**功能微生物多样性分析**

赵圣国

中国农业科学院北京畜牧兽医研究所

微生物多样性分析中，物种注释是最为关键的步骤。对于微生物多样性分析，常使用16S rRNA基因或ITS序列，利用RDP Classifier等通过朴素贝叶斯算法对序列进行物种注释。

功能微生物是指执行某一特定功能的一类微生物群体，比如产甲烷微生物、尿素分解微生物、氨氧化微生物、固氮微生物。与一般性微生物相比，功能微生物与生态位表型具有更直接的联系，更能反映出生态位的功能变化。因此研究功能微生物多样性，对于解析生态位的功能机制具有重要意义。功能微生物多样性研究中，常对某些关键功能基因进行测序分析。与16S rRNA基因或ITS基因相比，功能基因常具有多个不同拷贝，难以作为系统发育的标签基因，无法根据基因序列组成和相似特点直接进行物种注释，所以常用的RDP Classifier等算法无法适用于功能基因物种注释分析。

GraftM是用于功能基因注释的优秀软件，它通过对已知功能基因构建系统发育树（含物种信息），然后将查询功能基因定位到系统发育树，根据树上位置和距离，注释查询功能基因物种信息。

本章节介绍了功能微生物多样性研究，重点介绍了基于GraftM进行功能微生物的物种注释。功能微生物的OTU聚类、α和β多样性计算等可参考本书第二章第1/2节。

**一、安装方法**

方法1：通过conda安装(提前安装conda)

conda create -n graftm

conda activate graftm

conda install graftm -c bioconda # Examining @/linux-64::\_\_glibc==2.31=0

注：在Ubuntu 16.04/18.04中均顺序安装，在20.04环境下报错，可能是目前软件并没有在最新版系统中进行优化。

方法2：通过pip安装

pip install graftm

配套程序：

orfm v. >= 0.2.0 (https://github.com/wwood/OrfM)

hmmer v. >= 3.1b1 (http://hmmer.janelia.org/)

mfqe v. >= 0.5.0 (https://github.com/wwood/mfqe)

pplacer v. >= 2.6.32 (http://matsen.fhcrc.org/pplacer/)

krona v. >= 2.4 (http://sourceforge.net/p/krona/home/krona/)

mafft v. >= 7.22 (http://mafft.cbrc.jp/)

diamond v. >= 0.9 (https://github.com/bbuchfink/diamond)

FastTreeMP (http://www.microbesonline.org/fasttree/)

方法3：通过Docker安装

安装方法：

docker pull wwood/graftm:v0.11.1

使用方法：

docker run –u $(id -u):$(id -g) –v ‘pwd’ :/data wwood/graftm:v0.11.1 graft <option>

方法4：通过GNU Guix安装

git clone https://github.com/Ecogenomics/ace-guix

GUIX\_PACKAGE\_PATH=ace-guix guix package --install graftm

**二、实例解读**

尿素是牛的饲料氮来源之一，提高尿素的消化利用率，将有助于节约饲料蛋白质成本、维持动物健康、提高动物生产性能和减少氮排放。尿素消化是由牛瘤胃微生物产生的脲酶催化，分解产物氨被用于微生物蛋白质合成。因此，揭示牛瘤胃尿素分解菌多样性对于建立提升尿素利用率的调控技术至关重要。脲酶基因是尿素分解菌的重要功能基因，常用于表征尿素分解菌的多样性。脲酶基因是一类基因簇，由多个基因组成，其中ureC基因作为编码酶活性中心的基因，具有保守性和多变性的特点，是理想的多样性分析用功能基因。为了阐明牛瘤胃液相、固相和瘤胃上皮三个生态位中尿素分解菌多样性，研究学者设计了ureC基因引物，获得ureC基因扩增子，通过高通量测序获得ureC基因序列。在基于QIIME软件进行OTU聚类基础上，利用自己构建的已知菌ureC基因数据库，通过GraftM软件对OTU进行物种注释，然后开展多样性统计分析。

由图1可见，牛瘤胃尿素分解菌中有一半多无法被注释到科水平，表明瘤胃尿素分解菌与已知尿素分解菌分类距离较远，瘤胃中存在大量新的未被发现的尿素分解菌。由图2可见，瘤胃不同生态中尿素分解菌多样性各有不同，其中液相与固相比较相似，而瘤胃上皮与液相固相差异较大，表明生态位环境影响了尿素分解菌的生长。

该研究详情见：Jin D, Zhao S, Zheng N, Bu D, Beckers Y, Denman SE, McSweeney CS, Wang J. Differences in Ureolytic Bacterial Composition between the Rumen Digesta and Rumen Wall Based on ureC Gene Classification. Frontiers in Microbiology. 2017;8:385.

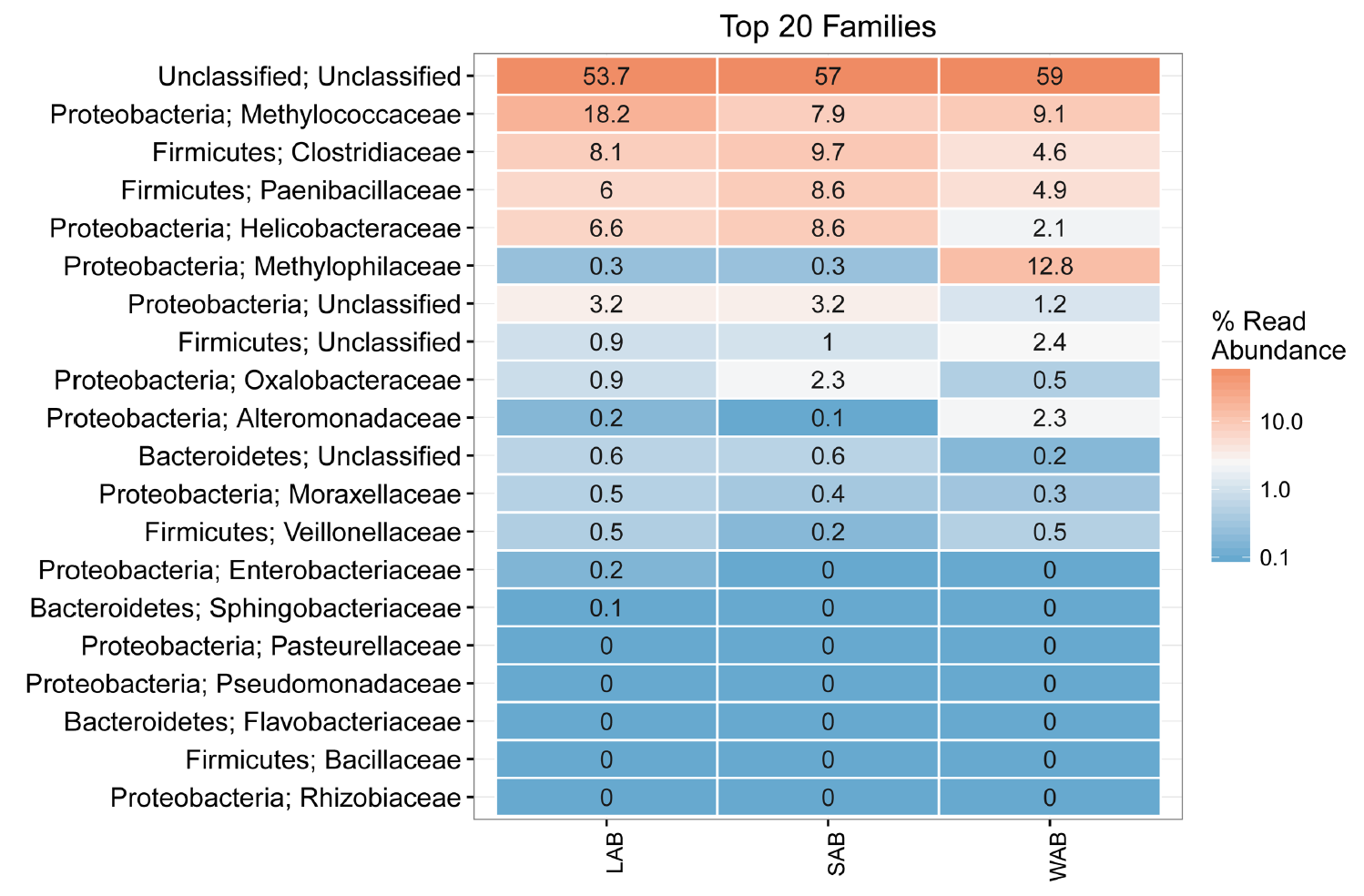


图1 奶牛瘤胃不同生态位中尿素分解菌相对丰度

注：LAB指液相，SAB指固相，WAB指上皮

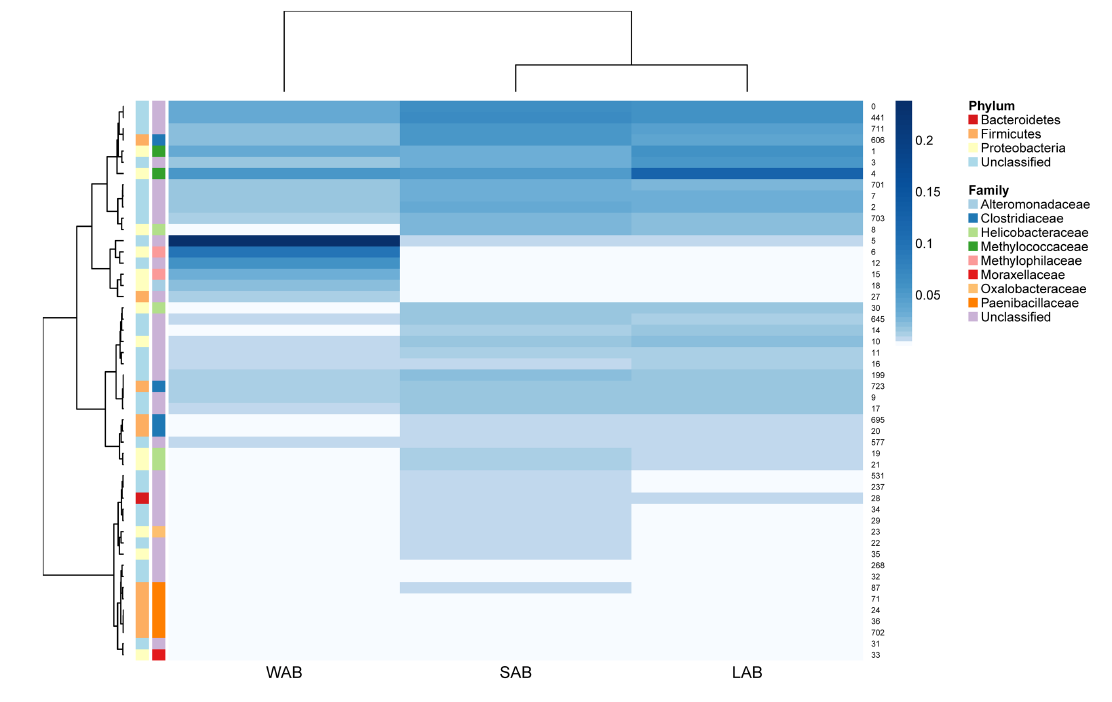


图2 奶牛瘤胃不同生态位中尿素分解菌多样性

注：LAB指液相，SAB指固相，WAB指上皮

**三、实战代码**

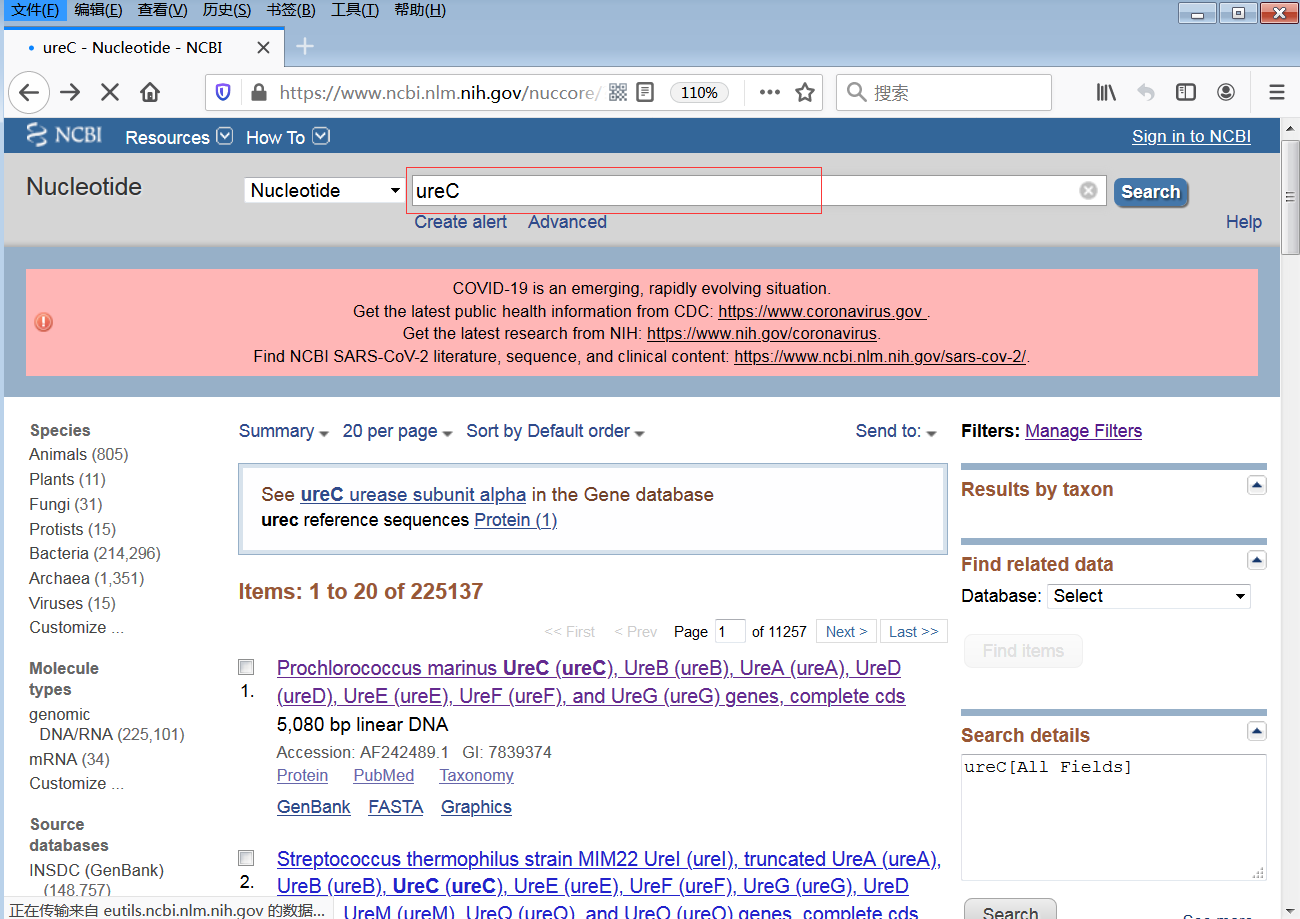
**1. 创建与更新功能基因数据库包：**

**（1）下载功能基因数据**

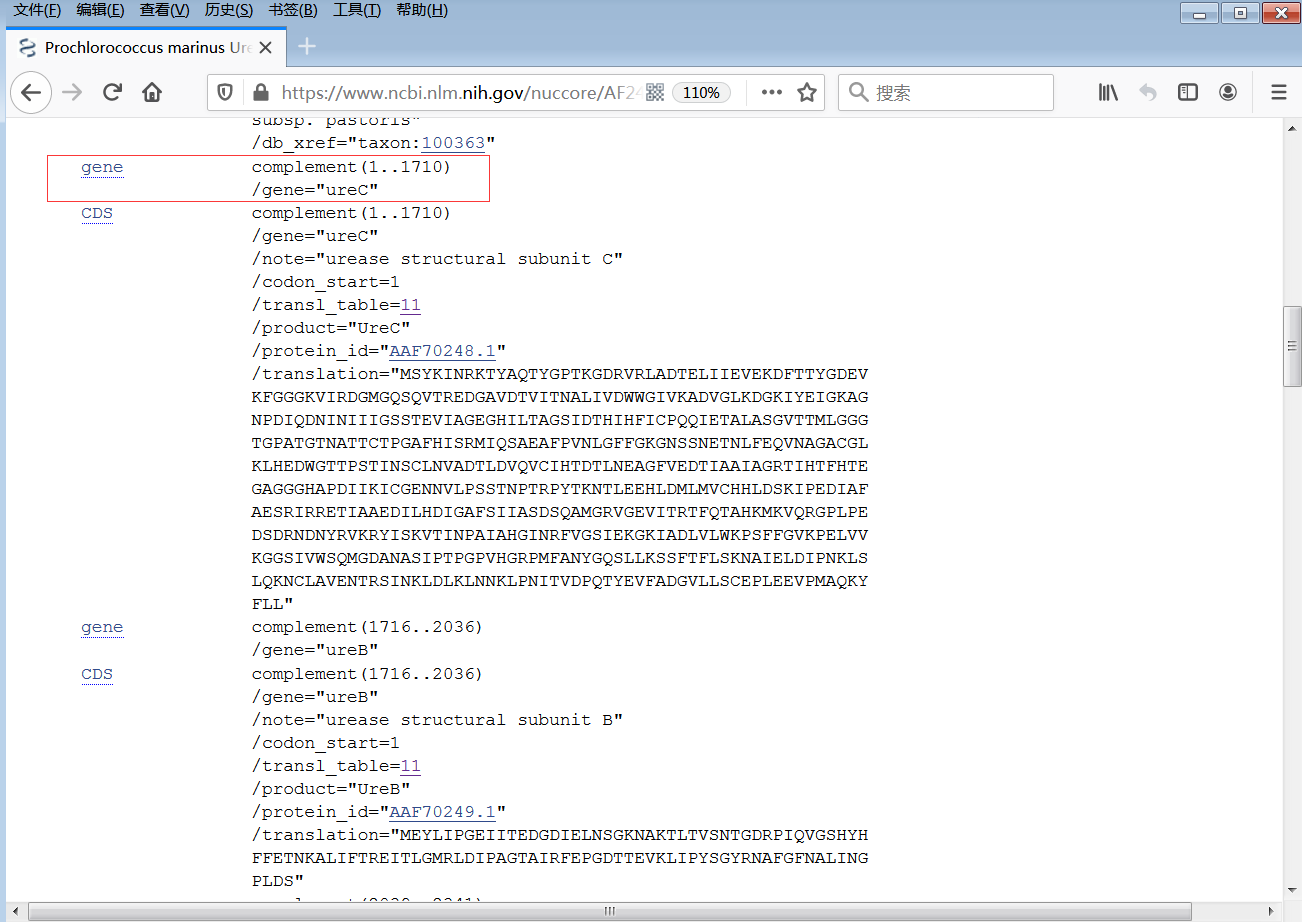
**登录NCBI核酸数据库（**[**https://www.ncbi.nlm.nih.gov/nuccore**](https://www.ncbi.nlm.nih.gov/nuccore)**），根据功能基因名称查询序列，下载目标功能基因序列和物种分类信息，分别整理成两个文件（marker.genes.fasta和marker.genes.taxonomy.txt。**

**例子：以搜索脲酶基因ureC为例**

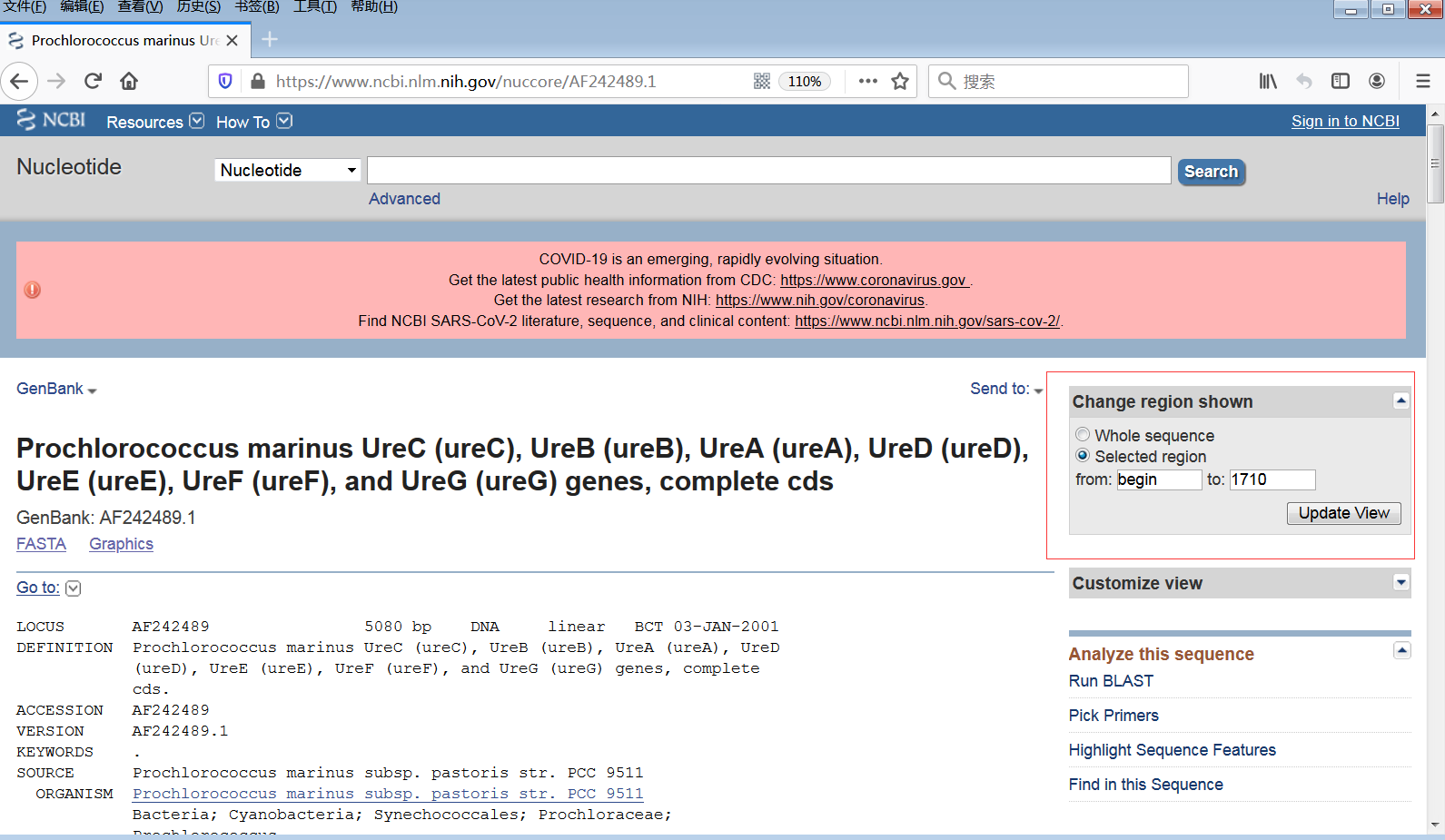
**1）登录NCBI核酸数据库，输入关键词“ureC”，检索后出现所有包含ureC基因的序列或基因组。点击需要下载的序列，进入信息页。**



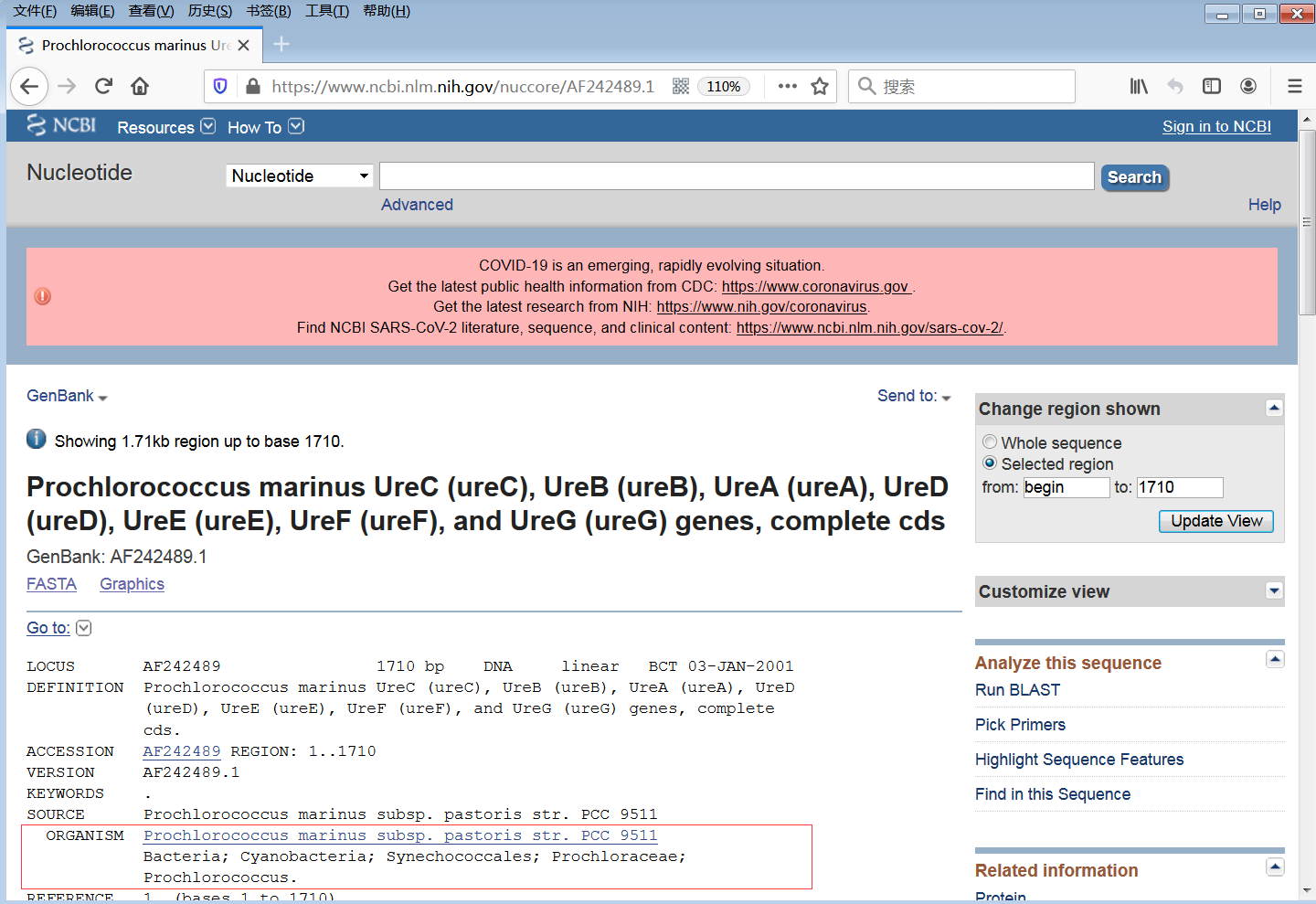
**2）找到ureC基因所在的编码位置，本例中是1 ~ 1710。**



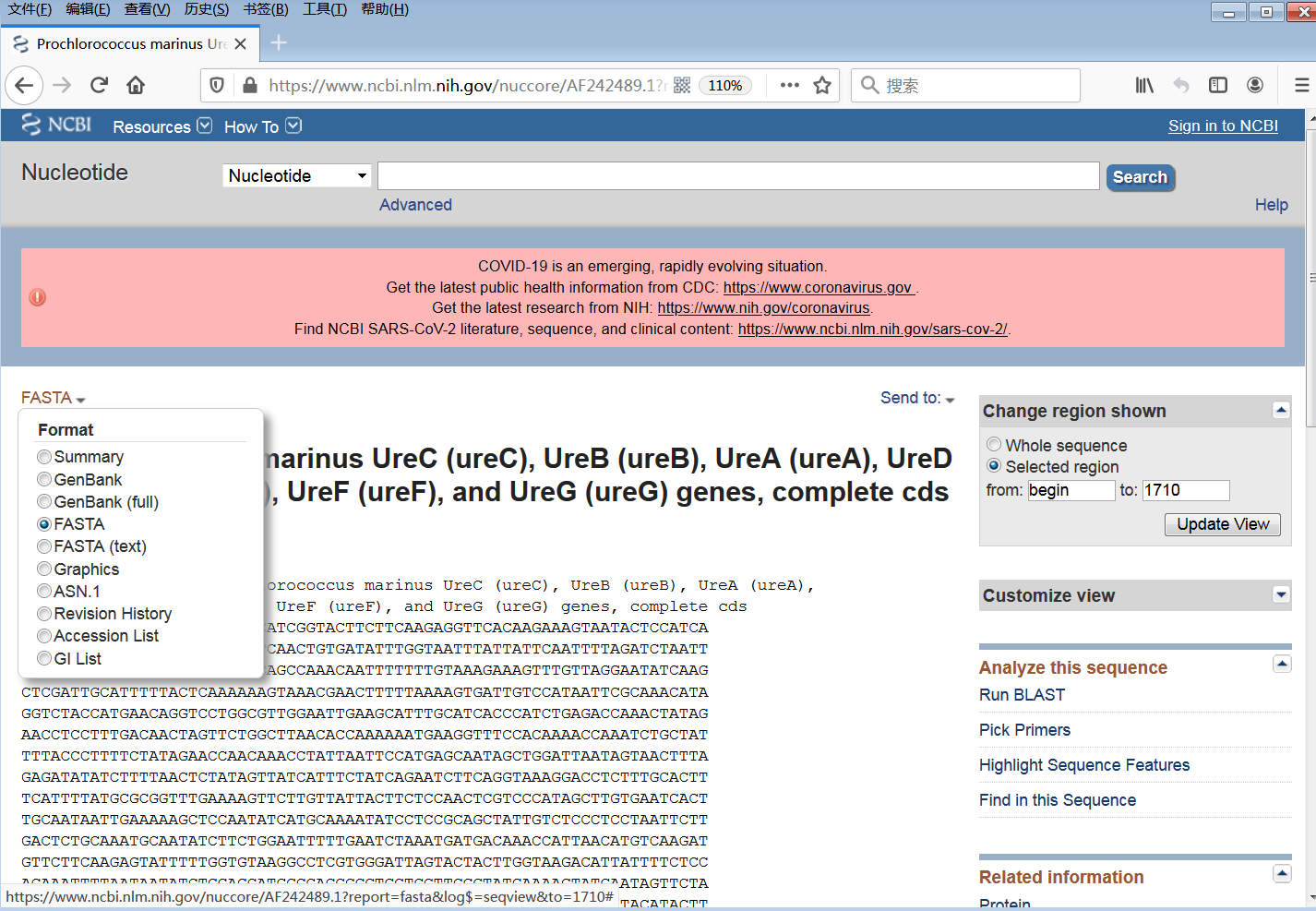
**3）鼠标滑轮上滑后，在“Change region shown”那里输入1 ~ 1710，点击update view.**



**4) 保存ORGANISM信息**



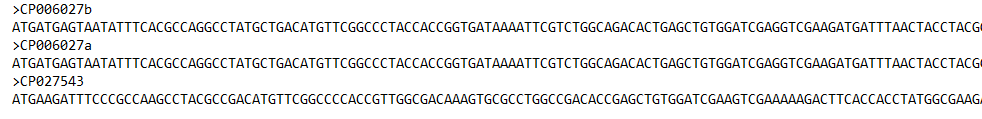
**5）点击显示方式为FASTA，将FASTA格式序列保存。**



**6）将所有下载的ureC基因FASTA序列复制到一个文件中，物种分类信息复制到另一个文件中。**

两个文件格式为：

文件1：参考功能基因文件，格式为FASTA：



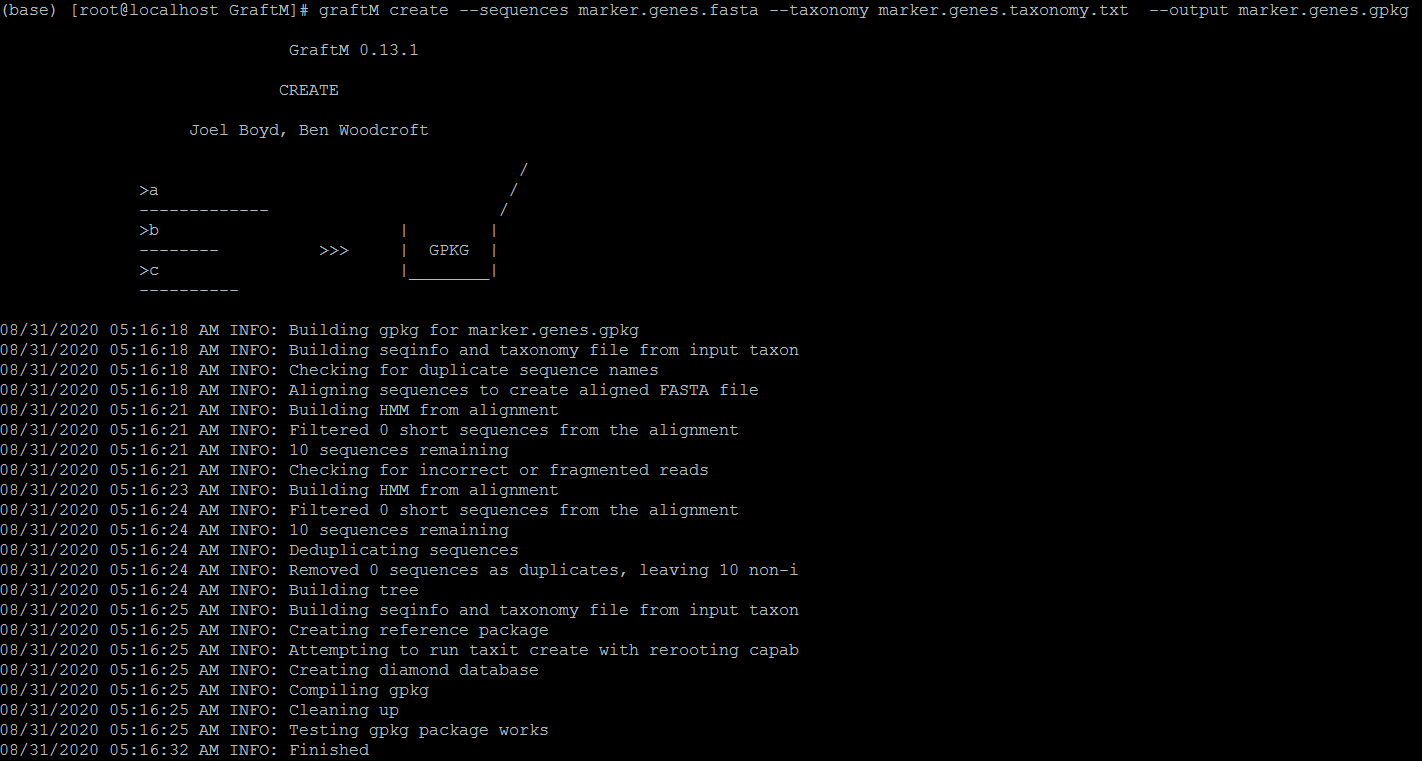
文件2：参考功能基因物种信息文件，文本文件（第一列为ID，第二列为分类信息，两列Tab隔开），格式如下：



**（2）创建数据库包**

运行程序：

graftM create --sequences marker.genes.fasta --taxonomy marker.genes.taxonomy.txt --output marker.genes.gpkg



该图为我运行成功的图，不太确定下面的错误具体什么原因，有可能是配套软件版本问题，我发邮件问问开发者。

# 显示如下报告

08/24/2020 10:31:45 PM ERROR:

taxit create failed to run in a small amount of time suggesting that

rerooting was unsuccessful. Unfortunately this tree will need to be rerooted

manually yourself using a tree editor such as ARB or FigTree.

Once you have a rerooted newick format tree, rerun graftm create

specifying the new tree with --rerooted\_tree. The tree file to be rerooted is 'graftm\_create\_tree.marker.tree'

08/24/2020 10:31:45 PM ERROR:

When rerunning, please use the following flags for the command line to account

for the fact that some sequences may have been removed during the deduplication

process.

graftM create --taxtastic\_taxonomy graftm\_create\_taxonomy.marker.csv --taxtastic\_seqinfo graftm\_create\_seqinfo.marker.csv --alignment graftm\_create\_alignment.marker.faa --rerooted\_tree <REROOTED\_TREE>

# 查看graftM tree --help 需要reference tree （GraftM 0.13.1）

**graftM create参数：**

--sequences；参考功能基因序列文件，必选

--taxonomy；参考功能基因物种信息文件，必选

--alignment；比对后文件，如果有可提交，以减少运行时间

--hmm；HMM文件，如果有可提交，以减少运行时间

--tree；newick格式的系统发育树文件，同时提供log文件

--tree\_log；系统发育树的log文件

--output；输出文件夹

--threads；线程数

--graftm\_package；需要更新的旧数据库包，仅更新数据库包时使用

**（3）更新数据库包**

运行程序：

$ graftM create --graftm\_package marker.genes.gpkg --sequences marker.genes.new.fasta --taxonomy marker.genes.new.taxonomy.txt --output marker.genes.updated.gpkg

**2. 功能基因物种注释**

运行程序：

$ graftM graft --forward query.fasta --graftm\_package marker.genes.gpkg/ --output\_directory query.graftm

**graftM graft参数：**

--forward；查询功能基因序列，fasta格式，必选

--graftm\_package；构建好的数据库包，必选

--output；输出文件夹

--threads；线程数（默认5）

--placements\_cutoff confidence；置信截取值（默认0.75）

**3. 结果查看**

导出文件夹query.graftm中query文件夹中query\_read\_tax.tsv文件。第一列为OTU（Feature）编号，第二列为分类信息。



**参考文献：**

1. Joel A Boyd Ben J Woodcroft Gene W Tyson. GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. Nucleic Acids Research, 2018; 46: e59.
2. Jin D, Zhao S, Zheng N, Bu D, Beckers Y, Denman SE, McSweeney CS, Wang J. Differences in Ureolytic Bacterial Composition between the Rumen Digesta and Rumen Wall Based on ureC Gene Classification. Frontiers in Microbiology. 2017;8:385.
3. <http://geronimp.github.io/graftM>