bio-101 www.bio-protocol.org/eXXX DOI:10.21769/BioProtoc.XXX

基于 GraftM 对功能基因进行物种注释

Taxonomic classification of microbes with a given function based on a specific functional gene

赵圣国 [1, *]

[1] 动物营养学国家重点实验，中国农业科学院北京畜牧兽医研究所院，北京

*通讯作者邮箱：zhaoshengguo@caas.cn

摘要：功能微生物是指执行某一特定功能的一类微生物群体。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。常用的 RDP Classifier 等算法无法适用于功能基因物种注释分析，因此本文介绍了基于 GraftM 的系统发育树原理对功能基因进行物种注释的方法。

关键词: GraftM，功能微生物，功能基因，物种注释

研究背景：

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[1]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[2]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。本文介绍了基于 GraftM 进行功能微生物的物种注释。

软件和数据库

Graftm (0.13.1) (https://pypi.org/project/graftm/)

Bioconda (https://bioconda.github.io/)

实验步骤

一、安装 Graftm 程序

通过 conda 安装:

```
conda create -n graftm
conda activate graftm
conda install graftm -c bioconda
```

二、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)，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)（图 1 和图 2）。

文件 1：参考功能基因文件，marker.genes.fasta，格式为 FASTA：

```
>CP006027b
ATGATGAGTAATATTTCACGCCAGGCCTATGCTGACATGTTCGGCCCTACCACCGGTGATAAAATTCGTCTGGCAGACACTGAGCTGTGGATCGAGGTCGAAGATGATTTAACTACCTACG
>CP006027a
ATGATGAGTAATATTTCACGCCAGGCCTATGCTGACATGTTCGGCCCTACCACCGGTGATAAAATTCGTCTGGCAGACACTGAGCTGTGGATCGAGGTCGAAGATGATTTAACTACCTACG
>CP027543
ATGAAGATTTCCCGCCAAGCCTACGCCGACATGTTCGGCCCCACCGTTGGCGACAAAGTGCGCCTGGCCGACACCGAGCTGTGGATCGAAGTCGAAAAAGACTTCACCACCTATGGCGAAG
```

图 1. 参考功能基因文件格式

文件 2：参考功能基因物种信息文件，marker.genes.taxonomy.txt，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：

```
CP006027b	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichiacoli
CP006027a	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichiacoli
CP027543	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas
```

图 2. 参考功能基因物种信息文件格式

例子：以搜索脲酶基因 ureC 为例[3]

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（图 3）。

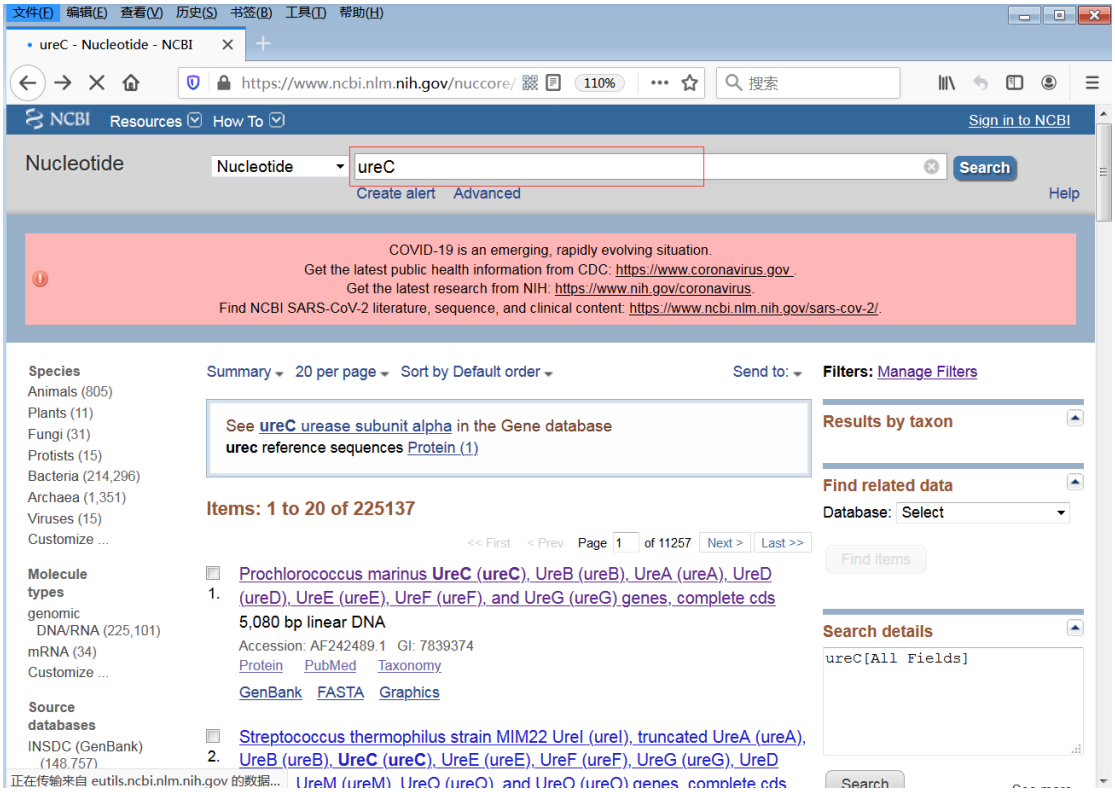

图 3：NCBI 核酸数据库，需要下载序列信息页

2) 找到 ureC 基因所在的编码位置，本例中是 1 – 1710（图 4）。

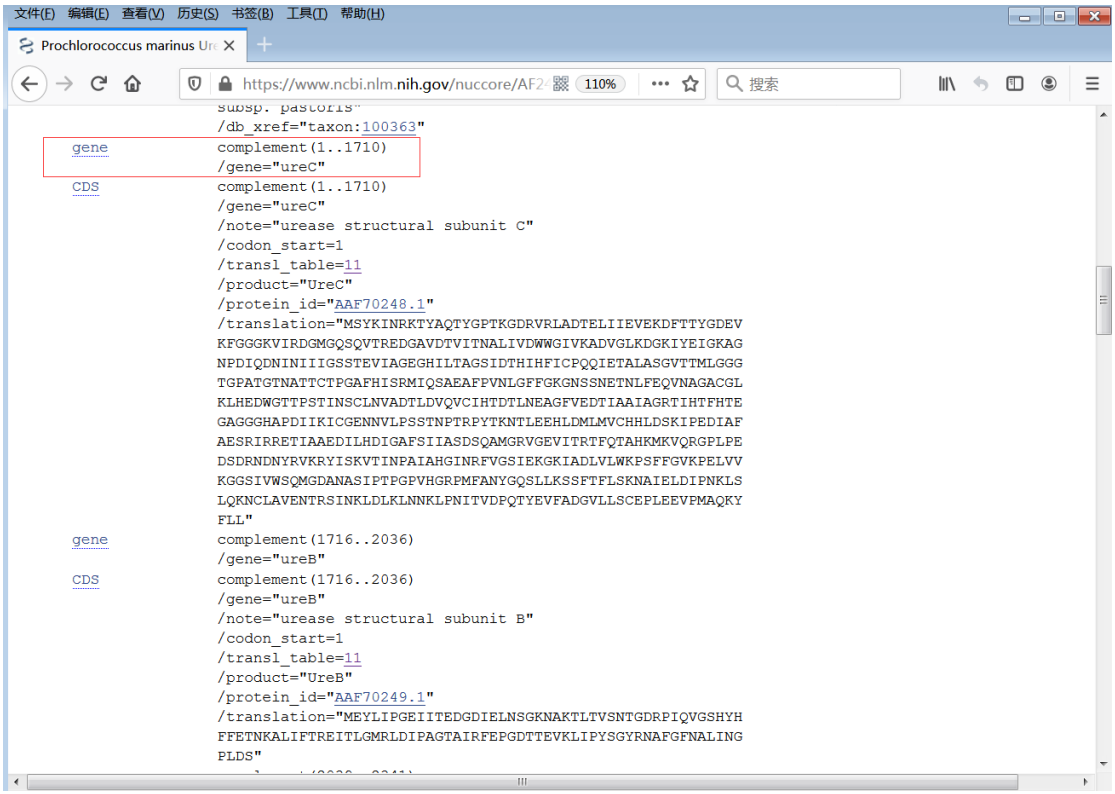

图 4：ureC 基因所在的编码位置

3) 鼠标滑轮上滑后，在“Change region shown”那里输入 1 - 1710，点击 update view（图 5）。

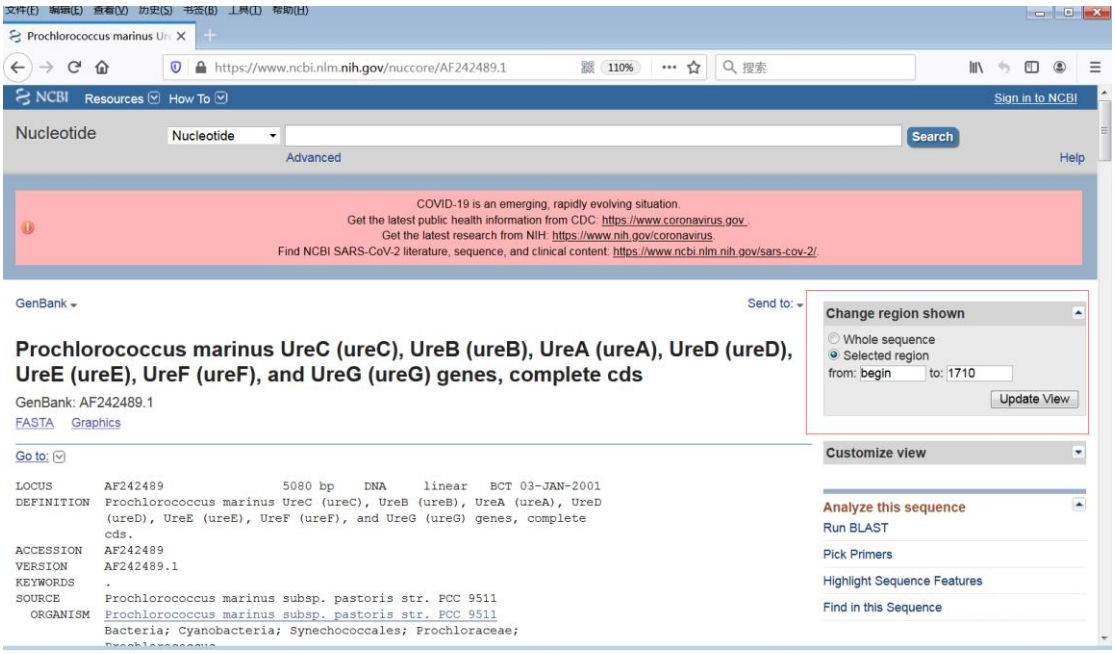

图 5：“Change region shown”界面

4) 保存 ORGANISM 信息（图 6）。

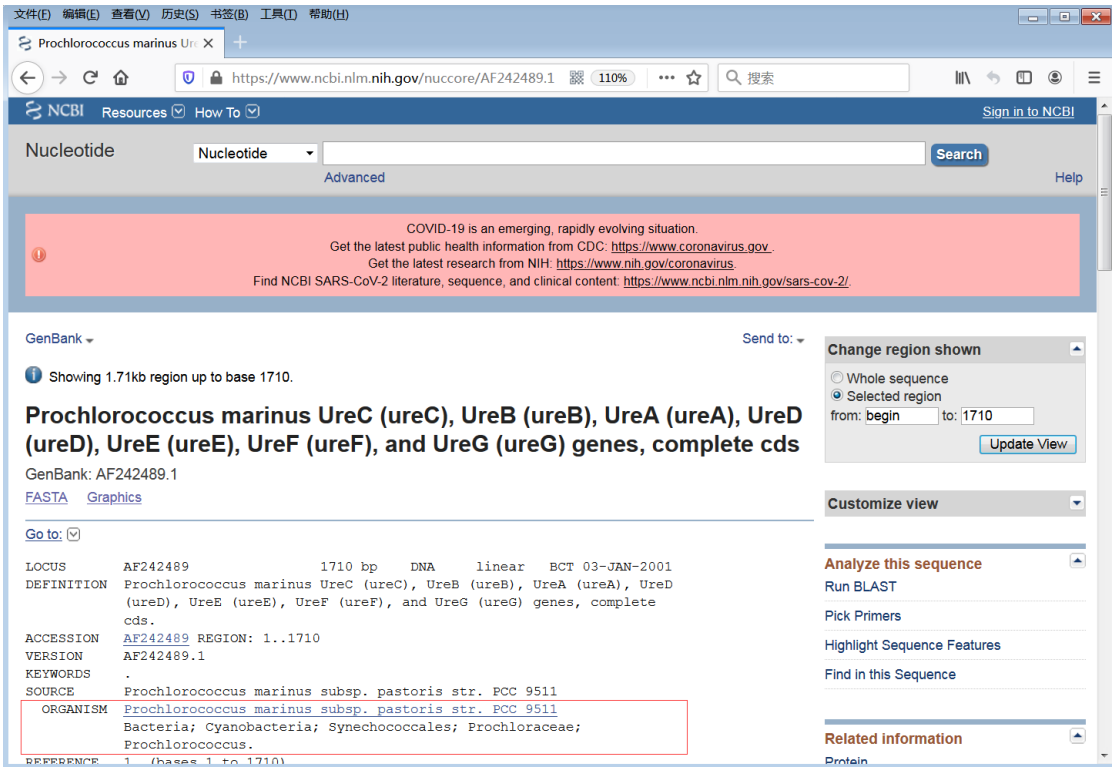

图 6：ORGANISM 信息界面

5) 点击显示方式为 FASTA，将 FASTA 格式序列保存（图 7）。

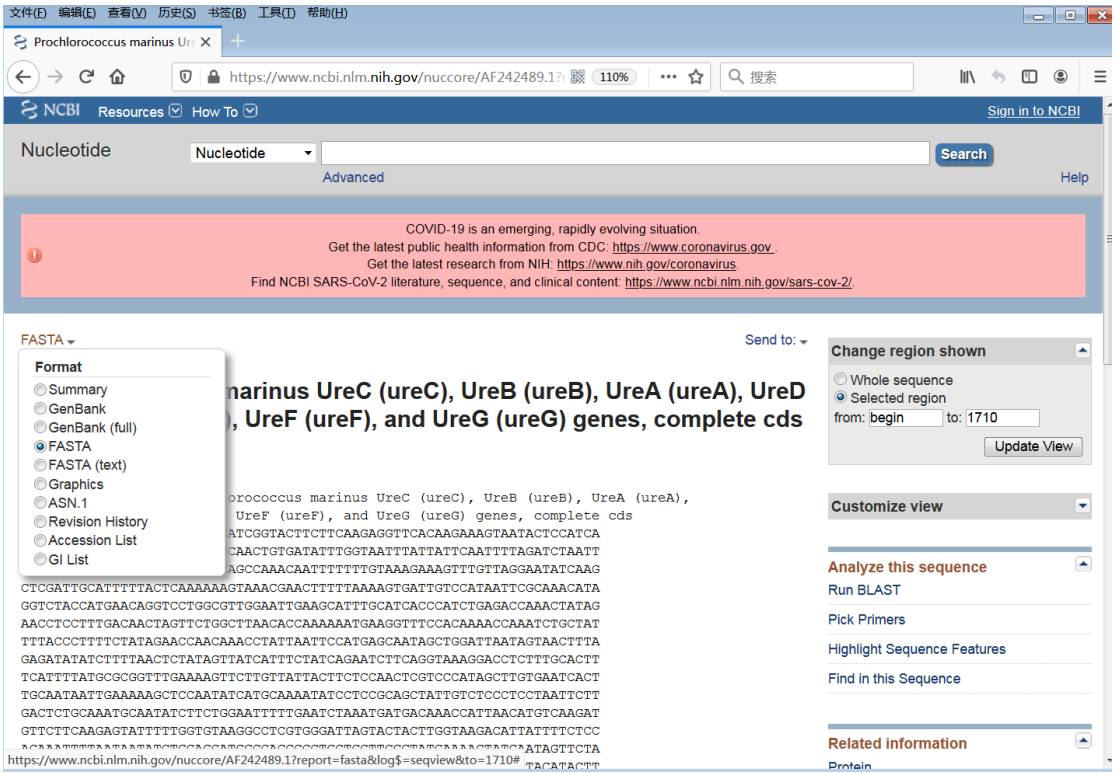

图 7：保存 FASTA 格式

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（图 8，9）：

文件 1：参考功能基因文件，格式为 FASTA：

```
>CP006027b
ATGATGAGTAATATTTCACGCCAGGCCTATGCTGACATGTTCGGCCCTACCACCGGTGATAAAATTCGTCTGGCAGACACTGAGCTGTGGATCGAGGTCGAAGATGATTTAACTACCTACG
>CP006027a
ATGATGAGTAATATTTCACGCCAGGCCTATGCTGACATGTTCGGCCCTACCACCGGTGATAAAATTCGTCTGGCAGACACTGAGCTGTGGATCGAGGTCGAAGATGATTTAACTACCTACG
>CP027543
ATGAAGATTTCCCGCCAAGCCTACGCCGACATGTTCGGCCCCACCGTTGGCGACAAAGTGCGCCTGGCCGACACCGAGCTGTGGATCGAAGTCGAAAAAGACTTCACCACCTATGGCGAAG
```

图 8：参考功能基因文件

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）：

```
CP006027b	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichiacoli
CP006027a	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichiacoli
CP027543	k__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas
```

图 9：参考功能基因物种信息文件

2. 创建功能基因数据库包

运行程序：

graftM create --sequences marker.genes.fasta --taxonomy marker.genes.taxonomy.txt --output marker.genes.gpkg（图 10）：

```
(base) [root@localhost GraftM]# graftM create --sequences marker.genes.fasta --taxonomy marker.genes.taxonomy.txt  --output marker.genes.gpkg

                          GraftM 0.13.1

                         CREATE

                Joel Boyd, Ben Woodcroft

                                                /
            >a                                 /
            -------------                     /
            >b                      |        |
            --------          >>>   |  GPKG  |
            >c                      |________|
            ----------


08/31/2020 05:16:18 AM INFO: Building gpkg for marker.genes.gpkg
08/31/2020 05:16:18 AM INFO: Building seqinfo and taxonomy file from input taxon
08/31/2020 05:16:18 AM INFO: Checking for duplicate sequence names
08/31/2020 05:16:18 AM INFO: Aligning sequences to create aligned FASTA file
08/31/2020 05:16:21 AM INFO: Building HMM from alignment
08/31/2020 05:16:21 AM INFO: Filtered 0 short sequences from the alignment
08/31/2020 05:16:21 AM INFO: 10 sequences remaining
08/31/2020 05:16:21 AM INFO: Checking for incorrect or fragmented reads
08/31/2020 05:16:23 AM INFO: Building HMM from alignment
08/31/2020 05:16:24 AM INFO: Filtered 0 short sequences from the alignment
08/31/2020 05:16:24 AM INFO: 10 sequences remaining
08/31/2020 05:16:24 AM INFO: Deduplicating sequences
08/31/2020 05:16:24 AM INFO: Removed 0 sequences as duplicates, leaving 10 non-i
08/31/2020 05:16:24 AM INFO: Building tree
08/31/2020 05:16:25 AM INFO: Building seqinfo and taxonomy file from input taxon
08/31/2020 05:16:25 AM INFO: Creating reference package
08/31/2020 05:16:25 AM INFO: Attempting to run taxit create with rerooting capab
08/31/2020 05:16:25 AM INFO: Creating diamond database
08/31/2020 05:16:25 AM INFO: Compiling gpkg
08/31/2020 05:16:25 AM INFO: Cleaning up
08/31/2020 05:16:25 AM INFO: Testing gpkg package works
08/31/2020 05:16:32 AM INFO: Finished
```

图 10. 运行结果

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

如果新下载功能基因需要补充到数据库中，则需要更新数据库包。

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
```

三、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)

结果与分析

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，如下所示（图 11）：

```
1       Root; k__Bacteria
2       Root; k__Bacteria; p__Actinobacteria
```

图 11. 运行结果

致谢

感谢中国农业科学院创新工程 (ASTIP-IAS12) 支持。

参考文献

[1] Wang, Q, G. M. Garrity, J. M. Tiedje, and J. R. Cole. (2007). Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy. *Appl Environ Microbiol.* 73: 5261-5267.

[2] Joel A Boyd, Ben J Woodcroft and Gene W Tyson. (2018). GraftM: a tool for scalable, phylogenetically informed classification of genes within metagenomes. *Nucleic Acids Research*. 46(10): e59.

[3] Jin, D., Zhao, S., Zheng, N., Bu, D., Beckers, Y., Denman, S. E., McSweeney, C. S. and Wang, J. (2017). Differences in ureolytic bacterial composition between the rumen digesta and rumen wall based on *urec* gene classification. *Front Microbiol* 8: 385.