1.测序结果预处理

使用fastp软件进行原始数据质控，选用软件默认参数，对 Illumina HiSeq 测序平台获得的原始数据(Raw Data)进行预处理，获取用于后续分析的有效数据(Clean Data)。如果样本中存在污染，需与宿主数据库进行比对，过滤掉可能来源于宿主的 reads。选用Bowtie2软件，参数设置为：--sensitive, -I 200, -X 400。

2. Metagenome 组装

1. 经过预处理后得到 Clean Data，使用 MEGAHIT 组装软件进行组装分析(Assembly Analysis)；组装参数： --k-min 35 --k-max 95 --k-step 20 --min-contig-len 500；
2. 将各样品质控后的 CleanData 采用 Bowtie2 软件比对至各样品组装后的 contigs 上，获取未被利用上的 PE reads；比对参数： -I 200, -X 400；
3. 将各样品未被利用上的 reads 放在一起，进行混合组装，组装参数与单样品组装参数相同；

3.基因预测及丰度分析

1. 从各样品及混合组装的 contigs（>=500bp）出发，采用 MetaGeneMark进行 ORF (Open Reading Frame) 预测，采用默认参数；
2. 从预测结果出发，过滤掉长度小于 100nt的预测基因；
3. 合并各样品及混合组装的 ORF 预测结果，利用 CD-HIT软件进行去冗余，以获得非冗余的初始 gene catalogue（此处，操作上，将非冗余的连续基因编码的核酸序列称之为 genes），默认以 identity 95%, coverage 90% 进行聚类，并选取最长的序列为代表性序列；采用参数：-c 0.95, -G 0, -aS 0.9, -g 1, -d 0；
4. 采用 Bowtie2，将各样品的 Clean Data 比对至初始 gene catalogue，计算得到基因在各样品中比对上的 reads 数目；比对参数：--end-to-end, --sensitive, -I 200, -X 400
5. 过滤掉在各个样品中支持 reads 数目<=2 的基因，获得最终用于后续分析的 gene catalogue（Unigenes）；
6. 从比对上的 reads 数目及基因长度出发，计算得到各基因在各样品中的丰度信息，计算公式如下所示；r 为比对上基因的 reads 数目，L 为基因的长度：

$$G_k = \frac{r_k}{L_k} \cdot \frac{1}{\sum_{i=1}^{n} \frac{r_i}{L_i}}$$

4. 物种注释

1. 使用genes 与物种数据库进行比对：DIAMOND软件将 Unigenes 与从 NCBI 的NR数据库中抽提出的细菌(Bacteria)、真菌(Fungi)、古菌(Archaea)和病毒(Viruses)序列进行比对（blastp，evalue <= 1e-5）；

2. LCA算法：由于每一条序列可能会有多个比对结果，得到多个不同的物种分类信息，为了保证其生物意义，利用 MEGAN软件的LCA算法，得到该序列最终的物种注释信息；

3. 从 LCA 注释结果及基因丰度表出发，获得各个样品在各个分类层级（界门纲目科属种）上的丰度信息，对于某个物种在某个样品中的丰度，等于注释为该物种的基因丰度的加和；

4. 从 LCA 注释结果及基因丰度表出发，获得各个样品在各个分类层级（界门纲目科属种）上的基因数目表，对于某个物种在某个样品中的基因数目，等于在注释为该物种的基因中，丰度不为 0 的基因数目；

5. 从各个分类层级（界门纲目科属种）上的丰度表出发，进行 Krona 分析，相对丰度概况展示，丰度聚类热图展示，PCA 和 NMDS 降维分析，Anosim组间（内）差异分析，组间差异物种的Metastat和LEfSe多元统计分析；

5. 常用功能数据库注释

1. 序列比对：使用DIAMOND软件将 Unigenes 与各功能数据库进行比对（blastp，evalue <= 1e-5）；

2. 比对结果过滤：对于每一条序列的 比对结果，DIAMOND 参数--max-target-seqs 1，保留唯一比对结果；

3. 从比对结果出发，统计不同功能层级的相对丰度（各功能层级的相对丰度等于注释为该功能层级的基因的相对丰度之和），其中，KEGG 数据库划分为 6 个层级，eggNOG 数据库划分为 3 个层级， CAZy 数据库划分为 3 个层级，

4. 从功能注释结果及基因丰度表出发，获得各个样品在各个分类层级上的基因数目表,对于某个功能在某个样品中的基因数目，等于在注释为该功能的基因中，丰度不为 0 的 基因数目。

5. 从各个分类层级上的丰度表出发，进行注释基因数目统计，相对丰度概况展示，丰度聚类热图展示，PCA 和 NMDS 降维分析，基于功能丰度的 Anosim 组间（内）差异分析，代谢通路比较分析，组间功能差异的 Metastat 和 LEfSe 分析。

6. 抗性基因注释

1. 使用 CARD 数据库提供的 Resistance Gene Identifier （RGI）软件将 Unigenes 与 CARD 数据库（https://card.mcmaster.ca/）进行比对（RGI 内置 blastp，默认

evalue < 1e-30）；

2. 根据 RGI 的比对结果，结合 Unigenes 的丰度信息，统计出各 ARO 的相对丰度；

3. 从 ARO 的丰度出发，进行丰度柱形图展示，丰度聚类热图展示，丰度分布圈图展示，组间 ARO 差异分析，抗性基因（注释到 ARO 的 unigenes）及抗性机制物种归属分析等（对部分名称较长的 ARO,用其前三个单词与下划线缩写的形式展示）。