1. Raw reads数据质控，输出：Clean reads: **fastp**
   1. 去adapter
   2. 去低质量碱基(q<20)
   3. 去短序列（<36bp）
   4. 去冗余序列
2. Clean reads去宿主，输出：unmapped clean reads：**bowtie2**

数据质控完成后，使用clean reads与宿主基因组（默认：人）进行比对，输出未必对上的clean reads

3-1．基于测序reads的物种分类：**kraken2**

去除宿主后的reads与NCBI RefSeq（archaea, bacteria, viral, plasmid, human

protozoa, fungi & plant等）物种基因组数据库进行比对，根据比对上的reads数目，来判定样本中可能包含的物种。

3-2．基于测序reads的物种分类：**metaphlan**

与kraken2分类方法不同，metaphlan使用的参考数据库更多是具有分类意义的基因marker组成，relies on ~7.3M unique clade-specific marker genes (the latest marker information file can be found [here](http://cmprod1.cibio.unitn.it/biobakery4/metaphlan_databases/mpa_vJun23_CHOCOPhlAnSGB_202307_marker_info.txt.bz2)) identified from >1.1M microbial genomes (>150k isolate genomes and >950k metagenome-assembled genomes) spanning 36,822 species-level genome bins 11,062 of them taxonomically unidentified at the species level (the full list of the species included in the latest database can be found [here](http://cmprod1.cibio.unitn.it/biobakery4/metaphlan_databases/mpa_vJun23_CHOCOPhlAnSGB_202307_species.txt.bz2)).

4．基因组组装：**metaSPAdes与MEGAHIT**

目前基于宏基因组测序的基因组拼接算法有很多，引用来自权威文献的比较结果，我们分别最终选择，**metaSPAdes**对单一样本进行组装拼接；**MEGAHIT**如果涉及到同一类型多个样本，选择**MEGAHIT**进行多样本合并组装拼接。并通过**MetaQUAST**对组装结果进行评估。

* *Breitwieser F P, Lu J, Salzberg S L. A review of methods and databases for metagenomic classification and assembly[J]. Briefings in bioinformatics, 2019, 20(4): 1125-1136.*
* *Zhang Z, Yang C, Veldsman W P, et al. Benchmarking genome assembly methods on metagenomic sequencing data[J]. Briefings in Bioinformatics, 2023, 24(2): bbad087.*

5．编码基因序列预测：**Prodigal**

6. 耐药基因序列片段预测: **staramr & CARD**

耐药基因的数据分析，采用2种不同的方式：a)基于测序reads比对的方式，这里我们使用的是CARD数据库；b)当采用双端测序，且样本在质控与去过宿主后数据量（>0.5M）时,使用staramr基于拼接的基因组片段序列进行耐药基因分析。

7. 毒力基因预测：**VFDB**

将步骤5中预测获得的编码序列，借助**BLASTP**与毒力基因蛋白质数据库进行比对，输出样本中肯能包含的毒力因子信息。

==================================================================

1. 宏基因组binning

在完成步骤4的基因组组装后，我们会得到一些较大的序列片段(contig or scaffold)序列，下一步我们需要对肯能来自同一物种的序列进行分类，这一过程可以简称为:binning.

本流程参考宏基因组研究文章，我们选取了三种binning分析算法：**metaBAT2, MaxBin2, CONCOCT**，然后使用**metaWRAP**算法将结果整合，以期得到最佳结果。

* *Banerjee G, Papri S R, Banerjee P. Protocol for the construction and functional profiling of metagenome-assembled genomes for microbiome analyses[J]. STAR protocols, 2024, 5(3): 103167.*

9. 单一bining完整性评估：**CheckM**

对每一个binning进行完整性评估，一般评价指标completeness (> 90%) and contamination (< 5%) for high-quality drafts of MAGs (metagenome-assembled genomes) ，则可认为是比较好的基因组序列。

|  |  |
| --- | --- |
| Criterion | Description |
| High-quality | Completeness > 90%, Contamination < 5%, presence of the 23S, 16S and 5S rRNA genes and at least 18 tRNAs |
| Near complete | Completeness > 90%, Contamination < 5% |
| Medium-quality | Completeness ≥ 50%, Contamination < 10% |
| Low-quality | Completeness < 50%, Contamination < 10% |

* *Jia L, Wu Y, Dong Y, et al. A survey on computational strategies for genome-resolved gut metagenomics[J]. Briefings in Bioinformatics, 2023, 24(3): bbad162.*

10．单一binnng分类信息: **GTDB-Tk**

对binning现有的基因组数据库进行比对，以获得对应的分类信息。