**全基因组测序分析SOP**

目录

[1. 概述（背景介绍） 2](#_Toc99702877)

[1.1质控 2](#_Toc99702878)

[1.2组装 2](#_Toc99702879)

[1.3评估 3](#_Toc99702880)

[1.4基于全基因组的物种分类注释 3](#_Toc99702881)

[1.5注释 3](#_Toc99702882)

[2. 软件安装 4](#_Toc99702883)

[2.1安装bioconda 4](#_Toc99702884)

[2.2安装分析软件 5](#_Toc99702885)

[3. 数据处理 6](#_Toc99702886)

[3.1激活基因组分析的conda环境 6](#_Toc99702887)

[3.2使用manifestGen.py脚本将数据存放路径写入到csv文件中 6](#_Toc99702888)

[3.3使用WGS-Pipeline.py脚本对数据进行质控、组装、评估、注释 6](#_Toc99702889)

[3.4 CheckM对基因组完整度和污染度进行评估 7](#_Toc99702890)

[3.5 GTDB-tk使用基因组数据确定物种分类地位 8](#_Toc99702891)

[3.6 功能注释分析 8](#_Toc99702892)

# 概述（背景介绍）

## 1.1质控

测序下机的FASTQ数据需要进行质控和预处理，以保证下游分析输入的数据都是干净可靠的。**使用工具为fastp**，该工具具有以下功能特点：

（1）对数据自动进行全方位质控，生成网页版交互式报告。

（2）过滤功能（去掉低质量，太短，太多N……）。

（3）对每一个序列的头部或尾部，计算滑动窗内的质量均值，并将均值较低的子序列进行切除（类似Trimmomatic的做法，但是快非常多）。

（4）全局剪裁 （在头/尾部，不影响去重），对于Illumina下机数据往往最后一到两个cycle需要这样处理。

（5）去除接头污染。不用输入接头序列，算法会自动识别接头序列并进行剪裁。

（6）对于双端测序（PE）的数据，会自动查找每一对read的重叠区域，并对该重叠区域（overlap）中不一致的碱基对，依据质量值进行校正。

（7）去除尾部的polyG。对于Illumina NextSeq/NovaSeq的测序数据，因为是两色法发光，polyG是常有的事，所以此选项对该两类测序平台默认打开。

（9）可以对带分子标签（UMI）的数据进行预处理，不管UMI在插入片段还是在index上，都可以轻松处理。

（10）可以将输出进行分拆，而且支持两种模式，分别是指定分拆的个数，或者分拆后每个文件的行数。

<https://github.com/OpenGene/fastp>

## 1.2组装

经过质控的数据，就可以用于基因组的组装，已经有大量用于二代测序数据拼接工具被开发出来，比如：Velvet、Spades、ABySS、SOAPdenovo等。这里**推荐使用Unicycler**，该工具分析流程的最大亮点在可以同时使用二代short reads和三代long reads测序数据混合拼接。对于仅有Illumina测序的双端序列数据，unicycler可以在spades初步组装的结果上进行优化，包括：（1）区分单拷贝的contigs和重复的contigs，选择合适的单拷贝contigs进行下一步处理；（2）为了尽可能的环化contigs，剔除contig间的重叠区域；（3）构建单拷贝contigs桥接图；（4）使用 Pilon将短序列mapping到组装好的基因组上进行矫正。因此使用Unicycler即可获得较高质量的基因组草图。(俊宇已经用模拟数据与Spades和MEGAHIT比较过。

[GitHub - fw1121/Unicycler: hybrid assembly pipeline for bacterial genomes](https://github.com/fw1121/unicycler)

## 1.3评估

对于**基因组组装效果的评估使用QUAST**，主要用于统计contigs数目、长度、GC含量、N50、L50等参数。一般认为N50越长，基因组拼接效果越好。对于拼接**基因组的污染度和完整度的评估使用CheckM**，CheckM 提供了一系列工具用于评估从分离培养、单细胞、宏基因组获得的基因组质量，可以根据基因组在参考基因组发育树中的位置来推断其精确的单拷贝标记基因集（lineage-specificmarker set），同时也提供数据库可用的基于分类学的基因集（taxonomic-specificmarker set）。CheckM利用基因的单拷贝性来有效的估计基因组完整度和污染，同时能绘制基因组关键特征（例如GC含量、编码率）的图像来评估基因组的质量。

[GitHub - Ecogenomics/CheckM: Assess the quality of microbial genomes recovered from isolates, single cells, and metagenomes](https://github.com/Ecogenomics/CheckM)

## 1.4基于全基因组的物种分类注释

**基于全基因组数据进行物种分类注释，使用较多的工具是GTDBtk**，此工具依赖GTDB（GENOME TAXONOMY DATABASE）的基因组数据库，这个数据库中包含了纯培养基因组和宏基因组组装的基因组，这个数据库最大的特点在于对基因组分类地位的准确定位，与NCBI相比，约58%在NCBI分类系统中已收录基因组的分类地位有变动。并且数据库会定期更新，添加新发现的基因组以及对已有基因组的重新命名。GTDBtk的分类系统以细菌中普遍存在的120个单拷贝蛋白质（bac120）为基础，在大量氨基酸水平差异的基础上构建新的分类系统，目前此工具支持**对细菌和古菌进行物种分类**，对于真菌没有涉及。

[Genome Taxonomy Database (ecogenomic.org)](https://gtdb.ecogenomic.org/)

## 1.5注释

基因组上ORF的预测普遍使用prodigal，该工具使用动态规划算法，特别关注改善基因结构预测，改善翻译起始位点识别和减少假阳性三个目标，预测准确度较好，被众多分析流程整合。

获得ORF后，可以进一步获得基因的功能信息。基因的功能注释指的是根据数据库中已知编码基因的注释信息（包括motif、domain），基于同源比对，对基因中的模序和结构域、基因编码的蛋白质功能、所参与的信号传导通路和代谢途径等的预测。常用的数据库有NR、SWISS-PROT、InterProScan、COG、eggNOG、KEGG、GO等。此外，基因组注释内容还可涉及蛋白激酶、病原与宿主互作、致病毒力因子预测、抗性基因等等。

**目前，对原核基因组进行基因预测和注释的主流工具是Prokka**，Prokka是一款简单、快速和高效的原核基因组注释流程（已经整合prodigal和相关数据库），其输出结果包含多种格式的文件：.gff 基因注释文件，.gbk Genebank格式，.faa 翻译CDS的氨基酸序列，.ffn所有转录本核酸序列，.tsv所有注释基因特征表格等。

[GitHub - tseemann/prokka: Rapid prokaryotic genome annotation](https://github.com/tseemann/prokka)

其它方面的注释，如：抗性基因（CARD）、毒力因子（VFDB）、病原与宿主互作、碳水化合物利用（CAZy）、转运蛋白（TCDB）、化合物基因簇（antiSMASH、BiG-SCAPE）、肠道微生物代谢（gutSMASH）等，可以根据需要使用单独工具进行注释。

# 软件安装

## 2.1安装bioconda

Bioconda是一个自动化管理生物信息软件的工具，就像APPstore、360软件管家一样。其优点是安装简单，各个软件依赖的环境一同打包且相互隔离，非常适合在服务器中建立自己的生物信息分析环境。目前已有的conda环境anno、antismash、bigscape、gtdbtk、gutsmash、hiprfish、PathoFact、qiime2-2021.2、rgi、roary、taxa、wgs，可以根据数据分析的需求选择使用不同的conda环境，这里我们需要对全基因组数据进行分析，因此选择“wgs”。

选读：

（1）下载biconda

wget <https://repo.continuum.io/miniconda/Miniconda3-latest-Linux-x86_64.sh>

（2）安装软件

sh Miniconda3-latest-Linux-x86\_64.sh

source ~/.bashrc

（3）添加软件源

conda config --add channels bioconda

conda config --add channels conda-forge

（4）基本使用

·搜索软件（以软件名“bwa”为例）

conda search bwa

·安装软件

conda install bwa #软件名后可以指定安装软件的版本号，如bwa=0.7.17

·查看已安装软件

conda list

·创建新的conda环境

conda creat -n your\_env\_name python=X.X #可以指定python版本，有些工具需要依赖<python3

·查看已有的conda环境

conda env list

·激活conda环境

conda activate your\_env\_name

## 2.2安装分析软件

（1）创建基因组分析的conda环境

conda create -n wgs

（2）安装分析需要的软件

conda install -y -c bioconda fastp unicycler quast bandage prokka diamond pandas checkm-genome

安装GTDBtk

conda create -n gtdbtk

conda activate gtdbtk

conda install -c bioconda gtdbtk

（3）配置数据库

Prokka：prokka --setupdb

CheckM：下载文件<https://data.ace.uq.edu.au/public/CheckM_databases/>，指定存放目录

checkm data setRoot <checkm\_data\_dir>

GTDB：download-db.sh 下载速度较慢的话可以手动下载放到gtdb软件安装路径下的db文件夹。

# 数据处理

推荐使用俊宇整合的pipline，里面整合了以上大部分软件。

## 3.1激活基因组分析的conda环境

conda activate /home/chenjunyu/miniconda3/envs/wgs/

## 3.2使用manifestGen.py脚本将数据存放路径写入到csv文件中

将所有要处理样本的二代测序数据.fq.gz格式文件放在同一个文件夹里，作为输入文件夹，运行以下命令，生成数据list：

python /home/licun/biosoft/bin/manifestGen.py -i input\_filedir -o output\_dir -F surfix\_of\_forward\_file -R surfix\_of\_ reverse \_file

manifestGen.py帮助文档：

-i FILEDIR, --input FILEDIR the path of the reads

-o OPDIR, --output OPDIR the output path of reads

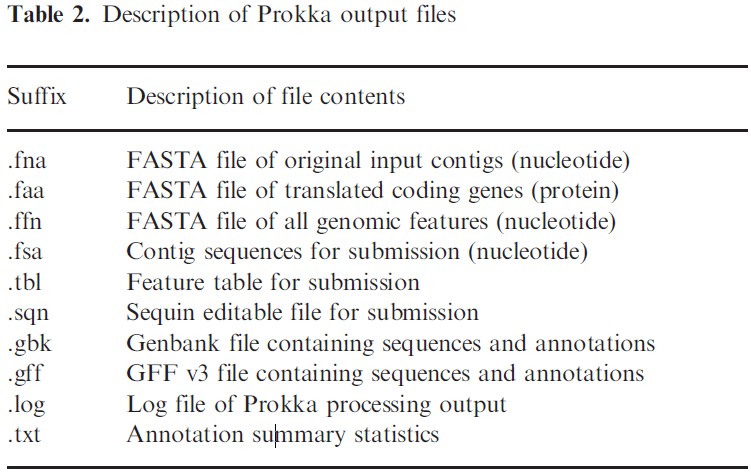
-F SP1, --sepF SP1 It is the surfix to recognize the forward info, default='\_1.clean.fq.gz'.

-R SP2, --sepR SP2 It is the surfix to recognize the reverse info, default='\_2.clean.fq.gz'.

## 3.3使用WGS-Pipeline.py脚本对数据进行质控、组装、评估、注释

python /home/licun/biosoft/bin//WGS-Pipeline.py -i PathTable.tsv -o outptudir -F surfix\_of\_forward\_file -R surfix\_of\_forward\_file

WGS-Pipeline.py整合了fastp、unicycler、quast、bandage、prokka等软件，在指定的输出结果路径下可以看到对应软件的分析结果文件夹。由于这一步包含了unicycler拼接组装，对于大批量数据此过程可能耗时较长，可调整-j -t参数调用更多计算资源。基因组注释结果可在生成的prokka文件夹中查阅，prokka软件对每个基因组的注释都会产生10个文件，其格式分别为：fna, faa, ffn, fsa, tbl, sqn, gbk, gff, log, txt. 具体内容如下图：



WGS-Pipeline.py帮助文档：

WGS Assembly & Annotation

optional arguments:

-h, --help show this help message and exit

-i INFILE, --input INFILE the path of the reads

-o OUTDIR, --output OUTDIR the output path of reads

-kd KRAKEN2\_DB, --database KRAKEN2\_DB the nr database path

-j JOBS, --jobs JOBS the number of jobs run in parallel

-t THREADS, --threads THREADS the number of threads run for a job

-F SP1, --sepF SP1 It is the surfix to recognize the forward info, default='\_1.clean.fq.gz'.

-R SP2, --sepR SP2 It is the surfix to recognize the reverse info, default='\_2.clean.fq.gz'.

## 3.4 CheckM对基因组完整度和污染度进行评估

使用CheckM lineage\_wf ，此流程包含其它几个模块：tree -> Place bins in the reference genome tree；lineage\_set -> Infer lineage-specific marker sets for each bin；analyze -> Identify marker genes in bins；qa -> Assess bins for contamination and completeness。

CheckM会把完整度和污染度等结果输出到屏幕，因此为了保存结果，建议将输出到屏幕的结果写入到文件中。

checkm lineage\_wf -t 8 -x fna bin folder output folder >checkm.out

根据结果对基因组进行筛选select only genomes that passed the following criteria: >50% genome completeness, <5% contamination and an estimated quality score (completeness – 5 × contamination) > 50。

## 3.5 GTDB-tk使用基因组数据确定物种分类地位

使用GTDB-tk的classify\_wf流程，该流程包括了identify鉴定单拷贝标记基因、align多序列对齐和classify物种分类鉴定三步。

conda activate /home/chenjunyu/miniconda3/envs/gtdbtk

gtdbtk classify\_wf --genome\_dir fna/ --out\_dir classify\_wf --extension fna --cpus 8

在输出的结果文件夹中会有gtdbtk.bac120.summary.tsv，gtdbtk.arc122.summary.tsv文件，分别对应着细菌和古菌分类地位。

注意：每个gtdbtk的环境，会固定对应某个gtdb的版本，注意不要弄混了。（因为该软件不支持另外指定数据库路径）

## 3.6 功能注释分析

### 泛功能注释

EggNOG数据库注释(包含了COG、KEGG、CAZy、Pfam和GO数据库的多层级内容)：

conda activate /home/licun/miniconda3/envs/eggnog/

emapper.py -m diamond --itype proteins -i sequence.faa -o prefix\_name --cpu 16 --excel --evalue 1e-5

除了diamond外，eggnog-mapper 同时也提供其它比对方式mmseqs, hmmer, no\_search, cache

输入数据格式也支持多种格式：CDS, proteins, genome, metagenome，其它详细用法见链接：<https://github.com/eggnogdb/eggnog-mapper/wiki/eggNOG-mapper-v2.1.5-to-v2.1.7>

Kofam-scan【待补充】

### 代谢基因

CAZy：碳水化合物活性酶注释

conda activate /home/chenjunyu/miniconda3/envs/anno

python /home/chenjunyu/Lab/Culturomics/scripts/run\_dbcan.py -i input\_dir -o dbcan\_out -t 8 -j 8

特定代谢基因（BileAcide, SCFA等）注释：

使用blastp比对KEGG数据库，根据结果匹配代谢基因KO号（后续需要进一步写脚本提取，这个是比较泛泛的分析方法）

diamond blastp --db /home/chenjunyu/databases/kegg/prokaryotes\_filter --query fasta --out OutDir prefix \_blasp.tsv --evalue 1e-05 --outfmt 6 --max-target-seqs 1 -threads 6

胆汁酸代谢基因建库：bai基因簇、BSH、7aHSDH、7bHSDH（子课题1）

diamond blastp --db /home/licun/Data/Blastp\_0829/ZCH/zch\_three\_strain/bile\_acid --query fasta --out OutDir prefix \_blasp.tsv --evalue 1e-05 --outfmt 6 --max-target-seqs 1 -threads 6

胆汁酸代谢基因建库：5AR、5BR、3bHSDH、3aHSDH（子课题2）

diamond blastp --db /home/licun/Data/Blastp\_0829/WGS\_LDlab/WGS3.0/muban\_seq/BileAcid --query fasta --out OutDir prefix \_blasp.tsv --evalue 1e-05 --outfmt 6 --max-target-seqs 1 -threads 6

也可以考虑把两个子课题合在一起建blastp数据库。

### 毒力基因

VFDB【待补充到github，实际上所有通过blastp进行注释的工具都通过这个流程】

### 抗性基因

RGI：抗生素抗性基因注释：

conda activate /home/chenjunyu/miniconda3/envs/rgi/

rgi main --input\_sequence sequence.faa --output\_file rgi.out --input\_type protein --alignment\_tool DIAMOND --clean -d wgs

主要涉及的参数设置：

--input\_sequence 必须是contig或者蛋白，fasta格式

--input\_type contig或者protein，默认是contig（和input的实际情况对应）

--alignment\_tool Blast或者DIAMOND，默认是Blast

--clean 是不保留中间文件

-d wgs 其他选择是plasmid,chromosome，一般我们都是用wgs

噬菌体抗性：PADLOC【待补充】

CRISPR：CRSPRDetect【待补充】（<https://github.com/ambarishbiswas/CRISPRDetect_2.2>）

参考脚本：[Gut-Metagenome-Pipeline-Based-on-Nanopore-Sequencing/4\_CRISPR\_spacers\_prediction.sh at main · chen318liang/Gut-Metagenome-Pipeline-Based-on-Nanopore-Sequencing (github.com)](https://github.com/chen318liang/Gut-Metagenome-Pipeline-Based-on-Nanopore-Sequencing/blob/main/scripts/4_CRISPR_spacers_prediction.sh)

### 水平转移相关

digIS转座酶(insertion sequence elements)注释：

conda activate /home/licun/miniconda3/digIS

python /home/licun/biosoft/digIS-digISv1.2/digIS\_search.py -i sequence.fna -g annotation.gbk -o digIS\_output

mgefinder【待补充】

<https://pypi.org/project/MobileElementFinder/>【待补充】

### 其他特性预测（如质粒、prophage、细胞内定位）：

质粒分析：plasforest

docker run -u $(id -u):$(id -g) --cpus=1 --rm -t -i -w / -v /data:/data dbest/plasforest:v1.0

cd PlasForest

mkdir /data/Xianjinyuan/tanyuxiang/1-projects/SZ\_child\_str/PlasForest\_17/

python3 PlasForest.py -i /data/Xianjinyuan/tanyuxiang/1-projects/SZ\_child\_str/str\_fa/17/BN-219.Draft.genome.fasta -o /data/Xianjinyuan/tanyuxiang/1-projects/SZ\_child\_str/PlasForest\_17/BN-219.csv

prophage预测：

conda activate /home/licun/miniconda3/prophage

PhiSpy.py -o outfile DA\*.gbk --output\_choice 512

蛋白的细胞内定位：psortb

转运蛋白（TCDB）【待补充】

### 基因簇预测：

antiSMASH：次级代谢基因簇（BGC）注释

conda activate /home/licun/miniconda3/antismash

antismash --cb-general --cb-knownclusters --cb-subclusters --asf --pfam2go --smcog-trees --genefinding-tool prodigal -c 12 DA\*.gbk

gutSMASH：肠道菌初级代谢基因簇（MGC）注释

conda activate /home/chenjunyu/miniconda3/envs/gutsmash

python run\_gutsmash.py --genefinding-tool prodigal --cb-knownclusters --enable-genefunctions fasta\_file\_input

还可以补充抗性基因的基因簇预测ARTS