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。

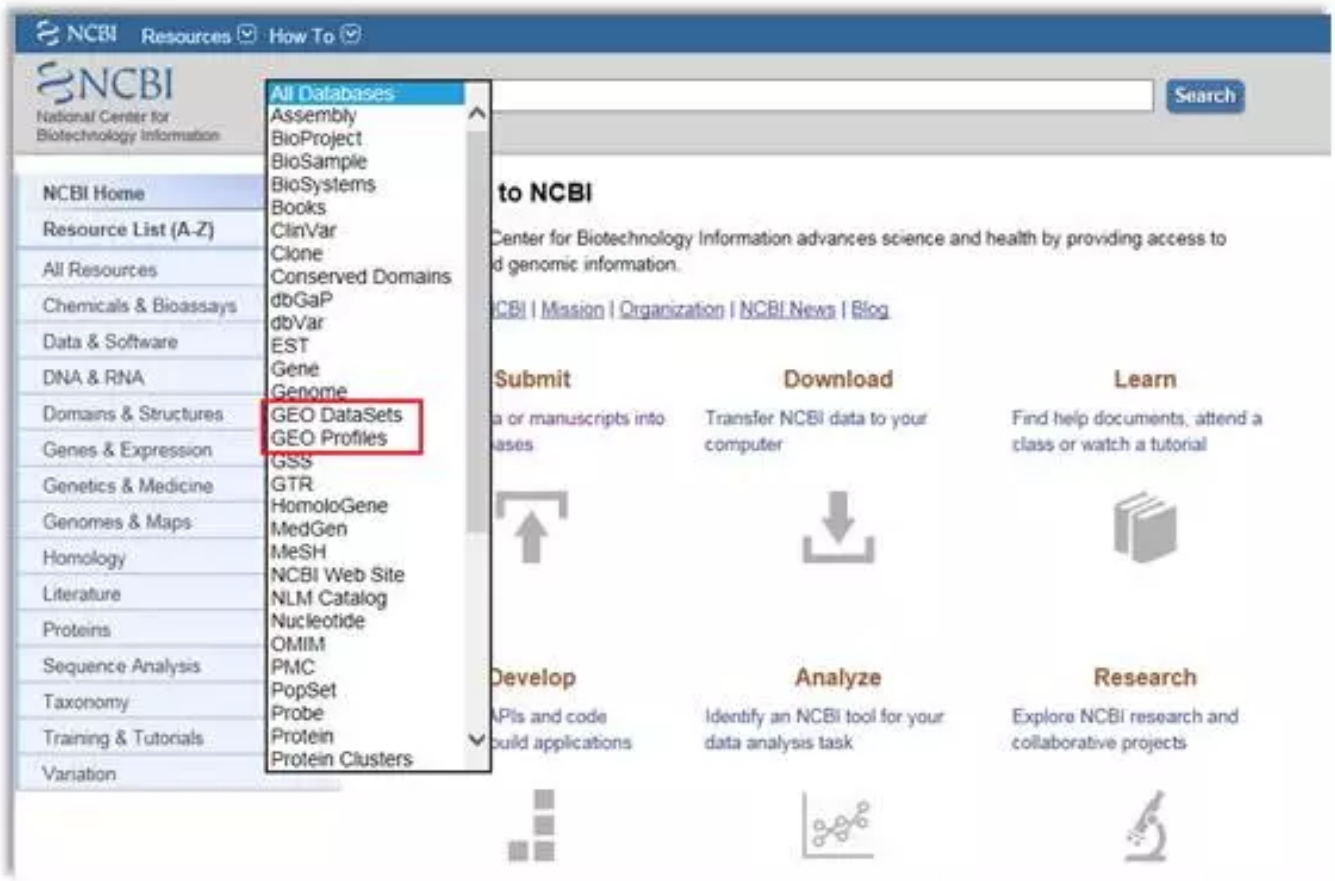
Gene Expression Omnibus（GEO）是一个储存高通量功能基因组学数据的数据库，这些高通量功能基因组学数据来自芯片和新一代的测序仪得到的试验数据。GEO除了收录基因表达数据之外还收录其它数据，例如基因组拷贝数变异数据、基因组-蛋白相互作用数据以及基因组甲基化数据等。该数据库既接受原始数据，也接受经过处理的数据，不过这些数据都要符合“有关芯片试验的最小信息（minimum information about a microarray experiment，MIAME）”标准。

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。

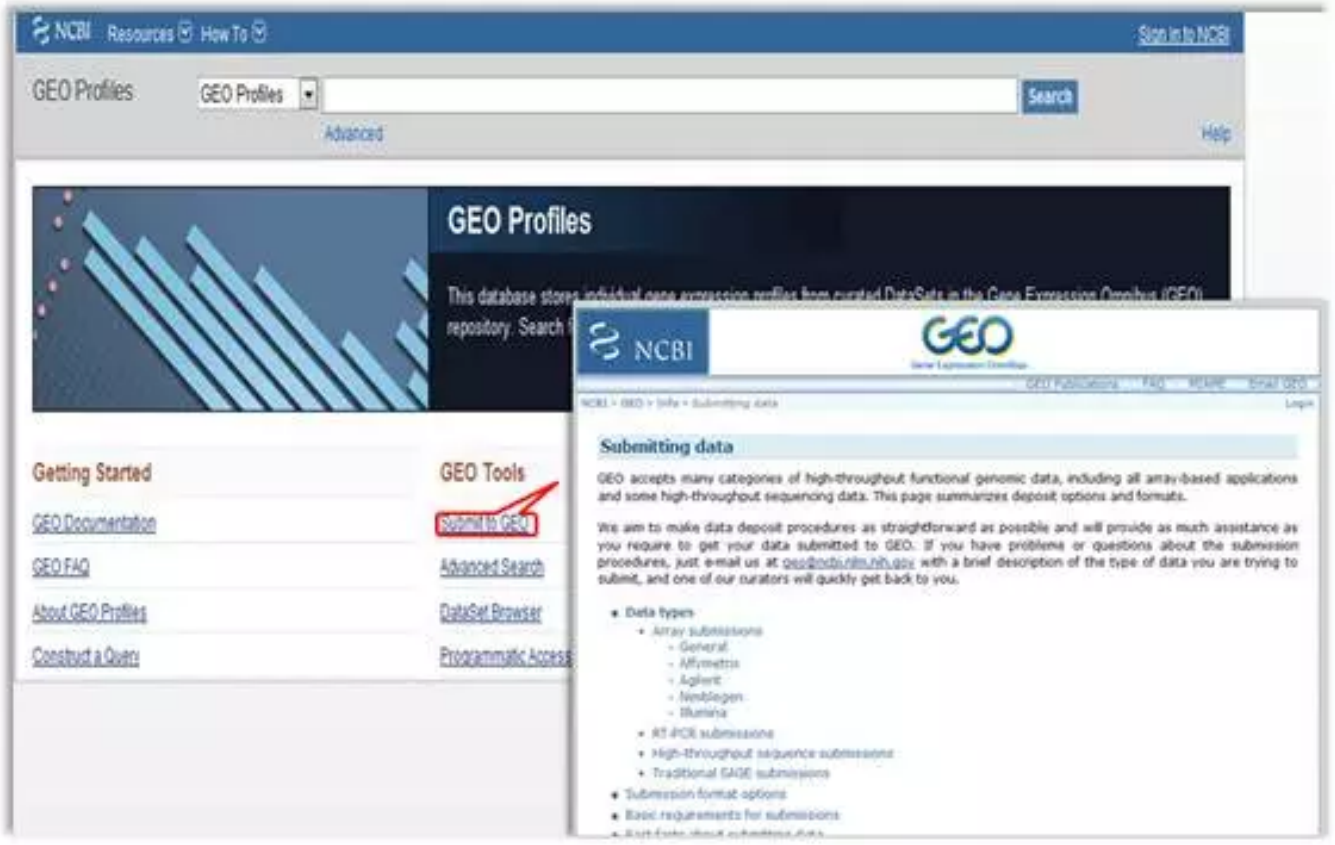
不管上传至哪个数据库，第一步都需要在NCBI上注册账号：

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：

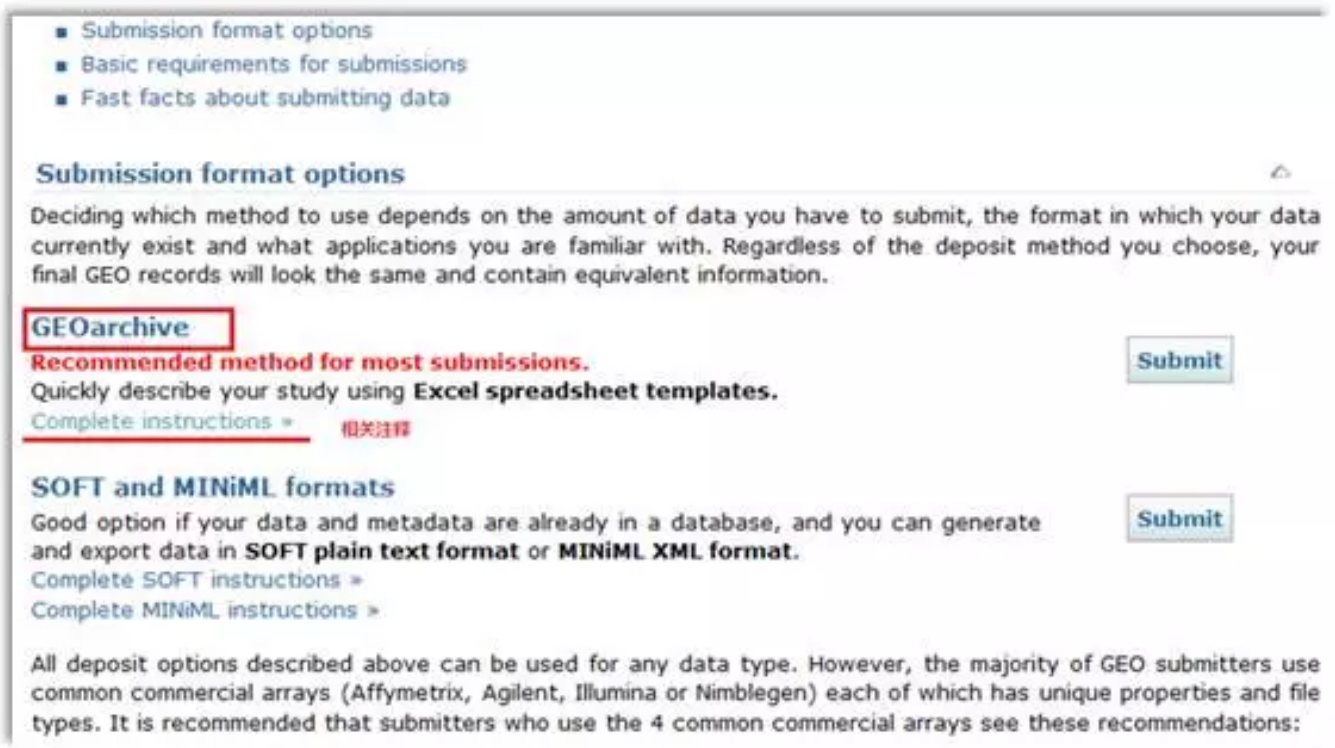


下拉框中选择“GEO Datasets”或者“GEO Profiles”，再点击“search”。

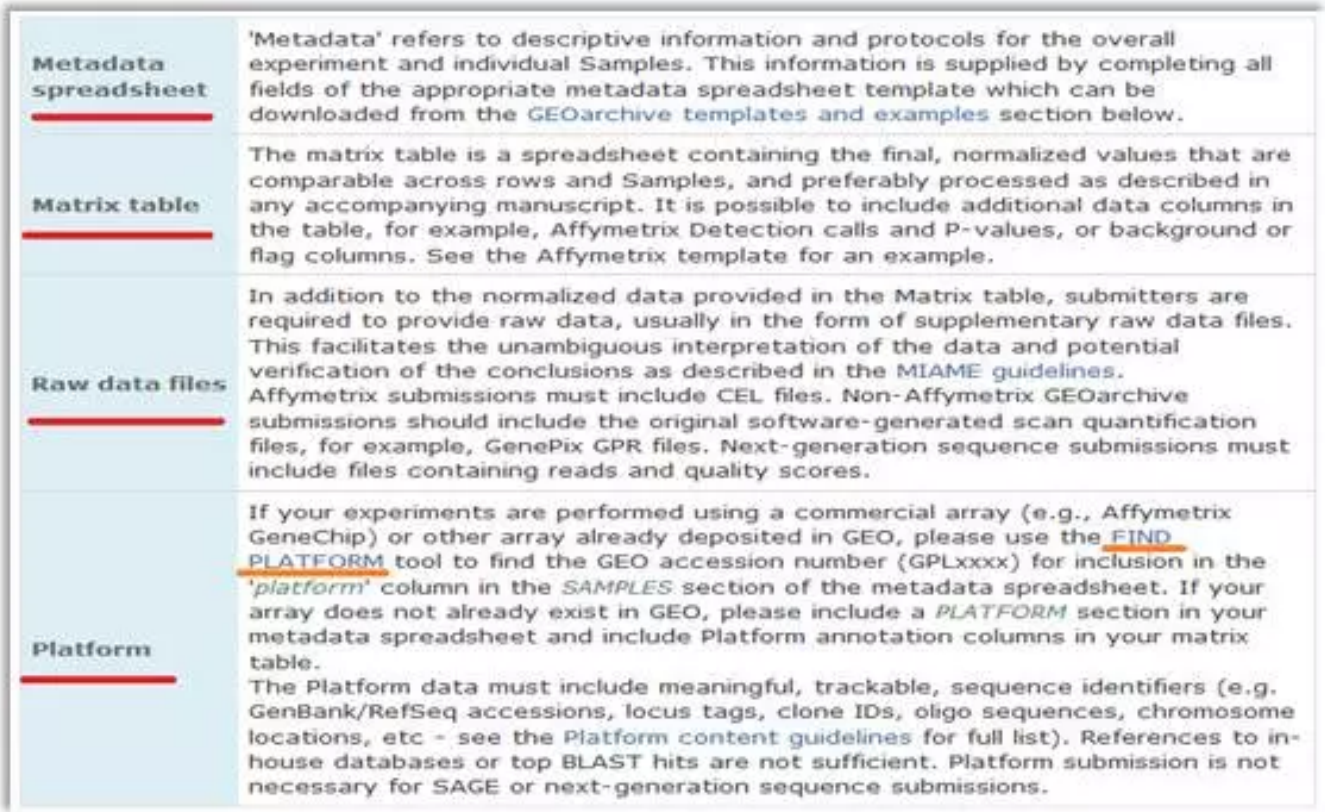


点击“submit to GEO”提交数据通道。

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）

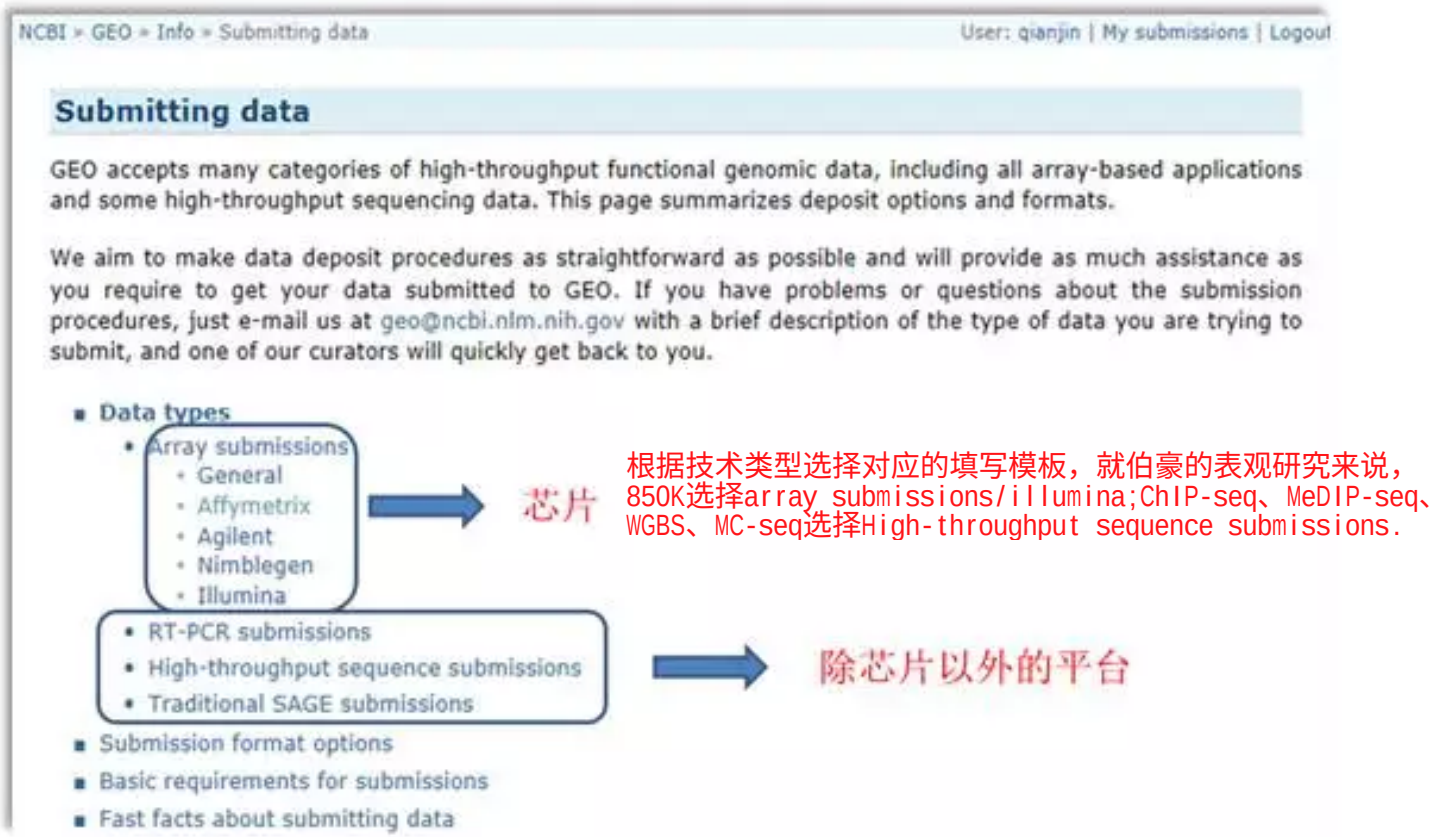


这边还可以了解一下上传需要的数据类型，点击“Complete instructions”，结果如下图：

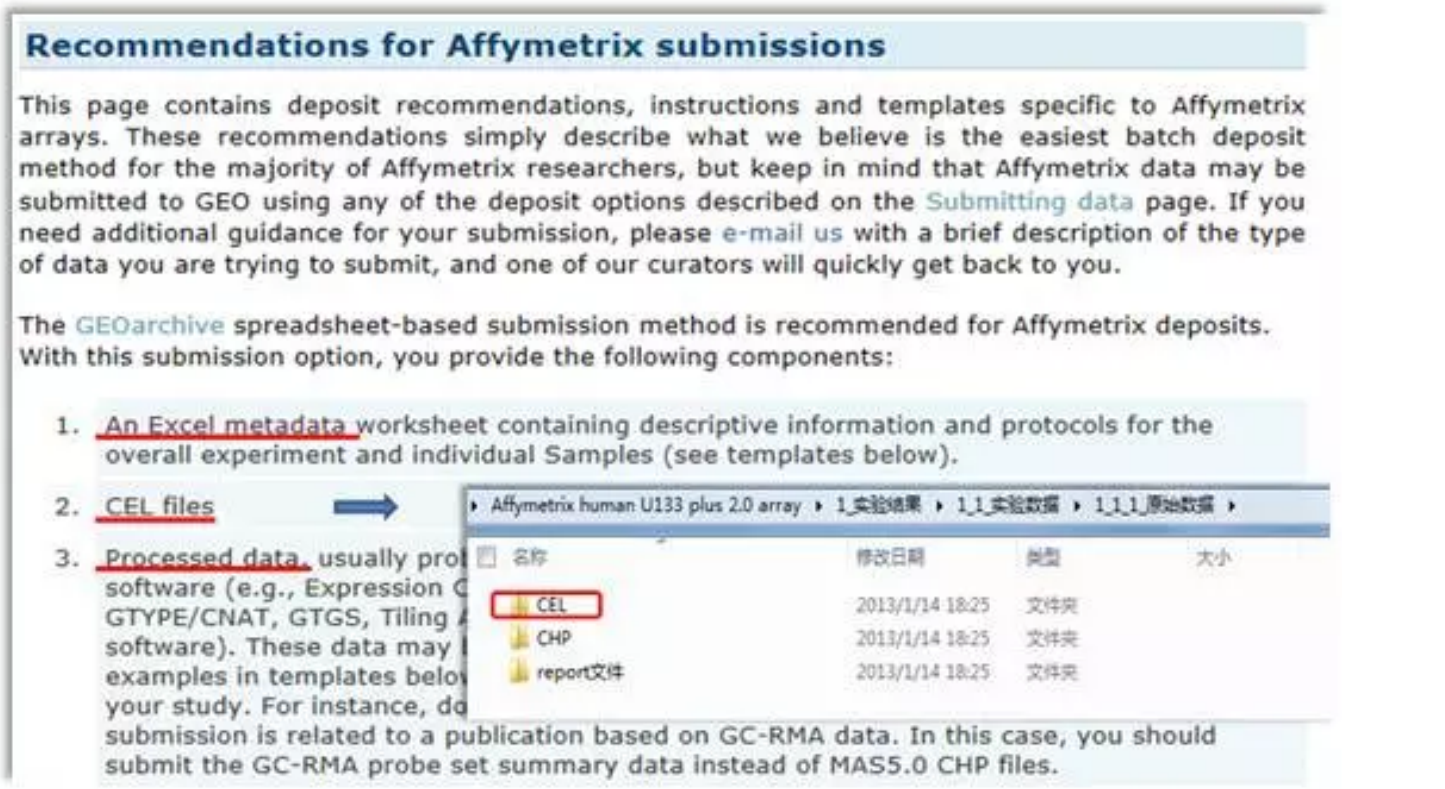


对于上传所需要的文件内容，每个标题后面都有详细的注解。

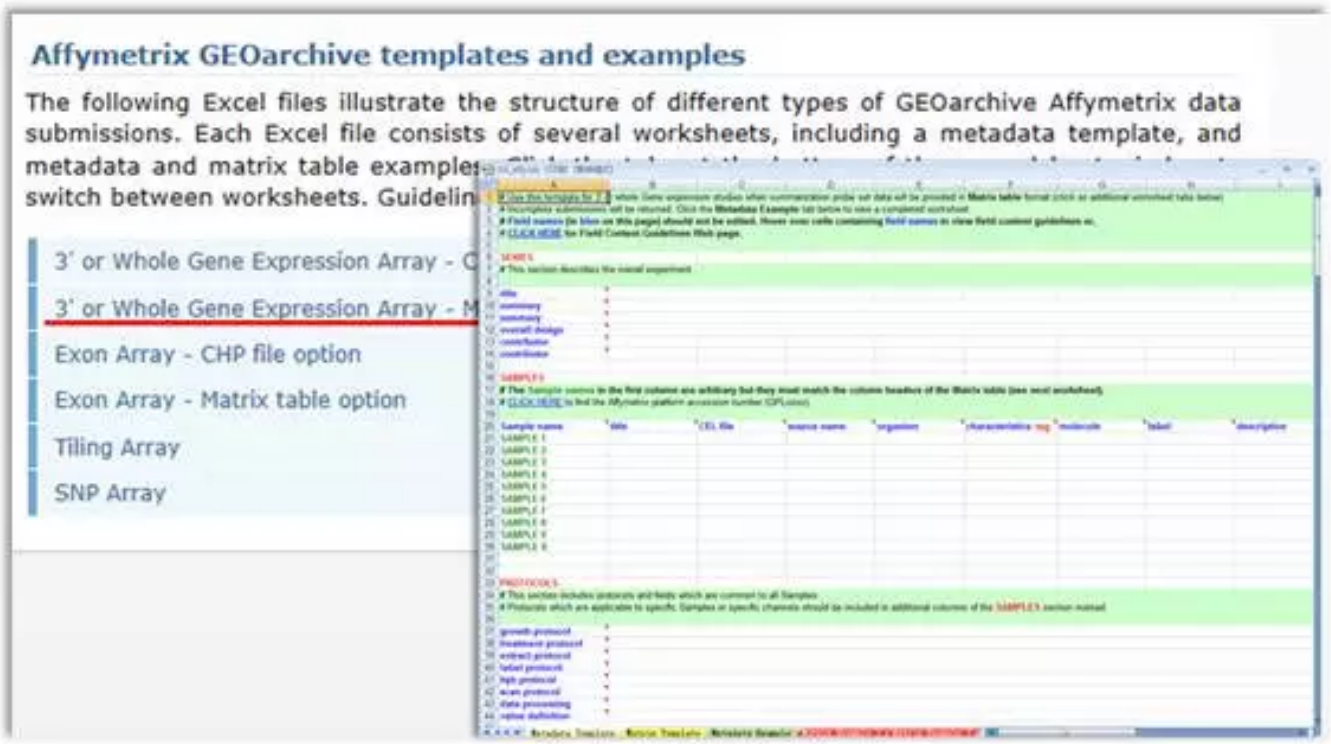
初次上传数据，首先选择您芯片的类型，并点击：



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

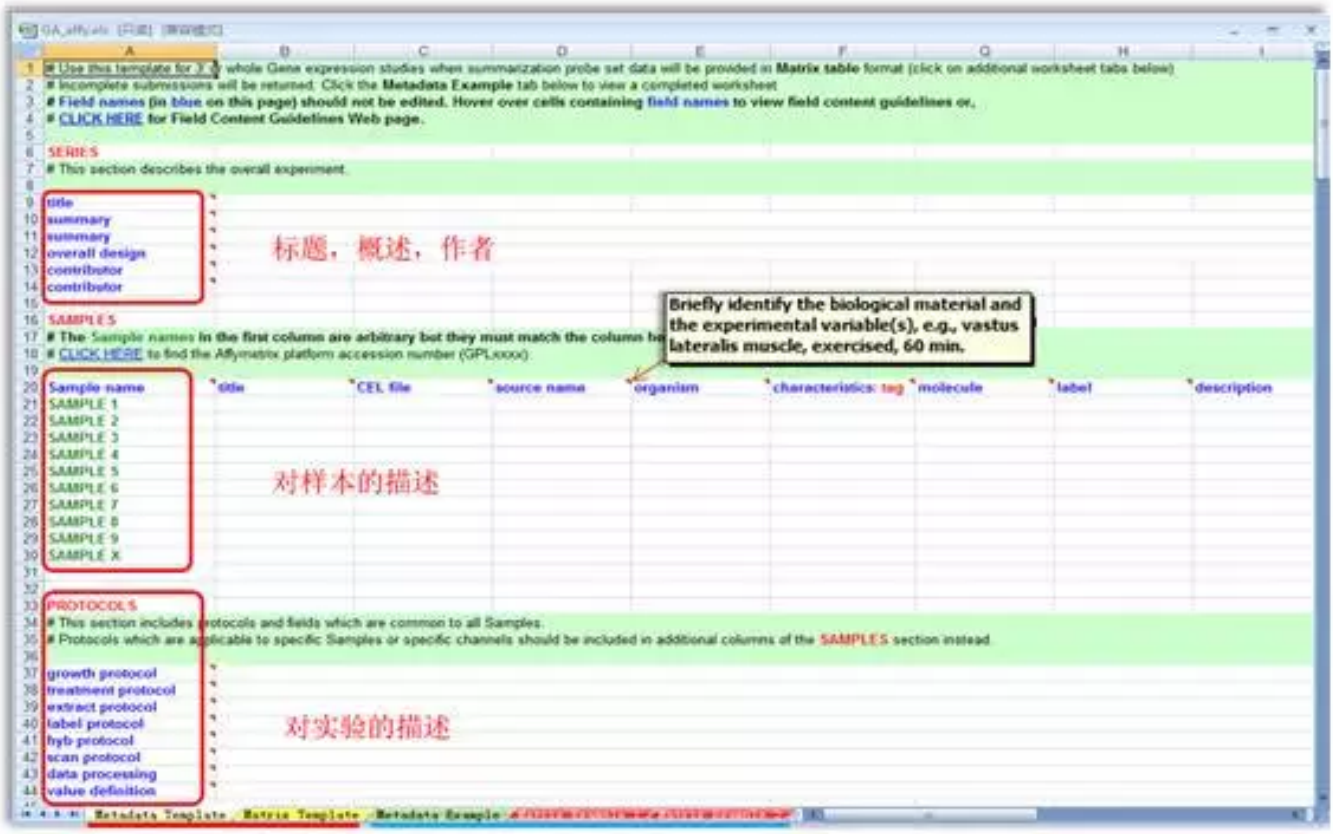


阅读上图内容，我们了解到需要上传芯片需要准备的文件类型，对于AFFY芯片，我们需要准备的是一张metadata表和processed表（即为matrix表，我后面会重点讲），原始数据（cel格式的文件）。



每种品牌都有不同类型的芯片，当然根据设计的不同，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 ExpressionArray-Matrix table option”

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

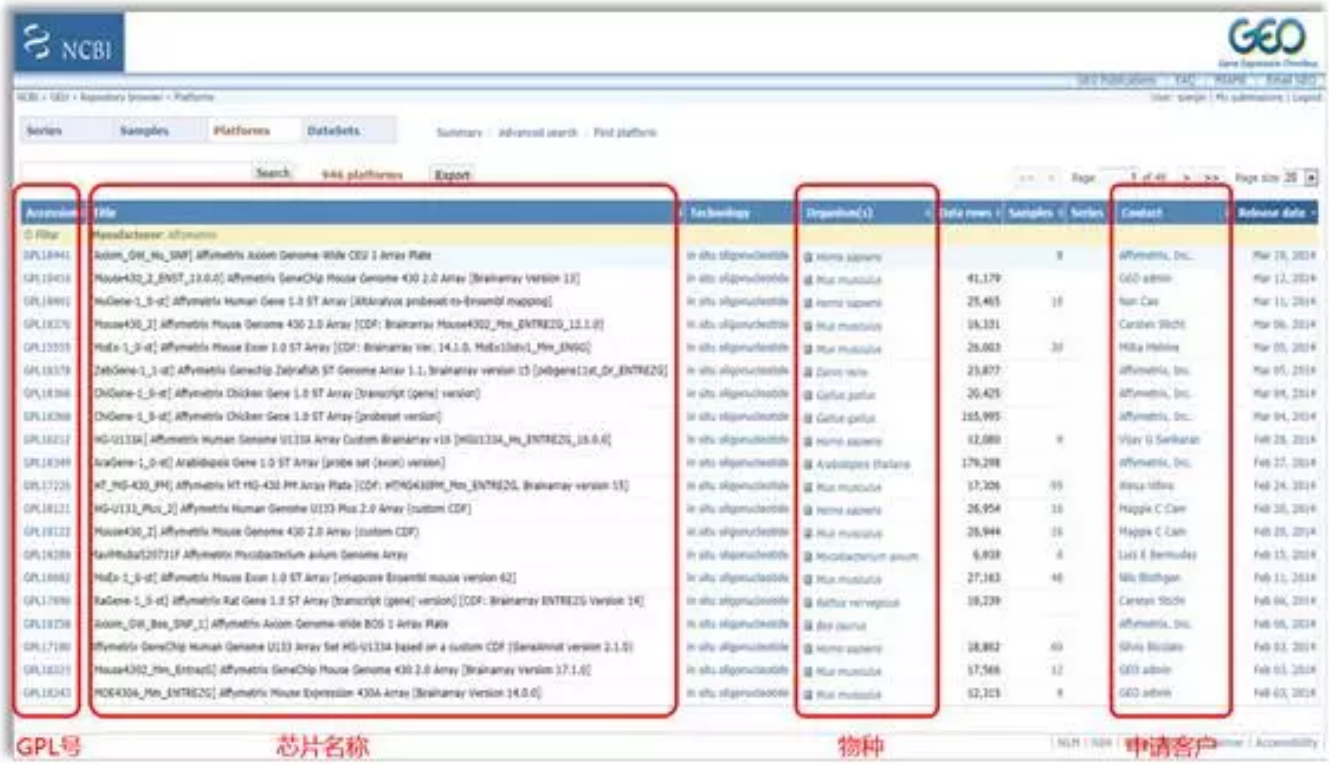


这是metadata表，主要填写跟文章相关的样本信息和实验信息，每个填写项都有备注，有助于您填写。在这个工作薄里还有关于这两个表的案例展示（example）。

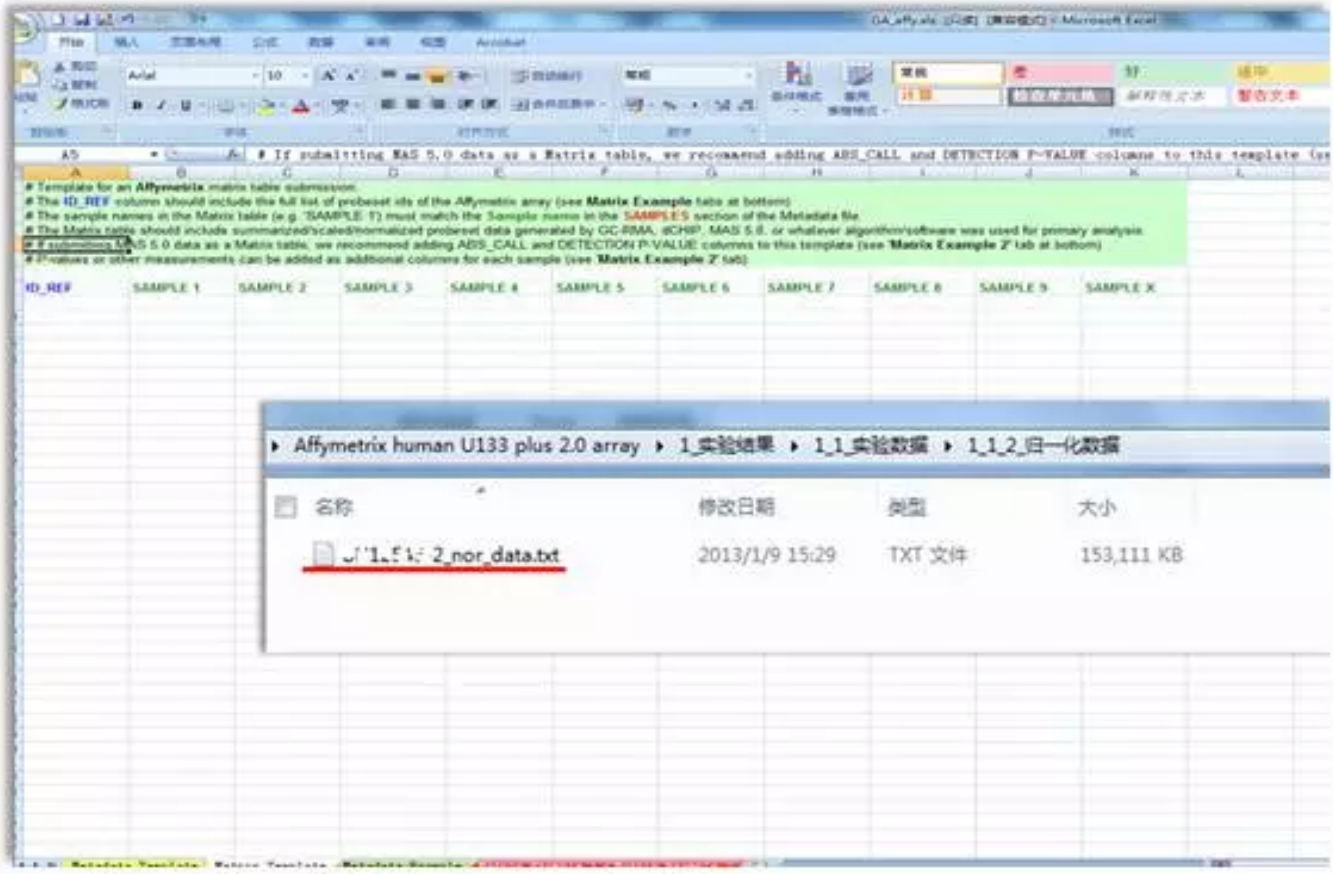
在metadata表中第二部分，查找芯片platform号，我们可以回到“GEOarchive”相关注释的界面上，如下图：

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”



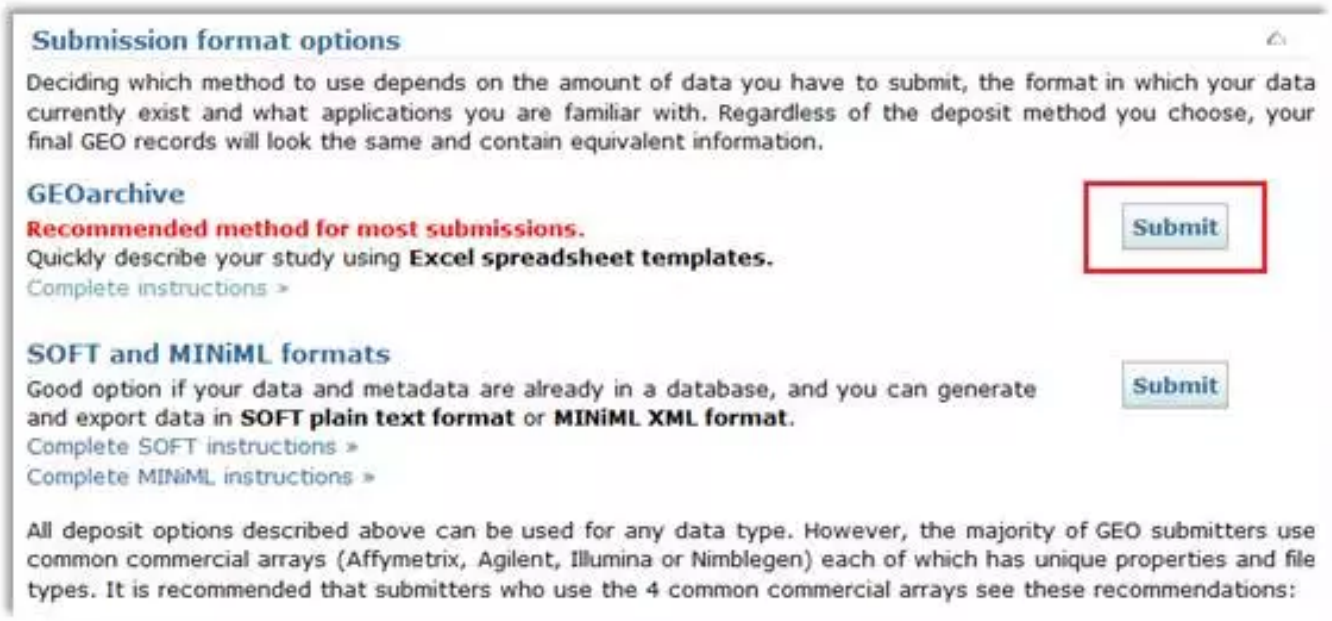
尽可能根据您知道的物种，芯片类型，生产厂家等参数进行筛查。有种最简单直接的是通过您归一化数据实际探针行数来找



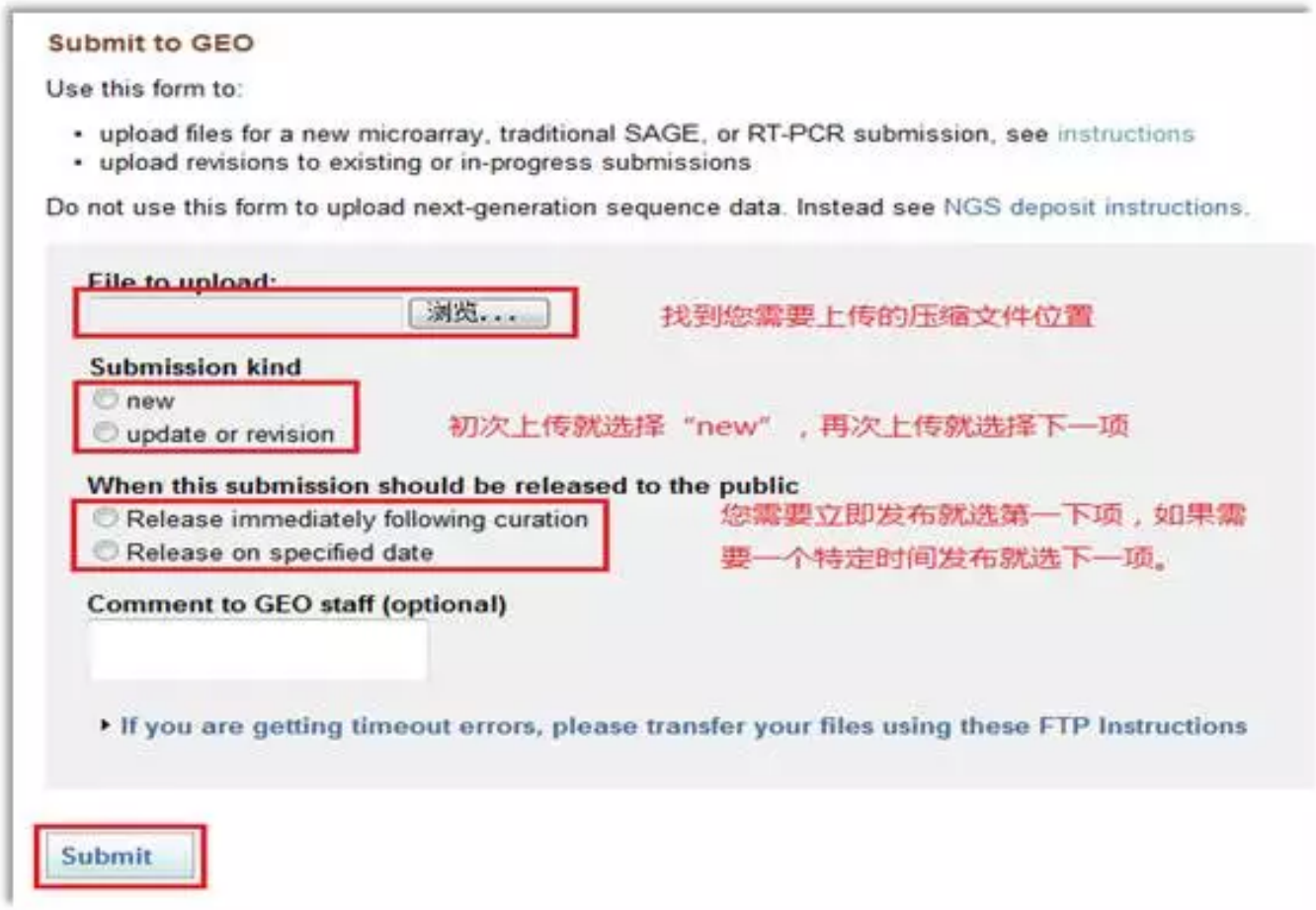
这是matrix表，将您的归一化数据复制粘贴就行了，请用office 07版以上的版本打开它，对于超过2万多个探针的数据03版行数不够。

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:



点击“submit”



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。您就可用于准备发表的文章中。