

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

B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
SI_Assembly	NCBI_Organism Name	NCBI_O	Genome Status	GenBank Access	Isolation Co	Geographic	Host Comm	Geno	Publication	Host Gend	Host Age	Warning	Reason	publ	article	
f165345v1	Helicobacter pylori	K26A1	Complete	CP011486	Angola	Africa	Human	Good							hspAfrica2	
f1449772v1	Helicobacter pylori	Arg-34UA	WGS	JACTGF00000000	Argentina	South Ameri	Human	Good	32879462		43					
f25884v1	Helicobacter pylori NCTC	CCUG 171	WGS	AIHX000000000	Australia	Oceania	Human	Good	22493206							
f49831v1	Helicobacter pylori BM012	BM012A	Complete	CP006888.1	Australia	Oceania	Human	Good	24340004							
f49833v1	Helicobacter pylori BM012	BM012S	Complete	CP006889.1	Australia	Oceania	Human	Good	24340004							
et	Helicobacter pylori Sahul64	Sahul64	WGS	ALWV000000000	Australia	Oceania	Human	Good	24375107							

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）

	A	B
1	Assembly	NCBI_Assembly NC
2	GCF_001653455.1	ASM165345v1 He
3	GCA_014497725.1	ASM1449772v1 He
4	GCA_000258845.1	ASM25884v1 He
5	GCA_000498315.1	ASM49831v1 He
6	GCA_000498335.1	ASM49833v1 He
7	GCA_000513515.1	Velvet He

An official website of the United States government [Here's how you know](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

Log in

All Databases  Search

**NCBI Home**  
Resource List (A-Z)  
All Resources  
Chemicals & Bioassays  
Data & Software  
DNA & RNA  
Domains & Structures  
Genes & Expression  
Genetics & Medicine  
Genomes & Maps  
Homology  
Literature  
Proteins  
Sequence Analysis  
Taxonomy  
Training & Tutorials  
Variation

**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.  
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

**Submit**  
Deposit data or manuscripts into NCBI databases

**Download**  
Transfer NCBI data to your computer

**Learn**  
Find help documents, attend a class or watch a tutorial

**Develop**  
Use NCBI APIs and code libraries to build applications

**Analyze**  
Identify an NCBI tool for your data analysis task

**Research**  
Explore NCBI research and collaborative projects

**Popular Resources**  
PubMed  
Bookshelf  
PubMed Central  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein  
PubChem

**NCBI News & Blog**  
GenBank Release 261.0 is Available!  
20 Jun 2024  
GenBank release 261.0 (6/18/2024) is now available on the NCBI FTP site. This release has 32.04 trillion bases and 4.51  
Upcoming Changes to NCBI Taxonomy

搜索结果如下：

An official website of the United States government [Here's how you know](#)

**NIH** National Library of Medicine  
National Center for Biotechnology Information

Log in

Search NCBI  Search

Results found in 2 databases

**GENOME**  
[Helicobacter pylori PMSS1 genome assembly ASM199109v1](#)  
Submitted by University of California, Davis (February 2017)  
RefSeq: GCF\_001991095.1

**Genomes**  
Browse all Helicobacter pylori PMSS1 genomes

**Genes**  
Browse and download annotated genes

Literature	Genes	Proteins
Bookshelf 0	Gene 0	Conserved Domains 0
MeSH 0	GEO DataSets 0	Identical Protein Groups 0

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

### Chromosomes

Download

step2:确认基因组是否含有质粒序列，如有，多是以p开头，例如pHPYLPMS1即为质粒序列，需要在excel中 genome status中注明complete, plasmid，表示该基因组为完整基因组且带有质粒

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
chromosome	CP018823.1	NZ_CP018823.1	1,618,480	39	0	⋮
pHPYLPMS1	CP018824.1	NZ_CP018824.1	6,058	37	0	⋮

### Revision history

This record has not been revised

step1:确认是否为完整基因组，complete genome即为完整，否则为 WGS

GenBank	RefSeq	Name	Level	Date	Action
GCA_001991095.1	GCF_001991095.1	ASM199109v1	Complete Genome	Feb 10, 2017	⋮

FOLLOW NCBI




A	B	C	D	E	F	G	H	I
Assembly	NCBI_Assembly	NCBI_Organism Name	NCBI_Organism	Genome Status	Gen Bank Accession	Isolation Country	Geographic Group	Host Comments
GCF_001653455.1	ASM165345v1	Helicobacter pylori	K26A1	Complete	CP011486	Angola	Africa	Human
GCA_014497725.1	ASM1449772v1	Helicobacter pylori	Arg-34UA	WGS	JACITGF0000000	Argentina	South America	Human
GCA_000258845.1	ASM25884v1	Helicobacter pylori NCTC	CCUG 17	WGS	AIBX000000000	Australia	Oceania	Human
GCA_000498315.1	ASM49831v1	Helicobacter pylori BM012	BM012A	Complete	CP006888.1	Australia	Oceania	Human
GCA_000498335.1	ASM49833v1	Helicobacter pylori BM012	BM012S	Complete	CP006889.1	Australia	Oceania	Human
GCA_000513515.1	Velvet	Helicobacter pylori Sahul64	Sahul64	WGS	ALVW000000000	Australia	Oceania	Human
GCA_001991095.1	ASM199109v1	Helicobacter pylori PMSS1	PMSS1	Complete,Plasmid	CP018823,CP018824	Australia	Oceania	Human
GCA_002191215.1	ASM219121v1	Helicobacter pylori	OND1954	WGS	MVFB000000000	Australia	Oceania	Mouse

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

# Sample details

另请参阅: [information](#)

BioSample ID	SAMN04362855
Description	Pathogen: clinical or host-associated sample from Helicobacter pylori PMSS1
Submitter	University of California, Davis
Strain	PMSS1
Collected by	Adrian Lee
Collection date	1994
Geographic location	Australia: Sydney
  	Homo sapiens
Host disease	Duodenal ulcer
Isolation source	Gastric tissue biopsy
Latitude and longitude	33 87041555 S 151 25076562 E

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Assembly	NCBI_Assembly	NCBI_Organism Name	NCBI_Or	Genome Status	GenBank Access	Isolation Country		Host Comm	Geno	Publication	Host Gen	Host Age	Wa
2	GCF_001653455.1	ASM165345v1	Helicobacter pylori	K26A1	Complete	CP011486	Angola	Africa	Human	Good				
3	GCA_014497725.1	ASM1449772v1	Helicobacter pylori	Arg-34UA	WGS	JACTGF00000000	Argentina	South America	Human	Good	32879462		43	
4	GCA_000258845.1	ASM25884v1	Helicobacter pylori NCTC	CCUG 17	WGS	AIHX000000000	Australia	Oceania	Human	Good	22493206			
5	GCA_000498315.1	ASM49831v1	Helicobacter pylori BM012	BM012A	Complete	CP006888.1	Australia	Oceania	Human	Good	24340004			
6	GCA_000498335.1	ASM49833v1	Helicobacter pylori BM012	BM012S	Complete	CP006889.1	Australia	Oceania	Human	Good	24340004			
7	GCA_000513515.1	Velvet	Helicobacter pylori Sahul64	Sahul64	WGS	ALWV000000000	Australia	Oceania	Human	Good	24375107			
8	GCA_001991095.1	ASM199109v1	Helicobacter pylori PMSS1	PMSS1	Complete, Plasmid	CP018823, CP018824	Australia	Oceania	Human	Good		female	42	
9	GCA_002191215.1	ASM219121v1	Helicobacter pylori	OND1954	WGS	MVFB000000000	Australia	Oceania	Mouse	Good				
10	GCA_003635545.1	ASM363554v1	Helicobacter pylori	HPJ165	WGS	MVRZ000000000	Australia	Oceania	Human	Good				
11	GCA_003635555.1	ASM363555v1	Helicobacter pylori	HPJ040	WGS	MVSM000000000	Australia	Oceania	Human	Good				
12	GCA_003635585.1	ASM363558v1	Helicobacter pylori	HPAS14	WGS	MVSU000000000	Australia	Oceania	Human	Good				
13	GCA_003635595.1	ASM363559v1	Helicobacter pylori	HPJ013	WGS	MVSR000000000	Australia	Oceania	Human	Good				

3. article

文章溯源:

## Chromosomes

Download

Chromosome	GenBank	RefSeq	Size (bp)	GC content (%)	Unlocalized count	Action
chromosome	CP018823.1	NZ_CP018823.1	1,618,480	39	0	⋮
pHPYLPMS1	CP018824.1	NZ_CP018824.1	6,058	37	0	⋮

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: <a href="#">PRJNA306775</a> BioSample: <a href="#">SAMN04362855</a>				
KEYWORDS	.				
SOURCE	Helicobacter pylori PMSS1				
ORGANISM	<a href="#">Helicobacter pylori PMSS1</a> Bacteria; Campylobacterota; Epsilonproteobacteria; Campylobacterales; Helicobacteraceae; Helicobacter.				
REFERENCE	1 (bases 1 to 1618480)				
AUTHORS	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., Pourmand, N., Solnick, J. V. and Karplus, K.				
TITLE	Fallacy of the unique genome: Sequence diversity within single Helicobacter pylori strains				
JOURNAL	MBio 8 (1) (2017) In press				
REFERENCE	2 (bases 1 to 1618480)				
AUTHORS	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.				
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)