bio-101 www.bio-protocol.org/exxxx DOI:10.21769/BioProtoc.xxxx

随机宏基因组测序数据质量控制和去宿主的分析流程和常见问题

Analysis pipeline and frequently asked questions of quality control and host removal in shotgun metagenomic sequencing

刘永鑫 [1, 2, 3, #, *]，刘芳 [1, 2, 3, #]，陈同 [4]，白洋 [1, 2, 3, 5, *]

[1] 中国科学院遗传与发育生物学研究所，植物基因组学国家重点实验室，北京；[2] 中国科学院大学，生物互作卓越创新中心，北京；[3] 中国科学院遗传与发育生物学研究所，中国科学院–英国约翰英纳斯中心植物和微生物科学联合研究中心，北京；[4] 中国中医科学院，中药资源中心，北京；[5] 中国科学院大学现代农学院，北京

*通讯作者邮箱: yxliu@genetics.ac.cn; ybai@genetics.ac.cn

[#]共同第一作者/同等贡献

摘要：随机宏基因组测序，也称鸟枪法宏基因组测序，是指对环境样品的总 DNA 进行高通量测序以获得微生物群落的物种组成及其潜在功能，抑或通过序列拼接和分箱得到其微生物的基因组。宏基因组测序数据预处理包括两方面：一方面，与转录组、基因组测序等分析相似的数据质量控制过程，包括质量评估，去除低质量、引物和接头序列；另一方面，涉及到宿主相关微生物的宏基因组样本易受宿主序列的污染，需要去除宿主序列并评估宿主比例，以获得高质量的微生物组相关数据以方便开展下游分析。本文主要介绍 FastQC、MultiQC、KneadData（涵盖并调用 Trimmomatic + Bowtie 2）等软件组合分析流程的安装、使用方法和结果解读，实现数据质量评估、质量控制和去宿主污染、质量再评估的分析过程，同时对各步骤常见问题和解决方法进行总结，方便同行更准确、高效地实现宏基因组数据的预处理，为下游分析提供高质量的宏基因组数据。

关键词：宏基因组测序，质量控制，去宿主，FastQC，KneadData

仪器设备

1. 计算服务器（操作系统：Linux 主流发行版本，如 CentOS 7+ / Ubuntu 16.04+；CPU：8 核+；内存：32G+；硬盘：> 30 GB，且大于原始数据大小 3 倍），网络访问畅通。
2. 个人电脑（Windows 用户需安装 XShell 或 Putty 等终端类软件，Mac 使用系统内置终端）即可远程访问计算服务器。

软件和数据库

1. 远程文件传输工具 FileZilla 客户端 3.49.1+：https://filezilla-project.org/
2. （可选）Windows 远程访问服务器终端工具 Xshell 6.0.0197p+：https://www.netsarang.com/zh/free-for-home-school/
3. 软件管理器 Miniconda2 Linux 64-bit（Python 2.7）：https://conda.io/miniconda.html
4. 测序数据质量评估 FastQC v0.11.9：https://www.bioinformatics.babraham.ac.uk/projects/download.html
5. 质量评估报告汇总 MultiQC version 1.6 （Ewels 等，2016）：https://multiqc.info/
6. 宏基因组质量控制和去宿主分析流程 KneadData v0.7.4: http://huttenhower.sph.harvard.edu/kneaddata
7. （可选）并行任务队列管理 Parallel 20200522 （Tange，2020）：https://www.gnu.org/software/parallel/
8. 常用宿主基因组下载 Ensembl Genome：http://ensemblgenomes.org/，如人类基因组（International Human Genome Sequencing，2001），拟南芥基因组（The Arabidopsis Genome，2000）。
9. 流程参考代码详见：https://github.com/YongxinLiu/MicrobiomeProtocol/blob/master/e1.KneadData/QualityControl_HostRemoval_Pipelie.sh

软件安装和数据库部署

Windows/Mac 用户安装 FileZilla 客户端，用于上传测序数据至服务器或数据中心，也可下载分析结果本地查看。Windows 用户安装 Xshell 用于远程访问服务器并开展分析，Mac 用户可使用系统自带 Terminal 中的 ssh 命令远程访问服务器。

在 Linux 系统的计算服务器端，以 Miniconda2 软件和 Python2 虚拟环境安装所需软件，在将来随着软件的更新可能需要新建 Python3 虚拟环境才能安装新版本；然后下载人类基因组索引，同时以拟南芥为例介绍下载基因组并建立索引的步骤。

注：代码行添加灰色底纹背景，其中需要根据系统环境修改的部分标为蓝色。

1. 安装 Miniconda2 Linux 64-bit（Python 2.7），已经安装 Conda 可跳过此步骤。

wget -c https://repo.continuum.io/miniconda/Miniconda2-latest-Linux-x86_64.sh

bash Miniconda2-latest-Linux-x86_64.sh

2. 配置 Conda 环境，添加 Bioconda 生物频道以方便安装生物学相关的分析软件。

conda config --add channels bioconda

conda config --add channels conda-forge

3. Conda 新建 Python 2.7 环境，命名为 qc2（quality control python2），然后进入。

conda create -n qc2 python=2.7

conda activate qc2

注：新建虚拟环境，然后在新建的环境下安装工作流程，可以防止新装的软件或者其依赖软件与系统默认环境中的版本相互冲突。另外，将整个分析流程的软件存放在虚拟环境并放置在指定目录下，不用时可以轻松移除，不会对系统产生任何影响。

4. Conda 安装相关软件，-y 默认同意直接安装，不再提示是否确认。

conda install fastqc -y

conda install multiqc -y

conda install kneaddata -y

conda install parallel -y

*注：如果软件下载慢或无法下载，详见**常见问题 1**。Conda 默认安装 Bioconda 中的最新版本或所处系统环境支持的最新版本；如果无法安装或安装后使用存在问题，可使用 conda remove xxx 移除某软件，再指定版本安装，如指定安装 KneadData 的 0.6.1 版本：conda install kneaddata=0.6.1。*

5. 宿主基因组数据库下载。

为了方便指定接下来的文件路径，我们首先使用 mkdir 命令为整个分析流程建立一个文件夹，并命名为 meta_preprocess（参数-p 允许建立多级文件夹、多个文件夹且不报错）。然后使用 cd 命令进入该文件夹。

mkdir -p meta_preprocess

cd meta_preprocess

为了去除宿主序列，我们需要建立宿主序列的索引以供 KneadData 通过序列比对找到并去除宿主序列。KneadData 提供了多个预先建立的常用的宿主序列索引。下面的命令可供我们查看 KneadData 软件整理好的可用的数据库索引，包括人类基因组、小鼠基因组、人类转录组和核糖体数据库等。

```
kneaddata_database
```

以人类基因组为例，下载 Bowtie 2 格式索引，此类索引文件通过包含多个文件，推荐建立文件夹并指定下载位置。

```
mkdir -p db
kneaddata_database --download human_genome bowtie2 db/
```

如果默认数据库下载速度慢或无法下载，可使用国内备份链接，详见**常见问题 2**。KneadData 包括的数据库种类有限，用户可自行下载参考基因组并建索引，以拟芥为例的实例详见**常见问题 3**。

6. 准备输入数据

通常测序公司会返回原始（raw）或纯净（clean）数据两类数据：原始数据为下机后按测序文库的索引（Index）拆分获得的样本序列，纯净数据是去除了明显的低质量、测序引物和接头污染序列后的结果。推荐大家使用体积更小、质量更高的纯净序列进行下游分析和提交数据中心。此外，涉及人类研究的数据，需要上传去除人类相关序列后再上传数据中心（即本文的输出结果）。

本文使用的数据来自人类口腔癌症研究的文章（Schmidt 等，2014），NCBI 的 SRA 项目号为 PRJEB4953。为方便演示流程的使用，我们从中选取 4 个样本，并且随机抽取了 75000 对序列作为软件的测序数据，可以从中国科学院基因组研究所的原始数据归档库（Genome Sequence Archive，GSA，https://bigd.big.ac.cn/gsa/）（Wang 等，2017）中按批次编号 CRA002355 搜索并下载，也可通过 wget 并结合 for 循环通过批次和样本编号批量下载至 seq 目录（代码如下）。

```
mkdir -p seq
```

使用 wget 下载单个样本，-c 为支持断点续传，-O 指定保存位置并可重命名，每个双端样本需要下载两个文件。

```
wget -c ftp://download.big.ac.cn/gsa/CRA002355/CRR117732/CRR117732_
f1.fq.gz \ -O seq/C2_1.fq.gz
wget -c ftp://download.big.ac.cn/gsa/CRA002355/CRR117732/CRR117732_
r2.fq.gz \ -O seq/C2_2.fq.gz
```

结合 for 循环再下载 3 个样本，seq 命令产生连续序列，$i 替换命令中可变部分，结尾加\保证变量名结束而被识别。

```
for i in `seq 3 5`;do
  wget -c
ftp://download.big.ac.cn/gsa/CRA002355/CRR11773$i/CRR11773$i\_f1.fq.gz \
    -O seq/C$i\_1.fq.gz
  wget -c
ftp://download.big.ac.cn/gsa/CRA002355/CRR11773$i/CRR11773$i\_r2.fq.gz \
    -O seq/C$i\_2.fq.gz
done
```

视频 1. 宏基因组测序数据分析流程演示视频和讲解

（https://v.qq.com/x/page/a3128efr2t3.html）

实验步骤

开始分析前，我们应处于项目所在目录（如 meta_preprocess），并启动软件所在的 Conda 环境。

```
cd meta_preprocess
conda activate qc2
```

1. FastQC 测序数据质量评估。

   ```
   fastqc seq/*.fq.gz -t 3
   ```

   *.fq.gz 代表所有以.fq.gz 结尾的文件，即所有测序数据；-t 3 指定 3 个线程，即同时对 3 个文件进行并行分析。

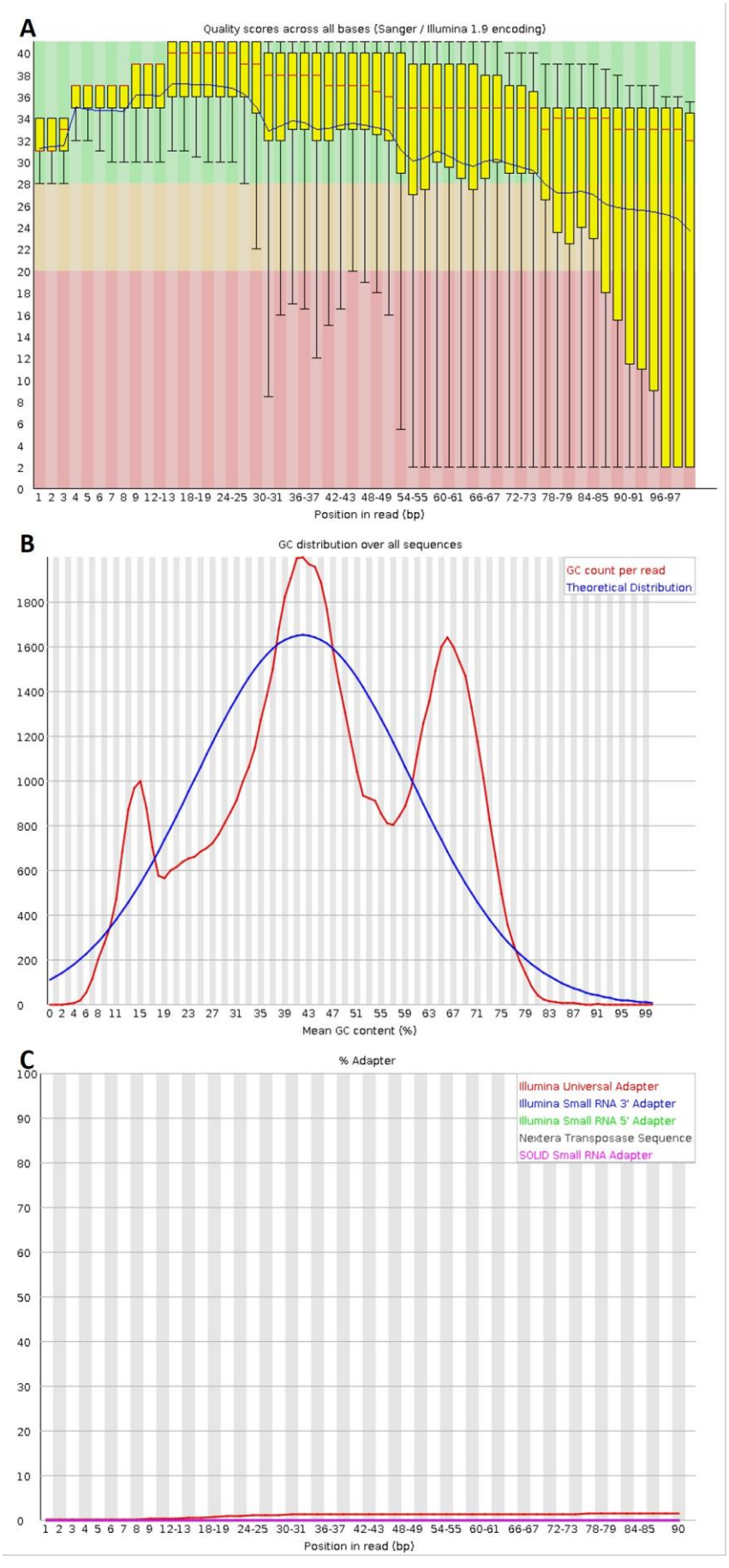
A
Quality scores across all bases (Sanger / Illumina 1.9 encoding)
Position in read (bp)
B
GC distribution over all sequences
GC count per read
Theoretical Distribution
Mean GC content (%)
C
% Adapter
Illumina Universal Adapter
Illumina Small RNA 3' Adapter
Illumina Small RNA 5' Adapter
Nextera Transposase Sequence
SOLID Small RNA Adapter
Position in read (bp)

图 1. FastQC 质量评估报告中的主要结果和注意事项。A. 序列中每个碱基的质量分布（Per base sequence quality）。B. 所有序列的 GC 含量（Per sequence GC content）分布（红色）与理论值分布（蓝色）曲线。C. 接头含量（Adapter content）。本图数据为样本 C3 右端序列为列对 fastQC 的评估结果进行说明，完整评估报告详见 seq/C3_2_fastqc.html。

FastQC 质量评估包括基本统计（比如对应样本总序列数，序列长度和 GC 含量等简要总结）、单碱基位点测序质量、GC 含量及接头含量等 10 大类的评估。我们以 C3 样本右端报告为例，首先查看基本统计中的总序列数（Total Sequences）和 GC 含量（%GC）等。其次查看每个碱基位点的质量分数的箱线图（**图 1A**），每个箱体中间的红线代表此位置上所有序列的测序质量的中位数，然后黄色箱体代表 25%-75%百分位数内的质量分布，而两端黑线顶端对应 10%和 90%百分位的质量数，另外连接每个箱体的蓝色线代表的是平均值。根据 Y 轴序列质量，整个图片区域被划分为高（绿色，得分>=28）、中（黄色，<=20 得分<28）、低（红色，得分<20）三个区域。通常 Illumina 测序数据质量从左往右逐渐降低，从**图 1A** 可以看到序列结尾的箱体进入红色区域，即序列末端存在大量低质量区，这是我们要质量控制中重点关注并需要去除的部分，待质量控制后再次查看此区域。其次查看所有序列的 GC 含量（Per sequence GC content）分布，经常会出现实际值与理论值存在明显差异无法通过评估（**图 1B**），因为理论值是基于单物种的估计结果，而宏基因组测序对象是多物种的混合物，出现分布明显偏移或多峰属于正常现象。过多的序列（Overrepresented sequences）处有时可以查看到污染的引物、接头序列（**常见问题 4**），或样本中特别丰富的序列。接头含量（Adapter Content）评估通用接头的比例，**图 1C** 显示 C3 样本中存在少量 Illumina 通用接头的污染。

2. MultiQC 对多样本的 FastQC 评估结果进行汇总。

研究中通常包含大量样本，而且单个样本又包括双端测序两个结果报告，分别查看每个报告是非常巨大的工作量，而且在缺少比较的条件下判断结果的优劣是比较困难的。MultiQC 可以将所有结果汇总为单个网页报告，实现了样本间的同屏比较，同时方便筛选异常样本。

multiqc -d seq/ -o ./

-d 指定输入目录，-o 指定输出目录，./代表当前目录。

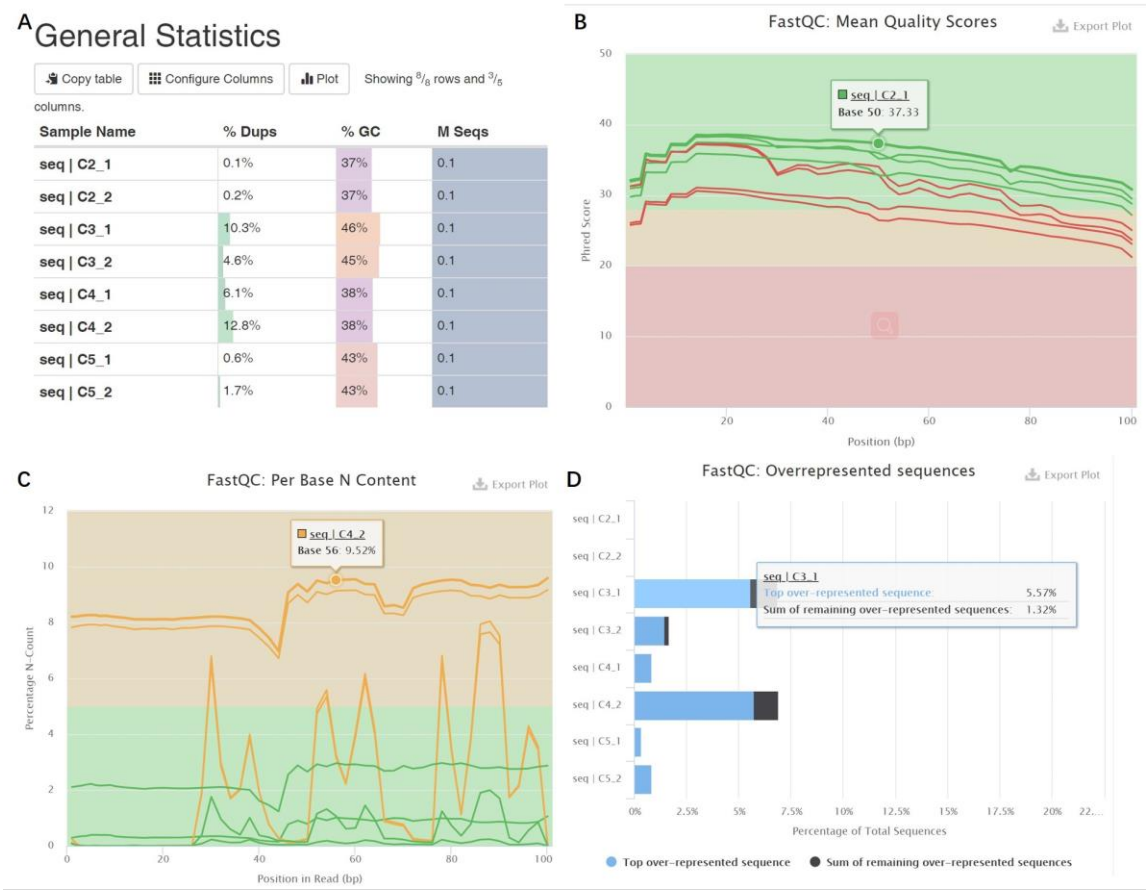

图 2. MultiQC 质量评估汇总报告中的重要结果。A. 综合统计（General Statistics）。B.单位点测序质量的平均值分布（Mean Quality Scores）。C. 单碱基位点 N 含量（Per Base N Content）。D. 过多序列的比例（Overrepresented sequences）。本报告汇总了样本 C2-5 共 4 个样本包含的 8 个序列评估报告的汇总，详见 multiqc_report.html。

我们对多样本质量评估汇总报告（multiqc_report.html）进行观察，发现样本 C3/C4 中有较高的重复序列（**图 2A**），可能原因是测序质量低、测序引物和接头序列污染、样本 DNA 含量低采用较多 PCR 循环扩增等原因。还发现 C3/C5 的 GC 含量明显更高（**图 2A**），可能存在微生物群落组成的差异。我们还可以通过移动鼠标交互地探索每个样本在每个碱基位置上的质量平均值（**图 2B**）。此外关注碱基中 N 的含量（**图 2C**），并记录存在较高 N 含量的样本。如果在下游分析中这些样本也异常时，

可以考虑制定质量筛选标准过滤部分低质量样本，以减少由于实验或测序过程引用的错误。最后重点关注过多序列的比例（**图 2D**），可能是测序引物和接头污染，也可能是微量 DNA 的 PCR 扩增导致，具体原因需要进一步查看过多序列含量其对应样本的 FastQC 报告，结合其对过多序列的详细信息进一步核实是否被标记为测序引物和接头，另外，未知序列也可在线 BLAST（https://blast.ncbi.nlm.nih.gov/Blast.cgi）分析来源（Altschul 等，1997）。

3. 检查测序双端序列标签是否对应且唯一。

zcat 查看样本压缩格式内容，head 显示文件前 10 行，注意观察标签是否重复(**图 3A**)。

```
zcat seq/C2_1.fq.gz|head
zcat seq/C2_2.fq.gz|head
```

双端序列标签对应且唯一是分析中保证准确识别每条序列的前提，通常测序下机数据符合序列名唯一的格式要求(**图 3B**)。但 NCBI 发布的数据为节约存储空间简化序列标签(**图 3A**)，下载的数据会出现双端序列标签完全相同而无法区分正反序列的问题。为保证下游分析的正常，需要修改双端序列标签使之对应且唯一(**图 3B**)。代码详见**常见问题 4**。

A. NCBI SRA sequence IDs are duplicated

```
(qc2) [yongxin@meta:~/KneadData]$zcat seq/C2_1.fq.gz|head
@SRR3586062.883556
CTTGGGGCTGCTGAGCTTCATGCTCCCCTCCTGCCTCAAGGACAATAAGGAGATCTTCG
+
CCCFFFFFHHHHHIJJJJJJIJIJJJJGIJDGIJEIIJIJJJJJJJJIJJJJIJJIJJ
@SRR3586062.3376311
GACGGTGTCCTCAGGACCCTTCAGTGCCTTCATGATCTGCTCAGAGGTGATGGAGTCAC
+
@@@DDDDAFF?DF;EH+ACHIIICHDEHGIGBFE@GCGDGG?D?G@BGHG@FHCGC;CC
@SRR3586062.3139527
CCTCTATTTCCAACTTGGTCATTTCTCTCTACATATGAAAATGAATTTACACATAAAAT
(qc2) [yongxin@meta:~/KneadData]$zcat seq/C2_2.fq.gz|head
@SRR3586062.883556
AAGCCAGGACCCGAGTAGTCTCGGCCAGCATAGTACTCCATACGCCAGCCCTTGGGCGC
+
CCCFFFFFHHHHHJJHIJEHIJJHIIIIIIIJJJHIJGIJIGIJJJJGGCHFEGCHGHGF
@SRR3586062.3376311
TGGCGTGCAGTTCAATACCCTCCTCGACACCGTCGTCGTCGTGCGTGTCATCGAGGGTG
+
=8+=?=)0<+CDDEAEEA<C>FD>F<EEDADDBD)@DA7BADDD8CDA(7.;@A@@>=;
@SRR3586062.3139527
ATTAATTTATTTTTATATTTACTTAGAATCTTTGGAAGAAAAAAATAATTACTTAGGAG
```

B. Modified NCBI SRA sequence IDs are unique

```
(qc2) [yongxin@meta:~/KneadData]$head seq/C2_1.fq
@SRR3586062.883556\1
CTTGGGGCTGCTGAGCTTCATGCTCCCCTCCTGCCTCAAGGACAATAAGGAGATCTTCG
+
CCCFFFFFHHHHHIJJJJJJIJIJJJJGIJDGIJEIIJIJJJJJJJJIJJJJIJJIJJ
@SRR3586062.3376311\1
GACGGTGTCCTCAGGACCCTTCAGTGCCTTCATGATCTGCTCAGAGGTGATGGAGTCAC
+
@@@DDDDAFF?DF;EH+ACHIIICHDEHGIGBFE@GCGDGG?D?G@BGHG@FHCGC;CC
@SRR3586062.3139527\1
CCTCTATTTCCAACTTGGTCATTTCTCTCTACATATGAAAATGAATTTACACATAAAAT
(qc2) [yongxin@meta:~/KneadData]$head seq/C2_2.fq
@SRR3586062.883556\2
AAGCCAGGACCCGAGTAGTCTCGGCCAGCATAGTACTCCATACGCCAGCCCTTGGGCGC
+
CCCFFFFFHHHHHJJHIJEHIJJHIIIIIIIJJJHIJGIJIGIJJJJGGCHFEGCHGHGF
@SRR3586062.3376311\2
TGGCGTGCAGTTCAATACCCTCCTCGACACCGTCGTCGTCGTGCGTGTCATCGAGGGTG
+
=8+=?=)0<+CDDEAEEA<C>FD>F<EEDADDBD)@DA7BADDD8CDA(7.;@A@@>=;
@SRR3586062.3139527\2
ATTAATTTATTTTTATATTTACTTAGAATCTTTGGAAGAAAAAAATAATTACTTAGGAG
```

图 3. NCBI SRA 序列标签修改前（重复）后（唯一）对比。A. NCBI SRA 下载双端序列双端标签完全相同。B. 修改后序列双端标签对应且唯一。蓝色代表命令行，其他颜色为 fastq 格式序列内容，其中序列标签标记为红色。

4. KneadData 流程实现数据质量控制和去宿主。

KneadData 流程主要依赖 Trimmomatic （Bolger 等，2014）进行质量控制和去除引物和接头，Bowtie 2 （Langmead and Salzberg，2012）用来比对宿主基因组，然后通过自定义脚本筛选未能比对到宿主的序列作为输出结果用于下游分析。软件的详细信息，运行 kneaddata -h 查看。序列接头可从测序供应商处获得，基于质量评估结果查找接头序列的方法详见**常见问题 5**，软件运行提示 Java 版本不支持的处理方法详见**常见问题 6**。

单个样本质控和去宿主，可逐个或结合 for 循环处理每个样本。

```
kneaddata -i seq/C2_1.fq.gz -i seq/C2_2.fq.gz \
  -o qc/ -v -t 8 --remove-intermediate-output \
  --trimmomatic ~/.conda/envs/qc2/share/trimmomatic \
  --trimmomatic-options
'ILLUMINACLIP:~/.conda/envs/qc2/share/trimmomatic/adapters/TruSeq3-
PE.fa:2:40:15 SLIDINGWINDOW:4:20 MINLEN:50' \
  --bowtie2-options '--very-sensitive --dovetail' \
  --bowtie2-options="--reorder" \
  -db db/Homo_sapiens
```

使用 parallel 管理队列，允许多个任务并行提高工作效率，详见**软件和数据库 7. 流程参考代码**。

KneadData 流程自带了 kneaddata_read_count_table 流程可完成多样本的质控结果汇总。

```
kneaddata_read_count_table \
--input qc \
--output kneaddata_sum.txt
```

提取原始（raw）、质量控制后（trim）和去宿主后（final）序列数量，详见**表 1**。

```
cut -f 1-5,12-13 kneaddata_sum.txt | sed 's/_1_kneaddata//;s/pair//g' \
  > kneaddata_report.txt
```

表 1. KneadData 流程质量控制和去宿主结果统计。

Sample	raw 1	raw 2	trimmed 1	trimmed 2	final 1	final 2
C2	75000	75000	65316	65316	64876	64876
C3	75000	75000	48082	48082	30897	30897
C4	75000	75000	50387	50387	29343	29343
C5	75000	75000	60959	60959	57379	57379

注：Sample 为样本名，raw 1/2 是双端测序的数据量，trimmed 1/2 是经 Trimmomatic 质量控制后仍成对的序列，final 1/2 是指经过质量控制和去宿主仍成对的序列。注意 1/2 必须一致，否则是程序出错，请检查上一步。

5. 质控后质量再评估。

```
fastqc qc/*_1_kneaddata_paired_*.fastq -t 3
multiqc -d qc/ -o ./
```

使用 fastqc 评估质控后的每对测序数据。然后再次使用 multiqc 进行结果汇总(**图 4**)。结果不仅有序列基本信息统计，还包括质控去除比例(%Dropped)和宿主污染比例(%Aligned)的信息(**图 4A**)。其中质控部分还采用堆叠柱状图展示质控后各部分的百分比(**图 4B**)。去宿主部分用堆叠柱状图展示了序列是否比对宿主基因组的读长数量(**图 4C**)。此外，我们还要重点关注质控后的整体质量分布，以均值位于绿色区间为宜(**图 4D**)。

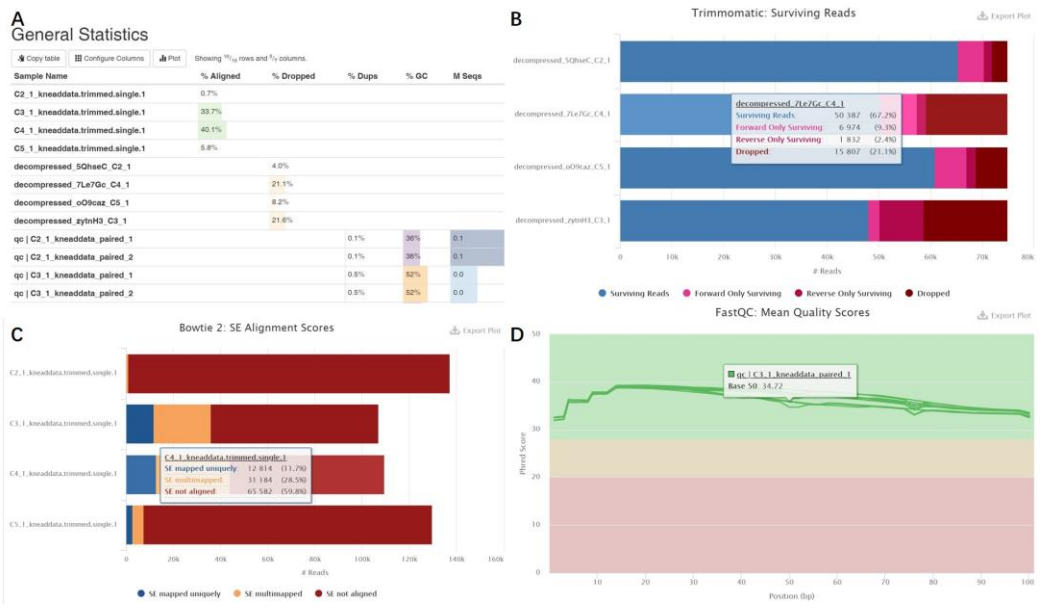

图 4. MultiQC 汇总质量控制、去宿主和最终序列的情况。A. 综合统计（General Statistics），%Aligned 是指比对至宿主基因组的比例，即宿主污染所占比例，%Dropped 为低质量或建库污染的比例。B. Trimmomatic 质量控制结果柱状图，蓝色为质控后结果，粉红为去除的低质量序列，可交互图片移动鼠标至目标区域可显示细节。C. 比对宿主后各部分序列的比例。蓝色为比对至宿主基因组且有唯一位置，橙色为比对至宿主中有多个位置，红色为非宿主序列。D. 质控后序列质量，一般全部在高质量区（绿色）。详见 multiqc_report_1.html。

常见问题

1. 软件下载慢或无法下载。

   大部分软件可通用 Conda（类似于 360 软件管家或腾讯软件管理）快速安装，有时会出现无法下载的问题，请检查网络是否正常，或换个时间再试。对于下载速度较慢的情况，也可以添加 Conda 国内镜像站点加速下载，如清华大学、中国科技大学镜像站等，以添加清华 Conda 镜像站为例：

   ```
   site=https://mirrors.tuna.tsinghua.edu.cn/anaconda
   conda config --add channels $site/pkgs/free/
   conda config --add channels $site/pkgs/main/
   conda config --add channels $site/cloud/conda-forge/
   conda config --add channels $site/pkgs/r/
   conda config --add channels $site/cloud/bioconda/
   ```

2. 数据库下载慢或无法下载。

   很多国外数据库下载缓慢，甚至托管于 Google 或 Dropbox 等国内无法访问的站点。宏基因组公众号团队建立了本领域常用数据库下载的国内备份链接和百度云链接，方便国内同行下载和使用，详见：https://github.com/YongxinLiu/MicrobiomeStatPlot/blob/master/Data/BigDataDownlaodList.md。

3. 物种参考基因组下载和建索引，以拟南芥为例。

   下载目标物种的参考基因组序列，如在 Ensembl Genomes 中按分类查找目标物种的基因组下载链接，使用 wget 下载。

   ```
   wget -c ftp://ftp.ensemblgenomes.org/pub/plants/release-
   ```

47/fasta/arabidopsis_thaliana/dna/Arabidopsis_thaliana.TAIR10.dna.toplevel.fa.gz \
-O ath.fa.gz

-c 实现断点续传，-O 实现文件重命名，“\”用于代码换行。

然后使用 bowtie2-build 建立索引，输入文件可以是 gz 压缩格式的 fasta 文件，并指定输出索引文件前缀。

```
bowtie2-build ath.fa.gz ath.bt2
```

4. 检查测序双端序列标签是否唯一

质控后双端序列数量不同，或双端文件标签不对应（**视频 1**），可能是输入序列标签不唯一，需要检查测序双端序列标签是否唯一。

```
zcat seq/C2_1.fq.gz|head
zcat seq/C2_2.fq.gz|head
```

如果标签重名，需要进行数据解压、对序列的左、右端标题行分别添\1、\2。

```
gunzip seq/*.gz
sed -i '1~4 s/$/\\1/g' seq/*_1.fq
sed -i '1~4 s/$/\\2/g' seq/*_2.fq
```

再次核对样本是否标签有重复。

```
head seq/C2_1.fq
head seq/C2_2.fq
```

结果压缩节省空间，同时与原始序列保持文件名一致。

```
gzip seq/*.fq
```

5. 根据质量评估报告确定接头序列

在 MultiQC 的汇总报告中记录每个过多序列较多的样本，如 C3/4/5，然后并别查看每个样本对应的 FastQC 报告中过多序列部分的序列，并复制部分注释为接头的序列，在 trimmomatic 的接头文件库中搜索。

使用 type 命令确定 trimmomatic 软件位置

```
type trimmomatic
```

根据上面显示的环境路径+share/trimmomatic/adapters 目录匹配接头序列的文件，本例为 C3 样本的右端 FastQC 评估报告中过多的序列栏目可查看到接头序列。

```
grep 'ATCGGAAGAGCACACGTCTGAAC'
```

~/.conda/envs/qc2/share/trimmomatic/adapters/*

6. KneadData 运行提示 Java 版本不支持

   尝试使用 conda 安装指定版本的 Java 开发环境即可。

   conda install openjdk=8.0.152

致谢

本项目由中国科学院战略先导专项(编号：XDA24020104)、中国科学院前沿科学重点研究项目(编号：QYZDB-SSW-SMC021)、国家自然科学基金项目(编号：31772400, 31761143017, 31801945, 31701997)和中国科学院青年创新促进会(编号：2020101)[Supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Precision Seed Design and Breeding, No. XDA24020104), the Key Research Program of Frontier Sciences of the Chinese Academy of Science (No. QYZDB-SSW-SMC021), the National Natural Science Foundation of China (No. 31772400, 31761143017, 31801945, 31701997), the Chinese Academy of Sciences Youth Innovation Promotion Association (No. 2020101)]支持。此分析流程在最近发表的综述中被提及（刘永鑫 等，2019; Liu 等，2020）。感谢西北农林科技大学席娇对本文的修改。

参考文献

1. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25 (17): 3389-3402.
2. Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30 (15): 2114-2120.
3. Ewels, P., Magnusson, M., Lundin, S. and Käller, M. (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32 (19): 3047-3048.
4. International Human Genome Sequencing, C. (2001). Initial sequencing and analysis of the human genome. *Nature* 409 (6822): 860-921.
5. Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9 (4): 357-359.
6. Liu, Y.-X., Qin, Y., Chen, T., Lu, M., Qian, X., Guo, X. and Bai, Y. (2020). A practical guide to amplicon and metagenomic analysis of microbiome data. *Protein Cell* 11.
7. Schmidt, B. L., Kuczynski, J., Bhattacharya, A., Huey, B., Corby, P. M., Queiroz, E. L. S., Nightingale, K., Kerr, A. R., DeLacure, M. D., Veeramachaneni, R., Olshen, A. B.,

Albertson, D. G. and Muy-Teck, T. (2014). Changes in abundance of oral microbiota associated with oral cancer. *PLoS One* 9 (6): e98741.

8. Tange, O. (2020). GNU Parallel 20200522 ('Kraftwerk'). Zenodo.
9. The Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408 (6814): 796-815.
10. Wang, Y., Song, F., Zhu, J., Zhang, S., Yang, Y., Chen, T., Tang, B., Dong, L., Ding, N., Zhang, Q., Bai, Z., Dong, X., Chen, H., Sun, M., Zhai, S., Sun, Y., Yu, L., Lan, L., Xiao, J., Fang, X., Lei, H., Zhang, Z. and Zhao, W. (2017). GSA: Genome Sequence Archive*. *Genom Proteom Bioinf* 15 (1): 14-18.
11. 刘永鑫, 秦媛, 郭晓璇 和白洋 (2019). 微生物组数据分析方法与应用. *遗传* 41 (9): 845-826.

请通过以下链接下载视频：

视频 1：

https://os.bio-protocol.org/doc/upprotocol/p3347/Abstract3347_20200803025729579/kneaddata%20pipeline.wmv