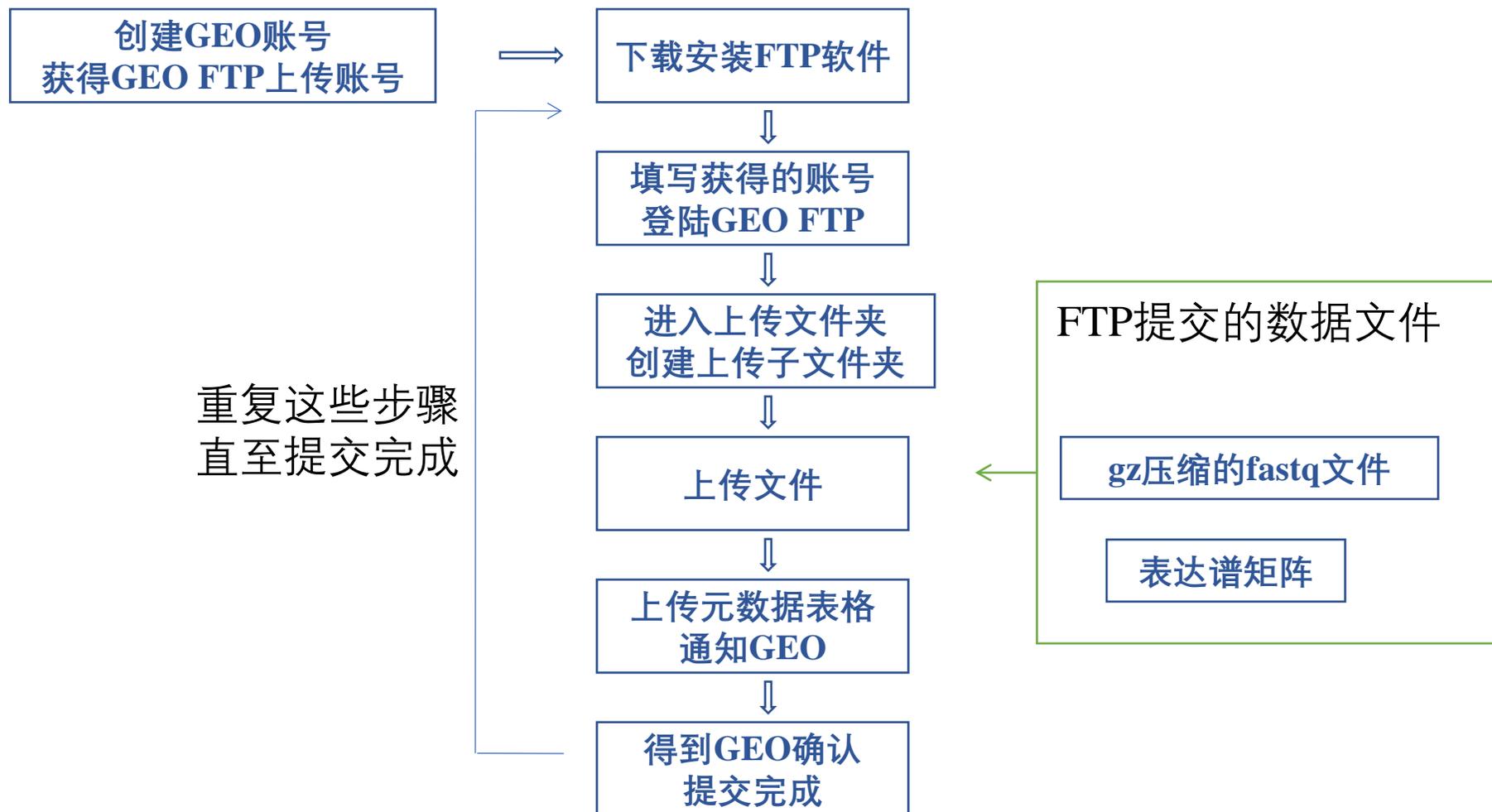


# RNA-seq原始数据 上传到GEO数据库详细步骤

陈明杰  
202411

# GEO数据提交流程

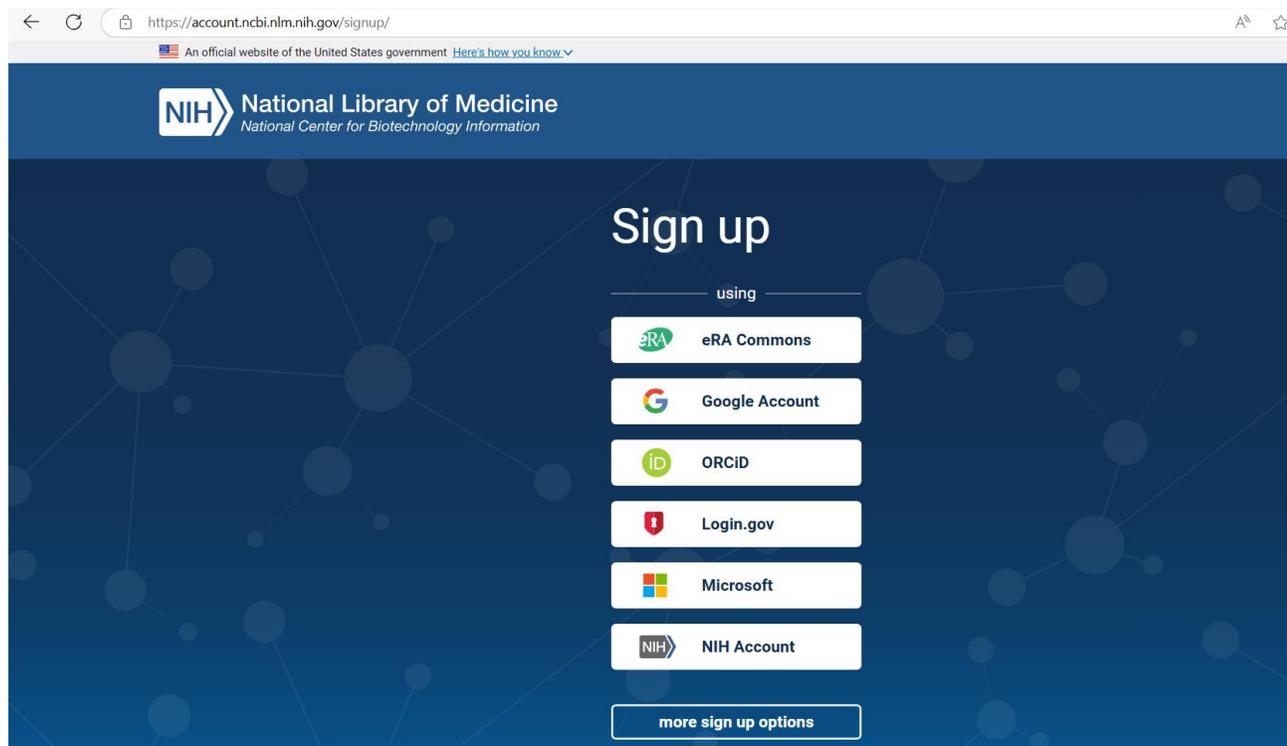


# 提交前的准备

- 1, 创建用户账号 <https://www.ncbi.nlm.nih.gov/account/>
- 2, FTP上传软件
  - 推荐winscp: <https://winscp.net/eng/index.php>
- 3, 三类文件
  - Raw data: gz压缩的FASTQ原始文件
  - Metadata: 元数据表格 (下载并填写)
  - Processed data: 表达谱数据 (count矩阵, FPKM矩阵或者TPM矩阵等)

# 注册账号

<https://account.ncbi.nlm.nih.gov/signup/>



请勿使用163，QQ邮箱，建议ORCID或者Microsoft创建新账号，接收邮件并点击激活链接，激活

## 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

### 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](#)  
[ENCODE Data Listings and Tracks](#)

### Information for Submitters

[Login to Submit](#)

[Submission Guidelines](#)

[Update Guidelines](#)



NCBI » GEO » Info » Submitting data [GEO Publications](#) [FAQ](#) [MIAME](#) [Email GEO](#) [Login](#)

### Submitting data

GEO is an open-access archive of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data.

### Data types

[Submit high-throughput sequencing \(HTS\)](#)  
[Submit microarray and other non-HTS data types](#)

**WARNING:** If you are submitting human data, it is your responsibility to comply with Human Subject Guidelines.

### Fast facts

- Your final GEO records will be organized as illustrated at [GEO Overview](#).
- See examples of the kinds of data GEO can accept.
- GEO accession numbers are normally approved within 5 business days after completion of submission. If you do not receive an e-mail from us within 5 business days of your submission, please first check your spam or junk e-mail folders because some systems recognize GEO e-mail correspondence as spam, then e-mail us to inquire about your submission.
- Your GEO submissions can remain private until a manuscript citing the data is published.
- You can allow reviewers anonymous access to your private records.
- You can update or edit your existing GEO records at any time.
- GEO supports MIAME- and MINSEQE-compliant data submissions.

## Submitting high-throughput sequence data to GEO

- Submission instructions [YouTube](#)
  - Metadata spreadsheet **REQUIRED**
  - Processed data files **REQUIRED**
  - Raw data files **REQUIRED**
- Tutorial video
- Data file compression
- Single-cell studies
- NanoString GeoMx Digital Spatial Profiling (DSP)
- Organizing your submission
- Uploading your submission
- General information
  - Data provisions, standards and administration
  - Categories of sequence submissions accepted by GEO

**WARNING:** If you are submitting human data, it is your responsibility

### Submission instructions

GEO accepts next generation sequence data that examine quantitative (e.g., RNA-seq) and other aspects of functional genomics using methods such as RNA-seq (e.g., RNA-seq), ChIP-seq, HIC-seq, methyl-seq, etc. We process all components of your submission, including processed data files, and we submit the raw data files to the Sequence Read Archive (SRA).

**Step 1.** Check that GEO accepts your data type.

**Step 2.** Gather raw files.

**Step 3.** Gather processed data files.

**Step 4.** Download metadata spreadsheet and fill in Metadata tab for your study. Use one spreadsheet per data type (e.g., ChIP-seq, RNA-seq).

**Step 5.** In the metadata spreadsheet file, list the MD5 checksum for all raw and processed data files in the 'MD5 Checksums' tab.

**Step 6.** Create single folder on your computer that contains all raw and processed data files for your experiment. If you have multiple data types, please use one folder per experiment.

**Step 7.** Transfer your data to GEO by FTP using these instructions. **ftp账号密码、路径等信息**

**Step 8.** After FTP transfer has completed, submit metadata file(s) on the [Submit to GEO](#) page.

More information on required components:

- **Metadata spreadsheet**

[Download metadata spreadsheet](#)

元数据表格（用前下载，保持最新）

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.

- **Processed data files**

GEO requires that submitters deposit the processed data that support the findings of their study. The processed data should have a quantitative component, such as gene abundances or other count data. Please do not submit alignment files (e.g., BAM, SAM, BED) as processed data, as these are considered intermediary files and do not include a quantitative component. When standard alignments are the only processed data available, please [write to us](#) to inquire about whether your data are suitable for submission to GEO.

Processed data format and content will depend on the data type: RNA-seq processed data can include raw and/or normalized counts (FPKM, TPM, etc) of sequencing reads for the features of interest (protein-coding genes, lncRNA, miRNA, circRNA, etc).

ChIP-Seq and ATAC-seq processed data can include peak files with quantitative data, tag density files, etc. Common formats include WIG, bigWig, bedGraph. Please leave files in native format and do not paste peak data into Excel.

- **Raw data files**

Raw data are a required part of GEO submissions. The raw data files should be the original files containing reads and quality scores, as generated by the sequencing instrument. Edited files may not be processed correctly by SRA.

Raw data for high throughput sequencing studies submitted to GEO will be brokered to SRA for you.

Raw data can instead be submitted directly to [SRA](#). After you have received the SRA accessions, please see [above](#) for instructions and [specific template](#) for this case. Please submit the metadata and processed data to GEO.

# 元数据表格 (应填尽填)

## 1, 实验设计, 人员信息

STUDY	LIBRARY	PROTOCOL	PAIRED-END EXPERIMENTS
9 # This section describes the overall study. 10 # Information provided in this section will be displayed in a GEO Series (GSE record) on public web pages. 11 title 12 summary (abstract) 13 experimental design 14 contributor 15 contributor 16 contributor 17 contributor 18 supplementary file	27 # Information provided in this section will be displayed in GEO Samples (GSM records) on public web pages. 28 # A GEO Sample record will be created from each row in this section. 29 # Biological replicates of the same sample, if provided, should be listed on different rows and listed accordingly (biol rep 1, biol rep 2, and so on). 30 # Technical replicates, eg. the same libraries were run in different lanes of a flow cell or sequenced multiple times, if provided, list all raw files in the same row, adding more "raw file" columns as needed to accommodate all raw files. 31 # All library names, titles and raw files must be unique. 32 library name title organism tissue cell line cell type genotype treatment molecule single or paired-end instrument model description processed data file processed data file raw file raw file raw file raw file	42 # Information provided in this section will appear in each GEO Sample (GSM record). 43 growth protocol 44 treatment protocol 45 reagent protocol 46 library construction protocol 47 library strategy 48 data processing step 49 data processing step 50 data processing step 51 data processing step 52 data processing step 53 data processing step 54 data processing step 55 genome build/assembly 56 processed data file format and content 57 processed data file format and content	59 # "paired-end" experiments are included, list the files for each paired-end run in a row. Each row will become one sequencing run on a GEO Sample (GSM record). 60 # Single-cell data: if applicable, list index files (1, 2, etc.) in "file name 3", "file name 4" columns. 61 # Provide more runs as raw files listed here are also listed in the "raw file" columns in the above SAMPLES section. 62 file name 1 file name 2 file name 3 file name 4 63 64 65 66 67 68 69

Legend:  
\* = required field  
\*\* = one of these fields must be completed.  
Include additional columns to thoroughly describe the samples: strain, genotype, sex, cultivar, breed, disease, etc.  
Include additional columns to describe experimental variables: treatment, time, antibody, condition, diet, etc.

Complete all columns in the SAMPLES section

All processed file names and raw file names must be listed here so that the Sample records can be linked to the correct files. The file names must exactly match the submitted file names (include extensions). Raw file names can only be listed once in the SAMPLES section.

## 2, 样品信息, 样品类型, 测序平台等

## 3, 样品抽提, 建库步骤

## 4, 数据分析相关信息

## 5, 原始数据文件

RAW FILES	PROCESSED DATA FILES
file name	file name
file checksum	file checksum
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	
17	
18	

## 6, 原始数据文件名及md5值

## 7, processed数据文件名及md5值

# PROCESSED DATA FILES

- 表达谱矩阵 (txt或者excel格式都行)

 mRNA Expression Profiling.xlsx      2020/2/20 8:45      Microsoft Excel ...      12,947 KB

	A	B	C	D	E	F	G
1	<b>Symbol</b>	<b>A1</b>	<b>A2</b>	<b>A3</b>	<b>B1</b>	<b>B2</b>	<b>B3</b>
2	Ndr4	154	167	164	25	41	13
3	Kcnk6	11	13	15	84	64	18
4	Ppp3r1	22	21	13	1	2	3
5	Nrip1	2	9	5	1	1	5
6	Agtpbp1	3	0	0	2	0	0
7	Parp6	33	33	3	1	1	0
8	Add1	0	5	0	4	0	3

建议标明注释版本

# 原始文件及md5值

gz压缩的FASTQ文件

A_R1.fastq.gz	2018/9/11 6:54	WinRAR 压缩文...	3,333,257...
A_R2.fastq.gz	2018/9/11 6:58	WinRAR 压缩文...	3,880,918...
B_R1.fastq.gz	2018/9/11 7:02	WinRAR 压缩文...	3,432,598...
B_R2.fastq.gz	2018/9/11 7:07	WinRAR 压缩文...	4,080,836...
C_R1.fastq.gz	2018/9/11 7:11	WinRAR 压缩文...	3,256,546...
C_R2.fastq.gz	2018/9/11 7:14	WinRAR 压缩文...	3,897,603...
D_R1.fastq.gz	2018/9/11 7:18	WinRAR 压缩文...	3,301,864...
D_R2.fastq.gz	2018/9/11 7:22	WinRAR 压缩文...	3,924,098...
E_R1.fastq.gz	2018/9/11 7:25	WinRAR 压缩文...	3,344,490...
E_R2.fastq.gz	2018/9/11 7:29	WinRAR 压缩文...	4,004,335...
F_R1.fastq.gz	2018/9/11 7:32	WinRAR 压缩文...	3,478,879...
F_R2.fastq.gz	2018/9/11 7:36	WinRAR 压缩文...	4,154,313...

校验文件正确性 `seqkit stats -a *.gz`

校验文件完整性

Win10系统: `Certutil -hashfile sample.fastq.gz md5`

Linux系统: `md5sum sample.fastq.gz`

Mac系统: `md5 sample.fastq.gz`

对  
暗  
号

# 用户信息表 (GEO页面上的信息)

## My GEO Profile

<b>Investigator</b> Use this section to provide details about the primary investigator. This information will be displayed on GEO records.	<b>Organization name*</b> Shanghai Newcore Biotechnol
<b>First name*</b> Jimmy	<b>Department</b> 
<b>Middle name</b> 	<b>Lab</b> 
<b>Last name*</b> Chen	<b>Street address*</b> Room 309, Building C, No.154
You may choose not to display your email or phone on GEO records by unchecking the corresponding checkbox.	<b>City*</b> shanghai
<b>E-mail(s)*</b> ding@bioinformatics.com.cn	<b>State/province</b> 
<b>Show e-mail</b> <input checked="" type="checkbox"/>	<b>ZIP/Postal code*</b> 200000
<b>Phone</b> 13917006049	<b>Country*</b> China
<b>Show phone</b> <input checked="" type="checkbox"/>	

**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.

**Name**  
jimmy2

**E-mail(s)**  
ding2@bioinformatics.com.cn

**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.

[Preview](#) [Save](#)

[New submission](#)

此信息在页面上实时更新

## GEO File Transfer Protocol (FTP)

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

上传路径

Select data type:

数据类型

- High-throughput sequencing
- Microarray and other (NanoString, RT-PCR, etc.)

**Step 2.** Transfer all your raw and processed data files to your personalized upload space according to FTP upload instructions below. **Do not upload the metadata file by FTP.**

▶ Transfer Files

上传ftp账户信息

**Step 3.** After FTP transfer of raw and processed data files is complete, upload Excel metadata file on the Submit Metadata page.

[Upload metadata](#)

元数据上传

e. For LINUX/UNIX users, we recommend transferring files with 'ncftp' or 'lftp', but you can also use 'ftp', 'sftp', or 'ncftpput'. Please see below for detailed examples.

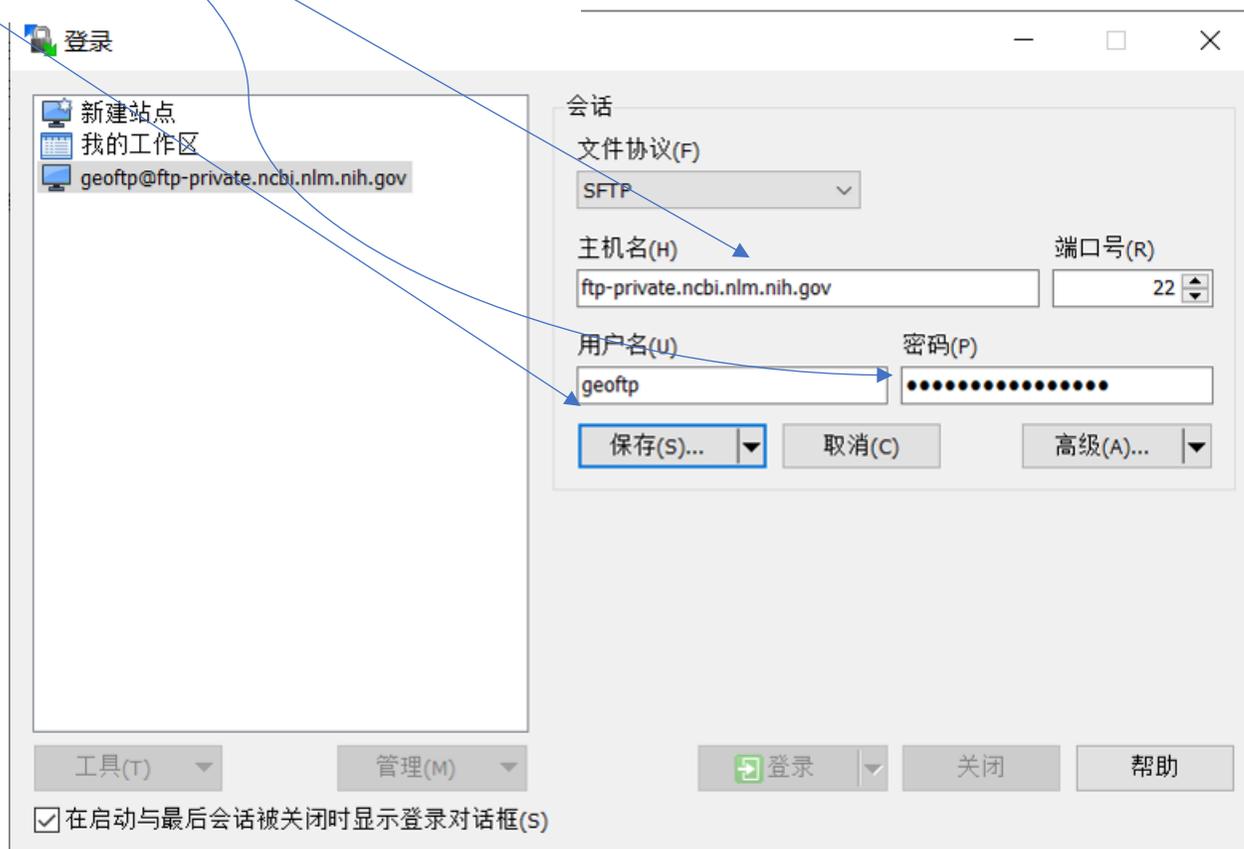
f. Our FTP server credentials are:

host address	ftp-private.ncbi.nlm.nih.gov
username	geoftp
password	XXXXXXXXXXXX

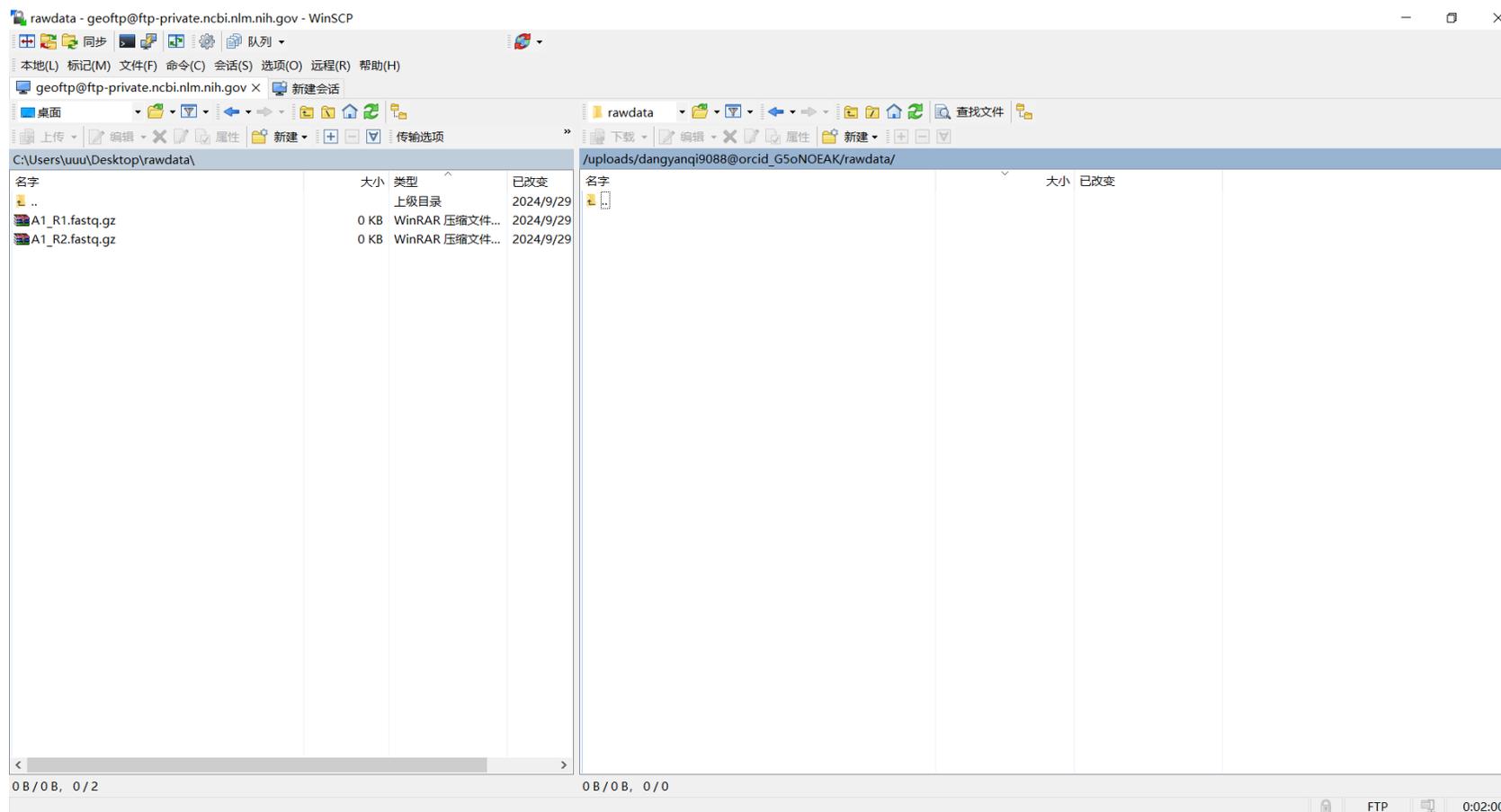
**Do not share these log-in credentials. Do not include these log-in credentials on a public page. These credentials are changed regularly, as per our security policies.**

g. After connecting, you **must navigate** to your personalized upload space:

uploads/XXXXXXXXXXXX@orcid\_XXXXXXXXXXXX



# FTP多线程上传



右侧点鼠标右键新建一个文件夹，例如raw\_data  
双击进去raw\_data，然后左侧点鼠标右键上传  
速度：每个线程越1.5G/h

# 通知GEO

## Submit to GEO



You are logged in under the ~~xxxxxx@xxxx~~ account. Messages from GEO regarding your submission will be sent to the following email address(es): [ding@bioinformatics.com.cn](mailto:ding@bioinformatics.com.cn), ~~xxxxxx@xxxx.com~~. If necessary, [visit your account](#) to edit your contact information. See [submitter accounts](#) for more details.

Use this page to upload Excel metadata file for a new sequence submission.

Instructions with [metadata template file](#) for submitting sequence submissions to GEO are available [here](#).

This page can accept only a single Excel metadata file at a time. If you have multiple Excel metadata files to upload, submit the second file after the first file has been successfully loaded, and so on.

### Select upload subfolder

Choose the subfolder that contains the raw and processed data files listed in the metadata file that you will upload below.

rawdata

### Excel metadata file to upload

选择文件 seq\_template.xlsx

### Submission release date (YYYY-MM-DD) ([more information about release dates](#))

2028-09-01

后续可以通过邮件让工作人员帮助修改

### Comment to GEO staff (optional)

备注信息

Submit

# GEO回复的邮件

NCBI My NCBI Sign Out

GEO Home Documentation Query & Browse Email GEO My GEO Submissions

## Submit to GEO

You are logged in under the [redacted] account. Messages from GEO regarding your submission will be sent to the following email address(es): [ding@bioinformatics.com.cn](mailto:ding@bioinformatics.com.cn), [redacted]. If necessary, visit your [account](#) to edit your contact information. See [submitter accounts](#) for more details.

### Submission Summary

Your metadata file has been successfully uploaded. Thank you for using the GEO Submission form.

Transferred files have been placed into the processing queue and will be reviewed within 5 business days. Expect to receive an email from GEO curators with your GEO accession numbers, or questions about your submission. We can be contacted at [geo@ncbi.nlm.nih.gov](mailto:geo@ncbi.nlm.nih.gov) if you do not hear from us within the allotted time, or if you require additional assistance.

Incomplete or incorrectly formatted submissions cannot be processed. A complete submission consists of:

1. Uploaded metadata file (Thank you!)
2. Raw data
3. Processed data

Please be aware that we do not have the resources to store files for incomplete submissions. If a submission has not been completed within two weeks, files will be removed from our servers.

Metadata file name	seq_template.xlsx
User ID	[redacted]
Public release date	2028-09-01
Comment	
Upload space subfolder	uploads/[redacted]

[Upload another metadata file](#)

### GEO submission summary

发件人: [geo@ncbi.nlm.nih.gov](mailto:geo@ncbi.nlm.nih.gov)

收件人: [redacted] <[redacted]> + 我 <[ding@bioinformatics.com.cn](mailto:ding@bioinformatics.com.cn)>

时间: 2024年09月29日 17:56 (星期日)

[sent to: "[redacted]" <[redacted]> "ding@bioinformatics.com.cn"]

Your metadata file has been successfully uploaded. Thank you for using the GEO Submission form.

Transferred files have been placed into the processing queue and will be reviewed within 5 business days. Expect to receive an email accession numbers, or questions about your submission. We can be contacted at [geo@ncbi.nlm.nih.gov](mailto:geo@ncbi.nlm.nih.gov) if you do not hear from us within additional assistance.

Incomplete or incorrectly formatted submissions cannot be processed. A complete submission consists of:

1. Uploaded metadata file (Thank you!)
2. Raw data
3. Processed data

Please be aware that we do not have the resources to store files for incomplete submissions. If a submission has not been completed, removed from our servers.

几分钟后会收到邮件

# GEO审核 (一般5个工作日)



----- MESSAGE BODY. YOU MAY CHANGE IT OR ADD COMMENTS ABOVE -----

Dear Submitter(s),

Thank you for your recent submission to the GEO repository.

However, the following files are corrupt:

```
rawdata[redacted]_1.fq.gz  unpigz: skipping: /panfs/traces01.be-md.ncbi.nlm.nih.gov/aspera/geo/
corrupted -- crc32 mismatch
```

```
rawdata[redacted].fq.gz  unpigz: skipping: /panfs/traces01.be-md.ncbi.nlm.nih.gov/aspera/geo/
corrupted -- crc32 mismatch
unpigz: abort: internal threads error
```

[redacted].fq.gz:

GEO computed: a04bc6f0267b7373e83b5a98daadbef5

meta sum: f6e9bc2bd739917af2445912af647688

文件损坏, 重新上传  
传好后, 回复下email即可

Thank you for the files. The records have been assigned GEO accession numbers as detailed below.

The records are scheduled to be publicly available on:

Sep 01, 2028

To change this release date, or to make other changes, please see:

<https://www.ncbi.nlm.nih.gov/geo/info/update.html>

\*\*\* It is your responsibility to keep track of the release date and to change it, when necessary change the release date of your private records are provided at <https://www.ncbi.nlm.nih.gov/geo>

\*\*\* If GEO accession numbers are quoted in any publicly available manuscript (including journal records must be released for public access, regardless of the scheduled release date (<https://www.ncbi.nlm.nih.gov/geo>

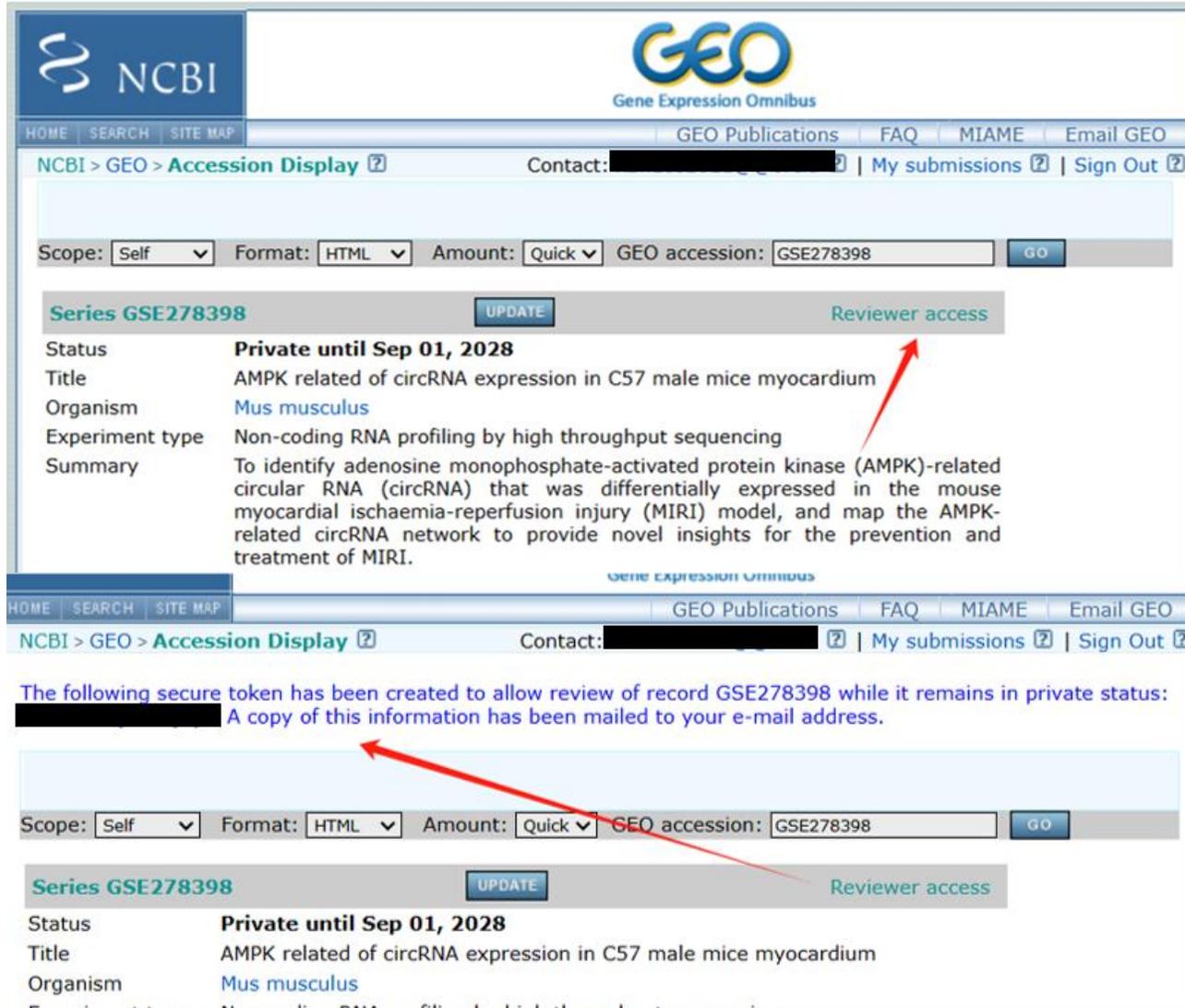
You can assist in keeping GEO up-to-date by informing us when any of your GEO accession numbers PubMed links and release data that is still private.

\* You may view your GSE278 study at:

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE278>

分配GEO号

# Reviewer token



The screenshot shows the NCBI GEO Accession Display page for GSE278398. The page is divided into two identical sections. The top section shows the series details, including the status 'Private until Sep 01, 2028' and a 'Reviewer access' link. A red arrow points from this link to the 'Reviewer access' link in the bottom section. Below the bottom section, a message states: 'The following secure token has been created to allow review of record GSE278398 while it remains in private status: [redacted]. A copy of this information has been mailed to your e-mail address.' A red arrow points from this message to the 'Reviewer access' link in the bottom section.

NCBI > GEO > **Accession Display** [?](#) Contact: [redacted] [?](#) | [My submissions](#) [?](#) | [Sign Out](#) [?](#)

Scope:  Format:  Amount:  GEO accession:

**Series GSE278398**  [Reviewer access](#)

Status **Private until Sep 01, 2028**

Title AMPK related of circRNA expression in C57 male mice myocardium

Organism [Mus musculus](#)

Experiment type Non-coding RNA profiling by high throughput sequencing

Summary To identify adenosine monophosphate-activated protein kinase (AMPK)-related circular RNA (circRNA) that was differentially expressed in the mouse myocardial ischaemia-reperfusion injury (MIRI) model, and map the AMPK-related circRNA network to provide novel insights for the prevention and treatment of MIRI.

NCBI > GEO > **Accession Display** [?](#) Contact: [redacted] [?](#) | [My submissions](#) [?](#) | [Sign Out](#) [?](#)

The following secure token has been created to allow review of record GSE278398 while it remains in private status: [redacted]. A copy of this information has been mailed to your e-mail address.

Scope:  Format:  Amount:  GEO accession:

**Series GSE278398**  [Reviewer access](#)

Status **Private until Sep 01, 2028**

Title AMPK related of circRNA expression in C57 male mice myocardium

Organism [Mus musculus](#)

Experiment type Non-coding RNA profiling by high throughput sequencing

# 总结

- 文件不完整会email告知，重新传，直到全部OK
- 全部传好后分配GSE123456编号
- GEO页面信息在user account中修改（实时更新）
- 最长5年不公开数据
- Reviewer，给个token
- 时差、周末不上班
- 上传后会移到SRA中，抹去read name信息
- GEO不检查内容，仅检查形式

# 其他上传数据库

## GSA国家生物信息中心



国家生物信息中心 Data Resources

**GSA**  
Genome Sequence Archive

GSA

e.g., CRA000112; CRX006656; SRX1053555; human

主页 数据提交 数据浏览 信息检索 数据统计 帮助和支持

新闻动态: 整合INSDC SRA库中全部元数据信息, 提供全局检索及热点数据下载服务。

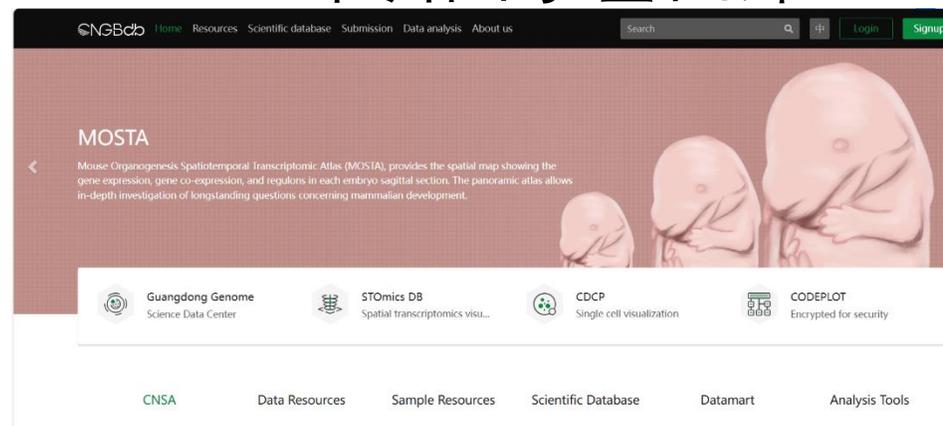
组学原始数据归档库

组学原始数据归档库 (Genome Sequence Archive) 是组学原始数据汇交、存储、管理与共享系统, 是国内首个被国际期刊认可的组学数据发布平台。目前已整合INSDC组学数据, 提供统一检索、数据下载及数据导向服务。

提交 浏览 下载 文档

提交数据到GSA 浏览已经公开的GSA信息 下载GSA数据 查找帮助信息和说明文档

## 深圳国家基因库



NGBdb Home Resources Scientific database Submission Data analysis About us

Search

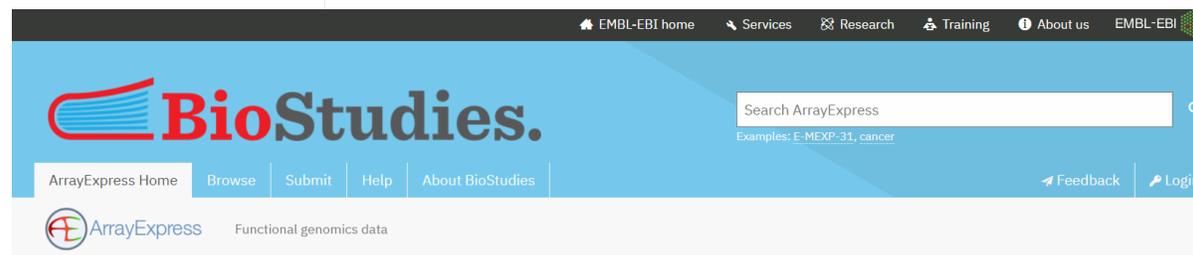
Login Signup

**MOSTA**

Mouse Organogenesis Spatiotemporal Transcriptomic Atlas (MOSTA), provides the spatial map showing the gene expression, gene co-expression, and regulons in each embryo sagittal section. The panoramic atlas allows in-depth investigation of longstanding questions concerning mammalian development.

Guangdong Genome Science Data Center STOmics DB Spatial transcriptomics visu... CDCP Single cell visualization CODEPLOT Encrypted for security

CNSA Data Resources Sample Resources Scientific Database Datamart Analysis Tools



EMBL-EBI home Services Research Training About us EMBL-EBI

**BioStudies.**

Search ArrayExpress

Examples: E-MEXP-31, cancer

ArrayExpress Home Browse Submit Help About BioStudies Feedback Login

ArrayExpress Functional genomics data

### BIOSTUDIES / ARRAYEXPRESS

#### ArrayExpress - Functional Genomics Data

The functional genomics data collection (ArrayExpress), stores data from high-throughput functional genomics experiments, and provides data for reuse to the research community. In line with community guidelines, a study typically contains metadata such as detailed sample annotations, protocols, processed data and raw data. Raw sequence reads from high-throughput sequencing studies are brokered to the European Nucleotide Archive (ENA), and links are provided to download the sequence reads from ENA. Data can be submitted to the ArrayExpress collection through its dedicated submission tool, Annotare. For more information about submissions, see our [submission guide](#).

 Browse ArrayExpress

 Submit an Experiment

## 欧洲EMBL ArrayExpress