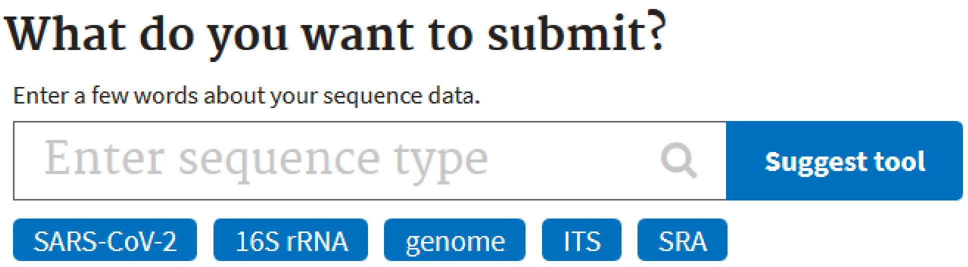
NCBI 数据上传指南

近年来，伴随着高通量测序的广泛应用，海量的测序数据也随之产生。高通量测序数据分析完成发表文章时，科研工作者需要将数据提交到一个公共平台（例如 NCBI 数据库）。下文中我们汇总了如何向 NCBI 平台提交数据，以及不同测序项目需要提交哪些数据，希望为您发表高分论文提供一些帮助。

一、 NCBI 数据库及数据类型

向 NCBI 数据库提交数据可参考 Submission Portal 网页中所列数据库与工具，按照网站提示及说明进行操作，可以在如下输入框中输入关键词查看相关信息。



What do you want to submit?

Enter a few words about your sequence data.

1. NCBI 常用数据库介绍

1）GenBank
网址：http://www.ncbi.nlm.nih.gov/genbank/

GenBank 是美国国立卫生研究院（NIH）基因序列数据库，包含所有公开的 DNA 序列和注释信息。GenBank 数据库也是世界上最大的、最重要的、最有影响力的生物全领域数据库，其数据正被全球数以百万计的研究人员获取与引用。

2）SRA
网址：http://www.ncbi.nlm.nih.gov/sra/

存储测序平台产生的测序数据。包括 Roche 454 GS System®, Illumina Genome Analyzer®, Applied BiosystemsSOLiD® System, HelicosHeliscope®, Complete Genomics®, and Pacific Biosciences SMRT®。

3）TSA（ Transcriptome Shotgun Assembly ）

网址：http://www.ncbi.nlm.nih.gov/genbank/tsa/

存储由第二代测序数据组装拼接得到的转录本序列。

2. 数据类型

上传到 NCBI 的数据，依据数据类型，大体可以分为测序原始数据和分析数据。

原始数据（Raw data）指未经任何处理的测序下机文件，包含二代及三代测序数据等。其中二代测序中最常见的是 illumina 测序仪产生的 fastq 文件，例如锐博在项目结题时为您提供的*fastq.gz 文件。这一类型的文件需要提交到 NCBI 的 SRA 数据库，具体上传方法我们将在下文中做详细介绍。

分析数据指原始数据在不同分析目标处理后得到的结果文件。不同的项目类型会产生不同的数据分析结果，相应的分析结果需要提交到不同的数据库。目前的高通量测序就项目类型而言可大体分为：基因组测序、转录组测序、16S/ITS 测序，宏基因组测序等。NCBI 中不同的数据对应的数据库及提交方法可参照如下链接：http://www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data/。

常见的高通量测序数据需要上传的数据库可参照下表：

数据类型	备注	数据库
高通量测序数据	包括二代和三代测序数据	SRA
功能基因组学研究	包括基因表达、调控及表观基因组	GEO
全基因组序列组装	包括叶绿体、线粒体、质粒、噬菌体和病毒	GeneBank WGS
大基因组完成图数据	包括细菌和真核生物	
不完整的基因组	whole genome shotgun (WGS) sequences.	
转录组组装序列	Transcript survey sequence assemblies	TSA
宏基因组	包含非人类及环境宏基因组	Metagenome

下文中我们就详细介绍一下如何将测序数据提交到 SRA 及 GEO 数据库。

二、 数据提交 SRA 数据库

向 SRA 提交数据一般分为以下几个步骤：

1) 注册 NCBI 账号；

2） 创建 BioProject 及 BioSample ID；

3） BioProject 和 BioSample 创建完成后，再转到 SRA 的网页，创建 New Submission，并完成信息填写；

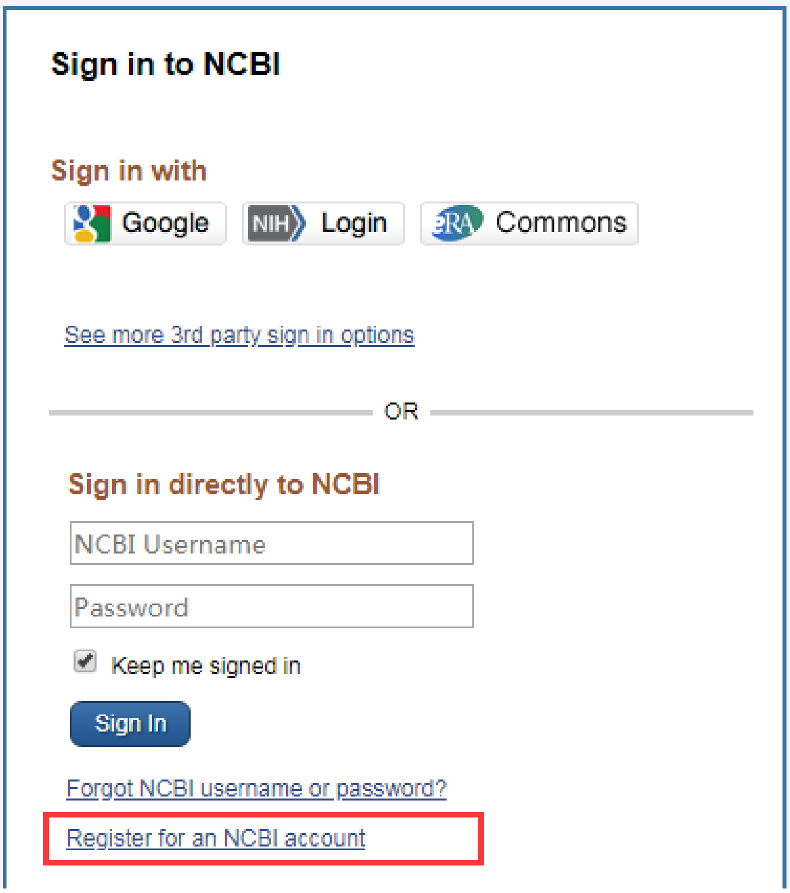
4） 完成上述步骤后，网页上 NCBI 会给出一个登陆 FTP 的账号和网址链接；

5） 登陆后用账号可直接上传（复制粘贴），或用软件 FileZilla 或 Aspera 上传；

6） 生成相应的数据编号，供发表文章使用。

1. 注册 NCBI 帐号

打开链接 https://www.ncbi.nlm.nih.gov/account/，如下图所示，点击标注的“Register for a NCBI account”，进入到注册页面，如实填写信息。帮助文档可参考：https://www.ncbi.nlm.nih.gov/books/NBK3842/#MyNCBI.Registering_with_My_NCBI



2. 创建 BioProject 及 BioSample ID

将数据传递到 NCBI，都需要对这份数据进行一个描述，包括前期项目情况、样本属性及制备情况等；BioProject 和 BioSample 即描述研究项目的、研究背景、材料属性等基本信息。

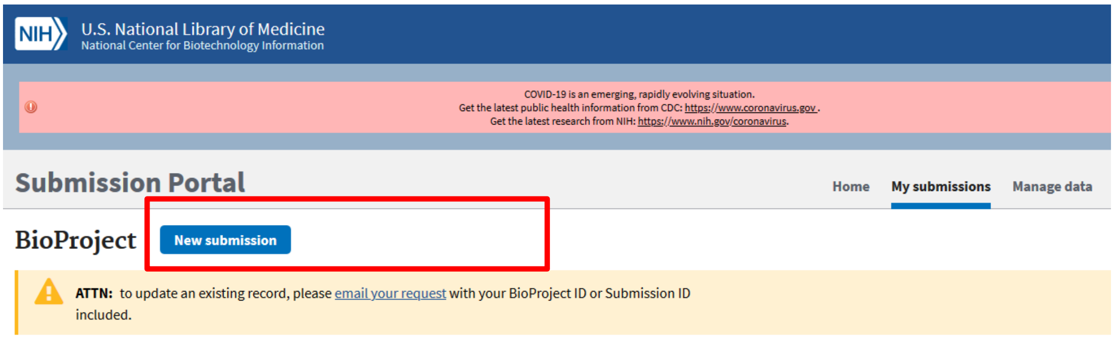
一个 BioProject 代表一项测序研究项目，可包含多个 BioSample，也可以包含多次实验 experiments，所以在提交数据前，先申请 BioProject 号和 BioSample 号。通常 BioSample 号以 SAMN 开头，如 SAMN*****；BioProject 号以 PRJNA 开头，如 PRJNA*****。这两个号需要在后续 SRA 提交过程中使用。

2.1 创建 BioProject ID

进入下方链接网页，点击 New submission：

https://submit.ncbi.nlm.nih.gov/subs/bioproject/

或登录 NCBI 之后点击页面左下角 Submit Data，在右下角选择 BioProject & BioSample，点击 Learn more，然后点击 submit。


NIH
U.S. National Library of Medicine
National Center for Biotechnology Information

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: https://www.coronavirus.gov .
Get the latest research from NIH: https://www.nih.gov/coronavirus.

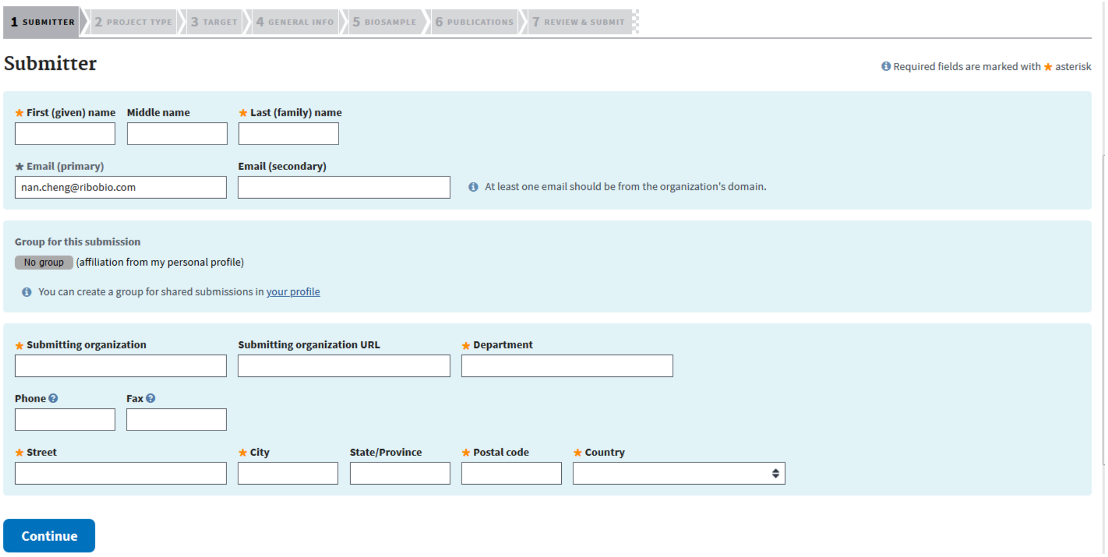
Submission Portal
Home
My submissions
Manage data
BioProject
New submission

ATTN: to update an existing record, please email your request with your BioProject ID or Submission ID included.

2.1.1 SUBMITTER

根据研究项目实际情况，填写一系列的信息，填完所有步骤后，要点击页面下方的 continue，保存已填写的信息。

提示：email 选项中，两个邮箱中要留一个该测序项目负责人的常用邮箱，因为后期如果想要修改数据信息或者释放时间，都需要该邮箱发送邮件到 NCBI 才会被受理。



Submitter

Required fields are marked with ★ asterisk

★ First (given) name　Middle name　★ Last (family) name

★ Email (primary): nan.cheng@ribobio.com　Email (secondary)

At least one email should be from the organization's domain.

Group for this submission
No group (affiliation from my personal profile)
You can create a group for shared submissions in your profile

★ Submitting organization　Submitting organization URL　★ Department

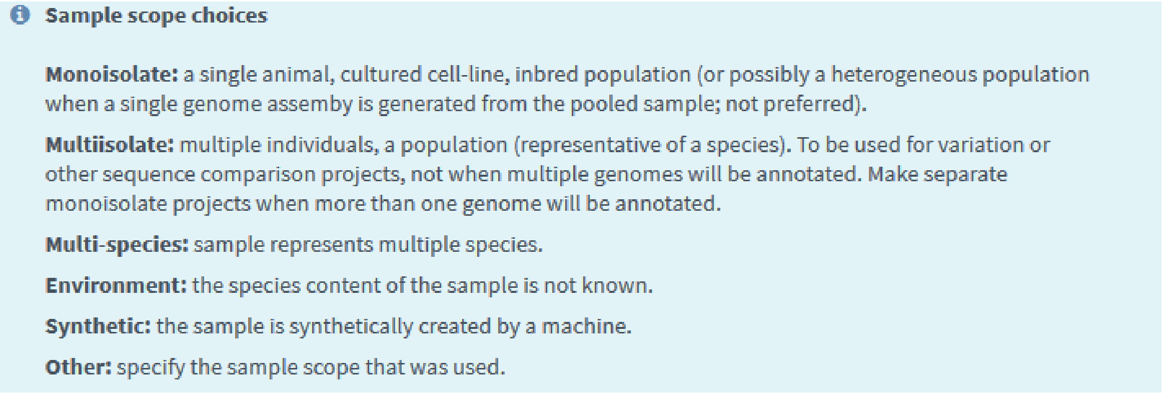
Phone　Fax

★ Street　★ City　State/Province　★ Postal code　★ Country

2.1.2 PROJECT TYPE

Project Type：可根据自己的项目类型选择，一般高通量测序数据可选择“Raw sequence reads”。

Sample scope：是对实验样品的简洁描述，根据不同选择会影响后面 TARGET 的填写，可选择 Monoisolate、Multiisolate、Monoisolate、Environment、Synthetic 或 others。对各种类型的说明如下截图:



Sample scope choices

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

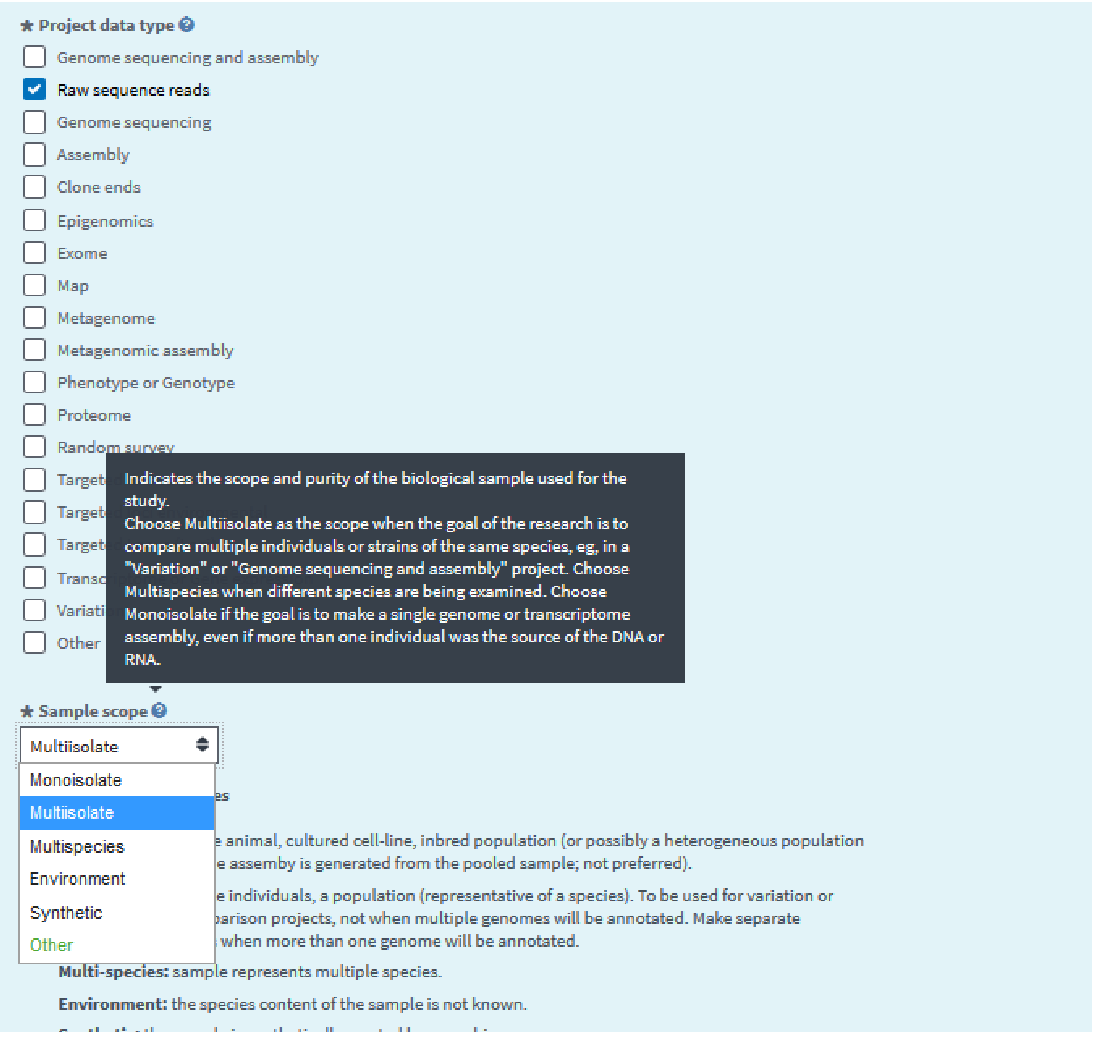
Multiisolate: multiple individuals, a population (representative of a species). To be used for variation or other sequence comparison projects, not when multiple genomes will be annotated. Make separate monoisolate projects when more than one genome will be annotated.

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.

Other: specify the sample scope that was used.



★ Project data type

□ Genome sequencing and assembly
☑ Raw sequence reads
□ Genome sequencing
□ Assembly
□ Clone ends
□ Epigenomics
□ Exome
□ Map
□ Metagenome
□ Metagenomic assembly
□ Phenotype or Genotype
□ Proteome
□ Random survey
□ Target
□ Target
□ Target
□ Transc
□ Variati
□ Other

Indicates the scope and purity of the biological sample used for the study.
Choose Multiisolate as the scope when the goal of the research is to compare multiple individuals or strains of the same species, eg, in a "Variation" or "Genome sequencing and assembly" project. Choose Multispecies when different species are being examined. Choose Monoisolate if the goal is to make a single genome or transcriptome assembly, even if more than one individual was the source of the DNA or RNA.

★ Sample scope

Multiisolate

Monoisolate
Multiisolate
Multispecies
Environment
Synthetic
Other

es

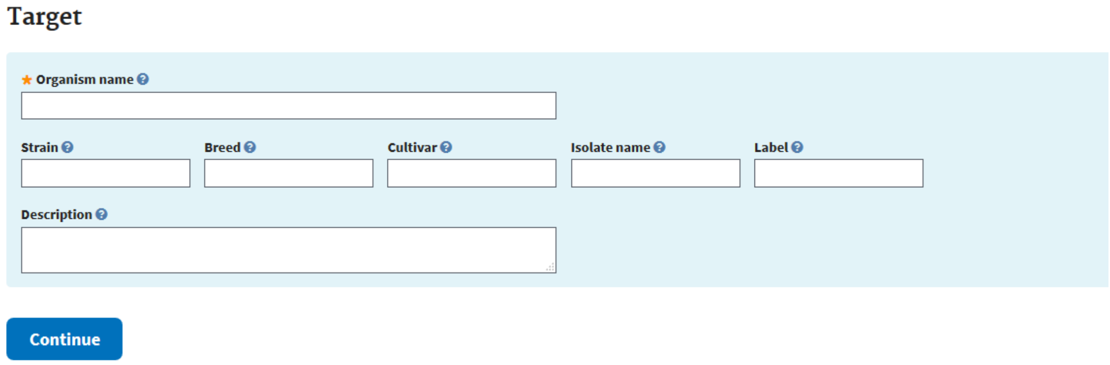
e animal, cultured cell-line, inbred population (or possibly a heterogeneous population e assemby is generated from the pooled sample; not preferred).

e individuals, a population (representative of a species). To be used for variation or parison projects, not when multiple genomes will be annotated. Make separate when more than one genome will be annotated.

Multi-species: sample represents multiple species.

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

2.1.3 TARGET



Target

★ Organism name

Strain	Breed	Cultivar	Isolate name	Label

Description

2.1.4 GENERAL INFO：基本信息

Release date：这个是您的数据公开日期，可以点击立即释放，也可以选择具体时间；

Project title：根据 TARGET 提供一个简短的标题，如：

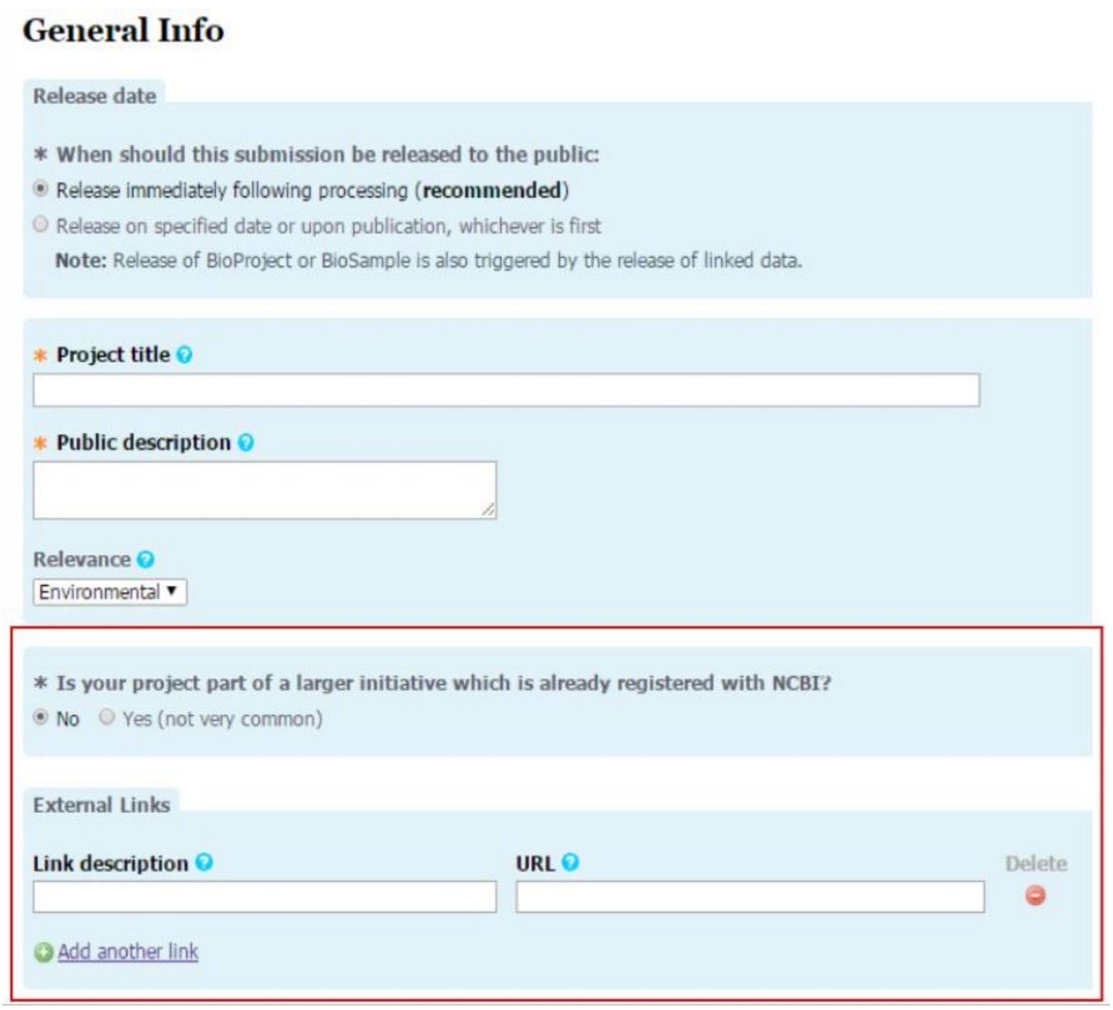
1）Chromosome Y sequencing;

2）Opportunistic pathogen that causes important food-born disease;

3）Global studies of microbial diversity on human skin.

注意：红框中要选择是否关联其他数据，若选择“No”则红框中的内容不进行填写；选择“Yes”，红框中的内容为必填项。

Public description：对研究目标及相关的内容进行一段描述。



General Info

Release date

∗ When should this submission be released to the public:

● Release immediately following processing (**recommended**)

○ Release on specified date or upon publication, whichever is first

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

∗ **Project title**

∗ **Public description**

Relevance

Environmental ▾

∗ **Is your project part of a larger initiative which is already registered with NCBI?**

● No ○ Yes (not very common)

External Links

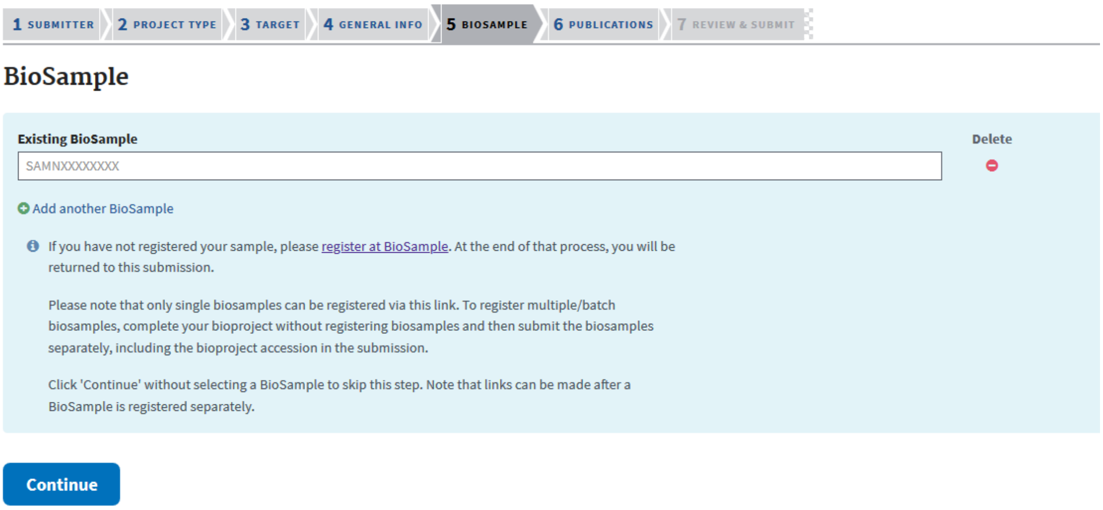
Link description | **URL** | Delete

Add another link

2.1.5 BIOSAMPLE

样品名称（编号 SAMNXXXXXXXXX），需要与创建 Biosample 时的样品名称一致。如果未创建 Biosample ID，可以点击 register at BioSample 进行创建，样品注册完成后会自动调回 BioProject 注册界面。多个样品可点击“Add another BioSample”增加样品信息。

如果您有多个样品,可以直接点击“Continue”,完成 BioProject 注册后再进行 BioSample 注册。



BioSample

Existing BioSample: SAMNXXXXXXXX

Delete

Add another BioSample

If you have not registered your sample, please register at BioSample. At the end of that process, you will be returned to this submission.

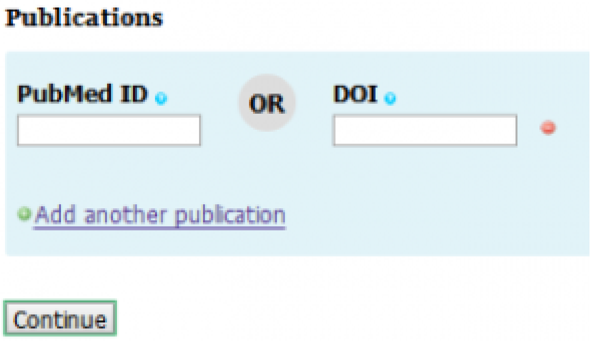
Please note that only single biosamples can be registered via this link. To register multiple/batch biosamples, complete your bioproject without registering biosamples and then submit the biosamples separately, including the bioproject accession in the submission.

Click 'Continue' without selecting a BioSample to skip this step. Note that links can be made after a BioSample is registered separately.

Continue

2.1.6 PUBLICATIONS

填写 PubMed ID 或 DOI 号。说明：BioSample 和 Publications 这两步可以省略，后期发邮件给 NCBI 进行修改。


Publications

PubMed ID OR DOI

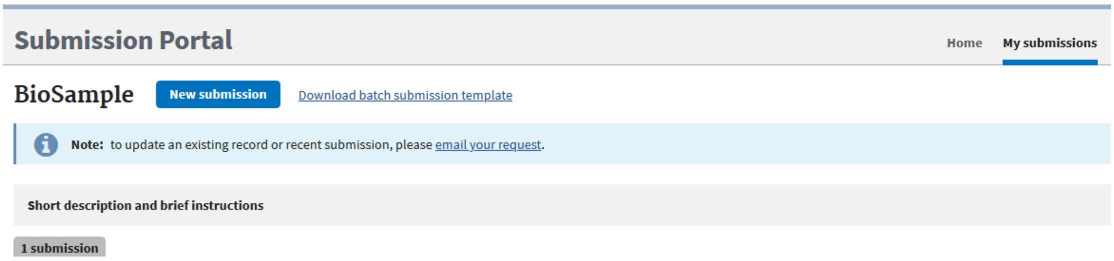
Add another publication

Continue

注：确认无误后，点击“Submit”按钮，创建该 Project。完成以上步骤，经过批准会发送到邮箱里面，获得以 PRJNA 开头的 BioProject ID。

2.2 创建 BioSample ID

1. 打开链接 https://submit.ncbi.nlm.nih.gov/subs/biosample/，点击 New submission。



Submission Portal

Home My submissions

BioSample New submission Download batch submission template

Note: to update an existing record or recent submission, please email your request.

Short description and brief instructions

1 submission

同样是根据项目研究的实际情况，填写信息；填写完成后，点击页面下方的 continue，保存已填写的信息。

2.2.1 SUBMITTER

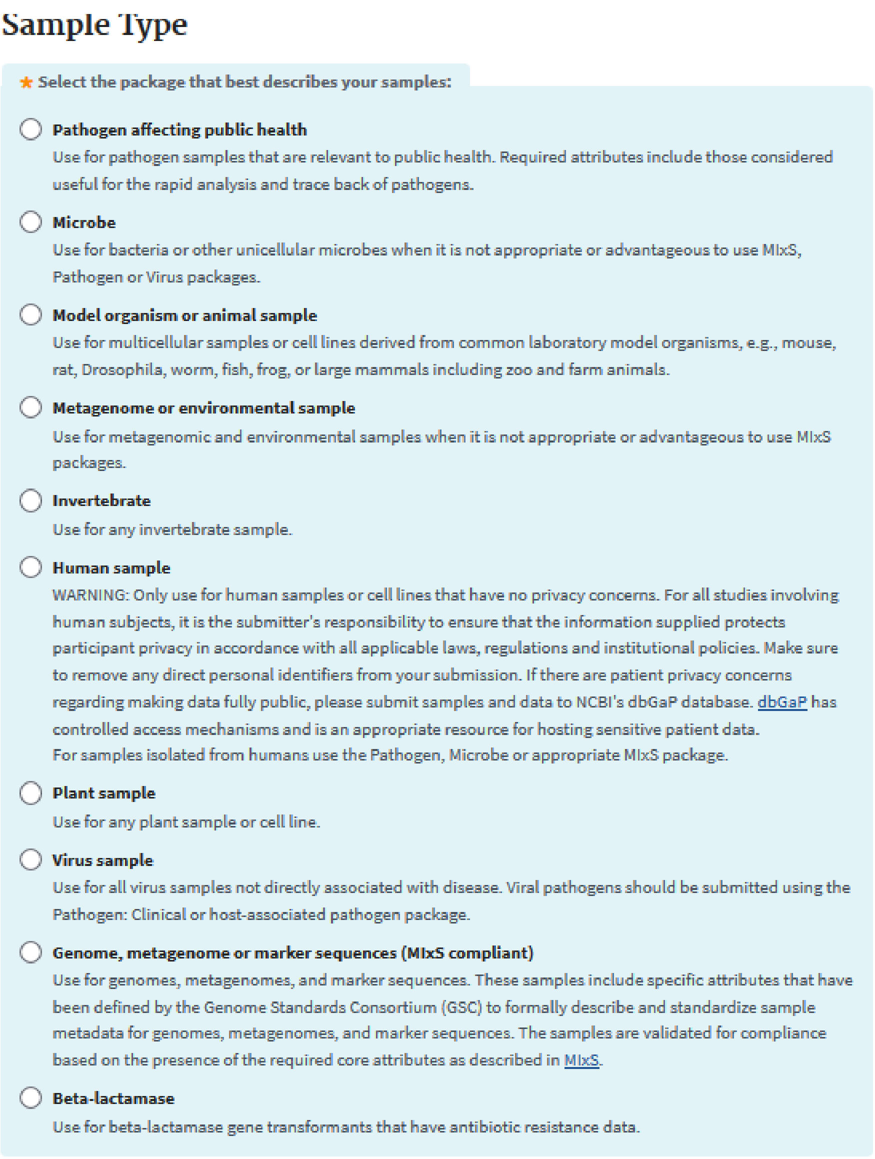
填写个人基本信息，如果已经成功提交 BioProject 会自动填补，无需修改。

2.2.2 GENERAL INFO

Release date，该信息与 BioProject 类似，数据释放时间；选择样本类型，是选择多样本还是单样本上传。

2.2.3 SAMPLE TYPE

根据样品实际情况选择。



Sample Type

★ Select the package that best describes your samples:

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

○ **Microbe**
Use for bacteria or other unicellular microbes when it is not appropriate or advantageous to use MIxS, Pathogen or Virus packages.

○ **Model organism or animal sample**
Use 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.

○ **Metagenome or environmental sample**
Use for metagenomic and environmental samples when it is not appropriate or advantageous to use MIxS packages.

○ **Invertebrate**
Use for any invertebrate sample.

○ **Human sample**
WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is the submitter's responsibility to ensure that the information supplied protects participant privacy in accordance with all applicable laws, regulations and institutional policies. Make sure to remove any direct personal identifiers from your submission. If there are patient privacy concerns regarding making data fully public, please submit samples and data to NCBI's dbGaP database. dbGaP has controlled access mechanisms and is an appropriate resource for hosting sensitive patient data.
For samples isolated from humans use the Pathogen, Microbe or appropriate MIxS package.

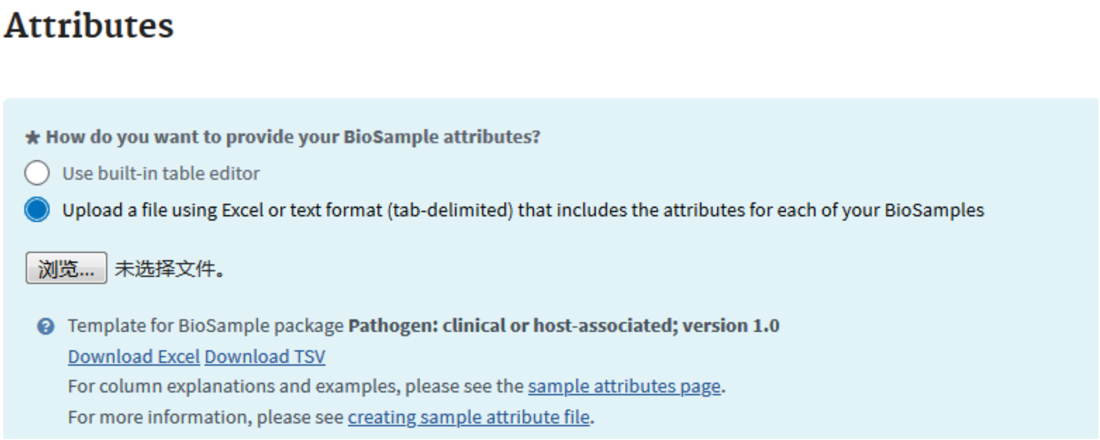
○ **Plant sample**
Use for any plant sample or cell line.

○ **Virus sample**
Use for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen: Clinical or host-associated pathogen package.

○ **Genome, metagenome or marker sequences (MIxS compliant)**
Use for genomes, metagenomes, and marker sequences. These samples include specific attributes that have been defined by the Genome Standards Consortium (GSC) to formally describe and standardize sample metadata for genomes, metagenomes, and marker sequences. The samples are validated for compliance based on the presence of the required core attributes as described in MIxS.

○ **Beta-lactamase**
Use for beta-lactamase gene transformants that have antibiotic resistance data.

2.2.4 ATTRIBUTES



Attributes

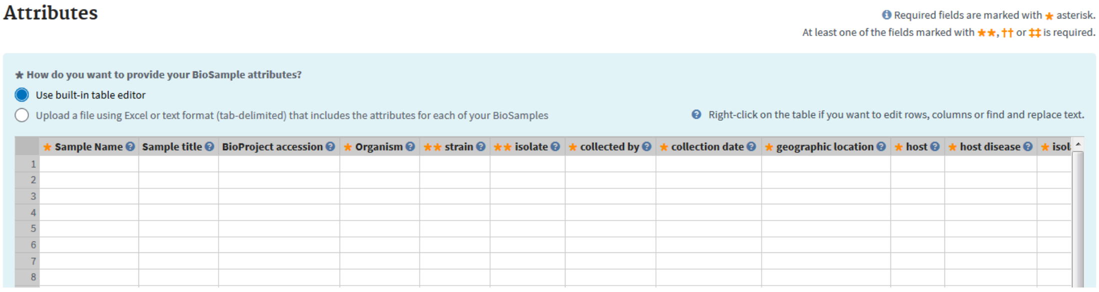
★ **How do you want to provide your BioSample attributes?**

○ Use built-in table editor

● Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples

Template for BioSample package **Pathogen: clinical or host-associated; version 1.0**
Download Excel Download TSV
For column explanations and examples, please see the sample attributes page.
For more information, please see creating sample attribute file.

有两种上传数据方式，点击“Use built-in tableeditor”，可在此直接进行编辑。



Attributes

ℹ Required fields are marked with ★ asterisk.
At least one of the fields marked with ★★, †† or ‡‡ is required.

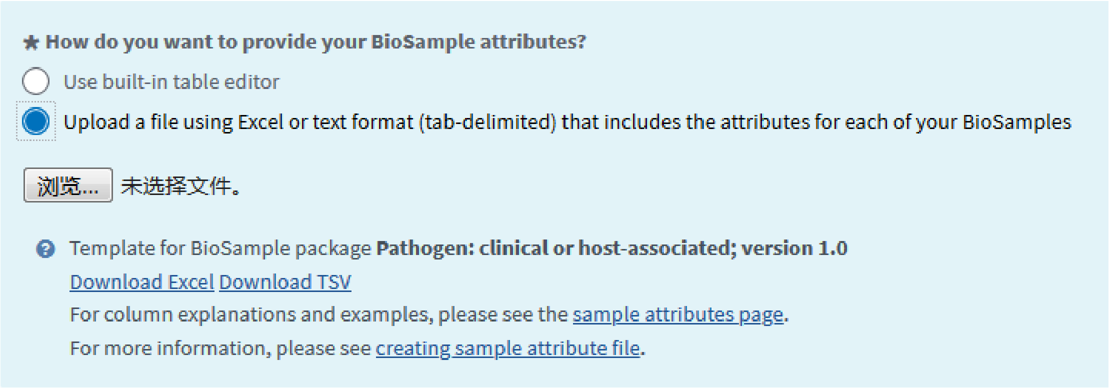
★ How do you want to provide your BioSample attributes?

● Use built-in table editor
○ Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples

❓ Right-click on the table if you want to edit rows, columns or find and replace text.

	★ Sample Name	Sample title	BioProject accession	★ Organism	★★ strain	★★ isolate	★ collected by	★ collection date	★ geographic location	★ host	★ host disease	★ isol
1												
2												
3												
4												
5												
6												
7												
8												

也可根据提示下载 excel 后填写。



★ How do you want to provide your BioSample attributes?

○ Use built-in table editor
● Upload a file using Excel or text format (tab-delimited) that includes the attributes for each of your BioSamples

浏览... 未选择文件。

❓ Template for BioSample package **Pathogen: clinical or host-associated; version 1.0**
Download Excel Download TSV
For column explanations and examples, please see the sample attributes page.
For more information, please see creating sample attribute file.

表格中的绿色是必填项，一定要保证至少一个因子可以区分各个样本（名字除外）。可参考下方链接：

https://submit.ncbi.nlm.nih.gov/biosample/template/?package=Microbe.1.0&action=definition

a. sample_name：样品名；

b. sample_title：每个处理可以写一个题目，可选；

c. description：处理的描述，可选；

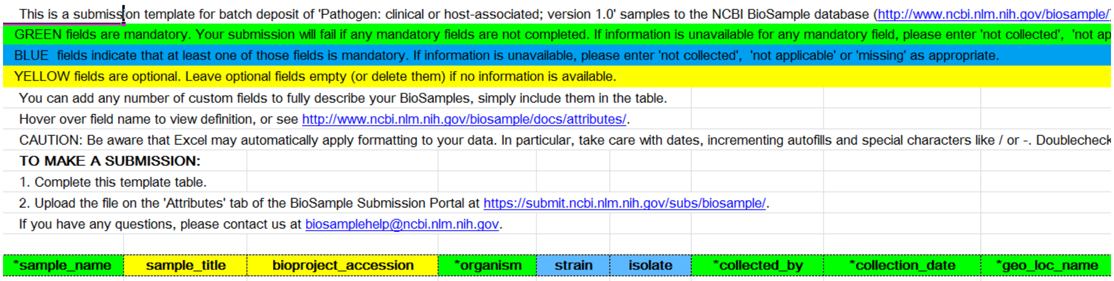
d. organism：优势物种名；

e. collection_date：采样时间，如：2012-08-16；

f. geo_loc_name：采样地，如：China:Beijing；

g. lat_lon：经纬度，如：39 N 116 E；

h. isolation_source：分离环境，如：Rhizosphere soil；



This is a submission template for batch deposit of 'Pathogen: clinical or host-associated; version 1.0' samples to the NCBI BioSample database (http://www.ncbi.nlm.nih.gov/biosample/
GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed. If information is unavailable for any mandatory field, please enter 'not collected', 'not ap
BLUE fields indicate that at least one of those fields is mandatory. If information is unavailable, please enter 'not collected', 'not applicable' or 'missing' as appropriate.
YELLOW fields are optional. Leave optional fields empty (or delete them) if no information is available.
You can add any number of custom fields to fully describe your BioSamples, simply include them in the table.
Hover over field name to view definition, or see http://www.ncbi.nlm.nih.gov/biosample/docs/attributes/.
CAUTION: 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
TO MAKE A SUBMISSION:
1. Complete this template table.
2. Upload the file on the 'Attributes' tab of the BioSample Submission Portal at https://submit.ncbi.nlm.nih.gov/subs/biosample/.
If you have any questions, please contact us at biosamplehelp@ncbi.nlm.nih.gov.

*sample_name	sample_title	bioproject_accession	*organism	strain	isolate	*collected_by	*collection_date	*geo_loc_name

注意：表格提交后，一定要保证没有任何 warnings，否则可能需要等待 2 个工作日才能重新进行该步骤。

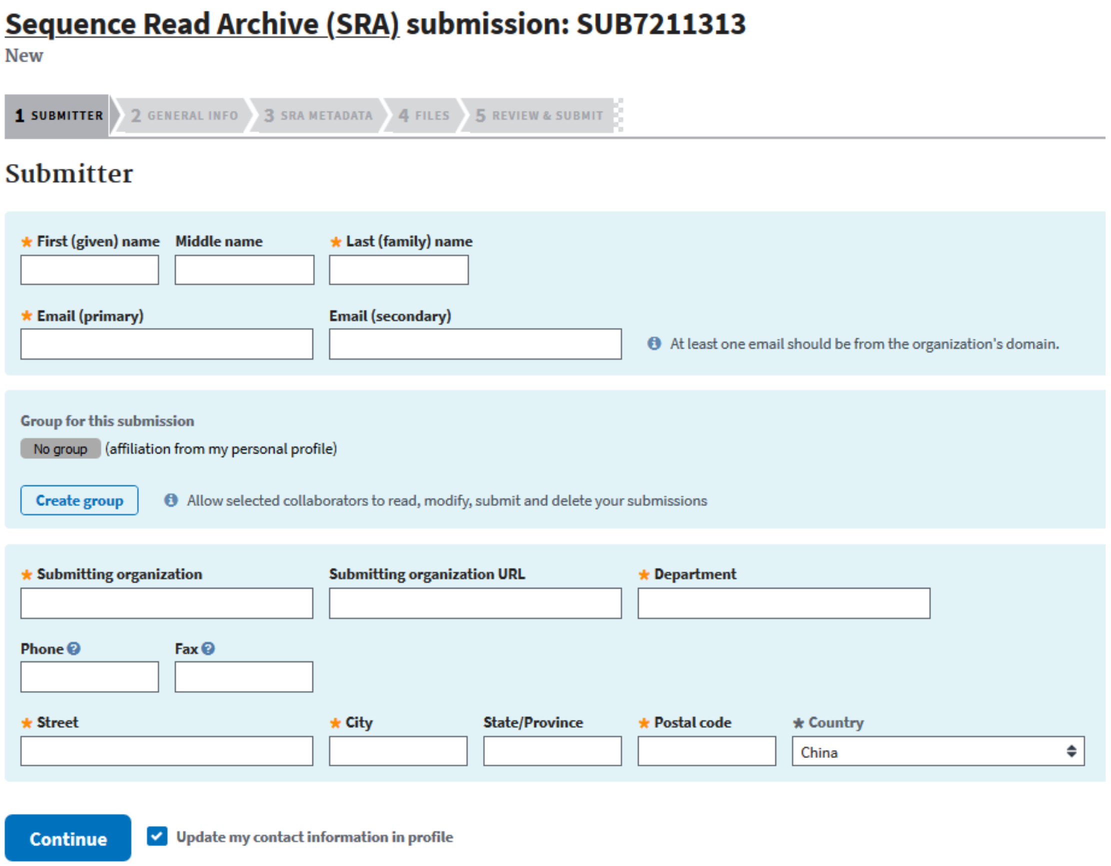
确认无误后，点击最后的“Submit”按钮。经过以上步骤，邮箱会收到以 SAMN 开头的 BioSample ID。

3. 创建 New Submission

直接登录 SRA 网址（https://submit.ncbi.nlm.nih.gov/subs/sra/），创建 New submission。

3.1 SUBMITTER

与 BioProject 相同，需要填写个人基本信息，如果已经成功提交 BioProject 会自动填补，无需修改。



Sequence Read Archive (SRA) submission: SUB7211313
New

Submitter

★ First (given) name
Middle name
★ Last (family) name
★ Email (primary)
Email (secondary)
At least one email should be from the organization's domain.

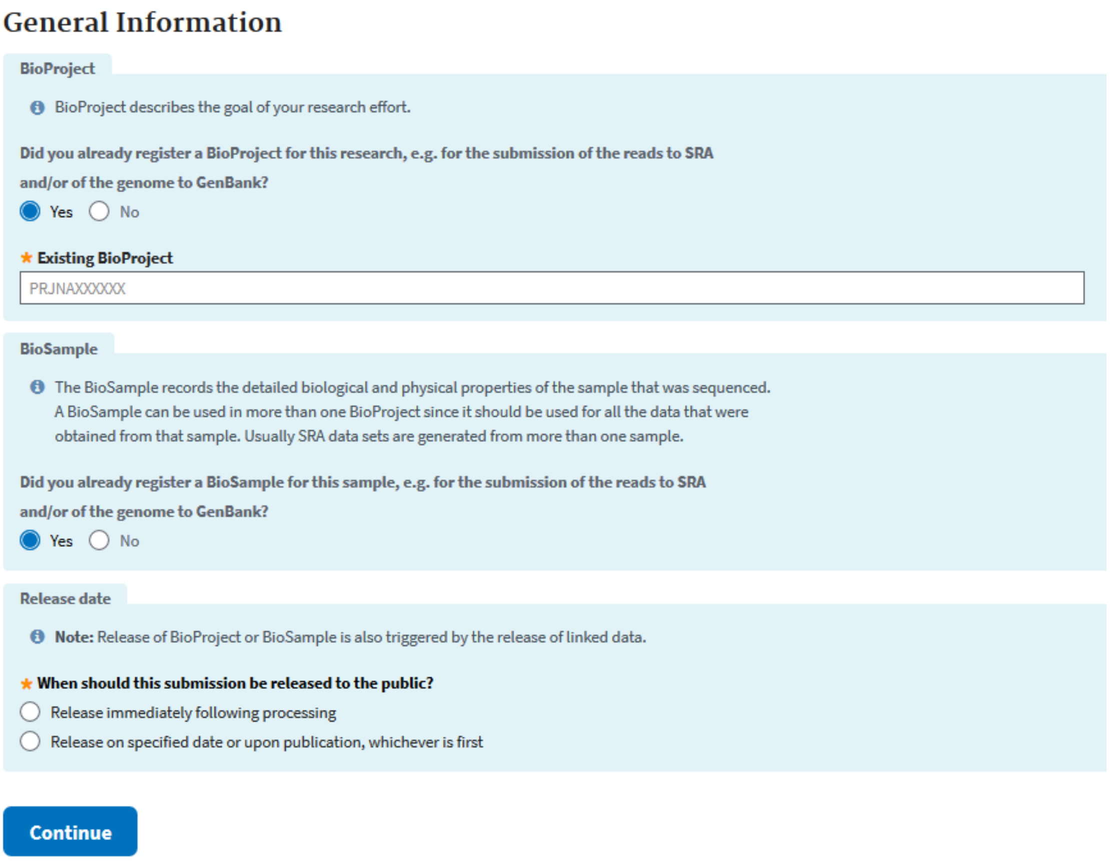
Group for this submission
No group (affiliation from my personal profile)
Create group
Allow selected collaborators to read, modify, submit and delete your submissions

★ Submitting organization
Submitting organization URL
★ Department
Phone
Fax
★ Street
★ City
State/Province
★ Postal code
★ Country: China

Continue
☑ Update my contact information in profile

3.2 GENERAL

直接引用上述创建的BioProject和BioSample ID就可以。此外还需要设置Release date，数据的释放时间一般尽可能选择文章发表之后，后续也可以根据实际需要进行更改。



General Information

BioProject

BioProject describes the goal of your research effort.

Did you already register a BioProject for this research, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?

● Yes ○ No

* Existing BioProject

PRJNAXXXXXX

BioSample

The BioSample records the detailed biological and physical properties of the sample that was sequenced. A BioSample can be used in more than one BioProject since it should be used for all the data that were obtained from that sample. Usually SRA data sets are generated from more than one sample.

Did you already register a BioSample for this sample, e.g. for the submission of the reads to SRA and/or of the genome to GenBank?

● Yes ○ No

Release date

Note: Release of BioProject or BioSample is also triggered by the release of linked data.

* When should this submission be released to the public?

○ Release immediately following processing

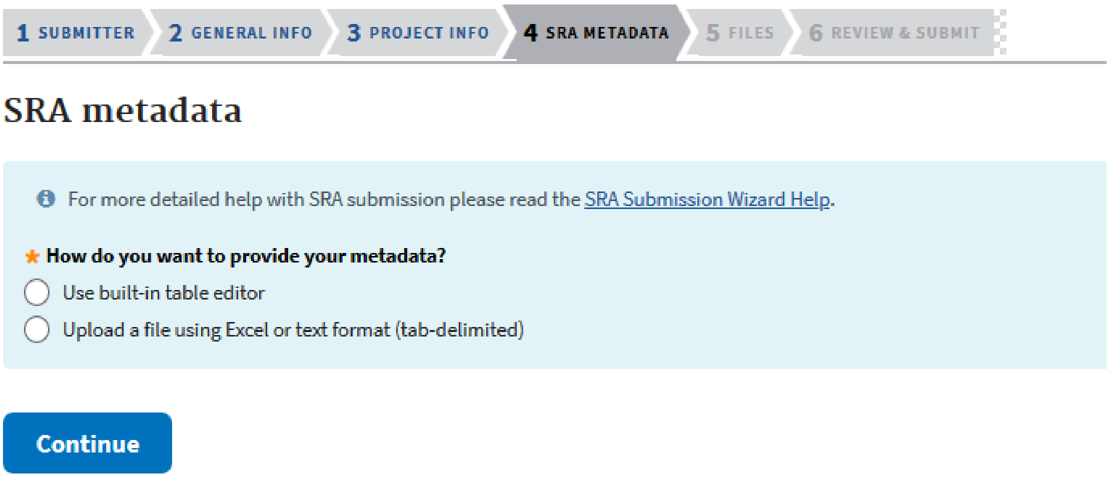
○ Release on specified date or upon publication, whichever is first

3.3 PROJECT INFO

与 BioProject 相同，填写 Project Title 与 Public description。

3.4 SRA METADATA

可选择在线填写或者下载 Excel 表格填写后上传。其中 BioProject 、BioSample 的登陆号 PRJNA#和 SAMN# 是必填的。



1 SUBMITTER 2 GENERAL INFO 3 PROJECT INFO 4 SRA METADATA 5 FILES 6 REVIEW & SUBMIT

SRA metadata

For more detailed help with SRA submission please read the SRA Submission Wizard Help.

★ How do you want to provide your metadata?

○ Use built-in table editor

○ Upload a file using Excel or text format (tab-delimited)

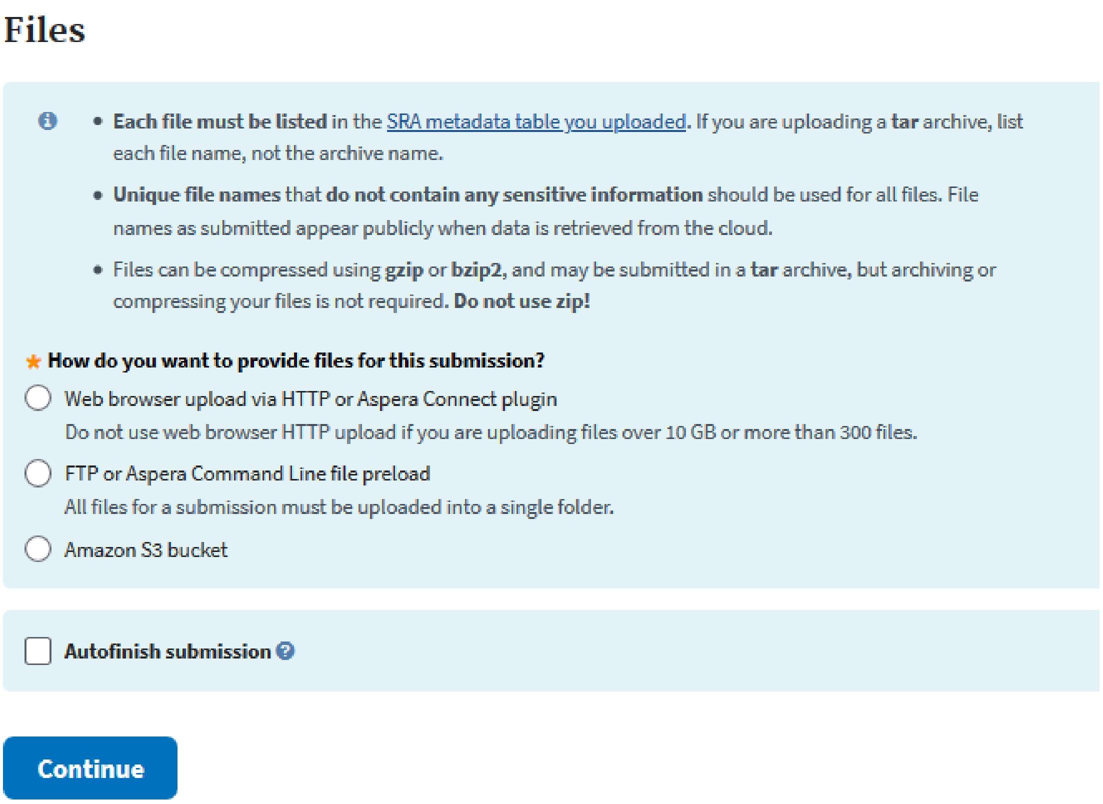
表格中需填写的内容说明如下：

1) library_strategy：测序策略，如 WGS、RNA-seq、miRNA-seq 等；
2) library_source：材料来源，如 GENOMIC、TRANSCRIPTOMIC、METAGENOMIC；
3) library_selection：富集方法，如 PCR、RANDOM 等；
4) library_layout：展示形式，如 Paired、Fragment；
5) platform: 测序平台，如 ILLUMINA、PACBIO_SMRT 等；
6) instrument_model:测序仪器型号，根据测序平台选择，Illumina HiSeq 3000、Illumina HiSeq X Ten、Illumina MiSeq 等；
7) Filetype: 上传数据形式，如 bam、fastq 等。

3.5 FILES

上传数据文件。如果数据量比较小，可以使用在线方式上传。对于数据量较大的项目，可以使用 NCBI 的 Aspera 软件，详细参见链接：

https://www.ncbi.nlm.nih.gov/sra/docs/submitfiles/



Files

- **Each file must be listed** in the SRA metadata table you uploaded. If you are uploading a **tar** archive, list each file name, not the archive name.
- **Unique file names** that **do not contain any sensitive information** should be used for all files. File names as submitted appear publicly when data is retrieved from the cloud.
- Files can be compressed using **gzip** or **bzip2**, and may be submitted in a **tar** archive, but archiving or compressing your files is not required. **Do not use zip!**

★ **How do you want to provide files for this submission?**

○ Web browser upload via HTTP or Aspera Connect plugin
Do not use web browser HTTP upload if you are uploading files over 10 GB or more than 300 files.

○ FTP or Aspera Command Line file preload
All files for a submission must be uploaded into a single folder.

○ Amazon S3 bucket

□ **Autofinish submission**

注意：上传文件支持 tar、tar.gz、tgz、tar.bz2、tbz2、gz 等格式。

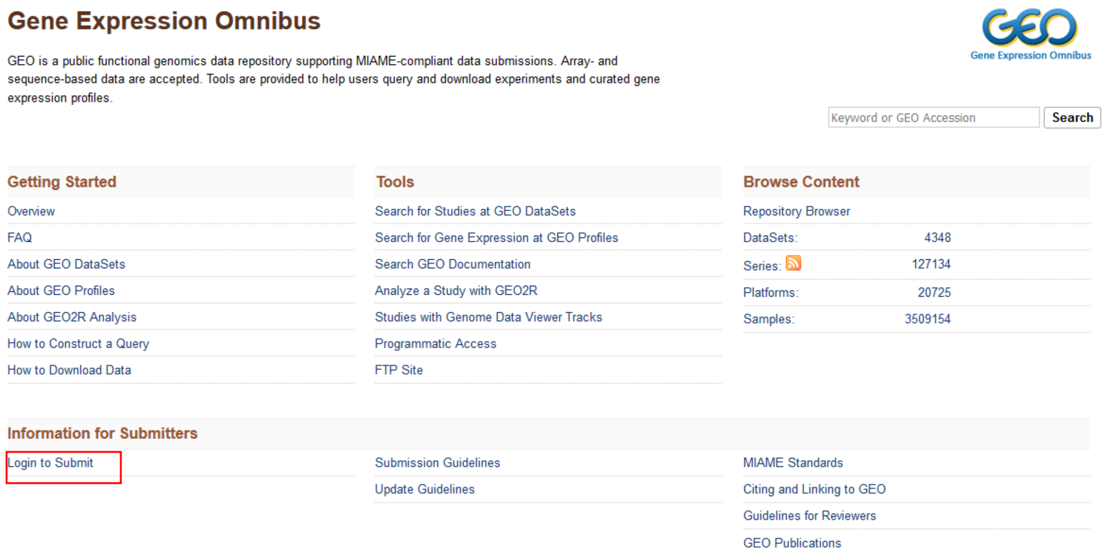
3.6 REVIEW & SUBMIT

核查提交信息，确认无误后，点击“Submit”。后续邮箱中会收到相应的 Accession number 的登录号（SRR*****或者 SRA*******），用于查询和检索。如果您在上传的过程中遇到技术问题，可以联系 sra@ncbi.nlm.nih.gov 寻求帮助。

三、 数据提交 GEO 数据库

1. 注册 GEO 账号

如果要上传 GEO 数据库，与提交 SRA 数据相同，首先要建立一个 NCBI 的账号。然后需要注册一个 GEO 的账号，可以从 GEO 首页（https://www.ncbi.nlm.nih.gov/geo/）左下角的 Login to Submit 进入创建。



Gene Expression Omnibus

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.

Keyword or GEO Accession | Search

Getting Started

- Overview
- FAQ
- About GEO DataSets
- About GEO Profiles
- About GEO2R Analysis
- How to Construct a Query
- How to Download Data

Tools

- Search for Studies at GEO DataSets
- Search for Gene Expression at GEO Profiles
- Search GEO Documentation
- Analyze a Study with GEO2R
- Studies with Genome Data Viewer Tracks
- Programmatic Access
- FTP Site

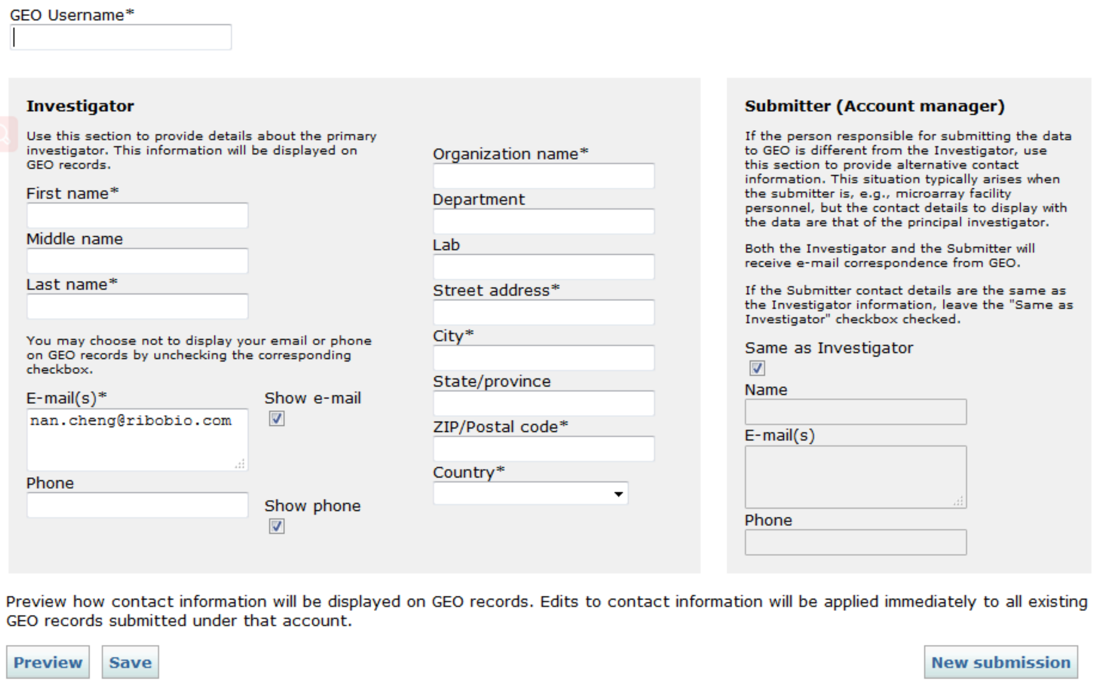
Browse Content

Repository Browser

DataSets:	4348
Series:	127134
Platforms:	20725
Samples:	3509154

Information for Submitters

- Login to Submit
- Submission Guidelines
- Update Guidelines
- MIAME Standards
- Citing and Linking to GEO
- Guidelines for Reviewers
- GEO Publications



GEO Username*

Investigator

Use this section to provide details about the primary investigator. This information will be displayed on GEO records.

First name*

Middle name

Last name*

You may choose not to display your email or phone on GEO records by unchecking the corresponding checkbox.

E-mail(s)*: nan.cheng@ribobio.com — Show e-mail ☑

Phone — Show phone ☑

Organization name*

Department

Lab

Street address*

City*

State/province

ZIP/Postal code*

Country*

Submitter (Account manager)

If the person responsible for submitting the data to GEO is different from the Investigator, use this section to provide alternative contact information. This situation typically arises when the submitter is, e.g., microarray facility personnel, but the contact details to display with the data are that of the principal investigator.

Both the Investigator and the Submitter will receive e-mail correspondence from GEO.

If the Submitter contact details are the same as the Investigator information, leave the "Same as Investigator" checkbox checked.

Same as Investigator ☑

Name

E-mail(s)

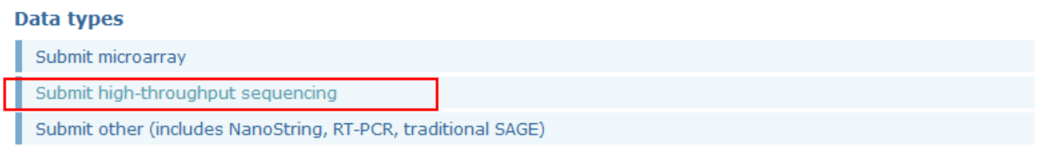
Phone

Preview how contact information will be displayed on GEO records. Edits to contact information will be applied immediately to all existing GEO records submitted under that account.

创建完成后，点击 Save 保存信息，再点击 New submission 进入 GEO 主页。

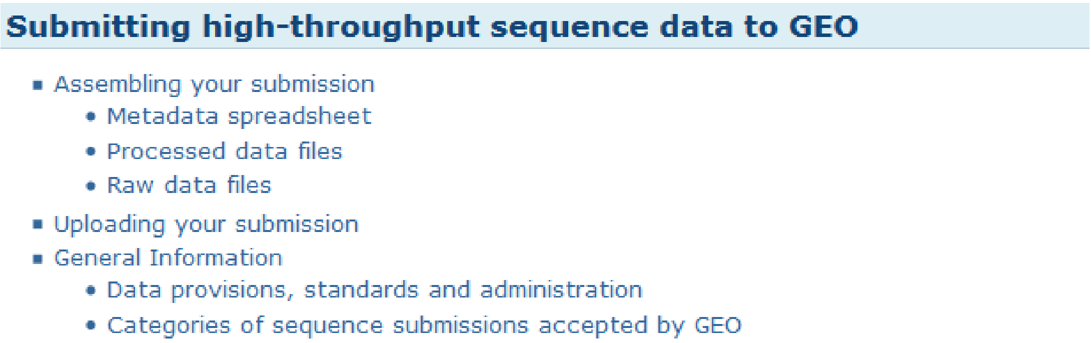
2. 上传数据

接下来，选择你要上传的数据类型，这里只介绍上传转录组测序数据。点击 High-throughput sequence submissions。



Data types

- Submit microarray
- Submit high-throughput sequencing
- Submit other (includes NanoString, RT-PCR, traditional SAGE)



Submitting high-throughput sequence data to GEO

- Assembling your submission
  - Metadata spreadsheet
  - Processed data files
  - Raw data files
- Uploading your submission
- General Information
  - Data provisions, standards and administration
  - Categories of sequence submissions accepted by GEO

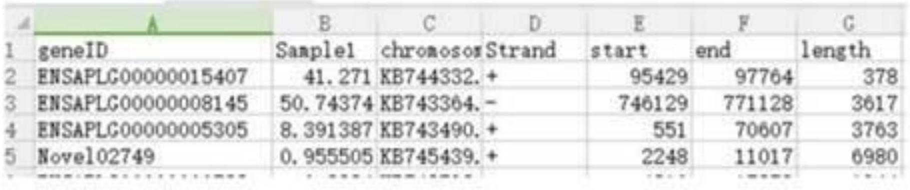
2.1 上传文件类型

上传总共需要 3 类文件：

1.Metadata spreadsheet 上传所需要填写的表格，将在下文中详细介绍；

2.Processed data files 基因表达量文件，如原始 count 文件、校正后的表达值文件（包含校正 count 值、RPKM、FPKM、TPM 等）；

如下图所示。如果有新预测基因或转录本，除 geneID 及样品表达值外，还需提供：1）Chromosome（染色体号）；2）Strand（链的正负）；3）start（起始位置）；4）end（终止位置）；5）长度 length（长度）等信息。如果没有新基因，只需要提供 A、B 列即可。



	A	B	C	D	E	F	G
1	geneID	Sample1	chromosom	Strand	start	end	length
2	ENSAPLG00000015407	41.271	KB744332.	+	95429	97764	378
3	ENSAPLG00000008145	50.74374	KB743364.	-	746129	771128	3617
4	ENSAPLG00000005305	8.391387	KB743490.	+	551	70607	3763
5	Novel02749	0.955505	KB745439.	+	2248	11017	6980

表达值文件可以以表达矩阵表格的形式或单独的文件形式提供。

CHIP 测序支持 WIG、bigWig、bedGraph 格式。

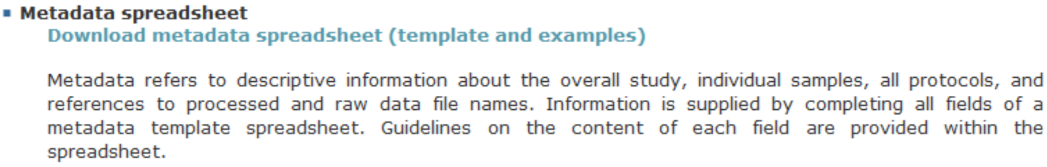
3.Raw data files（原始的测序数据，如 fastq 文件）

更多说明文件格式说明请参考：

https://www.ncbi.nlm.nih.gov/geo/info/seq.html

2.2 Metadata spreadsheet 介绍

进入 High-throughput sequence submissions 页面后，下载 Metadata spreadsheet（Download metadata spreadsheet (template and examples)）。



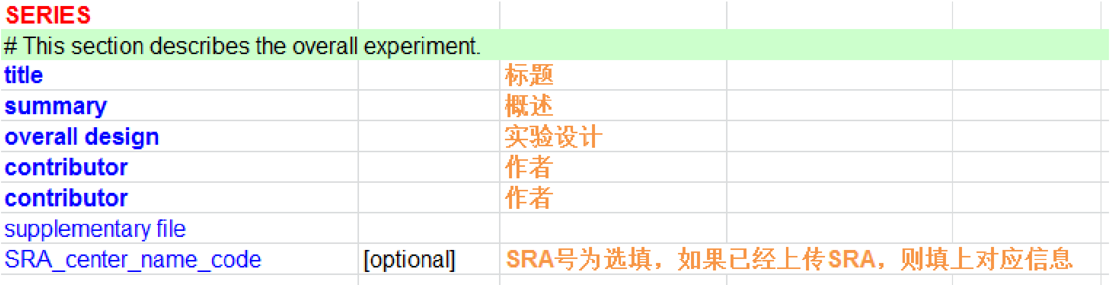
- **Metadata spreadsheet**

  Download metadata spreadsheet (template and examples)

  Metadata refers to descriptive information about the overall study, individual samples, all protocols, and references to processed and raw data file names. Information is supplied by completing all fields of a metadata template spreadsheet. Guidelines on the content of each field are provided within the spreadsheet.

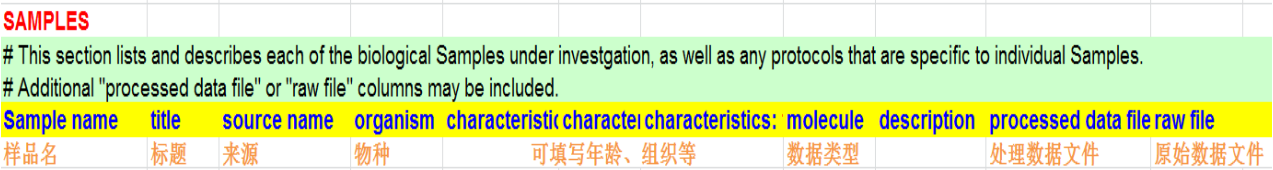
打开该表格后，共包含 7 个部分：

1）SERIES：跟文章相关的内容：标题，摘要，实验设计，参与者（根据自己情况填写）；



SERIES				
# This section describes the overall experiment.				
title		标题		
summary		概述		
overall design		实验设计		
contributor		作者		
contributor		作者		
supplementary file				
SRA_center_name_code	[optional]	SRA号为选填，如果已经上传SRA，则填上对应信息		

2）SAMPLES：跟样本信息相关的内容：样本名称，物种，特征信息及对应的处理文件（表达值数据文件等）和原始数据（fastq 等）；原始数据文件及处理文件同上文描述。



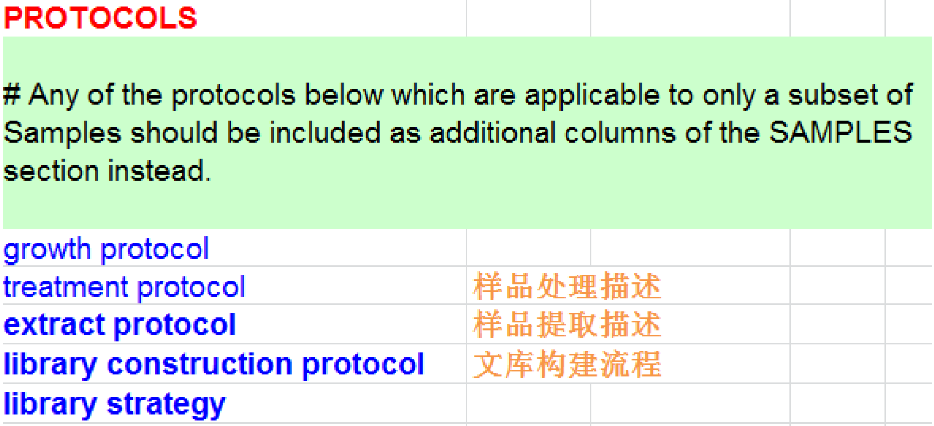
SAMPLES

This section lists and describes each of the biological Samples under investgation, as well as any protocols that are specific to individual Samples.

Additional "processed data file" or "raw file" columns may be included.

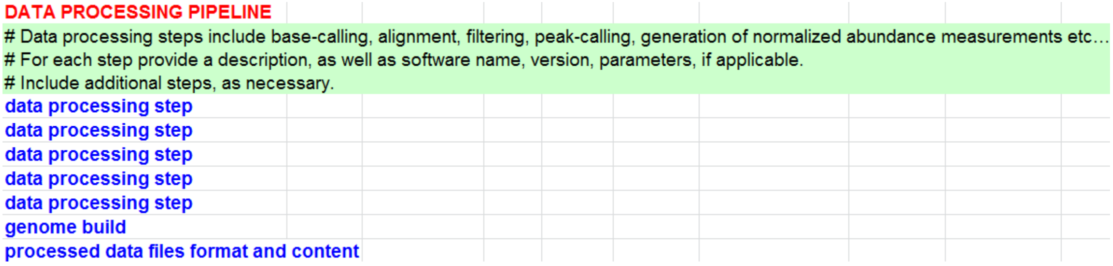
Sample name	title	source name	organism	characteristic	characte	characteristics:	molecule	description	processed data file	raw file
样品名	标题	来源	物种	可填写年龄、组织等			数据类型		处理数据文件	原始数据文件

3）PROTOCOLS：样本的处理提取和文库构建的描述，如样本提取及建库是锐博生物操作，该部分信息可提供给您参考；



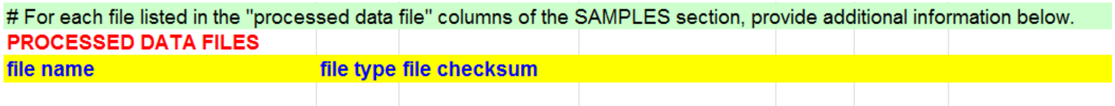
PROTOCOLS				
# Any of the protocols below which are applicable to only a subset of Samples should be included as additional columns of the SAMPLES section instead.				
growth protocol				
treatment protocol	样品处理描述			
extract protocol	样品提取描述			
library construction protocol	文库构建流程			
library strategy				

4）DATA PROCESSING PIPELINE:数据处理方面的描述，如数据预处理，数据比对，采用的基因组版本等；锐博生物会提供不同测序类型提供的描述文件供您参考。



DATA PROCESSING PIPELINE
Data processing steps include base-calling, alignment, filtering, peak-calling, generation of normalized abundance measurements etc…
For each step provide a description, as well as software name, version, parameters, if applicable.
Include additional steps, as necessary.
data processing step
data processing step
data processing step
data processing step
data processing step
genome build
processed data files format and content

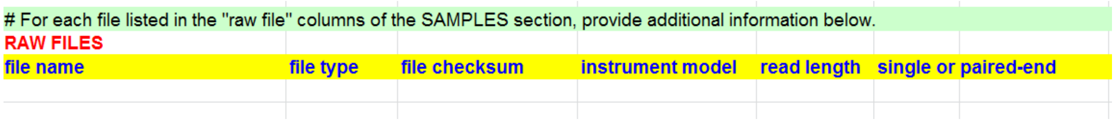
5）ROCESSED DATA FILES:处理后数据名称、格式及 MD5 码。RNA 测序中即可填写表达值文件，其中 file type 一列可以统一写成 abundance measurements。file checksum 列为 MD5 码。锐博提供了所有结果文件的 MD5 码文件（md5.txt），存放于 custom 文件夹下。



For each file listed in the "processed data file" columns of the SAMPLES section, provide additional information below.
PROCESSED DATA FILES

file name	file type	file checksum

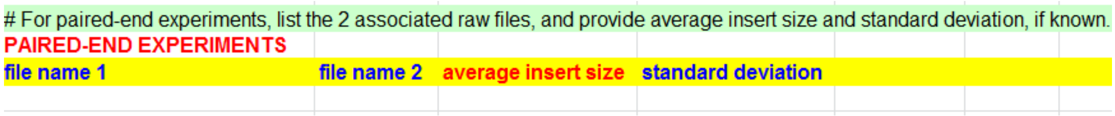
6）RAW FILES:原始数据名称、数据格式、MD5 码、平台类型、测序读长及单双端信息，平台类型、测序读长及单双端信息可由锐博生物提供；



For each file listed in the "raw file" columns of the SAMPLES section, provide additional information below.
RAW FILES

file name	file type	file checksum	instrument model	read length	single or paired-end

7）PAIRED-END EXPERIMENTS:如果是双端测序，还需要填写对应双端原始数据的名称、插入片段长度及插入长度的标准偏差，这部分内容是非必填项。



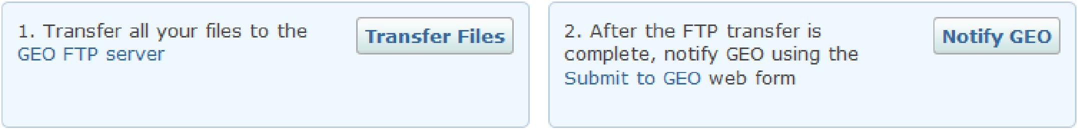
For paired-end experiments, list the 2 associated raw files, and provide average insert size and standard deviation, if known.
PAIRED-END EXPERIMENTS

file name 1	file name 2	average insert size	standard deviation

到这里 METADATA TEMPLATE 算是填写完成了，接下来就可以进行数据上传步骤。

2.3 数据上传

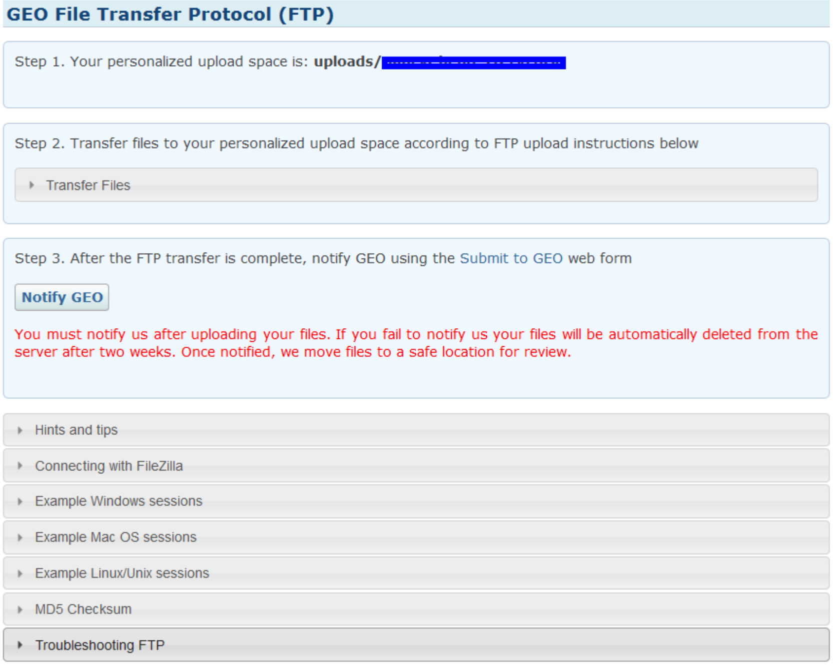
数据上传主要包含两个步骤：



1. Transfer all your files to the GEO FTP server

2. After the FTP transfer is complete, notify GEO using the Submit to GEO web form

点击 Transfer Files 进入数据传输页面（https://www.ncbi.nlm.nih.gov/geo/info/submissionftp.html），网页中针对 windows、MAC、Linux 系统上传数据都有详细说明，下面以 windows 系统为例进行说明。



GEO File Transfer Protocol (FTP)

Step 1. Your personalized upload space is: **uploads/**[redacted]

Step 2. Transfer files to your personalized upload space according to FTP upload instructions below

▸ Transfer Files

Step 3. After the FTP transfer is complete, notify GEO using the Submit to GEO web form

Notify GEO

You must notify us after uploading your files. If you fail to notify us your files will be automatically deleted from the server after two weeks. Once notified, we move files to a safe location for review.

▸ Hints and tips

▸ Connecting with FileZilla

▸ Example Windows sessions

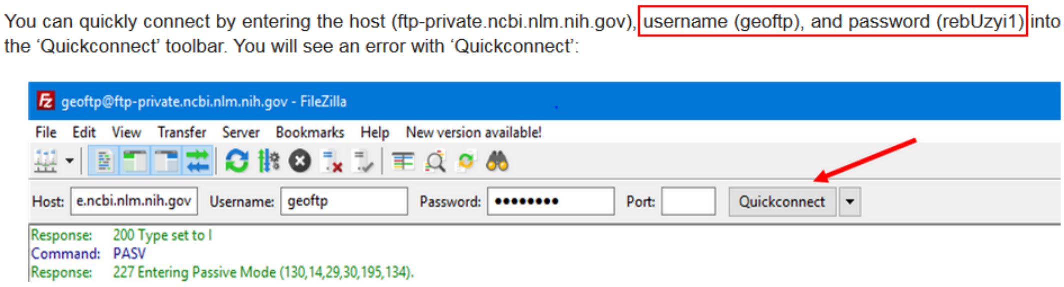
▸ Example Mac OS sessions

▸ Example Linux/Unix sessions

▸ MD5 Checksum

▸ Troubleshooting FTP

首先，需要下载Filezilla软件，然后在Filezilla中输入GEO地址：ftp-private.ncbi.hlm.nih.gov并登陆（用户名和密码参考上链接中Connecting with FileZilla部分），即可连接GEO数据库进行上传了。



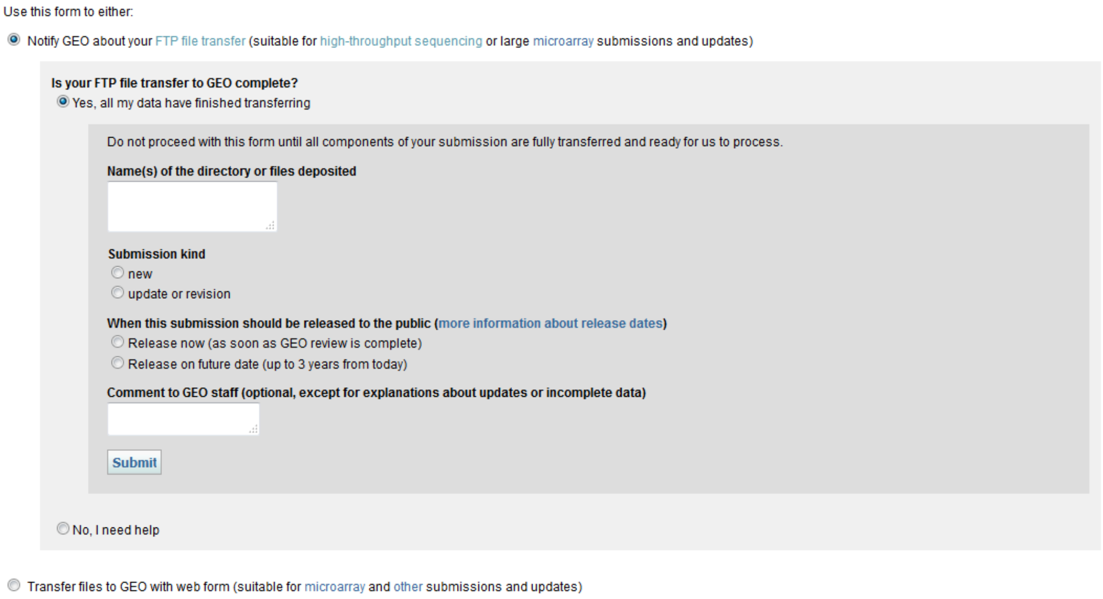
You can quickly connect by entering the host (ftp-private.ncbi.nlm.nih.gov), username (geoftp), and password (rebUzyi1) into the 'Quickconnect' toolbar. You will see an error with 'Quickconnect':

geoftp@ftp-private.ncbi.nlm.nih.gov - FileZilla

Host: e.ncbi.nlm.nih.gov Username: geoftp Password: •••••••• Port: Quickconnect

```
Response:   200 Type set to I
Command:    PASV
Response:   227 Entering Passive Mode (130,14,29,30,195,134).
```

等待数据都上传完成后就可以点击“Notify GEO”通知GEO数据上传完成。



Use this form to either:

● Notify GEO about your FTP file transfer (suitable for high-throughput sequencing or large microarray submissions and updates)

Is your FTP file transfer to GEO complete?

● Yes, all my data have finished transferring

Do not proceed with this form until all components of your submission are fully transferred and ready for us to process.

Name(s) of the directory or files deposited

Submission kind

○ new

○ update or revision

When this submission should be released to the public (more information about release dates)

○ Release now (as soon as GEO review is complete)

○ Release on future date (up to 3 years from today)

Comment to GEO staff (optional, except for explanations about updates or incomplete data)

○ No, I need help

○ Transfer files to GEO with web form (suitable for microarray and other submissions and updates)

接下来两天内应该会收到GEO的回复邮件，告知您数据对应的GEO号。

四、结语

感谢您选择锐博生物，如果您在上传数据方面遇到问题，也可以联系我们的技术支持或销售寻求帮助。祝您科研顺利！