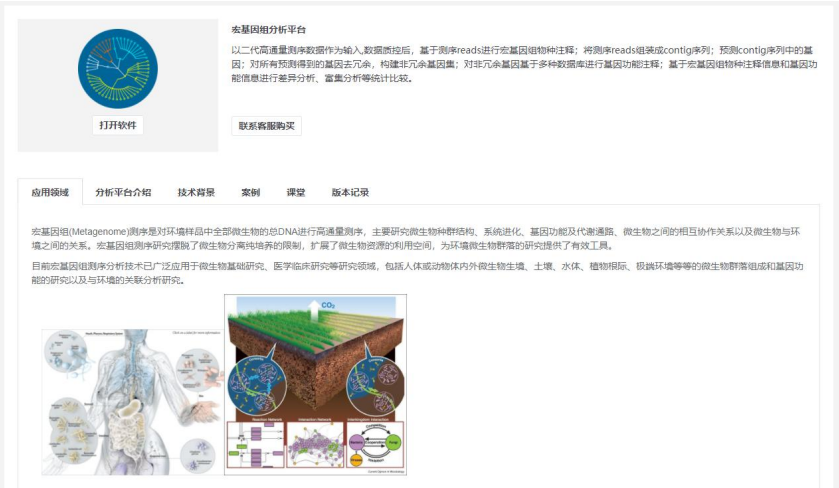
二代宏基因组分析流程参数



1 背景介绍

二代宏基因组测序是基于 Illumina 高通量测序，利用双末端测序（Paired-End）的方法， 构建小片段文库进行测序。通过生物信息分析获得环境中的基因功能信息和物种组成信息。通过基因功能的角度分析环境样品所具有的功能多样性；通过物种的角度分析环境样品中含有的物种的多样性。

2 信息分析流程

对测序得到的原始 reads 进行质量控制并过滤得到 Clean reads，用于后续生物信息学的分析。对 Clean reads 进行拼接组装、预测编码基因，并对编码基因进行通用数据库和专用数据库的功能注释；同时，对 Clean reads 进行分类学分析，统计样品物种组成和丰度信息，依托 BMKCloud 完成分析结题报告。

3 基本分析结果

3.1 测序数据质控

测序得到的原始序列(Raw reads)里面含有低质量的序列，为了保证信息分析质量，需要对 Raw reads 进行过滤，得到 Clean reads，用于后续信息分析。数据过滤的主要步骤如下：

1) 使用软件 Trimmomatic（版本 v0.33 参数 PE LEADING:3 TRAILING:3 SLIDINGWINDOW:50:20 MINLEN:120）软件，对 Raw Tags 进行过滤，得到高质量的测序数据(Clean Tags)；

2) 使用 bowtie2（版本 v2.2.4 参数--seed 123456 -I 200 -X 1000 --un-conc）同宿主基因组序列比对，去除宿主污染（如果提供了宿主参考基因组）。

3.2 宏基因组组装

使用软件 MEGAHIT[1]（版本 v1.1.2 默认参数）进行宏基因组组装，过滤短于 300bp 的 contig 序列。

采用 QUAST[2] （版本 v2.3 默认参数）软件对组装结果进行评估。

3.3 宏基因组组分分析

3.3.1 基因预测

采用 MetaGeneMark[3]（版本 v.3.26 默认参数）软件(http://exon.gatech.edu/meta_gmhmmp.cgi，Version 3.26），使用默认参数（参数-A -D -f G）来识别基因组中的编码区域。

3.4 非冗余基因集的构建

使用 MMseq2 软件（https://github.com/soedinglab/mmseqs2, Version 11-e1a1c）去除冗余，相似性阈值设置为 95%，覆盖度阈值设置为 90%。

3.5 功能注释分析

3.5.1 Nr 数据库注释

Nr[5]数据库的全称是 Non-Redundant Protein Database，是一个非冗余的蛋白质数据库,由 NCBI 创建并维护,该数据库含有全面的蛋白序列和注释信息，而且在注释信息中存在相应的物种信息。与使用其他数据库相比，使用 Nr 数据库一般能使基因组中更多的基因具有注释信息，即具有最高的注释比例。但该数据库中很多蛋白序列和注释信息未经过验证,可靠性有待提高。

具体注释时，通过将非冗余基因的蛋白序列和 Nr 数据库进行 BLAST 比对（diamond v0.9.29.130 比对筛选阈值 E-value 1e-5），找到在 Nr 数据库中最相似的序列，该序列对应的注释信息即为对应测序基因组基因的注释信息。

3.5.2 GO 数据库注释

GO[6](gene ontology)是基因本体联合会(Gene Onotology Consortium)所建立的数据库，旨在建立一个适用于各种物种的，对基因和蛋白质功能进行限定和描述的，并能随着研究不断深入而更新的语义词汇标准。

目前基因的相同功能在不同的功能数据库可能会使用不同的术语，好比是不同的方言。这样导致使用多种数据库注释时的分歧，不利于功能注释的长期发展和使用。Gene Ontology 就是为了解决上述问题而建立的，使各种数据库中基因产物功能描述相一致而发起的一个项目。

GO 数据库对于生物功能逐渐深入的倒树根形结构，最高级别功能节点包括三个：（1）细胞组分（Cellular Component）：用于描述亚细胞结构、位置和大分子复合物；（2）分子功能（Molecular Function）：用于描述基因、基因产物个体的功能；（3）生物过程（ Biological Process）：用来描述基因编码的产物所参与的生物过程。

在具体注释时，基于前面 Pfam 数据库注释结果，其中比对到 Pfam 中的基因对应于 GO 数据库中包含功能注释信息的节点上，这些节点对应的注释信息即为测序基因组中基因的注释信息。

3.5.3 kegg 功能注释

kegg[8]（Kyoto Encyclopedia of Genes and Genomes）是收集了生物的基因组、通路和化合物信息的综合性的数据库。该数据库对收录的序列进行分析聚类分析，形成直系同源蛋白群，对于不同的蛋白群指定不同的 KO 序列。根据已发表文献，kegg 人工绘制了大量的生物过程图（如代谢途径、信号传导途径等），在图中的特殊的矩形框或线条对应发挥相应功能的某种 KO 序号的蛋白群，通过这些生物过程图更能直观地体现生物的生命过程。

kegg（Kyoto Encyclopedia of Genes and Genomes）是收集基因组、生物通路、疾病、药物和化学物质的数据库。在 kegg 中有一个“专有名词”KO（kegg Orthology），它是蛋白质（酶）的一个分类体系，将序列高度相似并且在同一条通路上有相似功能的蛋白质序列归为一组，然后打上 KO 标签，对应每个 KO 标签添加相应的功能注释信息。而这些 KO 标签对应到基因组在 kegg 中对于许多代谢通路或者细胞过程。

具体注释时,通过将非冗余基因的蛋白序列和 kegg 数据库中收录的蛋白序列进行 BLAST 比对（diamond v0.9.29 比对筛选阈值 E-value 1e-5），找到在 kegg 数据库中最相似的序列，该序列的注释信息、对应的 KO 号、KO 对应的通路中的位置即为测序基因组中对应的基因的注释信息、KO 号以及在生物过程通路中的位置该序列对应具有 KO 序号。

3.5.4 eggNOG 数据库注释

eggNOG[9] 是收录生物直系同源基因簇的数据库，是在 COG 数据库的基础上的持续更新。构成每个直系同源基因簇(Cluster of Orthologous Groups)的蛋白都是被假定为来自于一个祖先蛋白，具有相同的功能。该数据库是通过比较完整基因组的蛋白序列而产生的。该数据库常被用来对新测序基因组的基因进行分类和注释。

具体注释时，通过将非冗余基因的蛋白序列和 eggNOG 数据库进行 BLAST 比对（diamond v0.9.29 比对筛选阈值 E-value 1e-5），找到在 eggNOG 数据库中最相似的序列，该序列对应的注释信息和分类信息即为对应测序基因组基因的注释信息和分类信息。

3.5.5 Pfam 数据库注释

Pfam[10] 数据库是一种包含注释信息和多序列比对信息的蛋白家族数据库，其中的多序列比对信息是由隐马尔科夫模型产生。该数据库提供了较为完整和精确的蛋白家族和功能域的分类信息。不像通常进行的 BLAST 搜索，Pfam 使用软件 MMseqs2（版本 11-e1a1c 比对筛选阈值 E-value 1e-5）进行未知功能序列和 Pfam 数据库的比较。

3.5.6 SwissProt 数据库注释

SwissProt[11] 数据库是一个人工注释的非冗余高质量蛋白序列数据库，其特点是注释结果有相应实验验证，可靠性较高。

在注释时，通过将非冗余基因的蛋白序列分别和 SWISS-PROT 和 TrEMBL 数据库中收录的蛋白序列进行 BLAST 比对（diamond v0.9.29 比对筛选阈值 E-value 1e-5），找到在两个数据库中最相似的序列，该序列对应的注释信息即为测序基因组中对应的基因的注释信息。

3.5.7 CAZy 数据库注释

CAZy[12]（Carbohydrate-active enzymes database）是收录碳水化合物活性酶的数据库，该数据库由已发表文献和相关蛋白的收集和分类形成，并由专家精心维护。该数据库中蛋白分为五大类功能类别：糖苷水解酶（glycoside hydrolases，GHs）、糖基转移酶（glycosyltransferases，GTs）、多糖裂解酶（polysaccharide lyases，PLs）、碳水化合物酯酶（carbohydrate esterases，CEs)和非催化的结合碳水化合物的功能域（CBMs）。上述每一种大的类别中都含有许多不同的家族。不过每个家族中都是由蛋白序列组成的，不能突出该家族共有的特征性的功能域，而细胞内的碳水化合物活性酶一般具有许多的结构域。dbCAN 具体分析 CAZy 数据库收集的每一个家族，并对每一个家族构建该家族特征性结构域的隐马尔可夫模型。软件 hmmer 使用这些隐马尔可夫模型可以识别出属于特定家族的保守功能域。

在具体注释时，使用 hmmer（版本 3.0）软件对于通过将非冗余基因的蛋白序列分别与 CAZy 数据库的每一个家族的隐马尔可夫模型比对（比对参数 默认，默认筛选阈值"if alignment > 80aa, use E-value < 1e-5, otherwise use E-value < 1e-3; covered fraction of HMM > 0.3"），找出所有满足过滤阈值的家族，这样可以找出基因组中的碳水化合物活性酶，也可以分析出它们的含有几个保守碳水化合物相关功能域。

3.5.8 CARD 数据库注释

CARD[13](Comprehensive Antibiotic Research Database)是一个持续更新的抗生素抗性相关信息的数据库。CARD 包含描述抗生素和它们的靶标，涉及抗生素抗性基因、相关蛋白、抗生素抗性机制信息。在 CARD 的核心是一个高度开发的抗生素抗性本体（Antibiotic Resistance Ontology (ARO)），用于抗生素抗性基因数据的分类。

在注释时，使用 CARD 数据库中的工具软件 rgi（版本 4.2.2 默认 Perfect, Strict 算法）将非冗余基因的蛋白序列分别数据库进行比对，找出比对上的数据库中对应序列，并得到对应的抗性基因和抗性相关信息。

3.5.9 VFDB 数据库注释

VFDB[14]（virulence factor database）是收集了已经充分描述的细菌致病菌的绝大多数的致病因子，这些致病因子帮助致病菌建立对宿主的感染、应对宿主的免疫系统在宿主中存活并引起宿主的疾病。该数据库主要包括两种毒力因子序列数据集，一个是已经实验确证的核心数据集 sets A，另一个是既包含 sets A 中的已实验确认的毒力因子也包含预测出的毒力因子的全数据集 set B。这个数据库可以帮助预测尚未详细研究的细菌基因组中可能的毒力因子。

在具体注释时，通过将非冗余基因的蛋白序列 Blast 比对（BLAST 2.2.31+，比对参数默认，筛选阈值 E-value 1e-5）核心数据集 sets A，在数据库中找到最相似的序列，该序列携带的相关信息即为测序基因组中对应基因的注释信息。虽然可能并不是明显的致病菌，但是，仍可以通过这个数据库注释获得相应基因可能的功能信息，为后续研究提供线索。

3.5.10 PHI-base 数据库注释

PHI-base[15]（Pathogen Host Interactions Database）是病原与宿主互作数据库，其中收录的基因经过实验验证，主要来源于真菌、卵菌和细菌病原，感染的宿主包括动物、植物、真菌以及昆虫。该数据库含有致病菌感染宿主过程中预测的相关蛋白的详细描述。

在具体注释时，通过将非冗余基因的蛋白序列和 PHI 数据库中收录的蛋白序列进行 Blast 比对（BLAST 2.2.31+，比对参数默认，筛选阈值 E-value 1e-5），找到在该数据库中最相似的序列，该序列对应的注释信息和分类信息即为测序基因组中对应的基因的注释信息。

3.5.11 CYPED 数据库注释

细胞色素 P450 单氧化酶（CYPs）是一种亚铁血红素，包含在许多微生物、植物、动物和人类的生理上代谢重要的物质。CYPs 催化在生物合成和生物降解通路中广泛的内源性物质的氧化，这些通路包括诸如药物和环境污染物的异源物质（xenobiotics）。因此，理解人类 CYPs 的特异性底物在药物开发中至关重要。CYPED[16](Cytochrome P450 Engineering Database)数据库是一个包含 CYPs 的序列、序列比对、注释和结构的信息的细胞色素 P450 数据库。对应的数据，序列、结构和注释信息从 GenBank 和 PDB 中提取出来，并被手动地确认。

在具体注释时，通过将非冗余基因的蛋白序列和 Cytochrome P450 Engineering Database 数据库中收录的蛋白序列进行 Blast 比对（BLAST 2.2.31+，比对参数默认，筛选阈值 E-value 1e-5），找到在该数据库中最相似的序列，从而得到对应的 P450 家族分类信息。

4 功能多样性分析

4.1 功能基因组成及丰度分析

4.1.1 碳水化合物酶组成分析

CAZy 功能注释中，计算各样品注释到的 Class 和 Family 个数并进行统计

4.1.2 CARD 功能基因组成分析

CARD 功能注释中，计算各样品注释到的 ARO 和 Resistance 个数并进行统计

4.2 功能基因的 Beta 多样性分析

4.2.1 功能基因的主成分分析

主成分分析(Principal Component Analysis, PCA)是一种分析和简化数据集的技术，通过将方差进行分解，将多组数据的差异反映在二维坐标图上，坐标轴取能够最大反映方差的两个特征值。通过分析不同样品功能基因组成可以反映样品间的差异和距离。PCA 图上两个样品距离越近，则表示这两个样品中功能基因的组成越相似。使用 R 语言工具分别绘制不同水平的 PCA 分析图

4.2.2 功能基因的 PCoA 分析

主坐标分析法[17](Principal coordinates analysis,PCoA)是一种与 PCA 类似的降维排序方法，原理是假设对 N 个样品有衡量它们之间差异或距离的数据，就可以用此方法找出一个直角坐标系，将 N 个样品表示成 N 个点，而使点间的欧式距离的平方正好等于原来的差异数据，实现定性数据的定量转换，从多维数据中提取出最主要的元素和结构。通过主坐标分析可以实现多个样品的分类，进一步展示样品间功能基因多样性差异。

基于 Beta 多样性分析得到的两种距离矩阵，使用 R 语言工具分别绘制的 PCoA 分析结果如下图：坐标图上距离越近的样品，功能组成相似性越大。

4.2.3 功能基因的 NMDS 分析

非度量多维标定法[18](Non-MetricMulti-Dimensional Scaling,NMDS)是一种适用于生态学研究的排序方法，主要是将多维空间的研究对象（样本或变量）简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。类似于 PCA 或者 PCoA，通过样本的分布可以看出组间或组内差异。NMDS 原设计的目的是为了克服以前排序方法中包括 PCA、PCoA 在内的缺点，即线性模型。NMDS 的模型是非线性的，能更好地反映生态学数据的非线性结构，有的研究认为 NMDS 的效果优于 PCA/PCoA。建议不分组时，样本数量不少于 10 个；多组样本时，每组样本数量不少于 5 个。分析软件版本参数 python 分析 R 语言绘图

4.2.4 功能基因的 UPGMA 分析

UPGMA（Unweighted Pair-group Method with Arithmetic Mean）即：非加权组平均法，也可理解为样品层次聚类，是一种常用的聚类分析方法。

基于 Beta 多样性分析得到的两种距离矩阵，通过 R 语言工具采用非加权配对平均法（UPGMA）对样品进行层次聚类，以判断各样品间功能基因组成的相似性。样品层次聚类树：样品越靠近，枝长越短，说明两个样品的功能基因组成越相似。

4.3 功能基因的组间差异分析

4.3.1 功能基因的组间差异参数检验

常用的统计方法分为两种：参数统计和非参数统计方法。参数统计是指总体分布类型已知，用样本指标对总体参数进行推断或作假设检验的统计分析方法；非参数统计是指不考虑总体分布类型是否已知，不比较总体参数，只比较总体分布的位置是否相同的统计方法。

Welch's t-test 即 t′检验，用于两组样本的差异检验。ANOVA（Analysis of Variance，即：方差分析）又称 F 检验（F test）。主要用于检验多个样本的均值是否相等，从而判断两个或两个以上样本的均值差异显著程度。

这里当分组个数为两组时，使用 Welch's t-test 进行组间差异参数检验；如果超过两组，使用 ANOVA 分析进行组间差异参数检验。

4.3.2 功能基因的 metagenomeSeq 分析

metagenomeSeq[19]【v1.22.0】是一个用于进行差异丰度基因或物种检验的 R 包，该软件应用了一个新的标准化(normalization)方法，这样可以避免不均匀测序深度的影响，还应用一种零膨胀高斯分布混合模型，针对解决因为抽样不足引起的丰度差异对检验的影响。

4.4 功能基因的关联及相关性分析

4.4.1 功能基因的相关性分析

相关性网络图是相关性分析的一种表现形式，筛选丰度最高的 80 个功能基因，根据各个功能基因在各个样品中的丰度以及变化情况，使用 spearman 算法进行相关分析（包括正相关和负相关）并进行统计检验，筛选相关性大于 0.5 且 p 值小于 0.05 的数据组，基于 python 绘制相关性网络图。

4.4.2 功能基因的 RDA/CCA 分析

RDA(Redundancy analysis)/CCA(Canonical Correspondence analysis)[20]是基于对应分析发展的一种排序方法，RDA 分析基于线性模型，CCA 分析基于单峰模型，主要用来反映功能或样品与环境因子之间的关系。根据功能分布变化，选择最佳的分析模型。

RDA 或 CCA 模型的选择原则：先用 function-sample 数据做 DCA 分析，看分析结果中 Lengths of gradient 的第一轴的大小，如果大于 4.0，就应该选 CCA，如果 3.0-4.0 之间，选 RDA 和 CCA 均可，如果小于 3.0，RDA 的结果要好于 CCA。

使用 R 语言 vegan 包中 rda 或者 cca 分析和作图。样品间功能多样性 RDA/CCA 分析结果如下：图中元素点与点之间的关系由距离表示，距离越近代表样品功能相近；射线与射线之间的 关系由夹角表示，钝角代表负相关，锐角代表正相关。

4.4.3 功能基因与环境因子相关性热图

基于 R 语言可以进行功能基因丰度和环境因子之间的相关性分析，可以分析出环境因子和功能基因之间的相关性关系。热图中方格中的*表示差异显著（p 小于 0.05） **表示差异极显著（p 小于 0.01）。

5 物种多样性分析

5.1 物种组成及丰度分析

5.1.1 物种注释及统计

根据上述非冗余基因比对到 Nr 中的序列的物种信息，得到样品的物种组成和相对丰度信息。

5.1.2 物种组成柱状图

使用 Python 绘制在界、门、纲、目、科、属、种分类学水平上的物种柱状图，从图中可以直观看出各样品的物种组成及不同物种在各样品中所占的比例。

5.2 物种的 Beta 多样性分析

5.2.1 物种组成的主成分分析

主成分分析(Principal Component Analysis, PCA)是一种分析和简化数据集的技术，通过将方差进行分解，将多组数据的差异反映在二维坐标图上，坐标轴取能够最大反映方差的两个特征值。通过分析不同样品物种组成可以反映样品间的差异和距离。PCA 图上两个样品距离越近，则表示这两个样品中物种的组成越相似。使用 R 语言工具分别绘制不同分类水平 PCA 分析图。

5.2.2 物种组成的 PCoA 分析

主坐标分析法[21](Principal coordinates analysis,PCoA)是一种与 PCA 类似的降维排序方法，原理是假设对 N 个样品有衡量它们之间差异或距离的数据，就可以用此方法找出一个直角坐标系，将 N 个样品表示成 N 个点，而使点间的欧式距离的平方正好等于原来的差异数据，实现定性数据的定量转换，从多维数据中提取出最主要的元素和结构。通过主坐标分析可以实现多个样品的分类，进一步展示样品间物种多样性差异。

基于 Beta 多样性分析得到的四种距离矩阵，使用 R 语言工具分别绘制的 PCoA 分析结果图：坐标图上距离越近的样品，相似性越大。

5.2.3 物种组成的 NMDS 分析

非度量多维标定法[22](Non-MetricMulti-Dimensional Scaling,NMDS)是一种适用于生态学研究的排序方法，主要是将多维空间的研究对象（样本或变量）简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。类似于 PCA 或者 PCoA，通过样本的分布可以看出组间或组内差异。NMDS 原设计的目的是为了克服以前排序方法中包括 PCA、PCoA 在内的缺点，即线性模型。NMDS 的模型是非线性的，能更好地反映生态学数据的非线性结构，有的研究认为 NMDS 的效果优于 PCA/PCoA。建议不分组时，样本数量不少于 10 个；多组样本时，每组样本数量不少于 5 个。

5.2.4 物种组成的 UPGMA 分析

UPGMA（Unweighted Pair-group Method with Arithmetic Mean）即：非加权组平均法，也可理解为样品层次聚类，是一种常用的聚类分析方法。其原理是：假定的前条件是在进化过程中，每一世系发生趋异的次数相同，即核苷酸或氨基酸的替换速率是均等且恒定的。通过 UPGMA 法所产生的系统发生树可以说是物种树的简单体现，在每一次趋异发生后，从共同祖先节点到 2 个 OTU 间的支的长度一样。因此，这种方法较多地用于物种树的重建。

基于 Beta 多样性分析得到的四种距离矩阵，通过 R 语言工具采用非加权配对平均法（UPGMA）对样品进行层次聚类，以判断各样品间物种组成的相似性。样品层次聚类树图：样品越靠近，枝长越短，说明两个样品的物种组成越相似。

5.3 物种组成的组间差异分析

5.3.1 物种组成的组间差异参数检验

常用的统计方法分为两种：参数统计和非参数统计方法。参数统计是指总体分布类型已知，用样本指标对总体参数进行推断或作假设检验的统计分析方法；非参数统计是指不考虑总体分布类型是否已知，不比较总体参数，只比较总体分布的位置是否相同的统计方法。

Welch's t-test 即 t′检验，用于两组样本的差异检验。ANOVA（Analysis of Variance，即：方差分析）又称 F 检验（F test）。主要用于检验多个样本的均值是否相等，从而判断两个或两个以上样本的均值差异显著程度。

这里当分组个数为两组时，使用 Welch's t-test 进行组间差异参数检验；如果超过两组，使用 ANOVA 分析进行组间差异参数检验。

5.3.2 metagenomeSeq 分析

metagenomeSeq[23]是一个用于进行差异丰度基因或物种检验的 R 包，该软件应用了一个新的标准化(normalization)方法，这样可以避免不均匀测序深度的影响，还应用一种零膨胀高斯分布混合模型，针对解决因为抽样不足引起的丰度差异对检验的影响。

5.3.3 物种 lefse 分析

lefse[24] (Line Discriminant Analysis (LDA) Effect Size)（lefse python） 能够在不同组间寻找具有统计学差异的 Biomarker。

5.3.4 物种随机森林分析

随机森林分析(Random Forest 分析)（R 语言），属于机器学习算法，是一个包含多棵决策树的分类器,可以基于原有样品分类高效快速挑选出对分类最为重要的物种类别（biomarker）。

5.4 关联及相关性分析

5.4.1 物种相关性分析

相关性网络图是相关性分析的一种表现形式，筛选丰度最高的 80 个物种，根据各个物种在各个样品中的丰度以及变化情况，使用 spearman 算法进行相关分析（包括正相关和负相关）并进行统计检验，筛选相关性大于 0.5 且 p 值小于 0.05 的数据组，基于 python 绘制相关性网络图。

5.4.2 物种 RDA/CCA 分析

RDA(Redundancy analysis)/CCA(Canonical Correspondence analysis)[25]是基于对应分析发展的一种排序方法，RDA 分析基于线性模型，CCA 分析基于单峰模型，主要用来反映菌群或样品与环境因子之间的关系。根据物种分布变化，选择最佳的分析模型。

RDA 或 CCA 模型的选择原则：先用 species-sample 数据做 DCA 分析，看分析结果中 Lengths of gradient 的第一轴的大小，如果大于 4.0，就应该选 CCA，如果 3.0-4.0 之间，选 RDA 和 CCA 均可，如果小于 3.0，RDA 的结果要好于 CCA。

使用 R 语言 vegan 包中 rda 或者 cca 分析和作图。属分类学水平样品间物种多样性 RDA/CCA 分析结果如下：图中元素点与点之间的关系由距离表示，距离越近代表样品组成相近；射线与射线之间的 关系由夹角表示，钝角代表负相关，锐角代表正相关。

5.4.3 物种与环境因子相关性热图

基于 R 可以进行物种丰度和环境因子之间的相关性分析，可以分析出环境因子和物种之间的相关性关系。基于相关性关系绘制热图。

参考文献

1. Li D, Liu C M, Luo R, et al. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph[J]. Bioinformatics, 2015, 31(10):1674-1676.
2. Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies[J]. Bioinformatics, 2013, 29(8):1072-1075.
3. Zhu W, Lomsadze A, Borodovsky M. Ab initio gene identification in metagenomic sequences. Nucleic acids research. 2010;38(12):132–132.
4. Fu L., et al. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics, 2012;28(23): 3150-3152.
5. Deng Y Y, Li J Q, Wu S F, et al. Integrated nr database in protein annotation system and its localization[J]. Comput Eng, 2006, 32(5): 71-74.
6. Ashburner M, Ball C A, Blake J A, et al. Gene Ontology: tool for the unification of biology[J]. Nature genetics, 2000, 25(1): 25-29.
7. Conesa A, Götz S, García-Gómez J M, et al. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research[J]. Bioinformatics, 2005,21(18), 3674-3676.
8. Kanehisa M, Goto S, Kawashima S, et al. The kegg resource for deciphering the genome[J]. Nucleic acids research, 2004, 32(suppl 1): D277-D280.
9. Powell S , Forslund K , Szklarczyk D , et al. eggNOG v4.0: nested orthology inference across 3686 organisms[J]. Nucleic Acids Research, 2014, 42(Database issue):231-9.
10. Finn R D , Bateman A , Clements J , et al. Pfam: the protein families database.[J]. Nucleic Acids Research, 2014, 42(Database issue):222-30.
11. Bairoch A , Apweiler R . The SWISS-PROT protein sequence data bank and its new supplement TREMBL.[J]. Nucleic Acids Research, 1996, 24(1):21.
12. Cantarel B L, Coutinho P M, Rancurel C, et al. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics[J]. Nucleic acids research, 2009, 37(suppl 1): D233-D238.
13. Jia B, Raphenya A R, Alcock B, et al. CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database[J]. Nucleic acids research, 2016: gkw1004.
14. Chen L, Yang J, Yu J, et al. VFDB: a reference database for bacterial virulence factors[J]. Nucleic acids research, 2005, 33(suppl 1): D325-D328.

15. Winnenburg R, Baldwin T K, Urban M, et al. PHI-base: a new database for pathogen host interactions[J]. Nucleic acids research, 2006, 34(suppl 1): D459-D464.
16. Fischer M , Knoll M , Sirim D , et al. The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family[J]. Bioinformatics, 2007, 23(15):2015-2017.
17. Sakaki T, Takeshima T, Tominaga M, Hashimoto H, Kawaguchi S: Recurrence of ICA-PCoA aneurysms after neck clipping. Journal of neurosurgery 1994, 80(1):58-63.
18. Looft T, Johnson TA, Allen HK, et al. In-feed antibiotic effects on the swine intestinal microbiome. Proceedings of the National Academy of Sciences 2012, 109(5): 1691-1696.
19. Paulson J N, Stine O C, Bravo H C, et al. Differential abundance analysis for microbial marker-gene surveys[J]. Nature Methods, 2013, 10(12):1200.
20. Lindström ES, Kamst-Van Agterveld MP, Zwart G: Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. Applied and environmental microbiology 2005, 71(12):8201-8206.
21. Sakaki T, Takeshima T, Tominaga M, Hashimoto H, Kawaguchi S: Recurrence of ICA-PCoA aneurysms after neck clipping. Journal of neurosurgery 1994, 80(1):58-63.
22. Looft T, Johnson TA, Allen HK, et al. In-feed antibiotic effects on the swine intestinal microbiome. Proceedings of the National Academy of Sciences 2012, 109(5): 1691-1696.
23. Paulson J N, Stine O C, Bravo H C, et al. Differential abundance analysis for microbial marker-gene surveys[J]. Nature Methods, 2013, 10(12):1200.
24. Segata N, Izard J, Waldron L, et al. Metagenomic biomarker discovery and explanation[J]. Genome Biology, 2011, 12(6):R60.
25. Lindström ES, Kamst-Van Agterveld MP, Zwart G: Distribution of typical freshwater bacterial groups is associated with pH, temperature, and lake water retention time. Applied and environmental microbiology 2005, 71(12):8201-8206.