# 宏基因组分析方法

# 宏基因组数据的质量控制和预处理

使用 FastQC 工具 [[17]](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0085) 评估原始测序读长的质量，所有样品都通过了常用的质量标准，包括[碱基](https://www.sciencedirect.com/topics/pharmacology-toxicology-and-pharmaceutical-science/base" \o "Learn more about base from ScienceDirect's AI-generated Topic Pages)质量、[GC 含量](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/gc-content" \o "Learn more about GC content from ScienceDirect's AI-generated Topic Pages)、核苷酸分布和重复率（平均 Phred 评分 >20;读长 >100 bp;核苷酸分布 %A = %T 和 %G = %C，在读长长度上比例恒定;[重复的序列](https://www.sciencedirect.com/topics/medicine-and-dentistry/segmental-duplication" \o "Learn more about duplicated sequences from ScienceDirect's AI-generated Topic Pages)应占总读取数的 <50%）。使用 Trimmomatic [[18]](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0090) 修剪掉低质量的读数和接头序列，然后使用 BowTie2 [19] 提取映射到人类基因组组装（GRch38 版本，从 NCBI GenBank 下载）的所有读数来去除人类 DNA 序列。其余读数随后用于表征口腔和粪便样本中的微生物组和[抵抗组](https://www.sciencedirect.com/topics/immunology-and-microbiology/resistome" \o "Learn more about resistome from ScienceDirect's AI-generated Topic Pages)。

# 分类学分析

使用 Kraken v.1 软件对成对的非宿主读数进行分类并分配分类标记 [[20]。](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0100)读数 （k-mer） 与预构建的 MiniKraken DB\_8GB 数据库进行比对，该数据库包含 NCBI 的 RefSeq（2017 年 10 月 18 日）中的完整古细菌、细菌和病毒基因组。Kraken 为给定宏[基因组](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metagenomics" \o "Learn more about metagenomic from ScienceDirect's AI-generated Topic Pages)读数集中的每个读数分配最特异性的分类标签进行分类。但是，有时标签分配不在物种级别。特别是在在多个物种中发现几乎相同的读数的情况下，Kraken 会将其分配给最低的共同祖先，这可能是属级别或更高。因此，在包含一个或多个高度相似物种的样品中，物种特异性读数的数量会被低估。为了克服这个问题，使用 Bracken（KrakEN 分类后丰度的贝叶斯重新估计）重新估计了 KrakEN 分类后的[物种丰度](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/population-abundance" \o "Learn more about species abundance from ScienceDirect's AI-generated Topic Pages) [[21]。](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0105)此外，在物种水平上计算口腔和粪便样本的 Shannon 和 Chao1 多样性指数。将样品稀薄至均匀的测序深度（每百万个读数）以进行比较分析。只有分类学分类的读长才用于下游分析。

# 功能基因分析

### 2.4. 生物信息学分析

对所有样品的原始读数进行修剪和质量控制，然后进行分类并映射到 Mengovirus 参考基因组（以确认使用 Kraken 分类进行 Mengovirus QC 实施的有效性）。

### 2.5. 质量控制和修剪

包括 Trimmomatic v0.38 （[Bolger等人，2014](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0020)） 在内的两步修剪策略，以去除接头序列和低质量碱基（设置“ILLUMINACLIP 2：30：10”、“领先：5”、“尾随：5”、“SLIDINGWINDOW：4：10”和“MINLEN：20”选项），然后使用 PRINSEQ v0.20.4 去除低复杂性序列信息（[Schmieder 和 Edwards，2011](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0160)）（设置 '-trim\_tail\_right 10' 和 '-trim\_tail\_left 10' 选项）已实现。仅保留成对读数以供进一步分析。

### 2.6. 分类学分析

使用 Kraken v1.1 对修剪读数进行分类 （[Wood 和 Salzberg， 2014](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0185)）。使用从 RefSeq 基因组（O'Leary等人，2016）（ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/）下载的六个目标分类群（古细菌、细菌、真菌、人类、原生动物和病毒）的所有可用 RefSeq“完整基因组”序列构建了一个定制的 Kraken 数据库2019 年 2 月 18 日。每个订单采样的序列数量概述见补充表 2。根据以下公式，将 Kraken 分类的每个分类水平的读取计数进一步归一化为每百万 （RPM） 总（修剪）读取数，以消除样本之间测序深度变化引入的技术偏差：RPM=Tot一个lnumberofre一个dscl一个ss我f我ed×1m我ll我onTot一个lnumberoftr我mmedre一个ds

值得注意的是，门戈病毒在 NCBI RefSeq 基因组数据库中由[脑心肌炎病毒](https://www.sciencedirect.com/topics/immunology-and-microbiology/encephalomyocarditis-virus" \o "Learn more about Encephalomyocarditis virus from ScienceDirect's AI-generated Topic Pages)（EMCV，最接近的 RefSeq 完整基因组和基于 NCBI 分类数据库的门戈病毒父分类单元，NCBI：txid12104）表示，本研究计算了 Kraken 命中率。出于计算原因，只有[人类基因组](https://www.sciencedirect.com/topics/immunology-and-microbiology/human-genome" \o "Learn more about human genome from ScienceDirect's AI-generated Topic Pages)作为脊椎动物物种包含在数据库中。代表噬菌体物种命中的数据总是在 Order 水平 （Caudovirales） 重新分组。对于重要的 mNGS 发现，使用了 RPM > 1 的任意接受标准。

#### 2.6.1. 使用 Bowtie2 在 Mengovirus 基因组上进行读长定位

由于 NCBI 核苷酸数据库 （L22089.1、DQ294633.1、KX231802.1、KU955338.1） 中存在 Mengovirus 基因组序列的不同候选条目，因此与 Mengovirus 菌株 vMC0 （Mengo Extraction control kit， CeraamTools， bioMérieux） 最相似的序列鉴定如下。在读取映射之前，使用 sdust v0.1-r2 （[https://github.com/lh3/sdust](https://github.com/lh3/sdust" \t "https://www.sciencedirect.com/science/article/pii/_blank)） 鉴定候选基因组的低复杂性区域，并使用 'maskfasta' 函数 （[Quinlan 和 Hall， 2010](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0135)） 用 bedtools v2.27.1 进行掩蔽，以避免产生虚假比对。NEC 样品的读数（分子生物学级水，加标 2.17 × 106Mengovirus 对照的拷贝）随后使用 Bowtie2 v2.3.4.3 （设置“--very-sensitive-local”选项）（[Langmead 和 Salzberg，2012](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0085) 年）与所有掩蔽的基因组进行比对。在 IGV 中目视验证了比对结果（[Robinson et al.， 2011](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0140)），表明参考文献 L22089.1 和 DQ294633.1 在整个基因组中具有完整和一致的覆盖度（补充图 1）。接下来，使用 samtools v1.9 和 bcftools v1.9 （[Li， 2011](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0095);[Li et al.， 2009](https://www.sciencedirect.com/science/article/pii/S0166093420301683?via=ihub" \l "bib0100)）使用选项 '--multiallelic-caller'、'--Variants-only'、'--ploidy 1' 和 '--prior 0.0011'。候选基因组 DQ294633.1 被选为已鉴定变异数量最少的 Mengovirus 参考基因组。因此，对于每个样品，使用 Bowtie2 （设置 '--very-sensitive-local' 选项），修剪的读数始终与 DQ294633.1 对齐。使用 samtools（设置 '-q 15' 选项），仅保留映射质量分数（MAPQ，读取错误比对概率的转换）至少 15 的高质量比对。每个读取对仅计数一次，并标准化为 RPM 以允许与 Kraken 进行比较（其中计数仅基于读取对）。

### 宏基因组读长分类

所有参与者都遵循标准化的规定数据分析程序（文件 S1），以排除使用不同的生物信息学方法引入的潜在变异，使用 Kraken 进行基于 k-mer 的分类学分类，并与提供的 NCBI RefSeq 微生物基因组数据库版本 （[28](https://pmc.ncbi.nlm.nih.gov/articles/PMC11448438/" \l "B28)） 进行分类，该数据库将分类和未分类的 PAstV 列为“物种”。此外，与 Kraken 提供的物种分类相比，使用 Bowtie2 （[29](https://pmc.ncbi.nlm.nih.gov/articles/PMC11448438/" \l "B29)） 与提供给参与者的参考基因组列表进行读取映射，用于不同星状病毒的半定量物种水平定量。对于 Bowtie2 读取比对数据分析，参与者收到了一个多 FASTA 文件，其中包含 8 种猪星状病毒物种和 1 种与 PAstVs 密切相关的单峰骆驼星状病毒的参考基因组（[图 S1](https://pmc.ncbi.nlm.nih.gov/articles/PMC11448438/" \l "SuF3)），其登录号见[表 2](https://pmc.ncbi.nlm.nih.gov/articles/PMC11448438/" \l "T2)，成对序列同一性见[表 S2](https://pmc.ncbi.nlm.nih.gov/articles/PMC11448438/" \l "SuF5)。

### 2.1. 全基因组测序和组装

基因组 DNA 是从 2021 年从卡洛纳岛（南纬 39°15′，东经 146°30′）采集的澳大利亚海豹胎盘拭子样本中提取的。胎盘是在该物种正常幼崽季节的高峰期推测的足月分娩产生的。没有关于母犬或幼崽的临床数据。该标本先前的贝氏柯克斯体 DNA 检测呈阳性，并按照前面所述进行了分子分型 [[14](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B14-pathogens-12-00893)]。

文库构建后，澳大利亚基因组研究设施（[https://www.agrf.org.au/](https://www.agrf.org.au/" \t "https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/_blank)，2022 年 7 月 18 日访问）使用 Novaseq 平台（Illumina Australia and New Zealand，澳大利亚墨尔本）进行基因组测序。将原始测序数据上传到短读存档 （Short Read Archive， PRJNA962036）。

使用 Trimgalore（版本 0.4.5，https://github.com/FelixKrueger/TrimGalore，2022 年 11 月 8 日访问）过滤和修剪 Illumina 150 bp 双端读数，以识别和去除测序接头和低质量碱基片段。Kraken（2.0.8 版）[[19](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B19-pathogens-12-00893)] 和 Bracken [[20]](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B20-pathogens-12-00893) 用于使用 NCBI 基因组和核苷酸数据库（2021 年 4 月 15 日下载）评估胎盘 DNA 提取物中存在的微生物群落的组成。

随后使用 SPAdes v 3.15.0 组装由 Kraken 鉴定的质量过滤的 Coxiella 读数 [[21](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B21-pathogens-12-00893)]。使用 Prokka3 对组装的重叠群进行功能注释（https://github.com/tseemann/prokka，2022 年 11 月 8 日访问）[[22](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B22-pathogens-12-00893)]。使用 quast

（第 3 版）[[23](https://pmc.ncbi.nlm.nih.gov/articles/PMC10386718/" \l "B23-pathogens-12-00893)] 评估组装的完整.

**数据分析**

对于测序数据的生物信息学分析，使用 cutadapt 1.16 进行接头剪辑。使用 fastp （fastp 0.20.0） 使用 75 bp 的长度过滤器进行读取修剪，并将其他参数设置为默认值。使用 bmtagger 1.1.0 对 hg38 参考基因组进行宿主去除。使用 Kraken 2.1.2 针对预制的 Kraken 2 的标准加原生动物和真菌数据库（创建于 2022 年 8 月 9 日）执行读取分配。Kraken 2 的置信度分数阈值设置为 0.1。使用 Bracken 2.7 使用读取数阈值 10 和读取长度参数 75 在属水平进行读取丰度重新估计。

为了评估和比较样本中的相对分类多样性，使用 RStudio （v2022.07.1） 和 ggplot2 库制备了条形图。使用Galaxy网络平台（[https://huttenhower.sph.harvard.edu/galaxy/](https://huttenhower.sph.harvard.edu/galaxy/" \t "https://www.microbiologyresearch.org/content/journal/jmm/10.1099/xrefwindow)）[[23](https://www.microbiologyresearch.org/content/journal/jmm/10.1099/jmm.0.001719" \l "R23)]上的线性判别分析效应大小（LEfSe）工具将样本组相互比较，以计算如何区分组以及哪些分类群（如果有）以差异表示。为了评估可能的样本分组和组差异，将它们与使用 RStudio 和 vegan 和 ggplot2 库的非度量多维缩放 （NMDS） 进行比较，以及通过使用 RStudio 和 vegan 库计算 alpha 多样性 （Shannon） 指数。

# 数据分析

使用FastQC（Version 0.12.1）工具评估原始测序读长的质量，所有样品都通过了常用的质量标准,使用Fastp（Version 0.23.2）,所有参数设置为默认值，进行reads过滤修剪,然后使用 BowTie2（Version 2.3.5.1） 提取未映射到人类基因组（GRch38 版本，从NCBI GenBank下载）的所有读数来去除人类DNA序列。

使用Kraken（Version 2.1.3）软件对成对的非宿主读数进行分类并分配分类标记 [。](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0100)读数 （k-mer） 与预构建的 Kraken\_DB\_standard数据库进行比对，Kraken 为给定宏[基因组](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/metagenomics" \o "Learn more about metagenomic from ScienceDirect's AI-generated Topic Pages)读数集中的每个读数分配最特异性的分类标签进行分类。但是，有时标签分配不在物种级别。特别是在在多个物种中发现几乎相同的读数的情况下，Kraken 会将其分配给最低的共同祖先，这可能是属级别或更高。因此，在包含一个或多个高度相似物种的样品中，物种特异性读数的数量会被低估。为了克服这个问题，使用 Bracken（Version 2.8）（Kraken2分类后丰度的贝叶斯重新估计）重新估计了 Kraken2分类后的[物种丰度](https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/population-abundance" \o "Learn more about species abundance from ScienceDirect's AI-generated Topic Pages)[。](https://www.sciencedirect.com/science/article/pii/S2213716519302024?via=ihub" \l "bib0105)

使用 megahit （Version 1.2.9）组装由fastp质量过滤的读数。使用 Prodigal（Version 2.6.3）对组装的重叠群进行功能预测，使用KEGG数据库进行功能注释。使用 quast（Version 5.0.2）评估基因组组装。

# 数据可视化

为了评估和比较样本中的相对分类多样性，使用 RStudio 和 ggplot2 库作物种组成柱状图，物种丰度变动diff图。使用VennDiagram作物种种类Venn图。

通过使用RStudio和vegan,ggplot2库计算alpha多样性（Shannon，Simpson，Richness）指数。

使用RStudio和vegan,ggplot2库计算beta多样性（非度量多维缩放（NMDS）,PCoA,基于Bray-Curtis相似性的ANOSIM检验）进行比较。

使用RStudio和ggplot2,DESeq2，EnhancedVolcano，pheatmap库进行差异物种分析和差异基因分析，作火山图和热图。并使用microeco库，使用线性判别分析效应大小（LEfSe）将样本组相互比较，以计算如何区分组以及哪些分类群（如果有）以差异表示，作LDA score图，丰度图和系统发育图。使用clusterProfiler库进行差异功能基因富集分析，作富集通路柱状图。

使用RStudio和psyc库计算差异物种，差异代谢物之间的相关性、p值等信息，使用pheatmap库作相关性热图，使用igraph库作网络图并使用Gephi软件进行修改。