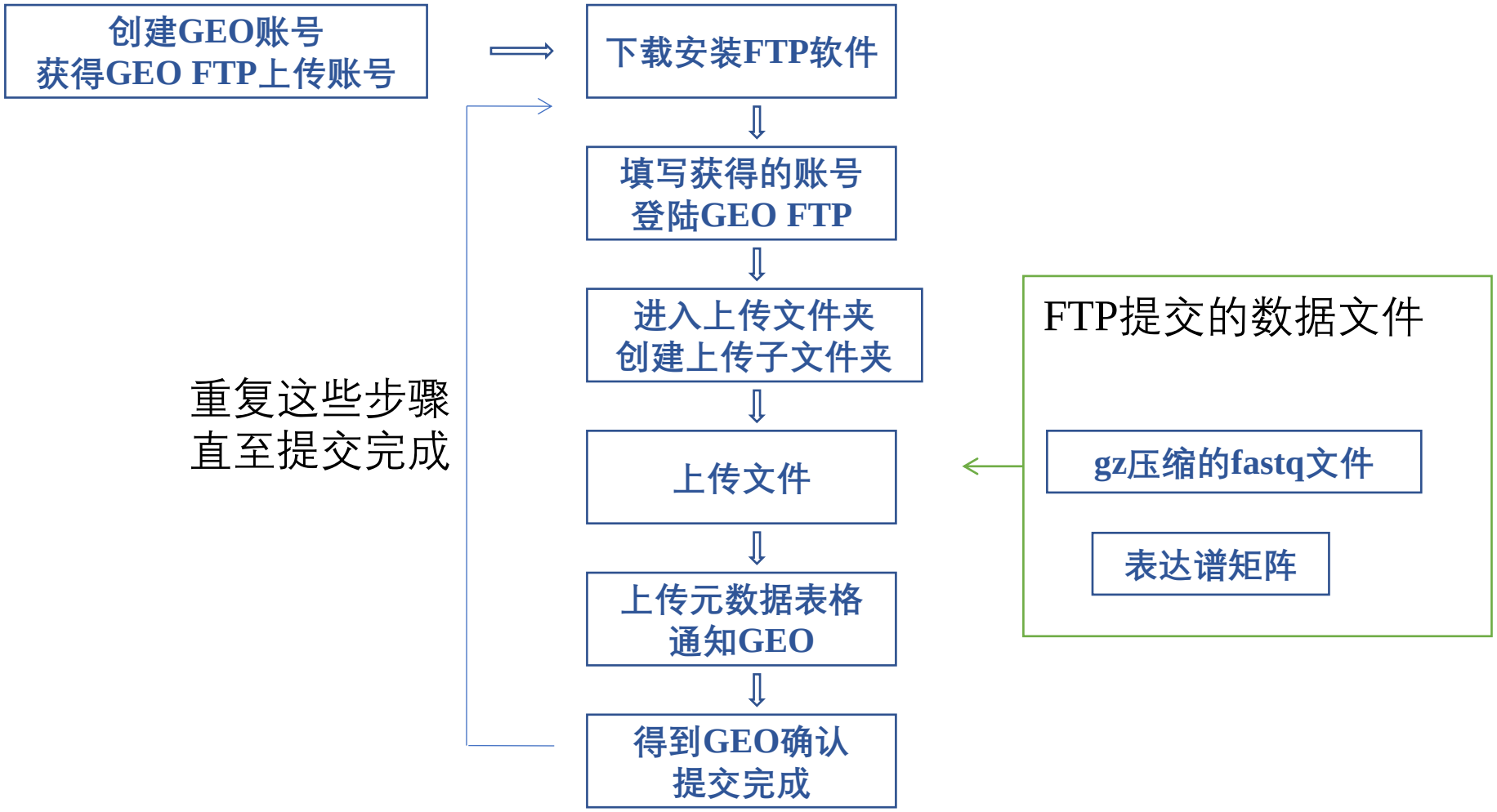
RNA-seq原始数据
上传到GEO数据库详细步骤

陈明杰
202411

GEO数据提交流程

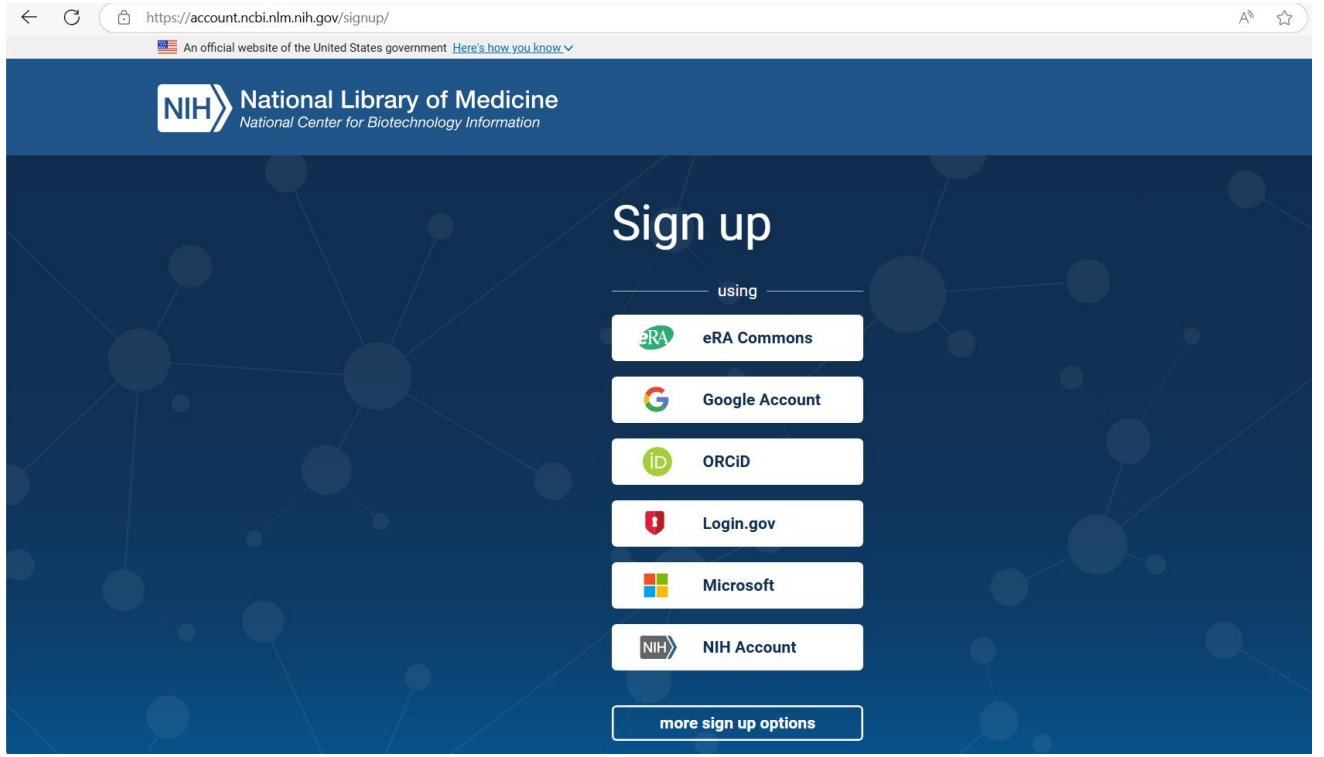

创建GEO账号
获得GEO FTP上传账号
下载安装FTP软件
填写获得的账号
登陆GEO FTP
进入上传文件夹
创建上传子文件夹
上传文件
上传元数据表格
通知GEO
得到GEO确认
提交完成
重复这些步骤
直至提交完成
FTP提交的数据文件
gz压缩的fastq文件
表达谱矩阵

提交前的准备

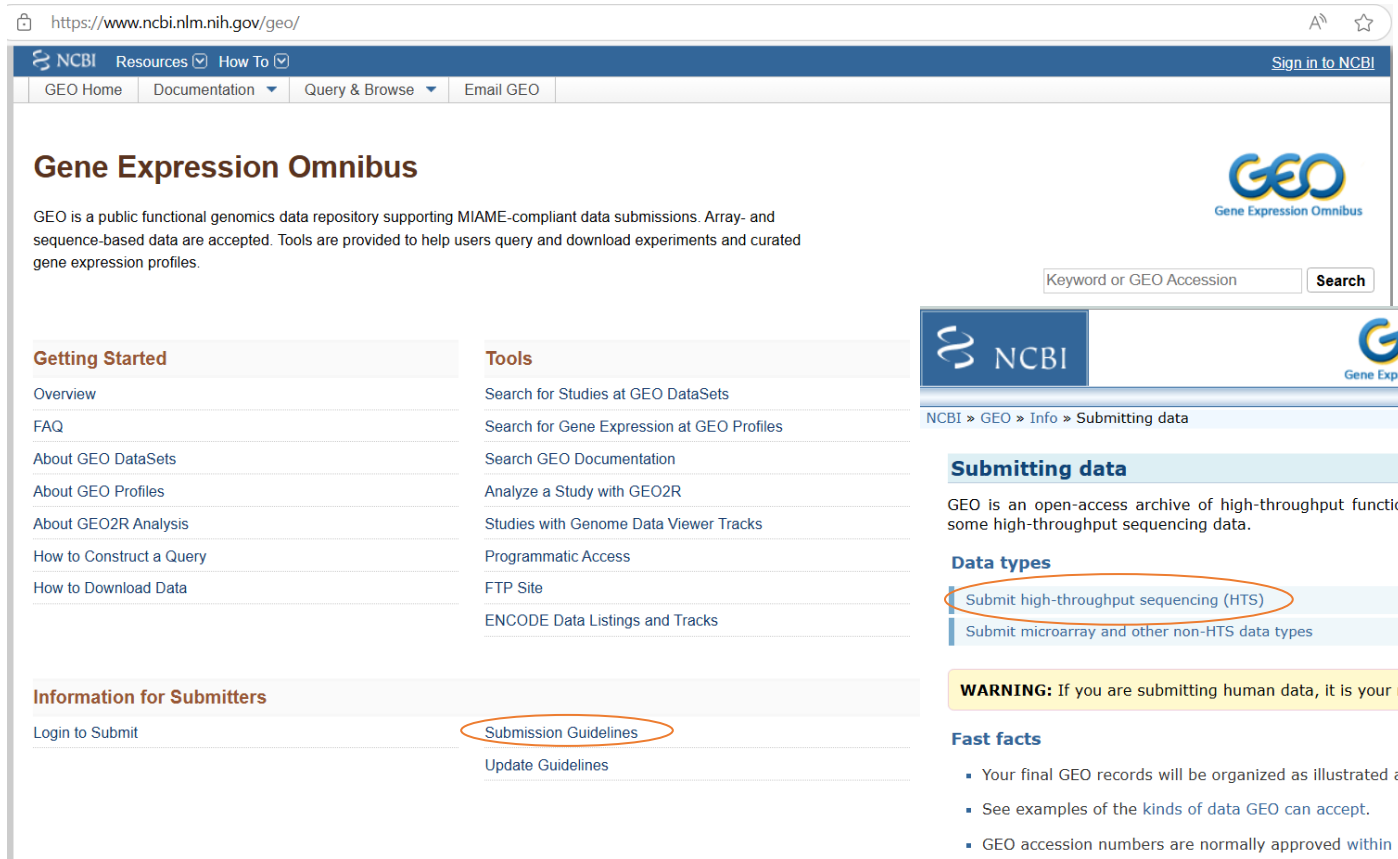
- 1，创建用户账号 https://www.ncbi.nlm.nih.gov/account/
- 2， FTP上传软件
  - 推荐winscp： https://winscp.net/eng/index.php
- 3， 三类文件
  - Raw data：gz压缩的FASTQ原始文件
  - Metadata：元数据表格（下载并填写）
  - Processed data：表达谱数据（count矩阵，FPKM矩阵或者TPM矩阵等）

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



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.

NCBI

Gene Expression Omnibus

GEO Publications | FAQ | MIAME | Email GEO

NCBI » GEO » Info » Submitting high-throughput sequence data to GEO Login

Submitting high-throughput sequence data to GEO

- Submission instructions
  - 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
other aspects of functional genomics using methods such as RNA-seq (
seq, HiC-seq, methyl-seq, etc. We process all components of your st
processed data files, and we submit the raw data files to the Sequence R

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**

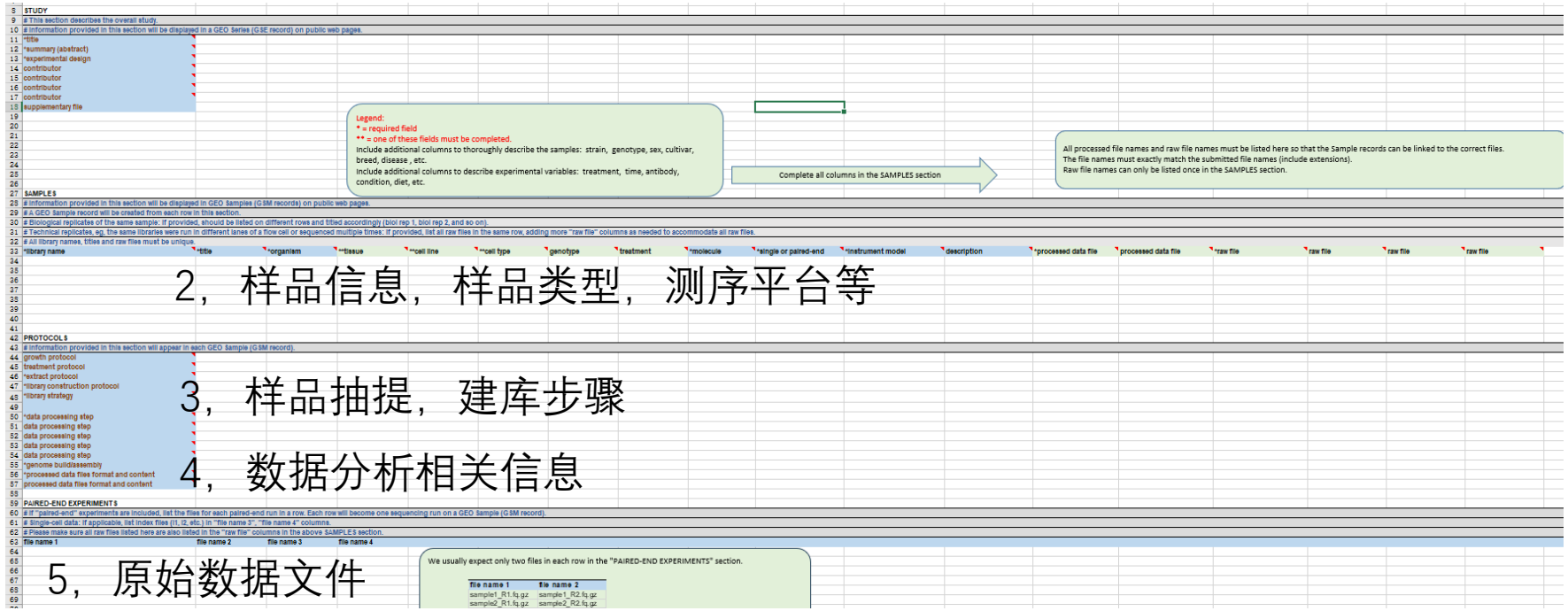
  Raw data are a required part of GEO submissions. The raw data files should be the original files containing reads and quality scores, as generated by the sequencing instrument. Edited files may not be processed correctly by SRA.

  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

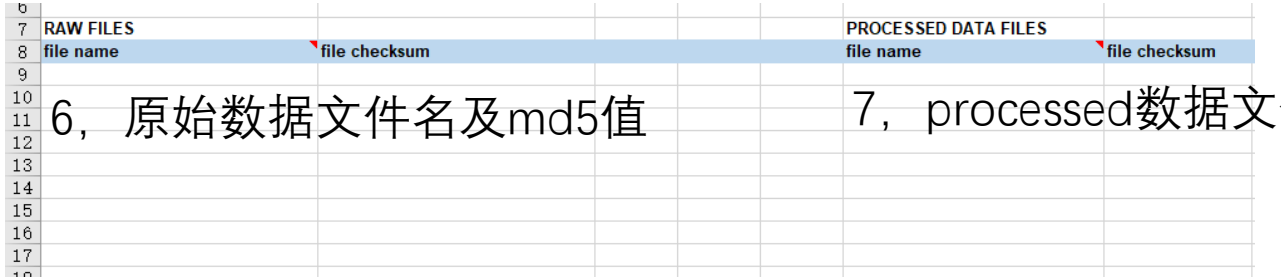


2，样品信息，样品类型，测序平台等

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



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）

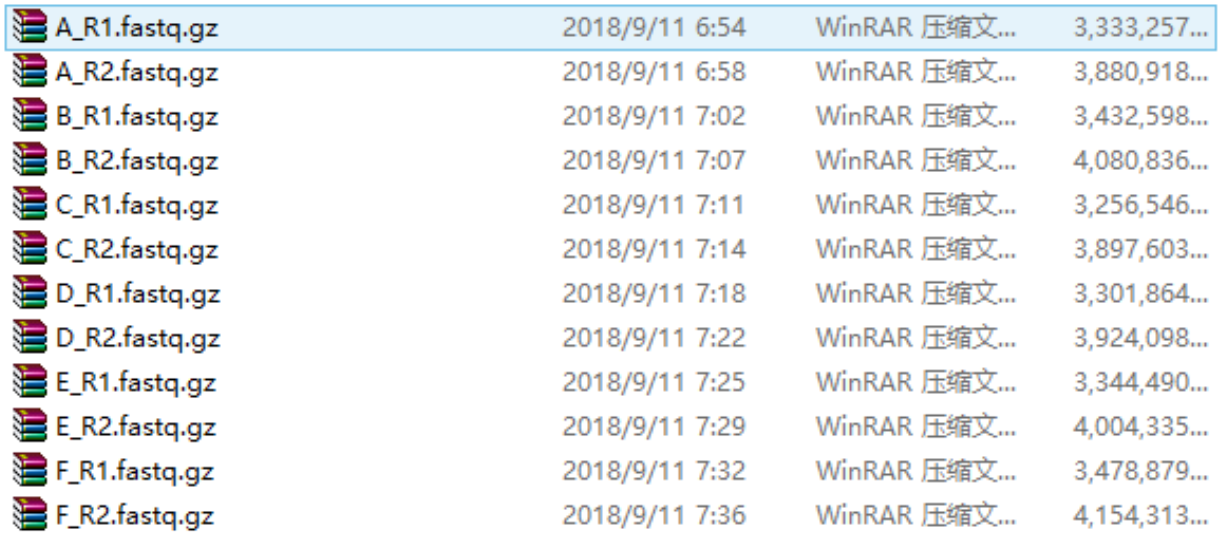
mRNA Expression Profiling.xlsx 2020/2/20 8:45 Microsoft Excel ... 12,947 KB

	A	B	C	D	E	F	G
1	**Symbol**	**A1**	**A2**	**A3**	**B1**	**B2**	**B3**
2	Ndrg4	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



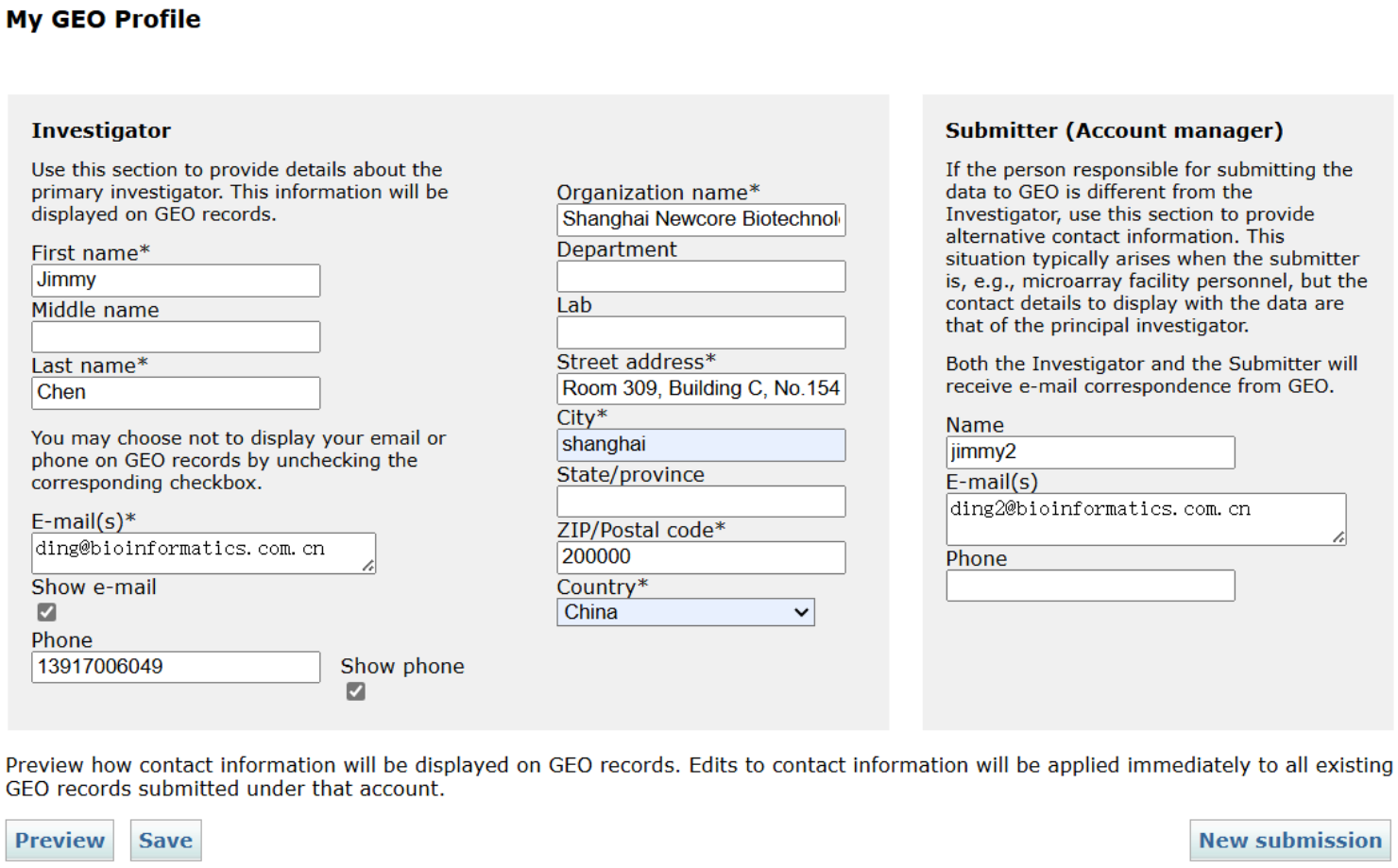
A_R1.fastq.gz	2018/9/11 6:54	WinRAR 压缩文...	3,333,257...
A_R2.fastq.gz	2018/9/11 6:58	WinRAR 压缩文...	3,880,918...
B_R1.fastq.gz	2018/9/11 7:02	WinRAR 压缩文...	3,432,598...
B_R2.fastq.gz	2018/9/11 7:07	WinRAR 压缩文...	4,080,836...
C_R1.fastq.gz	2018/9/11 7:11	WinRAR 压缩文...	3,256,546...
C_R2.fastq.gz	2018/9/11 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

https://mp.weixin.qq.com/s/teEbS3Nbcd4Y-XDQFHgTRw

用户信息表（GEO页面上的信息）



此信息在页面上实时更新

NCBI

Gene Expression Omnibus

GEO File Transfer Protocol (FTP)

Step 1. Your personalized upload space is: **uploads/x[redacted]0** 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

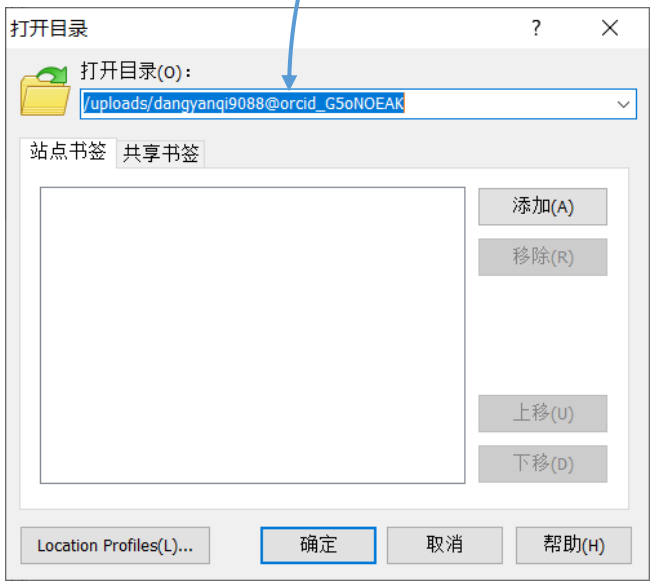
e. For LINUX/UNIX users, we recommend transferring files with 'ncftp' or 'lftp', but you can also use 'ftp', 'sftp', or 'ncftpput'. Please see below for detailed examples.

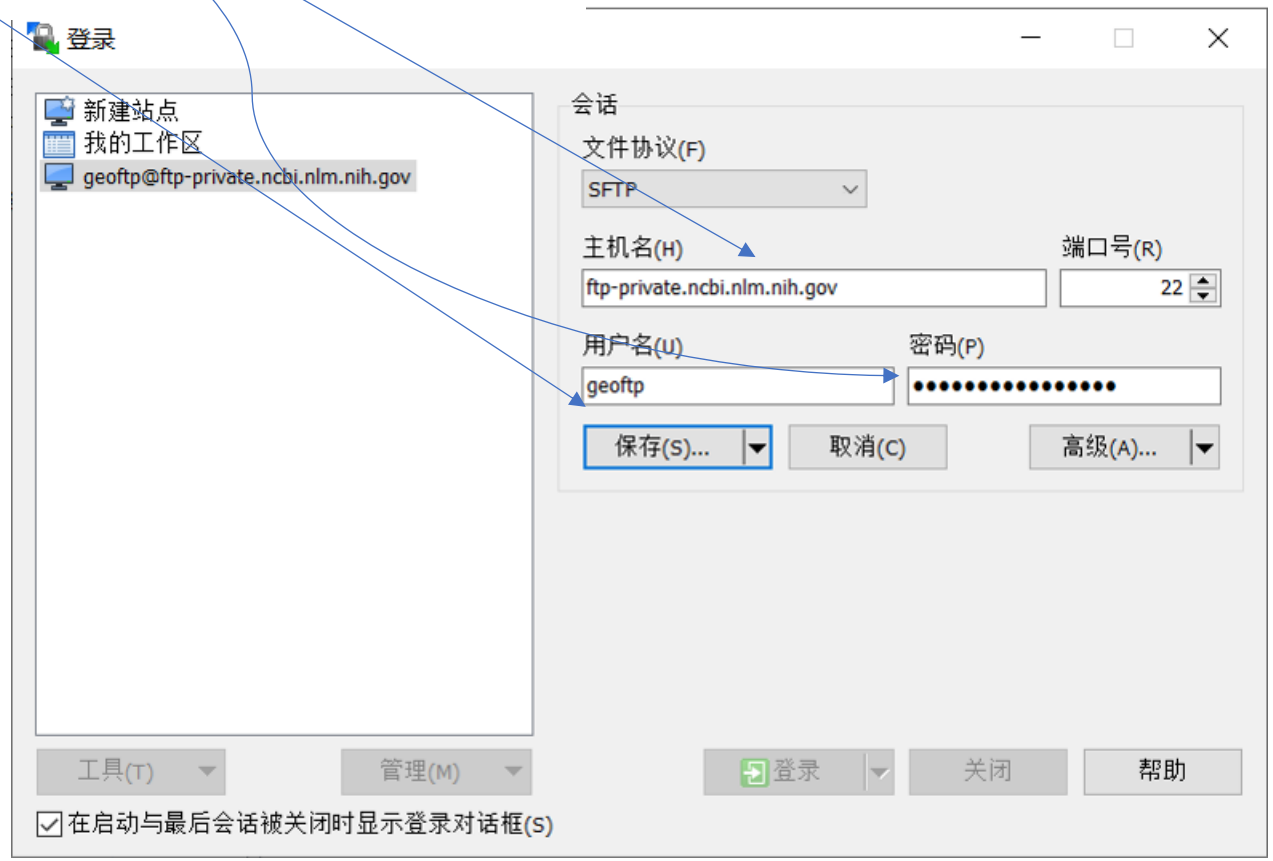
f. Our FTP server credentials are:

host address	ftp-private.ncbi.nlm.nih.gov
username	geoftp
password	[redacted]

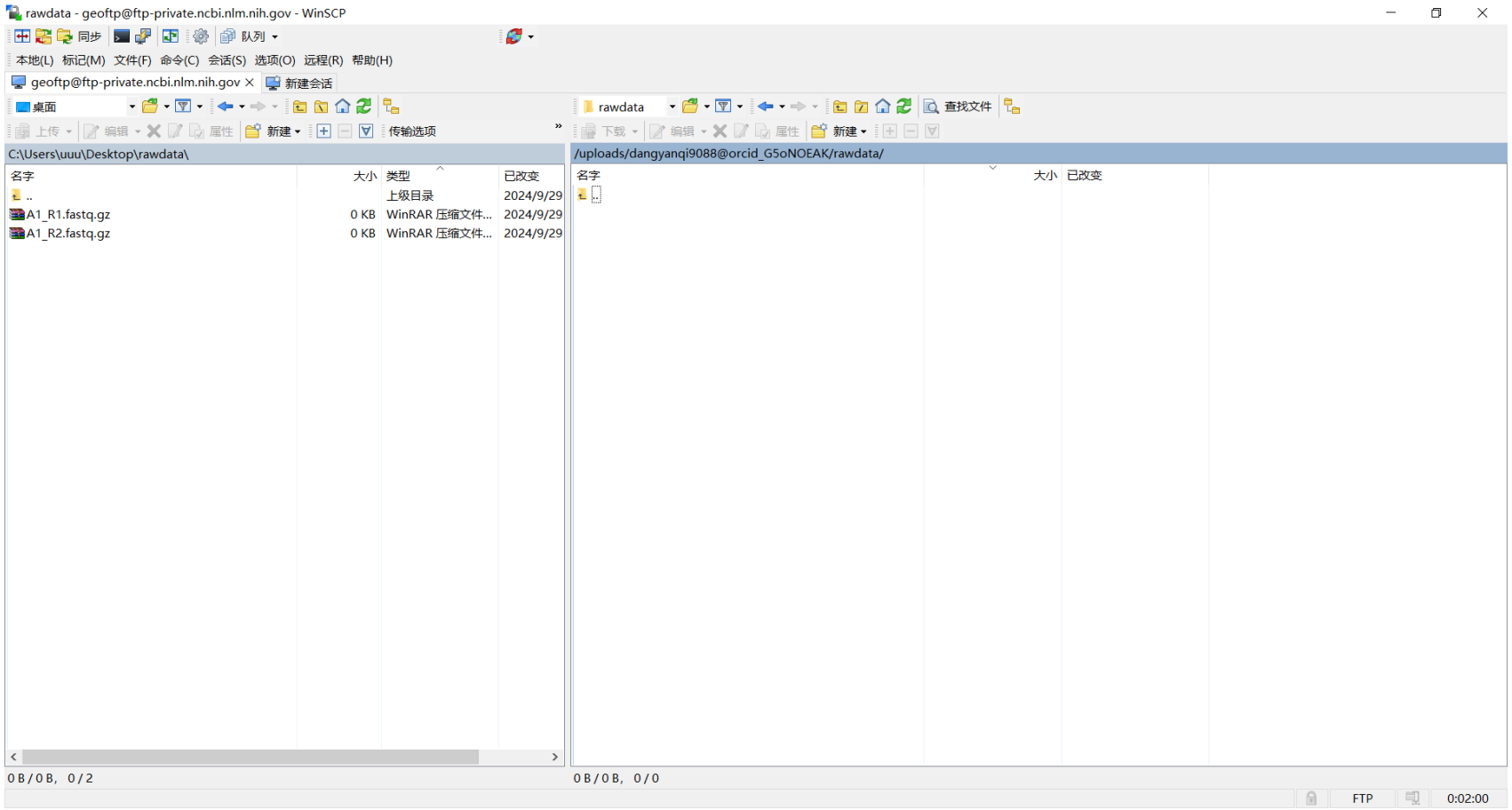
Do not share these log-in credentials. Do not include these log-in credentials on a public page.
These credentials are changed regularly, as per our security policies.

g. After connecting, you **must navigate** to your personalized upload space:
uploads/[redacted]





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 [redacted] account. Messages from GEO regarding your submission will be sent to the following email address(es): **ding@bioinformatics.com.cn,** [redacted]. 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)

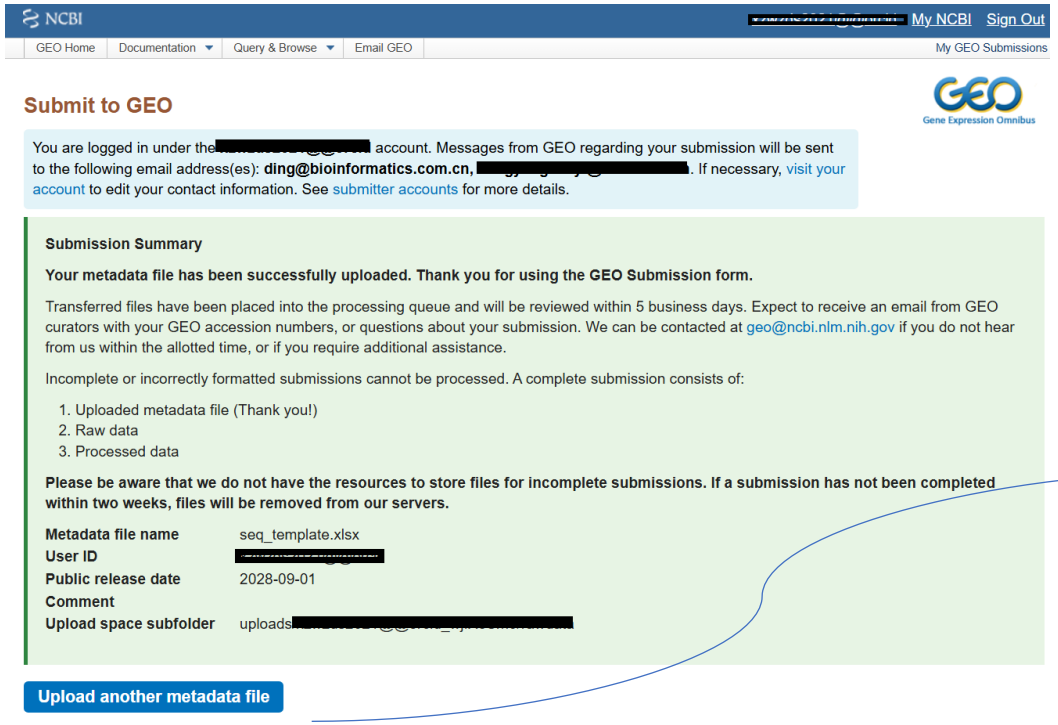
2028-09-01 后续可以通过邮件让工作人员帮助修改

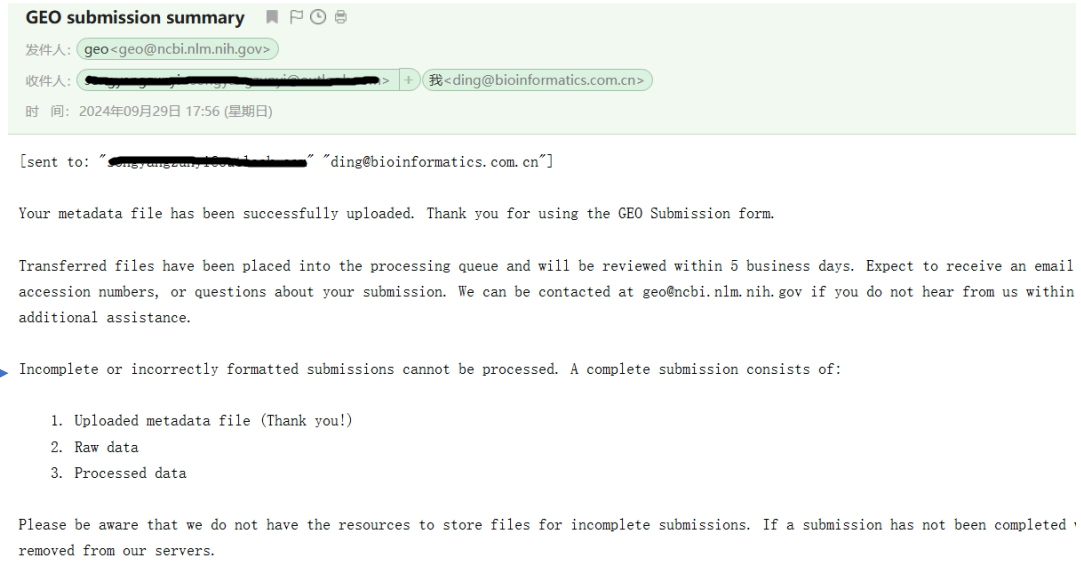
Comment to GEO staff (optional)

备注信息

Submit

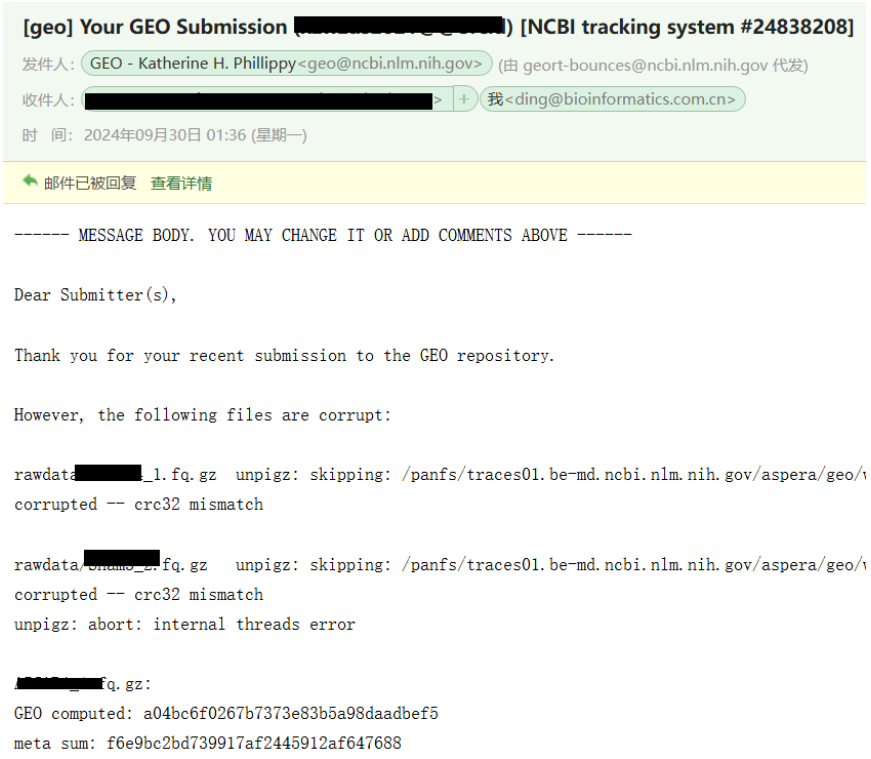
GEO回复的邮件





几分钟后会收到邮件

GEO审核（一般5个工作日）



[geo] Your GEO Submission [redacted]) [NCBI tracking system #24838208]

发件人：GEO - Katherine H. Phillippy<geo@ncbi.nlm.nih.gov> (由 geort-bounces@ncbi.nlm.nih.gov 代发)

收件人：[redacted] 我<ding@bioinformatics.com.cn>

时　间：2024年09月30日 01:36 (星期一)

邮件已被回复　查看详情

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

Dear Submitter(s),

Thank you for your recent submission to the GEO repository.

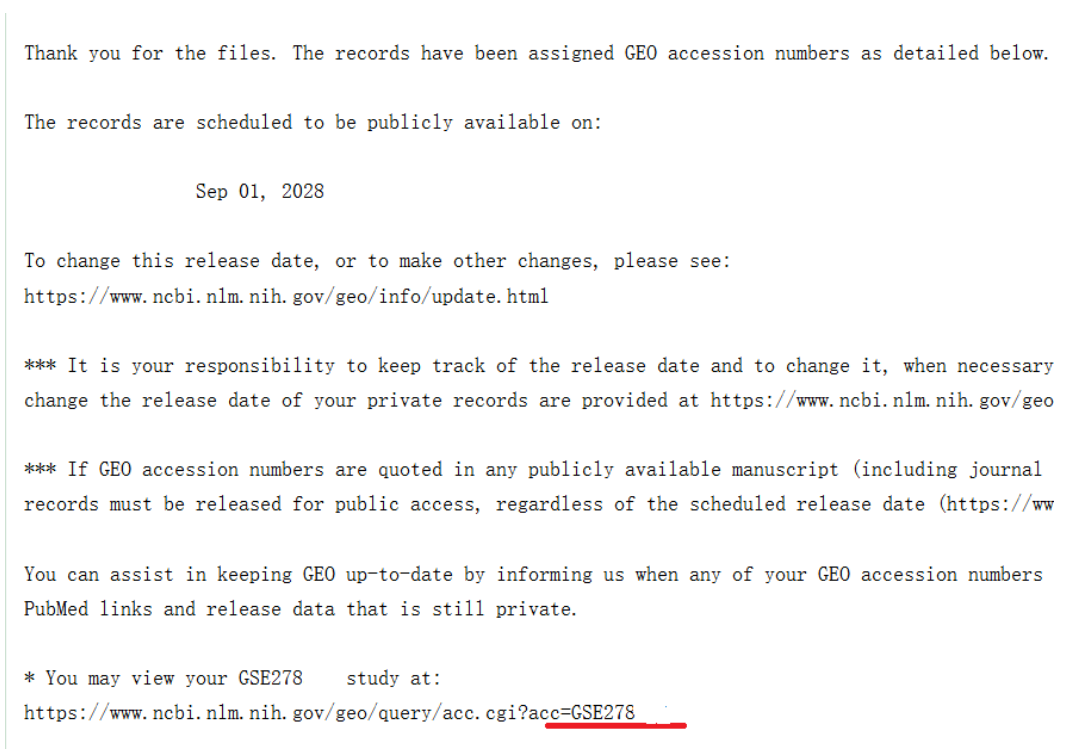
However, the following files are corrupt:

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

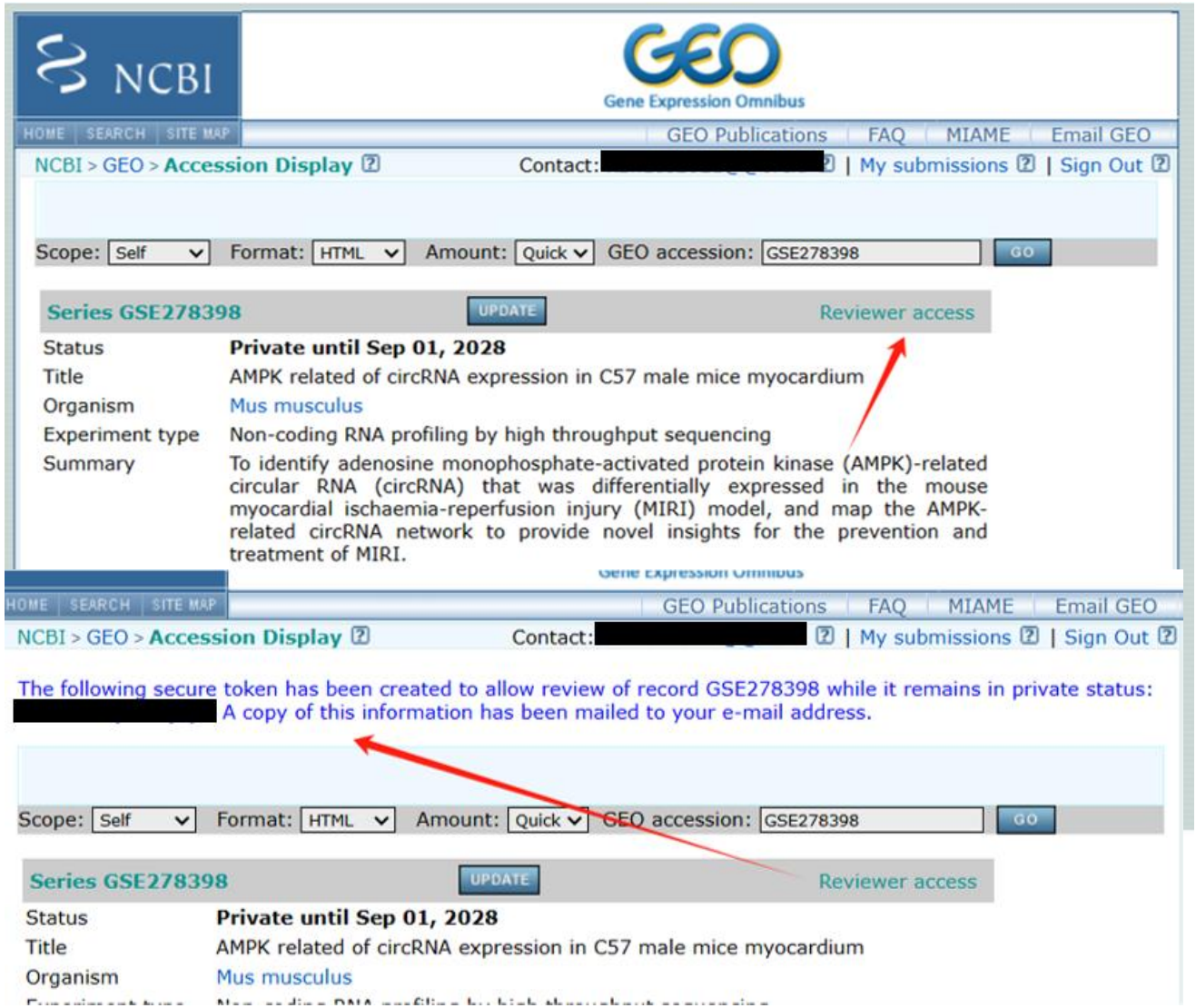
*** If GEO accession numbers are quoted in any publicly available manuscript (including journal
records must be released for public access, regardless of the scheduled release date (https://ww

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



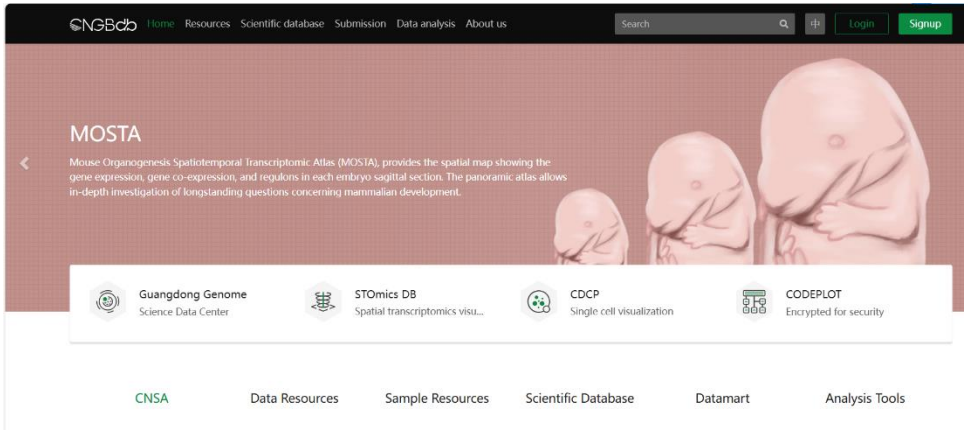
总结

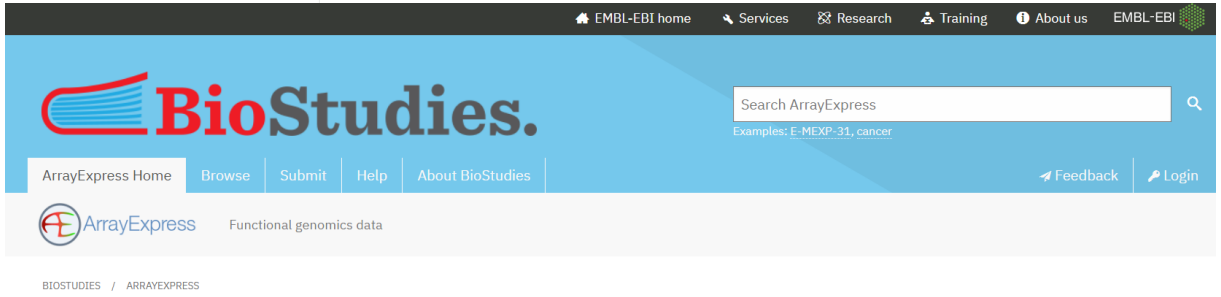
- 文件不完整会email告知，重新传，直到全部OK
- 全部传好后分配GSE123456编号
- GEO页面信息在user account中修改（实时更新）
- 最长5年不公开数据
- Reviewer，给个token
- 时差、周末不上班
- 上传后会移到SRA中，抹去read name信息
- GEO不检查内容，仅检查形式

其他上传数据库

GSA国家生物信息中心

深圳国家基因库





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