

微生物基因组数据上传指南

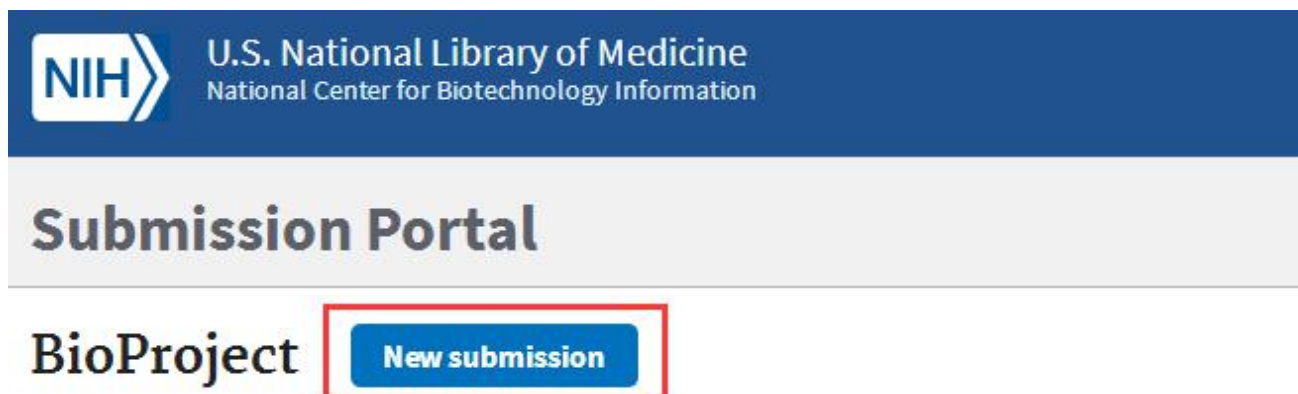
BioProject ID 的获得

1、在 NCBI 主页右上角 (<https://www.ncbi.nlm.nih.gov/>) My NCBI 登录系统中创建新的账号 (已有就不需要) , 点击 Register for an account 创建账号 :

2、登录 BioProject (<https://submit.ncbi.nlm.nih.gov/>) ,获取一个 BioProject ID

Sequence Data	Biological Research Project Data
GenBank Ribosomal RNA (rRNA), rRNA-ITS or Influenza sequences 3 Unassembled reads should be submitted to SRA . All other submission types should use one of the alternate submission tools (e.g. Bankit, Sequin, tbl2asn, etc.)	BioProject 1 A collection of biological data related to a single initiative, originating from a single organization or from a consortium.
Genomes (WGS or complete) 1 Prokaryotic and eukaryotic genomes that are either draft/incomplete (WGS) or complete	BioSample 1 Descriptions of biological source materials used in experimental assays.
	Microarray Data

3、点击 New submission , 进行提交



4、填写信息 (该步骤需要您填写的邮箱进行验证)

1 SUBMITTER 2 PROJECT TYPE 3 TARGET 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 OVERVIEW

Submitter

注意：标星部分为必填信息

★ First (given) name	Middle name	★ Last (family) name
<input type="text" value="LI"/>	<input type="text"/>	<input type="text" value="li"/>

★ E-mail (primary)	E-mail (secondary)	At least one e-mail should be from the organization's domain
<input type="text" value="986380125@qq.com"/>	<input type="text" value="clge@lc-bio.com"/>	

Group for this submission

(affiliation from my personal profile)

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

★ Submitting organization	Submitting organization URL	★ Department
<input type="text" value="Huazhong University of Science and Technology"/>	<input type="text"/>	<input type="text" value="College of Life Science & Technology"/>

Phone	Fax
<input type="text"/>	<input type="text"/>


★ Street	★ City	State/Province	★ Postal code	★ Country
<input type="text" value="Hongsan borough"/>	<input type="text" value="WUHAN"/>	<input type="text" value="HUBEI"/>	<input type="text" value="430074"/>	<input type="text" value="China"/>

填写完成后点击Continue



Update my contact information in profile


5、类型填写 (根据项目的实际情况进行填写)

Project Type

★ 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  ★ Target description 

Other  Streptococcus thermophilus

6、TARGET 填写


BioProject submission: SUB4202385
Genome sequencing






1 SUBMITTER 2 PROJECT TYPE **3 TARGET** 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 OVERVIEW


Target

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

填写详细信息，标星为必填项

★ Organism name 

Strain  Breed  Cultivar  Isolate name  Label 

Description 

Continue

7、General Info 信息填写

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?

选择数据释放时间，根据需求进行选择

- Release immediately following processing
- Release on specified date or upon publication, whichever is first

★ Project title

Streptococcus thermophilus Genome sequencing

★ Public description

信息描述

Relevance

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

- No Yes (not very common)

8、BIOSAMPLE 信息填写（点击 register at BioSample 进行详细信息填写）

1 SUBMITTER 2 PROJECT TYPE 3 TARGET 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 OVERVIEW

BioSample

Sample

Delete

+ Add another BioSample

Note: 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

1) 释放时间

BioSample submission: SUB4203158

New

1 GENERAL INFO 2 SAMPLE TYPE 3 ATTRIBUTES 4 DESCRIPTION 5 OVERVIEW

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

Continue

2) 样本类型 (根据样本来源情况进行填写)

Genome, metagenome or marker sequences (MxS 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 [MxS](#).

- Environmental/Metagenome Genomic Sequences [MIMS](#)
- Cultured Bacterial/Archaeal Genomic Sequences [MIGS](#)**
 - No environmental package
 - air
 - built
 - host-associated
 - human-associated
 - human-gut
 - human-oral
 - human-skin
 - human-vaginal
 - microbial mat/biofilm
 - miscellaneous or artificial
 - plant-associated
 - sediment
 - soil
 - wastewater/sludge
 - water
- Eukaryotic Genomic Sequences [MIGS](#)
- Viral Genomic Sequences [MIGS](#)
- Specimen Marker Sequences [MIMARKS](#)
- Survey-related Marker Sequences [MIMARKS](#)

3) 属性填写 (根据实际情况进行信息添加)

BioSample submission: SUB4203204

MIGS Cultured Bacterial/Archaeal sample from Streptococcus thermophilus

1 GENERAL INFO 2 SAMPLE TYPE 3 ATTRIBUTES 4 DESCRIPTION 5 OVERVIEW

Attributes

Package MIGS: cultured bacteria/archaea, water; version 4.0

★ Sample Name ?

★ Organism ?

★ strain ?

▼ Environment

★ collection date ?

★ environment biome ?

4) 描述信息添加

BioSample submission: SUB4203204

MIGS Cultured Bacterial/Archaeal sample from Streptococcus thermophilus

1 GENERAL INFO 2 SAMPLE TYPE 3 ATTRIBUTES 4 DESCRIPTION 5 OVERVIEW

Title and Comments

Sample Title ?

i 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 ?

Continue

5) 信息确认，确认无误后点击 submit

BioSample submission: SUB4203204

MIGS Cultured Bacterial/Archaeal sample from Streptococcus thermophilus






点击 Submit 后会自动跳回至 BioProject

BioProject submission: SUB4202385

Streptococcus thermophilus Genome sequencing



BioSample

Sample	Delete
Created: 2018-06-26	
SUB4203158	
CS20 Organism: Streptococcus thermophilus Created: 2018-06-26	

[+ Add another BioSample](#)

i 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

9、发表杂志的 PubMed ID 或 DOI 信息填写，若无可不填写

BioProject submission: SUB4202385

Streptococcus thermophilus Genome sequencing



Publications

PubMed ID  OR DOI 

[+ Add another publication](#)

Continue

10、信息确认，确认无误后点击 submit

BioProject submission: SUB4202385

Streptococcus thermophilus Genome sequencing

1 SUBMITTER 2 PROJECT TYPE 3 TARGET 4 GENERAL INFO 5 BIOSAMPLE 6 PUBLICATIONS 7 OVERVIEW

Overview

填写完成后，约几分钟的时间就收到 NCBI 审核的邮件，刷新上传的界面，状态从 Processing 变成 Processed

11、获得的 BioProject ID，以 PRJNA 字符为前缀，并且获得通过自动分配方式获得项目唯一的 Locus Tag Prefix 值，比如下面例子的 1306，该值用于注释结果 locus_tag 这一项的前缀（注意：注释文件必须含有该值！）如下图红框所示：

Submission	Title	Group	Status
SUB4202385	Streptococcus thermophilus Genome sequencing		✓ BioProject: Processed PRJNA477906 Streptococcus thermophilus Genome sequencing (TaxId: 1308)

组装结果的准备（该部分步骤先跳过，先进行序列的提交部分的上传步骤，如后续上传有问题再进行该步骤）

1、生成 template (*.sbt) 文件，填写 submission template form

(<http://www.ncbi.nlm.nih.gov/WebSub/template.cgi>)。 template form 如下图所示(需要填写提交人的各项信息，姓名，地址，单位，联系方式等等，还有文章的题目及签名获得 BioProject ID)：

GenBank Submission Template

Contact Information

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

LI li

★ E-mail (primary)

BioProject

BioSample

填写完成后，点击最下方的 Create Template，生成 sbt 文件，备用；



2、准备基因组文件，要求基因组序列文件，无 gap，即序列中不含 N，每个文件不超过 10，000 条序列。

格式：该文件为标准的 fasta 格式，第一行是描述信息，以 ">" 开头；第二行起始序列信息，每行长度不超过 80 个字符。如下图所示：

```
>Scaffold1_1
CTGTTCTTCGGTGAATCGTTTCTTCATGTCCAATCTCCATGTGGGGTTGATTGGACTCTA
AAGTGACGTGCTACTCAATTCTGGGGGGACGTCGCATGAGGCTGAGTCCCTACTCAAAGC
TGCCCTTGTGAGCATGGGCAGGGCCATTGGCTCCAACAACATGCTGGAGCGCCAGAAGGA
GGGGTTGGAAGGGCAATTCGGGCAGCCTCAATGGGCCATGCGGGGACCGCTGCTTAGTC
ATTTGTATAGGTGAGAAAATCTCTACCGGTATTTGGCGGCAGTAATGGCGTATCCAAATG
AGATGTCTACTATGGTTATTTTCGCATGGCCTGTTGTTGCATTACGAAAATAAAATGCTG
GGTCAACATGCGATGCCAATAGAGCTTGGATAAAAGCTGATAAATAGCGGATTTTGTCTG
GGTGGTCATTTTATCTGCTCATCATTGCCATTTTGGTTCGGATAGTAACCATTCTGGGTG
ACAGTTTTTATATATGTGAATACTTTTGATACTATTAGGTTGAGTGAATGAGAGTGTTGG
CGGTCCAATCATCTAGCGAACTATTTGCTGCCAATGGATGATTTTTTTGATTAAATCTTTT
TATGAAAAGATGGGATAATTGGTCGTGTGGTGAGGTTTAATTTTTAGATTGATAGATGAT
TGTATTTAAATGGATGGCATCCATTCCCTCGAACGGTACAAGAATATCGTCACTCAAACCTC
GGCAGCTTTGAGAGTTCTCAAAGGCTATCTGCGGGTTAATGCCTTTGCTGTATCGCCCA
```

序列的提交

1.上传生成得到的*.sqn 文件(具体生成步骤详见最后一部分**利用 tbl2asn 软件生成符合 NCBI 上传规则的 *.sqn 文件**) 或者 FASTA 文件，网址如下：<https://submit.ncbi.nlm.nih.gov/subs/wgs/>

2.按照上传引导填写相应信息，上传文件，整个过程需要使用一致的物种名称。

Submission Portal

Genome

New submission



Note: To find submissions started before Feb. 3, 2014, go to the [previous version](#) of the WGS submission wizard.

基于序列类型进行选择上传

Genome submission: SUB4204626

New

Submission Type

★ How do you want to submit your data?



Single genome

Manually complete a web form to describe one genome assembly and to upload its sequences.



Batch/multiple genomes

Upload the "Genome Info" file ([download template ahead of time](#) or during submission), a tab-delimited text file that describes each of your genome assemblies and their attributes/metadata, plus the genome sequences (one file per genome).

Information that must be common to all genomes in the batch are:

- BioProject
- (initial) release date
- assembly type (either WGS or non-wgs, not a mix of both types)
- file type (FASTA or SQN)
- gap/Ns details
- publication information (for FASTA submissions only)
- PGAP request status (Yes/No; for prokaryotic genomes only)
- [See more details here](#)

Continue

剩下的步骤按照跳转提示一步一步进行填写即可（**这里部分重复的步骤省略**）；

1) 其中信息部分，前面已经生成过 BioProject 和 BioSample 填写相应的 ID 即可，其中 BioSample 为 SAMN 开头的信息；

1 SUBMITTER 2 GENERAL INFO 3 SOURCE 4 FILES 5 ASSIGNMENT 6 REFERENCES 7 OVERVIEW

General Information

BioProject

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

Yes No

* Project

[PRJNA477906](#) Streptococcus thermophilus Genome sequencing

Organization: Huazhong University of Science and Technology

 The BioProject bundles the data for this research project.

BioSample

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

Yes No

* Sample

[SAMN09487225](#) CS20 Organism: Streptococcus thermophilus Tax ID: 1308 

Submitted: 2018-06-26

* Assembly method

ABySS

* Version or Date program was run


v2.0.2


Delete 

* Sequencing Technology

Illumina HiSeq 

PacBio 




 Add another sequencing technology

Delete

* Do you want to submit the PacBio reads and the motif_summary.csv analysis file that is produced by the standard PacBio methylation analysis program?

- Yes, reads and motif_summary.csv
 Yes, reads
 Yes, motif_summary.csv
 No, I already submitted those files to SRA
 No

 [Why is the PacBio base modification data important?](#)

2) Source 部分选择 No

1 SUBMITTER 2 GENERAL INFO 3 SOURCE 4 FILES 5 ASSIGNMENT 6 REFERENCES 7 OVERVIEW


Source

Prokaryote source

Bacteria and/or source DNA is available from 

* Annotate this prokaryotic genome in the [NCBI Prokaryotic Annotation Pipeline \(PGAP\)](#) before its release?

Yes No

 Note that annotation is not required, but you may request to have your prokaryotic genome annotated by PGAP before it is released. The pre-release annotated genome will be posted back to this submission for you to view. If no Release date has been requested in this submission, the PGAP-annotated genome will be released upon completion of processing.

Continue

3) 数据上传类型选择 (如果选择.sqn 格式的文件需要按照利用 **tbl2asn** 软件生成符合 NCBI 上传规则的

***.sqn 文件**步骤生成该格式文件, 选择 fasta 格式直接上传即可)

Files for submission

Which of these 3 options describes this genome submission?

- 1. Each chromosome is in a single sequence and there are no extra sequences
 - There can still be gaps within the sequences.
We will prompt you to provide the information for any Ns that represent gaps.
 - Internal sequences must be arranged in the correct order and orientation.
Sequences concatenated in unknown order are not allowed.
 - Plasmids and organelles can still be in multiple pieces.
 - If the sequences are assembled using an AGP file, choose the next option.
- 2. One or more chromosomes are still in multiple pieces and/or some sequences are not assembled into chromosomes
 - This will be processed as a WGS genome and may include AGP files in the submission
 - There can still be gaps within the sequences.
We will prompt you to provide the information for any Ns that represent gaps.
 - Internal sequences must be arranged in the correct order and orientation.
Sequences concatenated in unknown order are not allowed.
- 3. We are submitting just the AGP file(s) for a genome assembly; the components of the AGP file are already in GenBank

Select file type for the sequences

- ASN.1 (.sqn) FASTA

选择上传文件类型, sqn或者FASTA, 其中FASTA格式已有, sqn需要额外生成

Select upload type

- I have all files preloaded for this submission
 I will upload all the files now via HTTP/Aspera

Current versions of browsers Firefox, Chrome, Safari or Internet Explorer are recommended.
 To upload large eukaryotic files (larger than 2GB), please use [Aspera Connect plugin](#).
 Please note: in order to use Aspera for file upload with Chrome, you need to update Aspera Connect plugin to version 3.6 or newer. [More details...](#)

Upload FASTA

选择文件 | 未选择任何文件

选择需要上传的文件 (只能选择一个文件, 因此需要将多个染色体序列文件合并成一个文件), 注意当文件较大时不建议这种方式上传

4) 该部分根据样本实际情况进行填写 (是否含有质粒, 是否完整基因组是否为环状等等)

1 SUBMITTER 2 GENERAL INFO 3 SOURCE 4 FILES 5 ASSIGNMENT 6 REFERENCES

Assignment

Warning: Reminder: you selected option 1 in the Files tab, so each chromosome must be sequence, the chromosome(s) must be one of the sequences in this submission be assigned to a chromosome or plasmid (or organelle). Please provide that in the submission type to option 2 (WGS) in the Files tab.

★ Does any sequence belong to a plasmid?

- Yes No

Chromosomes

★ Does the organism have only one chromosome?

- Yes No

★ Sequence ID

CS20

Length Complete Circular

1936216

Continue

5) 文献情况，请根据实际情况填写

Genome submission: SUB4204626

CS20.fna genome submission

1 SUBMITTER 2 GENERAL INFO 3 SOURCE 4 FILES 5 ASSIGNMENT 6 REFERENCES

References

Sequence authors

★ First (given) name MI ⓘ ★ Last (family) name Delete

+ Add another sequence author

Reference

★ Publication status

Unpublished In-press Published

★ Reference title

Reference authors

Same as sequence authors Specify new authors

[Continue](#)

6) 信息确认，确认无误后点击 submit

上传完成后，需要等 NCBI 的审核，审核完成后会邮件通知上传者。

最后简单介绍下*.sqn 文件生成的步骤，如有需要请使用：

利用 tbl2asn 软件生成符合 NCBI 上传规则的*.sqn 文件

1.准备生成*.sqn 文件。该文件需要两个文件：1) 前面生成的*.sbt；2) 基因组序列文件--即 03.Assmebly 文件夹下的 fna 文件；

2.下载 tbl2asn 软件，下载地址如下：

ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn

转到高层目录

02/12/2012 12:00上午	19	DOCUMENTATION	
02/12/2012 12:00上午	26	README	
04/13/2016 12:00上午	18	linux.tbl2asn.gz	
03/31/2015 12:00上午	60	linux32.tbl2asn.gz	
05/16/2018 05:20下午	67	linux64.tbl2asn.gz	
05/16/2018 02:46上午	3,689,562	mac.tbl2asn.gz	
09/03/2014 12:00上午	44	solaris-x86.tbl2asn.gz	
03/31/2014 12:00上午	44	solaris.tbl2asn.gz	
05/16/2018 05:21下午	34	win.tbl2asn.zip	← windows版本

该软件说明如下：

<http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.html>

3.将以上两个文件置于 tbl2asn 软件目录下，进行命令行命令（开始-cmd），进入 tbl2asn 软件目录，输入如下命令后回车运行（注意空格和 "" 号）：

```
tbl2asn.exe -i *.fna -t *.sbt -a s -V v -Z log -j "[organism=*][strain=*]"
```

[organism=*][strain=]*部分内容需要自行添加；顺利运行后，tbl2asn 将出现由三个后缀名的文件

.sqn,.val,log。*.sqn 文件用于最后的提交作业；一般来说，*.val 文件大小为 0k 则整个转换过程无问题。

4.检查输出的*.val 文件和报告文件。查看*.val 文件内是否提示错误信息，如果有，找出并解决，以减少上传审核时间。