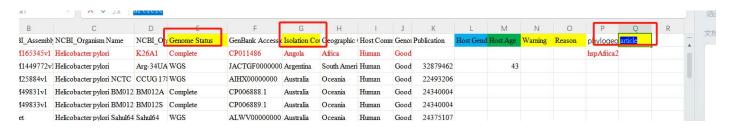
背景: 补充 hp_metadata.xlsx 表格信息

需求概括:

需要爬虫的内容: Genome Status; Isolation Country; article 三个字段填入表中



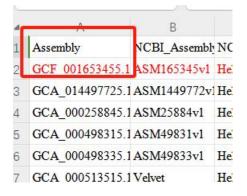
需求详述:

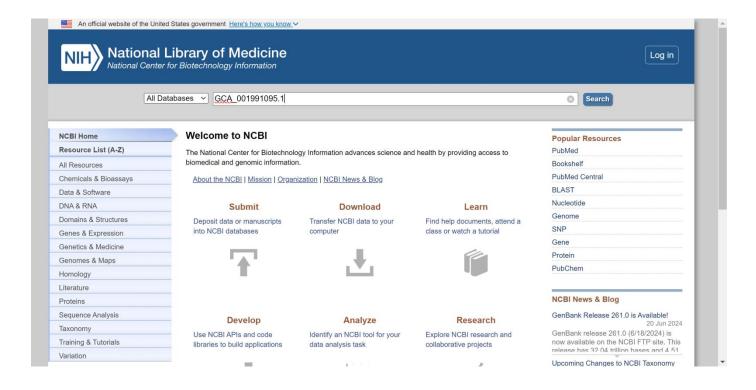
表格名 hp_metadata.xlsx, 共有 NG1-5 和 S1-5 十个 sheet 需要进行的处理相同

1. Genome Status

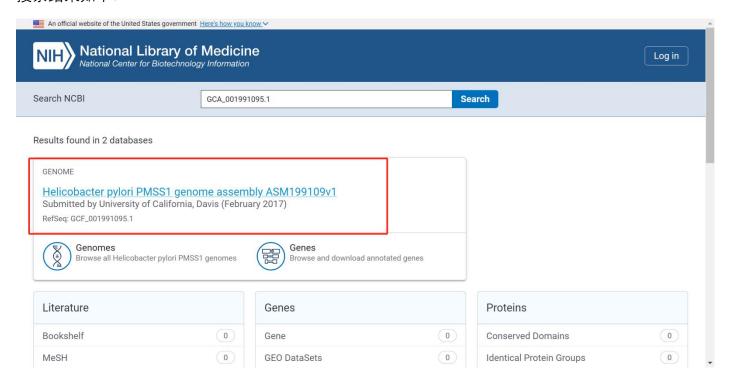
打开 ncbi 官网(https://www.ncbi.nlm.nih.gov/), 在搜索框输入基因组 assembly 编号,表格第一列的

值(如: GCA_001991095.1)



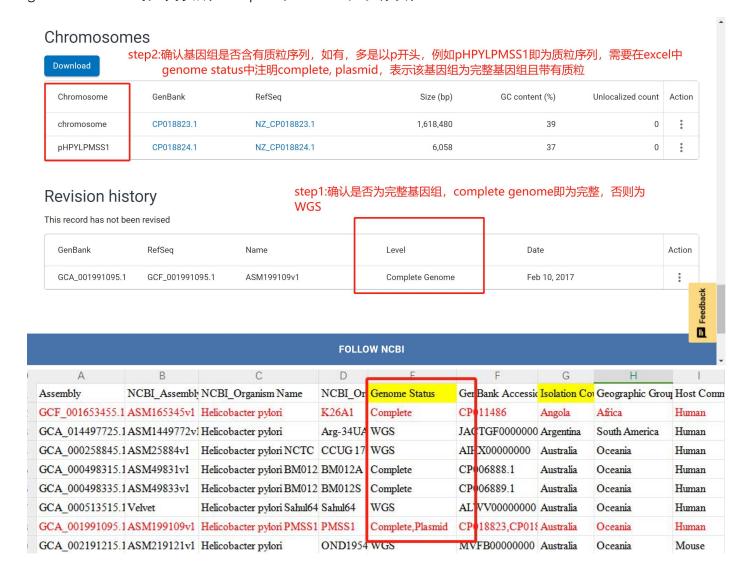


搜索结果如下:



点击进入词条,下拉至最低端,在 revision history 栏目下 level 列找是否为 complete genome,如果是的话在表格 genome status 列填入 complete genome 否则填入 wgs, 如果在 chromosomes 栏目下 chromosome 列出现了 p 开头的行,同时 genome status 列已经是 complete genome 的情况下,将

genome status 对应列变成 Complete, Plasmid 表示有质粒

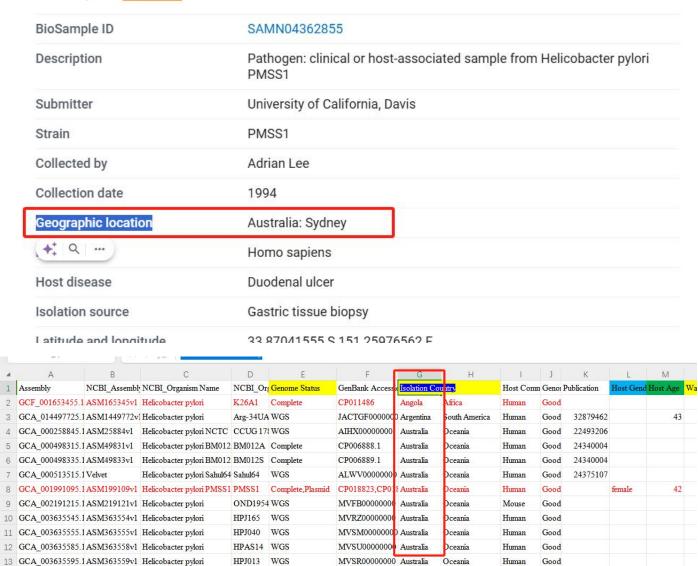


2. Isolation Country

上述检索词条页面(同上一字段爬虫所在页面),寻找 sample details 栏目,下方找到 geographic location 行,爬取对应的结果(如图中 australia: sydney)放入表格中的 Isolation Country 字段中,如果找不到,就忽略(部分样本没有)。其中有些样本是带有城市的,这样的格式都是 country:city,只取国家字段放入表中

另请参阅: information

Sample details



3. article

文章溯源:

Chromosomes



点击 genebank 编号, 查看序列提交信息,

Helicobacter pylori strain PMSS1 complete genome

GenBank: CP018823.1
FASTA Graphics

Go to: ✓ LOCUS CP018823 1618480 bp DNA circular BCT 10-FEB-2017 DEFINITION Helicobacter pylori strain PMSS1 complete genome. ACCESSION CP018823 VERSION CP018823.1 DBLINK BioProject: PRINA306775 BioSample: SAMNO4362855 KEYWORDS Helicobacter pylori PMSS1 ORGANISM Helicobacter pylori PMSS1
Bacteria; Campylobacterota; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter. 1 (bases 1 to 1618480) REFERENCE Draper, J. L., Hansen, L. M., Bernick, D., Abedrabbo, S., Underwood, J. G., Kong, N., Huang, C. B., Weis, A. M., Weimer, B. C., van Vliet, A. H. M., Pourmond, V., Solnick, J. V. and Korplus, K. AUTHORS TITLE Fallacy of the unique genome: Sequence diversity within single Helicobacter pylori strains MBio 8 (1) (2017) In press 2 (bases 1 to 1618480) JOURNAL Hansen, L. M., Karplus, K., Draper, J., Ottemann, K. M., Weis, A. M., Kong, N., Huang, C. B., Weimer, B. C. and Solnick, J. V. AUTHORS TITLE Direct Submission JOURNAL Submitted (04-FEB-2017) Departments of Medicine and Microbiology & Immunology, Center for Comparative Medicine, University of California at Davis, One Shields Avenue, Davis, CA 95616, USA

这里可以看到相应文章的信息,需要溯源回文章查找分离来源国家地区,宿主性别年龄以及进化地理分支等信息。(可能有若干个 TITLE 字段,一般最上方的那个是文章,也可能直接没有)