这两个脚本可以从LPSN上将菌种和对应的典型菌株号抓取下来，同时根据种属名和菌株号从NCBI的基因组信息表中将相应的基因组信息提取出来，最后可使用assembly号通过Batch Entrez网站 (<https://www.ncbi.nlm.nih.gov/sites/batchentrez>)或NCBI开发的datasets软件（https://www.ncbi.nlm.nih.gov/datasets/docs/v2/download-and-install/）批量下载典型菌株的基因组数据（核酸序列、氨基酸序列等）。

**准备工作：**

1、安装python（[www.python.org/](http://www.python.org/)）和Perl（<https://www.perl.org/>）；

2、安装pip（一般自带，在python的主目录Scripts文件夹下，需加入系统变量或切换到该目录运行）。

3、安装和验证Python模块：requests、bs4、os、re。建议在CSDN(https://www.csdn.net/)上进行检索，按照相应的命令行进行安装，操作简便。Tips：模块安装时，一般只需在windows的cmd()或者Linux的终端（Terminal）窗口输入命令：pip install 模块名称，即可自动下载安装。

**使用方法：**

**一、从LPSN（https://www.bacterio.net/domain）爬取信息，脚本名称：GetTypeStrainNumFromLPSN.py**

**步骤1：**进入脚本所在目录。例如，在window系统下，可以打开cmd命令提示符窗口，使用cd命令进入脚本所在目录。

**步骤2：**得到结果文件。例如，当使用者需要批量得到Rhodobacterales下的标准菌株信息时，在cmd命令提示符窗口输入：python GetTypeStrainNumFromLPSN.py --url <https://lpsn.dsmz.de/order/rhodobacterales> --output TEM。其中，--url后输入的是使用者感兴趣的分类所在的网址。按回车键，即可运行得到结果文件Species\_Numbers.txt（TEM文件夹中）。

若没有TEM文件夹，则将自动创建于脚本所在目录；若存在此文件，则会覆盖文件夹中的内容。因此需注意及时保存运行的结果。

**二、从NCBI基因组列表中提取所需基因组信息，脚本名称：GetItemsOfInterest.pl**

1、打开<https://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/> ，下载所有原核生物基因组信息“prokaryotes.txt”文件，放在包含脚本的文件夹内。

2、打开cmd或者终端，切换到包含脚本程序的文件夹为工作目录，输入“perl GetItemsOfInterest.pl prokaryotes.txt”，回车运行，即可在TEM文件夹下得到TypeStrainGenomeInfo.txt文件。可导入excel查看。

**结果分为三部分：**

表格左边三列是从LPSN上抓取的信息，后面的为NCBI提取的信息。

第一部分为高度可信结果；

第二部分Sharing genus or specific name部分的大部分结果是可用的，主要包括一些分类地位发生变动的菌株。需要人工核对。

第三部分Only sharing strain number大部分结果是不对的，包括很少部分分类地位发生变动的菌株。

第二第三部分的结果需要去核对分类地位变更历史，才能确定是否为所需菌种的基因组。

注意：1）如果作者提交的NCBI和发表文章中使用的菌株号不完全对应时（例如空格、“-”、“\_”、等其他符号标注不一致），会被认为不是type strain而漏掉。对于这些漏掉的，用户需要手动检索核对。2）目前NCBI会自动检查用户上传基因组时分类地位和种属名分配是否合理，后期发表了文章，种属名也会自动更新。因此这两个脚本能否抓取绝大多数的基因组。但是我仍然看到有些菌已经发表，但是NCBI的种属名记录还是作者上传的原始状态。综上，如果大家感兴趣的菌种的基因组没有被成功提取，那么还是建议手动检索排查一次。