**基于二代测序的真菌基因组组装和注释**

**Assembly and Annotation of Fungal Genome Based on Illumina Sequencing**

马紫英1, 2，吴琦1，周欣1, 2，李宽1，蔡磊1 \*

1 中国科学院微生物研究所真菌学国家重点实验室，北京；2中国科学院大学生命科学学院，北京

\*通讯作者邮箱：[cail@im.ac.cn](mailto:cail@im.ac.cn)

**摘要**: 高通量测序技术的飞速发展使得基因组测序成本不断降低，近几年来，越来越多的真菌物种完成了全基因组测序。对于真菌基因组，二代测序的低成本使得大规模测序成为可能，通过三代测序及Hi-C测序辅助组装等方式，可以使基因组组装到近染色体水平，极大的促进了人们对真菌的起源进化、群体遗传多样性、功能基因、致病机制、次级代谢产物等遗传及生物学特性的认识。基因组的组装、基因预测和注释是基因组学研究的基础，本流程以真菌基因组Illumina测序数据为例，详述了真菌基因组的组装、注释流程。

**关键词:** 真菌基因组，数据质控，基因组组装，基因预测注释

**仪器设备**

1. 服务器（型号：PowerEdge R940xa；系统：Ubuntu18.04.3；CPU：英特尔至强6248，80核，1T内存）

**软件和数据库**

1. SRA Toolkit (version2.10.8)
2. FastQC (version0.11.8)
3. Trimmomatic (version0.39)
4. FastUniq (version1.1)
5. SPAdes (version3.12.0)
6. QUAST (version5.0.2)
7. BUSCO (version3.0.2)
8. OrthoDB (version9)
9. Funannotate (version1.4.0)
10. GeneMark-ES (version4.38)
11. BLAST (version2.2.31)
12. eggNOG-mapper (version2.0.0)
13. Swiss-Prot database（2019.10, 264M）

**实验步骤**

一、软件安装

*注：本部分软件安装基于服务器中已安装好anaconda，python，java等基本软件。*

Conda是一个开源的软件包管理和环境管理系统，可以一键安装大多数生物学软件及其依赖关系，分为anaconda和miniconda，在此以anaconda为例简述其安装过程，如果空间有限，可使用精简版miniconda。

# anaconda下载安装，可根据官网或anaconda镜像自行下载需要的版本。

# 官网：<https://www.anaconda.com>

# 镜像：<https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/>

$ wget -c <https://repo.anaconda.com/archive/Anaconda2-5.2.0-Linux-x86_64.sh>

$ bash Anaconda2-5.2.0-Linux-x86\_64.sh

# 安装时许可协议输入yes，默认目录为~/anaconda2，默认不运行conda直接回车。

# 添加生物频道

$ conda config --add channels defaults

$ conda config --add channels conda-forge

$ conda config --add channels bioconda

# 详细安装说明可参考：<https://mp.weixin.qq.com/s/SzJswztVB9rHVh3Ak7jpfA>

* 1. 需要自行安装的软件，命令如下：

（1）SRA Toolkit

$ wget [https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.10.8/sratoolkit.2.10.8-ubuntu64.tar.gz -P /home/user/software/](https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/2.10.8/sratoolkit.2.10.8-ubuntu64.tar.gz%20-P%20/home/user/software/)

$ tar zxf sratoolkit.2.10.8-ubuntu64.tar.gz

（2）FastQC

$ wget [http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc\_v0.11.8.zip -P /home/user/software/](http://www.bioinformatics.babraham.ac.uk/projects/fastqc/fastqc_v0.11.8.zip%20-P%20/home/user/software/)

$ unzip /home/user/software/fastqc\_v0.11.8.zip

$ chmod 755 /home/user/software/FastQC/fastqc

（3）OrthoDB数据库

$ wget [http://busco.ezlab.org/v2/datasets/sordariomyceta\_odb9.tar.gz -P /home/user/software/BUSCO/](http://busco.ezlab.org/v2/datasets/sordariomyceta_odb9.tar.gz%20-P%20/home/user/software/BUSCO/)

$ tar zxf sordariomyceta\_odb9.tar.gz

（4）Funannotate

由于Funannotate 需要依赖特别多的软件，故基于anaconda为该软件新建一个环境。

$ conda create -n funannotate

$ conda activate funannotate

$ conda install -c bioconda funannotate

#安装funannotate\_db数据库

$ funannotate setup -d $HOME/funannotate\_db

#配置funannotate\_db环境

$ echo "export FUNANNOTATE\_DB=$HOME/funannotate\_db" > /conda/installation/path/envs/funannotate/etc/conda/activate.d/funannotate.sh

$ echo "unset FUNANNOTATE\_DB" > /conda/installation/path/envs/funannotate/etc/conda/deactivate.d/funannotate.sh

（5）Genemark（运行Funannotate依赖的软件）

软件及密钥下载需要注册，下载地址为<http://topaz.gatech.edu/GeneMark/license_download.cgi>

$ tar zxf /home/user/software/gm\_et\_linux\_64.tar.gz

$ gzip -dc /home/user/software/gm\_key\_64.gz > ~/.gm\_key

# 安装perl模块

$ cpan -i YAML Hash::Merge Logger::Simple Parallel::ForkManager

#配置genemark环境

$ echo "export GENEMARK\_PATH=/home/user/gm\_et\_linux\_64/gmes\_petap" > /conda/installation/path/envs/funannotate/etc/conda/deactivate.d/funannotate.sh

（6）eggNOG-mapper及数据库

#下载eggNOG-mapper安装包

$ git clone <https://github.com/jhcepas/eggnog-mapper.git>

$ cd eggnog-mapper/

#下载并安装数据库

$ ./download\_eggnog\_data.py

* 1. 其他软件anaconda安装

其他软件如Trimmomatic，FastUniq，SPAdes，QUAST，BUSCO，blast可使用anaconda安装，以Trimmomatic为例，命令如下：

$ conda install trimmomatic

* 1. 配置环境变量

自行安装的软件运行时可使用软件的绝对路径或将软件路径写入环境变量，以方便引用，以SRA Toolkit为例，命令如下：

$ echo 'PATH=$PATH:/home/user/software/sratoolkit/bin/' >> ~/.bashrc

$ source ~/.bashrc

二、数据获得

本文中的数据来自“The genome of opportunistic fungal pathogen Fusarium oxysporum carries a unique set of lineage-specific chromosomes” （Zhang等，2020）中的二代数据（SRX6453258）和基因组（GCA\_009746015.1），鉴于数据量较大，故选取部分数据进行示例。

* 1. 下载原始测序数据

$ prefetch SRR9694936

* 1. 下载的数据为sra格式，需要对数据进行拆分和转换

$ fastq-dump --split-files SRR9694936

#注：拆分结果为SRR9694936\_1.fastq和SRR9694936\_2.fastq

* 1. 下载基因组数据

$ wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/009/746/015/GCA\_009746015.1\_ASM974601v1/GCA\_009746015.1\_ASM974601v1\_genomic.fna.gz

$ gunzip GCA\_009746015.1\_ASM974601v1\_genomic.fna.gz

# 4）选取部分数据进行后续示例分析，实际分析中不用执行此步骤：

# $ head -n 20000000 SRR9694936\_1.fastq > illumina.1.fastq

# $ head -n 20000000 SRR9694936\_2.fastq > illumina.2.fastq

# $ seqtk sample GCA\_009746015.1\_ASM974601v1\_genomic.fna 3 > genome.fa

三、使用FastQC对数据进行质量评估

# FastQC是一款基于java的软件，用于对高通量数据进行质量控制。

$ mkdir result\_qc

$ fastqc -o result\_qc -t 10 illumina.1.fastq illumina.2.fastq

运行结束后，使用浏览器打开html文件，部分结果如图1：



**图1. FastQC质控结果.** A为报告目录，合格的部分用绿色对勾表示，不合格用红色叉号表示；B表示每个位置所有碱基的质量值的范围，在绿色区域表示测序质量高。

此份数据Adaper Content为绿色对勾，大部分reads在红色区域，表示此份数据已去掉接头，但测序质量不高。

更多参数的具体结果解读可参考官网：<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/>

四、使用Trimmomatic对数据进行质量控制

#Trimmomatic是针对Illumina测序平台的数据过滤工具，用来去除接头和低质量碱基，对于未去接头的原始下机数据（Anthony等，2014），命令如下：

$ trimmomatic PE -threads 4 illumina.1.fastq illumina.2.fastq illumina.1.clean.fastq illumina.1.unparied.fastq illumina.2.clean.fastq illumina.2.unparied.fastq ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:50 TOPHRED33

主要参数：

PE/SE 设定针对Paired-End 或者Single-Endd reads进行处理

-threads 运行线程数

ILLUMINACLIP:TruSeq3-PE.fa:2:30:10 切接头序列

LEADING:3 切除reads开头质量值低于3的碱基

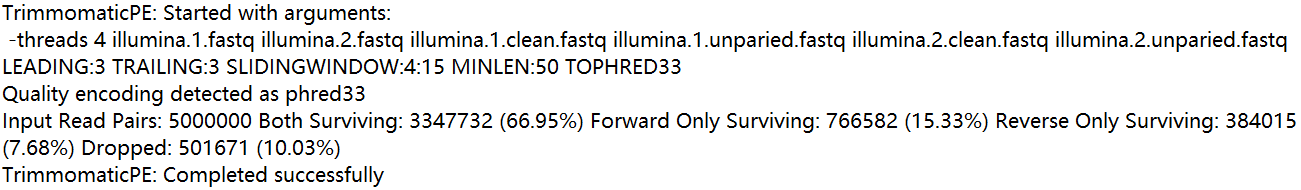
TRAILING:3 切除reads尾部质量值低于3的碱基

SLIDINGWINDOW:4:15 从 reads 的 5' 端开始进行滑窗质量过滤，切掉碱基质量平均值低于阈值（15）的滑窗（4个碱基）

MINLEN:50 保留剪切后reads长度的最小值

TOPHRED33 将碱基质量转换为phred-33格式

运行结束后，终端屏幕会出现如下过滤信息，图2：



**图2. Trimmomatic过滤结果**

五、使用FastUniq去除PCR重复

#illumina文库构建中pcr扩增会给测序结果引入PCR重复，可以使用FastUniq来去除（Xu等，2012）。

$ mkdir FastUniq

$ cd FastUniq/

#注：需要建立一个文本文件，写入需要处理的成对的fastq文件的路径。

$ ls illumina.1.clean.fastq illumina.2.clean.fastq > fragment.list

$ fastuniq -i fragment.list -o illumina.1.rd.fastq -p illumina.2.rd.fastq

主要参数：

-o 和-p 均为过滤后输出的序列文件

#注：对于自己测得的基因组数据，如果测序质控比较严格，对原始数据进行了很好的过滤，可以直接使用质控过的cleandata进行基因组拼接，而不必执行以上步骤四和五。

六、基因组*de novo*组装

二代数据的组装软件较多，如SPAdes，SOAPdenovo2，IDBA-UD，ALLPATHS-LG等，在此我们选用SPAdes软件进行*de novo*组装（Anton等，2012）。SPAdes适用于细菌/真菌等小型基因组的组装，不适用于大型基因组，输入数据可以是Ion Torrent, PacBio, Oxford Nanopore, 以及Illumina paired-end, mate-pairs和single reads。对于小型真菌基因组的Illumina paired-end reads，组装命令如下：

$ spades.py -k 21,33,55,77,99 --careful --cov-cutoff auto -1 illumina.1.rd.fastq -2 illumina.2.rd.fastq -o output

主要参数：

-k 设置kmer的大小

--careful 通过运行 MismatchCorrector 模块进行基因组上 mismatches 和 short indels 的修正

--continue从上一次终止处继续运行程序

--cov-cutoff auto 计算coverage值

#如果有多个 paired-end library 数据时，可使用如下参数：

$ spades.py -k 21,33,55,77,99 --careful --cov-cutoff auto --pe1-1 illumina1\_1.fastq --pe1-2 illumina1\_2.fastq --pe2-1 illumina2\_1.fastq --pe2-2 illumina2\_2.fastq -o output

运行结束后，在output文件夹中的scaffolds.fasta为拼接好的基因组文件。

注：该软件目前已更新到version 3.14.1，更多详细参数请参考官方说明文档：[http://cab.spbu.ru/files/release3.14.1/manual.html#sec1.1](http://cab.spbu.ru/files/release3.14.1/manual.html%23sec1.1)

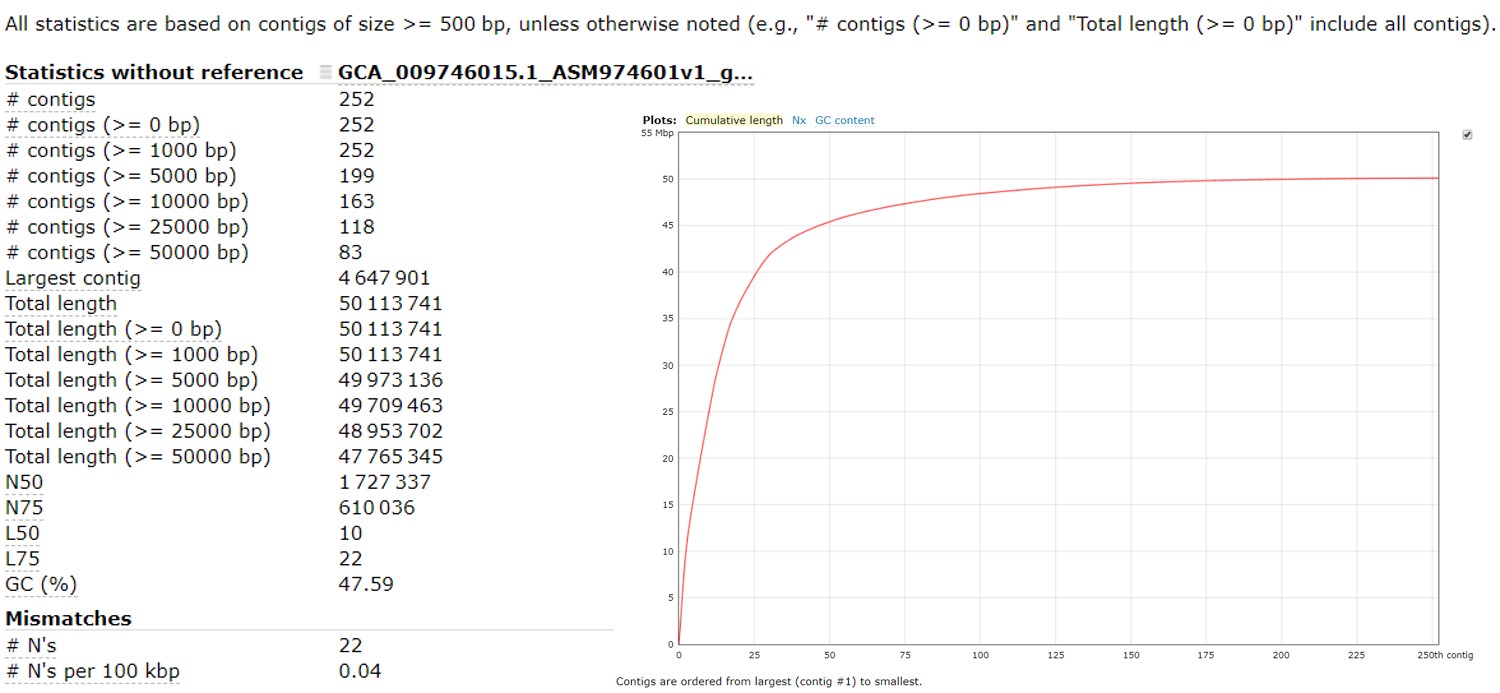
*注：由于在上述组装示例中使用的数据量少，组装结果不足以进行下一步组装质量评估和基因预测，故选取完整基因组GCA\_009746015.1\_ASM974601v1\_genomic.fna为例进行质量评估。*

七、用QUAST评估组装质量

#QUAST软件用来评估基因组的组装效果，有无参考基因组均可（Alexey等，2013）。

$ quast.py GCA\_009746015.1\_ASM974601v1\_genomic.fna -o output

运行结束后，在浏览器中打开html文件，结果如图3：



**图3. QUAST评估结果**

该基因组组装共252条contig，基因组大小为约50M，N50值为1727337bp，对于2代测序来说组装质量较好。

八、使用BUSCO评估组装质量

#BUSCO软件通过同源基因数据库从基因完整度来评价基因组组装结果（Mathieu等，2019）。

$ python run\_BUSCO.py -i GCA\_009746015.1\_ASM974601v1\_genomic.fna -o output -l /home/user/software//BUSCO/sordariomyceta\_odb9/ -m genome -c 10

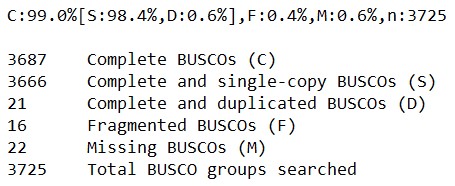
主要参数：

-l 指定单拷贝基因数据集

-m 评估模式，包含3种类型：geno or genome，基因组序列；tran or transcriptome，转录本序列；prot or proteins，蛋白氨基酸序列。

-c 程序运行线程数

运行结束后会在run\_output中生成一系列文件，其中short\_summary\_output文档中会显示整体评估结果，如图4：



**图4. BUSCO评估结果**

其中C值表示完整性，即被评估基因组与BUSCO基因相比的比例，S表示单拷贝基因的比例，D表示多拷贝基因的比例，M值表示可能缺少的基因的比例。如图4，C值和S值均大于98%，该基因组组装较好。

*注：BUSCO目前已更新到BUSCO version 5 beta，使用时可下载最新版及相应的OrthoDB（*[*https://busco-archive.ezlab.org/v3/*](https://busco-archive.ezlab.org/v3/)*）。*

更多帮助信息可参考官网：[https://busco.ezlab.org/busco\_userguide.html#manual-installation](https://busco.ezlab.org/busco_userguide.html%23manual-installation)

九、使用Funannotate进行基因预测

Funannotate是一个集基因组预测、注释和比较的综合软件（Palmer，2016），最初是为注释真菌基因组（小型真核生物，基因组大小约30 Mb）而写的，经过不断更新目前也可以用于较大的基因组。该软件的结果输出格式更易于NCBI数据库的数据提交。该预测软件整合了AUGUSTUS, GeneMark, Snap, GlimmerHMM, BUSCO, EVidence Modeler, tbl2asn, tRNAScan-SE, Exonerate, minimap2的预测结果。

# 首先软屏蔽基因组中的重复序列

$ funannotate mask -i genome.fa --cpus 12 -o genome\_masked.fasta

主要参数：

-i 输入文件，fasta基因组序列

-o 输出文件

-m 重复序列屏蔽方式，默认tantna，也可选repeatmasker或者repeatmodeler

-l, 如果选用repeatmodeler进行屏蔽，需设置本地repeat数据库的位置

# 使用屏蔽重复序列后的基因组进行基因预测

funannotate predict -i genome\_masked.fasta --species "Pseudogenus specicus" -o fun/ --busco\_seed\_species fusarium\_graminearum --busco\_db sordariomycetes --cpus 12

必须参数：

-i 输入文件，即mask后的输出结果

-o 输出文件夹

-s, --species 预测基因组的物种名

主要可选参数：

--maker\_gff 自行通过MAKER2预测的结果文件，直接用于EVM整合预测结果

-w, --weights设置EVM整合的权重

--busco\_seed\_species 选择BUSCO中Augustus软件进行训练的物种名，一般选择近缘物种，默认为anidulans

--busco\_db 选择BUSCO数据的模型，默认为dikarya

--ploidy 基因组的倍型，默认为1

--genemark\_gtf自行通过Genemark预测的结果文件

--min\_intronlen 最小内含子长度，默认为10

--max\_intronlen 最大内含子长度，默认为2000

--min\_protlen最小蛋白序列长度，默认为50

--cpus CPU默认为2

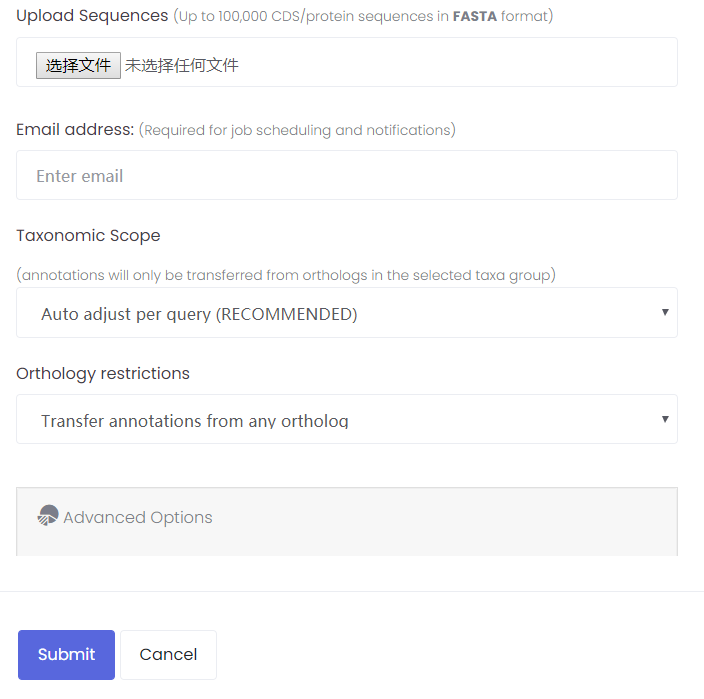
运行结束后，在输出文件夹fun内共3个子文件夹，其中logfiles文件夹中的文档记录了每一步的运行过程，可以查看结果和报错信息，predict\_results文件夹中为预测结果，包括Augustus的训练文件，预测基因的蛋白序列和CDS序列文件及gff和Genbank格式的预测文件，可以用于后续的基因注释。

十、基因功能注释

基因功能注释主要是基于同源序列比对进行的。根据预测结果中的蛋白序列和蛋白质数据库进行比对的结果，完成相应的功能注释。常用的数据库包括：Nr库（NCBI官方非冗余蛋白数据库）、Swiss-Prot数据库（蛋白序列得到实验验证）、KEGG数据库（代谢通路数据库）、eggNOG数据库（直系同源蛋白分组比对数据库，是NCBI的COG数据库的扩展）、InterPro数据库（由多种不同的数据库组成，如CDD，Pfam，ProDom，PROSITE，SMART，SUPERFAMILY等，两月更新一次）等，另外还有一些功能数据库如CAZyme、EffectorP数据库等，可以根据课题要求和分析目的进行选择。在此以eggNOG（Jaime等，2017&2019）和Swiss-Prot数据库两种不同的注释方式为例进行详述。

* 1. eggNOG mapper在线版：<http://eggnog-mapper.embl.de/>

在线版只需选择蛋白序列文件、设置邮箱、提交任务即可，如图5，操作简单方便，一次最多提交100000条蛋白序列，但是不适用于大规模序列注释。



**图5. eggNOG mapper在线版**

* 1. eggNOG mapper本地版

$ python emapper.py -i proteins.fa --output eggnog\_output -m diamond -d euk

主要参数：

-i输入文件

--output输出文件前缀

-m 设置比对算法。可以选择hmmer或diamond

-d指定数据库数据，真菌选euk

运行结束后会生产两个文件，eggnog\_output.emapper.seed\_orthologs和eggnog\_output.emapper.annotations，其中eggnog\_output.emapper.seed\_orthologs为每条序列的最佳比对结果列表，重点关注eggnog\_output.emapper.annotations，为对比结果，主要包括：

query\_name：序列名

seed\_eggNOG\_ortholog：最佳比对的eggNOG编号

seed\_ortholog\_evalue：最佳蛋白比对的E值

seed\_ortholog\_score：最佳蛋白比对得分

predicted\_gene\_name：预测基因名

GOs：GO注释

KEGG\_KO：KEGG注释的KO编号

BiGG\_Reactions：代谢反应

COG functional categories：COG功能分类

eggNOG free text desc：功能描述

* 1. Swiss-Prot注释

UniProtKB/Swiss-Prot数据库中的蛋白质功能经过了功能验证，注释准确，数据库较小，适合用于本地化blast进行的注释。

官网：<https://www.uniprot.org/>

# 下载Swiss-Prot的蛋白序列并构建Blast数据库

$ wget ftp://ftp.uniprot.org/pub/databases/uniprot/current\_release/knowledgebase/complete/uniprot\_sprot.fasta.gz -P /home/user/software/

$ gzip -dc /home/user/software/uniprot\_sprot.fasta.gz > uniprot\_sprot.fasta

$ makeblastdb -in uniprot\_sprot.fasta -dbtype prot -title uniprot\_sprot -parse\_seqids -out uniprot\_sprot -logfile uniprot\_sprot.log

#使用blastp进行Swiss-Prot注释

$ blastp -query proteins.fa -out swiss-prot.tab -db /home/user/software/uniprot\_sprot -evalue 1e-5 -outfmt 7

*注：结果中每个蛋白的详细信息可查询官网：<https://www.uniprot.org/>*

十一、下游分析

通过以上的基因组组装、基因预测和功能注释，我们已经获得了（1）目标物种的基因组信息，（2）蛋白序列和CDS序列及其在基因组上的位置，（3）大部分基因的功能预测，这些结果都为后续的个性化分析和深入研究如基因组共线性、同源蛋白分析、进化树构建、基因家族分析、功能基因验证等提供数据基础。

**失败经验**

1. Genemark-ES密钥存在一定的有效期，如果密钥过期，可以通过之前的网站重新注册下载新的密钥，替换原来的密钥即可。
2. Genemark 安装perl模块的过程中，可能会提示某些模块安装不成功，如hash-merge，最终使用anaconda安装完成。

$ conda install -c bioconda perl-hash-merge

**参考文献**

1. Zhang, Y., Yang, H., David, T., Zhou, S. G., Dilay, H. A., Gregory, A. D., Guo, L., Karen, B., Nathan, W., Jeffrey, J. C. et al. (2020). [The genome of opportunistic fungal pathogen fusarium oxysporum carries a unique set of lineage-specific chromosomes.](https://pubmed.ncbi.nlm.nih.gov/32005944/) *Commun Biol* 3(50): 1-12.
2. Anthony, M. B., Marc, L. and Bjoern, U. (2014). [Trimmomatic: A flexible trimmer for Illumina sequence data.](https://pubmed.ncbi.nlm.nih.gov/24695404/) *Bioinformatics* 30(15): 2114-2120.
3. Xu, H. B., Luo, X., Qian, J., Pang, X. H., Song, J. Y., Qian, G. R., Chen, J. H. and Chen, S.L. (2012). [FastUniq: A fast de novo duplicates removal tool for paired short reads.](https://pubmed.ncbi.nlm.nih.gov/23284954/) *PLoS ONE* 7(12): e52249.
4. Anton, B., Sergey, N., Dmitry, A., Alexey, A. G., Mikhail, D., Alexander, S. K., Valery, M. L., Sergey, I. N., Son, P., Andrey, D. P., Alexey, V. P., Alexander, V. S., Nikolay, V., Glenn, T., Max, A, A. and Pavel, A. P. (2012). [SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.](https://pubmed.ncbi.nlm.nih.gov/22506599/) *J Comput Biol* 19(5): 455-477.
5. Alexey, G., Vladislav, S., Nikolay, V. and Glenn, T. (2013). [QUAST: quality assessment tool for genome assemblies.](https://pubmed.ncbi.nlm.nih.gov/23422339/) *Bioinformatics* 29 (8): 1072-1075.
6. Mathieu S., Mosè, M. and Evgeny, M.Z. (2019). [BUSCO: assessing genome assembly and annotation completeness.](https://pubmed.ncbi.nlm.nih.gov/31020564/) *Methods Mol Biol* 1962 :227-245.
7. Palmer, J. (2016). [Funannotate: pipeline for genome annotation.](https://funannotate.readthedocs.io/en/latest/commands.html) <https://funannotate.readthedocs.io/en/latest/index.html.>
8. Jaime, H. C., Kristoffer, F., Luis, P. C., Damian, S., Lars, J. J., Christian, V. M. and Peer, B. (2017). [Fast genome-wide functional annotation through orthology assignment by eggnog-mapper.](https://pubmed.ncbi.nlm.nih.gov/28460117/) *Mol Biol Evol* 34(8): 2115-2122.
9. Jaime, H. C., Damian, S., Davide, H., Ana, H.P., Sofia, K. F., Helen, C., Daniel, R. M., Ivica, L., Thomas, R., Lars J. J., Christian, V. M. and Peer, B. (2019). [eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses.](https://pubmed.ncbi.nlm.nih.gov/30418610/) *Nucleic Acids Res* 47(Database issue): D309–D314.