高通量测序数据上传指南

很多老师在写文章的时候，精力通常会放在文章中的技术或者实验细节，而疏忽原始数据上传的工作。文章准备接收时，审稿人需要老师提供文章中使用数据的登录号，一般时间要求较紧，对于没接触过数据上传的老师来说，可能会成为文章接收最后的绊脚石，为此我们撰写了一份高通量测序数据上传指南，可以让老师快速掌握数据上传的技能，助力老师文章发表。

目前高通量测序原始数据通常上传到 NCBI 的 SRA(The Sequence Read Archive)数据库，部分老师会纠结是否需要把测序数据上传至 GEO 数据库，这里无需多虑，因为测序数据上传到 GEO 之后，最终也是存放在 SRA；而且 SRA 相对于 GEO 来说，上传过程更快速、便捷。

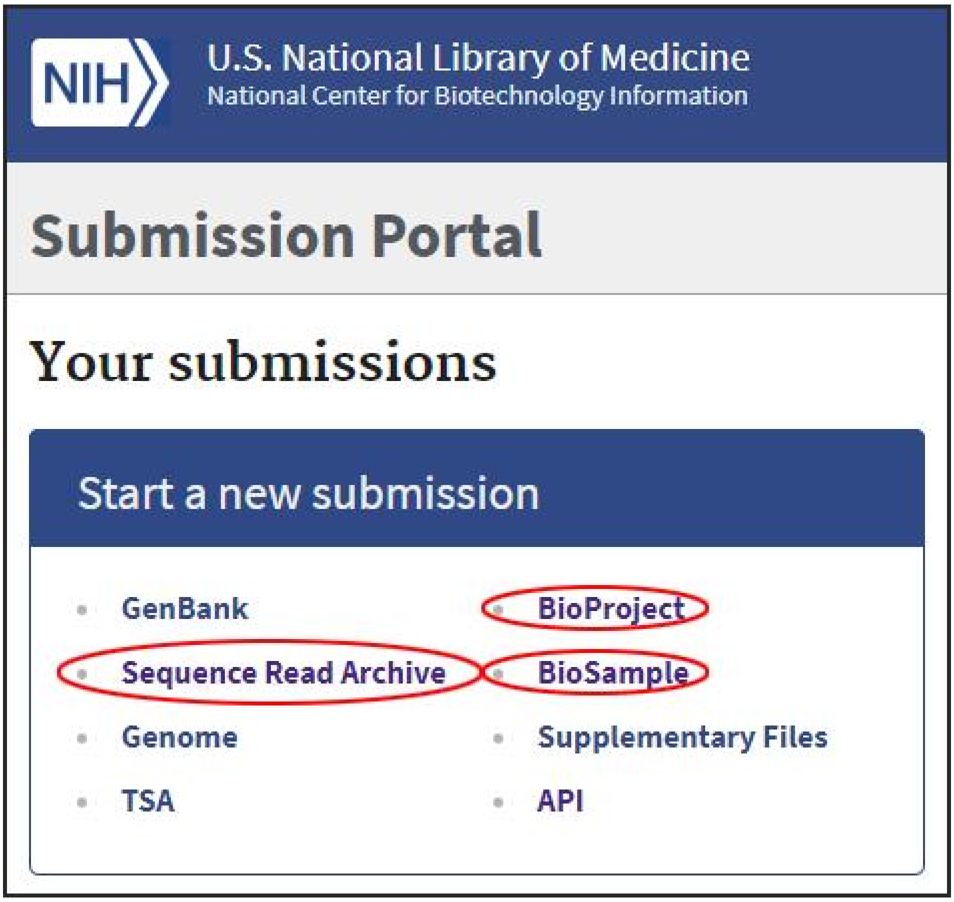
本指南以上传 SRA 数据库为例，整体篇幅以截图为主，并辅以文字描述，方便老师对照网页操作。

在数据上传过程中可能会要求填写实验细节，有些可能不明确或者不适用，如果没有特殊需要，部分细节可以不用过于纠结。数据成功上传后会生成一个 Accession Number，最终加入文章中，审稿人或者其他人能够通过该 Accession Number 查询、下载到对应数据。

1 登录/注册

1.1 进入 NCBI 数据上传主页

主页链接
https://submit.ncbi.nlm.nih.gov/subs/，如图：



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

Submission Portal

Your submissions

Start a new submission

- GenBank
- Sequence Read Archive
- Genome
- TSA
- BioProject
- BioSample
- Supplementary Files
- API

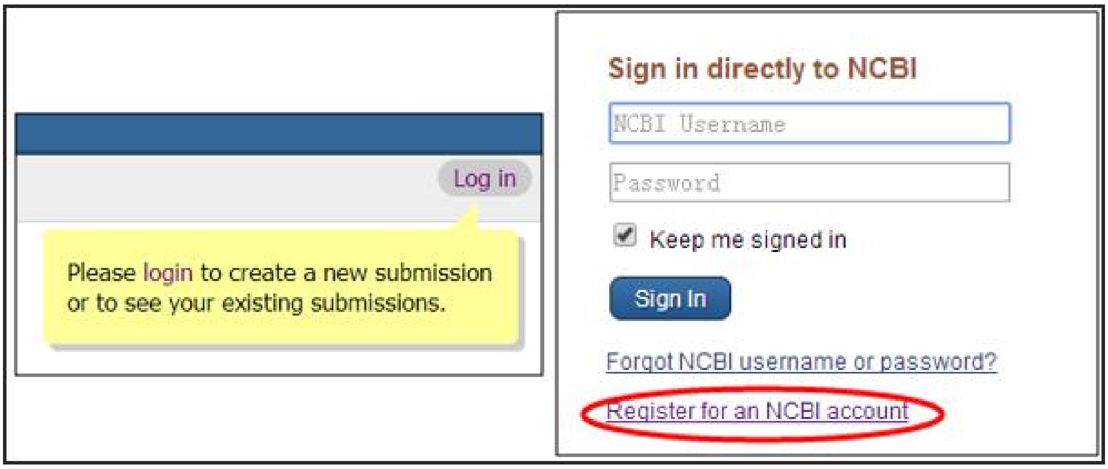
Sequence Read Archive：填写样本信息、上传样本数据

BioProject：填写项目信息

BioSample ：填写样本属性

1.2 Login in/Register (有账号可以跳过 1.2 和 1.3)

点击右上角 Login，如果在 NCBI 注册过，可以输入账号密码登陆，如果未注册可以点击 Register for NCBI account 注册：



Log in

Please login to create a new submission or to see your existing submissions.

Sign in directly to NCBI

NCBI Username

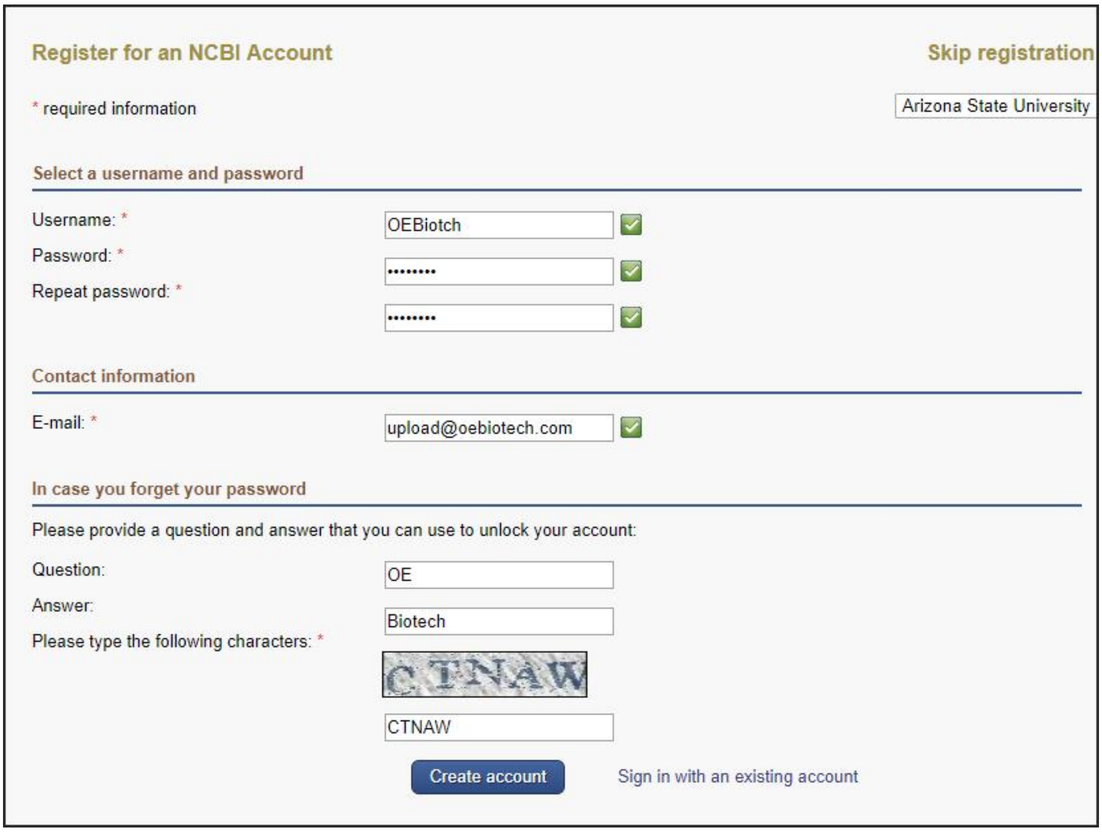
Password

☑ Keep me signed in

Sign In

Forgot NCBI username or password?

Register for an NCBI account



Register for an NCBI Account

Skip registration

* required information

Arizona State University

Select a username and password

Username: * OEBiotch ✓

Password: * •••••••• ✓

Repeat password: * •••••••• ✓

Contact information

E-mail: * upload@oebiotech.com ✓

In case you forget your password

Please provide a question and answer that you can use to unlock your account:

Question: OE

Answer: Biotech

Please type the following characters: * CTNAW

CTNAW

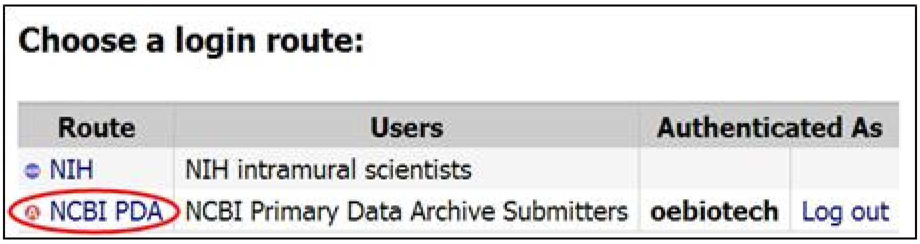
Create account

Sign in with an existing account

1.3 验证邮箱

注册完毕后会向所填写的邮箱中发送验证邮件，点击邮件中的链接即可激活账号。再次回到第二步之后的界面，此时会自动登录。

注：提交过程中如果出现以下的登陆方式选择，一律选择 NCBI PDA 登陆



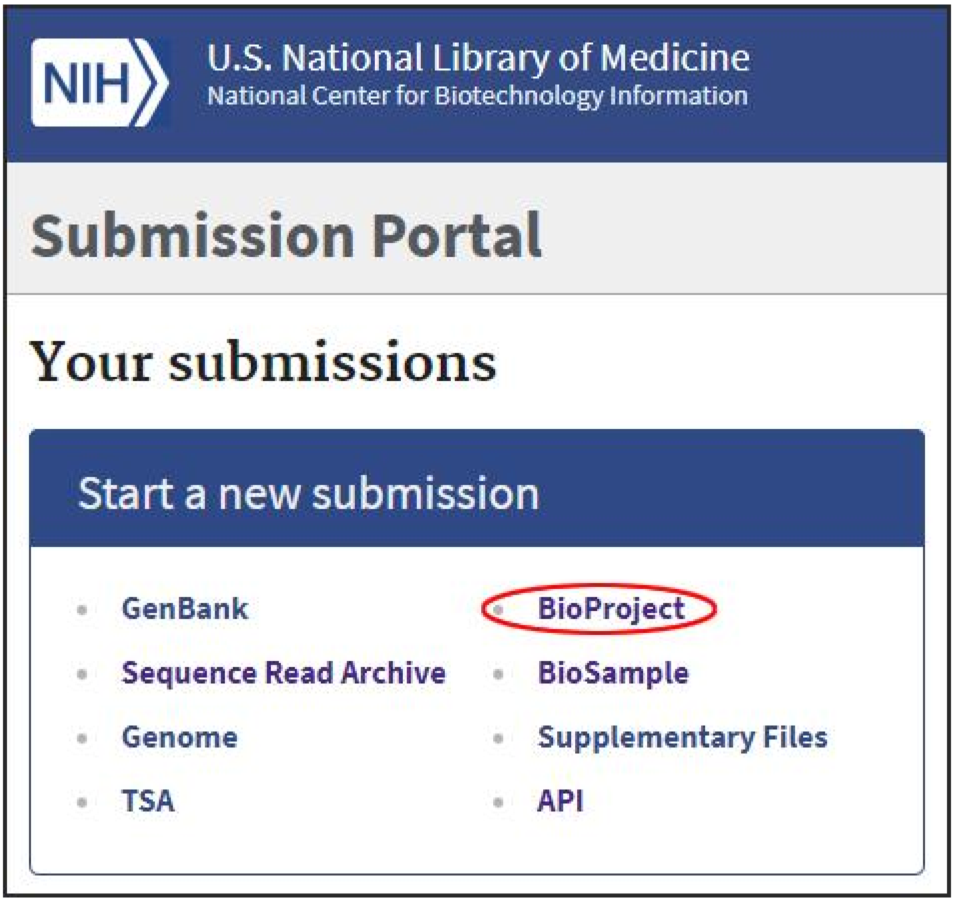
Choose a login route:

Route	Users	Authenticated As	
● NIH	NIH intramural scientists		
● NCBI PDA	NCBI Primary Data Archive Submitters	oebiotech	Log out

2 创建研究项目-BioProject

回到上传主页

https://submit.ncbi.nlm.nih.gov/subs/，点击 BioProject：



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

Submission Portal

Your submissions

Start a new submission

- GenBank
- Sequence Read Archive
- Genome
- TSA
- BioProject
- BioSample
- Supplementary Files
- API

页面跳转后点击 New submission：


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

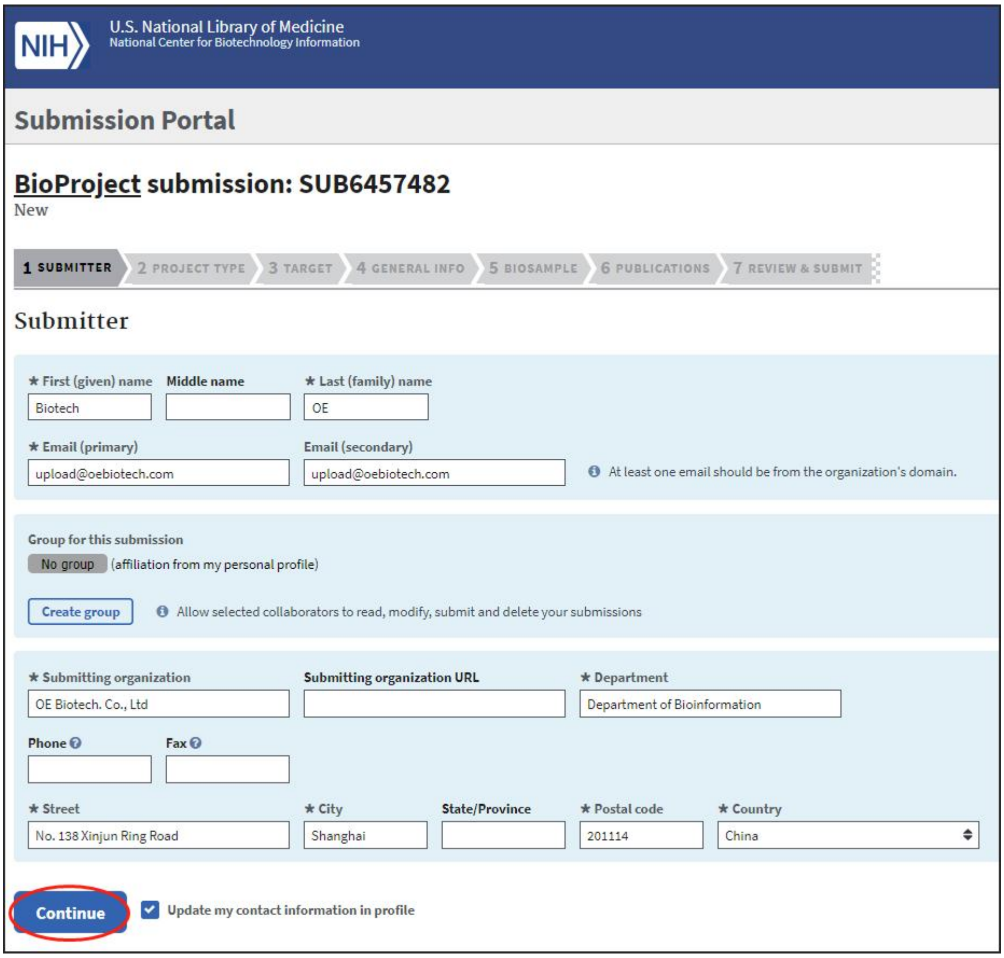
Submission Portal

BioProject

ATTN: to update an existing record or recent submission, please email your request with your BioProject ID or Submission ID included. **Do not** create new submission to update an existing submission!

2.1 SUBMITTER

进入创建 BioProject 页面 Submitter 选项卡，填写必要个人信息：


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

Submission Portal

BioProject submission: SUB6457482
New

Submitter

★ First (given) name: Biotech
Middle name:
★ Last (family) name: OE
★ Email (primary): upload@oebiotech.com
Email (secondary): upload@oebiotech.com
At least one email should be from the organization's domain.

Group for this submission
No group (affiliation from my personal profile)

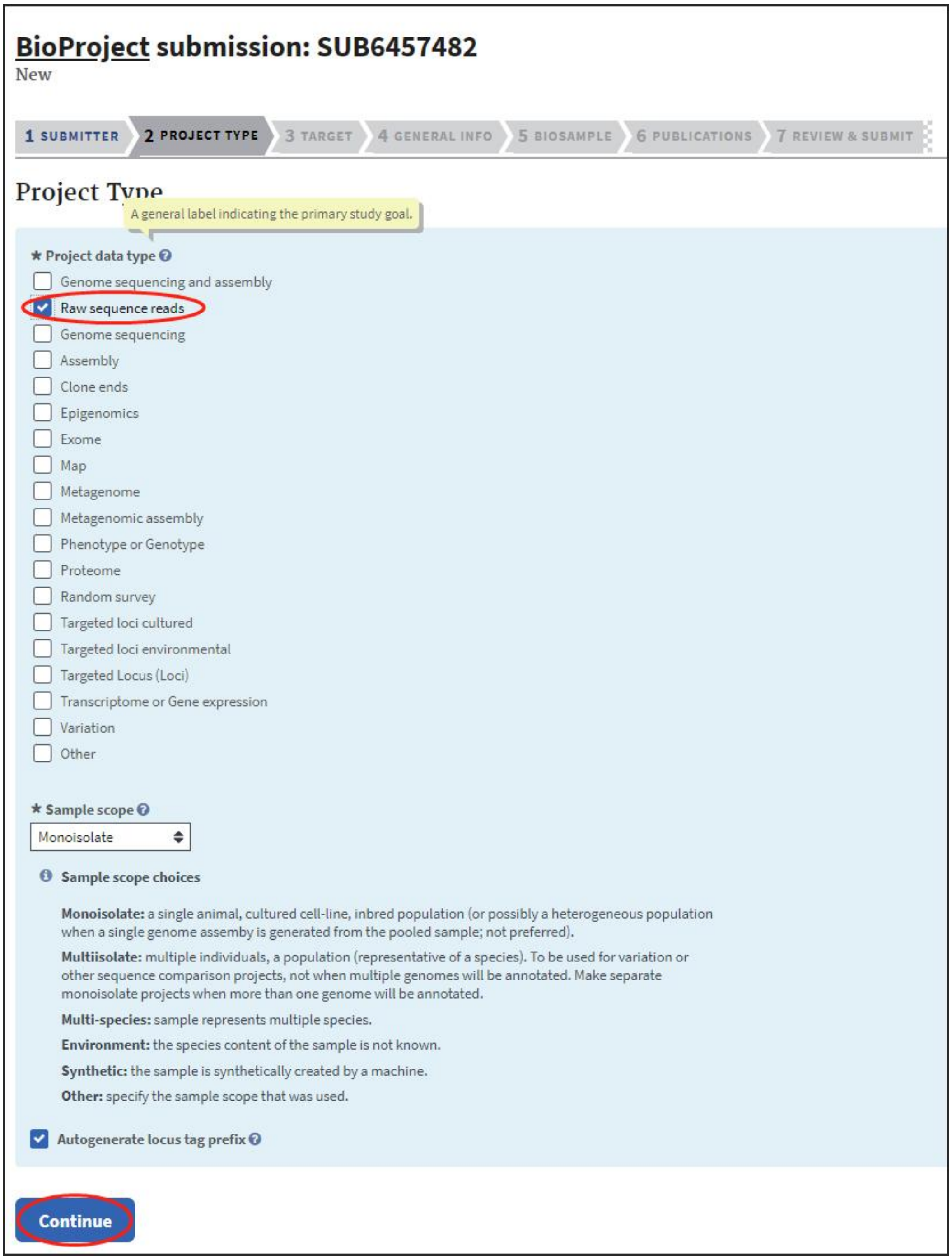
Allow selected collaborators to read, modify, submit and delete your submissions

★ Submitting organization: OE Biotech. Co., Ltd
Submitting organization URL:
★ Department: Department of Bioinformation
Phone:
Fax:
★ Street: No. 138 Xinjun Ring Road
★ City: Shanghai
State/Province:
★ Postal code: 201114
★ Country: China

☑ Update my contact information in profile

2.2 PROJECT TYPE

在 Project Type 选项卡中依次选择相应的类别，此处以普通转录组测序为例，勾选原始测序数据：



BioProject submission: SUB6457482

New

Project Type

A general label indicating the primary study goal.

★ 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
- ☐ Targeted loci cultured
- ☐ Targeted loci environmental
- ☐ Targeted Locus (Loci)
- ☐ Transcriptome or Gene expression
- ☐ Variation
- ☐ 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 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.

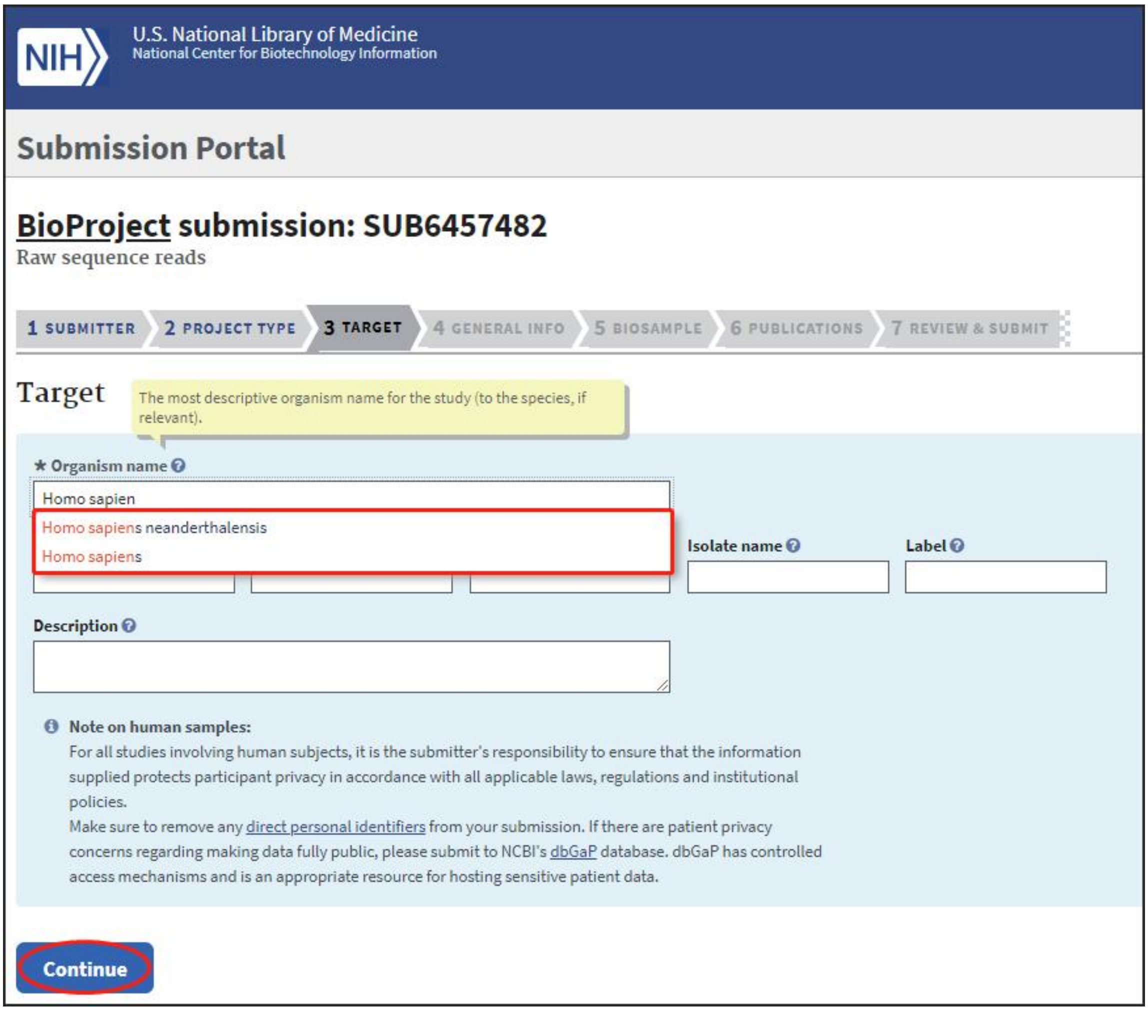
Synthetic: the sample is synthetically created by a machine.

Other: specify the sample scope that was used.

☑ **Autogenerate locus tag prefix**

2.3 TARGET

进入 Target 选项卡，填写物种名称（此处填写的物种拉丁文名称需在 NCBI 中有收录，输入关键词后选择弹窗的物种信息，否则会提示找不到该物种名），继续下一步：



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

Submission Portal

BioProject submission: SUB6457482
Raw sequence reads

Target

The most descriptive organism name for the study (to the species, if relevant).

★ Organism name: Homo sapien

Homo sapiens neanderthalensis
Homo sapiens

Isolate name

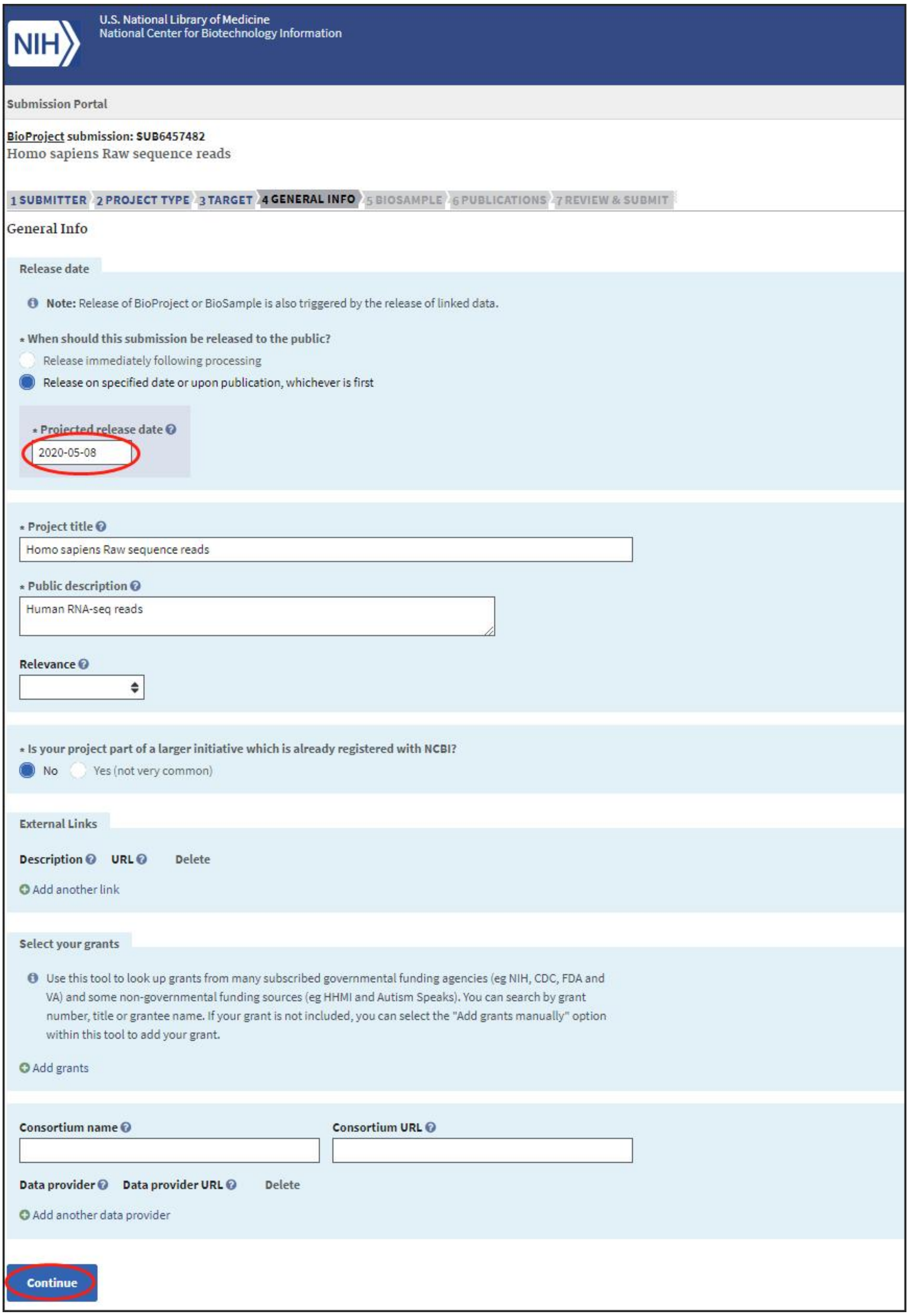
Label

Description

Note on human samples:
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 to NCBI's dbGaP database. dbGaP has controlled access mechanisms and is an appropriate resource for hosting sensitive patient data.

2.4 GENERAL INFO

进入 General Info 选项卡；第一个选项是问此数据是否马上发布，还是指定特定日期发布(释放日期可以给 NCBI 写邮件更改)。其他填写带星号的必填内容，继续下一步：



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

Submission Portal

BioProject submission: SUB6457482
Homo sapiens Raw sequence reads

General Info

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

* Projected release date: 2020-05-08

* Project title: Homo sapiens Raw sequence reads

* Public description: Human RNA-seq reads

Relevance:

* Is your project part of a larger initiative which is already registered with NCBI?
● No ○ Yes (not very common)

External Links

Description URL Delete

Add another link

Select your grants

Use this tool to look up grants from many subscribed governmental funding agencies (eg NIH, CDC, FDA and VA) and some non-governmental funding sources (eg HHMI and Autism Speaks). You can search by grant number, title or grantee name. If your grant is not included, you can select the "Add grants manually" option within this tool to add your grant.

Add grants

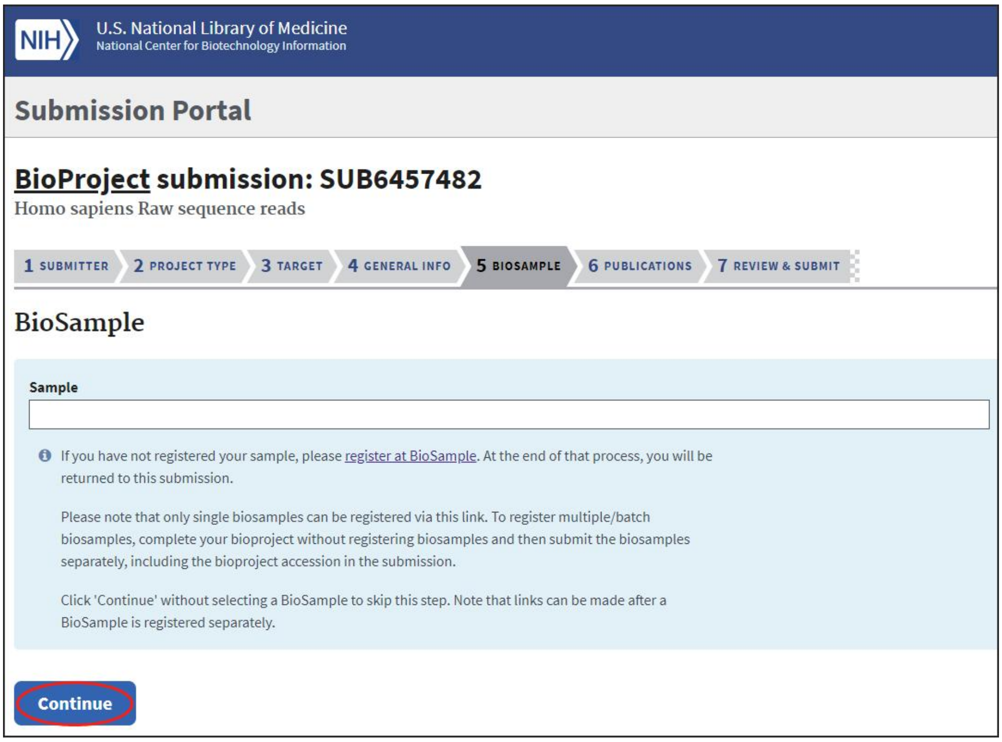
Consortium name:
Consortium URL:

Data provider Data provider URL Delete

Add another data provider

2.5 BIOSAMPLE

进入 BioSample 选项卡，需要指定项目中包含的 BioSample，可以先略过，也可以先去创建 BioSample。此处先略过，点击继续：



BioProject submission: SUB6457482

Homo sapiens Raw sequence reads

BioSample

Sample

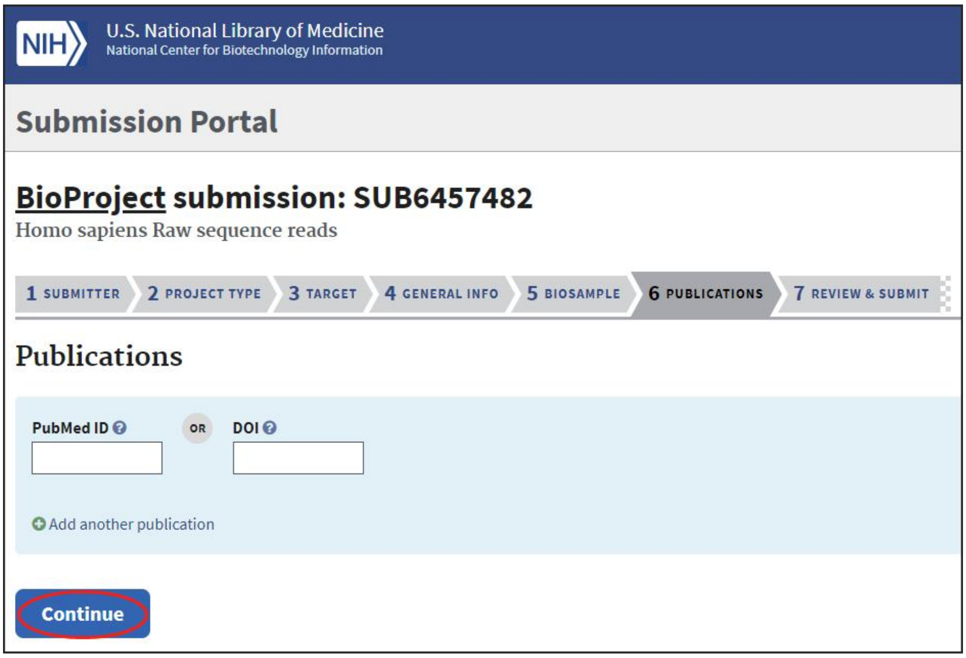
If you have not registered your sample, please register at BioSample. At the end of that process, you will be returned to this submission.

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.

2.6 PUBLICATIONS

进入 Publications 选项卡，如有已发表文章，可以指定 PubMedID 或 DOI，没有请略过：



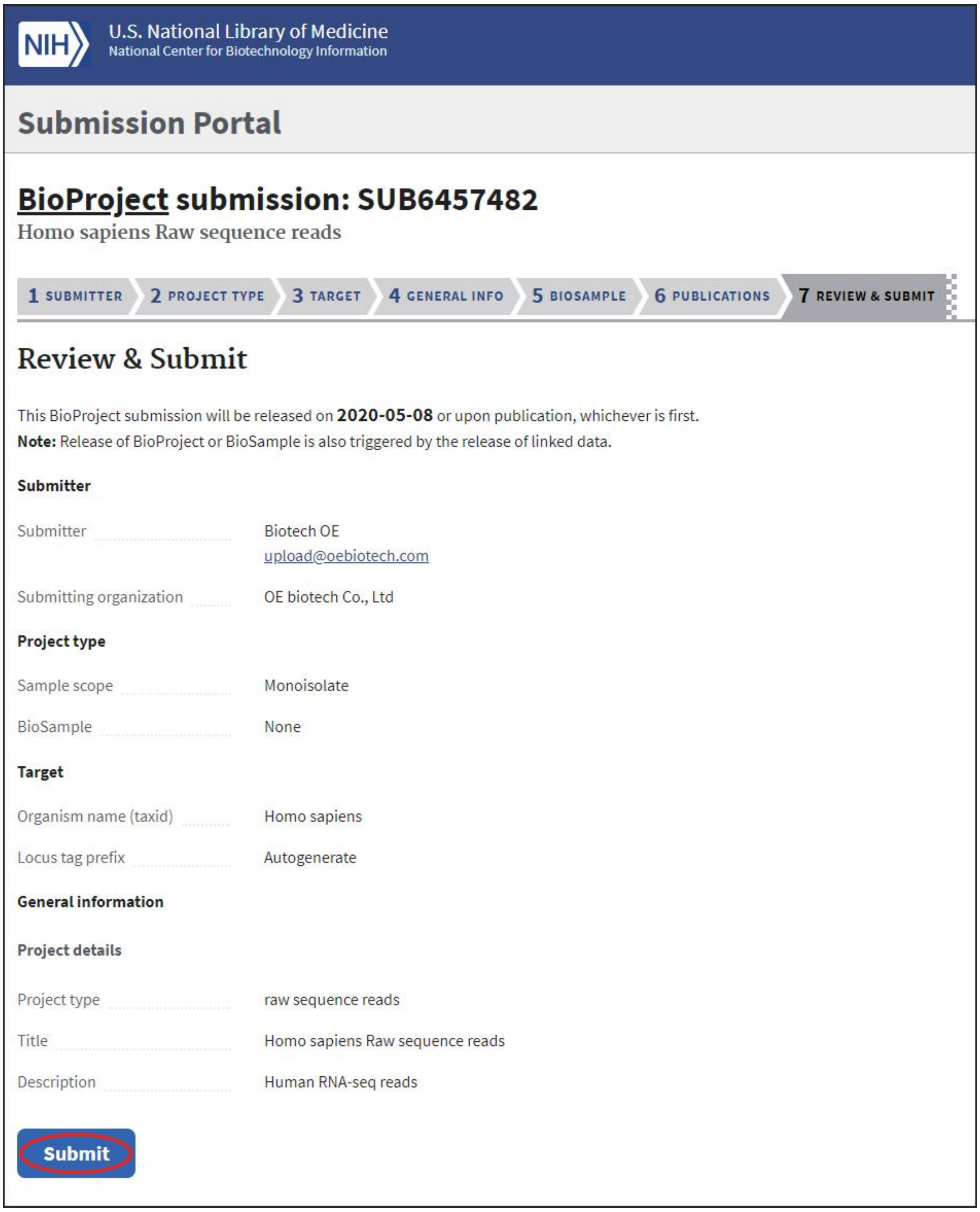
BioProject submission: SUB6457482

Homo sapiens Raw sequence reads

Publications

2.7 OVERVIEW

最后 Overview 中可以浏览概况，确认无误后点击 Submit 提交：



Submission Portal

BioProject submission: SUB6457482

Homo sapiens Raw sequence reads

Review & Submit

This BioProject submission will be released on **2020-05-08** or upon publication, whichever is first.
Note: Release of BioProject or BioSample is also triggered by the release of linked data.

Submitter

Submitter	Biotech OE upload@oebiotech.com
Submitting organization	OE biotech Co., Ltd

Project type

Sample scope	Monoisolate
BioSample	None

Target

Organism name (taxid)	Homo sapiens
Locus tag prefix	Autogenerate

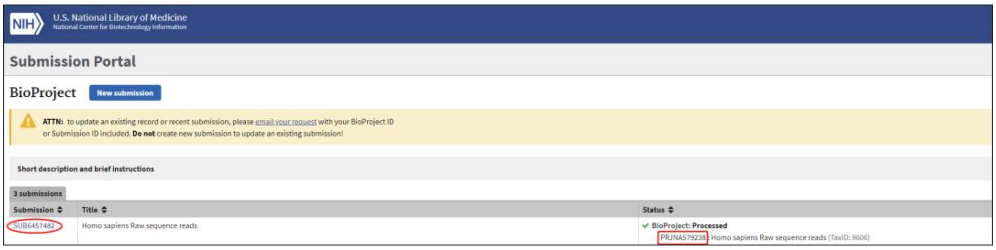
General information

Project details

Project type	raw sequence reads
Title	Homo sapiens Raw sequence reads
Description	Human RNA-seq reads

2.8 获得 PRJ ID

提交后会回到 BioProject 界面，列表中会显示刚才的项目已经提交正在审核(Awaiting processing)。一般几分钟就能审核完毕，刷新就可以看到 Processed，并可以看到 BioProject 编号 (PRJ 开头)：

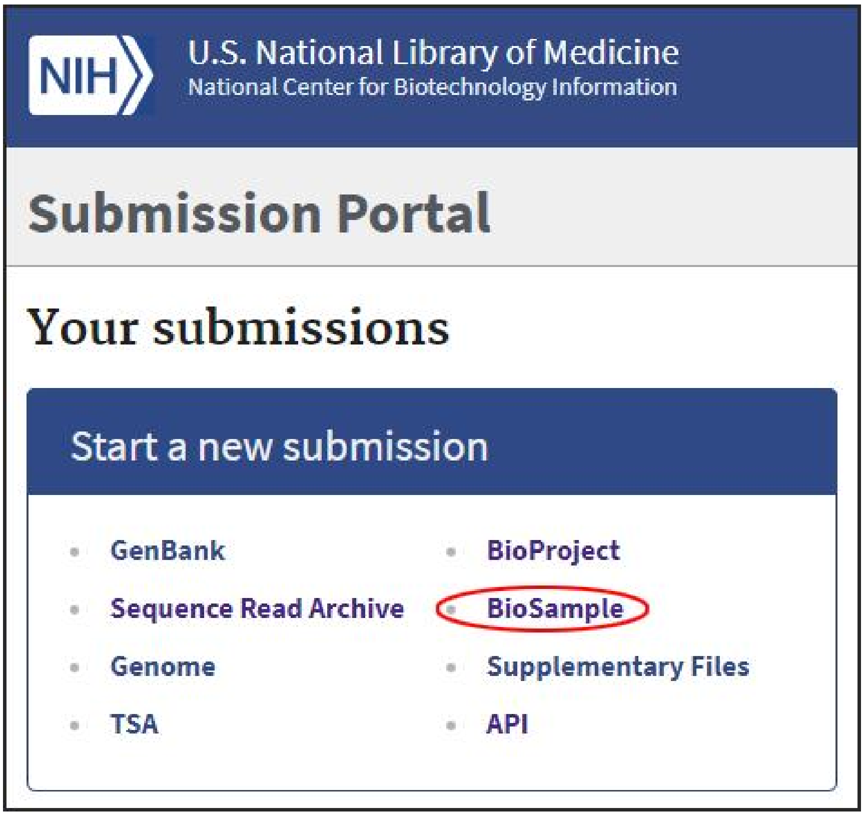


ATTN: to update an existing record or recent submission, please email your request with your BioProject ID or Submission ID included. **Do not** create new submission to update an existing submission!

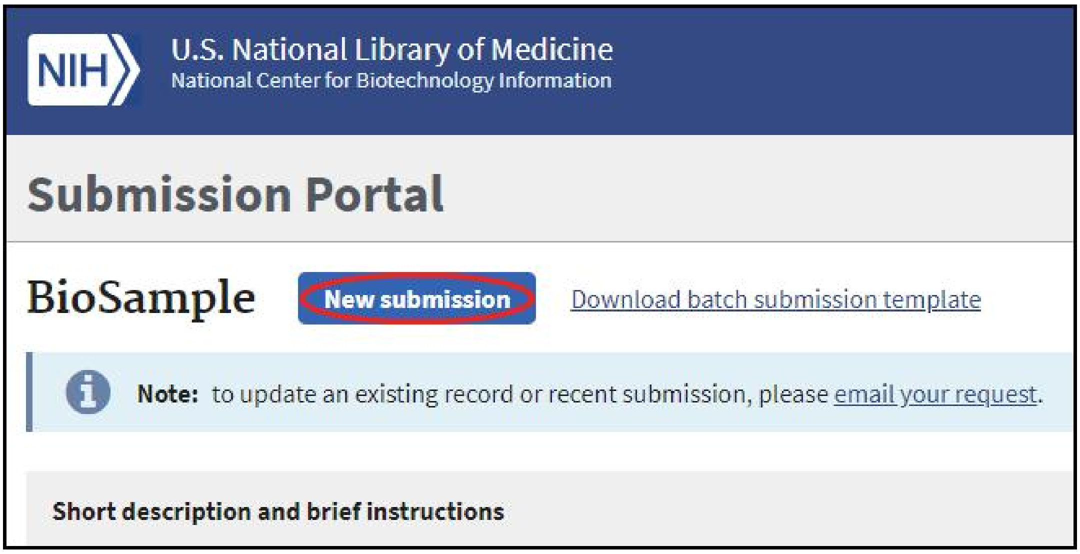
Submission	Title	Status
SUB6457482	Homo sapiens Raw sequence reads	✓ **BioProject: Processed** PRJNA579238: Homo sapiens Raw sequence reads (TaxID: 9606)

3 创建生物样品-BioSample

回到上传主页 https://submit.ncbi.nlm.nih.gov/subs/，点击 BioSample：



页面跳转后点击 New submission：



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

Submission Portal

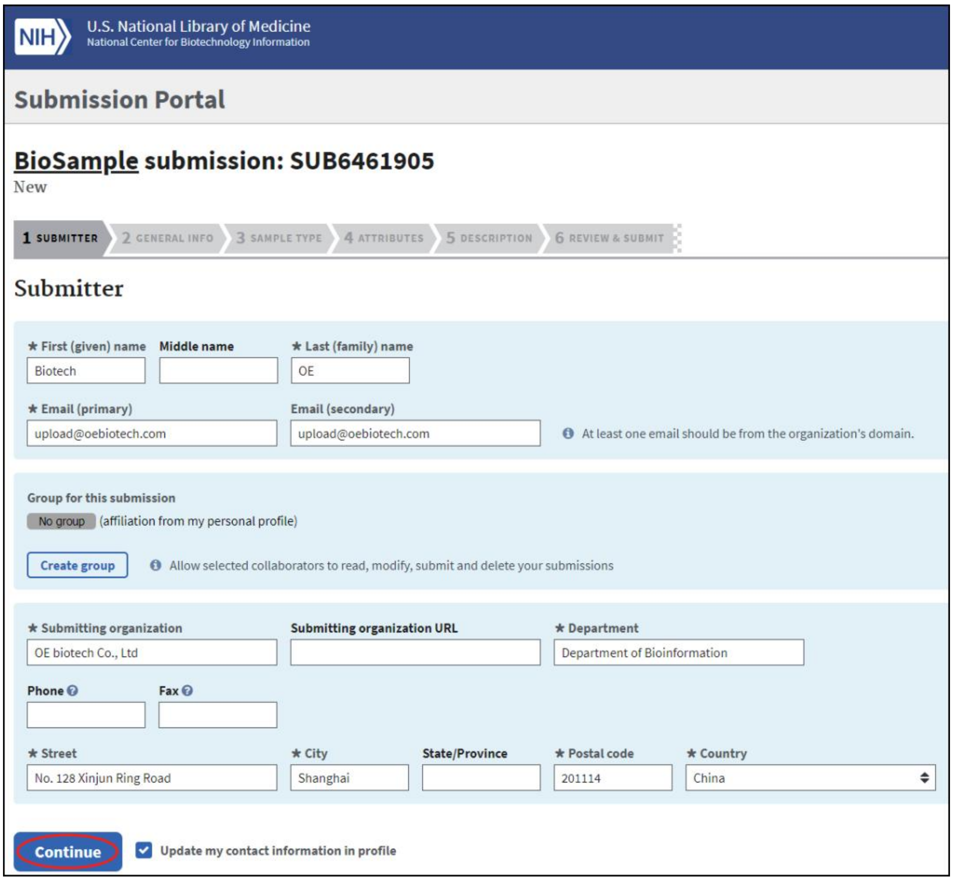
BioSample

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

Short description and brief instructions

3.1 SUBMITTER

进入创建 BioSample 页面 Submitter 选项卡，填写/完善必要的个人信息：



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

Submission Portal

BioSample submission: SUB6461905
New

Submitter

★ First (given) name: Biotech
Middle name:
★ Last (family) name: OE
★ Email (primary): upload@oebiotech.com
Email (secondary): upload@oebiotech.com
At least one email should be from the organization's domain.

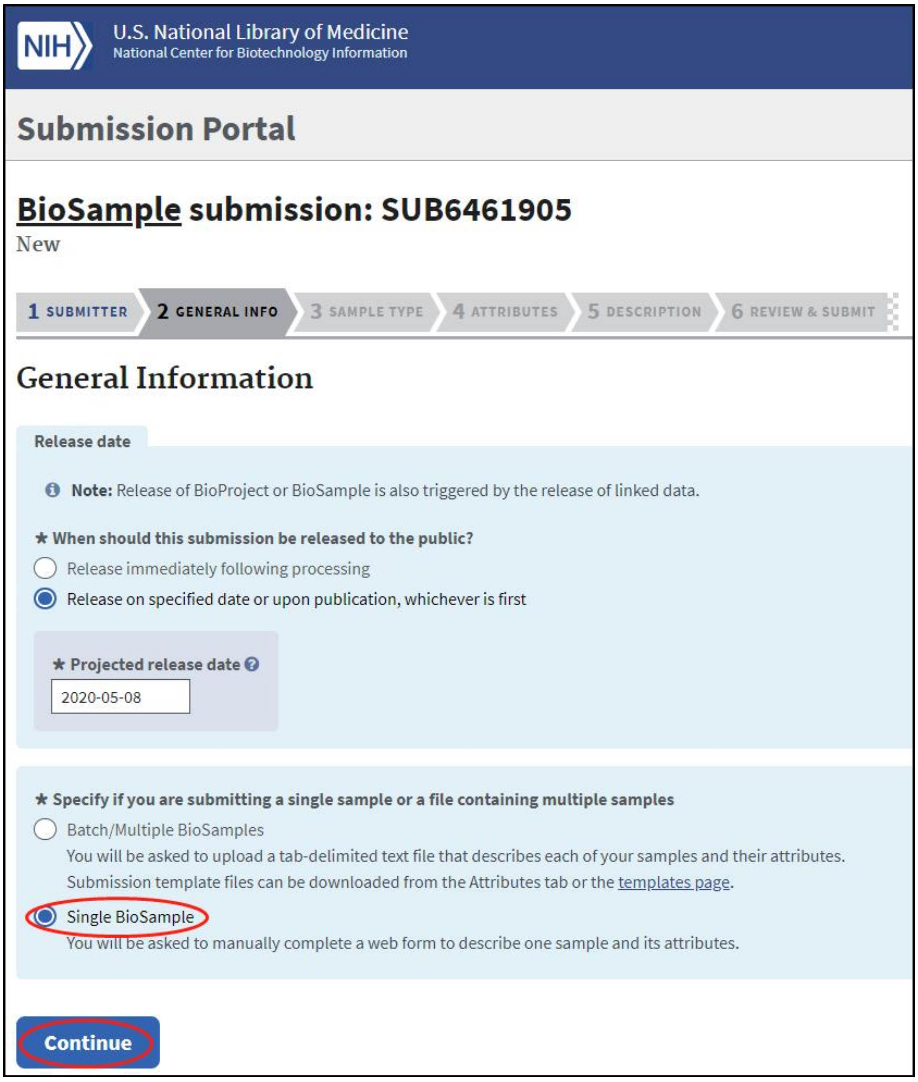
Group for this submission
No group (affiliation from my personal profile)
Allow selected collaborators to read, modify, submit and delete your submissions

★ Submitting organization: OE biotech Co., Ltd
Submitting organization URL:
★ Department: Department of Bioinformation
Phone:
Fax:
★ Street: No. 128 Xinjun Ring Road
★ City: Shanghai
State/Province:
★ Postal code: 201114
★ Country: China

☑ Update my contact information in profile

3.2 GENERAL INFO

在 General info 选项卡中，第一个是选择是否立即发布(和上面的 Bioproject 的类似)，第二个选择是问是否有多组样品，如果选多组样品的话，会要求上传表格文件，用于描述各样品的属性，比较繁琐；可选择单样品，不影响后续上传：



BioSample submission: SUB6461905
New

General Information

Release date

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

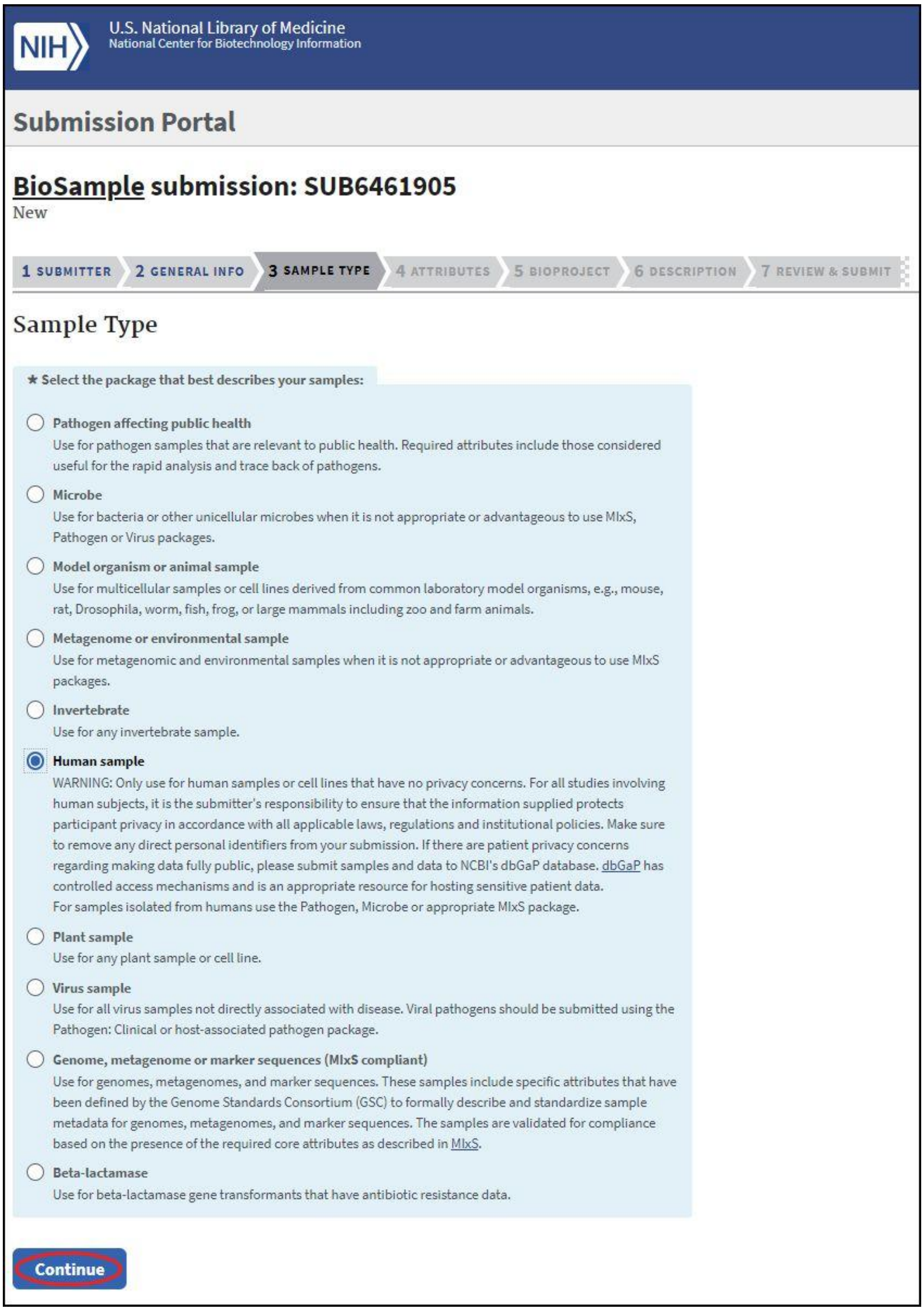
★ When should this submission be released to the public?
○ Release immediately following processing
● Release on specified date or upon publication, whichever is first

★ Projected release date
2020-05-08

★ Specify if you are submitting a single sample or a file containing multiple samples
○ Batch/Multiple BioSamples
You will be asked to upload a tab-delimited text file that describes each of your samples and their attributes. Submission template files can be downloaded from the Attributes tab or the templates page.
● Single BioSample
You will be asked to manually complete a web form to describe one sample and its attributes.

3.3 SAMPLE TYPE

在 Sample Type 选项卡中根据样本实际情况选择：



BioSample submission: SUB6461905

New

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.

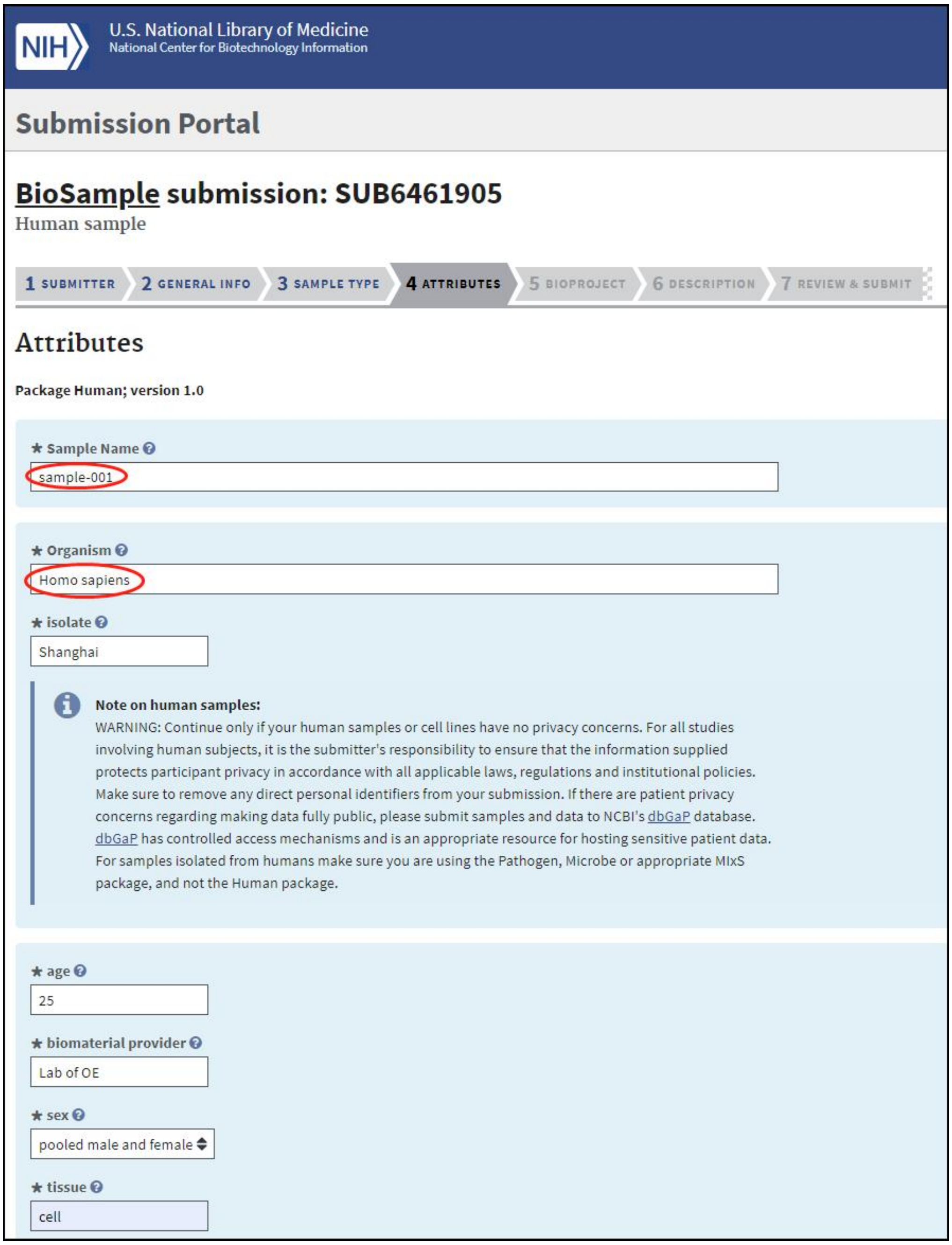
○ **Virus sample**
Use for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen: Clinical or host-associated pathogen package.

○ **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.

3.4 ATTRIBUTES

Attributes 选项卡，按情况填写星号内容(此处 Sample Name 可以自行填写，物种信息参考上面 Bioproject 的操作，填写完整的拉丁文名称-需要下拉列表有显示对应的物种名)：



Submission Portal

BioSample submission: SUB6461905

Human sample

Attributes

Package Human; version 1.0

★ Sample Name: sample-001

★ Organism: Homo sapiens

★ isolate: Shanghai

Note on human samples:
WARNING: Continue only if your human samples or cell lines 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 make sure you are using the Pathogen, Microbe or appropriate MIxS package, and not the Human package.

★ age: 25

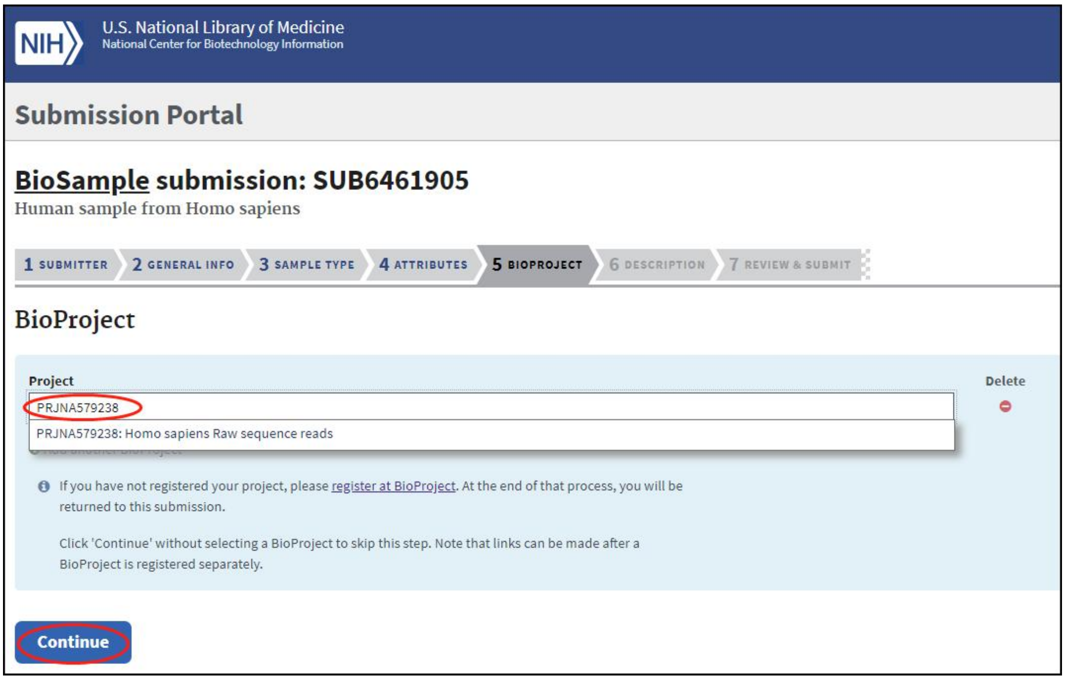
★ biomaterial provider: Lab of OE

★ sex: pooled male and female

★ tissue: cell

3.5 BIOPROJECT

BioProject 选项卡，输入刚才创建的 BioProject 编号：



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

Submission Portal

BioSample submission: SUB6461905

Human sample from Homo sapiens

BioProject

Project: PRJNA579238

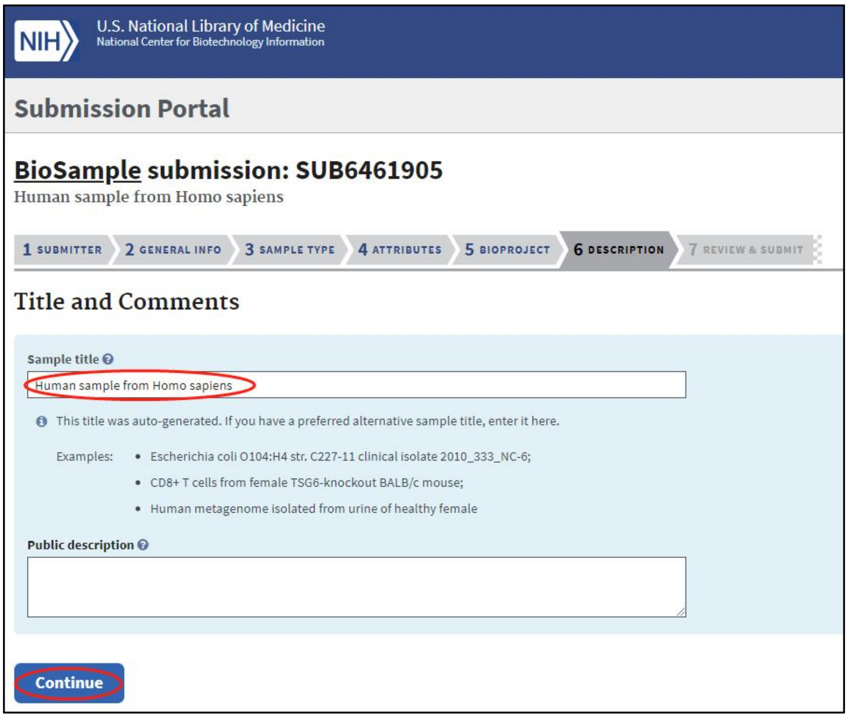
PRJNA579238: Homo sapiens Raw sequence reads

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

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

3.6 DESCRIPTION

Description 选项卡，信息如果不需要更改按照默认的即可：



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

Submission Portal

BioSample submission: SUB6461905

Human sample from Homo sapiens

Title and Comments

Sample title: Human sample from Homo sapiens

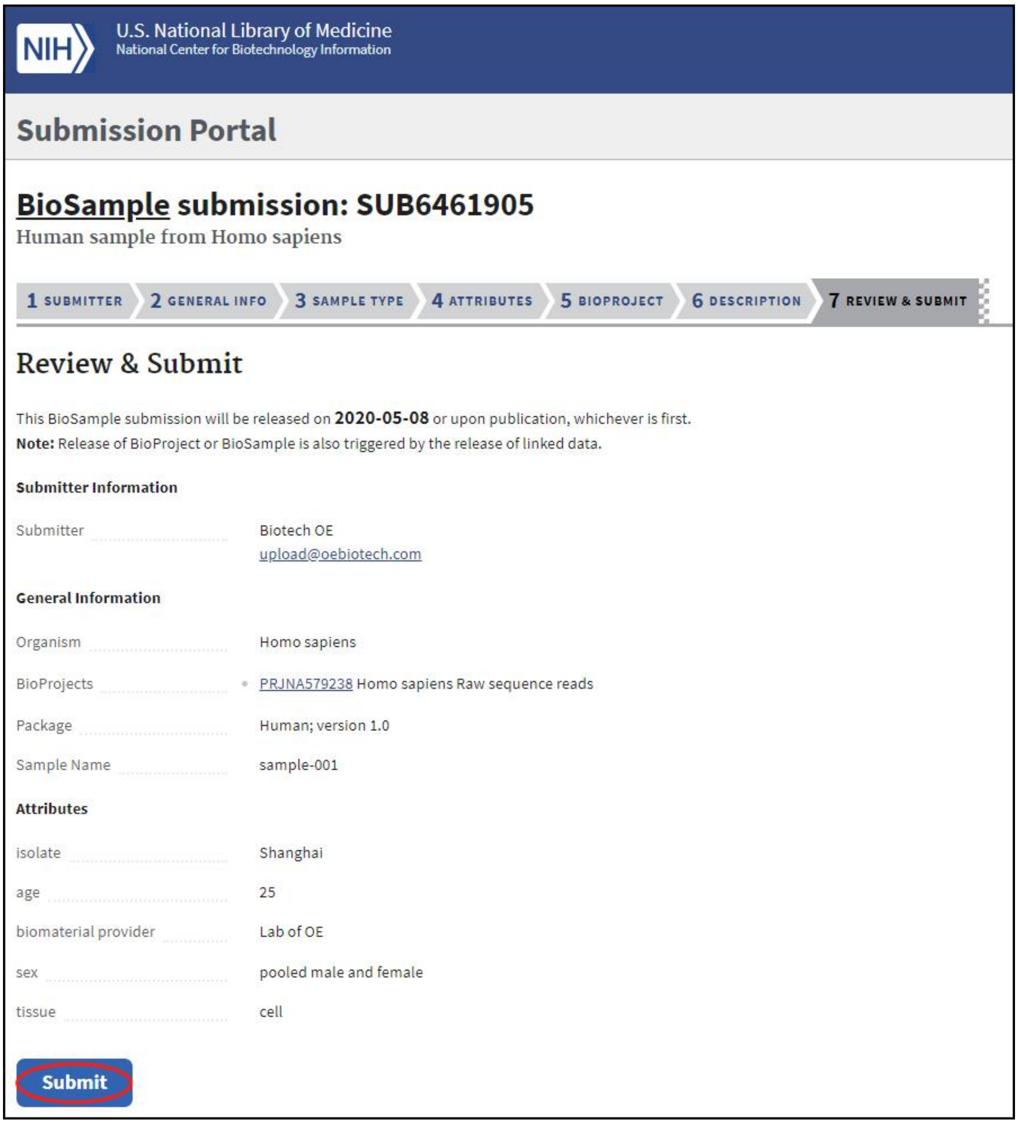
This title was auto-generated. If you have a preferred alternative sample title, enter it here.

Examples:
- Escherichia coli O104:H4 str. C227-11 clinical isolate 2010_333_NC-6;
- CD8+ T cells from female TSG6-knockout BALB/c mouse;
- Human metagenome isolated from urine of healthy female

Public description:

3.7 OVERVIEW

最后 Overview 中浏览概况，确认无误后点击 Submit 提交：



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

Submission Portal

BioSample submission: SUB6461905

Human sample from Homo sapiens

Review & Submit

This BioSample submission will be released on **2020-05-08** or upon publication, whichever is first.
Note: Release of BioProject or BioSample is also triggered by the release of linked data.

Submitter Information

Submitter	Biotech OE upload@oebiotech.com

General Information

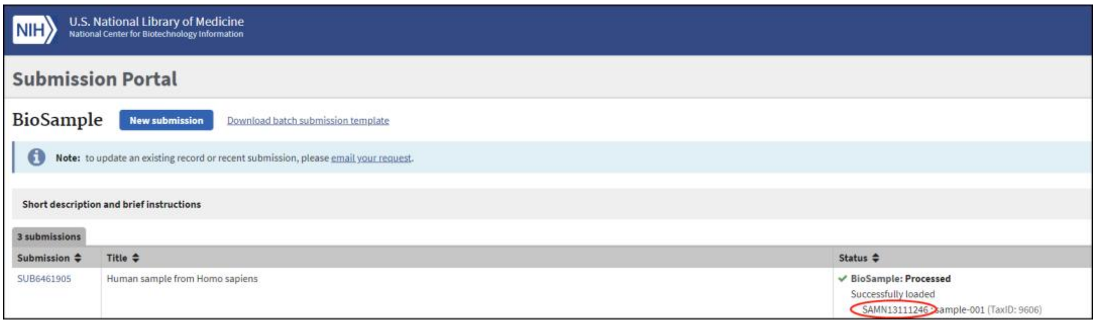
Organism	Homo sapiens
BioProjects	PRJNA579238 Homo sapiens Raw sequence reads
Package	Human; version 1.0
Sample Name	sample-001

Attributes

isolate	Shanghai
age	25
biomaterial provider	Lab of OE
sex	pooled male and female
tissue	cell

3.8 获得 SAM ID

提交后会回到 BioSample 界面，列表中会显示刚才的项目已经提交正在审核(Awaiting processing)。一般几分钟就能审核完毕，刷新就可以看到 Processed，并可以看到 BioSample 编号 (SAM 开头)：



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

Submission Portal

BioSample

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

Short description and brief instructions

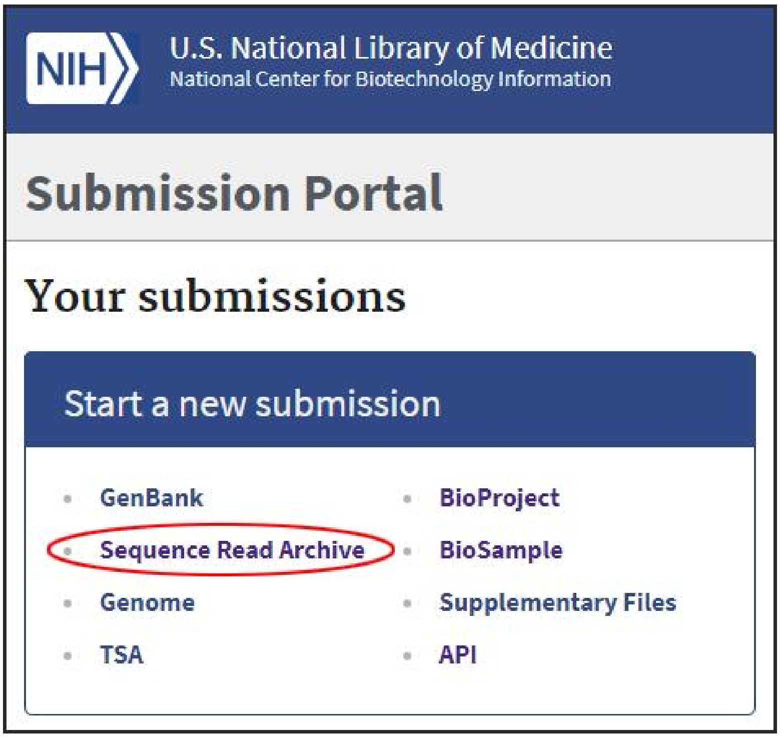
3 submissions

Submission	Title	Status
SUB6461905	Human sample from Homo sapiens	✓ BioSample: Processed Successfully loaded SAMN13111246 : sample-001 (TaxID: 9606)

4 创建 SRA

回到上传主页

https://submit.ncbi.nlm.nih.gov/subs/，点击 Sequence Read Archive：



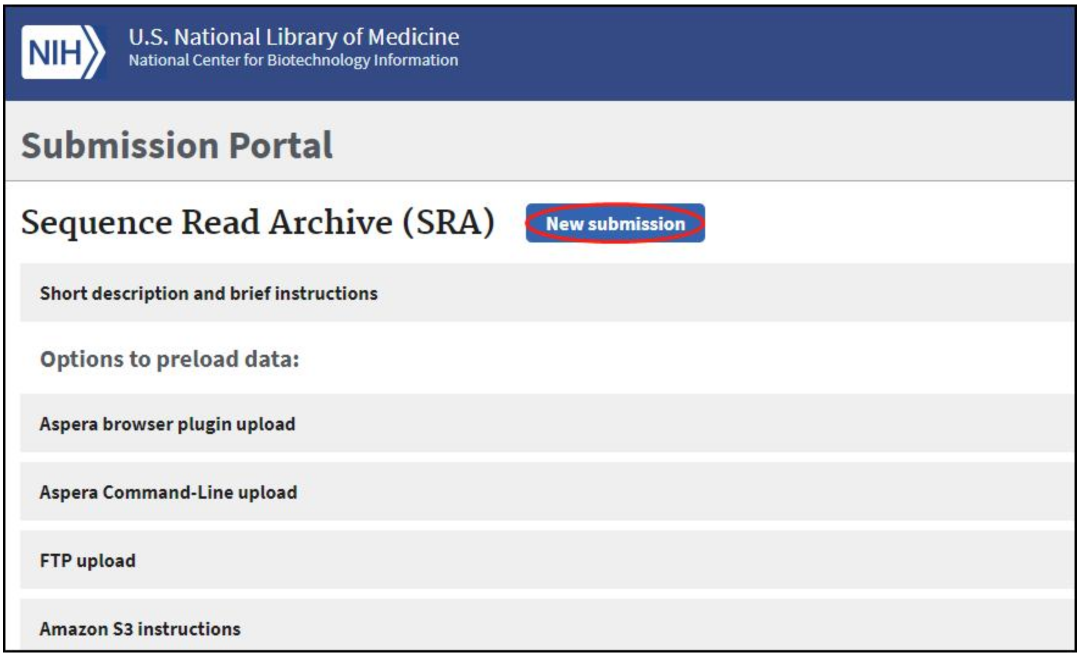
U.S. National Library of Medicine
National Center for Biotechnology Information

Submission Portal

Your submissions

Start a new submission

页面跳转后点击 New submission：



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

Submission Portal

Sequence Read Archive (SRA)

Short description and brief instructions

Options to preload data:

Aspera browser plugin upload

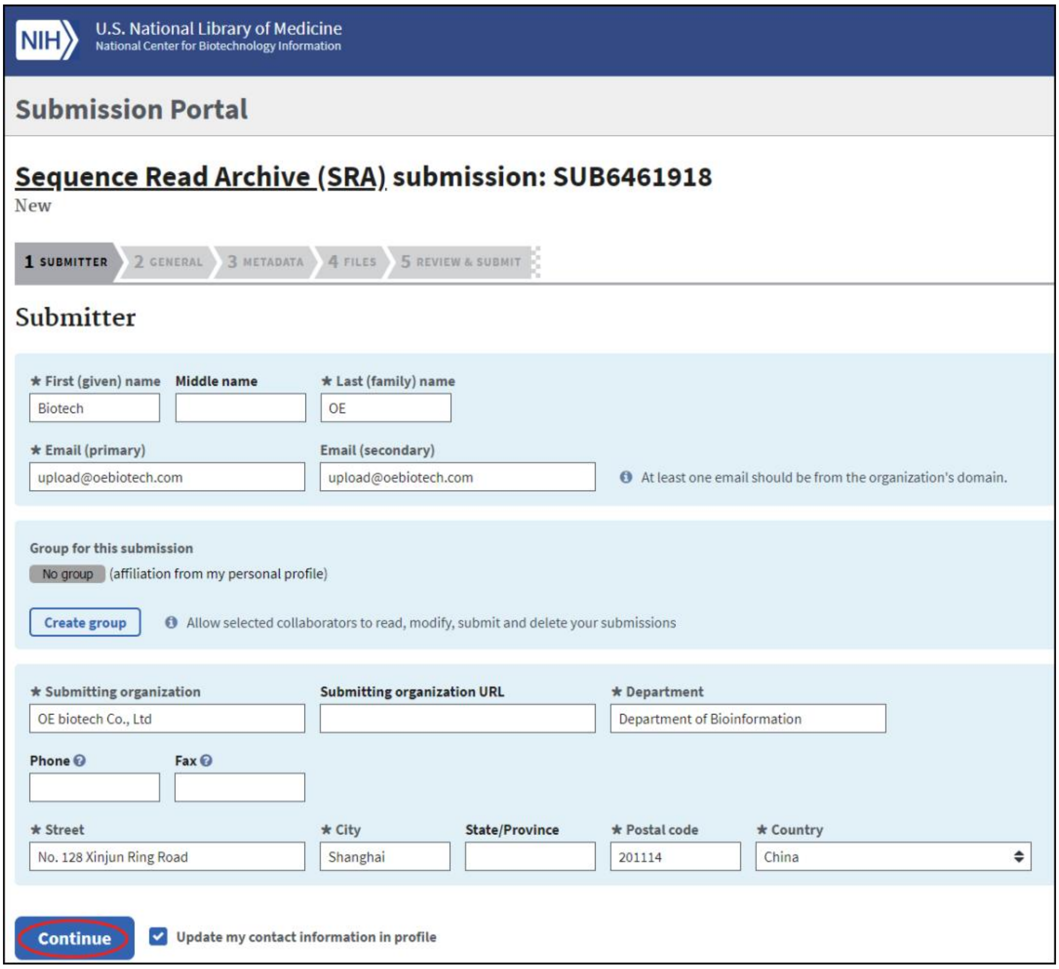
Aspera Command-Line upload

FTP upload

Amazon S3 instructions

4.1 SUBMITTER

进入创建 SRA 页面 Submitter 选项卡，填写/完善必要个人信息：



Sequence Read Archive (SRA) submission: SUB6461918

New

Submitter

★ First (given) name: Biotech

Middle name:

★ Last (family) name: OE

★ Email (primary): upload@oebiotech.com

Email (secondary): upload@oebiotech.com

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

Group for this submission

No group (affiliation from my personal profile)

Allow selected collaborators to read, modify, submit and delete your submissions

★ Submitting organization: OE biotech Co., Ltd

Submitting organization URL:

★ Department: Department of Bioinformation

Phone:

Fax:

★ Street: No. 128 Xinjun Ring Road

★ City: Shanghai

State/Province:

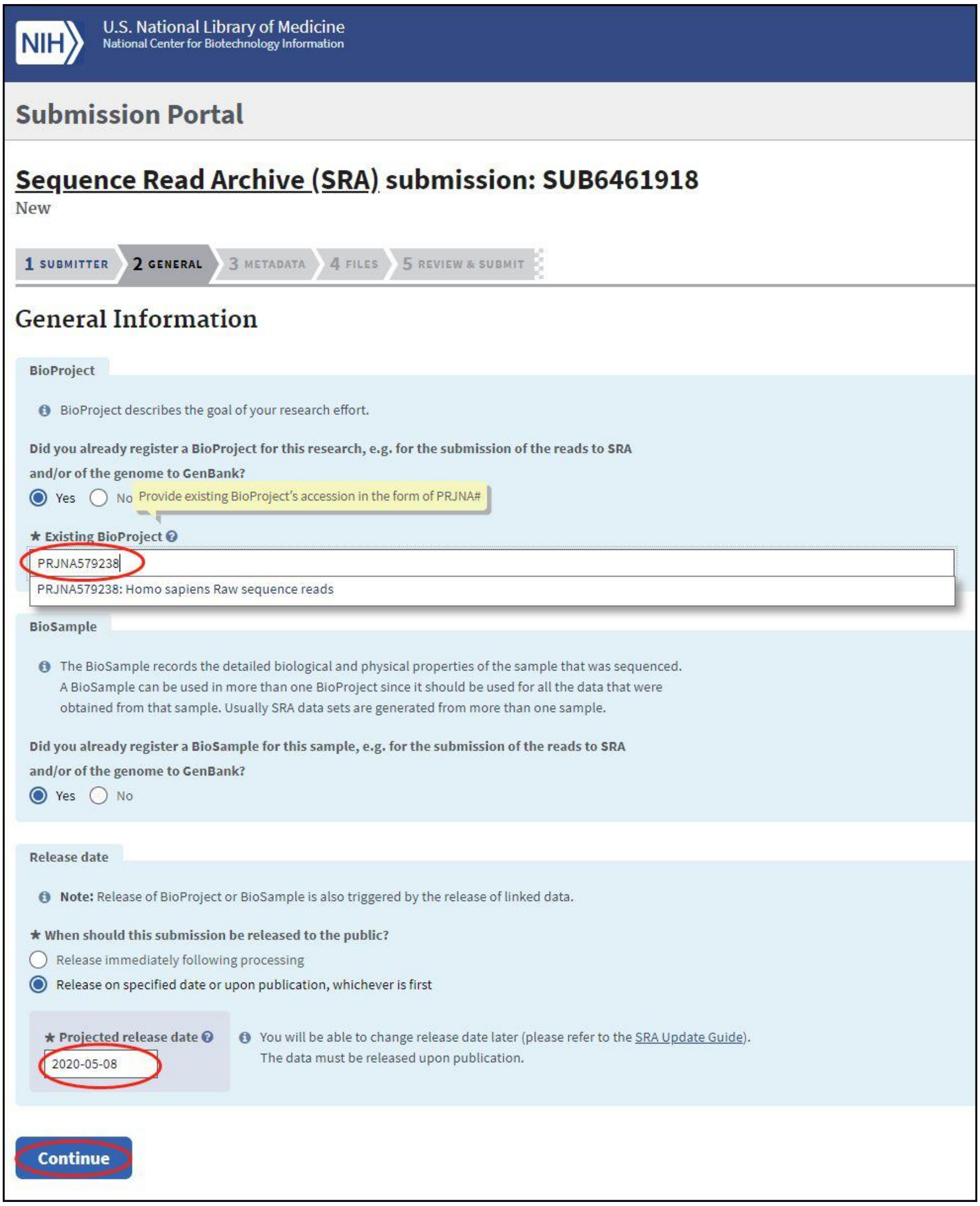
★ Postal code: 201114

★ Country: China

☑ Update my contact information in profile

4.2 GENERAL

在 General 选项卡里，选择前面创建的 BioProject，是否已经创建好了 BioSample 选择 Yes，发布日期根据实际情况选择：



Submission Portal

Sequence Read Archive (SRA) submission: SUB6461918

New

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 Provide existing BioProject's accession in the form of PRJNA#

★ **Existing BioProject**

PRJNA579238

PRJNA579238: Homo sapiens Raw sequence reads

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

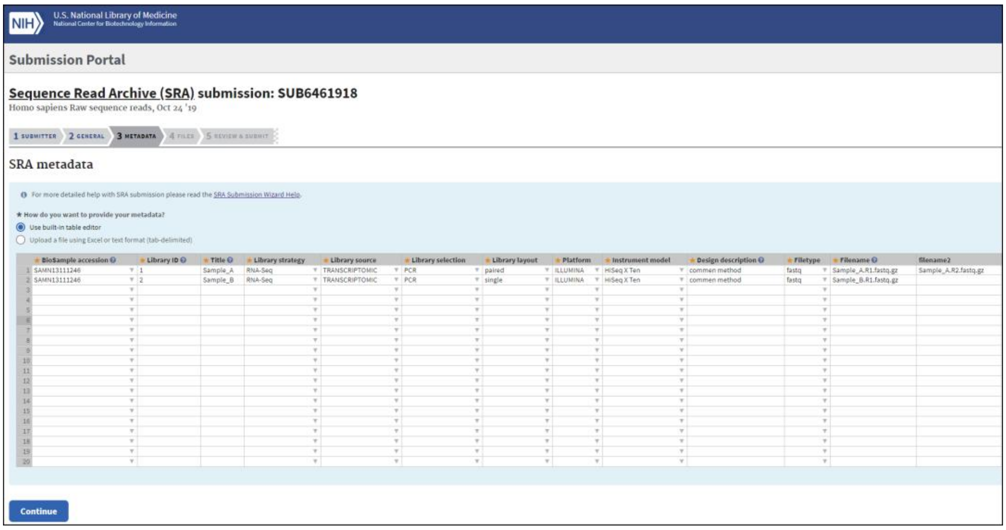
★ **Projected release date**

2020-05-08

You will be able to change release date later (please refer to the SRA Update Guide). The data must be released upon publication.

4.3 METADATA

在 Metadata 选项卡，填写提交数据的相关实验信息，这一步可以直接在网页上填写，示例如下：



Submission Portal

Sequence Read Archive (SRA) submission: SUB6461918

Homo sapiens Raw sequence reads, Oct 24 '19

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)

	★ BioSample accession	★ Library ID	★ Title	★ Library strategy	★ Library source	★ Library selection	★ Library layout	★ Platform	★ Instrument model	★ Design description	★ Filetype	★ Filename	filename2
1	SAMN13111246	1	Sample_A	RNA-Seq	TRANSCRIPTOMIC	PCR	paired	ILLUMINA	HiSeq X Ten	commen method	fastq	Sample_A.R1.fastq.gz	Sample_A.R2.fastq.gz
2	SAMN13111246	2	Sample_B	RNA-Seq	TRANSCRIPTOMIC	PCR	single	ILLUMINA	HiSeq X Ten	commen method	fastq	Sample_B.R1.fastq.gz	
3													
4													
5													
6													
7													
8													
9													
10													
11													
12													
13													
14													
15													
16													
17													
18													
19													
20													

★ BioSample accession	★ Library ID	★ Title	★ Library strategy	★ Library source	★ Library selection	★ Library layout	★ Platform
SAMN13111246	1	Sample_A	RNA-Seq	TRANSCRIPTOMIC	PCR	paired	ILLUMINA
SAMN13111246	2	Sample_B	RNA-Seq	TRANSCRIPTOMIC	PCR	single	ILLUMINA

★ Instrument model	★ Design description	★ Filetype	Reference FASTA File	Reference assembly	★ Filename	filename2
HiSeq X Ten	commen method	fastq			Sample_A.R1.fastq.gz	Sample_A.R2.fastq.gz
HiSeq X Ten	commen method	fastq			Sample_B.R1.fastq.gz	

也可以下载模板表格填好之后再上传：

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)

★ Metadata file

选择文件 未选择任何文件

Use one of these ways to fill out SRA metadata:
Download Excel spreadsheet designed to make it easier to select the correct metadata values), edit, save and then upload the modified Excel file.

表头	解释
BioSample accession	填写前创建的BioSample的ID，SAM开头的
Library ID	样本对应文库ID，根据实际情况填写，可自行命名，每行不重复即可
Title	描述样本信息，可以填写样本名称，每行不重复即可
Library strategy	建库方式，根据实际情况填写，这里示例填写"RNA_Seq"
Library source	建库来源，根据实际情况填写，这里示例填写"TRANSCRIPTOMIC"
Library selection	建库选择，根据实际情况填写，这里示例填写"PCR"
Library layout	测序方式，根据实际情况填"paired"或"single"
Platform	测序平台，根据实际情况填写，这里示例填写"ILLUMINA"
Instrument model	仪器型号，根据实际情况填写，这里示例填写"HiSeq X Ten"
Design description	方法描述，根据实际情况填写，这里示例填写"common method"
Filetype*	数据格式，根据情况填写(可接受fastq、fasta、bam等格式)，示例填写"fastq"
Filename*	上传数据文件名称(R1端)，根据实际情况填写
Filename2	上传数据文件名称(R2端)，如果是上传单端测序数据，这里不需要填写

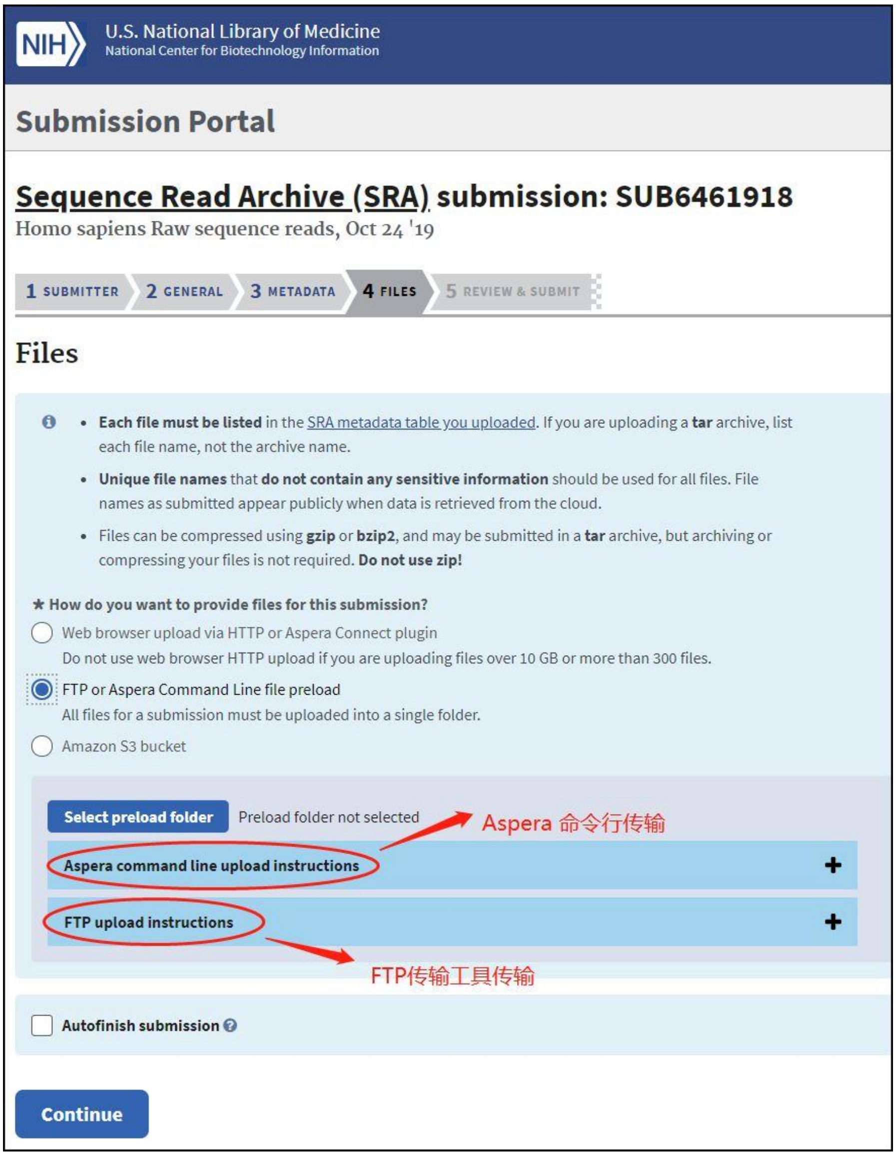
*Filetype 如果只有 fa 文件也可以选择 fastq 格式(NCBI 会默认上传的 fa 文件碱基质量值为 30)

*Filename 填写的文件名称必须和上传的序列文件名称一致(后台通过表格信息关联上传的数据，如果名称不一致会提示文件缺失)，如“Filename”、“Flename2”分别填写了 Sample_A.R1.fastq.gz、Sample_A.R2.fastq.gz，那么

该样本的 R1、R2 端数据的文件名也必须是 Sample_A.R1.fastq.gz、Sample_A.R2.fastq.gz。

4.4 FILES

在 Files 选项卡里，对于上传数据的方式，我们通常选择 FTP or Aspera Command Line file preload：即使用 FTP 工具(Filezilla)或命令行传输(ascp)：



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

Submission Portal

Sequence Read Archive (SRA) submission: SUB6461918
Homo sapiens Raw sequence reads, Oct 24 '19

1 SUBMITTER 2 GENERAL 3 METADATA 4 FILES 5 REVIEW & SUBMIT

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

Select preload folder   Preload folder not selected

Aspera command line upload instructions +

FTP upload instructions +

☐ Autofinish submission

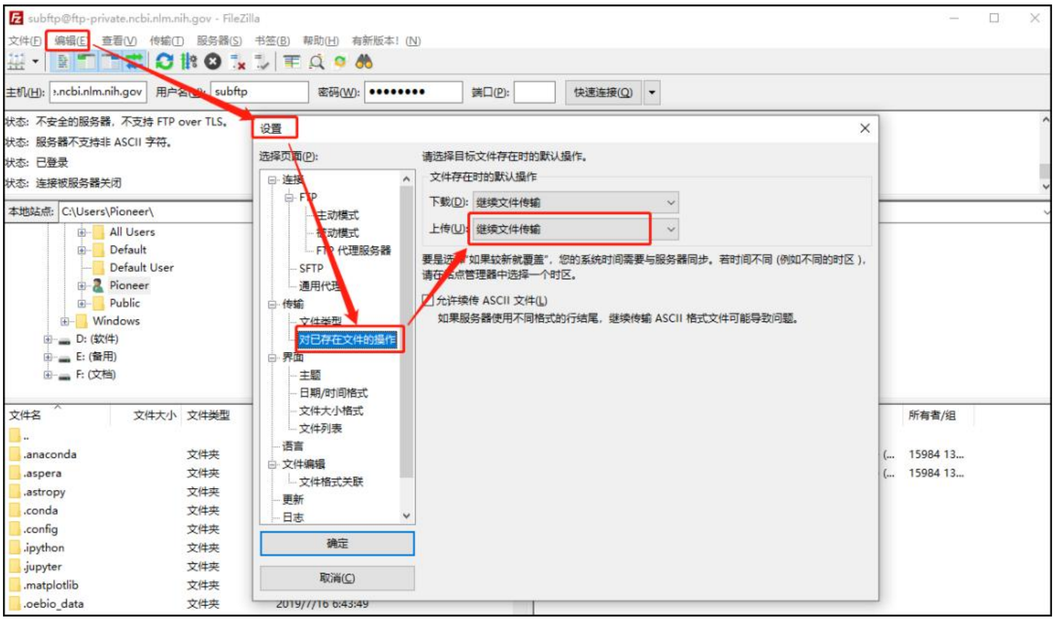
Continue

Aspera 命令行传输

FTP传输工具传输

4.4.1 FTP 传输-Filezilla

下载 Filezilla(https://www.filezilla.cn/download/) 安装好之后并做如下设置：



主机(H): ›.ncbi.nlm.nih.gov 用户名(U): subftp 密码(W): ••••••••• 端口(P):

状态: 不安全的服务器，不支持 FTP over TLS。
状态: 服务器不支持非 ASCII 字符。
状态: 已登录
状态: 连接被服务器关闭

本地站点: C:\Users\Pioneer\

设置

请选择目标文件存在时的默认操作。

文件存在时的默认操作

下载(D): 继续文件传输

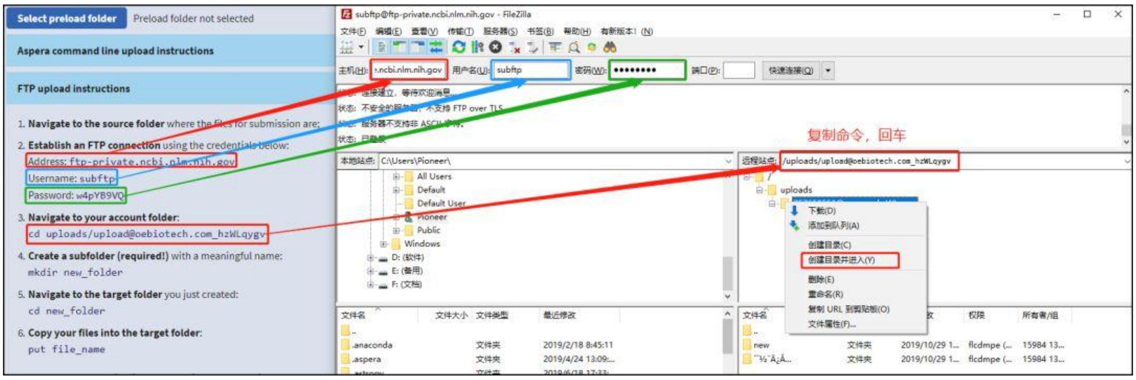
上传(U): 继续文件传输

要是选择"如果较新就覆盖"，您的系统时间需要与服务器同步。若时间不同 (例如不同的时区)，请在站点管理器中选择一个时区。

☐ 允许续传 ASCII 文件(L)

如果服务器使用不同格式的行结尾，继续传输 ASCII 格式文件可能导致问题。

根据网页提供的 FTP 地址、账号、密码用 Filezilla 连接上服务器。如果一直读取目录失败，可以直接将 3. Navigate to your account folder 下面的目录(如“uploads/upload@oebiotech.com_hzWLqygv”)复制进 Filezilla 的远程站点里(注意路径要以斜杠“/”开头)，然后回车键就能进入目录了，进入目录后再创建一个子目录并进入(如果上传的数据不在新创建的子目录中上传的文件不会被检测到！)：



1. **Navigate to the source folder** where the files for submission are:
2. **Establish an FTP connection** using the credentials below:
   Address: `ftp-private.ncbi.nlm.nih.gov`
   Username: `subftp`
   Password: `w4pYB9VQ`
3. **Navigate to your account folder:**
   `cd uploads/upload@oebiotech.com_hzWLqygv`
4. **Create a subfolder (required!)** with a meaningful name:
   `mkdir new_folder`
5. **Navigate to the target folder** you just created:
   `cd new_folder`
6. **Copy your files into the target folder:**
   `put file_name`

主机(H): ›.ncbi.nlm.nih.gov 用户名(U): subftp 密码(W): ••••••••

状态: 连接建立，等待欢迎消息...
状态: 不安全的服务器，不支持 FTP over TLS。
状态: 服务器不支持非 ASCII 字符。
状态: 已登录

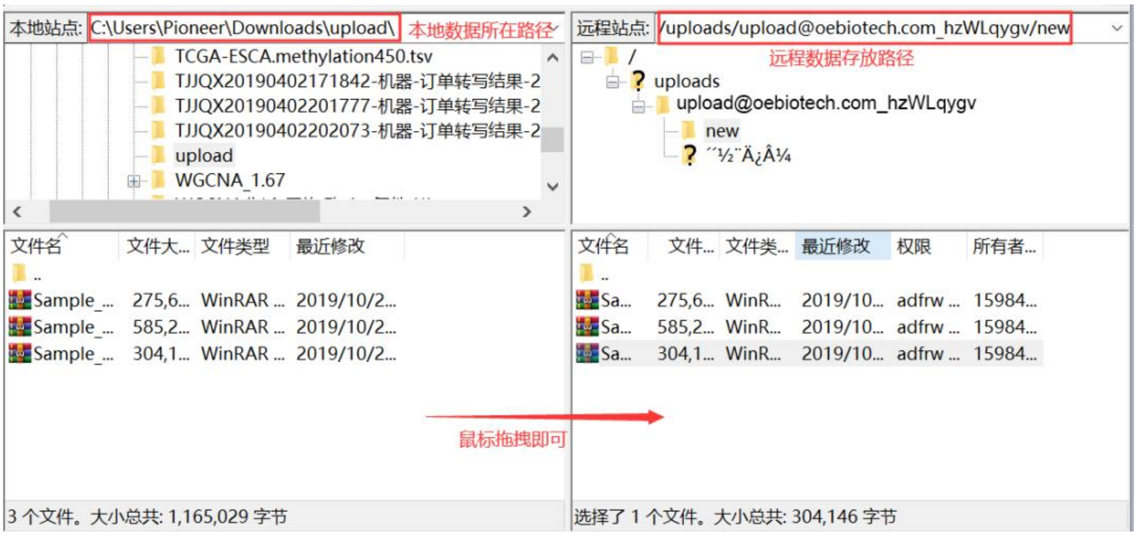
本地站点: C:\Users\Pioneer\

远程站点: /uploads/upload@oebiotech.com_hzWLqygv

文件名	文件类型	最近修改
.anaconda	文件夹	2019/2/18 8:45:11
.aspera	文件夹	2019/4/24 13:09:...

文件名	文件类型	最近修改	权限	所有者/组
new	文件夹	2019/10/29 1...	flcdmpe (...	15984 13...
"½"Ã¿Å...	文件夹	2019/10/29 1...	flcdmpe (...	15984 13...

将数据文件直接拖进右侧区域(新创建的子目录)等待上传完毕：



本地站点: C:\Users\Pioneer\Downloads\upload\ 本地数据所在路径

TCGA-ESCA.methylation450.tsv
TJJQX20190402171842-机器-订单转写结果-2
TJJQX20190402201777-机器-订单转写结果-2
TJJQX20190402202073-机器-订单转写结果-2
upload
WGCNA_1.67

文件名	文件大...	文件类型	最近修改
..			
Sample_...	275,6...	WinRAR ...	2019/10/2...
Sample_...	585,2...	WinRAR ...	2019/10/2...
Sample_...	304,1...	WinRAR ...	2019/10/2...

3 个文件。大小总共: 1,165,029 字节

远程站点: /uploads/upload@oebiotech.com_hzWLqygv/new

远程数据存放路径

/
uploads
upload@oebiotech.com_hzWLqygv
new
´´½¨Ä¿Â¼

文件名	文件...	文件类...	最近修改	权限	所有者...
..					
Sa...	275,6...	WinR...	2019/10...	adfrw ...	15984...
Sa...	585,2...	WinR...	2019/10...	adfrw ...	15984...
Sa...	304,1...	WinR...	2019/10...	adfrw ...	15984...

选择了 1 个文件。大小总共: 304,146 字节

鼠标拖拽即可

4.4.2 命令行传输-Aspera 上传/ascp

往 SRA 上传输数据的时候，有时会遇到 Filezilla 传输慢的情况，这时候就需要 Aspera 软件帮忙了。

软件官网下载链接：

https://downloads.asperasoft.com/en/downloads/62

网盘下载链接：

Windows：

https://pan.baidu.com/s/1OKgPKs2nEUH1ubuxAhXH8A 提取码: yw24

Mac OS：

https://pan.baidu.com/s/18oQ36MdNuMTm6OGUVJQJuw 提取码: gy55

下载好解压缩即可，解压缩的路径一会要用到，例如 D:\Aspera\cli\bin

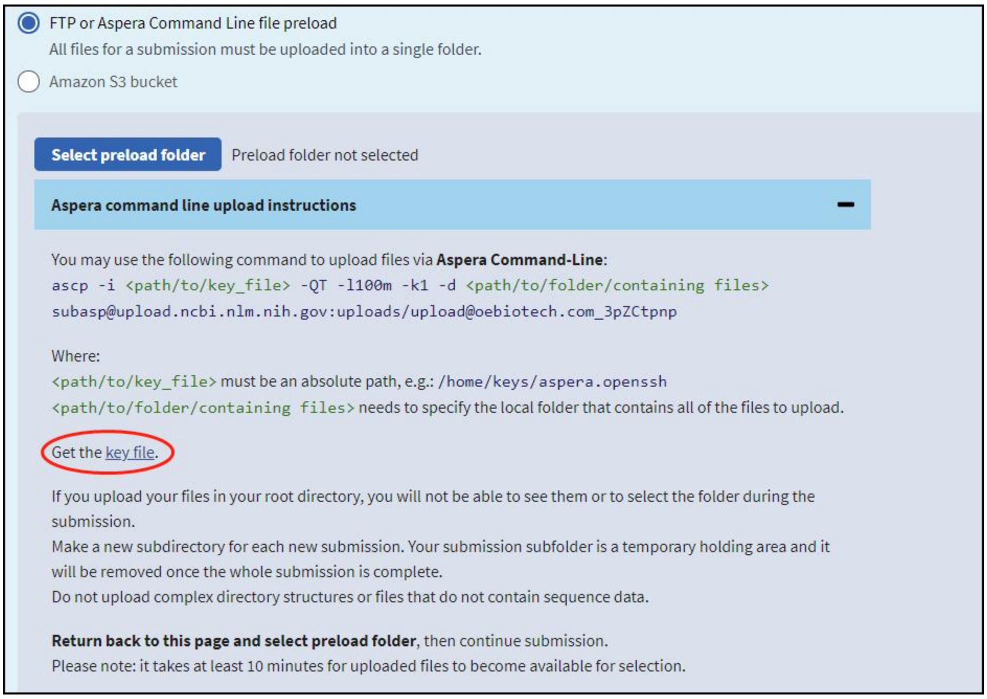
Aspera 环境变量设置(以 windows 为例)：右键我的电脑-属性-高级设置-环境变量设置-PATH 里添加软件的路径(D:\Aspera\cli\bin)

windows 键+R 弹出运行框，输入 cmd 回车进行进入 cmd 界面，输入 ascp –h 可正常显示 ：

```
C:\WINDOWS\system32\cmd.exe
Microsoft Windows [版本 10.0.17763.864]
(c) 2018 Microsoft Corporation。保留所有权利。

C:\Users\Pioneer>ascp -h
Usage: ascp.exe [OPTION] SRC... DEST
          SRC to DEST, or multiple SRC to DEST dir
          SRC, DEST format: [[user@]host:]PATH
  -h, --help                    Display usage
  -A, --version                 Display version.
  -T                            Disable encryption
  -d                            Create target directory, implied for file/file-pair lists
  -p                            Preserve file timestamp
  -q                            Disable progress display
  -v                            Verbose mode
  -6                            Use IPv6
  -D...                         Debug level
  -l MAX-RATE                   Max transfer rate
  -m MIN-RATE                   Min transfer rate
```

在 Aspera command line upload instructions， 可以查看到命令行：



● FTP or Aspera Command Line file preload

All files for a submission must be uploaded into a single folder.

○ Amazon S3 bucket

You may use the following command to upload files via **Aspera Command-Line**:

```
ascp -i <path/to/key_file> -QT -l100m -k1 -d <path/to/folder/containing files>
subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp
```

Where:

`<path/to/key_file>` must be an absolute path, e.g.: `/home/keys/aspera.openssh`

`<path/to/folder/containing files>` needs to specify the local folder that contains all of the files to upload.

Get the key file.

If you upload your files in your root directory, you will not be able to see them or to select the folder during the submission.

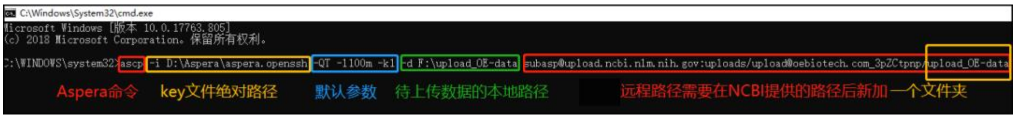
Make a new subdirectory for each new submission. Your submission subfolder is a temporary holding area and it will be removed once the whole submission is complete.

Do not upload complex directory structures or files that do not contain sequence data.

Return back to this page and select preload folder, then continue submission.

Please note: it takes at least 10 minutes for uploaded files to become available for selection.

注意命令行中提及的 key_file(上图红圈部分)点击下载 key 文件 aspera.openssh 并存放在指定位置，如 "D:\Aspera"；数据上传完整命令如下：



```
C:\Windows\System32\cmd.exe
Microsoft Windows [版本 10.0.17763.805]
(c) 2018 Microsoft Corporation。保留所有权利。

C:\WINDOWS\system32>ascp -i D:\Aspera\aspera.openssh -QT -l100m -k1 -d F:\upload_OE-data subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp/upload_OE-data
```

ascp -i D:\Aspera\aspera.openssh -QT -l100m -k1 -d F:\upload_OE-data subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp/upload_OE-data

-i 之后填写前面下载的 aspera.openssh 文件的绝对路径 ("D:\Aspera\aspera.openssh")

-d 之后填写需要上传的数据文件的路径，这个路径下除了待上传的原始数据最好不要存放其他文件("F:\upload_OE-data")

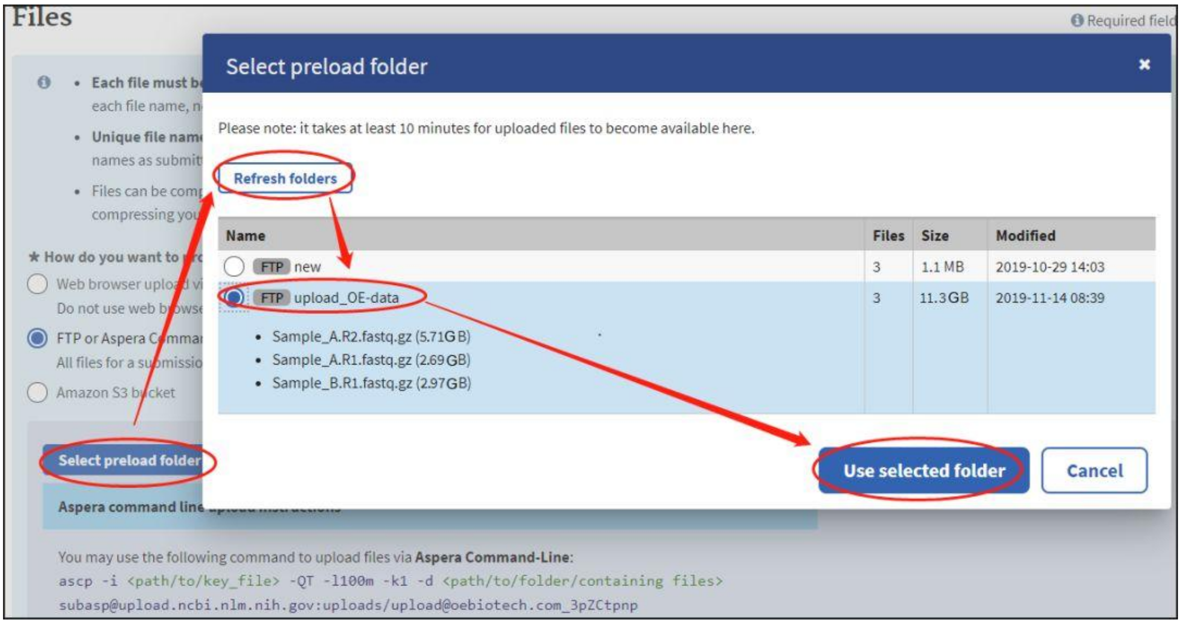
"空格"之后接 NCBI 的远程路径

("subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp/upload_OE-data")

远程路径后面需要添加一个子目录("upload_OE-data")，否则上传的数据检测不到!

命令输入后回车显示传输，等待提示传输完毕，传输中断可重复之前的命令。

数据上传成功之后(这一过程消耗的时间根据文件数量、大小及网速决定)做如下操作：



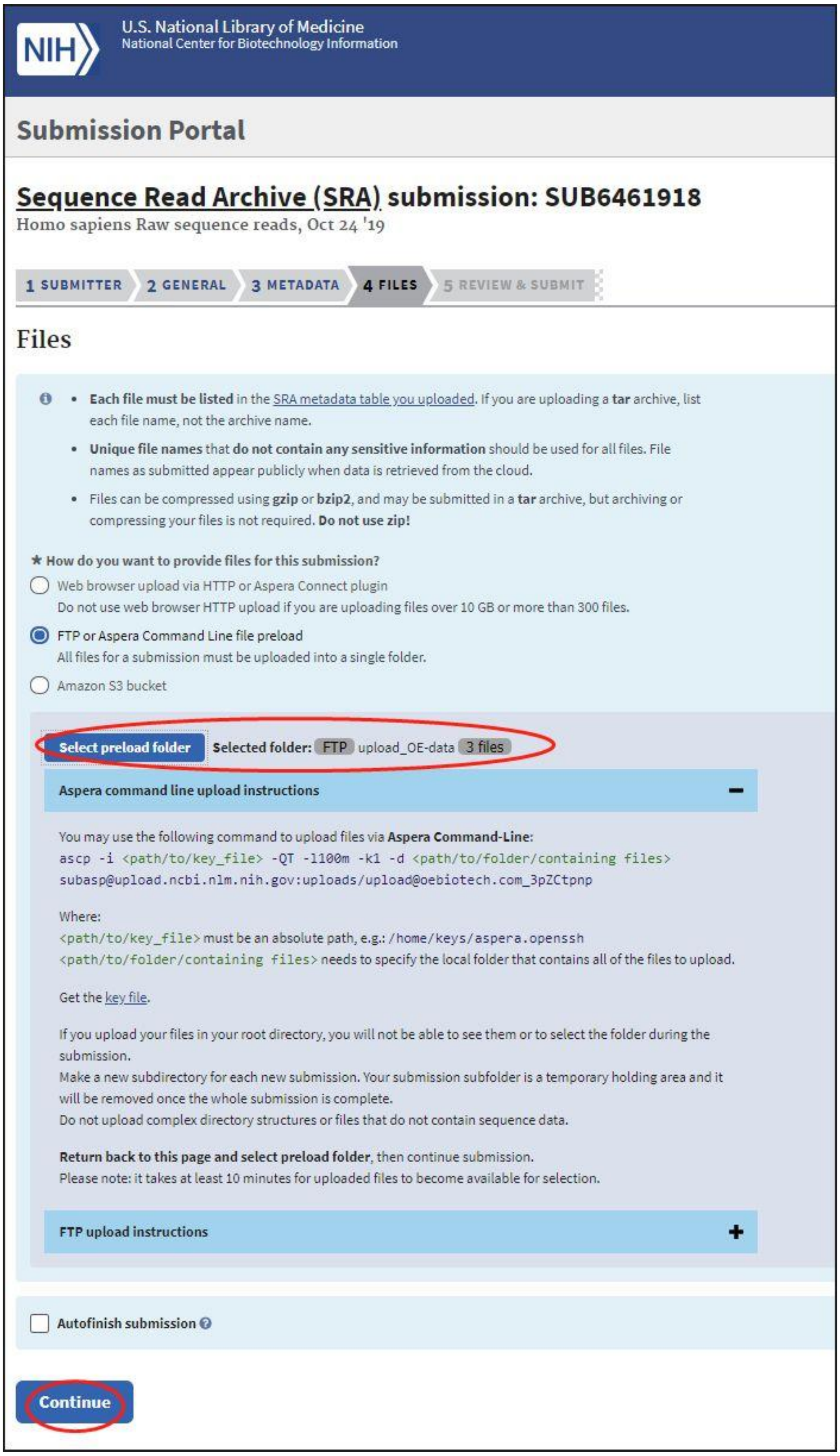
Please note: it takes at least 10 minutes for uploaded files to become available here.

Name	Files	Size	Modified
○ FTP new	3	1.1 MB	2019-10-29 14:03
● FTP upload_OE-data Sample_A.R2.fastq.gz (5.71GB) Sample_A.R1.fastq.gz (2.69GB) Sample_B.R1.fastq.gz (2.97GB)	3	11.3GB	2019-11-14 08:39

You may use the following command to upload files via **Aspera Command-Line**:

```
ascp -i <path/to/key_file> -QT -l100m -k1 -d <path/to/folder/containing files>
subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp
```

确认没问题，点击继续：



Submission Portal

Sequence Read Archive (SRA) submission: SUB6461918

Homo sapiens Raw sequence reads, Oct 24 '19

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

Selected folder: FTP upload_OE-data 3 files

Aspera command line upload instructions

You may use the following command to upload files via **Aspera Command-Line**:

```
ascp -i <path/to/key_file> -QT -l100m -k1 -d <path/to/folder/containing files>
subasp@upload.ncbi.nlm.nih.gov:uploads/upload@oebiotech.com_3pZCtpnp
```

Where:
`<path/to/key_file>` must be an absolute path, e.g.: `/home/keys/aspera.openssh`
`<path/to/folder/containing files>` needs to specify the local folder that contains all of the files to upload.

Get the key file.

If you upload your files in your root directory, you will not be able to see them or to select the folder during the submission.
Make a new subdirectory for each new submission. Your submission subfolder is a temporary holding area and it will be removed once the whole submission is complete.
Do not upload complex directory structures or files that do not contain sequence data.

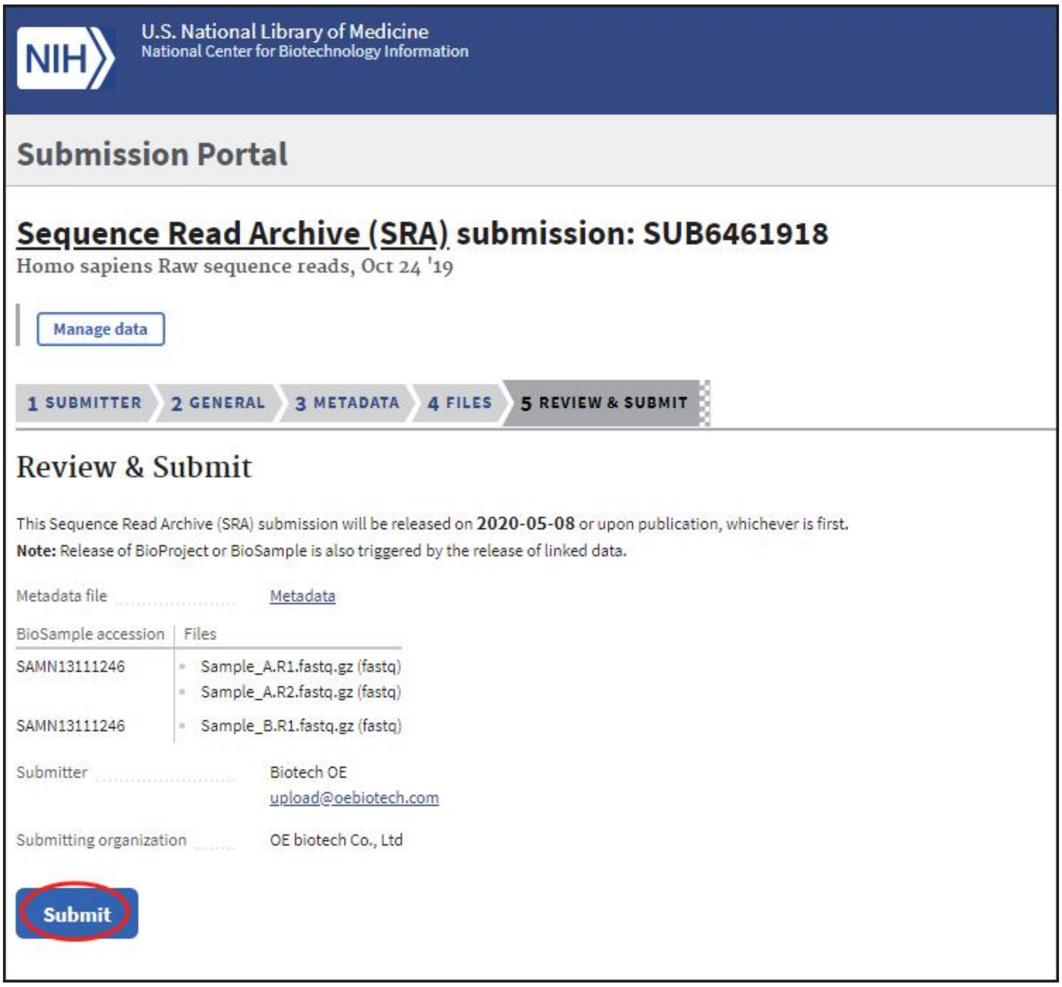
Return back to this page and select preload folder, then continue submission.
Please note: it takes at least 10 minutes for uploaded files to become available for selection.

FTP upload instructions

☐ **Autofinish submission**

4.5 OVERVIEW

最后 Overview 中浏览概况，确认无误后点击 Submit 提交：



Submission Portal

Sequence Read Archive (SRA) submission: SUB6461918

Homo sapiens Raw sequence reads, Oct 24 '19

Review & Submit

This Sequence Read Archive (SRA) submission will be released on **2020-05-08** or upon publication, whichever is first.

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

Metadata file: Metadata

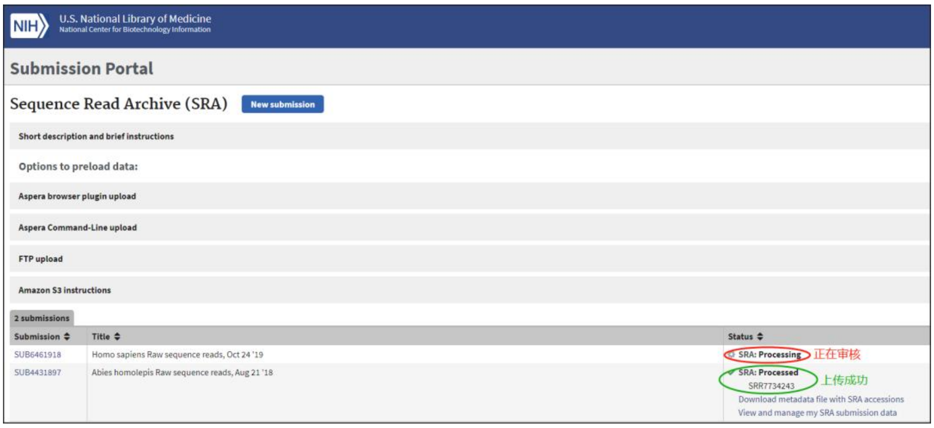
BioSample accession	Files
SAMN13111246	Sample_A.R1.fastq.gz (fastq) Sample_A.R2.fastq.gz (fastq)
SAMN13111246	Sample_B.R1.fastq.gz (fastq)

Submitter: Biotech OE
upload@oebiotech.com

Submitting organization: OE biotech Co., Ltd

5 完成上传

可以在 https://submit.ncbi.nlm.nih.gov/subs/sra/查看审核状态：



Submission Portal

Sequence Read Archive (SRA)

Short description and brief instructions

Options to preload data:

Aspera browser plugin upload

Aspera Command-Line upload

FTP upload

Amazon S3 instructions

2 submissions

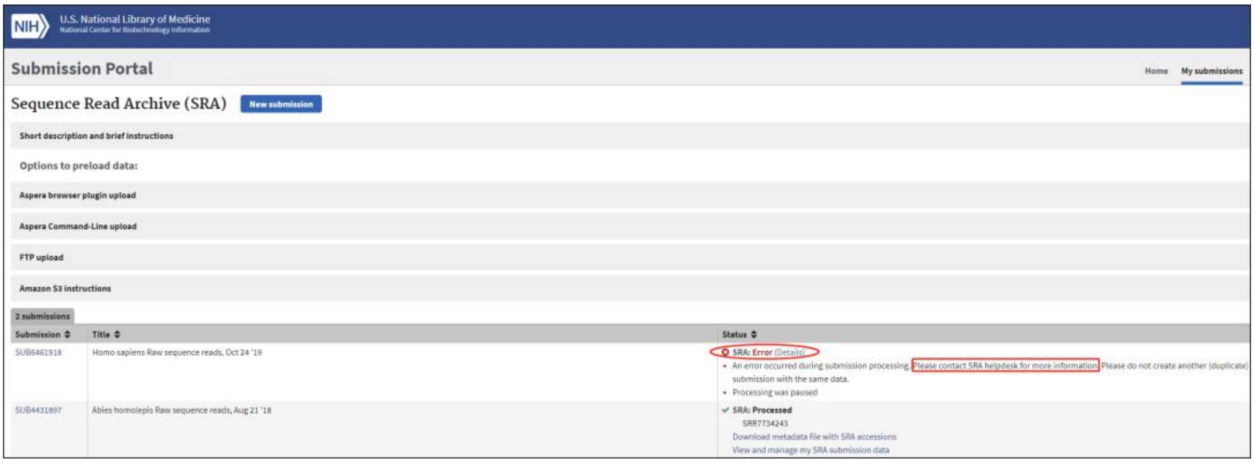
Submission	Title	Status
SUB6461918	Homo sapiens Raw sequence reads, Oct 24 '19	SRA: Processing
SUB4431897	Abies homolepis Raw sequence reads, Aug 21 '18	SRA: Processed SRR7734243 Download metadata file with SRA accessions View and manage my SRA submission data

目前测序数据上传完成。NCBI 会在后台审核数据，最后给出 Accession Number，一般一天左右。

审核完毕的数据可能不会马上被 NCBI 收录，一般需要 2-4 天才能被搜索到(如果设定了发布时间，则在发布时间之后才能搜到)。可以在 SRA 的 Submission 中(https://www.ncbi.nlm.nih.gov/Traces/sra_sub/?login=pda)查询到目前的审核进度。

注意最终使用的 Accession Number 一般为 Run 的编号，SRR 开头。可以点击 Submission Id 进入查看 Accession Number：

Tips：上传过程中遇到任何问题(如下图的报错信息)可以直接邮件联系 NCBI 请求帮助(邮箱：sra@ncbi.nlm.nih.gov)



Submission Portal

Sequence Read Archive (SRA)

Short description and brief instructions

Options to preload data:

Aspera browser plugin upload

Aspera Command-Line upload

FTP upload

Amazon S3 instructions

2 submissions

Submission ⇕	Title ⇕	Status ⇕
SUB6461918	Homo sapiens Raw sequence reads, Oct 24 '19	SRA: Error (Details) • An error occurred during submission processing. Please contact SRA helpdesk for more information. Please do not create another (duplicate) submission with the same data. • Processing was paused
SUB4431897	Abies homolepis Raw sequence reads, Aug 21 '18	✓ SRA: Processed SRR7734243 Download metadata file with SRA accessions View and manage my SRA submission data

至此，整个数据上传就已经全部完成了；Accession Number 放进文章，完成文章接收的最后一步。

SRA 数据上传的界面会时有更新，我们也会定期跟踪，及时更新，确保本指南的实用性。