

## 芯片数据如何上传GEO数据库？

原创：小Q编 伯豪生物 2016-08-12



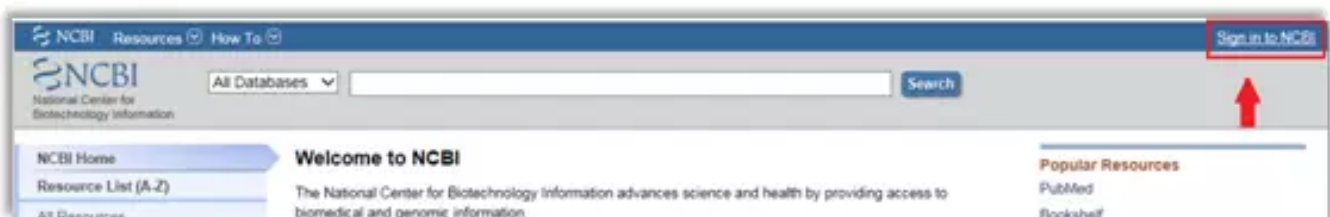
图片来源：GEO

首先对GEO数据库做一个介绍，它来源于美国国立生物技术信息中心（National Center for Biotechnology Information），即我们所熟知的NCBI是由美国国立卫生研究院（NIH）于1988年创办。是其下面的一个分支数据库。

Gene Expression Omnibus (GEO) 是一个储存高通量功能基因组学数据的数据库，这些高通量功能基因组学数据来自芯片和新一代的测序仪得到的试验数据。GEO除了收录基因表达数据之外还收录其它数据，例如基因组拷贝数变异数据、基因组-蛋白相互作用数据以及基因组甲基化数据等。该数据库既接受原始数据，也接受经过处理的数据，不过这些数据都要符合“有关芯片试验的最小信息（minimum information about a microarray experiment, MIAME）”标准。

该数据库能存储好几种格式的数据，包括web格式、spreadsheets格式、XML格式和纯文本格式。GEO数据库被分为两个部分收录在Entrez中，分别是GEO Profiles数据库（它负责收录一个基因在一次试验中的定量基因表达数据）和GEO DataSets 数据库（收录整个试验的数据）。目前，GEO数据库共收录了由世界各地的实验室提交的超过1871121个样本试验数据，16088个芯片平台记录，71339种实验项目以及3848种研究类型的基因表达谱数据。

不管上传至哪个数据库，第一步都需要在NCBI上注册账号：



NCBI Resources How To

### Sign in to NCBI

Sign in with

Google NIH Login eRA Commons

[See more 3rd party sign in options](#)

OR

### Sign in directly to NCBI

Password

Keep me signed in

Sign In

[Forgot NCBI username or password?](#)

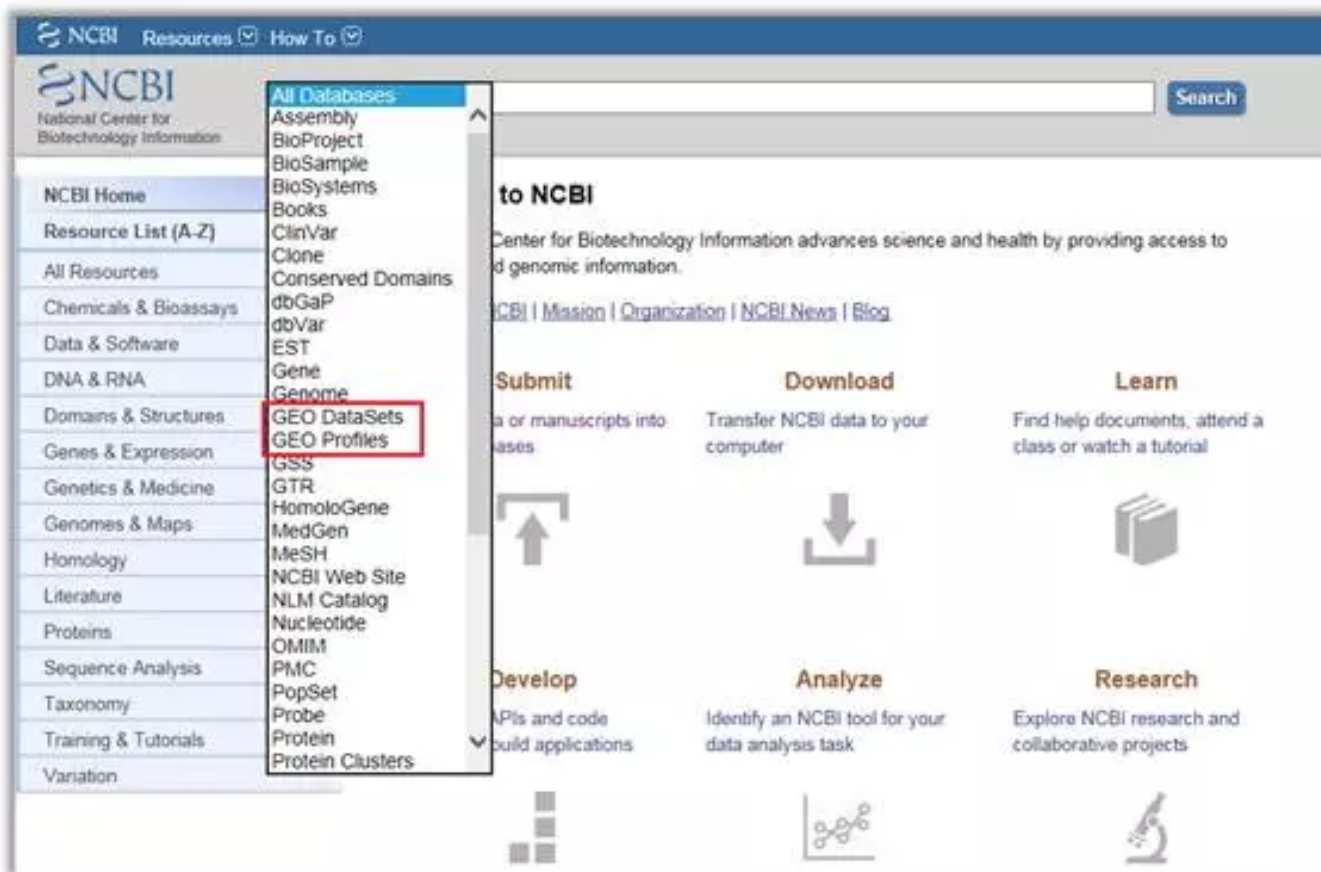
[Register for an NCBI account](#)

已有NCBI账户的，直接在此填写。

没有账户的，在此新建。

注意：新建的话，一定填写您常用的邮箱和基本信息，以免影响上传后的客服及时跟您沟通。

注册成功后回到主页：



下拉框中选择“GEO Datasets”或者“GEO Profiles”，再点击“search”。



点击“submit to GEO”提交数据通道。

如果对上传数据有了一定的了解，或是再次上传的，在准备工作做好的情况下，可以直接点击“GEOarchive”后面的“Submit”（如下图）

- Submission format options
- Basic requirements for submissions
- Fast facts about submitting data

### Submission format options

Deciding which method to use depends on the amount of data you have to submit, the format in which your data currently exist and what applications you are familiar with. Regardless of the deposit method you choose, your final GEO records will look the same and contain equivalent information.

**GEOarchive**  
**Recommended method for most submissions.** Submit  
 Quickly describe your study using **Excel spreadsheet templates**.  
[Complete instructions >](#) 相关注释

### SOFT and MINiML formats

Good option if your data and metadata are already in a database, and you can generate and export data in **SOFT plain text format** or **MINiML XML format**. Submit  
[Complete SOFT instructions >](#)  
[Complete MINiML instructions >](#)

All deposit options described above can be used for any data type. However, the majority of GEO submitters use common commercial arrays (Affymetrix, Agilent, Illumina or Nimblegen) each of which has unique properties and file types. It is recommended that submitters who use the 4 common commercial arrays see these recommendations:

这边还可以了解一下上传需要的数据类型，点击“Complete instructions”，结果如下图：

<b>Metadata spreadsheet</b>	'Metadata' refers to descriptive information and protocols for the overall experiment and individual Samples. This information is supplied by completing all fields of the appropriate metadata spreadsheet template which can be downloaded from the <a href="#">GEOarchive templates and examples</a> section below.
<b>Matrix table</b>	The matrix table is a spreadsheet containing the final, normalized values that are comparable across rows and Samples, and preferably processed as described in any accompanying manuscript. It is possible to include additional data columns in the table, for example, Affymetrix Detection calls and P-values, or background or flag columns. See the Affymetrix template for an example.
<b>Raw data files</b>	In addition to the normalized data provided in the Matrix table, submitters are required to provide raw data, usually in the form of supplementary raw data files. This facilitates the unambiguous interpretation of the data and potential verification of the conclusions as described in the MIAME guidelines. Affymetrix submissions must include CEL files. Non-Affymetrix GEOarchive submissions should include the original software-generated scan quantification files, for example, GenePix GPR files. Next-generation sequence submissions must include files containing reads and quality scores.
<b>Platform</b>	If your experiments are performed using a commercial array (e.g., Affymetrix GeneChip) or other array already deposited in GEO, please use the <a href="#">FIND PLATFORM</a> tool to find the GEO accession number (GPLxxxx) for inclusion in the 'platform' column in the <i>SAMPLES</i> section of the metadata spreadsheet. If your array does not already exist in GEO, please include a <i>PLATFORM</i> section in your metadata spreadsheet and include Platform annotation columns in your matrix table. The Platform data must include meaningful, trackable, sequence identifiers (e.g. GenBank/RefSeq accessions, locus tags, clone IDs, oligo sequences, chromosome locations, etc - see the Platform content guidelines for full list). References to in-house databases or top BLAST hits are not sufficient. Platform submission is not necessary for SAGE or next-generation sequence submissions.

对于上传所需要的文件内容，每个标题后面都有详细的注解。

初次上传数据，首先选择您芯片的类型，并点击：

NCBI > GEO > Info > Submitting data User: qianjin | My submissions | Logout

## Submitting data

GEO accepts many categories of high-throughput functional genomic data, including all array-based applications and some high-throughput sequencing data. This page summarizes deposit options and formats.

We aim to make data deposit procedures as straightforward as possible and will provide as much assistance as you require to get your data submitted to GEO. If you have problems or questions about the submission procedures, just e-mail us at [geo@ncbi.nlm.nih.gov](mailto:geo@ncbi.nlm.nih.gov) with a brief description of the type of data you are trying to submit, and one of our curators will quickly get back to you.

- Data types
  - Array submissions
    - General
    - Affymetrix
    - Agilent
    - Nimblegen
    - Illumina
  - RT-PCR submissions
  - High-throughput sequence submissions
  - Traditional SAGE submissions
- Submission format options
- Basic requirements for submissions
- Fast facts about submitting data

**芯片**

**除芯片以外的平台**

根据技术类型选择对应的填写模板，就伯豪的表观研究来说，850K选择array\_submissions/illumina; ChIP-seq、MeDIP-seq、WGBS、MC-seq选择High-throughput sequence submissions.

我们以Affymetrix human U133 plus2.0为例：点击“affymetrix”后：

## Recommendations for Affymetrix submissions

This page contains deposit recommendations, instructions and templates specific to Affymetrix arrays. These recommendations simply describe what we believe is the easiest batch deposit method for the majority of Affymetrix researchers, but keep in mind that Affymetrix data may be submitted to GEO using any of the deposit options described on the [Submitting data](#) page. If you need additional guidance for your submission, please e-mail us with a brief description of the type of data you are trying to submit, and one of our curators will quickly get back to you.

The [GEOarchive](#) spreadsheet-based submission method is recommended for Affymetrix deposits. With this submission option, you provide the following components:

1. An Excel metadata worksheet containing descriptive information and protocols for the overall experiment and individual Samples (see templates below).
2. CEL files
3. Processed data, usually produced by software (e.g., Expression Console, RMA, GTYPE/CNAT, GTGS, Tiling Analysis software). These data may be used in various ways, as shown in examples in templates below your study. For instance, data from a submission is related to a publication based on GC-RMA data. In this case, you should submit the GC-RMA probe set summary data instead of MAS5.0 CHP files.

Navigation: Affymetrix human U133 plus 2.0 array > 1\_实验结果 > 1\_1\_实验数据 > 1\_1\_1\_原始数据

名称	修改日期	类型	大小
CEL	2013/1/14 18:25	文件夹	
CHP	2013/1/14 18:25	文件夹	
report文件	2013/1/14 18:25	文件夹	

阅读上图内容，我们了解到需要上传芯片需要准备的文件类型，对于AFFY芯片，我们需要准备的是一张 metadata表和processed表（即为matrix表，我后面会重点讲），原始数据（cel格式的文件）。

## Affymetrix GEOarchive templates and examples

The following Excel files illustrate the structure of different types of GEOarchive Affymetrix data submissions. Each Excel file consists of several worksheets, including a metadata template, and metadata and matrix table examples. Guidelines for switching between worksheets. Guidelines:

- 3' or Whole Gene Expression Array - C
- 3' or Whole Gene Expression Array - M
- Exon Array - CHP file option
- Exon Array - Matrix table option
- Tiling Array
- SNP Array

每种品牌都有不同类型的芯片，当然根据设计不同，用途不同，需要的metadata表也不同，这点需要确认清楚，我们的举例芯片是3' IVT格式的芯片，则需要下载是“3' or Whole Gene Expression Array - Matrix table option”

下载刚才提到的metadata表和processed表（即为matrix表），他们是同一个工作簿中两个子表。

标题，概述，作者

对样本的描述

对实验的描述

这是metadata表，主要填写跟文章相关的样本信息和实验信息，每个填写项都有备注，有助于您填写。在这个工作簿里还有关于这两个表的案例展示（example）。

在metadata表中第二部分，查找芯片platform号，我们可以回到“GEOarchive”相关注释的界面上，如下图所示：

**Raw data files**

In addition to the normalized data provided in the Matrix table, submitters are required to provide raw data, usually in the form of supplementary raw data files. This facilitates the unambiguous interpretation of the data and potential verification of the conclusions as described in the MIAME guidelines. Affymetrix submissions must include CEL files. Non-Affymetrix GEOarchive submissions should include the original software-generated scan quantification files, for example, GenePix GPR files. Next-generation sequence submissions must include files containing reads and quality scores.

**Platform**

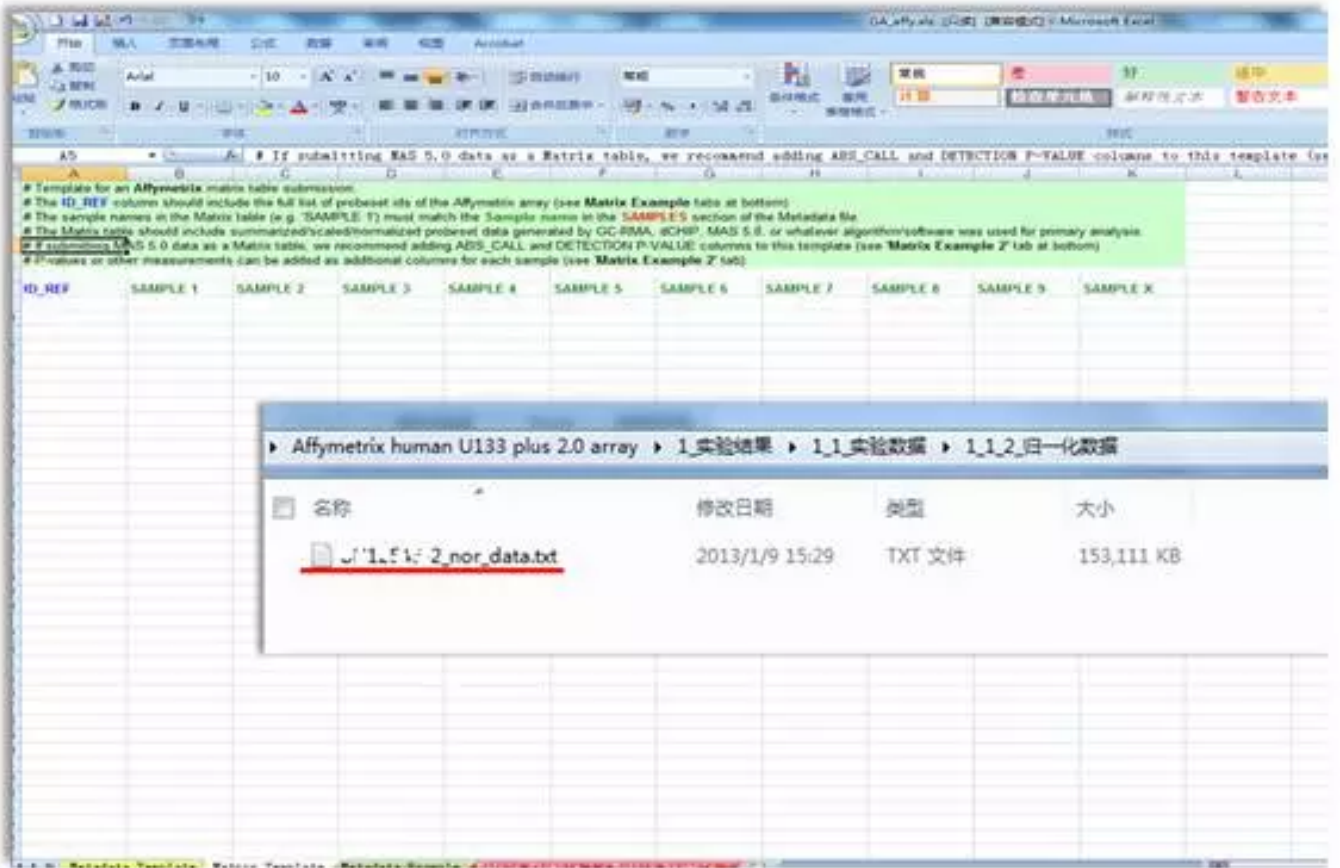
If your experiments are performed using a commercial array (e.g., Affymetrix GeneChip) or other array already deposited in GEO, please use the **FIND PLATFORM** tool to find the GEO accession number (GPLxxxx) for inclusion in the 'platform' column in the *SAMPLES* section of the metadata spreadsheet. If your array does not already exist in GEO, please include a *PLATFORM* section in your metadata spreadsheet and include Platform annotation columns in your matrix table.

The Platform data must include meaningful, trackable, sequence identifiers (e.g. GenBank/RefSeq accessions, locus tags, clone IDs, oligo sequences, chromosome locations, etc - see the Platform content guidelines for full list). References to in-house databases or top BLAST hits are not sufficient. Platform submission is not necessary for SAGE or next-generation sequence submissions.

点击“FIND PLATFORM”

Accession	Title	Technology	Organism(s)	Total rows	Samples	Series	Contact	Submit date
GPL18441	Axiom_GW_Hu_WV Affymetrix Axiom Genome-Wide CG2 1 Array Plate	in situ oligonucleotide	Homo sapiens		8		Affymetrix, Inc.	Mar 20, 2014
GPL10418	Mouse430_2_50K_12.0.0 Affymetrix GeneChip Mouse Genome 430 2.0 Array [Brainarray Version 13]	in situ oligonucleotide	Mus musculus	41,179			GEO archive	Mar 12, 2014
GPL18491	HuGene_1_9-af Affymetrix Human Gene 1.0 ST Array [RNAarray probeSet to-Ensembl mapping]	in situ oligonucleotide	Homo sapiens	25,461	18		Han Cao	Mar 11, 2014
GPL18276	Mouse430_2 Affymetrix Mouse Genome 430 2.0 Array [CDF: Brainarray Mouse430_2_50K_12.0.0]	in situ oligonucleotide	Mus musculus	38,331			Carsten Wotz	Mar 06, 2014
GPL12025	HuE_1_9-af Affymetrix Human Gene 1.0 ST Array [CDF: Brainarray Ver. 14.1.0. HuE130v1_Hu_ENHG]	in situ oligonucleotide	Mus musculus	28,003	30		Hiba Habbib	Mar 05, 2014
GPL18378	ZebGene_1_3-af Affymetrix GeneChip Zebrafish ST Genome Array 1.1, Brainarray version 13 [cdg:ncbi13k_0r_ENTREZG]	in situ oligonucleotide	Danio rerio	23,807			Affymetrix, Inc.	Mar 05, 2014
GPL18388	ChGene_1_9-af Affymetrix Chicken Gene 1.0 ST Array [transcript (gene) version]	in situ oligonucleotide	Gallus gallus	20,425			Affymetrix, Inc.	Mar 04, 2014
GPL18388	ChGene_1_9-af Affymetrix Chicken Gene 1.0 ST Array [probeSet version]	in situ oligonucleotide	Gallus gallus	165,995			Affymetrix, Inc.	Mar 04, 2014
GPL18217	HG-U133A Affymetrix Human Genome U133A Array Custom Brainarray v18 [HG-U133A_Hu_ENTREZG_13.0.0]	in situ oligonucleotide	Homo sapiens	42,080	9		Vijay G Sankaran	Feb 28, 2014
GPL18349	AraGene_1_9-af Arabidopsis Gene 1.0 ST Array [arabid set (arab) version]	in situ oligonucleotide	Arabidopsis thaliana	176,298			Affymetrix, Inc.	Feb 27, 2014
GPL17226	Ht_HG-430_3M Affymetrix Ht HG-430 HT Array Plate [CDF: yfHG430M_Hu_ENTREZG_Brainarray version 13]	in situ oligonucleotide	Mus musculus	37,304	99		Heena Mittal	Feb 24, 2014
GPL18121	HG-U133_Plus_2 Affymetrix Human Genome U133 Plus 2.0 Array [custom CDF]	in situ oligonucleotide	Homo sapiens	26,954	16		Haggie C Chan	Feb 20, 2014
GPL18122	Mouse430_2 Affymetrix Mouse Genome 430 2.0 Array [custom CDF]	in situ oligonucleotide	Mus musculus	28,944	16		Haggie C Chan	Feb 20, 2014
GPL14288	Na1000a21071F Affymetrix Pseudomonas axium Genome Array	in situ oligonucleotide	Pseudomonas aeruginosa	6,828	6		Lutz E Bernsdorf	Feb 13, 2014
GPL18063	HuE_1_9-af Affymetrix Human Gene 1.0 ST Array [mapcore Eszenti mouse version 62]	in situ oligonucleotide	Mus musculus	27,163	46		Wu Blotgen	Feb 11, 2014
GPL17090	RaGene_1_9-af Affymetrix Rat Gene 1.0 ST Array [transcript (gene) version] [CDF: Brainarray ENTREZG Version 14]	in situ oligonucleotide	Rattus norvegicus	18,239			Carsten Wotz	Feb 04, 2014
GPL18278	Axiom_GW_Hu_WV_1 Affymetrix Axiom Genome-Wide CG2 1 Array Plate	in situ oligonucleotide	Homo sapiens				Affymetrix, Inc.	Feb 04, 2014
GPL17190	Affymetrix GeneChip Human Genome U133 Array Set HG-U133A based on a custom CDF [Brainarray version 2.1.0]	in situ oligonucleotide	Homo sapiens	38,852	60		Olivia Bickel	Feb 03, 2014
GPL18029	Mouse430_2_Hu_EntrezG Affymetrix GeneChip Mouse Genome 430 2.0 Array [Brainarray Version 17.1.0]	in situ oligonucleotide	Mus musculus	37,546	12		GEO archive	Feb 01, 2014
GPL18243	MOE430A_Hu_ENTREZG Affymetrix Mouse Expression 430A Array [Brainarray Version 14.0.0]	in situ oligonucleotide	Mus musculus	32,323	8		GEO archive	Feb 01, 2014

尽可能根据您知道的物种，芯片类型，生产厂家等参数进行筛查。有种最简单直接的是通过您归一化数据实际探针行数来找



这是matrix表，将您的归一化数据复制粘贴就行了，请用office 07版以上的版本打开它，对于超过2万多个探针的数据03版行数不够。



最后将原始数据和表格一起打包并命名，以便GEO客服更快的找到您的数据进行审核。

回到之前“GEOarchive”界面：



### Submission format options ▲

Deciding which method to use depends on the amount of data you have to submit, the format in which your data currently exist and what applications you are familiar with. Regardless of the deposit method you choose, your final GEO records will look the same and contain equivalent information.

**GEOarchive**   
**Recommended method for most submissions.**  
 Quickly describe your study using **Excel spreadsheet templates**.  
[Complete instructions >](#)

**SOFT and MINiML formats**   
 Good option if your data and metadata are already in a database, and you can generate and export data in **SOFT plain text format** or **MINiML XML format**.  
[Complete SOFT instructions >](#)  
[Complete MINiML instructions >](#)

All deposit options described above can be used for any data type. However, the majority of GEO submitters use common commercial arrays (Affymetrix, Agilent, Illumina or Nimblegen) each of which has unique properties and file types. It is recommended that submitters who use the 4 common commercial arrays see these recommendations:

点击 “submit”

### Submit to GEO

Use this form to:

- upload files for a new microarray, traditional SAGE, or RT-PCR submission, see [instructions](#)
- upload revisions to existing or in-progress submissions.

Do not use this form to upload next-generation sequence data. Instead see [NGS deposit instructions](#).

**File to upload:**  
  找到您需要上传的压缩文件位置

**Submission kind**  
 new 初次上传就选择 “new” ，再次上传就选择下一项  
 update or revision

**When this submission should be released to the public**  
 Release immediately following curation 您需要立即发布就选第一下项，如果需  
 Release on specified date 要一个特定时间发布就选下一项。

**Comment to GEO staff (optional)**

► If you are getting timeout errors, please transfer your files using these [FTP Instructions](#)

选择压缩文件，上传类型，以及释放日期。

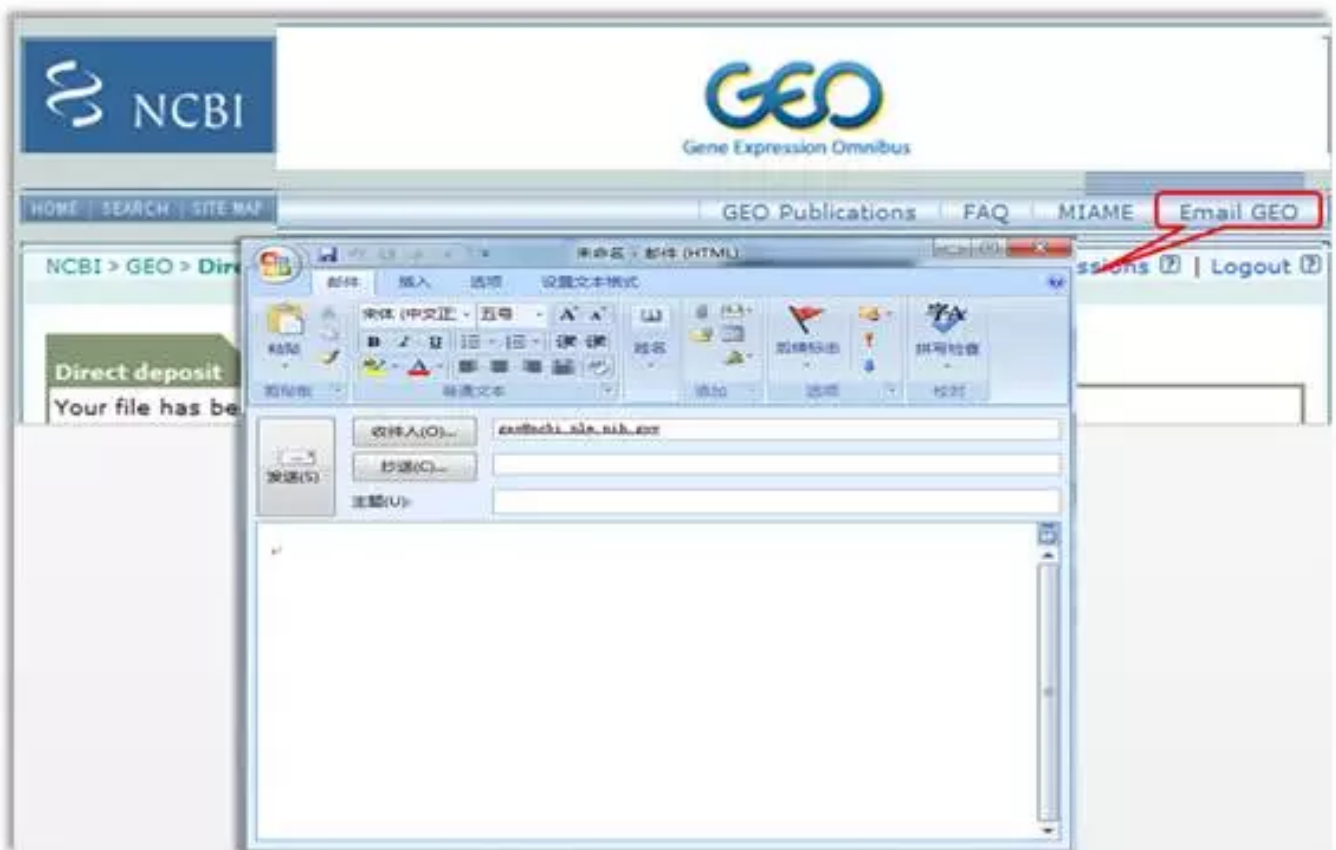
点击 “submit” 后，网页会不停的刷新，根据数据大小，刷新的时间不同。



成功后会出现以上网页，同时您的邮箱中也会接收到正式文件，如果不成功，GEO客服也会以邮件形式告诉您，数据需要补充的内容。

最后还需要给GEO客服写一封信，让他们知道你上传了数据，并让他们尽快发布：

这一步的作用是告诉GEO客服，您需要正式发布的数据。



信件内容，首先说明此次上传的芯片数据类型，存放的压缩包名字，上传的账户，以及压缩包里包含的数据内容。

一般大约2,3个工作日, 经GEO审核, 数据没有问题, 他们会以邮件形式告诉您数据的GSM、GSE。您就可以用于准备发表的文章中。



伯豪生物  
SHBIO

用我们的服务，加速您的研究！



长按二维码关注我们